UNIVERSITY OF CALIFORNIA
RIVERSIDE

New Methods for Solving Maximum Likelihood Estimating Equations of Logistic and Probit Regression Models

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Haoyu Wang

August 2011

Dissertation Committee:

Dr. Subir Ghosh, Chairperson
Dr. Daniel R. Jeske
Dr. Robert A. Hanneman

The Dissertation of Haoyu Wang is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

This dissertation would not have been possible without the support and help of many people. Firstly, I owe my deepest gratitude to my dissertation advisor, Professor Subir Ghosh, who is such a nice person. During the research process, he was abundantly helpful and offered me invaluable assistance, encouragement and guidance even when he was overwhelmed by his own work. Whenever I encountered difficult research problems, he was always there to encourage me, to support me and to guide me with great patience. Thank you very much. He is not only an excellent mentor in my research but also in my life. He gave me numerous treasurable advice in my life, helped me overcome difficulties and guided me to establish meaningful life goals. I really appreciate it. Working with you is one of the best memories I have ever had. I will definitely remember it forever.

I would like to thank my committee members, Professor Daniel R. Jeske and Professor Robert A. Hanneman. They provided invaluable suggestions to my research and contributed a lot. I am grateful to all the faculty members in Department of Statistics: Professor Barry C. Arnold, Professor Xinping Cui, Professor James Flegal, Professor Jun Li, Professor Keh-Shin Lii and Doctor Linda M. Penas. They helped me learn more deeply, build stronger interest and research more critically when I took courses with them. Thank all of you very much. I would also like to show my gratitude to Jill E. Smith, Analisa Flores, Barbara M. Beaver and Professor D.V.Gokhale. They offered me great experience and techniques in how to become a successful teacher. Thank you. In addition, I would like to take this opportunity to thank Perla Fabelo and Paula Lemire. You provided me a lot of assistances in enrolling courses, preparing necessary paperwork and most importantly, adapting to the new culture.

Lastly, I wish to express special thanks to my girlfriend, Yao Yu, my parents, my family members and my friends. Words alone cannot express what I owe them for their encouragement, support and trust. Every time I felt tired, they always stood by my side providing a safe harbor for me to rest. It is their patient love that enabled me to finish this dissertation. I love you all.

ABSTRACT OF THE DISSERTATION

New Methods for Solving Maximum Likelihood Estimating Equations of Logistic and Probit Regression Models

by

Haoyu Wang

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, August 2011
Dr. Subir Ghosh, Chairperson

Several iterative methods are available in literature for solving the Maximum Likelihood Estimating Equations (MLEEs) of logistic and probit regression models. Generalized Self Consistency (GSC) method is such an existing iterative method. We introduce a new idea using the paired observations and combine it with the GSC method for both logistic and probit regression models and propose several new methods for solving MLEEs. For probit regression model, we introduce a linear approximation method for finding the exact solution of MLEEs. We illustrate the proposed methods with a real data as well as a simulated data and compare their performances with the existing methods. We investigate some theoretical properties of our estimates. We also present a meaningful method of choosing the initial values of parameters for the iterative methods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Maximum Likelihood Estimate (MLE) is a popular method of estimation. The MLEs of unknown parameters are obtained from solving the Maximum Likelihood Estimating Equations (MLEEs). We often do not have an exact solution of MLEEs. There are several numerical methods available for finding the solution. The purpose of this dissertation is to propose new methods for solving MLEEs and compare them with the existing methods.

Generalized Self Consistency method (GSC) is an existing iterative method for solving the MLEEs in multinomial model. Tsodikov and Chefo (2008) introduced the GSC method based on the Quasi-EM algorithm idea in Tsodikov (2003). We apply the GSC method to logistic and probit regression models. For a real binomial data from Cornfield (1962), we present the estimating equations, describe the iterative methods for solving them and explain the results. We then introduce a new idea using the paired observations and combine it with the GSC method. On this basis we propose several new methods using the paired observations and apply them to the logistic and probit regres-

sion models. We analyze the Cornfield data as well as a simulated data and compare them with the existing methods. We also present a meaningful method of choosing the initial values of parameters for the iterative methods. We find that the efficiencies of the iterative methods would be substantially improved by the initial values provided by the method. Moreover, we investigate some theoretical properties of these methods. We demonstrate that the proposed methods perform as good as the existing methods with respect to all criteria considered, and even better with respect to some of them.

For the probit regression model, we do not have any closed form solution of MLEEs. We propose a new method by introducing a linear approximation which provides an exact solution of the MLEEs. We compare our estimates with the existing numerical method estimates through the Cornfield data as well as a simulated data. We find that our method performs as good as the existing methods with respect to all criteria considered, and even better with respect to some of them. The strength of our proposed method is that we have the closed form solution of MLEEs while the existing methods do not have. In addition, we establish some theoretical properties of our estimates.

In Chapter 2, we present the GSC method in general. In Chapter 3, we briefly recall the logistic regression model. In Chapter 4, we apply the GSC method to logistic regression model. In Chapter 5, we introduce a new method using the paired observations, PGSC method. In Chapter 6, we present some properties of the GSC method and PGSC method in logistic regression model. In Chapter 7, we introduce new methods based on PGSC method. In Chapter 8, we recall the probit regression model and propose a linear approximation method. In Chapter 9, we present some properties of the GSC method and PGSC method in probit regression model.

Chapter 2

Generalized Self Consistency (GSC) Method

2.1 Introduction

Generalized Self Consistency method is an iterative method for calculating the parameter estimates in multinomial model. Tsodikov and Chefo (2008) introduced the GSC method based on the Quasi-EM algorithm idea in Tsodikov (2003). In this chapter, we review the GSC method in general.

2.2 Artificial Variable

We study the dependence of a response variable y on some covariates represented in the vector $\boldsymbol{z}$. The $p(y, \boldsymbol{z})$ describes the distribution of y with $\boldsymbol{z}$. We now introduce an artificial non-observable random variable U into the distribution $p(y, \boldsymbol{z})$ by

$$p(y, \boldsymbol{z}) = E_{U|y,\boldsymbol{z}} \left[p(y, \boldsymbol{z}, U) \right]. \tag{2.1}$$

For the model in (2.1), Tsodikov and Chefo (2008) provided an EM-like algorithm with E-step finding the imputation of U and M-step maximizing the log-likelihood from $p(y, \boldsymbol{z}, U)$.

2.3 Artificial Variable in Multinomial Model

We consider the data $\{y_{ij}, \boldsymbol{z}_j\}, i = 1, \ldots, J; j = 1, \ldots, N$ given below in Table 2.1.

Table 2.1: Multinomial Data

		Covariate Value				
		1	$\cdots$	j	$\cdots$	N
		$\boldsymbol{z}_1$	$\cdots$	$\boldsymbol{z}_j$	$\cdots$	$\boldsymbol{z}_N$
Response Category	1	y_{11}	$\cdots$	y_{1j}	$\cdots$	y_{1N}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	y_{i1}	$\cdots$	y_{ij}	$\cdots$	y_{iN}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	J	y_{J1}	$\cdots$	y_{Jj}	$\cdots$	y_{JN}

In this data, there are J categories of response variable y and N independent realizations of covariate vector $\boldsymbol{z}$. The discrete random variable y_{ij} is the count of observations in the jth covariate group with covariate value $\boldsymbol{z}_j$ and in response category i. Many data in reality can be described in this form. For example, we consider a situation where we are interested in preferences for J different brands of burger. The covariate vector is $\boldsymbol{z} = (Gender, Age, Height, Weight)$ and we have N covariate groups where the jth group has the same values of covariates $\boldsymbol{z}_j$. Then for the jth covariate group, we will have y_{ij} responses for the ith brand of burger.

People often use multinomial model to describe the data in Table 2.1. Indeed, for

each covariate group represented by the columns of Table 2.1, we have a multinomial experiment. For the jth covariate group, let $p_i(\boldsymbol{z}_j)$ be the probability that the response will be in category i, and denote $y_{\cdot j} = \sum_{i=1}^{J} y_{ij}$. The likelihood for this group j is

$$f_j^* = \frac{y_{\cdot j}!}{y_{1j}! \cdots y_{ij}! \cdots y_{Jj}!} p_1(\boldsymbol{z}_j)^{y_{1j}} \cdots p_i(\boldsymbol{z}_j)^{y_{ij}} \cdots p_J(\boldsymbol{z}_j)^{y_{Jj}}, \tag{2.2}$$

where $0 < p_i(\boldsymbol{z}_j) < 1$ and $\sum_{i=1}^{J} p_i(\boldsymbol{z}_j) = 1$.

We now denote $p_i(\boldsymbol{z}_j)$ as

$$p_i(\boldsymbol{z}_j) = \frac{\theta_i^*(\boldsymbol{z}_j)}{\sum_{k=1}^{J} \theta_k^*(\boldsymbol{z}_j)}, \tag{2.3}$$

where $\theta_i^*(\boldsymbol{z}_j)$ is unknown parameter. By dividing the numerator and denominator of (2.3) by $\theta_1^*(\boldsymbol{z}_j)$, we get

$$p_i(\boldsymbol{z}_j) = \frac{\frac{\theta_i^*(\boldsymbol{z}_j)}{\theta_1^*(\boldsymbol{z}_j)}}{1 + \cdots + \frac{\theta_i^*(\boldsymbol{z}_j)}{\theta_1^*(\boldsymbol{z}_j)} + \cdots + \frac{\theta_J^*(\boldsymbol{z}_j)}{\theta_1^*(\boldsymbol{z}_j)}}.$$

We denote $\theta_i(\boldsymbol{z}_j) = \frac{\theta_i^*(\boldsymbol{z}_j)}{\theta_1^*(\boldsymbol{z}_j)}$, and note that $\theta_1(\boldsymbol{z}_j) = 1$. Thus we have

$$p_1(\boldsymbol{z}_j) = \frac{\theta_1(\boldsymbol{z}_j)}{\sum_{k=1}^{J} \theta_k(\boldsymbol{z}_j)} = \frac{1}{1 + \sum_{k=2}^{J} \theta_k(\boldsymbol{z}_j)}, p_i(\boldsymbol{z}_j) = \frac{\theta_i(\boldsymbol{z}_j)}{1 + \sum_{k=2}^{J} \theta_k(\boldsymbol{z}_j)}, i = 2, \ldots, J. \tag{2.4}$$

Moreover, we model $\theta_i(\boldsymbol{z}_j)$ as a function of a vector of unknown parameters $\boldsymbol{\beta}_i$

$$\theta_i(\boldsymbol{z}_j) = e^{\boldsymbol{\beta}_i' \boldsymbol{z}_j}. \tag{2.5}$$

Our goal is to make inferences on the unknown parameter vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J$. We want to maximize the likelihood under multinomial model to obtain their estimates.

Since the N covariate groups are independent, the likelihood for all the data is

$$L^* = \prod_{j=1}^{N} f_j^*, \tag{2.6}$$

where f_j^* is in (2.2). Note that the factorial terms in the likelihood do not contain unknown parameters, we consider the adjusted likelihood by ignoring these terms

$$L_A = \prod_{j=1}^{N} f_j = \prod_{j=1}^{N} p_1\left(\boldsymbol{z}_j\right)^{y_{1j}} \cdots p_i\left(\boldsymbol{z}_j\right)^{y_{ij}} \cdots p_J\left(\boldsymbol{z}_j\right)^{y_{Jj}}. \tag{2.7}$$

Now we will bring artificial random variables into (2.7) as what we did in (2.1).

For $U \sim \exp(1)$, we have (See Appendix A.1)

$$E_U\left[e^{-U\theta}\right] = \frac{1}{1+\theta}. \tag{2.8}$$

Now we introduce N artificial unobserved random variables $U_1, \cdots, U_j, \cdots, U_N$ that are i.i.d. exp(1) for each j, then (2.4) could be written as

$$p_i(\boldsymbol{z}_j) = E_{U_j}[p_i(\boldsymbol{z}_j, U_j)] = E_{U_j}\left[\theta_i(\boldsymbol{z}_j) \cdot e^{-U_j \sum_{k=2}^{J} \theta_k(\boldsymbol{z}_j)}\right],$$

where we denote

$$p_i(\boldsymbol{z}_j, U_j) = \theta_i(\boldsymbol{z}_j) \cdot e^{-U_j \sum_{k=2}^{J} \theta_k(\boldsymbol{z}_j)}, \tag{2.9}$$

and

$$p_i(\boldsymbol{z}_j) = E_{U_j}\left[p_i(\boldsymbol{z}_j, U_j)\right]. \tag{2.10}$$

Therefore, the adjusted likelihood (2.7) becomes

$$L_A = \prod_{i=1}^{J}\prod_{j=1}^{N} p_i\left(\boldsymbol{z}_j\right)^{y_{ij}} = \prod_{i=1}^{J}\prod_{j=1}^{N} \left\{E_{U_j}\left[p_i\left(\boldsymbol{z}_j, U_j\right)\right]\right\}^{y_{ij}}. \tag{2.11}$$

For our conveninece, we now consider the following likelihood

$$L = \prod_{i=1}^{J}\prod_{j=1}^{N} p_i\left(\boldsymbol{z}_j, U_j\right)^{y_{ij}}. \tag{2.12}$$

Now for simplicity, denote $\theta_i\left(\boldsymbol{z}_j\right) = \theta_{ij}$. Then the log-likelihood (See Appendix A.2) is

$$l = \log L = \sum_{i=2}^{J} l_i = \sum_{i=2}^{J}\sum_{j=1}^{N} \left(y_{ij} \cdot \log \theta_{ij} - y_{\cdot j} U_j \theta_{ij}\right). \tag{2.13}$$

2.4 Quasi-EM Algorithm in Multinomial Model

In (2.13), if we can find an appropriate imputation $\hat{U}_j \to U_j$ as the E-step, then the M-step is just to maximize (2.13) with respect to the unknown parameters $\boldsymbol{\beta}_i$. However, $p_i(\boldsymbol{z}_j, U_j)$'s in (2.9) are not real probabilities because they have to be between 0 and 1 as real probabilities. The actual range of $p_i(\boldsymbol{z}_j, U_j)$ is in fact $(0, +\infty)$. Recall the E-step in EM algorithm must be defined as a conditional expectation of the complete-data log-likelihood given observed data based on a real probabilistic model, and apparently in our case we do not have a real probabilistic model. In this section, we will introduce an EM-like algorithm called Quasi-EM (QEM) proposed by Tsodikov (2003) to deal with this problem.

Generally speaking, QEM algorithm is a generalization of EM algorithm by substituting integral operator "E" by some other operator called "Quasi-Expectation" operator (QE) and the "M" step remains the same. Recall that EM algorithm allows us to obtain the estimates of the unknown parameters $\boldsymbol{\beta}$ by maximizing a log-likelihood of the form

$$l(\boldsymbol{\beta}) = E_{U|\text{observed data}}[l(\boldsymbol{\beta}, U)], \tag{2.14}$$

where the log-likelihood is based on a real probabilistic model and U is an unobserved random variable. Since in our case we do not have a real probabilistic model but we also want to preserve the good properties of EM algorithm, we introduce QE operator such that we would also obtain estimates of the unknown parameters $\boldsymbol{\beta}$ by maximizing

$$l(\boldsymbol{\beta}) = QE_{U|\text{observed data}}[l(\boldsymbol{\beta}, U)], \tag{2.15}$$

where the log-likelihood is not based on a real probabilistic model and U is not necessarily a random variable because we are using $QE_{U|\text{observed data}}$ instead of $E_{U|\text{observed data}}$.

In summary, "E" operator requires a real probabilistic model and U to be a random variable while "QE" operator does not have such requirement. We might consider the Quasi-EM algorithm as "distribution free". The detail of how to apply this idea to multinomial model is discussed below.

Now we apply this Quasi-EM algorithm to the multinomial model. For multinomial model, Tsodikov (2003) and Tsodikov and Chefo (2008) defined the following basis functions

$$e_0(U,s) = e^{-Us}, e_1(U,s) = Ue^{-Us}, e_2(U,s) = U^2e^{-Us}, \tag{2.16}$$

and QE operator was defined as a linear operator on the linear span of $\{e_k\}$ such that

$$QE_U\left[e_0(U,s)\right] = \gamma\left(e^{-s}\right). \tag{2.17}$$

For each j, let

$$\gamma(e^{-s_j}) = \frac{1}{1+s_j}. \tag{2.18}$$

Therefore, from (2.16)–(2.18), we could rewrite (2.4) as

$$p_i(\boldsymbol{z}_j) = \theta_i(\boldsymbol{z}_j)\cdot\gamma(e^{-s_j}) = QE_{U_j}\left[\theta_i(\boldsymbol{z}_j)\cdot e_0(U_j,s_j)\right], \tag{2.19}$$

where $s_j = \sum_{k=2}^{J}\theta_k(\boldsymbol{z}_j)$.

From (2.19), note that $p_i(\boldsymbol{z}_j)$ only contains $e_0(U_j,s_j)$, which makes the adjusted likelihood L_A (representing the observed data) in (2.11) only contains $e_0(U_j,s_j)$, namely,

$$QE_{U_j|\text{observed data}}\left(U_j\right) = QE_{U_j|L_A}(U_j) = QE_{U_j|e_0(U_j,s_j)}\left(U_j\right). \tag{2.20}$$

Tsodikov (2003) and Tsodikov and Chefo (2008) also defined

$$\frac{\partial}{\partial s}QE_U\left\{e_k\left(U,s\right)\right\} = QE_U\left\{\frac{\partial e_k(U,s)}{\partial s}\right\},\ k=0,1,2. \tag{2.21}$$

and, for any function f in the $\{e_k\}$ linear span, conditional QE as

$$QE_{U|f}(U) = \frac{QE_U(Uf)}{QE_U(f)} \tag{2.22}$$

Therefore, from (2.20)–(2.22), we have the imputation

$$\begin{aligned} U_j^* &= QE_{U_j|e_0(U_j,s_j)}\left(U_j\right) \\ &= \frac{QE_{U_j}(U_j \cdot e_0(U_j,s_j))}{QE_{U_j}(e_0(U_j,s_j))} \\ &= \frac{QE_{U_j}(e_1(U_j,s_j))}{QE_{U_j}(e_0(U_j,s_j))} \\ &= \frac{QE_{U_j}\left(-\frac{\partial}{\partial s_j}e_0(U_j,s_j)\right)}{\gamma\left(e^{-s_j}\right)} \\ &= \frac{-\frac{\partial}{\partial s_j}QE_{U_j}(e_0(U_j,s_j))}{\gamma\left(e^{-s_j}\right)} \\ &= \frac{-\frac{\partial}{\partial s_j}\gamma\left(e^{-s_j}\right)}{\gamma\left(e^{-s_j}\right)} \\ &= \frac{-\frac{\partial}{\partial s_j}\frac{1}{1+s_j}}{\frac{1}{1+s_j}} \\ &= \frac{1}{1+s_j}. \end{aligned} \tag{2.23}$$

Then the algorithm for multinomial model should work in the following steps

1. Set initial values of $\boldsymbol{\beta}_i^{(0)}, i = 2, \ldots, J$ and set iteration step $m = 0$.

2. For each j , compute $U_j^{*(m)} = \frac{1}{1+s_j^{(m)}}$ where $s_j^{(m)} = \sum_{k=2}^{J} \theta_{kj}^{(m)}$ and $\theta_{kj}^{(m)} = \exp\left[\boldsymbol{\beta_k}^{(m)\prime}\boldsymbol{z}_j\right]$.

3. Solve $J-1$ separate MLE problems

$$\boldsymbol{\beta}_i^{(m+1)} = \arg\max_{\boldsymbol{\beta}_i} \sum_{j=1}^{N} \left[y_{ij}\log(\theta_{ij}) - y_{\cdot j}U_j^{*(m)}\theta_{ij}\right], i = 2, \ldots, J. \tag{2.24}$$

4. If the convergence criterion is reached, stop; else increase m and return to step 2.

From this procedure, it is clear that we can divide the entire unknown parameter set into some subsets in a clever way, and then solve the subsets respectively.

An important research question is on the best choice for the initial values of $\boldsymbol{\beta}_i^{(0)}$. The performance of the algorithm depends on the initial values. So we are interested in how to find the best choice of initial values, how the algorithm performs and what properties the algorithm has.

Chapter 3

Logistic Regression Model

3.1 Introduction

In this chapter, we present the model setup, assumptions and the estimating equation for logistic regression model. We also present an illustrative example of a real data from Cornfield (1962).

3.2 Model for General N

From now on, we consider a special case of multinomial model with $J = 2$. First of all, we introduce the data. In this data, we delete the first subscript of y_{ij} representing the response category since we only have 2 categories, and we introduce the total number of observations n_j for each group j. Thus we have Table 3.1 similar as Table 2.1.

Table 3.1: Binomial Data

		Covariate Value				
		1	$\cdots$	j	$\cdots$	N
		$\boldsymbol{z}_1$	$\cdots$	$\boldsymbol{z}_j$	$\cdots$	$\boldsymbol{z}_N$
Response Category	1	y_1	$\cdots$	y_j	$\cdots$	y_N
	2	$n_1 - y_1$	$\cdots$	$n_j - y_j$	$\cdots$	$n_N - y_N$

People often use logistic regression model to fit this data. As in multinomial model, for each covariate group, we have a binomial experiment. For the jth group, let $p(\boldsymbol{z}_j)$ and $1-p(\boldsymbol{z}_j)$ be the probabilities that the response variable will fall in response category 1 and 2 respectively. Notice we only have 2 categories, we further simplify the notation and denote $p(\boldsymbol{z}_j) = p_j$. And we have the likelihood for the jth group

$$f_j^* = \frac{n_j!}{y_j! \cdot (n_j - y_j)!} p_j^{y_j} \left(1 - p_j\right)^{n_j - y_j}. \tag{3.1}$$

Again, we define p_j as

$$p_j = \frac{\theta\left(\boldsymbol{z}_j\right)}{1 + \theta\left(\boldsymbol{z}_j\right)}, \tag{3.2}$$

where $\theta\left(\boldsymbol{z}_j\right) = \theta_j$ is modeled as a function of unknown parameters $\boldsymbol{\beta}$ by

$$\theta_j = e^{\boldsymbol{\beta}' \boldsymbol{z}_j}. \tag{3.3}$$

Here we recall the basic assumption of logistic regression model

$$\log \frac{p_j}{1 - p_j} = \boldsymbol{\beta}' \boldsymbol{z}_j. \tag{3.4}$$

In (3.4), we actually have a linear regression model of the logit of p_j. This function is also called the link function. The equation (3.4) is equivalent to (3.2) and (3.3). This unknown parameter vector $\boldsymbol{\beta}$ is again our major interest. We want to maximize the likelihood to obtain the estimates of $\boldsymbol{\beta}$.

Since we know that N covariate groups are independent, the likelihood of the whole data is

$$L^* = \prod_{j=1}^{N} f_j^*, \tag{3.5}$$

where f_j^* is in (3.1).

Note that the factorial terms in the likelihood do not contain unknown parameters, we consider the adjusted likelihood by ignoring these terms

$$L_A = \prod_{j=1}^{N} f_j = \prod_{j=1}^{N} p_j^{y_j} \, (1 - p_j)^{n_j - y_j}. \tag{3.6}$$

We substitute the expression of p_j from (3.6) and obtain

$$L_A = \prod_{j=1}^{N} f_j = \prod_{j=1}^{N} \left(e^{\boldsymbol{\beta}' \boldsymbol{z}_j} \right)^{y_j} \left(\frac{1}{1 + e^{\boldsymbol{\beta}' \boldsymbol{z}_j}} \right)^{n_j}. \tag{3.7}$$

The logarithm of adjusted likelihood is

$$l_A = \sum_{j=1}^{N} \log f_j = \sum_{j=1}^{N} \left[y_j \boldsymbol{\beta}' \boldsymbol{z}_j - n_j \log \left(1 + e^{\boldsymbol{\beta}' \boldsymbol{z}_j} \right) \right]. \tag{3.8}$$

Taking the derivative of (3.8) with respect to $\boldsymbol{\beta}$, we get the estimating equation (3.9)

$$\frac{\partial}{\partial \boldsymbol{\beta}} l_A = \sum_{j=1}^{N} \left[y_j \boldsymbol{z}_j - n_j \boldsymbol{z}_j \frac{e^{\boldsymbol{\beta}' \boldsymbol{z}_j}}{1 + e^{\boldsymbol{\beta}' \boldsymbol{z}_j}} \right] = \mathbf{0}. \tag{3.9}$$

Solving this equation (3.9), we obtain the MLE of the unknown parameter vector $\boldsymbol{\beta}$.

3.3 Example

We now present a real data from Cornfield (1962). We use this data to study the performance and property of all the methods and algorithms proposed in this dissertation.

A sample of male residents of Framingham, Massachusetts, aged 40 through 59, was classified on blood pressure. According to different blood pressure ranges, people were separated to 8 different groups. In each group, the total number of people and the number of people who were observed heart diseases were recorded. We take the middle point of each blood pressure interval to be the value of the covariate z and let the covariate vector be $\boldsymbol{z} = (1, z)'$. For the jth covariate group, we want to model the

Table 3.2: Real Example Data

	Covariate Value							
	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8
	111.5	121.5	131.5	141.5	151.5	161.5	176.5	191.5
Disease	3	17	12	16	12	8	16	8
No Disease	153	235	272	255	127	77	83	35

number of people with disease y_j based on the covariate vector $\boldsymbol{z}_j$. The data is shown in Table 3.2.

For this real example, we use logistic regression model discussed in the previous section to fit the data and obtain estimates of $\boldsymbol{\beta}$. From (3.9), the estimating equation becomes

$$\frac{\partial}{\partial \boldsymbol{\beta}} l_A = \sum_{j=1}^{N} \left[y_j \begin{pmatrix} 1 \\ z_j \end{pmatrix} - n_j \begin{pmatrix} 1 \\ z_j \end{pmatrix} \frac{e^{\beta_0+\beta_1 z_j}}{1+e^{\beta_0+\beta_1 z_j}} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Rightarrow \tag{3.10}$$

$$\begin{cases} \frac{\partial}{\partial \beta_0} l_A = \sum\limits_{j=1}^{N} \left[y_j - n_j \frac{\exp(\beta_0+\beta_1 z_j)}{1+\exp(\beta_0+\beta_1 z_j)} \right] = 0, \\ \frac{\partial}{\partial \beta_1} l_A = \sum\limits_{j=1}^{N} \left[y_j z_j - n_j z_j \frac{\exp(\beta_0+\beta_1 z_j)}{1+\exp(\beta_0+\beta_1 z_j)} \right] = 0. \end{cases}$$

Based on the data in Table 3.2, we solve (3.10) to find the estimates of β_0 and β_1.

