Generalized linear models with nonignorable missing covariates

by

Sonia Afroje Lopa

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfilment of the requirements for the degree of

Master of Science

in

Probability and Statistics

Carleton University
Ottawa, Ontario

Abstract

In this thesis, we present an overview of generalized linear models (GLMs) for binary and count data with missing covariates when the missing data mechanism is nonignorable. We study the maximum likelihood method to estimate the parameters in GLMs. Particularly, we consider joint estimation of the regression parameters and nuisance parameters by the maximum likelihood method, when some covariates are nonignorably missing. We study a set of estimating equations from the maximum likelihood method for fitting regression models to binary and Poisson data in the presence of missing covariates.

Simulations were carried out to observe the behaviour of the maximum likelihood estimates under both correctly specified and misspecified structures. Our simulation study shows that if the fitted model is correctly specified, then the maximum likelihood method generally provides unbiased and efficient estimators, where as a misspecified model provides biased and inefficient estimators. The simulation results also indicate that when the sample size is small, the empirical coverage probabilities of the parameter estimates are a bit apart from the nominal 95% level. But they tend to get closer to the nominal level when the sample size is larger. Also, the average lengths of the confidence intervals for the regression parameters tend to be smaller for larger sample size, as expected. Under misspecified missing data models, we observe systemic bias in the regression estimators and also poor coverage probabilities from the confidence intervals.

We conclude that when analyzing incomplete data with missing covariates, it is necessary to incorporate a suitable missing data model into the observed data likelihood function in order to obtain unbiased and efficient estimators of the model parameters. We also note that a misspecified missing data model can provide systematic bias in the maximum likelihood estimation. So it is important to assess the validity of a missing

data model when performing a likelihood inference based on the given observed data.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Sanjoy K. Sinha for his kind supervision, advice, and guidance. It has been an honour to get a chance of being his M.Sc. student. I appreciate all his encouragement, productive ideas, and support throughout this work. Completion of this work would have been impossible without his guidance.

I also thank many people in the School of Mathematics and Statistics at Carleton University for their help, continual support to continue my graduate study in Statistics. Specially I would like to thank Nicole Gaertner, the Graduate Administrator of the school, for her valuable help and suggestions.

I am grateful to Alia Alkhathami and all friends in my group specially Parvin Dehghani, Marie-Joe Nemnom, Touba Warsi for sharing their experience and knowledge during my graduate study.

It is an honour to thank my wonderful mother Shahera Akter who has dedicated her whole life to my sisters and myself and provided unconditional love and care. Also, she has helped me a lot in this very crucial time by looking after my first baby girl. I would like to extend my sincere thanks to my two sisters Lovely Akter and Lata Akter for their love, support, and encouragement in the higher education from my childhood. And most of all my loving, caring, encouraging and patient husband Musharraf whose support in every stage of this M.Sc is unforgettable. I also want to remember my nephew (Soad), nieces (Aboni, Tanha, Abriti) and my new born daughter Aritri (Mahdisha Afroje) who have enriched my life with joy and happiness.

Lastly, I offer my regards and blessing to all of those who supported me in any aspect

during the completion of my study. I am heartily thankful to all of you.

I dedicate this thesis to
My Daughter And My Mother
Mahdisha Afroje Aritri And Shahera Akter
for their unconditional love.

Contents

List of Tables

Abbreviations

CI	Confidence Interval
GLM	Generalized Linear Model
LM	Linear Model
MAR	Missing at Random
MCAR	Missing Completely at Random
NMAR	Nonignorable Missing at Random
MLE	Maximum Likelihood Estimates
MSE	Mean Squared Error
NI	Non-Ignorable
NR	Newton-Raphson

Chapter 1

Introduction

1.1 Generalized Linear Model

In biostatistics and other areas of applied statistics, analysis of generalized linear models (McCullagh and Nelder, 1989) with missing data is attracting more attention to the researchers in recent years. Missing values in the data could be either ignorable or nonignorable (Ibrahim, Lipsitz and Chen, 1999). An individual's unobserved response is called non-ignorable if it is related to the missing values of the response variable (Little, 1982). If there are non-ignorable missing responses, the distribution of the missing data needs to be correctly modelled to avoid any biases that may be resulted from the incomplete data analysis (Baker and Laird, 1988).

In regression analysis, the generalized linear model (GLM) has been considered one of the most significant advanced methods in the past decades (Hoffman, 2004). GLM is referred to as the most significant statistical tool since it is very flexible to address a variety of statistical problems and has the available software to fit the models.

There is a substantial confusion because of the abbreviation GLM, some practitioners use a term to refer this to the general linear model, which only comprises LMs, analysis of variance, and analysis of covariance (Lane, 2002). On the other hand, GLMs provide a unified framework, incorporating ordinary general linear models as well as many common nonlinear models (Wu, 2005).

The generalized linear model (GLM) is a common generalization of ordinary linear regression when the error distribution of response variables is not normal including count, binary, proportions and positive valued continuous distributions (Nelder and Wedderburn, 1972; Hilbe, 1994; Hoffman, 2004). Generalized linear models (GLMs) unify different approaches that can handle continuous variables as well as discrete variables when the error distribution of the variables belongs to the natural exponential family such as Poisson and binomial distributions.

One of the approaches to estimating the parameters in GLMs is obtaining the likelihood function by integrating out the random effects from the joint distribution of responses. Usually, Monte Carlo approximations are used to evaluate the integrations since it is quite demanding to evaluate the likelihood function in a closed form (Wei and Tanner, 1990). The likelihood function is maximized to estimate the parameters of interest.

1.2 Missing Data Problems

Missing data are common problems that arise very often in biological research. In common phenomenon, an individual's response might not be observed at one follow-up time

and might be measured at the next follow-up time, resulting in a large number of missingness patterns. When missing data occur, they must be taken into account in order to obtain valid statistical inference.

Missing data patterns are discussed in three different ways according to Little and Rubin (1987), which suggest the relationship between the missingness and observed values in the data. Suppose Y is the complete data that would occur in the absence of missing values. We can write $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{obs} denotes the observed values and Y_{mis} the missing values of Y. The density function of the joint distribution of Y_{obs} and Y_{mis} can be written as

$$f(Y|\theta) \equiv f(Y_{\text{obs}}, Y_{\text{mis}}|\theta),$$

depending on some vector of unknown parameters θ.

Consider a binary variable that indicates whether each component of Y is observed or missing. A vector of response indicators $R = (R_{ij})$ is defined in such a way that

$$R_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{if } y_{ij} \text{ is missing} \end{cases}$$

Here R is treated as a random vector in the model and we assume a joint distribution of R and Y. The density of the distribution can be expressed as

$$f(Y, R|\theta, \psi) = f(Y|\theta) f(R|Y, \psi),$$

where $f(R|Y,\psi)$ depends on some parameters ψ, which is the conditional distribution of R given Y for the missing data mechanism.

The data are considered to be missing completely at random (MCAR) when the missing data process does not depend on the observed values as well as the missing values of the data, which means that the missingness is independent of Y. It can be expressed as

$$f(R|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = f(R|\psi).$$

The missing data mechanism is considered to be missing at random (MAR) when the missingness depends only on the observed values of the variables in the data, which means that the missingness only depends on observed components Y_{obs} of Y. It can be expressed as

$$f(R|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = f(R|Y_{\text{obs}}, \psi).$$

An individual's non-response is considered to be missing not at random (NMAR), when the individual's response probability depends on the response variable itself, which means that the missingness depends on the unobserved values. It can be defined as

$$f(R, Y|X, \theta, \psi) = f(Y|X, \theta) f(R|Y, X, \psi),$$

where X represents a full set of covariates. The first component classifies the distribution of Y given X in the population and the second component models the occurrence of response as a function of Y and X.

Rubin (1976) discussed that when MAR occurs, the likelihood based inference doesn't depend on missing data mechanism. Baker and Laird (1988) pointed out that if missingness depends on individual's unobserved response, then it is considered to be nonignorable missing data. In such cases, the distribution needed to be modelled for the missing-data mechanism. Several examples that describe the likelihood-based inference using maximum likelihood and missing data mechanisms are discussed in Little (1995)

Likelihood function under certain modelling assumptions is mostly used as an estimation method for missing data. For the ML estimation, the EM algorithm is a well-known iterative algorithm. The M step of the EM algorithm creates a function for the ML estimation when data are complete. The conditional expectation of the current estimates of the parameters for missing data given the observed data is performed on the E step.

1.3 Statement of the Problems

In this thesis, we study the maximum likelihood method to make simultaneous inferences for generalized linear models with missing covariates occurring from the nonignorable missing-data mechanism.

We investigate a set of maximum likelihood equations for fitting regression models to both binary and count data when there are missing values. Under a given missing data mechanism, we study the finite sample properties of empirical biases and mean squared errors as well as coverage probabilities (CPs) and average lengths of the confidence intervals for regression and association parameters for both binary and Poisson models. We also investigate the impact of changing the coefficients of the missing values in terms of biases and mean squared errors as well as coverage probabilities (CPs) and average lengths under both models. Our interest also lies in studying the empirical properties of the parameter estimates in the generalized linear model under both correctly specified and misspecified models.

1.4 Organization of the Thesis

The thesis is organized as follows. In Chapter 2, we introduce the generalized linear models (GLMs) and review the literature on missing covariates for analyzing incomplete data. In Chapter 3, we discuss our studied methodology and estimation algorithm based on the maximum likelihood for fitting regression models to binary and Poisson data. In Chapter 4, we present a simulation study and discuss the results of the binary model. We present results from a simulation study for Poisson model in Chapter 5. Chapter 6 concludes the thesis with some discussion and direction for further research.

Chapter 2

Generalized Linear Models And Missing Covariates

2.1 Introduction

Before we present our methods for estimation in GLMs with nonignorable missing covariates, we give a brief introduction to LMs, GLMs, and models for the missing data in this chapter. In Section 2.2, we introduce LMs and the model. In Section 2.3, we discuss GLMs and methods of estimation for the parameters in GLMs. Section 2.4 presents some reviews of the literature on GLMs. In Section 2.5, we give a literature review of methods for handling non-ignorable missing covariates.

2.2 Linear Models

Linear models (LMs) are an extensive and venerable issue that provides a rich background of statistical concepts (Janke and Tinsley, 2005). The classical linear models are the most widely used statistical method in regression analysis since they are simple to construct and interpret. They can clarify the linear relationship between the expectation of a response (dependent) variable and a set of explanatory (independent) variables

very easily.

To estimate the regression coefficients of LMs, the least squares method is mostly used. It provides the most efficient estimator of all unbiased estimators, because it has various suitable statistical properties and thus termed as the "best linear unbiased estimator" or BLUE (Draper and Smith, 1981).

The basic assumption of LMs is the linear relationship between the response variable and the explanatory variables, when values of the response variable (Y) are continuous and follow a conditional normal distribution (given X) with a constant (error) variance.

Let the response variables $(Y_1, \ldots, Y_n)$ be independent normal with means $\mu_i (i = 1, 2 \ldots, n)$ and a constant variance σ^2. Suppose x_{ij} is the jth explanatory variable for the ith observation and β_j is the unknown parameter associated with the jth explanatory variable. Then the classical linear model can be written as

$$E(Y_i) = \mu_i = \sum_{j=1}^{p} \beta_j x_{ij}, i = 1, \ldots, n. \tag{2.1}$$

In matrix notation, it can be rewritten as

$$E(Y) = \mu = \mathbf{X}\boldsymbol{\beta}, \tag{2.2}$$

where $\mathbf{X}$ is the $n \times p$ design matrix, μ is the $n \times 1$, $\boldsymbol{\beta}$ is the $p \times 1$ vector and p is the number of unknown parameters including the intercept. According to McCullagh and Nelder's (1989) terminology, Y contains the random part, whereas $\mathbf{X}\boldsymbol{\beta}$ is the linear predictor.

2.3 Literature on Generalized Linear Models

The classical linear model is very popular to model a continuous response under certain assumptions. It is more appropriate when the response follows a normal distribution and has a linear relationship with the mean of the covariates. However, sometimes the relationship between the mean of the responses and the covariates cannot be expressed in a linear form. In this case, the standard linear model is unrealistic. Rather than classical linear models, one can consider generalized linear models (GLMs) which generalize the classical linear models, allowing non linear forms between the mean of responses and the covariates.

The generalized linear model for incomplete data is discussed in Ibrahim et al. (1990). A method for estimating parameters in binomial regression models was proposed by Ibrahim and Lipsitz (1996) when the missing data mechanism is non-ignorable. They proposed a conditional model for incomplete covariates in parametric regression models.

In Zhao and Shao (2015), a generalized linear model is considered for identifiability and estimation when the missing data mechanism is nonignorable and unspecified. They assumed a pseudo-likelihood approach by implementing an instrumental variable to support identifying unknown parameters in the presence of nonignorable missing data.

Shi, Zhu, and Ibrahim (2009) developed a method of general local influence to the generalized linear models in the presence of missing covariate data to convey sensitivity analyses of minor perturbations. They also developed various local influence measures

to identify influential points and test model misspecification. To assess appropriate perturbation schemes, they conducted influence measures based on two objective functions which are the maximum likelihood estimate and the likelihood ratio statistic. Furthermore, they modelled the missing data mechanism as a sequence of one-dimensional conditional distributions of binary logistic regressions.

Sinha (2008) used a generalized linear model for maximum likelihood estimation. He proposed a robust approach to downweight any influential observation in the presence of nonignorable missing covariates. He adopted a Metropolis-Hastings algorithm based on a Markov chain sampling method to carry out some simulations. He examined the behavior of the robust estimates and compared them to the classical maximum likelihood estimates. Finally, he applied his method to real life data of delirium patients.

2.3.1 Structure of GLMs

The generalized linear model was defined by Nelder and Wedderburn (1972), which is one of the widely used statistical models considered in regression analysis.

GLMs generalize LMs by relaxing the two assumptions: (i) the response variables are not necessarily to be continuous and to follow normal distributions. They can be categorical or ordinal. (ii) The nonlinear relationship may exist between the responses and explanatory variables.

Let $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ be independent observations, where y_i is the ith response variable and $\mathbf{x}_i$ is the ith random vector of p covariates. Assume that y_i is independently and identically distributed, which follows a distribution in the exponential family. Then

the density function of each observation y_i can be written in the form

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

for some specific functions a, b and c, and ϕ is a scaler dispersion parameter (McCullagh and Nelder, 1997). Here, $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^t$ are natural parameters.

Generally, a GLM contains three components including a random component, a systematic component, and a link function.

First, the random components of y_i have independent distributions with mean $E(y_i) = \mu_i$, which follow a distribution in the exponential family.

Second, the systematic component makes linear relationship between the covariates $\mathbf{x}_i$ and the linear predictors η_i in the form

$$\eta_i = \mathbf{x}_i^t\boldsymbol{\beta}; i = 1, \ldots, n, \tag{2.4}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients.

Third, the link component connects the random components and the systematic components of the model as a form

$$\eta_i = \mathbf{x}_i^t\boldsymbol{\beta} = g(\mu_i); i = 1, \ldots, n. \tag{2.5}$$

Here $g(.)$ is a monotonic and differentiable function, connecting the mean and the linear predictor. The link function is sometimes invertible and thus is called the inverse link function

$$g^{-1}(\eta_i) = \mu_i; i = 1, \ldots, n. \tag{2.6}$$

In the exponential family, if a link function $g(.)$ satisfies $g(\mu_i) = \theta_i(\mu_i)$, then the link is called the canonical link. Given an appropriate choice of the link functions, GLMs can

fit the regression models, where the underlying data may follow the normal, Poisson, or binomial distribution.

2.3.2 Maximum Likelihood Estimation in GLMs

The maximum likelihood method is the basic method to estimate the parameters used in GLMs. Assume that the dispersion parameter ϕ is known; then $c(y_i, \phi)$ is a constant in the log-likelihood function about θ_i. Applying the ML method to GLMs, the log-likelihood function for n independent observations can be written (without the constant term) as

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{y}) &= \sum_{i=1}^{n} \log f(y_i|\theta_i;\phi) \\ &= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)}. \end{aligned} \tag{2.7}$$

If we take the derivative of Eq.(2.7) with respect to θ_i, we get

$$\frac{\partial l}{\partial \theta_i} = \frac{1}{a(\phi)}\left(y_i - \frac{\partial b(\theta_i)}{\partial \theta_i}\right), \tag{2.8}$$

and

$$\frac{\partial^2 l}{\partial {\theta_i}^2} = -\frac{1}{a(\phi)}\frac{\partial^2 b(\theta_i)}{\partial {\theta_i}^2}. \tag{2.9}$$

We take the first derivative of the log-likelihood function with respect to the parameters of interest to obtain the MLEs. To estimate the parameters, we set the derivatives equal to 0, which are called the likelihood score equations. In general, the likelihood score equation does not have an explicit solution (Firth, 1991). Therefore, to solve the score equation, we often use the iteration method assisted by computer.

The second derivatives of the log-likelihood function with respect to the parameters of interest $\boldsymbol{\beta}$ provide the variance-covariance matrix of the ML estimator of $\boldsymbol{\beta}$ which is also known as the Fisher information matrix. The asymptotic variance-covariance matrix of the ML estimator $\hat{\boldsymbol{\beta}}$ is the inverse of the Fisher information matrix.