3.4 Estimates of Parameters for $N = 2$

The equation (3.10) is a transcendental equation. It does not have a closed form solution and thus needs numerical method for finding the solution. However, if we only consider a pair of observations, i.e., $N = 2$, we have a closed form solution. For example, we consider observations i and j. Denote the estimates of β_0 and β_1 by $\hat{\beta}_0$ and $\hat{\beta}_1$, and the

predicted values of y_i and y_j by $\hat{y}_i$ and $\hat{y}_j$. Then equation (3.10) becomes

$$\begin{cases} \left[y_i - n_i \frac{\exp(\beta_0+\beta_1 z_i)}{1+\exp(\beta_0+\beta_1 z_i)}\right] + \left[y_j - n_j \frac{\exp(\beta_0+\beta_1 z_j)}{1+\exp(\beta_0+\beta_1 z_j)}\right] = 0, \\ \left[y_i z_i - n_i z_i \frac{\exp(\beta_0+\beta_1 z_i)}{1+\exp(\beta_0+\beta_1 z_i)}\right] + \left[y_j z_j - n_j z_j \frac{\exp(\beta_0+\beta_1 z_j)}{1+\exp(\beta_0+\beta_1 z_j)}\right] = 0. \end{cases}$$

$$\Rightarrow$$

$$\begin{cases} \frac{\exp(\beta_0+\beta_1 z_i)}{1+\exp(\beta_0+\beta_1 z_i)} = \frac{y_i}{n_i}, \\ \frac{\exp(\beta_0+\beta_1 z_j)}{1+\exp(\beta_0+\beta_1 z_j)} = \frac{y_j}{n_j}. \end{cases}$$

and we get the exact solution as

$$\begin{cases} \hat{\beta}_0 = \frac{z_i \cdot \log \frac{y_j}{n_j} - z_j \cdot \log \frac{y_i}{n_i} + z_j \cdot \log\left(1-\frac{y_i}{n_i}\right) - z_i \cdot \log\left(1-\frac{y_j}{n_j}\right)}{z_i - z_j}, \\ \hat{\beta}_1 = \frac{\log \frac{y_i}{n_i} - \log \frac{y_j}{n_j} + \log\left(1-\frac{y_j}{n_j}\right) - \log\left(1-\frac{y_i}{n_i}\right)}{z_i - z_j}. \end{cases} \tag{3.11}$$

Consequently, we have

$$\hat{y}_i = y_i, \ \hat{y}_j = y_j. \tag{3.12}$$

Chapter 4

GSC Method for Logistic Regression Model

4.1 Introduction

In this chapter, we apply the GSC method to logistic regression model. We present the estimating equation and describe the iterative method for solving it.

4.2 Method for General N

Now we apply GSC method discussed in Chapter 2 to logistic regression model. Recall the algorithm for multinomial model discussed at the end of Chapter 2. Since logistic regression model is a special case of multinomial model with $J = 2$, the algorithm works in a very similar fashion as below

1. Set initial values of $\boldsymbol{\beta}^{(0)}$ and iteration step $m = 0$.

2. For each j, compute $U_j^{*(m)} = \frac{1}{1+s_j^{(m)}}$ where $s_j^{(m)} = \theta_j^{(m)} = \exp\left[\boldsymbol{\beta}^{(m)\prime}\boldsymbol{z}_j\right]$.

3. Solve an MLE problem

$$\boldsymbol{\beta}^{(m+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{N} \Big[y_j \log(\theta_j) - n_j U_j^{*(m)} \theta_j \Big]. \tag{4.1}$$

4. If the convergence criterion is reached, stop; else increase m and return to step 2.

We now use the Cornfield data introduced in Chapter 3 to study this method. In our example, remember $\theta_j = \exp(\beta_0 + \beta_1 z_j)$ and $\theta_j^{(m)} = \exp\left[\beta_0^{(m)} + \beta_1^{(m)} z_j\right]$, then the MLE problem becomes

$$\boldsymbol{\beta}^{(m+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{N} \left[y_j (\beta_0 + \beta_1 z_j) - n_j \frac{\exp(\beta_0 + \beta_1 z_j)}{1 + \exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)} \right]. \tag{4.2}$$

We take the derivative of objective function in (4.2) with respect to β_0 and β_1, and get

$$\begin{cases} \sum\limits_{j=1}^{N} \left[y_j - n_j \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_j\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)} \right] = 0, \\ \\ \sum\limits_{j=1}^{N} \left[y_j z_j - n_j z_j \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_j\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)} \right] = 0. \end{cases} \tag{4.3}$$

The algorithm works as follows:

1. Set initial values of $\beta_0^{(0)}$, $\beta_1^{(0)}$ and iteration step $m = 0$.

2. Solve an equation

$$\begin{cases} \sum\limits_{j=1}^{N} \left[y_j - n_j \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_j\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)} \right] = 0, \\ \\ \sum\limits_{j=1}^{N} \left[y_j z_j - n_j z_j \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_j\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)} \right] = 0. \end{cases}$$

3. If the convergence criterion is reached, stop; else increase m and return to step 2.

For general N, the regular MLE estimating equation is (3.10). Solving this (3.10) requires numerical method. A popular method to solve (3.10) is the Iterative Reweighted

Least Squares (IRLS) method, and there are several other similar methods. Under the GSC framework, Quasi-EM algorithm provides us another way to obtain the estimates from (4.3). The method using (4.3) is also an iterative numerical method for solving (3.10). Our goal is to compare the standard method with the method in (4.3).

4.3 Method for $N = 2$

We can not directly solve (4.3) for general N since it is a transcendental equation. However, if we only consider a pair, i.e., $N = 2$, then we are able to obtain the exact closed form solution. For example, we consider observations i and j. Directly solving (4.3) provides the exact solution as

$$\begin{cases} \beta_0^{(m+1)} = \dfrac{z_i \cdot \log \frac{y_j}{n_j} - z_j \cdot \log \frac{y_i}{n_i} + z_i \cdot \log\left[1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)\right] - z_j \cdot \log\left[1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_i\right)\right]}{z_i - z_j}, \\ \beta_1^{(m+1)} = \dfrac{\log \frac{y_i}{n_i} - \log \frac{y_j}{n_j} + \log\left[1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_i\right)\right] - \log\left[1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)\right]}{z_i - z_j}. \end{cases} \tag{4.4}$$

In other words, for $N = 2$, (4.3) has an exact solution and does not need any numerical method.

Chapter 5

A New Method: PGSC Method

5.1 Introduction

In this chapter, we propose a new method called PGSC method by applying the GSC method to all paired observations. We apply the new method to the logistic regression model and use it to analyze the Cornfield data. Then we compare it with the GSC method and standard method. Moreover, we present a method for the choice of the initial values of the parameters. We find that the efficiency of the standard method can be substantially improved by selecting better initial values provided by the new method.

5.2 Paired Observations

For a data with N observations, the standard method to obtain MLE is to consider these N observations together as the whole data and apply a specific algorithm. For the Cornfield data, the IRLS method gives us the estimates $\hat{\beta}_0 = -6.082033$ and $\hat{\beta}_1 = 0.02433824$ of β_0 and β_1 for the whole data. Now we take a pair containing observations 1 and 8. Then i, j in (4.4) become $i = 1,\ j = 8$, and (4.4) converges at $m = 12$ and result in

the estimates $\beta_0^{(m)} = -7.354763,\ \beta_1^{(m)} = 0.03069899$. We see that the solution obtained from the chosen pair is close to the solution obtained from (3.10). Our research question becomes: What happens if we use this solution to be the initial value of the GSC method (4.3) to obtain the solution for the whole data?

We know that, for N observations, we have $\frac{N(N-1)}{2} = M$ different pairs of data. The standard method to obtain MLE for the whole data requires numerical approximation because we do not have the exact closed form solution. Since we can obtain the exact solution of estimating equation by only using one pair, we apply the GSC method to each of the M pairs and choose the best pair with respect to a criterion. The estimates of the unknown parameters are based on this best pair. We present the details in next section.

5.3 PGSC Method

We define a criterion for selecting the best pair as follows:

Definition 1 *Denote the predicted value of an observation y_i by $\hat{y}_i$. The criterion we use to measure the goodness of fit is defined as*

$$\Delta = \sum_{i=1}^{N} |y_i - \hat{y}_i|.$$

We want to find an estimate of $\boldsymbol{\beta}$ which makes Δ as small as possible.

We now introduce the new method called PGSC method. For each of M pairs, we apply the GSC method until it converges and finally choose the best pair which gives us the smallest Δ. The method works as follows:

1. Set $u = 1$.

2. For the uth pair, give initial values of $\beta_0^{(0)}$ and $\beta_1^{(0)}$, compute $\beta_0^{(m+1)}$ and $\beta_1^{(m+1)}$ by using equation (4.4) until it converges. Calculate Δ and set $u = u + 1$.

3. Repeat step 2 until $u = M + 1$.

4. Choose the pair giving us the smallest Δ. Use the estimates of β_0 and β_1 obtained from this pair to be the final estimates.

We apply this method to the Cornfield data. We take $(0, 0)$ as the initial value for (4.4). Then we compare solutions from (3.11) and (4.4) for all pairs. In Table 5.1, we present the final estimates of (β_0, β_1) and the number of steps of convergence for all $M = 28$ pairs for both the PGSC method and standard method. In Table 5.1, we see that Δ attains its minimum value 14.17 for the 21st pair containing observations 4 and 7 for both methods. We take $\beta_0 = -7.30647$ and $\beta_1 = 0.032069227$ as the final estimates by the PGSC method.

5.4 GSC Method for General N

We recall that the GSC method for general N uses equations in (4.3) to numerically find the values of $\beta_0^{(m)}$ and $\beta_1^{(m)}$. We now directly solve (4.3) for general N. We use two different sets of initial values, one is (0,0) and the other one is the PGSC method solution described in the previous section, i.e., (-7.306470, 0.032069227). We use $\frac{\left|\Delta^{(m+1)} - \Delta^{(m)}\right|}{\Delta^{(m)}} < 10^{-5}$ as our stopping rule where $\Delta^{(m)}$ is the value of Δ at the mth stage. In Table 5.2, we present the comparison for these two different initial values. In

Table 5.1: Comparison Between PGSC Method and Standard Method with Paired Observations for All Pairs (Cornfield Data)

	PGSC Method			Standard Method	
Pair	$\left(\beta_0^{(m)}, \beta_1^{(m)}\right)$	Δ	m	$\left(\hat{\beta}_0, \hat{\beta}_1\right)$	Δ
(1,2)	(-18.487632, 0.130545346)	451.28	9	(-18.487632, 0.130545346)	451.28
(1,3)	(-8.452762, 0.040546511)	25.83	8	(-8.452762, 0.040546511)	25.83
(1,4)	(-8.254869, 0.038771694)	21.21	8	(-8.254869, 0.038771694)	21.21
(1,5)	(-8.315295, 0.039313630)	22.26	9	(-8.315295, 0.039313630)	22.26
(1,6)	(-7.650265, 0.033349235)	19.48	9	(-7.650265, 0.033349235)	19.48
(1,7)	(-7.852464, 0.035162673)	17.87	11	(-7.852464, 0.035162673)	17.87
(1,8)	(-7.354763, 0.030698989)	23.52	12	(-7.354763, 0.030698989)	23.52
(2,3)	(3.382085, -0.049452325)	62.79	8	(3.382085, -0.049452325)	62.79
(2,4)	(-1.761884, -0.007115133)	39.68	8	(-1.761884, -0.007115133)	39.68
(2,5)	(-3.708094, 0.008903058)	29.53	9	(-3.708094, 0.008903058)	29.53
(2,6)	(-3.725972, 0.009050207)	29.49	9	(-3.725972, 0.009050207)	29.49
(2,7)	(-4.791547, 0.017820369)	32.74	11	(-4.791547, 0.017820369)	32.74
(2,8)	(-4.623252, 0.016435224)	31.21	11	(-4.623252, 0.016435224)	31.21
(3,4)	(-7.752596, 0.035222059)	16.97	9	(-7.752596, 0.035222059)	16.97
(3,5)	(-8.128514, 0.038080749)	21.01	10	(-8.128514, 0.038080749)	21.01
(3,6)	(-6.875359, 0.028551051)	15.45	9	(-6.875359, 0.028551051)	15.45
(3,7)	(-7.430132, 0.032769856)	14.48	11	(-7.430132, 0.032769856)	14.48
(3,8)	(-6.726163, 0.027416482)	16.20	11	(-6.726163, 0.027416482)	16.20
(4,5)	(-8.561605, 0.040939439)	24.30	9	(-8.561605, 0.040939439)	24.30
(4,6)	(-6.336675, 0.025215547)	15.25	9	(-6.336675, 0.025215547)	15.25
(4,7)	(-7.306470, 0.032069227)	14.17	11	(-7.306470, 0.032069227)	14.17
(4,8)	(-6.427209, 0.025855366)	14.77	12	(-6.427209, 0.025855366)	14.77
(5,6)	(-3.797266, 0.009491656)	28.97	9	(-3.797266, 0.009491656)	28.97
(5,7)	(-6.680233, 0.028521142)	16.52	12	(-6.680233, 0.028521142)	16.52
(5,8)	(-5.705059, 0.022084348)	18.51	13	(-5.705059, 0.022084348)	18.51
(6,7)	(-8.919370, 0.041207466)	27.27	12	(-8.919370, 0.041207466)	27.27
(6,8)	(-6.508893, 0.026281912)	15.03	13	(-6.508893, 0.026281912)	15.03
(7,8)	(-3.650649, 0.011356358)	58.76	12	(-3.650649, 0.011356358)	58.76

Table 5.3, we present the computation times for two different initial values.

Table 5.2: GSC Method for Whole Data (Cornfield Data)

Stopping Rule	Initial Value	m	$\beta_0^{(m)}$	$\beta_1^{(m)}$	Δ	$-2\log L$
$\frac{\lvert\Delta^{(m+1)}-\Delta^{(m)}\rvert}{\Delta^{(m)}} < 10^{-5}$	(0,0)	6	-6.082	0.0243	17.14	38.62
	(-7.306,0.0321)	3	-6.082	0.0243	17.14	38.62

Table 5.3: Computation Time of GSC Method for Whole Data (Cornfield Data)

Initial Value	10 times	10^2 times	10^3 times	10^4 times	10^5 times
(0,0)	0.21 s	0.58 s	4.48 s	43.05 s	412.73 s
(-7.306,0.0321)	0.18 s	0.42 s	2.72 s	25.37 s	251.72 s

In Tables 5.2 and 5.3, we see that the GSC method converges to the same values no matter what initial values we choose. However, the GSC method using the PGSC method estimates as initial value converges faster than the $(0, 0)$ **initial value.**

5.5 Comparison Between GSC Method for General N and PGSC Method

We now compare the two methods in section 5.3 and 5.4. In Table 5.4, we present the comparison of two sets of estimates, and the corresponding Δ and $-2\log L$ values.

In Table 5.4, we observe that the PGSC method gives a better estimate than the GSC method for whole data with respect to the criterion function

Table 5.4: Comparison Between GSC Method for General N and PGSC Method (Cornfield Data)

Method	$\left(\hat{\beta}_0, \hat{\beta}_1\right)$	Δ	$-2\log L$
GSC (Whole Data)	(-6.082007, 0.02433807)	17.13631	38.61038
PGSC	(-7.306472, 0.03206920)	14.17023	41.56182

Δ. However, the GSC method for whole data performs better than the PGSC method with respect to the criterion function $-2\log L$. Figure 5.1 plots $\frac{y_i}{n_i}$ against z_i where $i = 1, \ldots, 8$ for PGSC and GSC methods.

Figure 5.1: Comparison Between GSC Method for General N and PGSC Method (Cornfield Data)

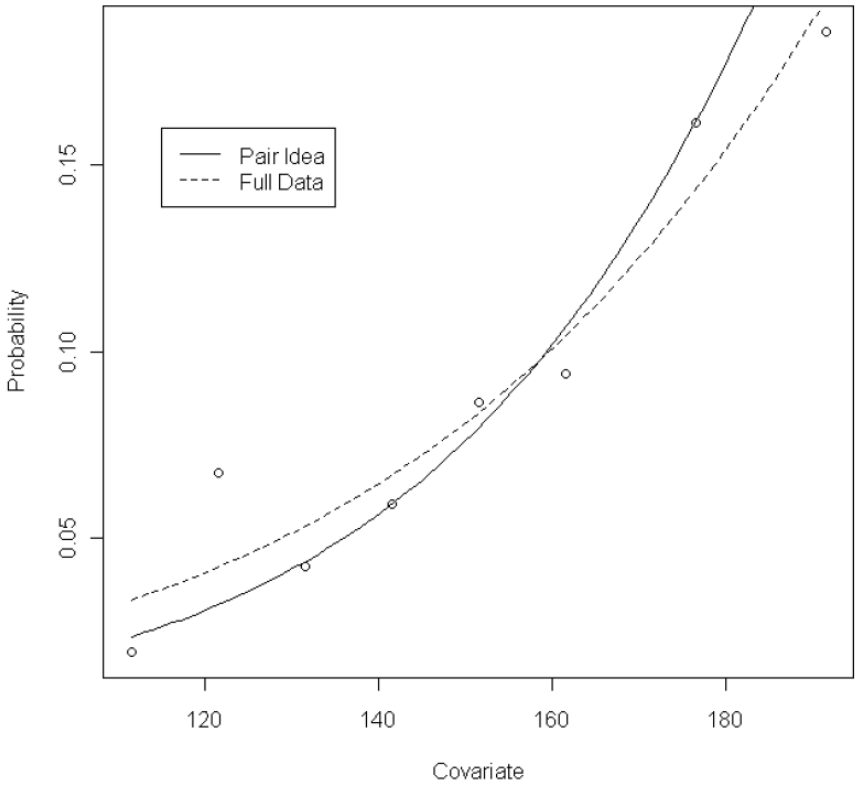
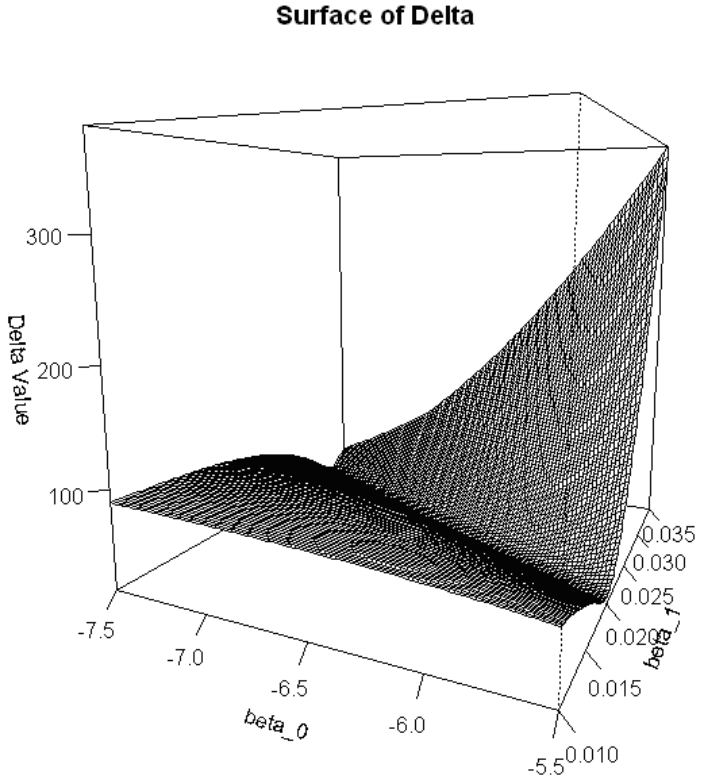


Figure 5.2 plots the Δ values against β_0 in $(-7.5, -5.5)$ and β_1 in $(0.010, 0.035)$ and Figure 5.3 provides the contour plot of Δ surface against β_0 in $(-7.5, -5.5)$ and β_1 in $(0.010, 0.035)$.

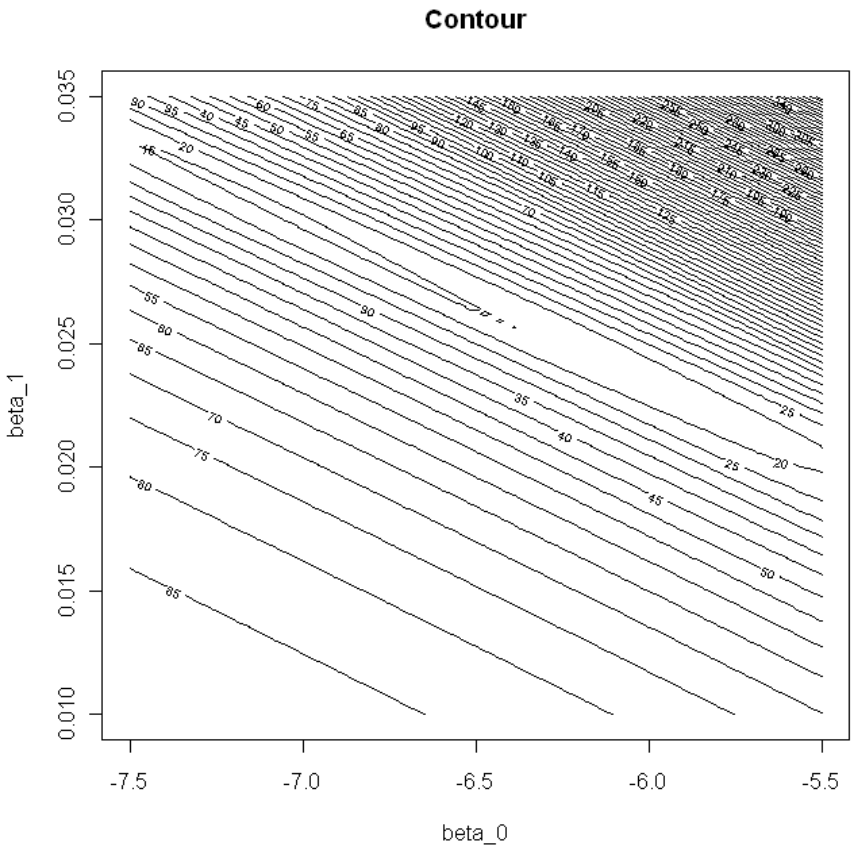
Figure 5.2: 3-D Surface Plot of Δ (Cornfield Data)



In Figure 5.2, we see that the location of (-7.306472, 0.0320692) provided by the PGSC method is closer to the "dip" which represents the lowest Δ values in the region, while the location of (-6.082007, 0.02433807) is far away from the "dip".

In Figure 5.3, we see that the estimate of (-7.306472, 0.0320692) provided by the PGSC method makes Δ fall in "15 contour". However, the estimate of (-6.082007, 0.02433807) provided by the GSC method to whole data only makes Δ fall in "20 contour". This demonstrates that the PGSC method is better with respect to Δ.

Figure 5.3: Contour Plot of Δ Surface (Cornfield Data)



Chapter 6

Properties of GSC and PGSC Methods

6.1 Introduction

In this chapter, we investigate the theoretical properties of the GSC and PGSC methods when applied to the logistic regression model. These properties include the convergence of the iteration process and the choice of the initial value.

6.2 GSC Method

We consider first the GSC method for general N to obtain the estimates of β_0 and β_1 from (4.3). We now study the theoretical properties of GSC method.

6.2.1 Property 1: Identical Solution

First, we compare the GSC method with the regular MLE estimating equation solving method. We find that GSC method gives the same solution as the regular MLE esti-

mating equation solving method. We denote the numerical solution of (3.10) by $\hat{\beta}_0$ and $\hat{\beta}_1$ and call them true solution. We have the following results.

Theorem 2 *In the GSC method (4.3), if two consecutive estimates at mth step and $(m+1)$th step are close enough, i.e., $\beta_0^{(m)}$ is close enough to $\beta_0^{(m+1)}$ and $\beta_1^{(m)}$ is close enough to $\beta_1^{(m+1)}$, then the estimates $\beta_0^{(m+1)}$ and $\beta_1^{(m+1)}$ will be identical with the true solution of (3.10), $\hat{\beta}_0$ and $\hat{\beta}_1$. In other words, if we define $\beta_0^{(m)} = \beta_0^{(m+1)}$, $\beta_1^{(m)} = \beta_1^{(m+1)}$ up to a certain decimal place, then we must have $\beta_0^{(m)} = \beta_0^{(m+1)} = \hat{\beta}_0$, $\beta_1^{(m)} = \beta_1^{(m+1)} = \hat{\beta}_1$ up to the same decimal place.*

Proof. From (3.10), we have

$$\begin{cases} \sum\limits_{j=1}^{N} y_j = \sum\limits_{j=1}^{N} n_j \frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_j\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_j\right)} & (1) \\ \sum\limits_{j=1}^{N} y_j z_j = \sum\limits_{j=1}^{N} n_j z_j \frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_j\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_j\right)} & (2) \end{cases}$$

Now we define $\beta_0^{(m)} = \beta_0^{(m+1)}$, $\beta_1^{(m)} = \beta_1^{(m+1)}$ up to a certain decimal place, then from (4.3) we have

$$\begin{cases} \sum\limits_{j=1}^{N} y_j = \sum\limits_{j=1}^{N} n_j \frac{\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)}{1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)} & (3) \\ \sum\limits_{j=1}^{N} y_j z_j = \sum\limits_{j=1}^{N} n_j z_j \frac{\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)}{1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)} & (4) \end{cases}$$

From $(1)-(3)$ and $(2)-(4)$ we have

$$\begin{cases} \sum\limits_{j=1}^{N} \frac{n_j e^{\hat{\beta}_0+\hat{\beta}_1 z_j}}{\left(1+e^{\hat{\beta}_0+\hat{\beta}_1 z_j}\right)\left(1+e^{\beta_0^{(m)}+\beta_1^{(m)} z_j}\right)} \left[1-\exp\left(\beta_0^{(m)}-\hat{\beta}_0+\beta_1^{(m)} z_j-\hat{\beta}_1 z_j\right)\right] = 0 & (5) \\ \sum\limits_{j=1}^{N} \frac{n_j z_j e^{\hat{\beta}_0+\hat{\beta}_1 z_j}}{\left(1+e^{\hat{\beta}_0+\hat{\beta}_1 z_j}\right)\left(1+e^{\beta_0^{(m)}+\beta_1^{(m)} z_j}\right)} \left[1-\exp\left(\beta_0^{(m)}-\hat{\beta}_0+\beta_1^{(m)} z_j-\hat{\beta}_1 z_j\right)\right] = 0 & (6) \end{cases}$$

Denote $\tilde{y}_j = \frac{n_j e^{\hat{\beta}_0+\hat{\beta}_1 z_j}}{\left(1+e^{\hat{\beta}_0+\hat{\beta}_1 z_j}\right)\left(1+e^{\beta_0^{(m)}+\beta_1^{(m)} z_j}\right)}$ which is > 0 for all j, then (5) and (6) become

$$\begin{cases} \sum_{j=1}^{N} \tilde{y}_j = \sum_{j=1}^{N} \tilde{y}_j e^{\beta_0^{(m)}-\hat{\beta}_0+\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_j} & (7) \\ \sum_{j=1}^{N} z_j\tilde{y}_j = \sum_{j=1}^{N} z_j\tilde{y}_j e^{\beta_0^{(m)}-\hat{\beta}_0+\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_j} & (8) \end{cases}$$

Multiplying LHS of (7) by RHS of (8) and LHS of (8) by RHS of (7), we obtain

$$\sum_{i=1}^{N} \tilde{y}_i \sum_{j=1}^{N} z_j\tilde{y}_j e^{\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_j} = \sum_{i=1}^{N} z_i\tilde{y}_i \sum_{j=1}^{N} \tilde{y}_j e^{\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_j}$$

$$\Rightarrow$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} (z_i - z_j)\,\tilde{y}_i\tilde{y}_j e^{\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_j} = 0$$

$$\Rightarrow$$

$$\sum_{i<j}\sum_{j=1}^{N} (z_i - z_j)\,\tilde{y}_i\tilde{y}_j e^{\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_j} + \sum_{i>j}\sum_{j=1}^{N} (z_i - z_j)\,\tilde{y}_i\tilde{y}_j e^{\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_j} = 0$$

$$\Rightarrow$$

$$\sum_{i>j}\sum_{j=1}^{N} (z_i - z_j)\,\tilde{y}_i\tilde{y}_j \left(e^{\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_i} - e^{\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_j}\right) = 0 \qquad (6.1)$$

Without loss of generality, we assume $z_j > 0$ for any j and $z_i < z_j$ for $i < j$. In (6.1), there are three possibilities:

1. If $\beta_1^{(m)} < \hat{\beta}_1$

   For $i > j$, we have $z_i > z_j$ and $z_i - z_j > 0$. Also $\tilde{y}_j > 0$ for any j. Furthermore, we have $\beta_1^{(m)} - \hat{\beta}_1 < 0$. Therefore, $\left(\beta_1^{(m)} - \hat{\beta}_1\right) z_i < \left(\beta_1^{(m)} - \hat{\beta}_1\right) z_j$ which implies $e^{\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_i} - e^{\left(\beta_1^{(m)}-\hat{\beta}_1\right)z_j} < 0$. Hence, every term of (6.1) is < 0. This is impossible.

2. If $\beta_1^{(m)} > \hat{\beta}_1$

   For $i > j$, we have $z_i > z_j$ and $z_i - z_j > 0$. Also $\tilde{y}_j > 0$ for any j. Furthermore, we have $\beta_1^{(m)} - \hat{\beta}_1 > 0$. Therefore, $\left(\beta_1^{(m)} - \hat{\beta}_1\right) z_i > \left(\beta_1^{(m)} - \hat{\beta}_1\right) z_j$ which implies $e^{\left(\beta_1^{(m)} - \hat{\beta}_1\right) z_i} - e^{\left(\beta_1^{(m)} - \hat{\beta}_1\right) z_j} > 0$. Hence, every term of (6.1) is > 0. This is impossible.

3. If $\beta_1^{(m)} = \hat{\beta}_1$

   Equation (6.1) holds.

To sum up, we see that only the case $\beta_1^{(m)} = \hat{\beta}_1$ is possible. Thus we have proved that $\beta_1^{(m)} = \hat{\beta}_1$. Then by plugging $\beta_1^{(m)} = \hat{\beta}_1$ in (7) or (8), we see that $\beta_0^{(m)} = \hat{\beta}_0$. Thus $\beta_0^{(m)} = \beta_0^{(m+1)} = \hat{\beta}_0$, $\beta_1^{(m)} = \beta_1^{(m+1)} = \hat{\beta}_1$ up to a certain decimal place. ■

6.2.2 Property 2: Initial Value

We now consider the issue of the initial value choice for the GSC method (4.3). We find that if we use the true solution of the regular estimating equation to be the initial value, the GSC method will stop at the 1st iteration step and give exactly the true solution as the estimates. The following property gives more details.

Theorem 3 *In the GSC method (4.3), if we use the true solution of (3.10) to be the initial value of (4.3), i.e., $\beta_0^{(0)} = \hat{\beta}_0$, $\beta_1^{(0)} = \hat{\beta}_1$, then (4.3) would stop at the 1st iteration step and give the same solution $\beta_0^{(1)} = \hat{\beta}_0$ and $\beta_1^{(1)} = \hat{\beta}_1$.*

Proof. Denote $\hat{y}_j = n_j \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)}$, then (3.10) becomes

$$\begin{cases} \sum_{j=1}^{N} y_j = \sum_{j=1}^{N} n_j \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)} = \sum_{j=1}^{N} \hat{y}_j & (1) \\ \sum_{j=1}^{N} y_j z_j = \sum_{j=1}^{N} n_j z_j \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)} = \sum_{j=1}^{N} \hat{y}_j z_j & (2) \end{cases}$$

If we use $\beta_0^{(0)} = \hat{\beta}_0,\ \beta_1^{(0)} = \hat{\beta}_1$, then at the 1st step (4.3) becomes

$$\begin{cases} \sum_{j=1}^{N} y_j = \sum_{j=1}^{N} n_j \frac{\exp\left(\beta_0^{(1)} + \beta_1^{(1)} z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)} & (3) \\ \sum_{j=1}^{N} y_j z_j = \sum_{j=1}^{N} n_j z_j \frac{\exp\left(\beta_0^{(1)} + \beta_1^{(1)} z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)} & (4) \end{cases}$$

From (1) to (4) we have

$$\begin{cases} \sum_{j=1}^{N} y_j = \sum_{j=1}^{N} n_j \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)} \times e^{\left(\beta_0^{(1)} - \hat{\beta}_0\right) + \left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} \\ \sum_{j=1}^{N} y_j z_j = \sum_{j=1}^{N} n_j z_j \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)} \times e^{\left(\beta_0^{(1)} - \hat{\beta}_0\right) + \left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} \end{cases}$$

$\Rightarrow$

$$\begin{cases} \sum_{j=1}^{N} \hat{y}_j = \sum_{j=1}^{N} \hat{y}_j e^{\left(\beta_0^{(1)} - \hat{\beta}_0\right) + \left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} & (5) \\ \sum_{j=1}^{N} z_j \hat{y}_j = \sum_{j=1}^{N} z_j \hat{y}_j e^{\left(\beta_0^{(1)} - \hat{\beta}_0\right) + \left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} & (6) \end{cases}$$

Multiplying LHS of (5) by RHS of (6), LHS of (6) by RHS of (5), we obtain

$$\sum_{i=1}^{N} \hat{y}_i \sum_{j=1}^{N} z_j \hat{y}_j e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} = \sum_{i=1}^{N} z_i \hat{y}_i \sum_{j=1}^{N} \hat{y}_j e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j}$$

$\Rightarrow$

$$\sum_{i=1}^{N} \sum_{j=1}^{N} z_i \hat{y}_i \hat{y}_j e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} = \sum_{i=1}^{N} \sum_{j=1}^{N} z_j \hat{y}_i \hat{y}_j e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j}$$

$\Rightarrow$

$$\sum_{i=1}^{N} \sum_{j=1}^{N} (z_i - z_j)\, \hat{y}_i \hat{y}_j e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} = 0$$

$\Rightarrow$

$$\sum_{i>j} \sum_{j=1}^{N} (z_i - z_j)\, \hat{y}_i \hat{y}_j e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} + \sum_{i<j} \sum_{j=1}^{N} (z_i - z_j)\, \hat{y}_i \hat{y}_j e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} = 0$$

$\Rightarrow$

$$\sum_{i>j} \sum_{j=1}^{N} (z_i - z_j)\, \hat{y}_i \hat{y}_j \left(e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_i} - e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} \right) = 0 \qquad (6.2)$$

Without loss of generality, we assume $z_j > 0$ for any j and $z_i < z_j$ for $i < j$. In (6.2), there are three possibilities:

1. If $\beta_1^{(1)} < \hat{\beta}_1$

   For $i > j$, we have $z_i > z_j$ and $z_i - z_j > 0$. Also $\hat{y}_j > 0$ for any j. Furthermore, we have $\beta_1^{(1)} - \hat{\beta}_1 < 0$. Therefore, $\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_i < \left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j$ which implies $e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_i} - e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} < 0$. Hence, every term of (6.2) is < 0. This is impossible.

2. If $\beta_1^{(1)} > \hat{\beta}_1$

   For $i > j$, we have $z_i > z_j$ and $z_i - z_j > 0$. Also $\hat{y}_j > 0$ for any j. Furthermore, we have $\beta_1^{(1)} - \hat{\beta}_1 > 0$. Therefore, $\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_i > \left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j$ which implies $e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_i} - e^{\left(\beta_1^{(1)} - \hat{\beta}_1\right) z_j} > 0$. Hence, every term of (6.2) is > 0. This is impossible.

3. If $\beta_1^{(1)} = \hat{\beta}_1$

   Equation (6.2) holds.

To sum up, we see that only the case $\beta_1^{(1)} = \hat{\beta}_1$ is possible. Thus we have proved that $\beta_1^{(1)} = \hat{\beta}_1$. Then by plugging $\beta_1^{(1)} = \hat{\beta}_1$ in (5) or (6), it is easy to show that $\beta_0^{(1)} = \hat{\beta}_0$. ■

6.3 PGSC Method

We recall that the PGSC method is to obtain the estimates of β_0 and β_1 by using (4.4). Now we study the theoretical properties of PGSC method.

6.3.1 Property 1: Identical Solution

We consider the uth pair containing observations i and j. First, we compare the PGSC method (4.4) with regular MLE estimating equation solving method (3.11). We find

that the PGSC method gives the same solution as the regular MLE estimating equation solving method. We denote the exact solution of (3.11) by $\hat{\beta}_0$ and $\hat{\beta}_1$ and call them true solution. We have the following property.

Theorem 4 *In the PGSC method (4.4), if two consecutive estimates at mth step and* $(m+1)$*th step are close enough, i.e.,* $\beta_0^{(m)}$ *is close enough to* $\beta_0^{(m+1)}$ *and* $\beta_1^{(m)}$ *is close enough to* $\beta_1^{(m+1)}$*, then the estimates* $\beta_0^{(m+1)}$ *and* $\beta_1^{(m+1)}$ *will be identical with the exact solution of (3.11),* $\hat{\beta}_0$ *and* $\hat{\beta}_1$*. In other words, if we define* $\beta_0^{(m)} = \beta_0^{(m+1)}$, $\beta_1^{(m)} = \beta_1^{(m+1)}$ *up to a certain decimal place, then we must have* $\beta_0^{(m)} = \beta_0^{(m+1)} = \hat{\beta}_0$, $\beta_1^{(m)} = \beta_1^{(m+1)} = \hat{\beta}_1$ *up to the same decimal place.*

We present two proofs of Theorem 4.

Proof 1. From (4.4), we have

$$\begin{cases} \frac{y_i}{n_i} = \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_i\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_i\right)} & (1) \\ \frac{y_j}{n_j} = \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_j\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)} & (2) \end{cases}$$

From (3.12), we have

$$\begin{cases} \frac{y_i}{n_i} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_i\right)} & (3) \\ \frac{y_j}{n_j} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)} & (4) \end{cases}$$

Dividing (1) by (3) and (2) by (4), we have

$$\begin{cases} 1 = e^{\left(\beta_0^{(m+1)} - \hat{\beta}_0\right) + \left(\beta_1^{(m+1)} - \hat{\beta}_1\right) z_i} \frac{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_i\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_i\right)}, \\ 1 = e^{\left(\beta_0^{(m+1)} - \hat{\beta}_0\right) + \left(\beta_1^{(m+1)} - \hat{\beta}_1\right) z_j} \frac{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)}. \end{cases}$$

For $\beta_0^{(m)} = \beta_0^{(m+1)},\ \beta_1^{(m)} = \beta_1^{(m+1)}$ up to a certain decimal place, we get

$$\begin{cases} 1 = \frac{\exp\left(\left(\beta_0^{(m)} - \hat{\beta}_0\right) + \left(\beta_1^{(m)} - \hat{\beta}_1\right) z_i\right) + \exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_i\right)}{1 + \exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_i\right)}, \\ 1 = \frac{\exp\left(\left(\beta_0^{(m)} - \hat{\beta}_0\right) + \left(\beta_1^{(m)} - \hat{\beta}_1\right) z_j\right) + \exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)}{1 + \exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)}. \end{cases}$$

$$\Rightarrow$$

$$\begin{cases} \left(\beta_0^{(m)} - \hat{\beta}_0\right) + \left(\beta_1^{(m)} - \hat{\beta}_1\right) z_i = 0, \\ \left(\beta_0^{(m)} - \hat{\beta}_0\right) + \left(\beta_1^{(m)} - \hat{\beta}_1\right) z_j = 0. \end{cases}$$

$$\Rightarrow$$

$$\begin{cases} \beta_0^{(m)} = \hat{\beta}_0, \\ \beta_1^{(m)} = \hat{\beta}_1. \end{cases}$$

Thus we have proved that $\beta_0^{(m)} = \beta_0^{(m+1)} = \hat{\beta}_0,\ \beta_1^{(m)} = \beta_1^{(m+1)} = \hat{\beta}_1$. ■

We now present the second proof.

Proof 2. From (4.4), we have

$$\begin{cases} y_i = n_i \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_i\right)}{1 + \exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_i\right)} \\ y_j = n_j \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_j\right)}{1 + \exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)} \end{cases}$$

And the predicted values are expressed as

$$\begin{cases} \hat{y}_i^{(m+1)u} = n_i \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_i\right)}{1 + \exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_i\right)} \\ \hat{y}_j^{(m+1)u} = n_j \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_j\right)}{1 + \exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_j\right)} \end{cases}$$

Denote $a_m = \exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_i\right)$ and we will have a sequence of $\{a_m\}$. This sequence has the following property

$$y_i = n_i \frac{a_{m+1}}{1 + a_m} \quad \Rightarrow \quad \frac{a_{m+1}}{1 + a_m} = \frac{y_i}{n_i} = q_i \leq 1$$

where q_i is a constant and usually $q_i < 1$.

Therefore we have

$$\begin{aligned} a_{m+1} &= q_i\,(1 + a_m) = q_i + q_i a_m \\ &= q_i + q_i^2\,(1 + a_{m-1}) = q_i + q_i^2 + q_i^2 a_{m-1} \\ &= q_i + q_i^2 + q_i^3 + \cdots + q_i^{m+1} + q_i^{m+1} a_0 \\ &= \tfrac{q_i\left(1-q_i^{m+1}\right)}{1-q_i} + q_i^{m+1} a_0 \\ &\Rightarrow \lim_{m\to\infty} a_m = \frac{q_i}{1-q_i} \end{aligned}$$

Plug the above limit in the expressions of the predicted values, and then we have

$$\lim_{m\to\infty} \hat{y}_i^{(m)u} = \lim_{m\to\infty} n_i \frac{\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_i\right)}{1 + \exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_i\right)} = \lim_{m\to\infty} n_i \frac{a_m}{1 + a_m} = n_i q_i = y_i$$

From (3.11) and (3.12), we have $\hat{y}_i = y_i$ and $\hat{y}_j = y_j$. If we continue the iteration for the PGSC method for a long time, then $\hat{y}_i^{(m)u}$ converges to y_i. This implies that $\lim_{m\to\infty} \hat{y}_i^{(m)u} = y_i$ and $\lim_{m\to\infty} \hat{y}_j^{(m)u} = y_j$. Therefore, in limit, the predicted y values for i and j approach to their true values. In other words, the solution of (4.4) converges to the solution of (3.11). ■

Note: There is an important strength regarding the initial values of the PGSC method. The initial values of $\beta_0^{(0)},\ \beta_1^{(0)}$ give us $a_0 = \exp\left(\beta_0^{(0)} + \beta_1^{(0)} z_i\right)$. According to Proof 2, no matter what initial values we choose, $\hat{y}_i^{(m)u}$ will always converge to the real value y_i. In other words, the choice of a_0 does not influence the convergence of $\hat{y}_i^{(m)u}$ to y_i.

6.3.2 Property 2: Initial Value

We now consider the issue of the initial value choice for the PGSC method (4.4). We find that if we use the exact solution of the regular estimating equation to be the initial value, the PGSC method will stop at the 1st iteration step and give exactly the exact

solution as the estimates. The following property gives more details.

Theorem 5 *In the PGSC method (4.4), if we use the exact solution of (3.11) to be the initial value of (4.4), i.e., let* $\beta_0^{(0)} = \hat{\beta}_0$, $\beta_1^{(0)} = \hat{\beta}_1$, *then (4.4) would stop at the 1st iteration step and give the same solution* $\beta_0^{(1)} = \hat{\beta}_0$ *and* $\beta_1^{(1)} = \hat{\beta}_1$.

Proof. From (4.4), we have

$$\begin{cases} \frac{y_i}{n_i} = \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_i\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_i\right)} \\ \frac{y_j}{n_j} = \frac{\exp\left(\beta_0^{(m+1)} + \beta_1^{(m+1)} z_j\right)}{1+\exp\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)} \end{cases}$$

Now we consider the first step, then we have

$$\begin{cases} \frac{y_i}{n_i} = \frac{\exp\left(\beta_0^{(1)} + \beta_1^{(1)} z_i\right)}{1+\exp\left(\beta_0^{(0)} + \beta_1^{(0)} z_i\right)} \\ \frac{y_j}{n_j} = \frac{\exp\left(\beta_0^{(1)} + \beta_1^{(1)} z_j\right)}{1+\exp\left(\beta_0^{(0)} + \beta_1^{(0)} z_j\right)} \end{cases}$$

We use the exact solution of (3.11) $\hat{\beta}_0$ and $\hat{\beta}_1$ to be the initial value, i.e., $\beta_0^{(0)} = \hat{\beta}_0$ and $\beta_1^{(0)} = \hat{\beta}_1$. Then we have

$$\begin{cases} \frac{y_i}{n_i} = \frac{\exp\left(\beta_0^{(1)} + \beta_1^{(1)} z_i\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_i\right)} & (1) \\ \frac{y_j}{n_j} = \frac{\exp\left(\beta_0^{(1)} + \beta_1^{(1)} z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)} & (2) \end{cases}$$

Also, from (3.11), we have

$$\begin{cases} \frac{y_i}{n_i} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_i\right)} & (3) \\ \frac{y_j}{n_j} = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_j\right)} & (4) \end{cases}$$

Dividing (1) by (3) and (2) by (4), we get

$$\begin{cases} 1 = \frac{\exp\left(\beta_0^{(1)}+\beta_1^{(1)}z_i\right)}{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)} \\ 1 = \frac{\exp\left(\beta_0^{(1)}+\beta_1^{(1)}z_j\right)}{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_j\right)} \end{cases}$$

$$\Rightarrow$$

$$\begin{cases} \beta_0^{(1)} + \beta_1^{(1)} z_i = \hat{\beta}_0 + \hat{\beta}_1 z_i \\ \beta_0^{(1)} + \beta_1^{(1)} z_j = \hat{\beta}_0 + \hat{\beta}_1 z_j \end{cases}$$

$$\Rightarrow$$

$$\begin{cases} \beta_0^{(1)} = \hat{\beta}_0 \\ \beta_1^{(1)} = \hat{\beta}_1 \end{cases}$$

■

6.3.3 Property 3: Monotonic Convergence

Now that we know the PGSC method will converge, we would like to investigate how well it performs. We find another good property of the PGSC method.

Theorem 6 *Consider the PGSC method (4.4) and the uth pair containing observations i and j. Every step of the PGSC method provides us a predicted value closer to the real value and monotonically converges to the real value. In other words, from (4.4), we have $(\beta_0^{(m)}, \beta_1^{(m)})$, $(\beta_0^{(m+1)}, \beta_1^{(m+1)})$ and $(\hat{y}_i^{(m)}, \hat{y}_i^{(m+1)})$. Then $\hat{y}_i^{(m)} < \hat{y}_i^{(m+1)} < y_i$ or $y_i < \hat{y}_i^{(m+1)} < \hat{y}_i^{(m)}$, and finally $\hat{y}_i^{(m+1)} = \hat{y}_i^{(m)} = y_i$. Similarly for j.*

Proof. Define $a_{0i} = 1 + \exp\left(\beta_0^{(0)} + \beta_1^{(0)} z_i\right)$. In Table 6.1, we present $\hat{y}_i^{(w)u}$ for $w = 0, 1, \ldots, m$.

Table 6.1: Progress of $\hat{y}_i^{(w)u}$ for the PGSC Method

w	$\exp\left(\beta_0^{(w)} + \beta_1^{(w)} z_i\right)$	$\hat{y}_i^{(w)u}$
0		
1	$a_{0i}\frac{y_i}{n_i}$	$\frac{a_{0i}y_i}{1+a_{0i}\frac{y_i}{n_i}}$
2	$a_{0i}\frac{y_i}{n_i}\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)$	$\frac{a_{0i}y_i\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)}{1+a_{0i}\frac{y_i}{n_i}\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)}$
3	$a_{0i}\frac{y_i}{n_i}\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(2)u}}\right)$	$\frac{a_{0i}y_i\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(2)u}}\right)}{1+a_{0i}\frac{y_i}{n_i}\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(2)u}}\right)}$
$\vdots$	$\vdots$	$\vdots$
m	$a_{0i}\frac{y_i}{n_i}\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(2)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)$	$\frac{a_{0i}y_i\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(2)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)}{1+a_{0i}\frac{y_i}{n_i}\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(2)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)}$

We have

$$\hat{y}_i^{(m)u} = \frac{a_{0i}y_i\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)}{1+a_{0i}\frac{y_i}{n_i}\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)} \tag{6.3}$$

$$= \frac{1}{\frac{1}{a_{0i}y_i\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)}+\frac{1}{n_i}},$$

and

$$\hat{y}_i^{(m+1)u} = \frac{a_{0i}y_i\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(m)u}}\right)}{1+a_{0i}\frac{y_i}{n_i}\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(m)u}}\right)} \tag{6.4}$$

$$= \frac{1}{\frac{1}{a_{0i}y_i\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(m)u}}\right)}+\frac{1}{n_i}}.$$

We now compare $\hat{y}_i^{(m)u}$ and $\hat{y}_i^{(m+1)u}$. There are three possibilities:

1. If $\left(\frac{y_i}{\hat{y}_i^{(m)u}}\right) < 1$, then $y_i < \hat{y}_i^{(m)u}$

   This means at the mth step we overestimate y_i. However, we observe that $\hat{y}_i^{(m+1)u}$ is decreasing compared with $\hat{y}_i^{(m)u}$, i.e., $\hat{y}_i^{(m+1)u} < \hat{y}_i^{(m)u}$. Furthermore, since $0 < y_i < \hat{y}_i^{(m)u} < n_i$, we get

   $$\hat{y}_i^{(m+1)u} = \frac{1}{\frac{1}{a_{0i}y_i\left(\frac{y_i}{\hat{y}_i^{(1)u}}\right)\cdots\left(\frac{y_i}{\hat{y}_i^{(m-1)u}}\right)\left(\frac{y_i}{\hat{y}_i^{(m)u}}\right)}+\frac{1}{n_i}} = \frac{1}{\left(\frac{1}{\hat{y}_i^{(m)u}}-\frac{1}{n_i}\right)\left(\frac{\hat{y}_i^{(m)u}}{y_i}\right)+\frac{1}{n_i}}$$

   $$= \frac{y_i}{1-\frac{\hat{y}_i^{(m)u}}{n_i}+\frac{y_i}{n_i}} = \frac{y_i}{1+\frac{\left(y_i-\hat{y}_i^{(m)u}\right)}{n_i}} > \frac{y_i}{1}$$

   Thus $y_i < \hat{y}_i^{(m+1)u} < \hat{y}_i^{(m)u}$. This indicates that $\hat{y}_i^{(m)u}$ will monotonically converge to y_i.

2. If $\left(\frac{y_i}{\hat{y}_i^{(m)u}}\right) > 1$, then $y_i > \hat{y}_i^{(m)u}$

   This means at the mth step we underestimate y_i. However, we observe that $\hat{y}_i^{(m+1)u}$ is increasing compared with $\hat{y}_i^{(m)u}$, i.e., $\hat{y}_i^{(m+1)u} > \hat{y}_i^{(m)u}$. Furthermore, since $0 < \hat{y}_i^{(m)u} < y_i < n_i$, we get

$$\hat{y}_i^{(m+1)u} = \frac{y_i}{1 + \frac{\left(y_i - \hat{y}_i^{(m)u}\right)}{n_i}} < \frac{y_i}{1}$$

   Thus $\hat{y}_i^{(m)u} < \hat{y}_i^{(m+1)u} < y_i$. This indicates that $\hat{y}_i^{(m)u}$ will monotonically converge to y_i.

3. If $\left(\frac{y_i}{\hat{y}_i^{(m)u}}\right) = 1$, then $y_i = \hat{y}_i^{(m)u}$

   This means at the mth step our estimate of y_i exactly equals y_i. By comparing $\hat{y}_i^{(m)u}$ with $\hat{y}_i^{(m+1)u}$, we observe that $\hat{y}_i^{(m)u} = \hat{y}_i^{(m+1)u}$. In other words, if at some step our estimate of y_i equals y_i, then from this step on every estimate of y_i will be equal to the real value of y_i. If this happens, the algorithm stops and converges.

■

Chapter 7

Our Methods

7.1 Introduction

In this chapter, we combine the paired observation procedure with the GSC method in a different way with in Chapter 5. This results in two new methods for solving MLEEs of logistic regression model. We present the details of these methods illustrating with the Cornfield data. We also compare these two methods with the standard method. We present a simulation study for our illustration.

7.2 Method 1

We first introduce four criteria based on (3.10) to measure the goodness of fit of logistic regression model.

Definition 7 *For a given estimate* $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ *of* (β_0, β_1)*, we define the first criterion* Δ

to be

$$\begin{cases} \Delta_1\left(\hat{\beta}_0,\hat{\beta}_1\right)=\sum\limits_{i=1}^{N}\left|y_i-n_i\frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}\right| \\ \Delta_2\left(\hat{\beta}_0,\hat{\beta}_1\right)=\sum\limits_{i=1}^{N}\left|y_i z_i-n_i z_i\frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}\right| \\ \Delta\left(\hat{\beta}_0,\hat{\beta}_1\right)=\Delta_1\left(\hat{\beta}_0,\hat{\beta}_1\right)+\Delta_2\left(\hat{\beta}_0,\hat{\beta}_1\right) \end{cases} \tag{7.1}$$

Definition 8 *For a given estimate* $\left(\hat{\beta}_0,\hat{\beta}_1\right)$ *of* (β_0,β_1)*, we define the second criterion* $\tilde{\Delta}$ *to be*

$$\begin{cases} \tilde{\Delta}_1\left(\hat{\beta}_0,\hat{\beta}_1\right)=\sum\limits_{i=1}^{N}\left|\frac{y_i}{n_i}-\frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}\right| \\ \tilde{\Delta}_2\left(\hat{\beta}_0,\hat{\beta}_1\right)=\sum\limits_{i=1}^{N}\left|\frac{y_i}{n_i}z_i-z_i\frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}\right| \\ \tilde{\Delta}\left(\hat{\beta}_0,\hat{\beta}_1\right)=\tilde{\Delta}_1\left(\hat{\beta}_0,\hat{\beta}_1\right)+\tilde{\Delta}_2\left(\hat{\beta}_0,\hat{\beta}_1\right) \end{cases} \tag{7.2}$$

Definition 9 *For a given estimate* $\left(\hat{\beta}_0,\hat{\beta}_1\right)$ *of* (β_0,β_1)*, we define the third criterion* Δ^2 *to be*

$$\begin{cases} \Delta_1^2\left(\hat{\beta}_0,\hat{\beta}_1\right)=\sum\limits_{i=1}^{N}\left(y_i-n_i\frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}\right)^2 \\ \Delta_2^2\left(\hat{\beta}_0,\hat{\beta}_1\right)=\sum\limits_{i=1}^{N}\left(y_i z_i-n_i z_i\frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}\right)^2 \\ \Delta^2\left(\hat{\beta}_0,\hat{\beta}_1\right)=\Delta_1^2\left(\hat{\beta}_0,\hat{\beta}_1\right)+\Delta_2^2\left(\hat{\beta}_0,\hat{\beta}_1\right) \end{cases} \tag{7.3}$$

Definition 10 *For a given estimate* $\left(\hat{\beta}_0,\hat{\beta}_1\right)$ *of* (β_0,β_1)*, we define the fourth criterion* $\tilde{\Delta}^2$ *to be*

$$\begin{cases} \tilde{\Delta}_1^2\left(\hat{\beta}_0,\hat{\beta}_1\right)=\sum\limits_{i=1}^{N}\left(\frac{y_i}{n_i}-\frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}\right)^2 \\ \tilde{\Delta}_2^2\left(\hat{\beta}_0,\hat{\beta}_1\right)=\sum\limits_{i=1}^{N}\left(\frac{y_i}{n_i}z_i-z_i\frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1 z_i\right)}\right)^2 \\ \tilde{\Delta}^2\left(\hat{\beta}_0,\hat{\beta}_1\right)=\tilde{\Delta}_1^2\left(\hat{\beta}_0,\hat{\beta}_1\right)+\tilde{\Delta}_2^2\left(\hat{\beta}_0,\hat{\beta}_1\right) \end{cases} \tag{7.4}$$

In (4.4), we have given the exact closed form solution of (4.3) for $N=2$. We now use it to introduce a new method to solve (4.3) for general N. We have M pairs of data and solve each pair separately. Then we get M pairs of $\left(\beta_0^{(m+1)},\ \beta_1^{(m+1)}\right)$. We select the pair that provides us the smallest value of the criterion function (defined above). We then use the estimate in (4.4) for this chosen pair as our new estimate for general

N. We perform the iteration process until it converges. The stopping rule we use is the same as in standard method "glm" in R, which is $\frac{|Deviance^{(m)} - Deviance^{(m+1)}|}{\left(|Deviance^{(m+1)}|+0.1\right)} < 10^{-8}$. The step-by-step description of this method is as follows:

1. Set initial values of $\left(\beta_0^{(0)},\ \beta_1^{(0)}\right) = (0, 0)$ and set iteration step $m = 0$.
2. In (4.4), plug $\left(\beta_0^{(0)},\ \beta_1^{(0)}\right)$ to obtain $\left(\beta_0^{(1)},\ \beta_1^{(1)}\right)$ for M pairs.
3. Calculate M values of the criterion function for M pairs.
4. Select the pair giving the smallest value of the criterion function.
5. Take $\left(\beta_0^{(1)},\ \beta_1^{(1)}\right)$ from the pair chosen in step 4 to calculate $\left(\beta_0^{(2)},\ \beta_1^{(2)}\right)$. Then repeat this process until the convergence criterion is reached.