We can show that

$$E\left(\frac{\partial l}{\partial \theta_i}\right) = 0, \tag{2.10}$$

$$\mathsf{var}\left(\frac{\partial l}{\partial \theta_i}\right) = -E\left(\frac{\partial^2 l}{\partial {\theta_i}^2}\right). \tag{2.11}$$

Substituting the values of Eq.(2.8) and Eq.(2.9) into Eq.(2.10) and Eq.(2.11), we get

$$E(y_i) = \mu_i(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}, \tag{2.12}$$

and

$$\begin{aligned} \mathsf{var}(y_i) &= \frac{1}{a(\phi)} \frac{\partial^2 b(\theta_i)}{\partial {\theta_i}^2} \\ &= \frac{1}{a(\phi)} \frac{\partial \mu_i(\theta_i)}{\partial \theta_i} \\ &= \frac{1}{a(\phi)} V(\mu_i), \end{aligned} \tag{2.13}$$

where $V(\mu_i) = \partial \mu_i(\theta_i)/\partial \theta_i$ is referred to as the variance function. Taking the derivatives on both sides of Eq.(2.5) with respect to $\boldsymbol{\beta}$, we get

$$\mathbf{x}_i = \frac{\partial g(\mu_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}}.$$

Also, it can be written as

$$\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \frac{1}{V(\mu_i) \partial g(\mu_i)/\partial \mu_i} \mathbf{x}_i. \tag{2.14}$$

To get the score function, we differentiate Eq.(2.7) and apply into Eq.(2.12), Eq.(2.13) and Eq.(2.14). Then we obtain the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ from the score function

$$\begin{aligned} \mathsf{S}(\boldsymbol{\beta}) &= \frac{\partial l(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \frac{\partial l_i(\theta_i|y_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{a(\phi)} \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)\partial g(\mu_i)/\partial \mu_i} \mathbf{x}_i. \end{aligned} \tag{2.15}$$

In matrix form, Eq.(2.15) can be rearranged as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathsf{S}(\boldsymbol{\beta}) = \frac{1}{a(\phi)} \mathbf{X} W D(\mathbf{Y} - \mu(\boldsymbol{\beta})), \tag{2.16}$$

where the design matrix,

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\},$$

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n),$$

$$W = diag^{-1} \left\{ V(\mu_1)(\partial g(\mu_1)/\partial \mu_1)^2, \ldots, V(\mu_n)(\partial g(\mu_n)/\partial \mu_n)^2 \right\},$$

and

$$D = diag \left\{ \partial g(\mu_1)/\partial \mu_1, \partial g(\mu_2)/\partial \mu_2 \ldots, \partial g(\mu_n)/\partial \mu_n \right\}.$$

To obtain the MLEs, Eq.(2.16) can be solved by performing Fisher scoring algorithm or Gauss-Newton algorithm. Both Fisher scoring and Newton-Raphson methods reduce to the iteratively re-weighted least squares algorithm in the presence of canonical links. To estimate the parameters in GLMs, the iteratively reweighted least squares and the

Newton-Raphson methods are the most common algorithms, which are used in most common statistical software packages such as SAS, S-Plus, and R.

Under certain conditions, MLEs of parameters in GLMs are asymptotically efficient and asymptotically normally distributed. That is,

$$\hat{\boldsymbol{\beta}} \rightarrow N(\boldsymbol{\beta}, a(\phi)(\mathbf{X}W\mathbf{X}^t)^{-1}).$$

To make the statistical inference, we can consider asymptotic normality property for large sample size.

2.4 Literature on Missing Data

We often experience the missing data problem in practice, which could be either in responses or in covariates. The missing data may lead to invalid inference if missingness is not properly addressed. An appropriate approach is needed to deal with the missing data. A wide range of analysis of missing data mechanism with incomplete data has been done by several authors in the literature (e.g., Baker and Laird, 1988; Wu and Carroll, 1988; Brown, 1990; Ibrahim, 1990; Schluchter, 1992; Little, 1993, 1994; Diggle and Kenward, 1994; Vach, 1994; Robins, Rotnitzky and Zhao, 1995; Ibrahim, Lipsitz and Chen, 1999; Robins, Greenland and Hu, 1999).

Baker and Laird (1988) developed a log-linear model for categorical data when there are nonignorable missing data. They obtained maximum likelihood estimators by using the EM algorithm. They considered two different log-linear models to describe two associated regressions. First, they fitted a marginal model to make inferences about the regression of Y on X, treating the nonresponse as a nuisance for the p-dimensional XY

margin. Second, they used a nonresponse model to describe the regression of R on X and Y for the full array assuming an interaction term of YR in the model.

Greenlees, Reece, and Zieschang (1982) proposed a parametric model using survey data for maximum likelihood estimators for both missing data mechanism and data-generating process to deal with nonignorable nonresponse.

Later on, Tang, Little, and Raghunathan (2003) suggested a pseudo-likelihood for nonignorable nonresponse, implementing a parametric model on the data generating process. Their work was slightly different from Greenlees, Reece, and Zieschang (1982) in the sense that the missing data mechanism needed to be unspecified. This approach estimates population parameters in the complete data case considering the missing data mechanism as a nuisance parameter.

2.5 Literature on Missing Covariates

There is extensive research done with missing covariates by many authors in the biological research area. Little (1992) defined four different types of missing covariates patterns. First, univariate missing data where only one covariate values are missing. Second, monotone or nested missing data. Third, a special pattern where two covariates cannot be observed together and fourth, a general pattern of missing data without any special structure.

He provided six methods of estimating parameters for the regression models which are different in assumptions made about the mechanism to deal with missing values. He discussed these statistical methods to make a comparative study with missing covari-

ates. He suggested three methods among six including the maximum likelihood method, Bayesian method and multiple imputation, which would be the better choice for dealing with missing covariate problems. However, for the large sample case, he preferred the maximum likelihood method, and small sample case, the Bayesian method or multiple imputation.

The generalized linear model is proposed by Ibrahim, Lipsitz and Chen (1999) for missing covariates with nonignorable missing data mechanism. They developed E step and M step of the EM algorithm for the covariates allowing them either to be categorical or continuous, where a Monte Carlo form of the EM algorithm is suggested through Gibbs sampler method for discrete covariates. Furthermore, they proposed a multinomial model that can be written as a logistic regression model or a sequence of one-dimensional conditional distributions with the missing data mechanism that may reduce the number of nuisance parameters. In Ibrahim, Chen, Lipsitz and Hemng (2005) a comparison of different methods is studied to make statistical inference in generalized linear models with missing covariates.

Yang, Belin and Boscardin (2005) proposed two alternative procedures to handle missing covariates and address the selection problem of a model in linear regression based on the Bayesian framework. One approach is called "impute then select" that produces multiplied imputed data sets first and then apply Bayesian variable selection to each of them subsequently. Another one is called "simultaneously impute and select", which inserts the steps for imputation and single Bayesian variable selection combining Gibbs sampling process. They used multivariate normal imputation models for incom-

plete data by implementing conditional distribution in linear regression.

Wu and Wu (2001) developed a three-step multiple imputation method to estimate the parameters in nonlinear mixed effect models in the presence of missing covariates (MAR) by implementing the Gibbs sampler method. First, they fitted a non-linear mixed-effects models without covariates. Then they developed a multivariate linear model assigning missing covariates implemented by the Gibbs sampler. Finally, to analyze each dataset, they used the complete data method.

Sinha (2008) considered analyzing some interesting clinical data obtained from the Mental Health Program at St. Boniface Hospital Research Centre in Winnipeg, Manitoba, Canada. The study involved a group of 102 patients between the ages of 40 and 89 who underwent elective surgeries to repair an aortic medical condition. The patients were recruited prospectively from consecutive cases seen in the St. Boniface Outpatient Clinic for elective management of abdominal aortic aneurysm (AAA) during the period December 2000 to December 2003. A goal of the study was to observe occurrences of delirium in the AAA patients and to investigate factors that are predictive of the delirium experienced by those patients. The covariates considered in the analysis include "age", "packyrs" (number of packs of cigarettes smoked per day multiplied by the number of years the patient smoked), "etoh" (number of alcoholic drinks consumed in one year), "ppsychom" (number of psychometric medications taken prior to operation), and "pvasoact" (number of vasoactive medications taken prior to operation). Among these covariates, packyrs contained a large proportion of missing values, where the missingness was considered nonigrobale. The author proposed an efficient robust method in

the framework of the maximum likelihood estimation by incorporating a missing data model into the observed data likelihood function. The method was shown to be robust against potential outliers in the data.

Chapter 3

Maximum Likelihood Estimation under Nonignorible Missing Covariates

3.1 Introduction

In this chapter, we discuss the maximum likelihood method for estimating parameters in GLMs involving nonignorable missing covariates. We assume a binomial response model for the missing data mechanism. In the next section, we introduce the model and notation, and describe the likelihood method for analyzing GLMs with nonignorable missing covariates.

3.2 Model and Notation

3.2.1 Full Model

Let $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ be independent observations, where y_i is the ith response variables and $\mathbf{x}_i$ is the ith random vector of p covariates. Conditional on $\mathbf{x}_i$, assume

that y_i follows a distribution in the exponential family:

$$f_{y_i|x_i}(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \phi) = \exp[\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i, \phi)] \tag{3.1}$$

for some functions a, b and c, and ϕ is a scaler dispersion parameter (McCullagh and Nelder, 1997). The canonical parameter is $\theta_i = \mathbf{x}_i^t\boldsymbol{\beta}$, where $\mathbf{x}_i$ is the ith row of the design matrix $\mathbf{X}$, which may contain constant 1 to incorporate an intercept term.

The log-likelihood for Eq.(3.1) is obtained as

$$l(\boldsymbol{\beta}, \phi|y, \mathbf{X}) = \sum_{i=1}^{n}[\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i, \phi)]. \tag{3.2}$$

Here the dispersion parameter ϕ is fixed at unity, for simplicity. The ML estimating equation for $\boldsymbol{\beta}$ can be obtained by taking the derivative of Eq.(3.2) with respect to $\boldsymbol{\beta}$,

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n}\frac{\partial}{\partial \boldsymbol{\beta}}[y_i(\mathbf{x}^T\boldsymbol{\beta}) - b(\theta_i)] \\
&= \sum_{i=1}^{n}[y_i\mathbf{x}_i - \frac{\partial}{\partial \boldsymbol{\beta}}\{b(\theta_i)\}] \\
&= \sum_{i=1}^{n}[y_i\mathbf{x}_i - b'(\theta_i)\mathbf{x}_i] \\
&= \sum_{i=1}^{n}[y_i - \mu_i(\boldsymbol{\beta}, \mathbf{x}_i)]\mathbf{x}_i = \mathbf{0},
\end{aligned} \tag{3.3}$$

where $\mu_i(\boldsymbol{\beta}, \mathbf{x}_i) = E(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = b'(\theta_i)$.

3.2.2 Model for Covariates

Specification of a parametric model plays an important role for the missing covariates (Ibrahim, Chen, Lipsitz and Herring, 1999). In a specified parametric model for the covariates, the indexing parameters of this distribution are usually observed as nuisance

parameters, which are not parameters of inferential interest. To reduce the number of nuisance parameters, some strategies are needed to be implemented when the covariate distribution is specified since the parameter estimation becomes computationally inefficient and intensive due to a large number of nuisance parameters.

When the missing covariates are categorical, Lipsitz and Ibrahim (1996) proposed a joint distribution of the covariates as a product of one-dimensional conditional distribution to reduce the number of nuisance parameters in the covariate distribution.
Here, we consider a joint distribution of the p-dimentional covariate vector $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^t$, denoted by $f(\mathbf{x}_i|\boldsymbol{\alpha})$, as used in Ibrahim, Lipsitz and Chen (1999). We can write

$$\begin{aligned} f(x_{1i}, \ldots\ x_{pi}|\boldsymbol{\alpha}) &= f(x_{pi}|x_{1i}, \ldots\ x_{p-1,i}\,, \boldsymbol{\alpha}_p) \\ &\quad \times f(x_{p-1,i}|x_{1i}, \ldots\ x_{p-2,i}\,, \boldsymbol{\alpha}_{p-1}) \\ &\quad \ldots \\ &\quad \times f(x_{2i}|x_{1i}, \boldsymbol{\alpha}_2) f(x_{1i}|\boldsymbol{\alpha}_1), \end{aligned} \tag{3.4}$$

where $\boldsymbol{\alpha}_j$ is a vector of parameters for the jth conditional distribution that are distinct and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p)$. The above model is quite useful for categorical covariates as well as continuous covariates when there is no natural joint distribution to specify. Eq.(3.4) indicates a sequence of one-dimensional conditional distributions for x_i. When the covariates are completely observed, models are not necessarily to be specified and then the above equation is used for the missing covariates only.

Eq.(3.4) is also useful for mixed covariate case, which often occurs in clinical trials (Ibrahim, Lipsitz and Chen, 1999). When the categorical and continuous covariates are both missing, specifying a joint distribution is quite unrealistic. In this situation, model (3.4) works well and reduce a large number of nuisance parameters. One approach for this mixed covariates case would be first specifying a one-dimensional distribution for the continuous covariates and then attaining the one-dimensional distribution for the categorical covariates by conditioning on the continuous covariates. Since each of the one dimensional conditional distributions in Eq.(3.4) has a distribution of the exponential family, the property of log-concavity still exists.

3.2.3 Model for Missing Data

Models are required when the missing data mechanism is nonignorable. Similarly to Ibrahim, Lipsitz and Chen (1999), here we consider a joint log-linear model for missing data mechanism $f(\mathbf{v}_i|y_i, \mathbf{x}_i, \boldsymbol{\tau})$ as the product of a sequence of one-dimensional conditional distributions. It can be written as

$$\begin{aligned} f(v_{1i}, \ldots\ v_{pi}|y_i, \mathbf{x}_i, \boldsymbol{\tau}) = f(v_{pi}|v_{1i}, \ldots\ v_{p-1,i}\ , y_i, \mathbf{x}_i, \boldsymbol{\tau}_p) \\ \times\ f(v_{p-1,i}|v_{1i}, \ldots\ v_{p-2,i}\ , y_i, \mathbf{x}_i, \boldsymbol{\tau}_{p-1}) \\ \ldots \\ \times f(v_{2i}|v_{1i}, y_i, \mathbf{x}_i, \boldsymbol{\tau}_2) f(v_{1i}|y_i, \mathbf{x}_i, \boldsymbol{\tau}_1), \end{aligned} \tag{3.5}$$

where the jth element of $\boldsymbol{\tau}_j$ in $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \ldots\ \boldsymbol{\tau}_p)$ represents a vector of parameters, and $v_{ji} (j = 1, 2, \ldots, p)$ is a binary variable indicating the missingness of the jth covariate x_{ji}.

For the missing data model (3.5), the likelihood score equations for estimating the nuisance parameters are obtained as

$$\sum_{i=1}^{n} E\left[\frac{\partial \log f\left(v_{1i}, \ldots\, v_{pi} | y_i, \mathbf{x}_i, \boldsymbol{\tau}\right)}{\partial \boldsymbol{\tau}} | y_i, \mathbf{x}_i, \boldsymbol{\tau}\right] = \mathbf{0}.$$

Assuming a logistic model for each of the conditional distribution in (3.5), we obtain the estimating equations for $\boldsymbol{\tau}_p$ as

$$\sum_{i=1}^{n} E\left[\left\{v_{pi} - \eta_{pi}\left(\boldsymbol{\tau}_p, \mathbf{x}_i{}^*\right)\right\} \mathbf{x}_i{}^* | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right] = \mathbf{0}, \tag{3.6}$$

where $\mathbf{x}_i{}^* = (v_{1i}, \ldots\, v_{p-1,i}\,, y_i, \mathbf{x}_i)$ and $\eta_{pi}(\boldsymbol{\tau}_p, \mathbf{x}_i{}^*)$ is the expected value of v_{pi} when $\mathbf{x}_i{}^*$ is given.

When the missingness of one variable affects the probability of missingness in others, the modelling approach (3.5) provides a flexible specification of the joint distribution of the missing-data indicators. It also enables the random sampling to approximate the conditional expectation for the score equation which is obtained from the conditional distribution of the missing covariates given the observed data (Sinha, 2008). The density functions of the right-hand side of Eq.(3.5) satisfy the log-concavity property since each of the univariate distribution can be expressed as a logistic regression. This property facilitates the computations of the maximum likelihood estimates.

In missing data problems, selection of appropriate covariates for the missing data

mechanism is a crucial issue. To check the adequacy of fits of various models in the presence of missing data, we can use the likelihood ratio or Akaike information criterion. Ibrahim, Lipsitz and Chen (1999) suggested that one should not use a very large model for the missing data mechanism, since the model can easily become unidentifiable due to overparameterization. Baker and Laird (1988) mentioned that when the missing data mechanism is nonignorable, the issue of estimability can often arise and characterization of the set of estimable parameters for a certain class of models is still under investigation.

3.3 Methods of Estimation

3.3.1 ML Estimation under Missing Covariates

Suppose $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ is a set of data that might occur when there is no missing value. Also, consider $\mathbf{x}_{\text{obs},i}$ is the observed values and $\mathbf{x}_{\text{mis},i}$ the missing values of $\mathbf{x}_i$. Assuming a nonmonotonic pattern of missing data in $\mathbf{x}_i$, some permutation can be written as $\mathbf{x}_i = (\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{mis},i})$, where $\mathbf{x}_{\text{mis},i}$ is the $p_i \times 1$ vector of missing values of $\mathbf{x}_i$.