7.2.1 Analysis of Real Example

For each criterion function, we present the convergence steps in Tables 7.1-7.4. In Table 7.5, we compare the criterion values for different estimates obtained by using our method.

Note: Criterion V is the standard method "glm" in R.

When we apply these five criteria to Method 1, Criterion III (Δ^2) has an interesting phenomenon. It does not converge to one solution but fluctuates between two solutions. In this case, we choose the solution which provides the smaller Δ^2 criterion value.

In Table 7.5, we have the smallest Δ value for Criterion I and the smallest $\tilde{\Delta}$

Table 7.1: Convergence Steps for Criterion I (Δ)

	Δ	$\hat{\beta}_0$	$\hat{\beta}_1$
Step 0	78963.75	0.00000	0.000000
Step 1	4936.55	-6.42119	0.028369
Step 2	2015.24	-6.79344	0.028345
Step 3	1964.32	-7.22182	0.031465
Step 4	1959.40	-7.29250	0.031970
Step 5	1958.78	-7.30419	0.032053
Step 6	1958.69	-7.30610	0.032067
Step 7	1958.68	-7.30641	0.032069
Step 8	1958.68	-7.30646	0.032069
Step 9	1958.68	-7.30647	0.032069
Step 10	1958.68	-7.30647	0.032069
Step 11	1958.68	-7.30647	0.032069
Step 12	1958.68	-7.30647	0.032069

Table 7.2: Convergence Steps for Criterion II ($\tilde{\Delta}$)

	$\tilde{\Delta}$	$\hat{\beta}_0$	$\hat{\beta}_1$
Step 0	480.10	0.00000	0.000000
Step 1	43.03	-6.42119	0.028369
Step 2	13.18	-6.47543	0.026376
Step 3	12.28	-6.44904	0.026020
Step 4	12.13	-6.43203	0.025890
Step 5	12.11	-6.42815	0.025862
Step 6	12.10	-6.42739	0.025857
Step 7	12.10	-6.42724	0.025856
Step 8	12.10	-6.42722	0.025855
Step 9	12.10	-6.42721	0.025855

Table 7.3: Convergence Steps for Criterion III (Δ^2)

	Δ^2	$\hat{\beta}_0$	$\hat{\beta}_1$
Step 0	972112351	0.00000	0.000000
Step 1	3257854	-6.42119	0.028369
Step 2	1134804	-6.01479	0.024279
Step 3	1137043	-6.58741	0.027922
Step 4	1111550	-5.88685	0.023285
Step 5	1122687	-6.55891	0.027720
Step 6	1113400	-5.87919	0.023234
Step 7	1122282	-6.55773	0.027713
Step 8	1113474	-5.87892	0.023232
Step 9	1122270	-6.55770	0.027712
Step 10	1113477	-5.87891	0.023232
Step 11	1122269	-6.55769	0.027712
Step 12	1113477	-5.87891	0.023232

Table 7.4: Convergence Steps for Criterion IV ($\tilde{\Delta}^2$)

	$\tilde{\Delta}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$
Step 0	28559.49	0.00000	0.000000
Step 1	387.78	-6.42119	0.028369
Step 2	35.53	-6.47543	0.026376
Step 3	32.53	-5.78911	0.022574
Step 4	33.58	-5.71601	0.022151
Step 5	33.76	-5.70670	0.022095
Step 6	33.80	-5.70533	0.022086
Step 7	33.80	-5.70511	0.022085
Step 8	33.80	-5.70507	0.022084
Step 9	33.80	-5.70506	0.022084
Step 10	33.80	-5.70506	0.022084
Step 11	33.80	-5.70506	0.022084
Step 12	33.80	-5.70506	0.022084

Table 7.5: Comparison of Estimates Obtained by Using Different Criteria for Method 1 (Cornfield Data)

Criterion	m	$\hat{\beta}_0$	$\hat{\beta}_1$	Δ	$\tilde{\Delta}$	Δ^2	$\tilde{\Delta}^2$
I	12	-7.30647	0.032069	1958.68	17.97	1417676	121.78
II	9	-6.42721	0.025855	2006.12	12.10	1253788	41.99
III	13	-5.87891	0.023232	2448.91	14.39	1113477	31.48
IV	12	-5.70506	0.022084	2494.43	13.75	1175245	33.80
V	5	-6.08203	0.024338	2319.25	14.11	1040658	30.79

value for Criterion II. Consequently, Method 1 using Criterion I is performing better than the standard method. The same is true for Method 1 using Criterion II. On the other hand, the standard method is performing better than Method I using Criterion III and IV.

Figure 7.1 provides the comparison among the proportions of the standard method versus Method 1 using criterion functions I-IV.

We now consider the issue of initial value choice. Notice that in Tables 7.1-7.4, no matter which criterion we use, the first step gives us the same estimates $(\hat{\beta}_0, \hat{\beta}_1) = (-6.42119, 0.028369)$. We use this estimate to be the initial value for Criterion V (standard method "glm" in R), and we also use $(0, 0)$ as initial value for Criterion V. Then we compare these two results by using stopping rule $\frac{|Deviance^{(m)} - Deviance^{(m+1)}|}{(|Deviance^{(m+1)}|+0.1)} < 10^{-8}$. In Tables 7.6 and 7.7, we present the detail of convergence for these two different initial values.

From Tables 7.6 and 7.7, we see that if we use the first step estimates $(\hat{\beta}_0, \hat{\beta}_1) =$

Figure 7.1: Comparison of Estimates Obtained by Using Different Criteria for Method 1 (Cornfield Data)

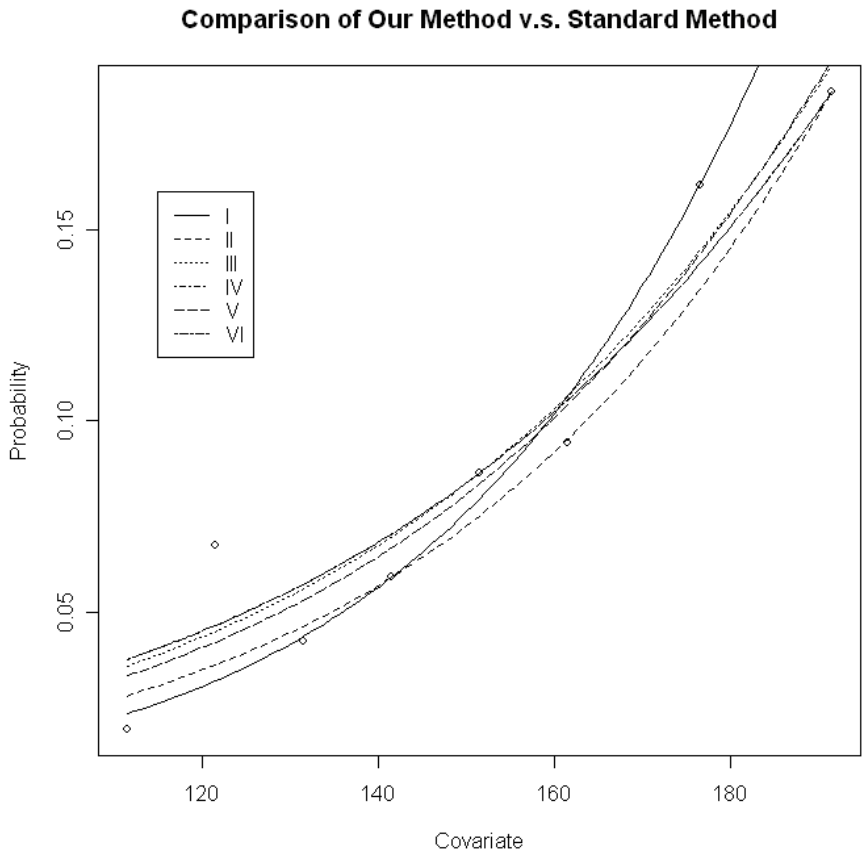


Table 7.6: Criterion V with (0,0) as Initial Value (Cornfield Data)

	$\hat{\beta}_0$	$\hat{\beta}_1$	Δ	$\tilde{\Delta}$	Δ^2	$\tilde{\Delta}^2$	$-2\log L$
Step 0	0.00000	0.000000	78963.75	480.10	972112351	28559.49	1236.28
Step 1	-2.71448	0.007153	14987.82	81.59	40217553	924.88	137.45
Step 2	-4.62996	0.016182	4095.01	23.61	3490647	83.49	47.56
Step 3	-5.83348	0.022904	2450.88	14.09	1120725	31.62	38.81
Step 4	-6.07510	0.024298	2322.85	14.11	1041768	30.78	38.61
Step 5	-6.08203	0.024338	2319.25	14.11	1040659	30.79	38.61
Step 6	-6.08203	0.024338	2319.25	14.11	1040658	30.79	38.61
Step 7	-6.08203	0.024338	2319.25	14.11	1040658	30.79	38.61

Table 7.7: Criterion V with (-6.42119,0.028369) as Initial Value (Cornfield Data)

	$\hat{\beta}_0$	$\hat{\beta}_1$	Δ	$\tilde{\Delta}$	Δ^2	$\tilde{\Delta}^2$	$-2\log L$
Step 0	-6.42119	0.028369	4936.55	43.03	3257850	387.77	45.32
Step 1	-6.13634	0.024872	2364.50	15.13	1055128	34.95	38.67
Step 2	-6.08310	0.024347	2319.66	14.12	1040552	30.81	38.61
Step 3	-6.08203	0.024338	2319.25	14.11	1040658	30.79	38.61
Step 4	-6.08203	0.024338	2319.25	14.11	1040658	30.79	38.61
Step 5	-6.08203	0.024338	2319.25	14.11	1040658	30.79	38.61

$(-6.42119, 0.028369)$ **obtained by using Method 1 to be the initial value, the convergence speed is somewhat faster than the initial value** $(0, 0)$.

7.2.2 Simulation Study

We use R to generate 100,000 observation vectors of $\boldsymbol{y}$ by assuming $\beta_0 = -6.5,\ \beta_1 = 0.02$, the same $\boldsymbol{n}$ and $\boldsymbol{z}$ as in the Cornfield data. For each vector $\boldsymbol{y}$, we apply Method 1. We record the mean and median values of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ for each criterion and obtain Table 7.8 and Table 7.9.

Table 7.8: Mean Value of 100,000 Pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$

Criterion	$\hat{\beta}_0$	$\hat{\beta}_1$
I	-6.52987	0.019725
II	-7.07977	0.022586
III	-6.51811	0.019860
IV	-7.27488	0.023735
V	-6.49773	0.019786
Truth	-6.5	0.02

We also record the descriptive statistics of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ for each criterion.

Table 7.9: Median Value of 100,000 Pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$

Criterion	$\hat{\beta}_0$	$\hat{\beta}_1$
I	-6.53243	0.020362
II	-6.59712	0.020468
III	-6.53230	0.020357
IV	-6.44183	0.019784
V	-6.48997	0.019928
Truth	-6.5	0.02

Consequently we obtain the histograms and tables. In Table 7.10, we present the descriptive statistics of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ for each criterion. In Figures 7.2-7.6, we present the corresponding histograms.

Table 7.10: Statistics of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ for 100,000 Simulations

Criterion	$\hat{\beta}_0$			$\hat{\beta}_1$		
	Q_1	Q_3	S.D.	Q_1	Q_3	S.D.
I	-7.51781	-5.49968	1.66702	0.012699	0.026712	0.011074
II	-8.06326	-5.20209	3.39466	0.011393	0.030594	0.019915
III	-7.33729	-5.54399	1.54122	0.013478	0.025973	0.010117
IV	-8.22197	-5.09072	4.05358	0.010321	0.030664	0.023151
V	-7.27609	-5.70537	1.18490	0.014604	0.025135	0.007963

Moreover, we use the counts of $\hat{\beta}_0$ and $\hat{\beta}_1$ in some chosen intervals to compare these different criteria and obtain Table 7.11 and Table 7.12.

Figure 7.2: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion I) (Method 1)

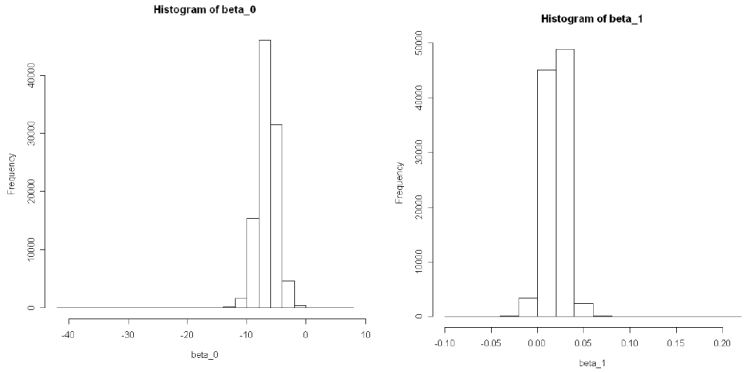


Figure 7.3: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion II) (Method 1)

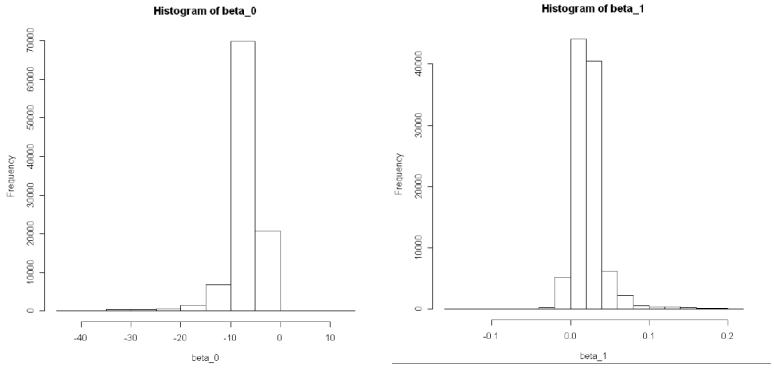


Figure 7.4: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion III) (Method 1)

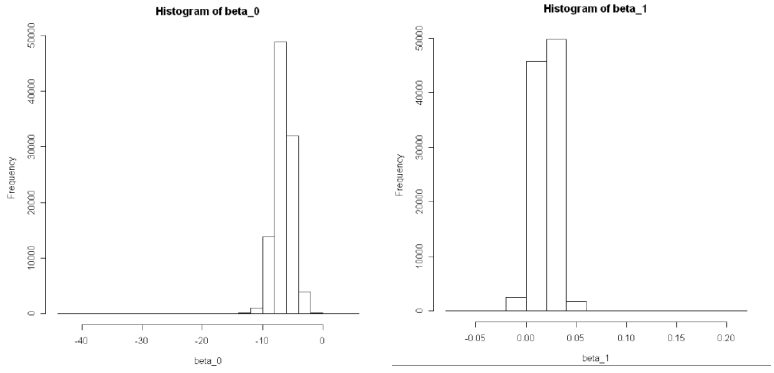


Figure 7.5: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion IV) (Method 1)

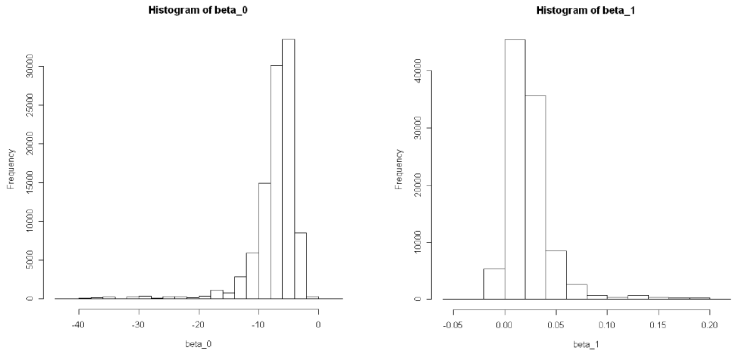


Figure 7.6: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion V) (Method 1)

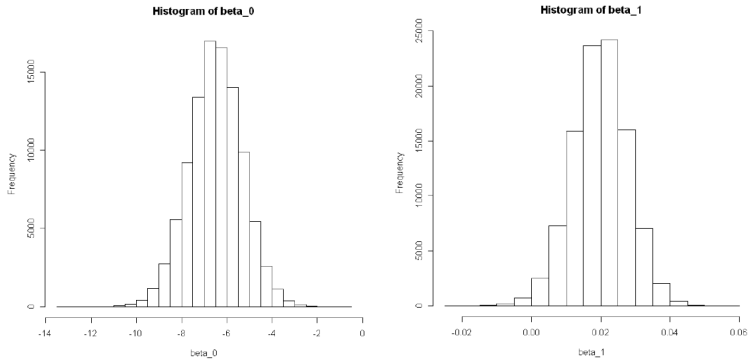


Table 7.11: Counts of 100,000 $\hat{\beta}_0$

Criterion	(-6.5-0.5,-6.5+0.5)	(-6.5-1.5,-6.5+1.5)	(-6.5-2.5,-6.5+2.5)
I	25027	66593	88807
II	20127	53241	77866
III	27517	70110	90726
IV	18528	48506	73050
V	33523	79980	96345

Table 7.12: Counts of 100,000 $\hat{\beta}_1$

Criterion	(0.02-0.005,0.02+0.005)	(0.02-0.01,0.02+0.01)	(0.02-0.015,0.02+0.015)
I	38596	66357	83190
II	30408	51688	71424
III	42124	70012	86775
IV	27801	48482	68334
V	47851	79705	93989

If, for some particular response vectors $\boldsymbol{y}$ out of 100,000 vectors, some criteria do not converge to one solution but fluctuate between several solutions, then we choose the solution giving us the smallest criterion value.

From all the above tables and figures we observe that our Method 1 for Criterion I and III are better than Criterion II and IV. This is consistent with our result in Section 7.2.1 and in Chapter 5. However, from Table 7.11 and Table 7.12, we observe the standard method performs better than Method 1. This motivates us to further improve our Method 1.

7.3 Method 2

In this section, we introduce our second method. We have two versions of Method 2. First, we introduce the first version, Method 2.1.

7.3.1 Method 2.1

We take $\left(\beta_0^{(0)}, \beta_1^{(0)}\right) = (-6.421191, 0.02836856)$ from the first iteration step estimate of Method 1. Then we plug this into (4.4) to obtain 28 pairs of $\left(\beta_0^{(1)}, \beta_1^{(1)}\right)$.

For each pair, we investigate the following equation (7.5) with $m = 0$.

$$\begin{cases} \sum_{i=1}^{8} y_i = \exp\left(\beta_0^{(m+1)}\right) \sum_{i=1}^{8} \frac{n_i \exp\left(\beta_1^{(m+1)} z_i\right)}{1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_i\right)} \\ \sum_{i=1}^{8} y_i z_i = \exp\left(\beta_0^{(m+1)}\right) \sum_{i=1}^{8} \frac{n_i z_i \exp\left(\beta_1^{(m+1)} z_i\right)}{1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_i\right)} \end{cases}$$

$$\Rightarrow \tag{7.5}$$

$$\begin{cases} \sum_{i=1}^{8} y_i = a_1 \sum_{i=1}^{8} \frac{n_i b^{z_i}}{1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_i\right)} \\ \sum_{i=1}^{8} y_i z_i = a_2 \sum_{i=1}^{8} \frac{n_i z_i b^{z_i}}{1+\exp\left(\beta_0^{(m)}+\beta_1^{(m)} z_i\right)} \end{cases}$$

We use $\left(\beta_0^{(0)}, \beta_1^{(0)}\right) = (-6.421191, 0.02836856)$ to plug in the denominator of (7.5) and use this pair's $\beta_1^{(1)}$ to plug in the numerator of (7.5) to get a_1 and a_2.

Now we have 28 pairs of $\left(\log a_1^{(1)}, \log a_2^{(1)}\right)$ which are based on all of 8 observations. Then we choose the pair giving us the smallest $\left|\log a_1^{(1)} - \log a_2^{(1)}\right|$. In our case, it is the 20th pair (Observations 4 and 6). Then we calculate two values of each criterion function by using $\left(\log a_1^{(1)}, \beta_1^{(1)}\right)$ and $\left(\log a_2^{(1)}, \beta_1^{(1)}\right)$ of the 20th pair. We select the smaller criterion which is $\left(\log a_2^{(1)}, \beta_1^{(1)}\right)$ in our case. Then we take this $\left(\log a_2^{(1)}, \beta_1^{(1)}\right)$ of the 20th pair as our $\left(\beta_0^{(1)}, \beta_1^{(1)}\right)$ for general N. And then we perform the iteration until it converges. The entire procedure is described step-by-step as follows:

1. Set initial value as $\left(\beta_0^{(0)}, \beta_1^{(0)}\right) = (-6.421191, 0.02836856)$ and iteration step as $m = 0$.

2. Plug $\left(\beta_0^{(0)},\ \beta_1^{(0)}\right)$ in (4.4) to obtain $M = 28$ pairs of $\left(\beta_0^{(1)},\ \beta_1^{(1)}\right)$.

3. For each pair in $M = 28$ pairs of $\left(\beta_0^{(1)},\ \beta_1^{(1)}\right)$, plug this pair's $\beta_1^{(1)}$ in the numerator of (7.5) and plug $\left(\beta_0^{(0)}, \beta_1^{(0)}\right)$ in the denominator of (7.5) to get a_1 and a_2.

4. Choose one pair from $M = 28$ pairs giving us the smallest $|\log a_1 - \log a_2|$.

5. Calculate two values of a criterion function (any criterion of four criteria defined in previous section) by using $\left(\log a_1, \beta_1^{(1)}\right)$ and $\left(\log a_2, \beta_1^{(1)}\right)$ from the pair chosen in Step 4. Select the one providing smaller criterion value.

6. Take the one chosen in Step 5 as $\left(\beta_0^{(1)},\ \beta_1^{(1)}\right)$ for general N. Perform the iteration process until the stopping rule is reached.

We use four different criteria at Step 5. Therefore, we have four different types of Method 2.1. We apply them to the Cornfield data using the same deviance stopping

rule as in previous sections. The results are given in Table 7.13.

Table 7.13: Method 2.1 (Cornfield Data)

Criterion	m	$\hat{\beta}_0$	$\hat{\beta}_1$	Δ	$\tilde{\Delta}$	Δ^2	$\tilde{\Delta}^2$	$-2\log L$
I	9	-6.25220	0.025461	2279.07	14.79	1028515	34.05	38.67
II	9	-6.25220	0.025461	2279.07	14.79	1028515	34.05	38.67
III	9	-6.25220	0.025461	2279.07	14.79	1028515	34.05	38.67
IV	9	-6.25220	0.025461	2279.07	14.79	1028515	34.05	38.67

In Table 7.13, we see that we have the same solution for Criteria I-IV.

7.3.2 Method 2.2

We still take $\left(\beta_0^{(0)}, \beta_1^{(0)}\right) = (-6.421191, 0.02836856)$ as initial value. Also, we plug this into (4.4) to obtain 28 pairs of $\left(\beta_0^{(1)}, \beta_1^{(1)}\right)$.

For each pair, we also investigate equation (7.5). This time we use this pair's $\beta_1^{(1)}$ to plug in the numerator of (7.5) and use this pair's $\left(\beta_0^{(1)}, \beta_1^{(1)}\right)$ to plug in the denominator of (7.5) to solve for a_1 and a_2.

Now we have 28 pairs of $\left(\log a_1^{(1)}, \log a_2^{(1)}\right)$ which are based on 8 observations. Then we choose the pair giving us the smallest $\left|\log a_1^{(1)} - \log a_2^{(1)}\right|$. In our case, it is the 18th pair (Observations 3 and 8) at the first step and the 20th pair (Observations 4 and 6) in the following steps. Then we calculate two values of a criterion function by using $\left(\log a_1^{(1)}, \beta_1^{(1)}\right)$ and $\left(\log a_2^{(1)}, \beta_1^{(1)}\right)$ of this chosen pair. We select the one with smaller criterion value and take that as our $\left(\beta_0^{(1)}, \beta_1^{(1)}\right)$ for general N. And then we perform the

iteration until it converges. The entire procedure is described step-by-step as follows:

1. Set initial value as $\left(\beta_0^{(0)}, \beta_1^{(0)}\right) = (-6.421191, 0.02836856)$ and iteration step as $m = 0$.

2. Plug $\left(\beta_0^{(0)},\ \beta_1^{(0)}\right)$ in (18) to obtain $M = 28$ pairs of $\left(\beta_0^{(1)},\ \beta_1^{(1)}\right)$.

3. For each pair in $M = 28$ pairs of $\left(\beta_0^{(1)},\ \beta_1^{(1)}\right)$, plug this pair's $\beta_1^{(1)}$ in the numerator of (7.5) and plug this pair's $\left(\beta_0^{(1)},\ \beta_1^{(1)}\right)$ in the denominator of (7.5) to solve for a_1 and a_2.

4. Choose one pair from $M = 28$ pairs giving us the smallest $|\log a_1 - \log a_2|$.

5. Calculate two values of a criterion function by using $\left(\log a_1, \beta_1^{(1)}\right)$ and $\left(\log a_2, \beta_1^{(1)}\right)$ of the pair chosen in Step 4. Select the one providing smaller criterion value.

6. Take the one chosen in Step 5 as $\left(\beta_0^{(1)},\ \beta_1^{(1)}\right)$ for general N. Perform the iteration process until the stopping rule is reached.

Similar as Method 2.1, we use four different criteria at Step 5. Therefore, we have four different types of Method 2.2. We apply them to the Cornfield data using the same deviance stopping rule as in previous sections. The results are given in Table 7.14.