The joint density function of the observed data may be obtained as

$$f\left(y_i, \mathbf{x}_{\text{obs},i}\right) = \int f\left(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{mis},i}, \right) d\mathbf{x}_{\text{mis},i}.$$

When the missing data mechanism is nonignorible, it is important to incorporate the missing data model into the observed data likelihood function.
Define

$$v_{ij} = \begin{cases} 1 & \text{if} \quad x_{ij} \text{ missing} \\ 0 & \text{if} \quad x_{ij} \text{ observed} , \end{cases}$$

where x_{ij} is the jth covariate in $\mathbf{x}_i$; $(j = 1, \ldots p)$ and $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots v_{ip})^t$ is the vector of missingness indicators.

The conditional distribution of $\mathbf{v}_i$ has a multinominal distribution $f(\mathbf{v}_i|y_i, \mathbf{x}_i, \boldsymbol{\tau})$ that depends on some parameters $\boldsymbol{\tau}$. Assume that the random vector $\mathbf{x}_i$ follows a density $f(\mathbf{x}_i|\boldsymbol{\alpha})$ that depends on some parameter $\boldsymbol{\alpha}$. Now, for the ith observation, we have the actual observed data that has the joint density $f(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i)$ which can be obtained by integrating out $f(y_i, \mathbf{x}_i, \mathbf{v}_i)$ with respect to $\mathbf{x}_{\text{mis},i}$. It can be written as

$$f(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = \int f(y_i|\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{mis},i}, \boldsymbol{\beta}) f(\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{mis},i}, |\boldsymbol{\alpha}) \times f(\mathbf{v}_i|y_i, \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{mis},i}, \boldsymbol{\tau}) \, d\mathbf{x}_{\text{mis},i}. \tag{3.7}$$

When the missing data mechanism is nonignorable, the inferences on $\boldsymbol{\beta}$ should be based on the full likelihood. This likelihood cannot normally be evaluated in a closed form since the ith unit involves an integral with dimension equal to the dimension of missing observation $\mathbf{x}_{\text{mis},i}$ when calculating the density $f(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\tau})$ for the actual observed data. The objective of the study is to review the algorithms for calculating the ML estimates under the full likelihood. We can write the full likelihood as

$$\begin{aligned} L_i &= f(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i) \\ &= \int f(y_i|\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{mis},i}, \mathbf{v}_i) \, d\mathbf{x}_{\text{mis},i}. \end{aligned} \tag{3.8}$$

The log-likelihood can be written as

$$l_i = \log L_i = \log f\left(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right).$$

Suppose that $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\tau})$. The score equations can be obtained by taking derivatives of the log-likelihood with respect to $\boldsymbol{\gamma}$. We can write

$$\begin{aligned}
\frac{\partial li}{\partial \boldsymbol{\gamma}} &= \frac{1}{f\left(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right)} \frac{\partial}{\partial \boldsymbol{\gamma}} f\left(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right) \\
&= \frac{1}{f\left(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right)} \frac{\partial}{\partial \boldsymbol{\gamma}} \int f\left(y_i, \mathbf{x}_i, \mathbf{v}_i\right) d\mathbf{x}_{\text{mis},i} \\
&= \frac{1}{f\left(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right)} \int \frac{\partial \log f\left(y_i, \mathbf{x}_i, \mathbf{v}_i\right)}{\partial \boldsymbol{\gamma}} f\left(y_i, \mathbf{x}_i, \mathbf{v}_i\right) d\mathbf{x}_{\text{mis},i} \\
&= \int \frac{\partial \log f\left(y_i, \mathbf{x}_i, \mathbf{v}_i\right)}{\partial \boldsymbol{\gamma}} \frac{f\left(y_i, \mathbf{x}_i, \mathbf{v}_i\right)}{f\left(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right)} d\mathbf{x}_{\text{mis},i} \\
&= \int \frac{\partial \log f\left(y_i, \mathbf{x}_i, \mathbf{v}_i\right)}{\partial \boldsymbol{\gamma}} f\left(\mathbf{x}_{\text{mis},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right) d\mathbf{x}_{\text{mis},i} \\
&= E\left[\frac{\partial \log f\left(y_i, \mathbf{x}_i, \mathbf{v}_i\right)}{\partial \boldsymbol{\gamma}} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right] \\
&= E\left[\frac{\partial}{\partial \boldsymbol{\gamma}} \log\left\{f\left(y_i | \mathbf{x}_i\right) f\left(\mathbf{x}_i\right) f\left(\mathbf{v}_i | y_i, \mathbf{x}_i\right)\right\} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right] \\
&= E\left[\frac{\partial}{\partial \boldsymbol{\gamma}} \left\{\log f\left(y_i | \mathbf{x}_i\right) + \log f\left(\mathbf{x}_i\right) + \log f\left(\mathbf{v}_i | y_i, \mathbf{x}_i\right)\right\} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right].
\end{aligned} \tag{3.9}$$

When the observed data $(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i)$ are given, the conditional expectations are taken with respect to $\mathbf{x}_{\text{mis},i}$. So we can find the score equations as

$$E\left[\frac{\partial \log f(y_i|\mathbf{x}_i)}{\partial \boldsymbol{\beta}}|y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right] = \mathbf{0}, \tag{3.10}$$

$$E\left[\frac{\partial \log f(\mathbf{x}_i)}{\partial \boldsymbol{\alpha}}|y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right] = \mathbf{0}, \tag{3.11}$$

$$E\left[\frac{\partial \log f(v_i|y_i, \mathbf{x}_i)}{\partial \boldsymbol{\tau}}|y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\right] = \mathbf{0}. \tag{3.12}$$

Our primary interest lies in the estimation of $\boldsymbol{\beta}$ while $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ being viewed as nuisance parameters. From the above equations, Eq.(3.10) can be solved by using Newton-Raphson or scoring approach as used for complete data in generalized linear models. For the exponential family (3.1), Eq.(3.10) for the ML estimates of $\boldsymbol{\beta}$ takes the following form

$$\sum_{i=1}^{n} E[\{y_i - \mu_i(\boldsymbol{\beta}, \mathbf{x}_i)\}\mathbf{x}_i|y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i] = \mathbf{0}, \tag{3.13}$$

where $\mathbf{x}_i = (\mathbf{x}_{obs,i}, \mathbf{x}_{mis,i})$. This score function is proportional to the observed values of y and x. For the missing data mechanism, the ML estimate of $\boldsymbol{\tau}$ in Eq.(3.12) can be solved in a similar way. Since Eq.(3.11) involves only the distribution of $\mathbf{x}$, typically it is easy to solve for the ML estimates of $\boldsymbol{\alpha}$.

3.3.2 Asymptotic Variance of the ML Estimator

We can show that

$$
\begin{aligned}
E(y_i|\mathbf{x}_i,\boldsymbol{\beta}) &= E\left(\frac{\partial l}{\partial \theta_i}\right) \\
&= \frac{\partial b(\theta_i)}{\partial \theta_i} \\
&= b'(\theta_i) \\
&= \mu_i(\boldsymbol{\beta},\mathbf{x}_i), \qquad (3.14)
\end{aligned}
$$

where $\theta_i = \mathbf{x}_i^t\boldsymbol{\beta}$. The asymptotic variance of the ML estimator of $\boldsymbol{\beta}$ can be obtained from Eq.(3.10), using the Fisher information which takes the form

$$
\begin{aligned}
\mathbf{I}_0(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} E\{\partial \mathbf{U}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^T | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\} - \sum_{i=1}^{n} E\{\mathbf{U}(\boldsymbol{\beta})\mathbf{U}(\boldsymbol{\beta})^T | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\} \\
&\quad + \sum_{i=1}^{n} E\{\mathbf{U}(\boldsymbol{\beta}) | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\} E\{\mathbf{U}(\boldsymbol{\beta})^T | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i\}, \qquad (3.15)
\end{aligned}
$$

where $\mathbf{U}(\boldsymbol{\beta}) = \partial \log f(y_i|\mathbf{x}_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

3.3.3 Newton-Raphson Method

Solving a system of nonlinear equations is not feasible as it can in the case of linear equations since the solution cannot be derived algebraically. The solution must be numerically estimated using an iterative process. The most popular method for solving these nonlinear equations is Newton-Raphson method.

The Newton-Raphson method is a powerful technique for solving nonlinear equations numerically. To obtain the maximum likelihood estimator of $\boldsymbol{\beta}$, here we use a Newton-Raphson algorithm to solve Eq.(3.13). This method is a linear Tylor series approximation where we expand the left side of Eq.(3.13) as a function of $\boldsymbol{\beta}$.
We can write the left side of Eq.(3.13) without the expectation in the form

$$\sum_{i=1}^{n}[\{y_i - \mu_i(\boldsymbol{\beta},\mathbf{x}_i)\}\mathbf{x}_i].$$

Using first order Tylor series around some initial value $\boldsymbol{\beta}_0$ and also using the Fisher scoring technique, we can write

$$\begin{aligned}
&\sum_{i=1}^{n}\{y_i-\mu_i(\boldsymbol{\beta},\mathbf{x}_i)\}\mathbf{x}_i \\
&\quad\cong \sum_{i=1}^{n}\{y_i-\mu_i(\boldsymbol{\beta}_0,\mathbf{x}_i)\}\mathbf{x}_i \\
&\quad+\sum_{i=1}^{n}\left[(\partial/\partial\theta_i)\left\{y_i-\mu_i(\boldsymbol{\beta},\mathbf{x}_i)\right\}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right] \\
&\quad\times\mathbf{x}_i\mathbf{x}_i^t(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \\
&\quad=\sum_{i=1}^{n}d_i(\boldsymbol{\beta}_0,y_i,\mathbf{x}_i)\mathbf{x}_i-\sum_{i=1}^{n}d_i'(\boldsymbol{\beta}_0,y_i,\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^t(\boldsymbol{\beta}-\boldsymbol{\beta}_0), \qquad (3.16)
\end{aligned}$$

where $d_i(\boldsymbol{\beta},y_i,\mathbf{x}_i)=y_i-\mu_i(\boldsymbol{\beta},\mathbf{x}_i)$ and $d_i'(\boldsymbol{\beta}_0,y_i,\mathbf{x}_i)=-(\partial/\partial\theta_i)d_i(\boldsymbol{\beta},y_i,\mathbf{x}_i)$ evaluated at $\boldsymbol{\beta}_0$.

Approximation of Eq.(3.16) leads to an iterative equation for $\boldsymbol{\beta}$, in the form

$$\begin{aligned}
\boldsymbol{\beta}^{(j+1)} &= \boldsymbol{\beta}^{(j)}+\left[\sum_{i=1}^{n}E\left\{d_i'(\boldsymbol{\beta}^{(j)},y_i,\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^t|y_i,\mathbf{x}_{\mathrm{obs},i},\mathbf{v}_i\right\}\right]^{-1} \\
&\quad\times\sum_{i=1}^{n}E\left\{d_i(\boldsymbol{\beta}^{(j)},y_i,\mathbf{x}_i)\mathbf{x}_i|y_i,\mathbf{x}_{\mathrm{obs},i},\mathbf{v}_i\right\}. \qquad (3.17)
\end{aligned}$$

To get the maximum likelihood estimates of $\boldsymbol{\beta}$, Eq.(3.17) requires the calculation of conditional expectations.

3.3.4 Asymptotics

Based on the maximum likelihood, asymptotic existence, consistency and asymptotic normality of fixed effect estimators are proved by Haberman (1977) and Fahrmeir and Kaufmann (1985) under suitable conditions. Some conditions are provided in Haberman (1977) for the asymptotic existence of the MLE by the use of the general theory

for exponential models derived in Berk (1972) and Barndorff-Nielsen (1983).

Under suitable regularity conditions, $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically distributed as Gaussian with mean zero and covariance matrix consistently estimated by $\hat{\mathbf{C}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{C}}^{-1}$, where

$$\hat{\mathbf{B}} = n^{-1}\sum_{i=1}^{n}\Phi_i(\boldsymbol{\beta})\Phi_i(\boldsymbol{\beta})^t,$$

with

$$\begin{aligned}\Phi_i(\boldsymbol{\beta}) &= E\left[\{y_i - \mu_i(\boldsymbol{\beta}, \mathbf{x}_i)\}\,\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right] \\ &= E\left[d_i(\boldsymbol{\beta}, \mathbf{x}_i)\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right],\end{aligned}$$

and

$$\begin{aligned}\hat{\mathbf{C}} &= -n^{-1}\sum_{i=1}^{n} E\left[\left(\frac{\partial d_i(\boldsymbol{\beta}, \mathbf{x}_i)}{\partial \boldsymbol{\beta}}\right)\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right] \\ &\quad -n^{-1}\sum_{i=1}^{n} E\left[d_i(\boldsymbol{\beta}, \mathbf{x}_i)\mathbf{x}_i\left(\frac{\partial \log f(y_i|\mathbf{x}_i)}{\partial \boldsymbol{\beta}}\right) | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right] \\ &\quad +n^{-1}\sum_{i=1}^{n} E\left[d_i(\boldsymbol{\beta}, \mathbf{x}_i)\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right] \\ &\quad \times E\left[\left(\frac{\partial \log f(y_i|\mathbf{x}_i)}{\partial \boldsymbol{\beta}}\right) | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right],\end{aligned}$$

where $d_i(\boldsymbol{\beta}, \mathbf{x}_i) = y_i - \mu_i(\boldsymbol{\beta}, \mathbf{x}_i)$.

Under regularity conditions and using a Taylor series expansion, $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ can be approximated from Eq.(3.13) by

$$\left[-n^{-1}\partial\boldsymbol{\Phi}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\right]^{-1}\left[n^{-1/2}\boldsymbol{\Phi}(\boldsymbol{\beta})\right], \tag{3.18}$$

where $\boldsymbol{\Phi}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Phi_i(\boldsymbol{\beta})$.

In Eq.(3.18), the function $n^{-1/2}\boldsymbol{\Phi}(\boldsymbol{\beta})$ generally has an asymptotic normal distribution as $n \to \infty$ with mean zero and covariance matrix

$$\mathbf{B} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \text{var}\{\Phi_i(\boldsymbol{\beta})\} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E\{\Phi_i(\boldsymbol{\beta})\Phi_i(\boldsymbol{\beta})^t\}, \tag{3.19}$$

where the expectation is with respect to the marginal distribution of y_i.

Now we can show that $-n^{-1}\partial\boldsymbol{\Phi}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ converges in probability to its asymptotic mean when $n \to \infty$. We have

$$\begin{aligned}
&n^{-1}\partial\boldsymbol{\Phi}(\boldsymbol{\beta})/\partial\boldsymbol{\beta} \\
&\quad = n^{-1} \sum_{i=1}^{n} E\left[\left(\frac{\partial d_i(\boldsymbol{\beta}, \mathbf{x}_i)}{\partial\boldsymbol{\beta}}\right)\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right] \\
&\quad + n^{-1} \sum_{i=1}^{n} E\left[d_i(\boldsymbol{\beta}, \mathbf{x}_i)\mathbf{x}_i \left(\frac{\partial \log f(y_i|\mathbf{x}_i)}{\partial\boldsymbol{\beta}}\right) | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right] \\
&\quad - n^{-1} \sum_{i=1}^{n} E\left[d_i(\boldsymbol{\beta}, \mathbf{x}_i)\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right] \\
&\quad \times E\left[\left(\frac{\partial \log f(y_i|\mathbf{x}_i)}{\partial\boldsymbol{\beta}}\right) | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i, \boldsymbol{\gamma}\right],
\end{aligned}$$

where $d_i(\boldsymbol{\beta}, \mathbf{x}_i) = y_i - \mu_i(\boldsymbol{\beta}, \mathbf{x}_i)$. The matrix $-n^{-1}\partial\boldsymbol{\Phi}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ converges in probability as $n \to \infty$ to the matrix

$$\mathbf{C} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E\{\dot{\Phi}_i(\boldsymbol{\beta})\}, \tag{3.20}$$

where

$$\begin{aligned}\dot{\Phi}_i(\boldsymbol{\beta}) &= E\left[\left(\frac{\partial d_i(\boldsymbol{\beta},\mathbf{x}_i)}{\partial\boldsymbol{\beta}}\right)\mathbf{x}_i|y_i,\mathbf{x}_{\text{obs},i},\mathbf{v}_i,\boldsymbol{\gamma}\right]\\ &\quad +E\left[d_i(\boldsymbol{\beta},\mathbf{x}_i)\mathbf{x}_i\left(\frac{\partial\log f(y_i|\mathbf{x}_i)}{\partial\boldsymbol{\beta}}\right)|y_i,\mathbf{x}_{\text{obs},i},\mathbf{v}_i,\boldsymbol{\gamma}\right]\\ &\quad -E\left[d_i(\boldsymbol{\beta},\mathbf{x}_i)\mathbf{x}_i|y_i,\mathbf{x}_{\text{obs},i},\mathbf{v}_i,\boldsymbol{\gamma}\right]\\ &\quad \times E\left[\left(\frac{\partial\log f(y_i|\mathbf{x}_i)}{\partial\boldsymbol{\beta}}\right)|y_i,\mathbf{x}_{\text{obs},i},\mathbf{v}_i,\boldsymbol{\gamma}\right].\end{aligned}$$

By combining (3.19) and (3.20), $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is asymptotically normally distributed with mean zero and covariance matrix $\mathbf{C}^{-1}\mathbf{B}\mathbf{C}^{-1}$. This covariance matrix is estimated by $\hat{\mathbf{C}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{C}}^{-1}$, where $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ were defined earlier, with $\boldsymbol{\beta}$ being replaced by its ML estimator $\hat{\boldsymbol{\beta}}$.

Chapter 4

Binary Regression

4.1 Introduction

A binomial logistic regression is often referred to simply as logistic regression. The logistic regression is a powerful and widely used statistical way of modelling a binary outcome which takes the value "0" or "1" like success or failure with one or more explanatory variables.

The logistic regression is used to predict an outcome variable that is categorical from predictor variables which may be continuous or categorical. Since categorical outcome variable violates the assumption of linearity in normal regression, one can consider the logistic regression rather than the linear regression. Also, when the outcome variable is dichotomous and the predictors tend to show a linear relationship, the logistic regression is used. Suppose that we are interested in predicting whether a patient has a given disease (e.g., diabetes, coronary heart disease), based on observed characteristics of the patient such as age, sex, body mass index, results of various blood tests, etc.. The outcome (response) variable is binary (0/1) and the predictor variables of interest are

several observed characteristics of the patient. In this case, the logistic regression can be used for modelling the binary response as a function of the predictors (Freedman, D. A., 2009).

The logistic regression is frequently used in social and medical sciences, as the outcomes from these experiments are usually binary. When the outcome variable is not discrete, the logistic regression is used by taking a logarithmic transformation on the outcome variable, which allows us to model a nonlinear association in a linear way. It shows the linear regression equation in logarithmic terms which is called the logit. When there is more than one risk factors and the odds ratio needs to be adjusted, it is better to use the logistic regression.

In this chapter, we introduce the response model and notation to define the generalized linear models for binary data and describe the maximum likelihood method for estimating parameters in the binary logistic model. We also present results from a simulation study, which was carried out to investigate the empirical properties of the maximum likelihood approach.

4.2 Model and Notation

4.2.1 Binary Logistic Model

In generalized linear models (GLMs), the logistic regression analysis is often used to investigate the relationship between a binary response variable and a set of explanatory variables. A binary response consists, for example, of success or failure. In the case of disease studies, the outcome is denoted as $Y_1 = 1$ if the disease is present or $Y_2 = 0$,

otherwise. To analyze this kind of variable which does not follow a normal distribution, usually a link function (LOGIT) is used.

Let y_i be the binary response variable that contains only two values,"0" and "1". Then we can define it as

$$y_i = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure} \end{cases}$$

We treat y_i as a realization of a random variable Y_i, which takes the values one and zero with probabilities p_i and $1 - p_i$, respectively. Here, p_i indicates the probability of success and $1 - p_i$ indicates the probability of failure. The distribution of Y_i is called a Bernoulli distribution with parameter p_i, and can be written as

$$Pr(Y_i = y_i) = {p_i}^{y_i}(1 - p_i)^{1-y_i}, \tag{4.1}$$

where $y_i = 0, 1$. The expected value and variance of Y_i can be obtained in the form

$$E(Y_i) = \mu_i = p_i,$$

and

$$\text{var}(Y_i) = {\sigma_i}^2 = p_i(1 - p_i),$$

where the mean and variance depend on the underlying parameter p_i.

Again, suppose that y_i can be treated as a realization of Y_i which takes values from 1 to n_i. Assume that p_i is the probability of success in a given trial, n_i is the number of ith independent observations and y_i is the number of units that succeed. Then, the

binomial distribution of Y_i with parameters n_i and p_i can be written as

$$Y_i \sim B(n_i, p_i). \tag{4.2}$$

The general form of the binomial probability density function is

$$\begin{aligned} f(y_i) &= \binom{n_i}{y_i} {p_i}^{y_i}(1-p_i)^{n_i-y_i} \\ &= \binom{n_i}{y_i}\left(\frac{p_i}{1-p_i}\right)^{y_i}(1-p_i)^{n_i}, \end{aligned} \tag{4.3}$$

where $y_i = 0, 1, \ldots, n_i$. The mean and variance can be obtained as

$$E(Y_i) = \mu_i = n_i p_i,$$

and

$$\text{var}(Y_i) = {\sigma_i}^2 = n_i p_i (1-p_i).$$

Again, the mean and variance depend on the underlying parameter p_i.

Suppose that the logit of the underlying probability p_i is a linear function of the predictors. The systematic structure of the model can be shown as

$$\text{logit}(p_i) = {\mathbf{x}_i}^t \boldsymbol{\beta}, \tag{4.4}$$

where x_i is a vector of covariates and $\boldsymbol{\beta}$ is a vector of regression coefficients.
The model defined in Eq.(4.2) and Eq.(4.4) is the generalized linear model with binomial response and logit link. Taking the exponential in Eq.(4.4),the odds for the ith unit are

$$\frac{p_i}{1-p_i} = \exp\{{\mathbf{x}_i}^t \boldsymbol{\beta}\}. \tag{4.5}$$

Solving Eq.(4.5) for the probability p_i in the logit model, we get

$$p_i = \frac{\exp\{\mathbf{x}_i{}^t\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i{}^t\boldsymbol{\beta}\}} = \frac{e^{\mathbf{x}_i{}^t\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i{}^t\boldsymbol{\beta}}}.$$

There is no explicit solution for the individual effects of predictors, as the right-hand side is a non-linear function of the predictors. To obtain an approximate answer, we can take derivatives with respect to x_j, that is

$$\frac{dp_i}{dx_{ij}} = \beta_j p_i(1-p_i).$$

Thus, the effect of the jth predictor on the probability p_i depends on the coefficient β_j and the value of the probability.

4.2.2 Maximum Likelihood Estimation

For n independent binomial observations, the likelihood function is a product of densities given by Eq.(4.3). The likelihood function of $\boldsymbol{\beta}$ for given data (y_i, n_i, x_i) is

$$L = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i{}^{y_i}(1-p_i)^{n_i-y_i}, \tag{4.6}$$

where p_i depends on the covariates x_i and a vector $\boldsymbol{\beta}$ of p parameters. The maximum likelihood estimates are the values for $\boldsymbol{\beta}$ which maximize the likelihood function (4.6). Taking logarithms on both sides of Eq.(4.6), we get the log-likelihood function

$$\begin{aligned} \log L &= \prod_{i=1}^{n} \log \binom{n_i}{y_i} p_i{}^{y_i}(1-p_i)^{n_i-y_i} \\ &= \sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i \log \frac{p_i}{1-p_i} + n_i \log(1-p_i) \right\}. \end{aligned} \tag{4.7}$$

To obtain the score equation, we take the first derivative of Eq.(4.7) with respect to $\boldsymbol{\beta}$. That is,

$$\begin{aligned}\frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left\{ y_i x_i - n_i \frac{1}{1-p_i} \frac{\partial p_i}{\partial \boldsymbol{\beta}} \right\} \\ &= \sum_{i=1}^{n} \left\{ y_i x_i - n_i \frac{1}{1-p_i} p_i(1-p_i) x_i \right\} \\ &= \sum_{i=1}^{n} \{ y_i x_i - n_i p_i x_i \} \\ &= \sum_{i=1}^{n} (y_i - n_i p_i)\, x_i. \end{aligned} \tag{4.8}$$

The maximum likelihood estimators of $\boldsymbol{\beta}$ are obtained by solving the ML estimating equations, which can be written as

$$\sum_{i=1}^{n} (y_i - n_i \hat{p}_i)\, x_i = 0.$$

Here, $\hat{\boldsymbol{\beta}}$ cannot be obtained explicitly. Some numerical algorithm such as the iteratively reweighted least squares (IRWLS) or Newton-Raphson method can be used to solve the likelihood equations.

4.3 Simulation Study

To assess the performance of the maximum likelihood method, we ran sixteen sets of simulations under four scenarios using the binary regression model with missing covariates. In the first two scenarios, the estimates were studied under correctly specified MAR models and NMAR models, respectively. In the second two scenarios, the estimates were studied under misspecified models for the missing data. We investigated

the empirical properties of the ML estimators of $\hat{\boldsymbol{\beta}}$ which may be adequate to allow the inferential procedures under normal theory for moderate sample sizes.

We computed 95% confidence intervals for the regression coefficients as well as for the nuisance parameters for each of the sixteen sets of simulations to investigate if departures from normality are sufficiently severe to adversely affect normal-theory parametric inferences. If the normality assumption is satisfied, then $t = (\hat{\theta_n} - \theta)/\text{s.e.}(\hat{\theta_n})$ approximately follows a Student's t_{n-p} distribution, where n is the total number of observations in the data set and p is the number of parameters in the model. Then the nominal level of the confidence interval for θ is $100(1-\alpha)$ as obtained from the confidence interval $\hat{\theta_n} \pm t_{n-p;\alpha/2} s.e.(\hat{\theta_n})$.

4.3.1 Investigative Methods

We studied the empirical properties of the maximum likelihood estimators in terms of biases and mean squared errors (MSEs) of the estimators, as well as coverage probabilities (CPs) and average lengths of the individual confidence intervals for the regression parameters.

Bias: The bias of an estimator $\hat{\theta}$ of a parameter θ is obtained as the difference between the expected value of $\hat{\theta}$ and the true value of the parameter θ, given by

$$\begin{aligned} Bias(\hat{\theta}) &= E(\hat{\theta}) - \theta \\ &\approx \left\{ \frac{1}{S} \sum_{s=1}^{S} \hat{\theta}^{(s)} \right\} - \theta \end{aligned}$$

where $\hat{\theta}^{(s)}$ is the estimate of θ obtained at the s^{th} simulation and S is the simulation size.

Mean Square Error: The mean squared error (MSE) of an estimator $\hat{\theta}$ of a parameter θ can be obtained as

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &\approx \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\theta}^{(s)} - \theta\right)^2 \end{aligned}$$

where $\hat{\theta}^{(s)}$ is the estimate of θ obtained at the s^{th} simulation and S is the simulation size.

Coverage Probability: The coverage probability of an estimator $\hat{\theta}$ for a 95% confidence interval on θ is obtained from

$$CP(\hat{\theta}) = \frac{1}{S}\sum_{s=1}^{S} I\left\{|\hat{\theta}^{(s)} - \theta| \leq 1.96 \times \mathrm{SE}(\hat{\theta}^{(s)})\right\},$$

where $\mathrm{SE}(\hat{\theta})$ is an estimate of the standard errors of $\hat{\theta}$ and I $\{\}$ is an indicator variable.

The 95% confidence interval for θ is obtained as

$$CI(\hat{\theta}) = \hat{\theta} \pm 1.96 \times \mathrm{SE}(\hat{\theta})$$

Average Length: The length of a confidence interval for θ is obtained as the difference between two confidence limits:

$$\mathrm{L}^c(\hat{\theta}) = \{\hat{\theta} + 1.96 \times \mathrm{SE}(\hat{\theta})\} - \{\hat{\theta} - 1.96 \times \mathrm{SE}(\hat{\theta})\}$$

The the average length is obtained by

$$\text{Ave.L}^c(\hat{\theta}) \quad = \quad \frac{1}{S}\sum_{s=1}^{S} \text{L}^c(\hat{\theta}^{(s)}),$$

where S is the simulation size.

4.3.2 Binary Logistic Model for Simulation

Consider a binary-logit model with two continuous covariates x_1 and x_2. The covariates (x_{i1}, x_{i2}) for the ith individual are assumed to be independent normal with mean $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_x$. For the ith response variable y_i, assume

$$y_i|x_{1i}, x_{2i} \sim \text{independent Bernoulli}(\mu_i), i = 1, 2 \ldots n;$$

and

$$\theta_i = \text{logit}(\mu_i) = \log\{\mu_i/(1-\mu_i)\} = \mathbf{x}_i{}^t\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \tag{4.9}$$

In this setting, we have $\mu_i = \mu_i(\boldsymbol{\beta}, \mathbf{x}_i) = E\{y_i|\mathbf{x}_i, \boldsymbol{\beta}\} = \exp(\mathbf{x}_i{}^t\boldsymbol{\beta})/\{1+\exp(\mathbf{x}_i{}^t\boldsymbol{\beta})\}$ and $\text{var}\{y_i|\mathbf{x}_i, \boldsymbol{\beta}\} = \sigma_i^2(\boldsymbol{\beta}, \mathbf{x}_i) = \exp(\mathbf{x}_i{}^t\boldsymbol{\beta})/\{1+\exp(\mathbf{x}_i{}^t\boldsymbol{\beta})\}^2$. Also, assume a logistic regression model for the missing data mechanism in the form

$$f_{v_{2i}|y_i,x_{1i},x_{2i}}(v_{2i}|y_i, x_{1i}, x_{2i}, \boldsymbol{\tau}) = \pi_i{}^{v_{2i}}(1-\pi_i)^{(1-v_{2i})};$$

$$\text{logit}(\pi_i) = \text{logit}\{P(v_{2i} = 1|y_i, x_{1i}, x_{2i}, \boldsymbol{\tau})\} = \tau_0 + \tau_1 x_{1i} + \tau_2 x_{2i} + \tau_3 y_i, \tag{4.10}$$

where the values of the covariate x_1 are completely observed ($v_{1i} = 0$ for all i) and the values of x_2 are missing according to the missing data model (4.10).

The following methods are used in the simulations:

MAR: When the data are generated by considering the missing data mechanism $\tau_3 = 0$ and the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution are estimated by considering $\tau_3 = 0$.

NMAR: When the data are generated by considering the missing data mechanism $\tau_3 \neq 0$ and the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution are estimated by considering $\tau_3 \neq 0$.

Correctly Specified: When the data sets are generated either considering the MAR method and the estimates are obtained under the MAR method or the data sets are generated by considering the NMAR method and the estimates are obtained under the NMAR method.

Misspecified: When the data sets are generated either considering the MAR method and the estimates are obtained under the NMAR method or the data sets are generated by considering the NMAR method and the estimates are obtained under the MAR method.

For the binary regression, a series of 1000 data were generated from the binary

model (4.9) when the sample sizes are n=100 and 200. We generated the values of the covariates (x_{i1}, x_{i2}) from a bivariate normal distribution with mean $\boldsymbol{\mu}_x = (2,1)^t$ and covariance matrix

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} 1 & .2 \\ .2 & 1 \end{pmatrix}.$$

In our notation, $\boldsymbol{\alpha} = (\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2})^t = (2,1,1,1,0.2)^t$. The regression coefficients were fixed at $\boldsymbol{\beta} = (-2,1,0.5)^t$ and $\boldsymbol{\beta} = (-2,0.5,1)^t$.

We obtain the ML estimates of the regression parameters $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ by the iterative Newton-Raphson method described earlier.

4.3.3 Results and Discussion for the Binary Model

In this section, we studied empirical biases and mean square errors of the ML estimators under four scenarios, where each scenario contains four sets of simulations conducted under the binary regression model with missing covariates. We also computed 95% confidence intervals and average lengths of the confidence intervals for the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\tau}$ of the missing data mechanism for different sample sizes.

When the missing data mechanism is MAR, for the two choices of $\boldsymbol{\beta} = (-2,1,0.5)^t$ and $\boldsymbol{\beta} = (-2,0.5,1)^t$, the data contained roughly 34.8% and 33.7% missing values, respectively. On the other hand, when the missing data mechanism is NMAR, for the two choices of $\boldsymbol{\beta} = (-2,1,0.5)^t$ and $\boldsymbol{\beta} = (-2,0.5,1)^t$, the data contained roughly 44.2% and 40% missing values, respectively. The simulation results are discussed here under four scenarios as presented bellow.

Table 4.1: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Correctly specified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	-2	-0.13830	0.49592	-0.06633	0.21766
β_1	1	0.13338	0.12268	0.09268	0.05195
β_2	0.5	0.03445	0.12408	0.00766	0.05328
μ_1	2	-0.00264	0.01069	0.00069	0.00502
μ_2	1	-0.25879	0.08037	-0.25189	0.07036
σ_1^2	1	-0.00952	0.01949	0.00013	0.01039
σ_2^2	1	-0.15111	0.04830	-0.15835	0.03964
σ_{12}	0.2	-0.13801	0.03454	-0.13462	0.02640

Scenario 1: True model: MAR ($\tau_3 = 0$); fitted model: MAR ($\tau_3 = 0$).

In this scenario, the data were generated by using the MAR model, where the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-4, 1, 1, 0)^t$ and the data were fitted by using the MAR model as well. In that sense, the model is correctly specified. The regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and $\boldsymbol{\beta} = (-2, 0.5, 1)^t$. As the model is correctly specified, we expect to get negligible biases, small MSEs and good coverage probabilities for the estimators.