Table 7.14: Method 2.2 (Cornfield Data)

Criterion	m	$\hat{\beta}_0$	$\hat{\beta}_1$	Δ	$\tilde{\Delta}$	Δ^2	$\tilde{\Delta}^2$	$-2\log L$
I	5	-6.25934	0.025439	2254.88	14.50	1027700	32.84	38.68
II	5	-6.25934	0.025439	2254.88	14.50	1027700	32.84	38.68
III	5	-6.25642	0.025448	2264.75	14.62	1027453	33.28	38.67
IV	5	-6.25934	0.025439	2254.88	14.50	1027700	32.84	38.68

In Table 7.14, we see that we have the same solution for Criterion I, II and IV. We observe that the solution of Criterion I, II and IV is better than that of Criterion III with respect to Δ and $\tilde{\Delta}$, and worse with respect to Δ^2 and $\tilde{\Delta}^2$.

We see that the solution of our Method 2 is very close to the solution of standard method, which indeed improves Method 1.

7.3.3 Simulation Study

We use R to generate 100,000 observation vectors of $\boldsymbol{y}$ by assuming $\beta_0 = -6.5,\ \beta_1 = 0.02$, the same $\boldsymbol{n}$ and $\boldsymbol{z}$ as in the Cornfield data. For each vector $\boldsymbol{y}$, we apply both versions of Method 2. We record exactly the same values as in simulation of Method 1 and present them in the following tables and figures.

Tables 7.15-7.19 and Figures 7.7-7.11 are simulation results for Method 2.1. Table 7.15 presents the mean value of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. Table 7.16 presents the median value of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. Table 7.17 presents the descriptive statistics of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. Table 7.18 presents the counts of 100,000 $\hat{\beta}_0$ in some chosen intervals centered at true value. Table 7.19 presents the counts of 100,000 $\hat{\beta}_1$ in some chosen intervals centered at true value. Figures 7.7-7.11 present the histograms of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$.

Table 7.15: Mean Value of 100,000 Pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Method 2.1)

Criterion	$\hat{\beta}_0$	$\hat{\beta}_1$
I	-6.49922	0.019791
II	-6.49932	0.019791
III	-6.49899	0.019791
IV	-6.49924	0.019791
V	-6.49894	0.019793
Truth	-6.5	0.02

Table 7.16: Median Value of 100,000 Pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Method 2.1)

Criterion	$\hat{\beta}_0$	$\hat{\beta}_1$
I	-6.50815	0.020269
II	-6.50829	0.020269
III	-6.50732	0.020269
IV	-6.50770	0.020268
V	-6.49424	0.019953
Truth	-6.5	0.02

Table 7.17: Statistics of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ for 100,000 Simulations (Method 2.1)

Criterion	$\hat{\beta}_0$			$\hat{\beta}_1$		
	Q_1	Q_3	S.D.	Q_1	Q_3	S.D.
I	-7.28656	-5.69561	1.200911	0.014336	0.025205	0.008086
II	-7.28673	-5.69588	1.200965	0.014336	0.025206	0.008086
III	-7.28656	-5.69504	1.201030	0.014341	0.025204	0.008086
IV	-7.28684	-5.69571	1.201047	0.014338	0.025204	0.008086
V	-7.27575	-5.71581	1.185476	0.014581	0.025162	0.007981

Figure 7.7: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion I) (Method 2.1)

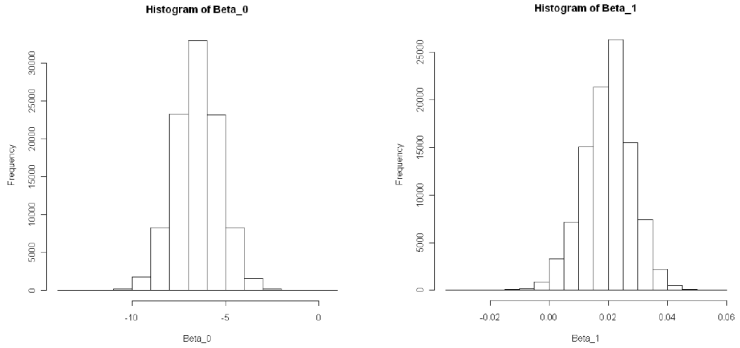


Figure 7.8: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion II) (Method 2.1)

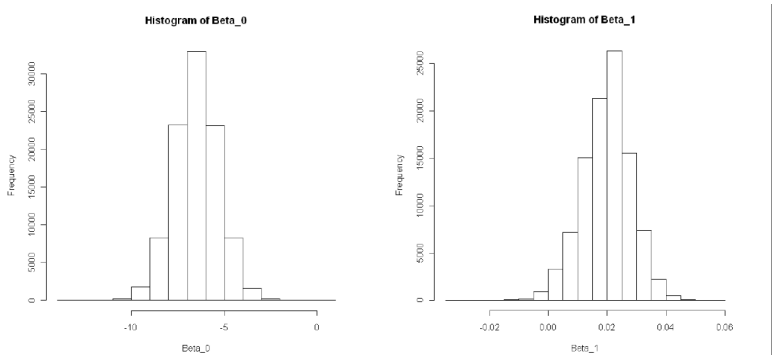


Figure 7.9: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion III) (Method 2.1)

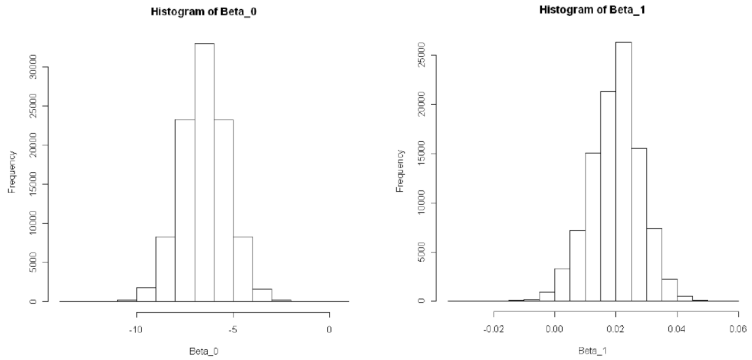


Figure 7.10: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion IV) (Method 2.1)

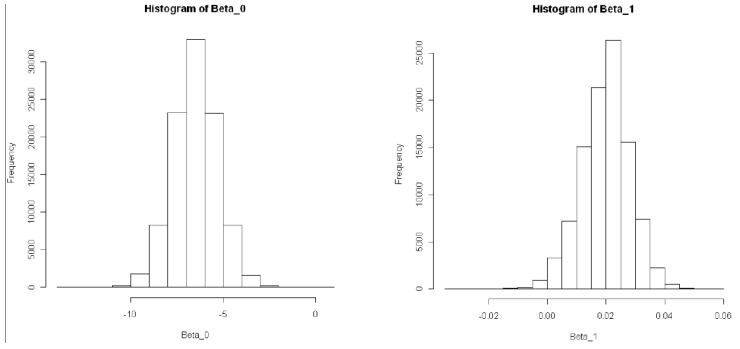


Figure 7.11: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion V) (Method 2.1)

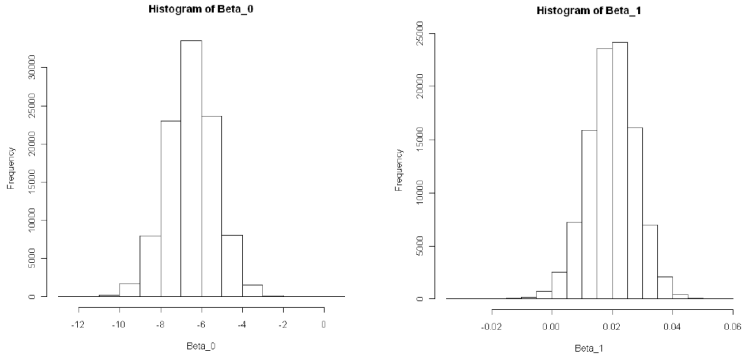


Table 7.18: Counts of 100,000 $\hat{\beta}_0$ (Method 2.1)

Criterion	(-6.5-0.5,-6.5+0.5)	(-6.5-1.5,-6.5+1.5)	(-6.5-2.5,-6.5+2.5)
I	32974	79393	96008
II	32979	79391	96007
III	32963	79397	96011
IV	32982	79385	96004
V	33506	80140	96266

Table 7.19: Counts of 100,000 $\hat{\beta}_1$ (Method 2.1)

Criterion	(0.02-0.005,0.02+0.005)	(0.02-0.01,0.02+0.01)	(0.02-0.015,0.02+0.015)
I	47709	78314	92877
II	47707	78309	92873
III	47701	78309	92876
IV	47708	78305	92874
V	47741	79702	93927

From the above simulation result, we observe that Method 2.1 performs very good. The mean and median values of $(\hat{\beta}_0, \hat{\beta}_1)$ are very close to the true value as well as the solution of standard method (Criterion V). The histograms of $(\hat{\beta}_0, \hat{\beta}_1)$ and counts of $(\hat{\beta}_0, \hat{\beta}_1)$ demonstrate Method 2.1 performs as good as the standard method.

Tables 7.20-7.24 and Figures 7.12-7.16 are simulation results for Method 2.2. Table 7.20 presents the mean value of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. Table 7.21 presents the median value of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. Table 7.22 presents the descriptive statistics of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$. Table 7.23 presents the counts of 100,000 $\hat{\beta}_0$ in some chosen intervals centered at true value. Table 7.24 presents the counts of 100,000 $\hat{\beta}_1$ in some chosen intervals centered at true value. Figures 7.12-7.16 present the histograms of 100,000 pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$.

Table 7.20: Mean Value of 100,000 Pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Method 2.2)

Criterion	$\hat{\beta}_0$	$\hat{\beta}_1$
I	-6.51384	0.019903
II	-6.51395	0.019903
III	-6.51365	0.019903
IV	-6.51372	0.019902
V	-6.50144	0.019807
Truth	-6.5	0.02

Table 7.21: Median Value of 100,000 Pairs of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Method 2.2)

Criterion	$\hat{\beta}_0$	$\hat{\beta}_1$
I	-6.51765	0.020307
II	-6.51765	0.020307
III	-6.51741	0.020308
IV	-6.51753	0.020307
V	-6.49651	0.019956
Truth	-6.5	0.02

Table 7.22: Statistics of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ for 100,000 Simulations (Method 2.2)

Criterion	$\hat{\beta}_0$			$\hat{\beta}_1$		
	Q_1	Q_3	S.D.	Q_1	Q_3	S.D.
I	-7.29808	-5.71425	1.211656	0.014667	0.025253	0.008169
II	-7.29793	-5.71418	1.211722	0.014667	0.025253	0.008169
III	-7.29802	-5.71394	1.211707	0.014668	0.025252	0.008168
IV	-7.29770	-5.71394	1.211704	0.014666	0.025252	0.008169
V	-7.27755	-5.71581	1.18619	0.014665	0.025135	0.007982

Figure 7.12: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion I) (Method 2.2)

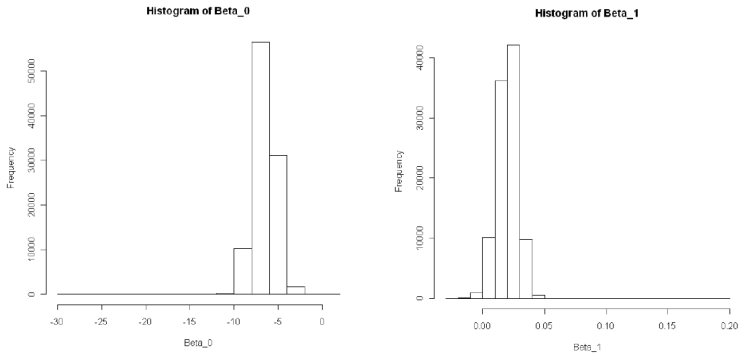


Figure 7.13: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion II) (Method 2.2)

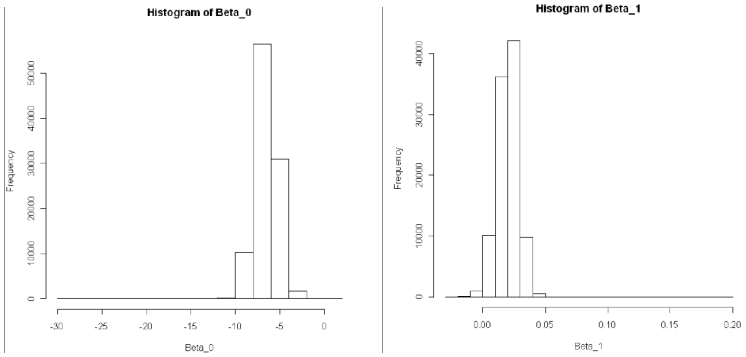


Figure 7.14: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion III) (Method 2.2)

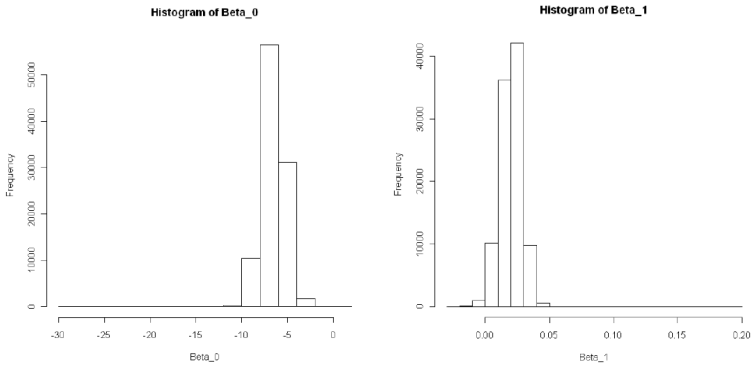


Figure 7.15: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion IV) (Method 2.2)

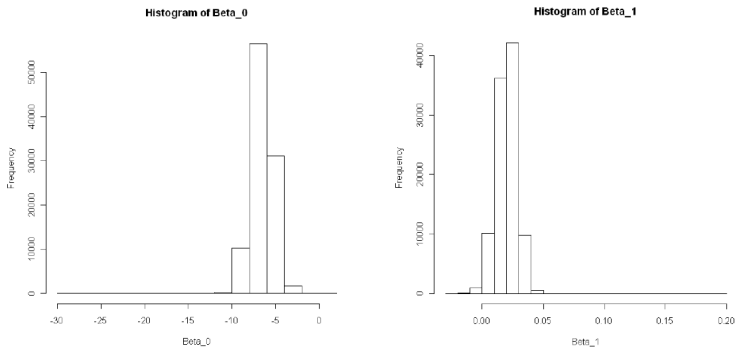


Table 7.23: Counts of 100,000 $\hat{\beta}_0$ (Method 2.2)

Criterion	(-6.5-0.5,-6.5+0.5)	(-6.5-1.5,-6.5+1.5)	(-6.5-2.5,-6.5+2.5)
I	33259	79338	96029
II	33267	79339	96030
III	33236	79327	96035
IV	33262	79339	96033
V	33640	80107	96330

Figure 7.16: Histogram of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ (Criterion V) (Method 2.2)

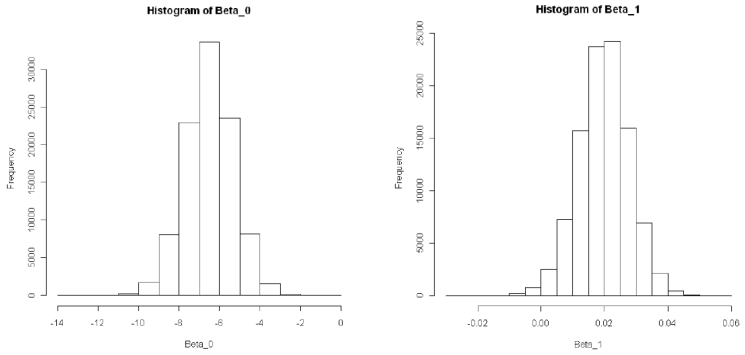


Table 7.24: Counts of 100,000 $\hat{\beta}_1$ (Method 2.2)

Criterion	(0.02-0.005,0.02+0.005)	(0.02-0.01,0.02+0.01)	(0.02-0.015,0.02+0.015)
I	48034	78332	92839
II	48035	78329	92840
III	48028	78334	92839
IV	48033	78330	92840
V	48021	79727	93859

From the above simulation result, we observe that Method 2.2 also performs very good. The mean and median values of $(\hat{\beta}_0, \hat{\beta}_1)$ are very close to the true value as well as the solution of standard method (Criterion V). The histograms of $(\hat{\beta}_0, \hat{\beta}_1)$ and counts of $(\hat{\beta}_0, \hat{\beta}_1)$ demonstrate Method 2.2 performs as good as the standard method. Moreover, from the histograms, we see that Method 2.2 does not perform well in some cases, but the count tables indicate this happens rarely.

Both Method 2 and standard method are iterative methods. One advantage of Method 2 is that we have exact expressions of $(\hat{\beta}_0, \hat{\beta}_1)$.

7.4 Comparison of Two Methods

We present Table 7.25 where we compare the estimates of parameters and the corresponding values of all criterion functions.

Table 7.25: Comparison of Method 1 and Method 2 (Cornfield Data)

Method 1 (I.V.)	m	$(\hat{\beta}_0, \hat{\beta}_1)$	Δ	$\tilde{\Delta}$	Δ^2	$\tilde{\Delta}^2$	$-2\log L$
C-I (0,0)	12	(-7.30647, 0.03207)	1958.68	17.97	1417676	121.78	41.56
C-II (0,0)	9	(-6.42721, 0.02586)	2006.12	12.10	1253788	42.00	39.88
C-III (0,0)	13	(-5.87891, 0.02323)	2448.91	14.39	1113477	31.48	38.80
C-IV (0,0)	12	(-5.70506, 0.02208)	2494.43	13.75	1175245	33.80	39.00
C-V (0,0)	7	(-6.08203, 0.02434)	2319.25	14.11	1040658	30.79	38.61
C-V (*)	5	(-6.08203, 0.02434)	2319.25	14.11	1040658	30.79	38.61
Method 2 (I.V.)	m	$(\hat{\beta}_0, \hat{\beta}_1)$	Δ	$\tilde{\Delta}$	Δ^2	$\tilde{\Delta}^2$	$-2\log L$
2.1: C-I-IV (0,0)	9	(-6.25220, 0.02546)	2279.07	14.79	1028516	34.05	38.67
2.2: C-I, II, IV (*)	5	(-6.25934, 0.02544)	2254.88	14.50	1027700	32.84	38.68
2.2: C-III (*)	5	(-6.25642, 0.02545)	2264.75	14.62	1027453	33.28	38.67

In Table 7.25, (*)=(-6.42119, 0.028369). "C" stands for "Criterion"."I.V." stands for "Initial Value". We observe that Method 2 improves Method 1 with respect to Δ^2 and $-2\log L$, and Method 2 also performs better than the standard method (Criterion V) with respect to Δ and Δ^2.

We select the first iteration step estimate of Method 1 which is (-6.42119, 0.028369) as the initial value for Method 2 or the standard method. Results illustrate that this initial value is closer to the final estimate and the number of iteration steps is smaller. It demonstrates that our Method 1 identifies a

better initial value than (0,0). A better initial value increases the efficiency of an iterative method. Method 2 gives us exact closed form solutions of $(\hat{\beta}_0, \hat{\beta}_1)$ **which standard method is unable to provide**.

Chapter 8

Probit Regression Model

8.1 Introduction

In this chapter, we consider the probit regression model for binomial data. Since standard method in obtaining the MLEs of parameters requires numerical method, the final parameter estimates do not have any closed form expressions. We introduce a new method by using a linear approximation which provides an exact solution of the MLEEs. We compare our estimates with the standard numerical method estimates with the Cornfield data as well as a simulated data, and find that our method performs as good as the standard method, even better with respect to some criteria. We also present some theoretical properties of our estimates.

8.2 Probit Regression Model

8.2.1 Model for General N

We now consider the binomial model and the same data shown in Table 3.1.

Using the same notation, we also have the likelihood for the jth group

$$f_j^* = \frac{n_j!}{y_j! \cdot (n_j - y_j)!} p_j^{y_j} \left(1 - p_j\right)^{n_j - y_j}. \tag{8.1}$$

Since we know that N covariate groups are independent, the likelihood of the whole data is

$$L^* = \prod_{j=1}^{N} f_j^*. \tag{8.2}$$

For the probit regression model, p_j is defined as

$$p_j = \Phi(\boldsymbol{\beta}' \boldsymbol{z}_j), \tag{8.3}$$

where $\Phi(\cdot)$ is the distribution function of standard normal random variable, and $\boldsymbol{\beta}$ is the vector of unknown parameters. We want to maximize the likelihood in (8.2) to obtain the estimates of $\boldsymbol{\beta}$.

Ignoring all factorial terms which do not include unknown parameters, the adjusted likelihood becomes

$$L_A = \prod_{j=1}^{N} f_j = \prod_{j=1}^{N} p_j^{y_j} \left(1 - p_j\right)^{n_j - y_j} = \prod_{j=1}^{N} \left[\frac{\Phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)}{1 - \Phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)}\right]^{y_j} \left(1 - \Phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)\right)^{n_j}. \tag{8.4}$$

The logarithm of adjusted likelihood is

$$l_A = \sum_{j=1}^{N} \log f_j = \sum_{j=1}^{N} \left\{ y_j \log \left[\frac{\Phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)}{1 - \Phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)}\right] + n_j \log \left[1 - \Phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)\right] \right\}. \tag{8.5}$$

Taking the derivative of (8.5) with respect to $\boldsymbol{\beta}$, we get the estimating equation (8.6)

$$\frac{\partial}{\partial \boldsymbol{\beta}} l_A = \sum_{j=1}^{N} \left(y_j \boldsymbol{z}_j \frac{\phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)}{\Phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)\left(1 - \Phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)\right)} - n_j \boldsymbol{z}_j \frac{\phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)}{1 - \Phi\left(\boldsymbol{\beta}' \boldsymbol{z}_j\right)} \right) = 0. \tag{8.6}$$

Solving this equation (8.6), we obtain the MLE of the unknown parameter vector $\boldsymbol{\beta}$.

For the Cornfield data introduced in Chapter 3, the estimating equation (8.6) becomes

$$\frac{\partial}{\partial \boldsymbol{\beta}} l_A = \sum_{j=1}^{N} \left(y_j \begin{pmatrix} 1 \\ z_j \end{pmatrix} \frac{\phi(\beta_0+\beta_1 z_j)}{\Phi(\beta_0+\beta_1 z_j)(1-\Phi(\beta_0+\beta_1 z_j))} - n_j \begin{pmatrix} 1 \\ z_j \end{pmatrix} \frac{\phi(\beta_0+\beta_1 z_j)}{(1-\Phi(\beta_0+\beta_1 z_j))} \right) = 0$$

$$\Rightarrow$$

$$\begin{cases} \frac{\partial l}{\partial \beta_0} = \sum_{j=1}^{N} \left(y_j \frac{\phi(\beta_0+\beta_1 z_j)}{\Phi(\beta_0+\beta_1 z_j)(1-\Phi(\beta_0+\beta_1 z_j))} - n_j \frac{\phi(\beta_0+\beta_1 z_j)}{(1-\Phi(\beta_0+\beta_1 z_j))} \right) = 0, \\ \frac{\partial l}{\partial \beta_1} = \sum_{j=1}^{N} \left(y_j z_j \frac{\phi(\beta_0+\beta_1 z_j)}{\Phi(\beta_0+\beta_1 z_j)(1-\Phi(\beta_0+\beta_1 z_j))} - n_j z_j \frac{\phi(\beta_0+\beta_1 z_j)}{(1-\Phi(\beta_0+\beta_1 z_j))} \right) = 0. \end{cases} \tag{8.7}$$

8.2.2 Estimates of Parameters for $N = 2$

As in Chapter 3, (8.7) is a transcendental equation and does not have a closed form solution and needs a numerical approximation for finding the solution. However, if we only consider a pair of 8 observations, i.e., $N = 2$, we have the exact closed form solution. For example, we consider the observations i and j. Denote the estimates of β_0 and β_1 by $\hat{\beta}_0$ and $\hat{\beta}_1$, and the predicted values of y_i and y_j by $\hat{y}_i$ and $\hat{y}_j$. Then equation (8.7) becomes

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 z_i = \Phi^{-1}\left(\frac{y_i}{n_i}\right), \\ \hat{\beta}_0 + \hat{\beta}_1 z_j = \Phi^{-1}\left(\frac{y_j}{n_j}\right). \end{cases} \tag{8.8}$$

which gives

$$\hat{y}_i = y_i,\ \hat{y}_j = y_j. \tag{8.9}$$

and the exact solution is

$$\begin{cases} \hat{\beta}_0 = \frac{z_i \Phi^{-1}\left(\frac{y_j}{n_j}\right) - z_j \Phi^{-1}\left(\frac{y_i}{n_i}\right)}{z_i - z_j}, \\ \hat{\beta}_1 = \frac{\Phi^{-1}\left(\frac{y_i}{n_i}\right) - \Phi^{-1}\left(\frac{y_j}{n_j}\right)}{z_i - z_j}. \end{cases} \tag{8.10}$$

In other words, for $N = 2$, (8.7) has an exact solution and does not need any numerical approximation.

8.3 Linear Approximation

In this section, we present a linear approximation method to solve the equation (8.7) which provides an exact solution of β_0 and β_1. We also have some good properties that the standard numerical method does not have. For our estimates of β_0 and β_1 using this linear approximation, we find the expectations.

We now present this approximation method. In the estimating equation (8.7), we write $x_j = \beta_0 + \beta_1 z_j$, and define

$$A(x_j) = \frac{\phi(x_j)}{\Phi(x_j) \cdot (1 - \Phi(x_j))} \tag{8.11}$$

and

$$B(x_j) = \frac{\phi(x_j)}{1 - \Phi(x_j)}. \tag{8.12}$$

Then (8.7) becomes

$$\begin{cases} \sum_{j=1}^{N} y_j A(x_j) - \sum_{j=1}^{N} n_j B(x_j) = 0, \\ \sum_{j=1}^{N} z_j y_j A(x_j) - \sum_{j=1}^{N} z_j n_j B(x_j) = 0. \end{cases} \tag{8.13}$$

The graphs of the two functions $A(x)$ and $B(x)$ are shown in Figure 8.1 and Figure 8.2 for the values of x satisfying $-3 \leq x \leq 3$.