Table 4.1 presents the simulated biases and mean squared errors, and Table 4.2 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes, where the regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 1, 0.5)^t$

Table 4.2: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Correctly specified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	-2	0.95401	2.64735	0.95910	1.80479
β_1	1	0.94652	1.19274	0.94581	0.80706
β_2	0.5	0.95401	1.29181	0.95092	0.88247

Table 4.3: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Correctly specified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	-2	-0.15705	0.56732	-0.06617	0.23958
β_1	0.5	0.15070	0.10462	0.12908	0.053273
β_2	1	0.08728	0.16743	0.03122	0.07741
μ_1	2	0.00533	0.010915	0.00110	0.00479
μ_2	1	-0.22314	0.07167	-0.22548	0.05755
σ_1^2	1	-0.01026	0.02384	-0.00662	0.0098
σ_2^2	1	-0.15040	0.05244	-0.14562	0.03354
σ_{12}	0.2	-0.11989	0.02844	-0.11791	0.02175

Table 4.4: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Correctly specified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	-2	0.95685	2.73537	0.95563	1.85475
β_1	0.5	0.93851	1.07735	0.91022	0.73181
β_2	1	0.95793	1.47877	0.94634	0.99954

and the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-4, 1, 1, 0)^t$.

On the other hand, to study the impact of changes to the coefficients of missing covariates, we consider the regression coefficients $\boldsymbol{\beta} = (-2, 0.5, 1)^t$ keeping the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism similar to those as used for correctly specified models under the MAR setting.

Table 4.3 presents the simulated biases and mean squared errors, and Table 4.4 presents the coverage probabilities and mean lengths of the confidence intervals for different sample sizes under the same setting as above but the regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 0.5, 1)^t$.

The maximum likelihood method appears to perform well for the correctly specified MAR mechanism under different sample sizes. The method provides small biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution in both Tables 4.1 and 4.3. The confidence intervals

have good coverages, which are close to the nominal 95% confidence level.

As expected, the biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ become smaller when we increase the sample size. The coverage probabilities for the parameter estimates become closer to the nominal levels and the mean lengths based on confidence interval become smaller when we increase the sample size. The confidence intervals for the regression coefficients $\boldsymbol{\beta}$ slightly lose coverages due to the choice of different values of $\boldsymbol{\beta} = (-2, 0.5, 1)^t$. As shown in Table 4.4, β_1 loses its coverage to 0.91022 from 0.94581 when the sample size is $n = 200$, but it still has the good coverages. The slight loss of coverages is due to the small bias in estimation under moderate sample sizes. For larger sample sizes, the bias decreases and coverage probability improves.

Scenario 2: True model: NMAR ($\tau_3 = 1$); fitted model: NMAR ($\tau_3 \neq 0$).
In this scenario, the data were generated by using the NMAR model where the parameters of the missing data mechanism were chosen as $\boldsymbol{\tau} = (-4, 1, 1, 1)^t$ and the data were fitted by using the NMAR model as well. In this case, the model is correctly specified. The regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and $\boldsymbol{\beta} = (-2, 0.5, 1)^t$. Here we expect to get unbiased and efficient estimators.

Table 4.5 presents the simulated biases and mean squared errors and Table 4.6 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes, where the regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and the parameters of the missing-data model were chosen as $\boldsymbol{\tau} = (-4, 1, 1, 1)^t$.

Table 4.5: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Correctly specified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	-2	-0.13911	0.56093	-0.05877	0.20660
β_1	1	0.05266	0.10679	0.019907	0.04151
β_2	0.5	0.03752	0.11575	0.02235	0.04981
τ_0	-4	-0.70373	4.25333	-0.30515	1.05603
τ_1	1	0.13530	0.21796	0.05548	0.08324
τ_2	1	0.23324	0.54397	0.12389	0.19967
τ_3	1	0.09170	1.28212	0.01686	0.19807
μ_1	2	0.00281	0.01061	0.00045	0.00521
μ_2	1	0.02093	0.02167	0.01216	0.01187
σ_1^2	1	0.00268	0.02095	0.00183	0.00999
σ_2^2	1	0.03100	0.05036	0.01929	0.02174
σ_{12}	0.2	0.00675	0.01867	0.01053	0.01006

Table 4.6: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Correctly specified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	-2	0.94769	2.61472	0.967	1.78360
β_1	1	0.94869	1.16784	0.966	0.79773
β_2	0.5	0.94668	1.21611	0.944	0.83585
τ_0	-4	0.96881	5.026281	0.943	3.15704
τ_1	1	0.95775	1.55262	0.952	1.00601
τ_2	1	0.95875	2.12204	0.941	1.37445
τ_3	1	0.95573	2.54862	0.957	1.68455

On the other hand, to study the impact of changes to the coefficients of missing covariates, we consider the regression coefficients $\boldsymbol{\beta} = (-2, 0.5, 1)^t$ keeping the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism similar to those as used for correctly specified models under the NMAR setting.

Table 4.7 presents the simulated biases and mean squared errors and Table 4.8 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes under the same setting but the regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 0.5, 1)^t$.

The maximum likelihood method appears to perform well for the correctly specified model under the NMAR mechanism for different sample sizes. The method provides

Table 4.7: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Correctly specified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	-2	-0.17290	0.65864	-0.06385	0.24297
β_1	0.5	0.03860	0.09034	0.00566	0.03502
β_2	1	0.07162	0.17567	0.03740	0.07135
τ_0	-4	-0.68744	3.22699	-0.38205	1.13386
τ_1	1	0.13598	0.21250	0.07739	0.08731
τ_2	1	0.26980	0.67595	0.13740	0.23439
τ_3	1	-0.01712	0.51816	0.00523	0.20235
μ_1	2	-4.8048E-05	0.00965	0.00132	0.00488
μ_2	1	0.02246	0.02071	0.01498	0.01216
σ_1^2	1	0.00507	0.02077	0.00452	0.00964
σ_2^2	1	0.04821	0.05101	0.02184	0.02353
σ_{12}	0.2	0.01728	0.01878	0.00399	0.00966

Table 4.8: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Correctly specified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	-2	0.94294	2.71037	0.955	1.84205
β_1	0.5	0.94795	1.05521	0.951	0.71442
β_2	1	0.93794	1.42141	0.946	0.97385
τ_0	-4	0.97498	5.02201	0.955	3.18308
τ_1	1	0.96397	1.50315	0.94	0.97763
τ_2	1	0.96096	2.28808	0.927	1.48164
τ_3	0	0.94595	2.50098	0.945	1.65639

unbiased estimators and small mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution. But estimates of $\boldsymbol{\tau}$ for the missing data mechanism show slightly large biases in both Tables 4.5 and 4.7. The parameter estimates have good coverages which are close to the nominal level. Here τ_2 loses slight coverages with the value of 0.927, as shown in Table 4.8 when the sample size is $n = 200$.

As expected, the biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ become smaller when we increase the sample size. The coverage probabilities for the parameter estimates become closer to the nominal levels and the mean lengths based on confidence intervals become smaller when we increase the sample size.

Table 4.9: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Misspecified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	-2	-0.11786	0.52308	-0.09645	0.24979
β_1	1	0.05453	0.10202	0.04459	0.05005
β_2	0.5	0.02849	0.09784	0.01212	0.04478
τ_0	-4	-0.80274	4.07318	-0.42001	1.30592
τ_1	1	0.14492	0.24338	0.08468	0.09054
τ_2	1	0.28534	0.74611	0.13966	0.27604
τ_3	0	-0.04386	0.55968	-0.04802	0.26663
μ_1	2	-0.00016	0.010008	-0.00113	0.00502
μ_2	1	0.01652	0.02074	0.00961	0.01149
σ_1^2	1	0.00174	0.02050	-0.00514	0.01008
σ_2^2	1	0.02826	0.04614	0.02421	0.02519
σ_{12}	0.2	0.00267	0.01865	0.00564	0.00899

Scenario 3: True model: MAR ($\tau_3 = 0$); fitted model: NMAR($\tau_3 \neq 0$).

In this scenario, the data were generated by using the MAR model where the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-4, 1, 1, 0)^t$ and the data were fitted by using the NMAR model $\boldsymbol{\tau} = (-4, 1, 1, 1)^t$. In this case, the model is misspecified. The regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and $\boldsymbol{\beta} = (-2, 1, 0.5)^t$. Here we expect to get larger biases, MSEs and bad coverages for the estimators under the misspecified model as compared to the correctly specified models.

Table 4.9 presents the simulated biases and mean squared errors and Table 4.10 presents the coverage probabilities and the mean lengths of the confidence intervals for

Table 4.10: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Misspecified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	-2	0.94183	2.57221	0.938	1.77996
β_1	1	0.94784	1.15772	0.941	0.80284
β_2	0.5	0.95687	1.16200	0.951	0.80151
τ_0	-4	0.96891	5.36726	0.953	3.33533
τ_1	1	0.95487	1.58073	0.943	1.02386
τ_2	1	0.95988	2.33606	0.922	1.49519
τ_3	0	0.96891	2.74035	0.941	1.82256

different sample sizes, where the regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-4, 1, 1, 0)^t$.

On the other hand, to study the impact of changes to the coefficients of missing covariates, we consider the regression coefficients $\boldsymbol{\beta} = (-2, 0.5, 1)^t$ keeping the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism similar to those as used for misspecified models under the same setting.

Table 4.11 presents the simulated biases and mean squared errors, and Table 4.12 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes under same setting, but the regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 0.5, 1)^t$.

Table 4.11: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Misspecified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	-2	-0.12326	0.53448	-0.04423	0.22837
β_1	0.5	0.03932	0.078824	0.01717	0.03534
β_2	1	0.03724	0.14559	0.01346	0.06095
τ_0	-4	-0.87216	4.50733	-0.34783	1.14227
τ_1	1	0.15858	0.23369	0.07499	0.07674
τ_2	1	0.32285	0.86813	0.11979	0.25882
τ_3	0	-0.11847	0.75405	-0.05043	0.26919
μ_1	2	0.00165	0.01039	0.00046	0.00528
μ_2	1	0.02229	0.02078	0.01005	0.00942
σ_1^2	1	-0.00656	0.02065	-0.00326	0.01025
σ_2^2	1	0.03859	0.05044	0.01707	0.02140
σ_{12}	0.2	0.00537	0.01829	0.00265	0.00826

Table 4.12: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Misspecified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	-2	0.95436	2.61883	0.949	1.79373
β_1	0.5	0.96045	1.03547	0.955	0.70848
β_2	1	0.94118	1.32528	0.955	0.91731
τ_0	-4	0.98174	5.38553	0.952	3.20270
τ_1	1	0.96450	1.53669	0.953	0.96499
τ_2	1	0.95740	2.46991	0.942	1.54279
τ_3	0	0.94929	2.78605	0.939	1.81549

The maximum likelihood method still appears to perform well for the misspecified model under different sample sizes. The method still provides small biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution. The method produces larger biases and mean squared errors for $\boldsymbol{\tau}$ of the missing data mechanism compared to Scenario 2 for both choices of $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and $\boldsymbol{\beta} = (-2, 0.5, 1)^t$. The parameter estimates have good coverages that are close to their nominal levels.

As expected, the biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ become smaller when we increase sample size. Although biases are large for $\boldsymbol{\tau}$ with n=100, they become smaller when we increase the sample size. The coverage probabilities for the parameter estimates become closer to the nominal levels, and mean lengths of confidence intervals become smaller when we

Table 4.13: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Misspecified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	-2	-0.03009	0.55848	0.07051	0.22132
β_1	1	0.14316	0.13654	0.09799	0.05506
β_2	0.5	-0.12182	0.16052	-0.15326	0.08051
μ_1	2	0.00060	0.00970	0.00232	0.00513
μ_2	1	-0.33276	0.12984	-0.33644	0.12185
σ_1^2	1	-0.01189	0.01973	-0.00161	0.00978
σ_2^2	1	-0.15891	0.05522	-0.17576	0.048134
σ_{12}	0.2	-0.15666	0.04368	-0.16947	0.03918

increase the sample size. The estimates of regression coefficients $\boldsymbol{\beta}$ and nuisance parameters $\boldsymbol{\tau}$ lose small amount of coverages for the choice of $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ as shown in Table 4.10.

Scenario 4: True model: NMAR ($\tau_3 = 1$); fitted model: MAR ($\tau_3 = 0$).
In this scenario, the data were generated by using the NMAR model where the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-4, 1, 1, 1)^t$ and the data were fitted by using the MAR model $\boldsymbol{\tau} = (-4, 1, 1, 0)^t$. In this case, the model is misspecified. The regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and $\boldsymbol{\beta} = (-2, 1, 0.5)^t$.

Table 4.13 presents the simulated biases and mean squared errors, and Table 4.14 presents the coverage probabilities and the mean lengths of the confidence intervals for

Table 4.14: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Misspecified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	-2	0.94405	2.62845	0.94344	1.77738
β_1	1	0.94286	1.19478	0.93810	0.80189
β_2	0.5	0.91548	1.37637	0.89861	0.93329

Table 4.15: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Misspecified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	-2	0.02397	0.53173	0.12072	0.23781
β_1	0.5	0.13680	0.10290	0.12353	0.04917
β_2	1	-0.06889	0.19882	-0.13774	0.09662
μ_1	2	-0.00035	0.010798	0.00428	0.00491
μ_2	1	-0.28008	0.09548	-0.28859	0.09117
σ_1^2	1	-0.00441	0.02022	-0.00634	0.01041
σ_2^2	1	-0.17319	0.05536	-0.18470	0.04887
σ_{12}	0.2	-0.12516	0.03246	-0.14027	0.02935

Table 4.16: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in binary logistic regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Misspecified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	-2	0.94158	2.69793	0.92300	2.04204
β_1	0.5	0.93471	1.06092	0.92405	0.80070
β_2	1	0.92554	1.58067	0.88713	1.19307

different sample sizes, where the regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-4, 1, 1, 1)^t$.

On the other hand, to study the impact of changes to the coefficients of missing value covariates, we consider the regression coefficients $\boldsymbol{\beta} = (-2, 0.5, 1)^t$ keeping the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism similar to those as used for misspecified models under the same setting.

Table 4.15 presents the simulated biases and mean squared errors, and Table 4.16 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes under same setting, but the regression coefficients were fixed at $\boldsymbol{\beta} = (-2, 0.5, 1)^t$.

The maximum likelihood method doesn't appear to perform well under this misspecified method for different sample sizes. The method provides slightly large biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance

parameters $\boldsymbol{\alpha}$ of the covariate distribution. Also the coverages are not very close to the nominal level 95%.

As expected, the biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ become larger when we increase the sample sizes. The coverage probabilities for the parameter estimates slightly moved away from the nominal levels. The mean lengths based on confidence interval and coverage probabilities become smaller when we increase the sample size. The regression coefficients β_2 loses it's coverages for both choices of $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and $\boldsymbol{\beta} = (-2, 0.5, 1)^t$. As seen in Table 4.14 and 4.16 the coverages for β_2 decreases when the sample size increases from n=100 to n=200.

It is clear from above four scenarios that Scenario 2 and 3 provide approximately unbiased estimates of the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism under all simulation configurations considered. On the other hand, Scenario 1 and 4 provide large biases of the estimators in both regression parameters and nuisance parameters. For example, in Tables 4.1 and 4.3, under correctly specified method, the estimator of β_1 gives large bias 0.13338 and 0.15070 respectively when $n = 100$. As shown in Tables 4.13 and 4.15, under the misspecified method, the estimator of β_1 gives large biases 0.14316 and 0.13680 respectively when $n = 100$.

When comparing four scenarios, our studied method works better for Scenario 2 and worst for Scenario 4. The Scenario 2 provides small biases for all the parameter

estimates under the correctly specified model. For example, in Tables 4.5 and 4.7, the estimator of β_1 gives smaller biases 0.05266 and 0.03860, respectively when $n = 100$. Also, Scenario 3, provides small biases for all the parameter estimates under the mis-specified model. For example, in Tables 4.9 and 4.11, the estimator of β_1 gives smaller biases 0.05453 and 0.03932 respectively when $n = 100$.

Scenario 2 also gives the better coverage probabilities and mean lengths for the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\tau}$ of the missing data mechanism as shown in Tables 4.6 and 4.8 under the correctly specified NMAR mechanism. As we expected, the coverage probabilities are very close to the nominal level 95% which gradually improve and the mean lengths gradually decrease when we increase the sample size for both choice of $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and $\boldsymbol{\beta} = (-2, 0.5, 1)^t$. Although Scenario 3 gives good coverages for all parameter estimates, it loses a small amount of coverages when we increase the sample size for the choice of $\boldsymbol{\beta} = (-2, 0.5, 1)^t$. The coverages decrease for both $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ accordingly with increased sample sizes.

Again, in Scenario 1, the estimators of β_1 gives empirical coverage probabilities of 0.94652 and 0.93851 respectively in Tables 4.2 and 4.4. In Scenario 4, the estimators of β_2 gives empirical coverage probabilities 0.91548 and 0.92554 in Tables 4.14 and 4.16 respectively, which are quite smaller than other scenarios.

Here we should note that there is no big bias for all estimates under the four scenarios. For the MSEs and mean lengths of the estimators of model parameters with different sample sizes, the empirical study shows that the MSEs and mean lengths decrease as

the number of sample size n increases. In terms of CPs, Scenarios 1, 2 and 3 give good coverages under both specified and misspecified models.

Chapter 5

Poisson Regression

5.1 Introduction

In an experiment, under some ideal conditions, if independent successive events occur in the same rate, the Poisson model is appropriate for the number of events observed (McCullagh and Nelder, 1989). A Poisson regression model is sometimes known as a log-linear model. The log-linear model for Poisson regression is commonly used to model the mean response of counts as a function of covariates.

The analysis of count data dominates an important place in applied statistics in many areas. The Poisson regression is used to model count data. We experience count data when the outcome of interest takes only non-negative integer values, such as the number of children in a family, the number of accidents at 10 different intersections, the number of domestic violence incidents in a month, and so on. Also, we may be interested in identifying factors or covariates that are predictive of the outcome of interest. To model these type of count data we can use Poisson regression where the numbers are considered as counts rather than ranking.

In medical and epidemiologic studies, data are often obtained in which the dependent variable is a count such as the number of cancer deaths, that is described by a set of predictor variables. A particular interest can occur in epidemiologic follow-up studies where data are organized into a format similar to that of a life table. One or more factors may affect the survival experience under the cohort study that could include categorical variables (e.g. race, sex) or grouped values of exposure variables (Frome, 1983). To deal with these situations, Poisson models can be used.

An example is discussed in Frome (1983) using the best dose-response data for human cancer that are obtained by Doll (1971) in a study of cigarette smoking in British physicians. These data can be used to illustrate the Poisson regression methods using both log-linear and nonlinear regression models. Another common approach is to categorize the data into two categories; (1): one of the eggs hatched or none did, (2): more ordered categories such as no eggs hatched, 1-2 eggs hatched, 3 or more eggs hatched. Thus, this approach leads a logistic regression model but it throws away real information and often gives lower power.

In this chapter, we introduce the response model and notation to define the generalized linear model for count data and describe the maximum likelihood method for estimating parameters in a log-linear model. We also present results from a simulation study, which was carried out to investigate the empirical properties of the maximum likelihood approach.

5.2 Model and Notation

5.2.1 Poisson Model for Count Data

The Poisson distribution of a random variable Y with parameter μ can be expressed by the probability density function

$$Pr(Y_i = y_i) = \frac{e^{-\mu}\mu^y}{y!}, \tag{5.1}$$

where $y = 0, 1, \ldots$ and $\mu > 0$. The mean and variance of this distribution can be shown as

$$E(Y) = \text{var}(Y) = \mu.$$

Since the mean and variance are equal, any factor that affects one will also affect the other.

Log-Linear Model

Suppose that we have a sample of n observations. Treating y_i as a realization of independent Poisson random variable Y_i which takes the values $1, 2, \ldots, n$ with mean μ_i, we can write

$$\mu_i = \mathbf{x}_i{}^t\boldsymbol{\beta}, \tag{5.2}$$

where the expected count μ_i depends on a vector of explanatory variables x_i. There are some limitations of this model. The linear predictor on the right hand side of this Eq.(5.2) can assume any real value, whereas the Poisson mean on the left hand side has to be non-negative. To deal with this problem, we can consider logarithm of the mean on the left hand side, assuming that the transformed mean follows a linear model. It

can be written as

$$\log(\mu_i) = \mathbf{x}_i{}^t\boldsymbol{\beta}, \tag{5.3}$$

where $\eta_i = \mathbf{x}_i{}^t\boldsymbol{\beta}$ is considered as a generalized linear model with the log link.

5.2.2 Maximum Likelihood Estimation

For n independent Poisson observations, the likelihood function is a product of probabilities given by Eq.(5.1). The likelihood function is

$$L = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!}, \tag{5.4}$$

where μ_i depends on the covariates x_i and a vector $\boldsymbol{\beta}$ of p parameters. The maximum likelihood estimates are the values for $\boldsymbol{\beta}$ which maximize the likelihood function in Eq.(5.4). Taking logarithms on both sides of Eq.(5.4) and ignoring the constant term, we get the log-likelihood function

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\{y_i\log(\mu_i) - \mu_i\} \\ &= \sum_{i=1}^{n}\{y_i x_i{}^t\boldsymbol{\beta} - \exp(\mathbf{x}_i{}^t\boldsymbol{\beta})\}. \end{aligned} \tag{5.5}$$

To obtain the score equation, we take the first derivative of Eq.(5.5) with respect to $\boldsymbol{\beta}$, which gives

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\{y_i x_i - x_i\mu_i\}.$$

The maximum likelihood estimators of $\boldsymbol{\beta}$ are obtained by solving the ML estimating equations. It can be shown that the maximum likelihood estimates in log-linear Poisson

models satisfy the estimating equations

$$\mathbf{X}^t\mathbf{X} = \mathbf{X}^t\hat{\boldsymbol{\mu}},$$

where $\mathbf{X}$ is the design matrix, $\mathbf{y}$ is the response vector, and $\hat{\boldsymbol{\mu}}$ is a vector of fitted values, which is obtained from the MLEs $\hat{\boldsymbol{\beta}}$ by exponentiating the linear predictor $\hat{\eta}_i = \mathbf{x}_i{}^t\hat{\boldsymbol{\beta}}$.

In general, $\hat{\boldsymbol{\beta}}$ cannot be obtained in a closed form. Some numerical algorithm such as the iteratively reweighted least squares (IRWLS) or Newton-Raphson method can be used to solve the equation.

When the events of interest follow the Poisson distribution, the maximum likelihood estimates can be obtained by the IRWLS algorithm which is equivalent to using the method of scoring (Frome, 1983). Under regularity conditions, Poisson regression models include not only log-linear models but also quasilinear and intrinsically nonlinear models. The selection of an approach enables one to describe the relation between the dependent variable and the predictor variables through the regression model.

5.3 Simulation Study

To assess the performance of the maximum likelihood method, we ran sixteen sets of simulations under four scenarios using the Poisson regression model with missing covariates. In the first two scenarios, the estimates were studied under correctly specified MAR models and NMAR models respectively. In the second two scenarios, the estimates were studied under misspecified models for the missing data. We investigated the empirical properties of the ML estimators of $\hat{\boldsymbol{\beta}}$ which may be adequate to allow the

inferential procedures under normal theory for moderate sample sizes.

We computed 95% confidence intervals for the regression coefficients as well as for the nuisance parameters for each of the sixteen sets of simulations to investigate if departures from normality are sufficiently severe to adversely affect normal-theory parametric inferences. If the normality assumption is satisfied, then $t = (\hat{\theta_n} - \theta)/\text{s.e.}(\hat{\theta_n})$ approximately follows a Student's t_{n-p} distribution, where n is the total number of observations in the data set and p is the number of parameters in the model. Then the nominal level of the confidence interval for θ is $100(1-\alpha)$ as obtained from the confidence interval $\hat{\theta_n} \pm t_{n-p;\alpha/2} s.e.(\hat{\theta_n})$.

5.3.1 Poisson Model for Simulation

Consider a Poisson model with two covariates x_1 and x_2. The covariates (x_{i1}, x_{i2}) for the ith individual are assumed to be independent normal with mean $\boldsymbol{\mu}_x$ and covariance matrix $\boldsymbol{\Sigma}_x$. For the ith response variable y_i, assume

$$y_i | x_{1i}, x_{2i} \sim \text{independent Poisson}(\mu_i), i = 1, 2 \ldots n;$$

and

$$\theta_i = \log(\mu_i) = \mathbf{x_i}^t \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \tag{5.6}$$

In this setting, we have $\mu_i = \mu_i(\boldsymbol{\beta}, \mathbf{x_i}) = E\{y_i | \mathbf{x_i}, \boldsymbol{\beta}\} = \exp(\mathbf{x_i}^t \boldsymbol{\beta})$ and $\text{var}\{y_i | \mathbf{x_i}, \boldsymbol{\beta}\} = \sigma_i^2(\boldsymbol{\beta}, \mathbf{x_i}) = \exp(\mathbf{x_i}^t \boldsymbol{\beta})$. Here, the values of the covariate x_1 are completely observed ($v_{1i} = 0$ for all i) and some values of x_2 are missing according to the missing data model

(4.10).

For the Poisson regression, a series of 1000 data were generated from Poisson model (5.6) when the sample sizes are n=100 and 200. We generated the values of the covariates (x_{i1}, x_{i2}) from a bivariate normal distribution with mean $\boldsymbol{\mu}_x = (0.5, 0.5)^t$ and covariance matrix

$$\boldsymbol{\Sigma}_x = \begin{pmatrix} .5 & .1 \\ .1 & .5 \end{pmatrix}.$$

In our notation, $\boldsymbol{\alpha} = (\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1 x_2})^t = (.5, .5, .5, .5, 0.1)^t$. The regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$.

We obtain the ML estimates of the regression parameters $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ by the iterative Newton-Raphson method described earlier.

5.3.2 Results and Discussion for the Poisson Model

In this section, we studied empirical biases and mean square errors of the ML estimators under four scenarios, where each scenario contains four sets of simulations conducted under the Poisson regression model with missing covariates. We also computed 95% confidence intervals and average lengths of the confidence intervals for the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\tau}$ of the missing data mechanism for different sample sizes.

When the missing data mechanism is MAR, for the two choices of $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$, the data contained roughly 21.1% and 19% missing values, respectively. On the other hand, when the missing data mechanism is NMAR, for the two

Table 5.1: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Correctly specified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	0.5	0.01193	0.01145	0.01292	0.00493
β_1	0.5	-0.00201	0.00653	0.000845	0.00286
β_2	1	0.00113	0.00596	0.00036	0.00272
μ_1	0.5	0.00366	0.00468	0.00032	0.00256
μ_2	0.5	-0.01108	0.00546	-0.01602	0.00298
σ_1^2	0.5	-0.00583	0.00506	-0.00045	0.00243
σ_2^2	0.5	0.00162	0.00554	-0.00299	0.00293
σ_{12}	0.1	0.00152	0.00271	-0.00104	0.00144

choices of $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$, the data contained roughly 42.2% and 40.3% missing values, respectively. The simulation results are discussed here under four scenarios as presented bellow.

Scenario 1: True model: MAR ($\tau_3 = 0$); fitted model: MAR ($\tau_3 = 0$).

In this scenario, the data were generated by using the MAR model where the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-2, 0.5, 0.5, 0)^t$ and the data were fitted by using the MAR model as well. In that sense, the model is correctly specified. The regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$. As the model is correctly specified, we expect to get negligible biases, small MSEs and good coverage probabilities for the estimators.

Table 5.1 presents the simulated biases and mean squared errors, and Table 5.2

Table 5.2: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Correctly specified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	0.5	0.93763	0.39242	0.94294	0.27354
β_1	0.5	0.93763	0.30333	0.94895	0.20904
β_2	1	0.95297	0.29974	0.95295	0.20760

Table 5.3: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Correctly specified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	0.5	0.00778	0.00901	0.00954	0.00453
β_1	1	0.00292	0.00578	0.00074	0.00272
β_2	0.5	0.00241	0.00593	0.00586	0.00267
μ_1	0.5	-0.00111	0.00557	-0.00191	0.00239
μ_2	0.5	-0.02698	0.00608	-0.03061	0.00367
σ_1^2	0.5	-0.00326	0.00473	-0.00149	0.00256
σ_2^2	0.5	-0.00950	0.00609	-0.00851	0.00299
σ_{12}	0.1	-0.00321	0.00274	-0.00206	0.00145

Table 5.4: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Correctly specified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	0.5	0.95282	0.38143	0.94394	0.26615
β_1	1	0.94974	0.29508	0.94995	0.20328
β_2	0.5	0.94667	0.29112	0.94995	0.20126

presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes, where the regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-2, 0.5, 0.5, 0)^t$.

On the other hand, to study the impact of changes to the coefficients of missing covariates, we consider the regression coefficients $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$ keeping the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism similar to those as used for correctly specified models under the MAR setting.

Table 5.3 presents the simulated biases and mean squared errors, and Table 5.4 presents the coverage probabilities and mean lengths of the confidence intervals for different sample sizes under the same setting as above but the regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$.

The maximum likelihood method appears to perform well for the correctly specified MAR mechanism under different sample sizes. The method provides unbiased estima-

tors and small mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution in both Tables 5.1 and 5.3. The confidence intervals have good coverages, which are close to the nominal 95% confidence level.

As expected, the biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ become smaller when we increase the sample size. The coverage probabilities for the parameter estimates become closer to the nominal levels and the mean lengths based on confidence interval become smaller when we increase the sample size. The confidence intervals for the regression coefficients $\boldsymbol{\beta}$ slightly improve coverages due to the choice of different values of $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$. As shown in Table 5.4, β_1 improves its coverage to 0.94974 from 0.93763 when the sample size is $n = 100$.

Scenario 2: True model: NMAR ($\tau_3 = 0.25$); fitted model: NMAR($\tau_3 \neq 0$).
In this scenario, the data were generated by using the NMAR model where the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-2, 0.5, 0.5, 0.25)^t$ and the data were fitted by using the NMAR model as well. In this case, the model is correctly specified. The regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$. Here, we expect to get unbiased and efficient estimators.

Table 5.5 presents the simulated biases and mean squared errors, and Table 5.6 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes, where the regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$

Table 5.5: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Correctly specified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	0.5	0.00267	0.01770	0.00341	0.00847
β_1	0.5	0.00568	0.01034	0.00605	0.00505
β_2	1	0.00033	0.01108	6.65815E-05	0.00503
τ_0	-2	-0.38703	0.66805	-0.16094	0.17269
τ_1	0.5	0.05145	0.39414	-0.01083	0.14341
τ_2	0.5	-0.04150	2.09604	-0.08365	0.86136
τ_3	0.25	0.05866	0.06127	0.03447	0.02397
μ_1	0.5	-0.00192	0.00476	0.00032	0.00252
μ_2	0.5	-0.01166	0.00887	-0.00899	0.00423
σ_1^2	0.5	-0.00419	0.00544	0.00165	0.00245
σ_2^2	0.5	0.00422	0.00800	0.00391	0.00356
σ_{12}	0.1	-0.00378	0.00328	-0.00040	0.00168

Table 5.6: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Correctly specified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	0.5	0.88648	0.43370	0.90010	0.30510
β_1	0.5	0.91847	0.37311	0.93170	0.26211
β_2	1	0.92260	0.36548	0.94190	0.25669
τ_0	-2	0.96594	2.31066	0.94801	1.40205
τ_1	0.5	0.93086	1.98681	0.92966	1.31357
τ_2	0.5	0.86068	3.93896	0.85831	2.61866
τ_3	0.25	0.89164	0.66541	0.87462	0.44538

and the parameters of the missing-data model were chosen as $\boldsymbol{\tau} = (-2, 0.5, 0.5, 0.25)^t$.

On the other hand, to study the impact of changes to the coefficients of missing covariates, we consider the regression coefficients $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$ keeping the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism similar to those as used for correctly specified models under the NMAR setting.

Table 5.7 presents the simulated biases and mean squared errors, and Table 5.8 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes under the same setting but the regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$.

The maximum likelihood method appears to perform well for the correctly specified

Table 5.7: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Correctly specified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	0.5	0.00500	0.01225	0.00439	0.00660
β_1	1	-0.000890	0.00870	0.00260	0.00428
β_2	0.5	-0.00533	0.00831	-0.00333	0.00379
τ_0	-2	-0.29387	0.48109	-0.13770	0.17931
τ_1	0.5	0.05645	0.58071	-0.00401	0.25321
τ_2	0.5	-0.00776	0.94030	-0.03022	0.45248
τ_3	0.25	0.04216	0.03283	0.02315	0.01399
μ_1	0.5	-0.00112	0.00512	0.00129	0.00245
μ_2	0.5	-0.01241	0.01162	-0.00565	0.00566
σ_1^2	0.5	-0.00071	0.00490	0.00083	0.00228
σ_2^2	0.5	0.00109	0.00856	0.00321	0.00409
σ_{12}	0.1	0.00039	0.00420	-0.00285	0.02531

Table 5.8: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Correctly specified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	0.5	0.91203	0.40589	0.92764	0.28741
β_1	1	0.92518	0.33497	0.93266	0.23707
β_2	0.5	0.92720	0.31939	0.92965	0.22441
τ_0	-2	0.96967	2.22625	0.95378	1.41763
τ_1	0.5	0.94034	2.59135	0.92663	1.75404
τ_2	0.5	0.90495	2.85656	0.86231	1.90343
τ_3	0.25	0.92417	0.56316	0.90854	0.37863

model under the NMAR mechanism in terms of bias and MSE for different sample sizes. The method provides unbiased estimators and small mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism. But τ_0 shows slightly large biases in both Tables 5.5 and 5.7. The confidence intervals have coverages which are not very close to the nominal level. The estimates of nuisance parameter $\boldsymbol{\tau}$ slightly improve coverages due to the choice of different values of $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$. As shown in Table 5.8, τ_2 improves its coverage to 0.90495 from 0.86068 when the sample size is $n = 100$.

As expected, the biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ become smaller when we increase the sample size. The coverage probabilities for the parameter estimates of $\boldsymbol{\beta}$ tend to be closer to the nominal level and the mean lengths based on confidence interval become smaller when we increase the sample size. Although we expected good coverages for the confidence intervals of nuisance parameters $\boldsymbol{\tau}$ but it loses slight coverage for both choices of $\boldsymbol{\beta} = (-2, 1, 0.5)^t$ and $\boldsymbol{\beta} = (-2, 0.5, 1)^t$ as shown in both Tables 5.6 and 5.8 when we increase the sample size.