We now approximate $A(x)$ on each side of $x = 0$ by two linear functions and similarly for $B(x)$. In other words, we approximate $A(x)$ and $B(x)$ by

$$A(x) = \frac{\phi(x)}{\Phi(x) \cdot (1 - \Phi(x))} = \begin{cases} \gamma_0 + \gamma_1 x, & \text{if } x > 0 \\ \gamma_0 - \gamma_1 x. & \text{if } x \leq 0 \end{cases} \tag{8.14}$$

Figure 8.1: Graph of Function $A(x)$

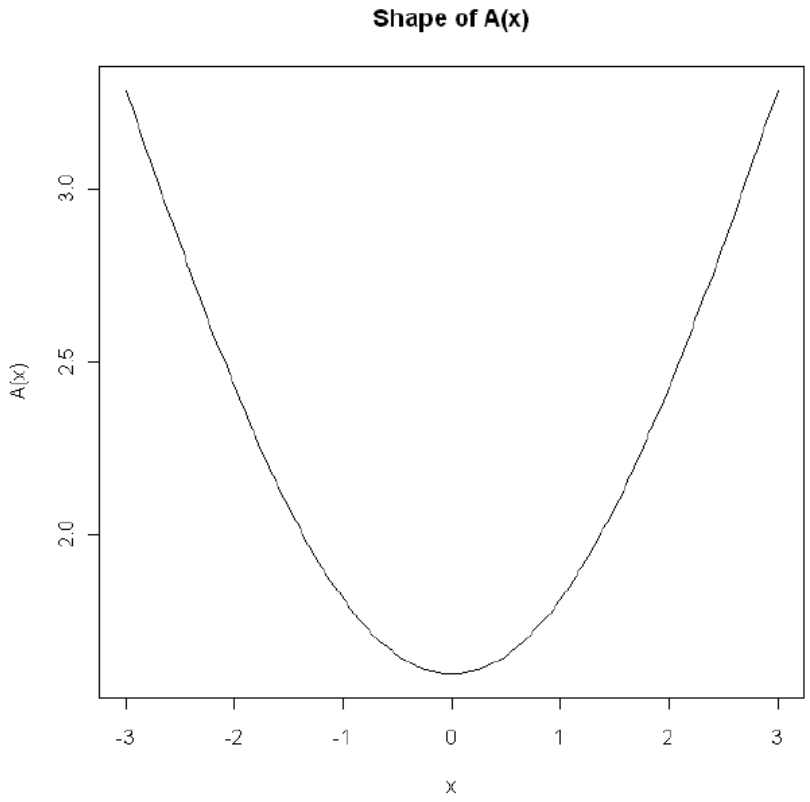
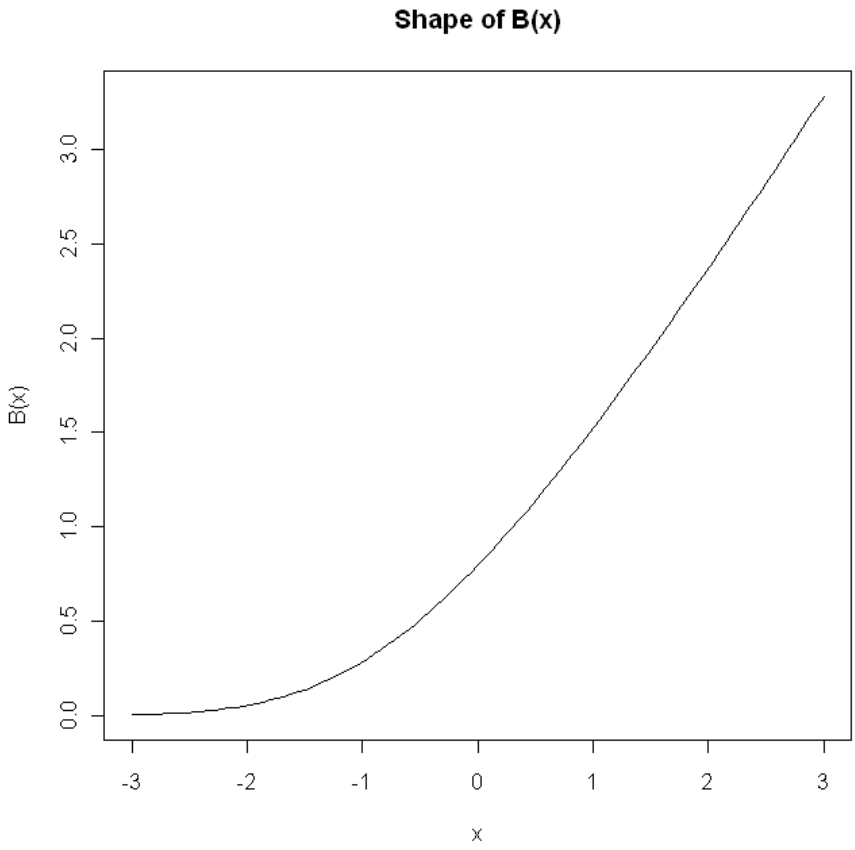


Figure 8.2: Graph of Function $B(x)$



and

$$B(x) = \frac{\phi(x)}{1 - \Phi(x)} = \begin{cases} \delta_0 + \delta_1 x, & \text{if } x > 0 \\ \delta_0 + \delta_2 x. & \text{if } x \leq 0 \end{cases} \tag{8.15}$$

To use (8.14) and (8.15) in (8.13), we need the signs of x_j's.

8.3.1 The Signs of x_j's

We now present a method to determine the signs of x_j's. The sample proportion $\frac{y_j}{n_j}$ is an estimator of $\Phi(x_j)$. We know that $\Phi(x_j) > 0.5$ when $x_j > 0$ and $\Phi(x_j) < 0.5$ when $x_j < 0$. Therefore, the method is as follows:

1. We take $x_j > 0$ if $\frac{y_j}{n_j} > 0.5$.

2. We take $x_j < 0$ if $\frac{y_j}{n_j} < 0.5$.

8.3.2 Values of γ_0, γ_1, δ_0, and δ_2

We also need the values of γ_0, γ_1, δ_0, and δ_2. We now present three methods to determine these values.

Method 1

Note that function $A(x)$ is symmetric about 0, we use two points $(0, A(0))$ and $(2, A(2))$ to fit the linear function. Similarly, we use two points $(0, B(0))$ and $(2, B(2))$ to fit the linear function when $x > 0$, and use two points $(0, B(0))$ and $(-2, B(-2))$ to fit the linear function when $x < 0$. Then we get $\gamma_0 = 1.5957691$, $\gamma_1 = 0.4163471$, $\delta_0 = 0.7978846$, $\delta_1 = 0.7876655$, $\delta_2 = 0.3713183$.

Method 2

1. Calculate $\hat{x}_j = \Phi^{(-1)}(\frac{y_j}{n_j})$ for all j's;

2. Find all N pairs of $(\hat{x}_j, A(\hat{x}_j))$ and locate their positions on the curve of $A(x)$. Use Least Squares Method to fit the line and get γ_0 and γ_1;

3. Find all N pairs of $(\hat{x}_j, B(\hat{x}_j))$ and locate their positions on the curve of $B(x)$. Use Least Squares Method to fit the line and get δ_0 and δ_2;

Method 3

1. Choose the uth pair containing (y_i, n_i) and (y_j, n_j), where $u = 1, \ldots, \frac{N(N-1)}{2}$;

2. Using this chosen pair to obtain $(\hat{\beta}_0^{(u)}, \hat{\beta}_1^{(u)})$ by using (8.10);

3. Calculate $\hat{x}_j^{(u)} = \hat{\beta}_0^{(u)} + \hat{\beta}_1^{(u)} z_j$ for all j's;

4. Find all N pairs of $(\hat{x}_j^{(u)}, A(\hat{x}_j^{(u)}))$ and locate their positions on the curve of $A(x)$. Use Least Squares Method to fit the line and get γ_0 and γ_1;

5. Find all N pairs of $(\hat{x}_j^{(u)}, B(\hat{x}_j^{(u)}))$ and locate their positions on the curve of $B(x)$. Use Least Squares Method to fit the line and get δ_0 and δ_2;

6. Repeat 1-6 until all $\begin{pmatrix} N \\ 2 \end{pmatrix} = \frac{N(N-1)}{2}$ pairs are calculated;

7. Choose the best pair according to a specified criterion.

8.3.3 Analysis of Real Example

We now use our method to analyze the Cornfield data. For the Cornfield data, we observe that all $\frac{y_j}{n_j}$'s are far less than 0.5, thus we have $x_j < 0$ for all j's. Then (8.13)

becomes

$$\begin{cases} \sum_{j=1}^{N} y_j \left(\gamma_0 - \gamma_1 \cdot (\beta_0 + \beta_1 z_j)\right) - \sum_{j=1}^{N} n_j \left(\delta_0 + \delta_2 \cdot (\beta_0 + \beta_1 z_j)\right) = 0, \\ \sum_{j=1}^{N} z_j y_j \left(\gamma_0 - \gamma_1 \cdot (\beta_0 + \beta_1 z_j)\right) - \sum_{j=1}^{N} z_j n_j \left(\delta_0 + \delta_2 \cdot (\beta_0 + \beta_1 z_j)\right) = 0. \end{cases}$$

$\Rightarrow$

$$\begin{pmatrix} \left(\gamma_1 \sum_{j=1}^{N} y_j + \delta_2 \sum_{j=1}^{N} n_j\right) & \left(\gamma_1 \sum_{j=1}^{N} y_j z_j + \delta_2 \sum_{j=1}^{N} n_j z_j\right) \\ \left(\gamma_1 \sum_{j=1}^{N} y_j z_j + \delta_2 \sum_{j=1}^{N} n_j z_j\right) & \left(\gamma_1 \sum_{j=1}^{N} y_j z_j^2 + \delta_2 \sum_{j=1}^{N} n_j z_j^2\right) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \tag{8.16}$$

$$= \begin{pmatrix} \gamma_0 \sum_{j=1}^{N} y_j - \delta_0 \sum_{j=1}^{N} n_j \\ \gamma_0 \sum_{j=1}^{N} y_j z_j - \delta_0 \sum_{j=1}^{N} n_j z_j \end{pmatrix}$$

Equation (8.16) gives us

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \left(\gamma_1 \sum_{j=1}^{N} y_j + \delta_2 \sum_{j=1}^{N} n_j\right) & \left(\gamma_1 \sum_{j=1}^{N} y_j z_j + \delta_2 \sum_{j=1}^{N} n_j z_j\right) \\ \left(\gamma_1 \sum_{j=1}^{N} y_j z_j + \delta_2 \sum_{j=1}^{N} n_j z_j\right) & \left(\gamma_1 \sum_{j=1}^{N} y_j z_j^2 + \delta_2 \sum_{j=1}^{N} n_j z_j^2\right) \end{pmatrix}^{-1} \begin{pmatrix} \gamma_0 \sum_{j=1}^{N} y_j - \delta_0 \sum_{j=1}^{N} n_j \\ \gamma_0 \sum_{j=1}^{N} y_j z_j - \delta_0 \sum_{j=1}^{N} n_j z_j \end{pmatrix}$$

$$= \frac{\begin{pmatrix} \left(\gamma_1 \sum_{j=1}^{N} y_j z_j^2 + \delta_2 \sum_{j=1}^{N} n_j z_j^2\right) & -\left(\gamma_1 \sum_{j=1}^{N} y_j z_j + \delta_2 \sum_{j=1}^{N} n_j z_j\right) \\ -\left(\gamma_1 \sum_{j=1}^{N} y_j z_j + \delta_2 \sum_{j=1}^{N} n_j z_j\right) & \left(\gamma_1 \sum_{j=1}^{N} y_j + \delta_2 \sum_{j=1}^{N} n_j\right) \end{pmatrix} \begin{pmatrix} \gamma_0 \sum_{j=1}^{N} y_j - \delta_0 \sum_{j=1}^{N} n_j \\ \gamma_0 \sum_{j=1}^{N} y_j z_j - \delta_0 \sum_{j=1}^{N} n_j z_j \end{pmatrix}}{\left(\gamma_1 \sum_{j=1}^{N} y_j + \delta_2 \sum_{j=1}^{N} n_j\right)\left(\gamma_1 \sum_{j=1}^{N} y_j z_j^2 + \delta_2 \sum_{j=1}^{N} n_j z_j^2\right) - \left(\gamma_1 \sum_{j=1}^{N} y_j z_j + \delta_2 \sum_{j=1}^{N} n_j z_j\right)^2} \tag{8.17}$$

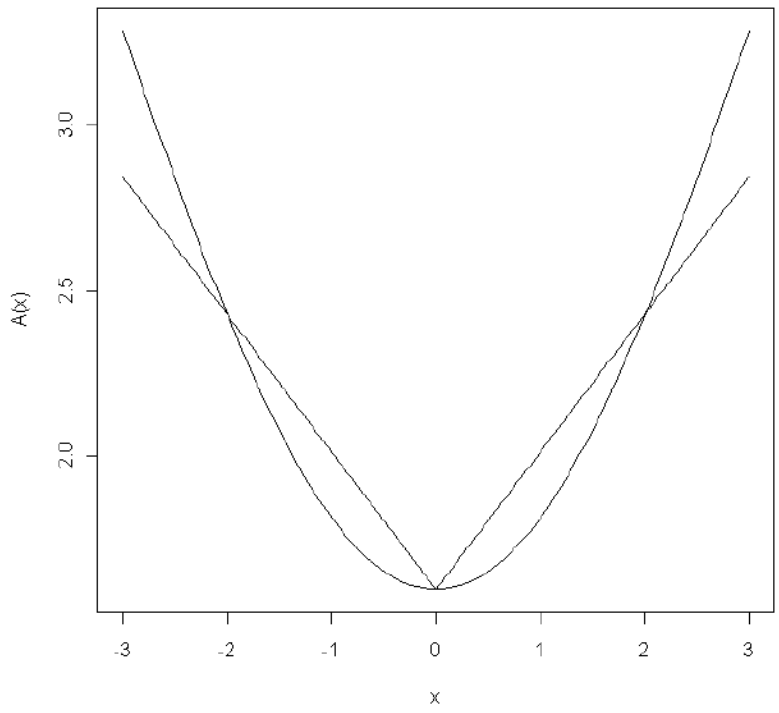
The solutions $\hat{\beta}_0$ and $\hat{\beta}_1$ in (8.17) are the estimators of β_0 and β_1 by our proposed method.

Since we have three methods to obtain the values of γ_0, γ_1, δ_0, and δ_2, we present three results of the analysis.

Method 1

We use (8.17) to get $\hat{\beta}_0 = -3.09881$ and $\hat{\beta}_1 = 0.009909$. Figure 8.3 and 8.4 demonstrate the fit of function $A(x)$ and $B(x)$. Figure 8.5 demonstrates the comparison of our method 1 with the standard method.

Figure 8.3: Fit of Function $A(x)$ of Method 1



Method 2

We use (8.17) to get $\hat{\beta}_0 = -3.24837$ and $\hat{\beta}_1 = 0.012232$. Figure 8.6 and 8.7 demonstrate the fit of function $A(x)$ and $B(x)$. Figure 8.8 demonstrates the comparison of our method 2 with the standard method.

Before we present the result of Method 3, we present our list of criterion functions. We

Figure 8.4: Fit of Function $B(x)$ of Method 1

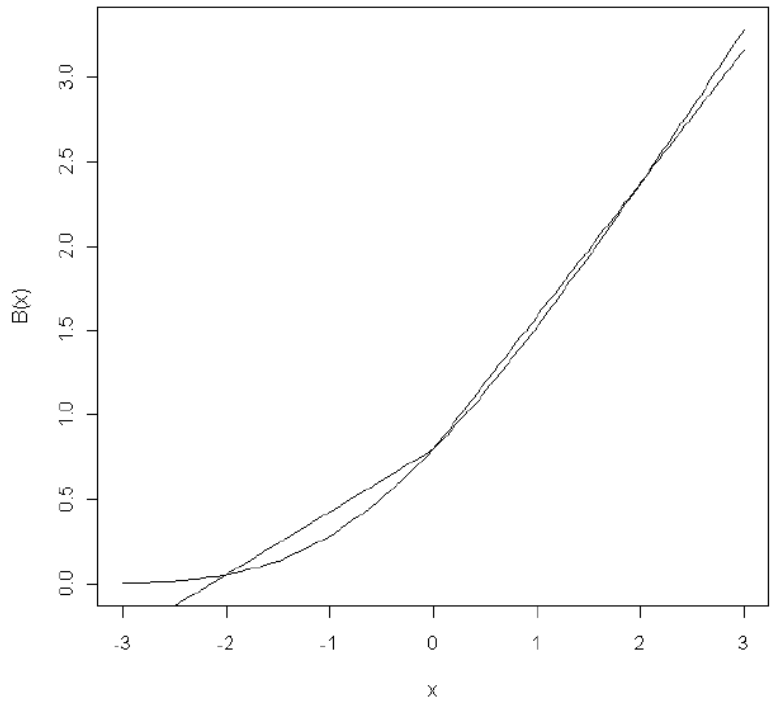


Figure 8.5: Comparison of Standard Method with Method 1 (Cornfield Data)

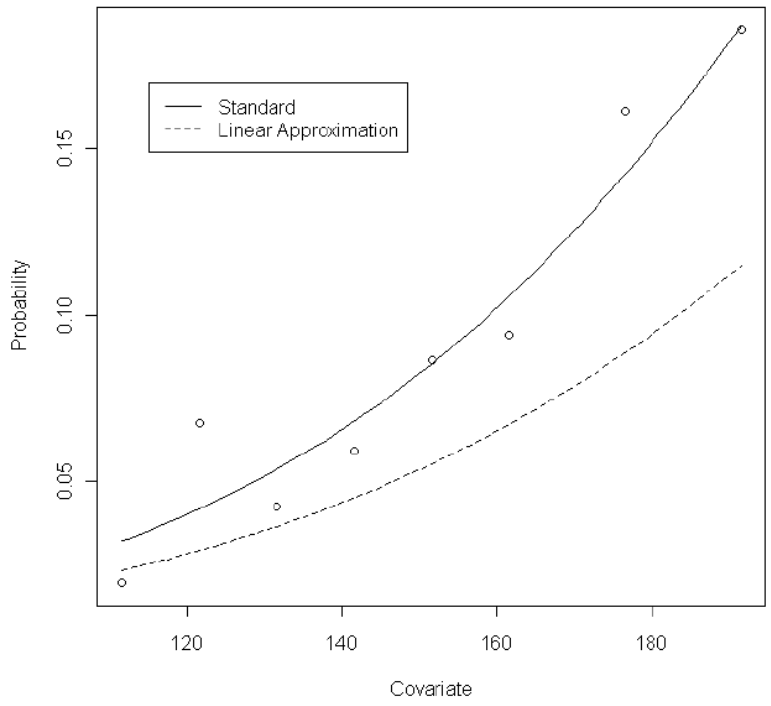


Figure 8.6: Fit of Function $A(x)$ of Method 2 (Cornfield Data)

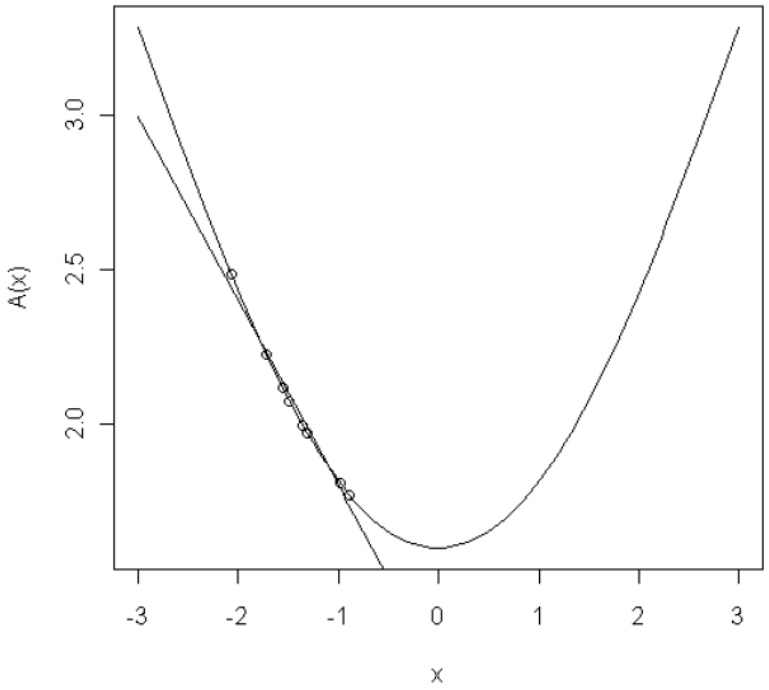


Figure 8.7: Fit of Function $B(x)$ of Method 2 (Cornfield Data)

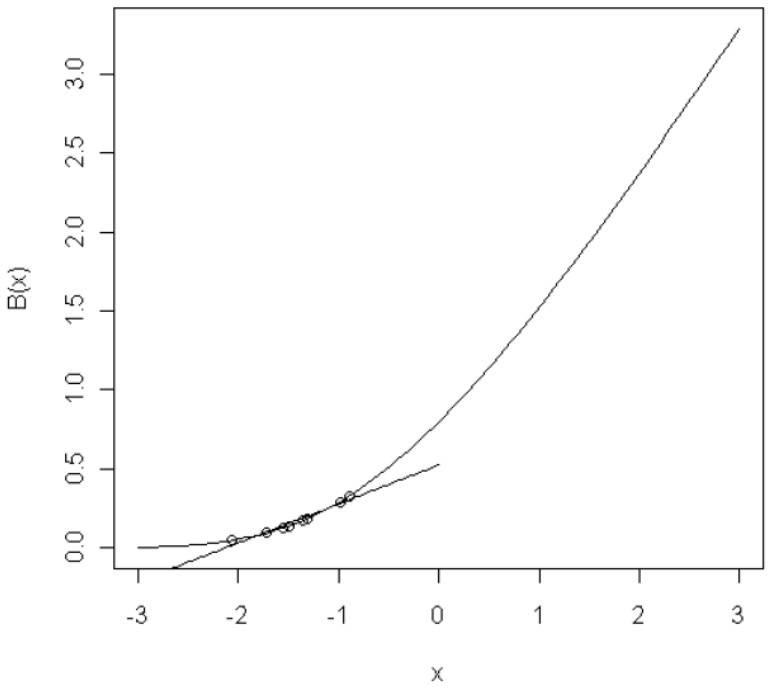
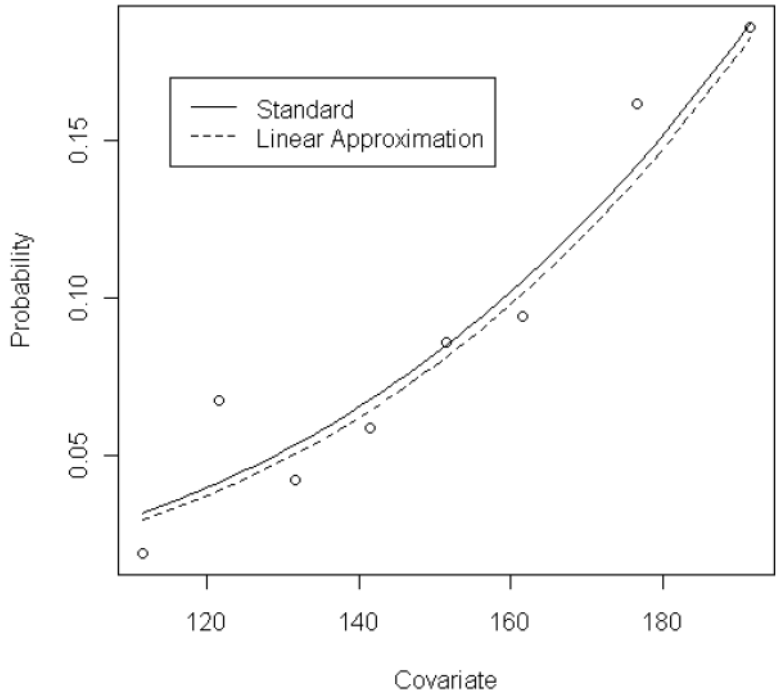


Figure 8.8: Comparison of Standard Method with Method 2 (Cornfield Data)



are interested in Δ, $-2\log L$, Deviance(D), Pearson's χ^2 and the goodness of fit for $A(x)$ and $B(x)$. The definitions are

$$\Delta = \sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{8.18}$$

and

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right], \tag{8.19}$$

and

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i\,(1-\hat{p}_i)}, \tag{8.20}$$

and

$$A^{(f)} = \sum_{i=1}^{N} |A_i - \hat{A}_i|, \tag{8.21}$$

and

$$B^{(f)} = \sum_{i=1}^{N} |B_i - \hat{B}_i|, \tag{8.22}$$

where $\hat{y}_i = n_i\hat{p}_i = n_i\Phi(\hat{\beta}_0 + \hat{\beta}_1 z_i)$, A_i and B_i are the values on the curve, $\hat{A}_i$ and $\hat{B}_i$ are the corresponding values of the fitted line.

Method 3

Method 3 for the Cornfield data gives $\frac{8(8-1)}{2} = 28$ sets of values of the above criteria. We find the 1^{st} pair containing observation (y_1, n_1) and (y_2, n_2) (denoted by (1,2) for simplicity) is unusually bad. From the remaining 27 results, by selecting the minimum $-2\log L$, D, Pearson's χ^2 and Δ, we find that the (5,8) pair gives us the best result. We have $\hat{\beta}_0 = -3.11237$ and $\hat{\beta}_1 = 0.011427$. Figure 8.9 and 8.10 demonstrate the fit of function $A(x)$ and $B(x)$ for (5,8) pair. Figure 8.11 demonstrates the comparison of our method 3 for (5,8) pair with the standard method.

Figure 8.9: Fit of Function $A(x)$ of Method 3 (Pair (5,8)) (Cornfield Data)

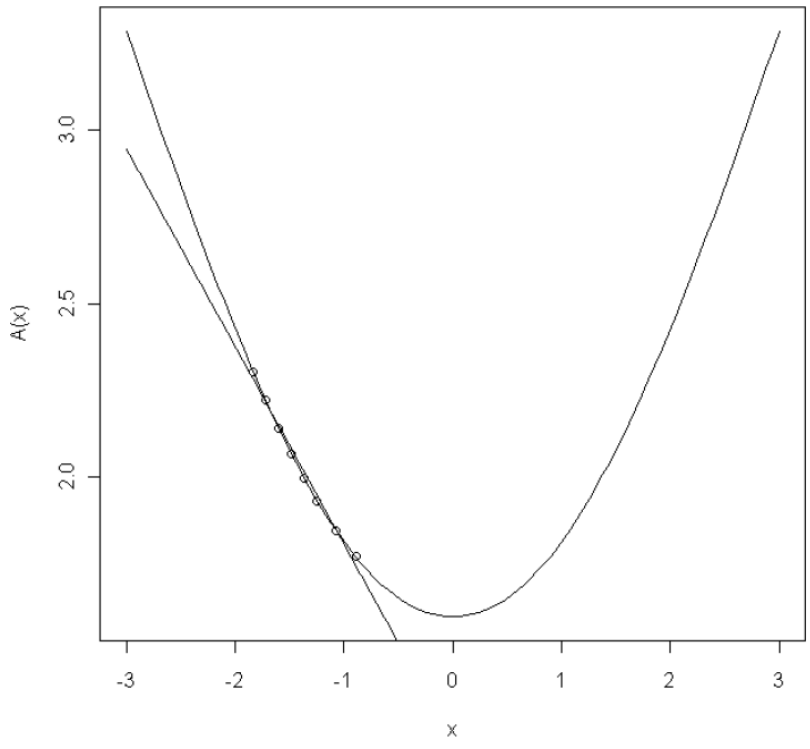
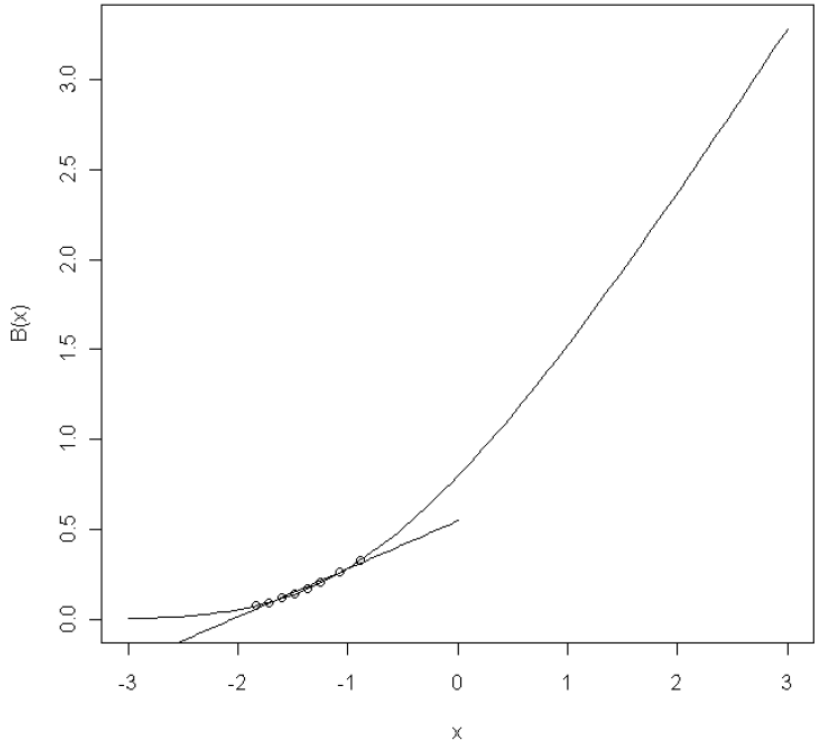


Figure 8.10: Fit of Function $B(x)$ of Method 3 (Pair (5,8)) (Cornfield Data)



We now present Table 8.1 where we compare the estimates of β_0, β_1 and the corresponding values of those criteria for these three different methods.

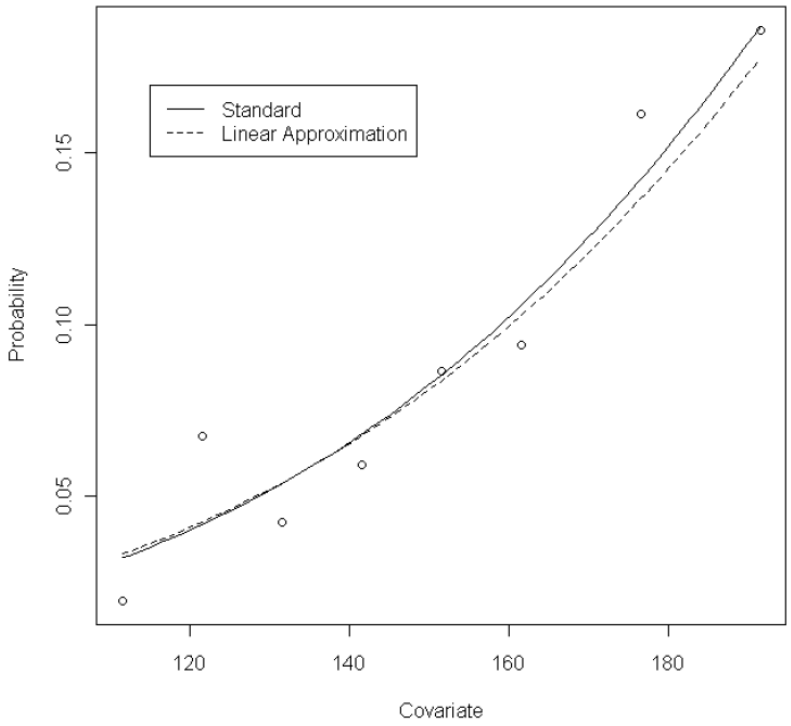
Note: In Table 8.1, SM$(\cdot)$ stands for "Standard Method $(\cdot)$" where we minimize the criterion function "$\cdot$" to get estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

From Table 8.1, we see that Method 2 and Method 3 for the majority of pairs perform as good as the standard method by all criterion functions, even better with respect to some criteria. The strength of our methods is that we have the exact closed form solution of the MLEEs by using the proposed linear approximation.

Table 8.1: Comparison of Standard Method with Our Linear Approximation Methods (Cornfield Data)

Method	$(\hat{\beta}_0, \hat{\beta}_1)$	Δ	$-2\log L$	D	χ^2	$A^{(f)}$	$B^{(f)}$
SM($-2\log L$)	(-3.19699, 0.012053)	17.17	38.76	6.06	6.51	NA	NA
SM(D)	(-3.19699, 0.012053)	17.17	38.76	6.06	6.51	NA	NA
SM(χ^2)	(-3.07957, 0.011359)	18.63	38.94	6.24	6.31	NA	NA
M1	(-3.09881, 0.009909)	32.76	54.28	21.58	27.14	0.82347	0.58769
M2	(-3.24837, 0.012232)	16.45	38.99	6.28	7.14	0.12307	0.09250
M3 (1,3)	(-3.19863, 0.011734)	16.86	39.56	6.86	7.90	0.16925	0.12601
M3 (1,4)	(-3.21296, 0.011843)	16.72	39.50	6.80	7.85	0.15920	0.11949
M3 (1,5)	(-3.18117, 0.011601)	17.02	39.64	6.93	7.96	0.18396	0.13678
M3 (1,6)	(-3.29146, 0.012456)	15.98	39.24	6.54	7.61	0.14087	0.10589
M3 (1,7)	(-3.22376, 0.011926)	16.62	39.46	6.76	7.82	0.15697	0.11785
M3 (1,8)	(-3.30747, 0.012584)	15.82	39.20	6.50	7.58	0.13933	0.10515
M3 (2,3)	(-4.21542, 0.019536)	33.44	49.83	17.12	18.50	0.89010	0.66147
M3 (2,4)	(-3.37477, 0.013493)	19.21	39.43	6.73	7.16	0.23688	0.17809
M3 (2,5)	(-3.06176, 0.011294)	19.23	39.09	6.38	6.34	0.11085	0.08319
M3 (2,6)	(-3.05797, 0.011267)	19.26	39.09	6.39	6.34	0.10984	0.08244
M3 (2,7)	(-2.89590, 0.010158)	20.70	39.63	6.93	6.61	0.11042	0.08237
M3 (2,8)	(-2.91557, 0.010291)	20.52	39.54	6.84	6.55	0.10660	0.07965
M3 (3,4)	(-3.21355, 0.011883)	16.74	39.35	6.65	7.63	0.15056	0.11304
M3 (3,5)	(-3.18690, 0.011616)	16.95	39.76	7.06	8.15	0.19488	0.14483
M3 (3,6)	(-3.25442, 0.012281)	16.39	38.97	6.27	7.13	0.12391	0.09346
M3 (3,7)	(-3.21115, 0.011859)	16.76	39.38	6.68	7.67	0.15237	0.11439
M3 (3,8)	(-3.25056, 0.012244)	16.43	39.00	6.30	7.16	0.12402	0.09320
M3 (4,5)	(-3.21365, 0.011682)	16.94	40.47	7.77	9.18	0.24578	0.18227
M3 (4,6)	(-3.22257, 0.012127)	16.76	38.84	6.13	6.81	0.11426	0.08607
M3 (4,7)	(-3.21334, 0.011871)	16.73	39.39	6.69	7.70	0.15317	0.11499
M3 (4,8)	(-3.21913, 0.012057)	16.76	38.92	6.22	6.96	0.11687	0.08780
M3 (5,6)	(-3.06344, 0.011298)	19.16	39.07	6.37	6.33	0.10904	0.08184
M3 (5,7)	(-3.14759, 0.011517)	17.45	39.08	6.38	7.07	0.12973	0.09711
M3 (5,8)	(-3.11237, 0.011427)	17.93	38.83	6.13	6.50	0.10474	0.07852
M3 (6,7)	(-3.48619, 0.013330)	17.18	42.71	10.01	12.84	0.27785	0.20882
M3 (6,8)	(-3.26623, 0.012337)	16.26	39.05	6.34	7.27	0.12752	0.09619
M3 (7,8)	(-2.78598, 0.009564)	22.98	40.78	8.07	7.32	0.13946	0.10386

Figure 8.11: Comparison of Standard Method with Method 3 (Pair (5,8)) (Cornfield Data)



For Cornfield data, we now perform a detailed study of Method 1. We use two points $(0, A(0))$ and $(x_0, A(x_0))$ to fit a line and we also use two points $(0, B(0))$ and $(x_2, B(x_2))$ to fit another line when $x < 0$. However, instead of using $x_0 = 2$ and $x_2 = -2$, this time we search for x_0, x_2 within the range $(1.5, 2.5)$ and $(-2.5, -1.5)$. We use step=0.2 to find γ_1 and δ_2 for every pair of x_0, x_2, then we use equation (8.17) to obtain the estimates of β_0 and β_1. We want to find how our choice of x_0, x_2 will influence the estimates of β_0 and β_1. Table 8.2 gives the details.

In Table 8.2, we observe the following:

1. When the absolute value of x is increasing, the estimates of both β_0 and β_1 are closer to the standard method estimates. In terms of the closeness of our estimates of β_0 and β_1 to the standard method estimates, the better absolute values of x_0

Table 8.2: Region Search of Linear Approximation Method 1 (Cornfield Data)

	$x_2 = -1.5$	$x_2 = -1.7$	$x_2 = -1.9$
$x_0 = 1.5$	(-2.57564, 0.007811)	(-2.75709, 0.008423)	(-2.95857, 0.009111)
$x_0 = 1.7$	(-2.58643, 0.007948)	(-2.76903, 0.008577)	(-2.97179, 0.009283)
$x_0 = 1.9$	(-2.59622, 0.008075)	(-2.77982, 0.008718)	(-2.98369, 0.009441)
$x_0 = 2.1$	(-2.60503, 0.008191)	(-2.78949, 0.008847)	(-2.99431, 0.009584)
$x_0 = 2.3$	(-2.61288, 0.008295)	(-2.79808, 0.008963)	(-3.00371, 0.009713)
$x_0 = 2.5$	(-2.61984, 0.008390)	(-2.80567, 0.009068)	(-3.01197, 0.009829)
	$x_2 = -2.1$	$x_2 = -2.3$	$x_2 = -2.5$
$x_0 = 1.5$	(-3.17951, 0.009874)	(-3.41856, 0.010710)	(-3.67369, 0.011613)
$x_0 = 1.7$	(-3.19414, 0.010067)	(-3.43472, 0.010925)	(-3.69146, 0.011854)
$x_0 = 1.9$	(-3.20725, 0.010243)	(-3.44911, 0.011122)	(-3.70719, 0.012072)
$x_0 = 2.1$	(-3.21889, 0.010403)	(-3.46182, 0.011300)	(-3.72100, 0.012270)
$x_0 = 2.3$	(-3.22914, 0.010547)	(-3.47295, 0.011460)	(-3.73302, 0.012446)
$x_0 = 2.5$	(-3.23811, 0.010675)	(-3.48265, 0.011602)	(-3.74343, 0.012604)

and x_2 are close to 2.

2. As the value of $\hat{\beta}_0$ increases with x, the corresponding value of $\hat{\beta}_1$ decreases.

Finally, we consider the initial value choice for the standard method. Method 3 gives us the best pair (5,8). We use the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained from the best pair as the initial value for the standard method. Table 8.3 gives the result.

In Table 8.3, we see that our linear approximation also provides a better initial value for the standard method with respect to the number of iteration steps. It is a substantial improvement on the efficiency of the standard method.

Table 8.3: Different Initial Values for Standard Method (Cornfield Data)

Method	Initial Value	m	$(\hat{\beta}_0, \hat{\beta}_1)$
SM $(-2\log L)$	(0,0)	6	(-3.19699, 0.012053)
	(-3.11237, 0.011427)	3	(-3.19699, 0.012053)
SM (D)	(0,0)	6	(-3.19699, 0.012053)
	(-3.11237, 0.011427)	3	(-3.19699, 0.012053)
SM (χ^2)	(0,0)	8	(-3.07957, 0.011359)
	(-3.11237, 0.011427)	2	(-3.07957, 0.011359)

8.3.4 Simulation Study

We now present a simulation study of our approximation method. Let $\beta_0 = -3.2$, $\beta_1 = 0.01$ and we use the same z_i's and n_i's as in Cornfield data shown in Table 3.2. For each covariate group corresponding to z_i, we use R to randomly generate a y_i based on probit regression assumption. Then we use both our linear approximation method and the standard method to obtain the estimates of β_0 and β_1. We do this for T times and compare the results. Table 8.4 shows the simulation result for Method 2.

Table 8.4: Simulation of Method 2

			$T = 25,000$	$T = 50,000$	$T = 100,000$
Linear Approximation	$\hat{\beta}_0$	Mean	-3.25569	-3.26397	-3.26209
		Median	-3.20971	-3.22060	-3.21683
	$\hat{\beta}_1$	Mean	0.010134	0.010188	0.010170
		Median	0.009898	0.009964	0.009939
Standard Method	$\hat{\beta}_0$	Mean	-3.20167	-3.20786	-3.20616
		Median	-3.19632	-3.20577	-3.20116
	$\hat{\beta}_1$	Mean	0.009966	0.010007	0.009991
		Median	0.009980	0.010034	0.010005

Note: The true values of β_0 and β_1 are $\beta_0 = -3.2$ and $\beta_1 = 0.01$.

We now present the histograms of the bias of $\hat{\beta}_0$ and $\hat{\beta}_1$ for both our proposed linear approximation method and the standard method.

Figure 8.12: Histogram of Bias of $\hat{\beta}_0$ of Simulation for Method 2 ($T = 25,000$)

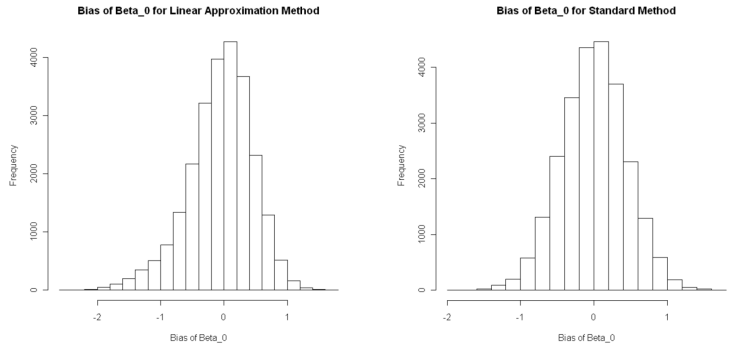
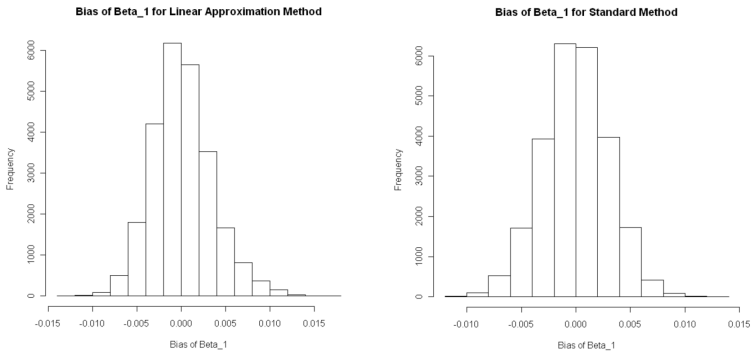


Figure 8.13: Histogram of Bias of $\hat{\beta}_1$ of Simulation for Method 2 ($T = 25,000$)



Moreover, we present in Table 8.5 the counts of estimates falling in some chosen intervals centered at the true values for both methods.

In Table 8.4, Table 8.5 and Figures 8.12-8.17, we see that for the simulated

Figure 8.14: Histogram of Bias of $\hat{\beta}_0$ of Simulation for Method 2 ($T = 50,000$)

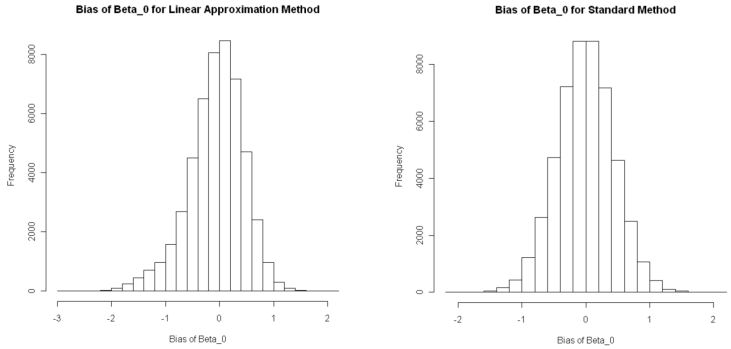


Figure 8.15: Histogram of Bias of $\hat{\beta}_1$ of Simulation for Method 2 ($T = 50,000$)

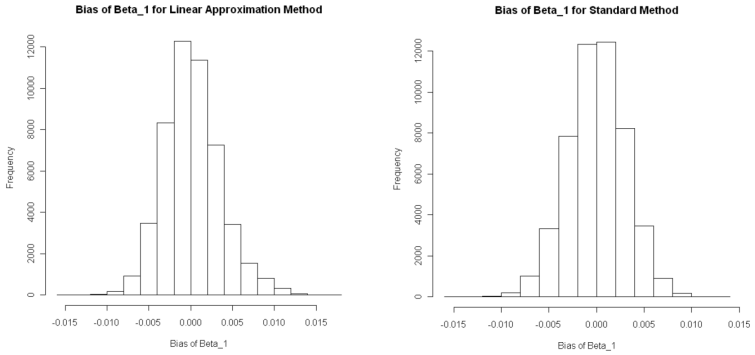


Figure 8.16: Histogram of Bias of $\hat{\beta}_0$ of Simulation for Method 2 ($T = 100,000$)

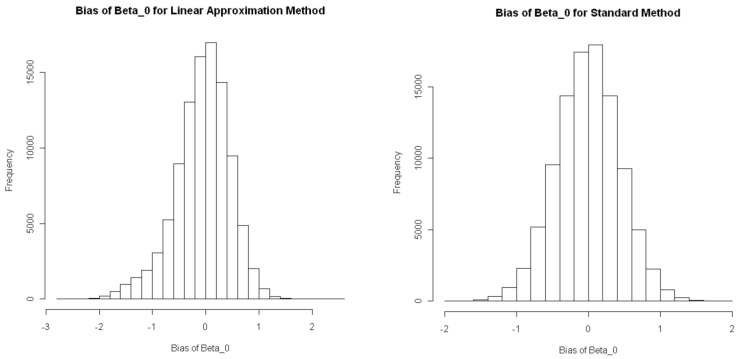


Figure 8.17: Histogram of Bias of $\hat{\beta}_1$ of Simulation for Method 2 ($T = 100,000$)

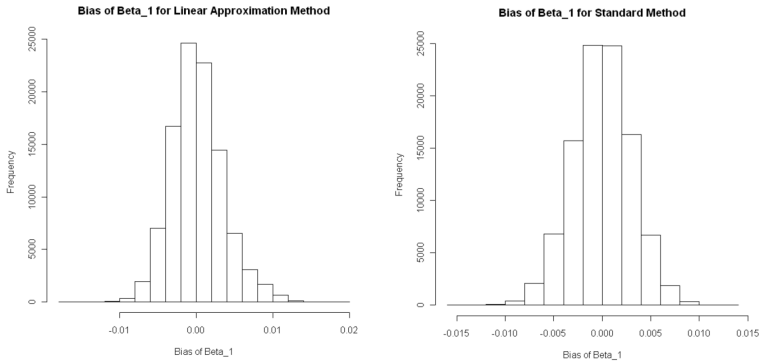


Table 8.5: Counts of Estimates in Certain Intervals for Both Methods of Simulation for Method 2

Estimate	Interval	$T = 25,000$		$T = 50,000$		$T = 100,000$	
		LA	SM	LA	SM	LA	SM
$\hat{\beta}_0$	(-3.2-0.4,-3.2+0.4)	15140	15967	30204	32050	60410	64073
	(-3.2-0.8,-3.2+0.8)	22253	23265	44513	46532	88958	93018
	(-3.2-1.2,-3.2+1.2)	24220	24814	48329	49641	96615	99253
$\hat{\beta}_1$	(0.01-0.003,0.01+0.003)	16315	17052	32772	34208	65567	68501
	(0.01-0.006,0.01+0.006)	23023	23830	46107	47653	92083	95215
	(0.01-0.009,0.01+0.009)	24632	24918	49190	49822	98304	99679

data our proposed linear approximation method 2 performs as good as the standard method.

Similarly, we also present the simulation result for Method 3.

Table 8.6: Simulation of Method 3

			$T = 25,000$	$T = 50,000$	$T = 100,000$
Linear Approximation	$\hat{\beta}_0$	Mean	-3.18819	-3.18699	-3.19041
		Median	-3.18539	-3.18136	-3.18556
	$\hat{\beta}_1$	Mean	0.009881	0.009877	0.009897
		Median	0.009896	0.009864	0.009912
Standard Method	$\hat{\beta}_0$	Mean	-3.19979	-3.19808	-3.20198
		Median	-3.19910	-3.19453	-3.20025
	$\hat{\beta}_1$	Mean	0.009952	0.009945	0.009968
		Median	0.009979	0.009949	0.009997

Note: The true values of β_0 and β_1 are $\beta_0 = -3.2$ and $\beta_1 = 0.01$.

Figure 8.18: Histogram of Bias of $\hat{\beta}_0$ of Simulation for Method 3 ($T = 25,000$)

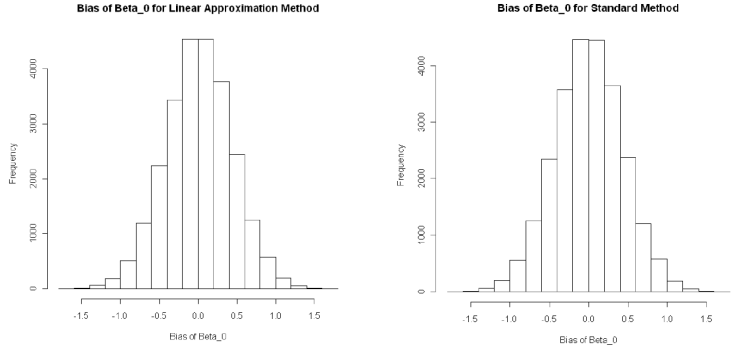


Figure 8.19: Histogram of Bias of $\hat{\beta}_1$ of Simulation for Method 3 ($T = 25,000$)

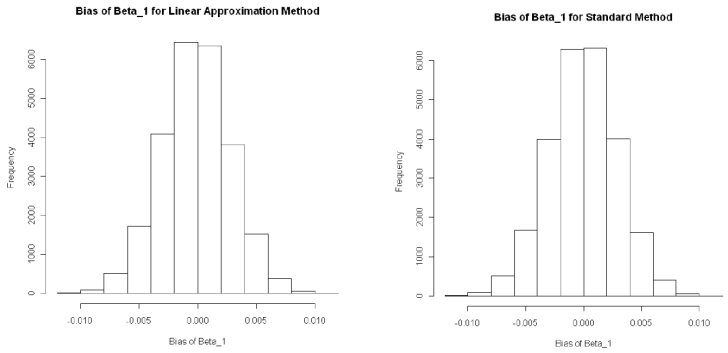


Figure 8.20: Histogram of Bias of $\hat{\beta}_0$ of Simulation for Method 3 ($T = 50,000$)

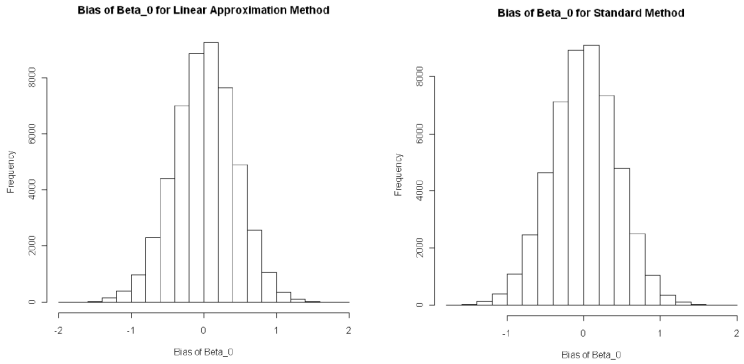


Figure 8.21: Histogram of Bias of $\hat{\beta}_1$ of Simulation for Method 3 ($T = 50,000$)

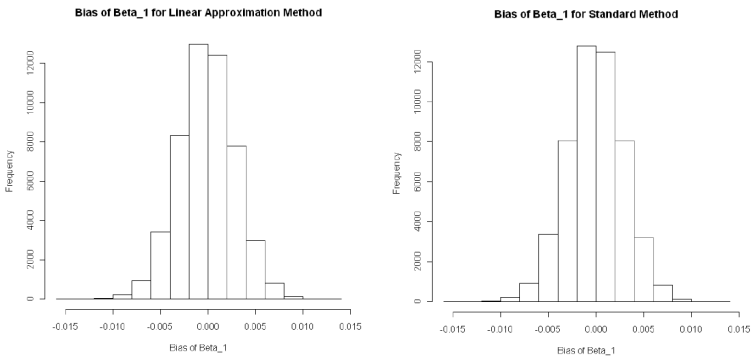


Figure 8.22: Histogram of Bias of $\hat{\beta}_0$ of Simulation for Method 3 ($T = 100,000$)

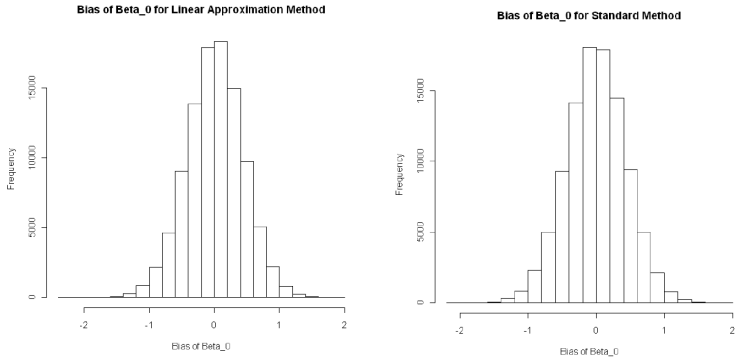


Figure 8.23: Histogram of Bias of $\hat{\beta}_1$ of Simulation for Method 3 ($T = 100,000$)

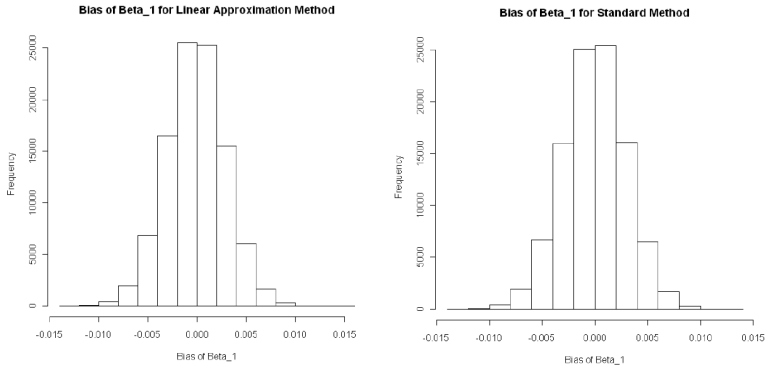


Table 8.7: Counts of Estimates in Certain Intervals for Both Methods of Simulation for Method 3

Estimate	Interval	$T = 25,000$		$T = 50,000$		$T = 100,000$	
		LA	SM	LA	SM	LA	SM
$\hat{\beta}_0$	(-3.2-0.4,-3.2+0.4)	16267	16148	32735	32482	65073	64699
	(-3.2-0.8,-3.2+0.8)	23395	23330	46900	46837	93512	93400
	(-3.2-1.2,-3.2+1.2)	24858	24861	49679	49704	99363	99363
$\hat{\beta}_1$	(0.01-0.003,0.01+0.003)	17418	17306	34939	34705	69425	68905
	(0.01-0.006,0.01+0.006)	23930	23910	47897	47911	95622	95575
	(0.01-0.009,0.01+0.009)	24932	24940	49859	49862	99701	99725

In Table 8.6, Table 8.7 and Figures 8.18-8.23, we see that for the simulated data our proposed linear approximation method 3 also performs as good as the standard method. Moreover, by comparing Table 8.5 and Table 8.7, we find Method 3 is closer to standard method than Method 2.