Scenario 3: True model: MAR ($\tau_3 = 0$); fitted model: NMAR ($\tau_3 \neq 0$).
In this scenario, the data were generated by using the MAR model where the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-2, 0.5, 0.5, 0)^t$ and the data were fitted by using the NMAR model as well. In this case, the model is misspecified. The regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$. Here, we expect to get larger biases, MSEs and bad coverages for the estimators under the

Table 5.9: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Misspecified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	0.5	-0.01081	0.01258	-0.00138	0.00515
β_1	0.5	0.00613	0.00652	-0.000468	0.00288
β_2	1	-0.00303	0.00683	-0.00012	0.00266
τ_0	-2	-0.41179	1.11462	-0.16063	0.21130
τ_1	0.5	0.15571	0.45363	0.06695	0.13993
τ_2	0.5	0.29816	3.14971	0.12177	0.89027
τ_3	0	-0.02528	0.04094	-0.01057	0.011035
μ_1	0.5	-0.00063	0.00491	0.00306	0.00243
μ_2	0.5	0.00221	0.00706	0.00245	0.00343
σ_1^2	0.5	-0.00010	0.00502	-0.00110	0.00240
σ_2^2	0.5	0.00995	0.00589	0.00202	0.00285
σ_{12}	0.1	0.00280	0.00305	-0.00173	0.00134

misspecified model as compared to the correctly specified models.

Table 5.9 presents the simulated biases and mean squared errors, and Table 5.10 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes, where the regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-2, 0.5, 0.5, 0)^t$.

On the other hand, to study the impact of changes to the coefficients of missing covariates, we consider the regression coefficients $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$ keeping the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism similar

Table 5.10: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Misspecified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	0.5	0.91174	0.38585	0.93494	0.26996
β_1	0.5	0.91976	0.28920	0.93193	0.20296
β_2	1	0.91575	0.28957	0.94695	0.20043
τ_0	-2	0.97192	2.59766	0.96697	1.48941
τ_1	0.5	0.93882	2.12699	0.95596	1.36049
τ_2	0.5	0.87162	4.38257	0.89489	2.74803
τ_3	0	0.89569	0.49687	0.90891	0.29726

Table 5.11: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Misspecified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	0.5	-0.00595	0.01140	-0.00247	0.00542
β_1	1	0.00649	0.00625	0.00513	0.00278
β_2	0.5	-0.00879	0.00636	-0.00635	0.00279
τ_0	-2	-0.51473	1.45177	-0.21195	0.36304
τ_1	0.5	0.24751	1.09658	0.10283	0.37407
τ_2	0.5	0.28992	2.14289	0.08173	0.80420
τ_3	0	-0.02978	0.03239	-0.00926	0.01053
μ_1	0.5	0.00104	0.00528	0.00029	0.00229
μ_2	0.5	0.00617	0.00851	-0.00144	0.00468
σ_1^2	0.5	-0.00276	0.00486	-0.00096	0.00249
σ_2^2	0.5	0.02336	0.00707	0.01116	0.00336
σ_{12}	0.1	0.00066	0.00332	8.65232E-07	0.00148

Table 5.12: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Misspecified NMAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	0.5	0.91592	0.37541	0.925	0.26227
β_1	1	0.90591	0.27683	0.928	0.19482
β_2	0.5	0.90591	0.27235	0.924	0.19259
τ_0	-2	0.96797	2.88646	0.945	1.62977
τ_1	0.5	0.93594	2.94996	0.925	1.82153
τ_2	0.5	0.86386	3.56772	0.808	2.22634
τ_3	0	0.91291	0.44219	0.889	0.26171

to those as used for misspecified models under the same setting.

Table 5.11 presents the simulated biases and mean squared errors, and Table 5.12 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes under same setting, but the regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$.

The maximum likelihood method doesn't appear to perform well for this misspecified model in terms of coverage probability for different sample sizes. The method still provides small biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution in both Tables 5.9 and 5.11 . But the method produces larger biases and mean squared errors for $\boldsymbol{\tau}$ of the missing data mechanism compared to Scenario 2 for both choices of $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$

and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$. The parameter estimates have bad coverages that moved away from their nominal levels. The confidence intervals for the nuisance parameter $\boldsymbol{\tau}$ lose coverages due to the choice of different values of $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$. As shown in Table 5.12, τ_2 loses its coverage to 0.808 from 0.89489 when the sample size is $n = 200$.

Based on our simulation, the biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ become smaller when we increase sample size. Although biases are large for $\boldsymbol{\tau}$ with n=100, they become smaller when we increase the sample size. The coverage probabilities for the parameter estimates $\boldsymbol{\beta}$ become close to their nominal levels and the coverage probabilities for the parameter estimates $\boldsymbol{\tau}$ moved away from their nominal levels when we increase the sample size. Also, the mean lengths based on confidence interval become smaller with the large sample size.

Scenario 4: True model: NMAR ($\tau_3 = 0.25$); fitted model: MAR ($\tau_3 = 0$).
In this scenario, the data were generated by using the NMAR model where the parameters of the missing-data mechanism were chosen as $\boldsymbol{\tau} = (-2, 0.5, 0.5, 0.25)^t$ and the data were fitted by using the MAR model. In this case, the model is misspecified. The regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$.

Table 5.13 presents the simulated biases and mean squared errors, and Table 5.14 presents the coverage probabilities and the mean lengths of the confidence intervals for different sample sizes, where the regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and the parameters of the missing data mechanism were chosen as $\boldsymbol{\tau} = (-2, 0.5, 0.5, 0.25)^t$.

Table 5.13: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Misspecified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	0.5	0.01645	0.014816	0.01686	0.00698
β_1	0.5	0.00621	0.01259	0.014430	0.00575
β_2	1	0.01576	0.01373	0.01030	0.00581
μ_1	0.5	-0.00378	0.00527	-0.00035	0.00231
μ_2	0.5	-0.03423	0.00752	-0.02947	0.00384
σ_1^2	0.5	-0.00353	0.00487	-0.00316	0.00244
σ_2^2	0.5	-0.01272	0.00741	-0.01565	0.00408
σ_{12}	0.1	-0.00603	0.00315	-0.00556	0.00170

Table 5.14: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 0.5$. Misspecified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	0.5	0.92612	0.44654	0.94194	0.31297
β_1	0.5	0.93653	0.41384	0.93894	0.28798
β_2	1	0.92092	0.41008	0.93894	0.28403

Table 5.15: Empirical biases and mean squared errors (MSEs) of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Misspecified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	Bias	MSE	Bias	MSE
β_0	0.5	0.02366	0.01247	0.02035	0.00583
β_1	1	0.02202	0.010450	0.01687	0.00530
β_2	0.5	-0.00454	0.01114	0.00262	0.00362
μ_1	0.5	-0.00018	0.00524	0.00031	0.00238
μ_2	0.5	-0.06373	0.01248	-0.06281	0.00771
σ_1^2	0.5	0.00038	0.00516	-0.00310	0.00226
σ_2^2	0.5	-0.02717	0.00845	-0.02487	0.00427
σ_{12}	0.1	-0.02227	0.00452	-0.01601	0.00236

Table 5.16: Coverage probabilities (CPs), average lengths of maximum likelihood estimators in Poisson regression models with missing covariates for different sample sizes. Slope parameter $\beta_1 = 1$. Misspecified MAR model is assumed for missing data.

		n=100		n=200	
Parameter	True value	CP	Length	CP	Length
β_0	0.5	0.93569	0.41527	0.93594	0.29119
β_1	1	0.93891	0.38320	0.93594	0.26961
β_2	0.5	0.94427	0.34806	0.95996	0.24165

On the other hand, to study the impact of changes to the coefficients of missing covariates, we consider the regression coefficients $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$ keeping the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism similar to those as used for misspecified models under the same setting.

Table 5.15 presents the simulated biases and mean squared errors, and Table 5.16 presents the coverage probabilities and the mean lengths of the parameter estimates for different sample sizes under same setting, but the regression coefficients were fixed at $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$.

The maximum likelihood method still performs well under this misspecified model for different sample sizes. The method provides small biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution of the missing data mechanism in both Tables 5.13 and 5.15. The parameter estimates $\boldsymbol{\beta}$ lose slight coverages due to the choice of $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ in Table 5.14 but still they are close to their nominal levels 95%.

Based on our simulation, the biases and mean squared errors for all the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ stay very close when we increase the sample size. The coverage probabilities for the parameter estimates slightly moved away from the nominal level but still the parameter estimates have the good coverages. The mean lengths based on confidence intervals become smaller and coverage probabilities stay very close when we increase the sample size. As seen, the CPs of β_1 are 0.93653 and 0.93894 in Table 5.14 as well as 0.93891 and 0.93594 in Table 5.16 when the sample

sizes are $n = 100$ and $n = 200$ respectively.

It is clear from the above four scenarios that Scenario 1, 2 and 4 provide approximately unbiased estimates for all regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ of the covariate distribution and $\boldsymbol{\tau}$ of the missing data mechanism under all simulation configurations considered. On the other hand, Scenario 3 provides approximately unbiased estimates of the regression coefficients $\boldsymbol{\beta}$ as well as the nuisance parameters $\boldsymbol{\alpha}$ but slightly large biases for the estimates of nuisance parameters $\boldsymbol{\tau}$. As shown in Table 5.9, under misspecified method, the estimators of τ_1 and τ_2 give large biases as 0.15571 and 0.29816 respectively. Also as shown in Table 5.11, 0.24751 and 0.28992 respectively when the sample size is $n = 100$.

When comparing four scenarios, the maximum likelihood method performs better for Scenario 1, which is correctly specified model under the MAR mechanism and worst for Scenario 3, which is misspecified model under NMAR mechanism. As shown in Tables 5.1 and 5.3, Scenario 1 provides approximately unbiased estimates. Also, Scenario 2, 3 and 4 provide quite small biases and MSEs for all the parameter estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ under both correctly specified and misspecified models for both choices of $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$.

Scenario 1 also gives the better coverage probabilities and mean lengths for the regression coefficients $\boldsymbol{\beta}$ as shown in Table 5.2 and 5.4 under the correctly specified MAR mechanism. As we expected, the coverage probabilities are very close to the nominal level 95% which gradually improve and the mean lengths gradually decrease when we

increase the sample size for both choices of $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$ and $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$. Scenario 2 and 3 give bad coverages for all parameter estimates of $\boldsymbol{\tau}$ and lose slight coverages when we increase the sample size to $n = 200$ from $n = 100$ due to the choice of $\boldsymbol{\beta} = (0.5, 1, 0.5)^t$. The coverages of $\boldsymbol{\tau}$ decrease for both Scenarios 2 and 3 accordingly with increased sample sizes.

In Scenario 4, the parameter estimates of $\boldsymbol{\beta}$ have good coverages and they stay close due to an increased sample sizes in both Tables 5.14 and 5.16, but they lose slight coverages for the choice of $\boldsymbol{\beta} = (0.5, 0.5, 1)^t$. For example, as in Table 5.14, β_2 gives empirical coverage probabilities to 0.92092 from 0.94427 when the sample size is $n = 100$.

Here, we should note that there is no big bias for all estimates under the four scenarios. For the MSEs and mean lengths of the estimators of model parameters with different sample sizes, the empirical study shows that the MSEs and mean lengths decrease as the number of sample size n increases. In terms of CPs, Scenarios 1 and 4 give good coverages under both specified and misspecified models.

Chapter 6

Conclusion

In this thesis, we have studied the generalized linear model with missing covariates for analyzing binary and count data. We particularly considered joint estimation of the regression parameters and the association parameters by using the maximum likelihood method. We studied a set of maximum likelihood estimating equations for fitting regression models to binary data as well as Poisson data. We used the Newton-Raphson algorithm to estimate the parameters for both regressions. Our simulation study demonstrates that the missing data mechanism generally provides unbiased and efficient estimators when it follows correctly specified models under the NMAR method for binary data and MAR method for count data.

The purpose of the thesis was to study the empirical properties of the maximum likelihood estimator for assessing the significance of regression parameter $\boldsymbol{\beta}$ at a given level of significance. We also study the biases, MSEs, average lengths and empirical coverage probabilities of the estimators of model parameters under both correctly specified and misspecified models, when the true distribution is either the binomial or Poisson distribution.

It is apparent that when we use the NMAR model to fit the data, the models provide unbiased and efficient estimators for binary regression. On the other hand, when we use the MAR model to fit the data, the models provide unbiased and efficient estimators for the Poisson regression.

From the numerical study, we conclude that when analyzing incomplete data with missing covariates, it is necessary to incorporate a suitable missing data model into the observed data likelihood function in order to obtain unbiased and efficient estimators of the model parameters. We also note that a misspecified missing data model can provide systematic bias in the maximum likelihood estimation. So it is important to assess the validity of a missing data model when performing a likelihood inference based on the given observed data.

6.1 Future Research

There are many current and future research considerations for missing data problems when the mechanism is nonignorable. In this thesis, we used univariate missing covariate, where only one covariate was considered missing. Analysis with multivariate missing covariates is computationally extensive. To reduce the computational burden, as a future research we plan to use an approximate method based on Markov Chain Monte Carlo approximation for fitting generalized linear models under nonignorable multivariate missing covariates.

Bibliography

[1] Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonresponse. *Journal of the American Statistical Association*, 83, 62-49.

[2] Barndorff Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimate. *Biometrika*, 70, 343-365.

[3] Berk, R. H. (1972). Consistency and asymptotic normality of MLEs for exponential models. *The Annals of Mathematical Statistics*, 43, 193-204 .

[4] Doll, R. (1971). The age distribution of cancer: implications for models of carcinogenesis. *Journal of the Royal Statistical Society*, A, 134, 133-166.

[5] Draper, N. and Smith, H. (1981). Applied regression analysis (2nd ed.). New York: Wiley.

[6] Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimators in generalized linear models. *The Annals of Statistics*, 13, 342-368.

[7] Firth, D. (1991). Generalized linear models. In D. V. Hinkley, N. Reid and E. J. Snell (Eds.), *Statistical theory and modelling*, 55-86. Chapman and Hall, London.

[8] Freedman, D. A. (2009). *Statistical models: Theory and practice*. Cambridge University Press. p. 128.

[9] Frome, E. L. (1983). The analysis of rates using Poisson regression model. *Biometrics*, 39, 665-674.

[10] Greenlees, W. S., Reece, J. S. and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

[11] Huang, L., Chen, M. H. and Ibrahim, J. G. (2005). Bayesian analysis for generalized linear model with nonignorably missing covariates. *Biometrics*, 61, 767-780.

[12] Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics*, 5, 815-841 .

[13] Hilbe, J. M. (1994). Generalized linear models. *The American Statistician*, 48, 255-265.

[14] Hoffmann, J. P. (2004). *Generalized Linear Models; an applied approach*, Boston: Allyn and Bacon.

[15] Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85, 765-769.

[16] Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Hemng, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100, 332-346.

[17] Ibrahim, J. G. and Lipsitz, S. R. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society*, B, 61:173-190.

[18] Lane, P. W. (2002). Generalized linear models in soil science. *European Journal of Soil Science*, 53, 241-251.

[19] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*, Second Edition. Wiley, New York.

[20] Little, R. J. A. (1992) Regression with missing X's: a review. *Journal of the American Statistical Association*, 87:1227-1237.

[21] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition. Chapman and Hall, London.

[22] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, A, 135, 370-384.

[23] Janke, S. J. and Tinsley, F. C. (2005). Book Review: Introduction to Linear Models and Statistical Inference. NJ : Wiley.

[24] Sinha, S. K.(2008). Robust methods for generalized linear models with nonignorable missing covariates. *The Canadian Journal of Statistics*, 36, 277-299.

[25] Shi, X., Zhu, H. and Ibrahim, J. G. (2009). Local Influence for generalized linear models with Missing Covariates. *Biometrics*, 65, 1164-1174 .

[26] Tang, G., Little, R. J. A., and Raghunathan, T. E. (2003). Analysis of multivariate missing data With nonignorable nonresponse. *Biometrika*, 90, 747-764.

[27] Wu, H. and Wu, L. (2001). A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics. *Statistics in medicine*, 20: 1755-1769 .

[28] Wei, G. C. G. and Tanner, M. A. (1990): A Monte Carlo Implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the*

American Statistical Association, 85, 699-704.

[29] WU, Z. (2005). Generalized linear models in family studies. *Journal of Marriage and Family* 67, 1029-1047 .

[30] Yang, X., Belin, T. R. and Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, 61, 498-506.

[31] Zhao, J. and Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing Data. *Journal of the American Statistical Association*, 110, 1577-1590.