Our analysis of the Cornfield data and simulated data demonstrate our Method 2 and Method 3 perform as good as the standard method, even better with respect to some criteria. These two methods also provide a better initial value for the standard method, which would improve the efficiency of the iterative method in terms of the number of iteration steps. Most importantly, these two methods give us the exact closed form solutions of MLEs under our proposed linear approximation framework, while the standard method gives only the numerical solutions. With the closed form solutions, we plan to perform deeper research and obtain more information and properties of the estimators.

8.3.5 Properties

We now present some theoretical properties of our estimators. In (8.17), we denote

$$u_j = \gamma_1 y_j + \delta_2 n_j, v_j = \gamma_0 y_j - \delta_0 n_j. \tag{8.23}$$

Then our estimators become

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} u_i z_i v_j (z_i - z_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j z_i (z_i - z_j)} \\ \\ -\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} u_i v_j (z_i - z_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j z_i (z_i - z_j)} \end{pmatrix}. \tag{8.24}$$

Let $f_1(y_1,\ldots,y_N)=\sum_{i=1}^{N}\sum_{j=1}^{N}u_iz_iv_j(z_i-z_j)$, $f_2(y_1,\ldots,y_N)=-\sum_{i=1}^{N}\sum_{j=1}^{N}u_iv_j(z_i-z_j)$ and $g(y_1,\ldots,y_N)=\sum_{i=1}^{N}\sum_{j=1}^{N}u_iu_jz_i(z_i-z_j)$. We have

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}=\begin{pmatrix}\frac{f_1(y_1,\ldots,y_N)}{g(y_1,\ldots,y_N)}\\ \frac{f_2(y_1,\ldots,y_N)}{g(y_1,\ldots,y_N)}\end{pmatrix}=\begin{pmatrix}h_1(y_1,\ldots,y_N)\\ h_2(y_1,\ldots,y_N)\end{pmatrix}. \tag{8.25}$$

Since we have the closed form expressions of these estimators, we are able to calculate their expectations and variances. Directly calculating the expectation and variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ is difficult. Thus we consider the first order approximation of Taylor series expansion.

We take $\hat{\beta}_0$ as an example. We have

$$\begin{aligned}h_1(y_1,\ldots,y_N)\approx{}& h_1(0,\ldots,0)+y_1\cdot\tfrac{\partial}{\partial y_1}h_1(y_1,\ldots,y_N)\,|_{(0,\ldots,0)}+y_2\cdot\tfrac{\partial}{\partial y_2}h_1(y_1,\ldots,y_N)\,|_{(0,\ldots,0)}\\ &+\cdots+y_N\cdot\tfrac{\partial}{\partial y_N}h_1(y_1,\ldots,y_N)\,|_{(0,\ldots,0)}.\end{aligned} \tag{8.26}$$

Denote $a_0=h_1(0,\ldots,0)$ and $\frac{a_i}{n_i}=\frac{\partial}{\partial y_i}h_1(y_1,\ldots,y_N)\,|_{(0,\ldots,0)}$. Then we have

$$E\left(\hat{\beta}_0\right)=E\left(h_1(y_1,\ldots,y_N)\right)\approx a_0+\frac{a_1}{n_1}E(y_1)+\frac{a_2}{n_2}E(y_2)+\cdots+\frac{a_N}{n_N}E(y_N). \tag{8.27}$$

Since $E(y_i)=n_ip_i=n_i\cdot\Phi(\beta_0+\beta_1z_i)$, we write

$$E\left(\hat{\beta}_0\right)\approx a_0+a_1\Phi(\beta_0+\beta_1z_1)+a_2\Phi(\beta_0+\beta_1z_2)+\cdots+a_N\Phi(\beta_0+\beta_1z_N). \tag{8.28}$$

Now we use the Taylor series expansion again to approximate $\Phi(\beta_0+\beta_1z_i)$. The first order approximation is

$$\Phi(\beta_0+\beta_1z_i)\approx\frac{1}{2}+\frac{1}{\sqrt{2\pi}}(\beta_0+\beta_1z_i). \tag{8.29}$$

Plugging (8.29) in (8.28), we have

$$E\left(\hat{\beta}_0\right)\approx\left(a_0+\frac{\sum_{i=1}^{N}a_i}{2}\right)+\frac{\sum_{i=1}^{N}a_i}{\sqrt{2\pi}}\beta_0+\frac{\sum_{i=1}^{N}a_iz_i}{\sqrt{2\pi}}\beta_1. \tag{8.30}$$

We observe from (8.30) that $E(\hat{\beta}_0) = \beta_0$ when

1. $a_0 + \frac{\sum_{i=1}^{N} a_i}{2} = 0$, and
2. $\sum_{i=1}^{N} a_i = \sqrt{2\pi}$, and
3. $\sum_{i=1}^{N} a_i z_i = 0$.

Since all a_i's depend on the values of γ_0, γ_1, δ_0 and δ_2, we are able to select these values such that $E(\hat{\beta}_0) = \beta_0$ holds.

Furthermore, we present the exact expressions of a_i's in (8.30),

$$a_0 = h_1(0, \ldots, 0) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} u_i z_i v_j (z_i - z_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j z_i (z_i - z_j)}|_{(0,\ldots,0)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} -\delta_0\delta_2 n_i n_j z_i (z_i - z_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \delta_2^2 n_i n_j z_i (z_i - z_j)}, \tag{8.31}$$

and

$$\frac{a_k}{n_k} = \frac{\partial}{\partial y_k} h_1(y_1, \ldots, y_N)\,|_{(0,\ldots,0)}$$

$$= \frac{\sum_{i=1}^{N} n_i(z_k - z_i)(-\gamma_1\delta_0 z_k - \gamma_0\delta_2 z_i) \sum_{i=1}^{N}\sum_{j=1}^{N} \left(\delta_2^2 n_i n_j\right) z_i(z_i - z_j) + \sum_{i=1}^{N}\sum_{j=1}^{N} (\delta_0\delta_2 n_i n_j) z_i(z_i - z_j) \sum_{i=1}^{N} \gamma_1\delta_2 n_i (z_k - z_i)^2}{\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \left(\delta_2^2 n_i n_j\right) z_i(z_i - z_j)\right)^2}. \tag{8.32}$$

We have the similar derivation for $\hat{\beta}_1$.

Chapter 9

GSC and PGSC Methods for Probit Regression Model

9.1 Introduction

In this chapter, we apply the GSC and PGSC methods to probit regression model. We present the details of the algorithms, analyze the real data and compare them with the standard method. We find that all the properties of these two methods established for logistic regression model in Chapter 6 also hold for probit regression model.

9.2 GSC Method

We recall the GSC method introduced in Chapter 2. Similar as the procedure in Chapter 4, we have the GSC method for probit regression model as follows:

1. Set initial values of $\boldsymbol{\beta}^{(0)}$ and iteration step $m = 0$.

2. For each j, compute $U_j^{*(m)} = \frac{1}{1+s_j^{(m)}}$ where $s_j^{(m)} = \theta_j^{(m)} = \frac{\Phi\left(\boldsymbol{\beta}^{(m)\prime}\boldsymbol{z}_j\right)}{1-\Phi\left(\boldsymbol{\beta}^{(m)\prime}\boldsymbol{z}_j\right)}$.

3. Solve an MLE problem

$$\boldsymbol{\beta}^{(m+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{N} \Big[y_j \log(\theta_j) - n_j U_j^{*(m)} \theta_j \Big]. \tag{9.1}$$

4. If the convergence criterion is reached, stop; else increase m and return to step 2.

For the Cornfield data, we have $\theta_j^{(m)} = \frac{\Phi\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)}{1-\Phi\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)}$. If we denote $\Phi_j = \Phi\left(\beta_0 + \beta_1 z_j\right)$ and $\Phi_j^{(m)} = \Phi\left(\beta_0^{(m)} + \beta_1^{(m)} z_j\right)$, then the MLE problem becomes

$$\boldsymbol{\beta}^{(m+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{N} \left(y_j \cdot \log \frac{\Phi_j}{1-\Phi_j} - n_j \left(1 - \Phi_j^{(m)}\right) \frac{\Phi_j}{1-\Phi_j} \right). \tag{9.2}$$

We take the derivative of objective function in (9.2) with respect to β_0 and β_1, and get

$$\begin{cases} \sum\limits_{j=1}^{N} n_j \cdot \frac{\phi_j^{(m+1)}}{\Phi_j^{(m+1)}\left(1-\Phi_j^{(m+1)}\right)} \left(\frac{y_j}{n_j} - \Phi_j^{(m+1)} \frac{\left(1-\Phi_j^{(m)}\right)}{\left(1-\Phi_j^{(m+1)}\right)} \right) = 0, \\ \sum\limits_{j=1}^{N} n_j z_j \cdot \frac{\phi_j^{(m+1)}}{\Phi_j^{(m+1)}\left(1-\Phi_j^{(m+1)}\right)} \left(\frac{y_j}{n_j} - \Phi_j^{(m+1)} \frac{\left(1-\Phi_j^{(m)}\right)}{\left(1-\Phi_j^{(m+1)}\right)} \right) = 0. \end{cases} \tag{9.3}$$

Therefore, for the Cornfield data, the algorithm works as follows:

1. Set initial values of $\beta_0^{(0)}$, $\beta_1^{(0)}$ and iteration step $m = 0$.

2. Solve an equation

$$\begin{cases} \sum\limits_{j=1}^{N} n_j \cdot \frac{\phi_j^{(m+1)}}{\Phi_j^{(m+1)}\left(1-\Phi_j^{(m+1)}\right)} \left(\frac{y_j}{n_j} - \Phi_j^{(m+1)} \frac{\left(1-\Phi_j^{(m)}\right)}{\left(1-\Phi_j^{(m+1)}\right)} \right) = 0, \\ \sum\limits_{j=1}^{N} n_j z_j \cdot \frac{\phi_j^{(m+1)}}{\Phi_j^{(m+1)}\left(1-\Phi_j^{(m+1)}\right)} \left(\frac{y_j}{n_j} - \Phi_j^{(m+1)} \frac{\left(1-\Phi_j^{(m)}\right)}{\left(1-\Phi_j^{(m+1)}\right)} \right) = 0. \end{cases}$$

3. If the convergence criterion is reached, stop; else increase m and return to step 2.

Furthermore, for $N = 2$, we find the exact solution of (9.3) as

$$\begin{cases} \beta_0^{(m+1)} = \dfrac{z_i \Phi^{-1}\left(\frac{y_j}{y_j + n_j\left(1-\Phi\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)\right)} \right) - z_j \Phi^{-1}\left(\frac{y_i}{y_i + n_i\left(1-\Phi\left(\beta_0^{(m)}+\beta_1^{(m)} z_i\right)\right)} \right)}{z_i - z_j}, \\ \beta_1^{(m+1)} = \dfrac{\Phi^{-1}\left(\frac{y_i}{y_i + n_i\left(1-\Phi\left(\beta_0^{(m)}+\beta_1^{(m)} z_i\right)\right)} \right) - \Phi^{-1}\left(\frac{y_j}{y_j + n_j\left(1-\Phi\left(\beta_0^{(m)}+\beta_1^{(m)} z_j\right)\right)} \right)}{z_i - z_j}. \end{cases} \tag{9.4}$$

9.3 PGSC Method

We now apply the PGSC method to probit regression model with the Cornfield data. We present Table 9.1 where we see that Δ attains its minimum value 13.55 for the 14th pair containing observations 3 and 4. We also apply the GSC method and the standard method to the Cornfield data and present the comparison in Table 9.2. In Table 9.3, we present the comparison of different initial values for the standard method.

From Tables 9.1-9.3, we get similar results as in Chapter 5. In Table 9.1, we observe that the solution of the PGSC method is identical with the solution of standard method with paried observations. In Table 9.2, we observe that the PGSC method gives a better estimate than the GSC method for whole data with respect to the criterion function Δ. However, the GSC method for whole data performs better than the PGSC method with respect to the criterion function $-2\log L$. In Table 9.3, we observe that, for the standard iterative method, the number of iteration steps by using the PGSC method solution as the initial value is smaller than using (0,0).

9.4 Properties of the GSC and PGSC Methods

The properties of the GSC and PGSC methods for logistic regression model also hold for probit regression model. In other words, we have the following theorems similar as in Chapter 6.

Theorem 11 *In the GSC method (9.3), if two consecutive estimates at mth step and*

Table 9.1: Comparison Between PGSC Method and Standard Method with Paired Observations for All Pairs (Cornfield Data)

	PGSC Method			Standard Method	
Pair	$(\beta_0^{(m)},\beta_1^{(m)})$	Δ	m	$(\hat{\beta}_0,\hat{\beta}_1)$	Δ
(1,2)	(-8.480322, 0.057492560)	363.95	7	(-8.480322, 0.057492560)	363.95
(1,3)	(-3.992097, 0.017239423)	15.54	7	(-3.992097, 0.017239423)	15.54
(1,4)	(-3.954339, 0.016900779)	14.54	7	(-3.954339, 0.016900779)	14.54
(1,5)	(-4.038438, 0.017655036)	17.22	8	(-4.038438, 0.017655036)	17.22
(1,6)	(-3.751510, 0.015081692)	17.59	7	(-3.751510, 0.015081686)	17.59
(1,7)	(-3.926059, 0.016647146)	14.29	9	(-3.926059, 0.016647146)	14.29
(1,8)	(-3.710823, 0.014716783)	18.83	9	(-3.710823, 0.014716783)	18.83
(2,3)	(1.301190, -0.023013714)	63.21	7	(1.301190, -0.023013714)	63.21
(2,4)	(-1.082470, -0.003395111)	39.75	7	(-1.082470, -0.003395111)	39.75
(2,5)	(-2.026644, 0.004375866)	29.71	7	(-2.026644, 0.004375866)	29.71
(2,6)	(-2.039172, 0.004478975)	29.66	7	(-2.039172, 0.004478975)	29.66
(2,7)	(-2.615293, 0.009220708)	33.40	10	(-2.615293, 0.009220708)	33.40
(2,8)	(-2.540600, 0.008605959)	32.07	9	(-2.540600, 0.008605951)	32.07
(3,4)	(-3.858502, 0.016223501)	13.55	6	(-3.858502, 0.016223490)	13.55
(3,5)	(-4.101404, 0.018070648)	17.76	8	(-4.101404, 0.018070648)	17.76
(3,6)	(-3.519194, 0.013643196)	15.27	8	(-3.519194, 0.013643196)	15.27
(3,7)	(-3.879598, 0.016383916)	13.56	9	(-3.879598, 0.016383916)	13.56
(3,8)	(-3.549794, 0.013875941)	14.82	8	(-3.549794, 0.013875941)	14.82
(4,5)	(-4.381248, 0.019917807)	22.02	8	(-4.381248, 0.019917807)	22.02
(4,6)	(-3.310835, 0.012353049)	15.69	8	(-3.310835, 0.012353049)	15.69
(4,7)	(-3.887688, 0.016429752)	13.61	9	(-3.887688, 0.016429752)	13.61
(4,8)	(-3.459888, 0.013406431)	14.24	8	(-3.459881, 0.013406431)	14.24
(5,6)	(-2.089140, 0.004788395)	28.93	6	(-2.089126, 0.004788289)	28.93
(5,7)	(-3.641437, 0.015034536)	15.49	8	(-3.641437, 0.015034524)	15.49
(5,8)	(-3.148148, 0.011778521)	17.64	9	(-3.148148, 0.011778519)	17.64
(6,7)	(-4.847073, 0.021865352)	31.74	10	(-4.847073, 0.021865347)	31.74
(6,8)	(-3.594358, 0.014108596)	15.82	11	(-3.594358, 0.014108595)	15.82
(7,8)	(-2.108941, 0.006351866)	56.18	9	(-2.108941, 0.006351866)	56.18

Table 9.2: Comparison Between GSC Method for General N, PGSC Method and Standard Method (Cornfield Data)

Method	$\left(\hat{\beta}_0, \hat{\beta}_1\right)$	Δ	$-2\log L$
GSC (Whole Data)	(-3.196911, 0.01205205)	17.17103	38.75740
PGSC	(-3.858502, 0.01622350)	13.55034	42.68684
Standard (Whole Data)	(-3.196988, 0.01205270)	17.17140	38.75740

Table 9.3: Different Initial Values of Standard Method (Cornfield Data)

Initial Value	m	$\beta_0^{(m)}$	$\beta_1^{(m)}$	Δ	$-2\log L$
(0,0)	6	-3.197	0.0121	17.17	38.76
(-3.859, 0.0162)	3	-3.197	0.0121	17.17	38.76

$(m+1)$th step are close enough, i.e., $\beta_0^{(m)}$ is close enough to $\beta_0^{(m+1)}$ and $\beta_1^{(m)}$ is close enough to $\beta_1^{(m+1)}$, then the estimates $\beta_0^{(m+1)}$ and $\beta_1^{(m+1)}$ will be identical with the true solution of (8.7), $\hat{\beta}_0$ and $\hat{\beta}_1$. In other words, if we define $\beta_0^{(m)} = \beta_0^{(m+1)}$, $\beta_1^{(m)} = \beta_1^{(m+1)}$ up to a certain decimal place, then we must have $\beta_0^{(m)} = \beta_0^{(m+1)} = \hat{\beta}_0$, $\beta_1^{(m)} = \beta_1^{(m+1)} = \hat{\beta}_1$ up to the same decimal place.

Theorem 12 *In the GSC method (9.3), if we use the true solution of (8.7) to be the initial value of (9.3), i.e., $\beta_0^{(0)} = \hat{\beta}_0$, $\beta_1^{(0)} = \hat{\beta}_1$, then (9.3) would stop at the 1st iteration step and give the same solution $\beta_0^{(1)} = \hat{\beta}_0$ and $\beta_1^{(1)} = \hat{\beta}_1$.*

Theorem 13 *In the PGSC method (9.4), if two consecutive estimates at mth step and $(m+1)$th step are close enough, i.e., $\beta_0^{(m)}$ is close enough to $\beta_0^{(m+1)}$ and $\beta_1^{(m)}$ is close enough to $\beta_1^{(m+1)}$, then the estimates $\beta_0^{(m+1)}$ and $\beta_1^{(m+1)}$ will be identical with the true solution of (8.10), $\hat{\beta}_0$ and $\hat{\beta}_1$. In other words, if we define $\beta_0^{(m)} = \beta_0^{(m+1)}$, $\beta_1^{(m)} = \beta_1^{(m+1)}$ up to a certain decimal place, then we must have $\beta_0^{(m)} = \beta_0^{(m+1)} = \hat{\beta}_0$, $\beta_1^{(m)} = \beta_1^{(m+1)} = \hat{\beta}_1$ up to the same decimal place.*

Theorem 14 *In the PGSC method (9.4), if we use the exact solution of (8.10) to be*

the initial value of (9.4), i.e., let $\beta_0^{(0)} = \hat{\beta}_0$, $\beta_1^{(0)} = \hat{\beta}_1$, *then (9.4) would stop at the 1st iteration step and give the same solution* $\beta_0^{(1)} = \hat{\beta}_0$ *and* $\beta_1^{(1)} = \hat{\beta}_1$.

Theorem 15 *Consider the PGSC equation (9.4) and the uth pair containing observations i and j. Every step of the PGSC method provides us a predicted value closer to the real value and monotonically converges to the real value. In other words, from (9.4), we have* $(\beta_0^{(m)}, \beta_1^{(m)})$, $(\beta_0^{(m+1)}, \beta_1^{(m+1)})$ *and* $(\hat{y}_i^{(m)}, \hat{y}_i^{(m+1)})$. *Then* $\hat{y}_i^{(m)} < \hat{y}_i^{(m+1)} < y_i$ *or* $y_i < \hat{y}_i^{(m+1)} < \hat{y}_i^{(m)}$, *and finally* $\hat{y}_i^{(m+1)} = \hat{y}_i^{(m)} = y_i$. *Similarly for j.*

All the proofs are exactly the same as in Chapter 6 except we use a different θ_j and p_j.

Chapter 10

Conclusions

In this dissertation, we propose several new methods for solving the Maximum Likelihood Estimating Equations (MLEEs) of logistic and probit regression models. We study the performances of these methods by analyzing the Cornfield data (Cornfield (1962)) as well as a simulated data. We compare them with the existing methods. We also investigate theoretical properties of these methods.

For probit regression model, we introduce a linear approximation method to solve MLEEs. We find our method performs as good as the existing methods and **even better with respect to the Δ criterion function defined in (8.18) and the Pearson's χ^2 criterion function defined in (8.20). The advantage of our method is that we find the exact closed form solution of MLEEs**. With the closed form expressions, we calculate the expectations of our estimates. This calculation of expectation is **not possible for the standard method**. Moreover, we study the properties of the existing Generalized Self Consistency (GSC) method. We then introduce a new method by the repeated application of the GSC method to all paired observations. We call this

method the PGSC method. This method provides a better initial value of parameters for iterative methods. **The number of steps for convergence to the final solution using the better initial value is smaller than using (0,0)**. In addition, we find that the PGSC method monotonically converges and its solution is identical with the solution for standard method.

For logistic regression model, we also implement the PGSC method which has **exactly the same goodness properties as in probit regression model**. Moreover, we combine the paired observation procedure with the GSC method in another way to propose two new methods. These two new methods also provide us a meaningful way to choose the initial value of parameters for the iterative methods.

Bibliography

[1] Alan Agresti. *Categorical Data Analysis*. WILEY INTERSCIENCE, 2nd edition, 2002.

[2] N.M. Laird A.P. Dempster and D.B. Rubin. Maximum likelihood from incomplete data via the em algorith. *Journal Of The Royal Statistical Society Series B-Methodological*, 39(1):1–38, 1977.

[3] Stuart G. Baker. The multinomial-poisson transformation. *The Statistician*, 43(4):495–504, 1994.

[4] Solomon Chefo and Alex Tsodikov. Stage-specific cancer incidence: An artificially mixed multinomial logit model. *Statistics in Medicine*, 28(15):2054–2076, 2009.

[5] Jerome Cornfield. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure - a discriminant function analysis. *Federation Proceedings*, 21(4):58, 1962.

[6] Peter J. Diggle, Patrick J. Heagerty, Kung-Yee Liang, and Scott L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 2nd edition, 2002.

[7] James W. Hardin and Joseph M. Hilbe. *Generlaized Estimating Equations*. Chapman and Hall/CRC, 2003.

[8] Joseph B. Lang. On the comparison of multimomial and poisson log-linear models. *Journal of the Royal Statistical Society. Series B(Statistical Methodology)*, 58(1):253–266, 1996.

[9] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, April 1986.

[10] Charles E. McCulloch. Generalized linear models. *Journal of the American Statistical Association*, 95(452):1320–1324, December 2000.

[11] Juni Palmgren. The fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika*, 68(2):563–566, 1981.

[12] P.McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[13] Kung-Yee Liang Scott L. Zeger and Steven G. Self. The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, 72(1):31–38, 1985.

[14] Thaddeus Tarpey and Bernard Flury. Self-consistency: A fundamental concept in statistics. *Statistical Science*, 11(3):229–243, August 1996.

[15] Alex Tsodikov. Semiparametric models: A generalized self-consistency approach. *Journal of the Royal Statistical Society. Series B(Statistical Methodology)*, 65(3):759–774, 2003.

[16] Alex Tsodikov and Solomon Chefo. Generalized self-consistency: Multinomial logit model and poisson likelihood. *Journal of Statistical Planning and Inference*, 138:2380–2397, 2008.

[17] Shufang Wang and Alex Tsodikov. A self-consistency approach to multinomial logit model with random effects. *Journal of Statistical Planning and Inference*, 140(7):1939–1947, 2010.

[18] R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61(3):439–447, December 1974.

[19] Scott L. Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, March 1986.

[20] A. Ziegler, C. Kastner, U. Gromping, and M. Blettner. The generalized estimating equations in the past ten years: An overview and a biomedical application. *Biometrika*, 73, April 1996.

A.1: Proof of equation (2.8)

$$\begin{aligned} E_U\left[e^{-US}\right] &= \int_0^\infty e^{-u}\cdot e^{-us}\mathrm{d}u = \int_0^\infty e^{-u(1+s)}\mathrm{d}u \\ &= -\tfrac{1}{1+s}\int_0^\infty e^{-u(1+s)}\mathrm{d}\ -u\,(1+s) \\ &= -\tfrac{1}{1+s}e^{-u(1+s)}|_0^\infty \\ &= \tfrac{1}{1+s} \end{aligned}$$

A.2: Proof of equation (2.13)

$$
\begin{aligned}
l &= \sum_{i=1}^{J}\sum_{j=1}^{N} y_{ij}\cdot\log\left[p_i(z_j,U_j)\right] = \sum_{i=1}^{J}\sum_{j=1}^{N} y_{ij}\cdot\log\left[\theta_{ij}\cdot e^{-U_j\sum_{k=2}^{J}\theta_{kj}}\right] \\
&= \sum_{i=2}^{J}\sum_{j=1}^{N} y_{ij}\cdot\log\left[\theta_{ij}\cdot e^{-U_j\sum_{k=2}^{J}\theta_{kj}}\right] + \sum_{j=1}^{N} y_{1j}\cdot\log\left[\theta_{1j}\cdot e^{-U_j\sum_{k=2}^{J}\theta_{kj}}\right] \\
&= \sum_{i=2}^{J}\sum_{j=1}^{N} y_{ij}\cdot\left[\log\theta_{ij} - U_j\sum_{k=2}^{J}\theta_{kj}\right] + \sum_{j=1}^{N} y_{1j}\cdot\left[\log\theta_{1j} - U_j\sum_{k=2}^{J}\theta_{kj}\right] \\
&= \sum_{i=2}^{J}\sum_{j=1}^{N} y_{ij}\cdot\log\theta_{ij} - \sum_{i=2}^{J}\sum_{j=1}^{N}\left(y_{ij}\cdot U_j\sum_{k=2}^{J}\theta_{kj}\right) - \sum_{j=1}^{N}\left(y_{1j}\cdot U_j\sum_{k=2}^{J}\theta_{kj}\right) \\
&= \sum_{i=2}^{J}\sum_{j=1}^{N} y_{ij}\cdot\log\theta_{ij} - \sum_{i=1}^{J}\sum_{j=1}^{N}\left(y_{ij}\cdot U_j\sum_{k=2}^{J}\theta_{kj}\right) \\
&= \sum_{i=2}^{J}\sum_{j=1}^{N} y_{ij}\cdot\log\theta_{ij} - \sum_{j=1}^{N}\left(\left(\sum_{i=1}^{J} y_{ij}\right)U_j\sum_{k=2}^{J}\theta_{kj}\right) \\
&= \sum_{i=2}^{J}\sum_{j=1}^{N} y_{ij}\cdot\log\theta_{ij} - \sum_{j=1}^{N}\left(y_{\cdot j}U_j\sum_{k=2}^{J}\theta_{kj}\right) \\
&= \sum_{i=2}^{J}\sum_{j=1}^{N}\left(y_{ij}\cdot\log\theta_{ij} - y_{\cdot j}U_j\theta_{ij}\right) \\
&= \sum_{i=2}^{J} l_i
\end{aligned}
$$