Appendix A

```
library(mvtnorm)

library(MASS)

binary.dat <- function(n=100, beta=c(-2,1,0.5), tau=c(-4,1,1,0), mx=c(2,1),

              Dx=matrix(c(1,.2,.2,1),2,2), eps=.05, out.x=0, out.y=0 ,iter=5)

  # generating missing data with outliers

{

  xx0 <- round(rmvnorm(n, mean=mx, sigma=Dx), digits=3)

  x1 <- xx0[ ,1]

  x2 <- xx0[ ,2]


  intercept <- rep(1, n)

  xx <- cbind(intercept, x1, x2)


  eta <- c(xx %*% beta)

  mu <- round(exp(eta)/(1+exp(eta)), digits=3)

  y <- rbinom(n, 1, mu)

  v1 <- rep(0, n)


  z <- cbind(intercept, x1, x2, y)

  zeta <- c(z %*% tau)

  pi <- round(exp(zeta)/(1+exp(zeta)), digits=3)
```

```
  v2 ← rbinom(n, 1, pi)

  n1 ← as.integer(eps*n)

  if(n1 > 0)

  {

    out.x ← rep(out.x, n1)

    out.y ← rep(out.y, n1)


    y[c(n-n1+1):n] ← ifelse(out.y==0, y[c(n-n1+1):n], rbinom(n1,1,.5))

    v2[c(n-n1+1):n] ← ifelse(out.y==0, v2[c(n-n1+1):n], rbinom(n1,1,.5))


    x1[c(n-n1+1):n] ← ifelse(out.x==0, x1[c(n-n1+1):n], x1[c(n-n1+1):n] + out.x)

    x2[c(n-n1+1):n] ← ifelse(out.x==0, x2[c(n-n1+1):n], x2[c(n-n1+1):n] + out.x)

  }
  data ← data.frame(y, x1, x2, mu, v1, v2, pi)

  data
}

data0←binary.dat(100)



psi.binom ← function(y=c(0,1), mu=c(2,1), c0=999)

  # This program computes the psi function, its derivatives

  # and the expected values for the binomial distribtuion

{

  n ← length(y)
```

```
r ← (y–mu)/sqrt(mu*(1–mu))

psi.r ← ifelse(abs(r) <= c0, r, sign(r) * c0)

deriv.psi.r ← ifelse(abs(r) <= c0, 1, 0)


y1 ← rep(0, n)

y2 ← rep(1, n)


r1 ← (y1–mu)/sqrt(mu*(1–mu))

psi.r1 ← ifelse(abs(r1) <= c0, r1, sign(r1) * c0)

deriv.psi.r1 ← ifelse(abs(r1) <= c0, 1, 0)


r2 ← (y2–mu)/sqrt(mu*(1–mu))

psi.r2 ← ifelse(abs(r2) <= c0, r2, sign(r2) * c0)

deriv.psi.r2 ← ifelse(abs(r2) <= c0, 1, 0)


f.y1 ← dbinom(y1, 1, mu)

f.y2 ← dbinom(y2, 1, mu)

Epsi.r ← psi.r1 * f.y1 + psi.r2 * f.y2

Epsi.rsq ← (psi.r1^2) * f.y1 + (psi.r2^2) * f.y2

Er.psi.r ← r1 * psi.r1 * f.y1 + r2 * psi.r2 * f.y2


Ederiv.psi.r ← deriv.psi.r1 * f.y1 + deriv.psi.r2 * f.y2

Er.deriv.psi.r ← r1 * deriv.psi.r1 * f.y1 + r2 * deriv.psi.r2 * f.y2
```

```
  list(r=r, psi.r=psi.r, deriv.psi.r=deriv.psi.r,
       Epsi.r=Epsi.r, Epsi.rsq=Epsi.rsq, Ederiv.psi.r=Ederiv.psi.r,
       Er.deriv.psi.r=Er.deriv.psi.r, Er.psi.r=Er.psi.r)
}


simpson <- function(vec, a, b)
{
  # Uses Simpson's rule; function values are elements of vec
  # length of vec should be odd
  n <- length(vec)
  m <- n - 1       # m is the number of intervals
  h <- (b - a)/m  # h is the length of each interval
  d <- c(1, rep(c(4, 2), m/2 - 1), 4, 1)
  integral <- (h/3) * sum(d * vec)
  list(integral = integral)
}


rglm.exact <- function(dat=data0, beta0=c(-2,1,0.5), tau0=c(-4,1,1,0),
 mx0=c(2,1), Dx0=matrix(c(1,.2,.2,1),2,2), c0=999, gam=0, iter=20)
  # fitting GLMs by rubust method
  # c0 is a constant for Huber's psi function
{
```

```
y ← dat$y

n ← length(y)


x1 ← dat$x1

x2 ← dat$x2

v1 ← dat$v1

v2 ← dat$v2

xx ← cbind(x1, x2)

a1 ← −7    # min(x)

a2 ← 7     # min(x)

x.mis ← seq(a1, a2, length=251)

len1 ← length(x.mis)

cutoff1 ← qchisq(.95, 1)

cutoff2 ← qchisq(.95, 2)

beta ← NULL

mx ← NULL

Dx ← NULL

tau ← NULL

for(it in 1:iter)

{

  cat(".")
```

```
M ← 0

q ← 0

Q ← 0

MM ← 0

wx ← NULL

sum.xi ← 0

sum.prod.xi ← 0

M.v ← 0

q.v ← 0

Q.v ← 0

MM.v ← 0

for (i in 1:n)

{

  x1i ← x1[i]

  v2i ← v2[i]

  if(v2i==0)

    # ---------------------------------
  {

    yi ← y[i]

    x2i ← x2[i]

    xxi ← c(x1i, x2i)

    xi ← c(1, xxi)
```

```
etai <- c(t(xi) %*% beta0)

mui <- exp(etai)/(1+exp(etai))


disi <- c(t(xxi - mx0) %*% solve(Dx0) %*% (xxi - mx0))

wxi <- ifelse(disi <= cutoff2, 1, (cutoff2/disi)^gam)


std.yi <- sqrt(mui * (1-mui))

wi <- std.yi * wxi

psii <- psi.binom(y=yi, mu=mui, c0=c0)

di <- psii$psi.r - psii$Epsi.r

d1i <- sqrt(mui * (1 - mui)) * (psii$deriv.psi.r - psii$Ederiv.psi.r) +

  (1/2 - mui) * (psii$r * psii$deriv.psi.r - psii$Er.deriv.psi.r) +

  sqrt(mui * (1 - mui)) * psii$Er.psi.r


Mi <- c(wi*d1i) * (xi %*% t(xi))

qi <- c(wi*di) * xi

Qi <- qi %*% t(qi)

MMi <- - Mi

# ------------------ for density of x --------------------------


sum.xi0 <- (wxi * xxi)

sum.prod.xi0 <- wxi^2 * (xxi %*% t(xxi))
```

```
  # ------------------- for tau -------------------------

  zi <- c(xi, yi)

  zetai <- c(t(zi) %*% tau0)

  pii <- exp(zetai)/(1+exp(zetai))

  std.vi <- sqrt(pii * (1-pii))

  # wyi <- psii$psi.r/psii$r

  # wyi <- (psii$psi.r+.000000001)/(psii$r+.000000001) # to avoid NA

  # w.vi <- std.vi * wxi * wyi

  w.vi <- std.vi * wxi

  psi.vi <- psi.binom(y=v2i, mu=pii, c0=c0)

  d.vi <- psi.vi$psi.r - psi.vi$Epsi.r

  d1.vi <- sqrt(pii * (1 - pii)) * (psi.vi$deriv.psi.r - psi.vi$Ederiv.psi.r) +

    (1/2 - pii) * (psi.vi$r * psi.vi$deriv.psi.r - psi.vi$Er.deriv.psi.r) +

    sqrt(pii * (1 - pii)) * psi.vi$Er.psi.r

  M.vi <- c(w.vi*d1.vi) * (zi %*% t(zi))

  q.vi <- c(w.vi*d.vi) * zi

  Q.vi <- q.vi %*% t(q.vi)

  MM.vi <- - M.vi
}

else

   # ------------------------------------
```

```
{
  x2i ← x.mis
  xxi ← cbind(rep(x1i,len1), x2i)

  xi ← cbind(rep(1,len1), xxi)
  yi ← rep(y[i],len1)
  etai ← c(xi %*% beta0)
  mui ← exp(etai)/(1+exp(etai))
  disi ← (x1i - mx0[1])^2/Dx0[1,1]
  wxi ← ifelse(disi <= cutoff1, 1, (cutoff1/disi)^gam)
  std.yi ← sqrt(mui * (1-mui))
  wi ← std.yi * wxi
  psii ← psi.binom(y=yi, mu=mui, c0=c0)
  di ← psii$psi.r - psii$Epsi.r
  d1i ← sqrt(mui * (1 - mui)) * (psii$deriv.psi.r - psii$Ederiv.psi.r) +
    (1/2 - mui) * (psii$r * psii$deriv.psi.r - psii$Er.deriv.psi.r) +
    sqrt(mui * (1 - mui)) * psii$Er.psi.r

  f.yi ← (mui^yi) * (1-mui)^(1-yi)
  f.xi ← dmvnorm(xxi, mean=mx0, sigma=Dx0)
  zi ← cbind(xi, yi)
  zetai ← c(zi %*% tau0)
```

```
pii ← exp(zetai)/(1+exp(zetai))

f.vi ← pii  # since v2i is always 1 here

fyi.xi ← f.yi * f.xi * f.vi

ave.fyi.xi ← simpson(fyi.xi, a1, a2)$integral

fyobsi ← ave.fyi.xi

fxmis.obsi ← fyi.xi/fyobsi

density.i ← fxmis.obsi

wwi ← c(wi*d1i)

Mi11 ← simpson(wwi * xi[,1] * xi[,1] * density.i, a1, a2)$integral

Mi12 ← simpson(wwi * xi[,1] * xi[,2] * density.i, a1, a2)$integral

Mi13 ← simpson(wwi * xi[,1] * xi[,3] * density.i, a1, a2)$integral

Mi22 ← simpson(wwi * xi[,2] * xi[,2] * density.i, a1, a2)$integral

Mi23 ← simpson(wwi * xi[,2] * xi[,3] * density.i, a1, a2)$integral

Mi33 ← simpson(wwi * xi[,3] * xi[,3] * density.i, a1, a2)$integral

Mi ← matrix(c(Mi11,Mi12,Mi13,Mi12,Mi22,Mi23,Mi13,Mi23,Mi33),3,3)

qqi ← c(wi*di) * xi

qi1 ← simpson(qqi[,1] * density.i, a1, a2)$integral

qi2 ← simpson(qqi[,2] * density.i, a1, a2)$integral

qi3 ← simpson(qqi[,3] * density.i, a1, a2)$integral

qi ← c(qi1, qi2, qi3)

Qi ← qi %*% t(qi)
```

```
    # ----------------

M1.i11 <- simpson(c(wi*di) * c(yi-mui) * xi[,1] * xi[,1] * density.i, a1, a2)$integral

M1.i12 <- simpson(c(wi*di) * c(yi-mui) * xi[,1] * xi[,2] * density.i, a1, a2)$integral

M1.i13 <- simpson(c(wi*di) * c(yi-mui) * xi[,1] * xi[,3] * density.i, a1, a2)$integral

M1.i22 <- simpson(c(wi*di) * c(yi-mui) * xi[,2] * xi[,2] * density.i, a1, a2)$integral

M1.i23 <- simpson(c(wi*di) * c(yi-mui) * xi[,2] * xi[,3] * density.i, a1, a2)$integral

M1.i33 <- simpson(c(wi*di) * c(yi-mui) * xi[,3] * xi[,3] * density.i, a1, a2)$integral

 M1i <- matrix(c(M1.i11,M1.i12,M1.i13,M1.i12,M1.i22,M1.i23,M1.i13,M1.i23,M1.i33),3,3)

    # ----------------

    M2.i1 <- simpson(c(yi-mui) * xi[,1] * density.i, a1, a2)$integral

    M2.i2 <- simpson(c(yi-mui) * xi[,2] * density.i, a1, a2)$integral

    M2.i3 <- simpson(c(yi-mui) * xi[,3] * density.i, a1, a2)$integral

    M2.i0 <- c(M2.i1, M2.i2, M2.i3)

    M2i <- qi %*% t(M2.i0)

    # ----------------

    MMi <- - Mi + M1i - M2i

    # ------------------ for density of x --------------------------

    ave.xmisi <- simpson(x.mis * density.i, a1, a2)$integral

    ave.xi <- c(x1i, ave.xmisi)

    sum.xi0 <- (wxi * ave.xi)

    prod.11 <- x1i^2
```

```
prod.12 ← x1i*ave.xmisi

prod.22 ← simpson(x.mis*x.mis * density.i, a1, a2)$integral

prod.mat ← matrix(c(prod.11,prod.12,prod.12,prod.22),2,2)

sum.prod.xi0 ← wxi^2 * prod.mat

# ------------------- for tau --------------------------

v2i ← rep(v2i, len1)

zi ← cbind(xi, yi)

zetai ← c(zi %*% tau0)

pii ← exp(zetai)/(1+exp(zetai))

std.vi ← sqrt(pii * (1-pii))

# wyi ← psii$psi.r/psii$r

# wyi ← (psii$psi.r+.000000001)/(psii$r+.000000001) # to avoid NA

# w.vi ← std.vi * wxi * wyi

w.vi ← std.vi * wxi

psi.vi ← psi.binom(y=v2i, mu=pii, c0=c0)

d.vi ← psi.vi$psi.r - psi.vi$Epsi.r

d1.vi ← sqrt(pii * (1 - pii)) * (psi.vi$deriv.psi.r - psi.vi$Ederiv.psi.r) +

  (1/2 - pii) * (psi.vi$r * psi.vi$deriv.psi.r - psi.vi$Er.deriv.psi.r) +

  sqrt(pii * (1 - pii)) * psi.vi$Er.psi.r

ww.vi ← c(w.vi*d1.vi)


M.vi11 ← simpson(ww.vi * zi[,1] * zi[,1] * density.i, a1, a2)$integral
```

```
M.vi12 <- simpson(ww.vi * zi[,1] * zi[,2] * density.i, a1, a2)$integral

M.vi13 <- simpson(ww.vi * zi[,1] * zi[,3] * density.i, a1, a2)$integral

M.vi14 <- simpson(ww.vi * zi[,1] * zi[,4] * density.i, a1, a2)$integral

M.vi22 <- simpson(ww.vi * zi[,2] * zi[,2] * density.i, a1, a2)$integral

M.vi23 <- simpson(ww.vi * zi[,2] * zi[,3] * density.i, a1, a2)$integral

M.vi24 <- simpson(ww.vi * zi[,2] * zi[,4] * density.i, a1, a2)$integral

M.vi33 <- simpson(ww.vi * zi[,3] * zi[,3] * density.i, a1, a2)$integral

M.vi34 <- simpson(ww.vi * zi[,3] * zi[,4] * density.i, a1, a2)$integral

M.vi44 <- simpson(ww.vi * zi[,4] * zi[,4] * density.i, a1, a2)$integral

M.vi <- matrix(c(M.vi11, M.vi12, M.vi13, M.vi14,
                 M.vi12, M.vi22, M.vi23, M.vi24,
                 M.vi13, M.vi23, M.vi33, M.vi34,
                 M.vi14, M.vi24, M.vi34, M.vi44),4,4)


qq.vi <- c(w.vi*d.vi) * zi

q.vi1 <- simpson(qq.vi[,1] * density.i, a1, a2)$integral

q.vi2 <- simpson(qq.vi[,2] * density.i, a1, a2)$integral

q.vi3 <- simpson(qq.vi[,3] * density.i, a1, a2)$integral

q.vi4 <- simpson(qq.vi[,4] * density.i, a1, a2)$integral

q.vi <- c(q.vi1, q.vi2, q.vi3, q.vi4)

Q.vi <- q.vi %*% t(q.vi)
```

```
        # ----------------

M1.vi11 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,1] * zi[,1] * density.i, a1, a2)$integral

M1.vi12 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,1] * zi[,2] * density.i, a1, a2)$integral

M1.vi13 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,1] * zi[,3] * density.i, a1, a2)$integral

M1.vi14 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,1] * zi[,4] * density.i, a1, a2)$integral


M1.vi22 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,2] * zi[,2] * density.i, a1, a2)$integral

M1.vi23 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,2] * zi[,3] * density.i, a1, a2)$integral

M1.vi24 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,2] * zi[,4] * density.i, a1, a2)$integral

M1.vi33 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,3] * zi[,3] * density.i, a1, a2)$integral

M1.vi34 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,3] * zi[,4] * density.i, a1, a2)$integral


M1.vi44 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,4] * zi[,4] * density.i, a1, a2 $integral

        M1.vi <- matrix(c(M1.vi11, M1.vi12, M1.vi13, M1.vi14,
                          M1.vi12, M1.vi22, M1.vi23, M1.vi24,
                          M1.vi13, M1.vi23, M1.vi33, M1.vi34,
                          M1.vi14, M1.vi24, M1.vi34, M1.vi44),4,4)

        # ----------------

        M2.vi1 <- simpson(c(v2i-pii) * zi[,1] * density.i, a1, a2)$integral

        M2.vi2 <- simpson(c(v2i-pii) * zi[,2] * density.i, a1, a2)$integral

        M2.vi3 <- simpson(c(v2i-pii) * zi[,3] * density.i, a1, a2)$integral

        M2.vi4 <- simpson(c(v2i-pii) * zi[,4] * density.i, a1, a2)$integral
```

```
    M2.vi0 ← c(M2.vi1 , M2.vi2 , M2.vi3 , M2.vi4 )
    M2.vi ← q.vi %*% t(M2.vi0 )
    # ----------------
    MM.vi ← - M.vi + M1.vi - M2.vi
  }
  M ← M + Mi
  q ← q + qi
  Q ← Q + Qi
  MM ← MM + MMi
  wx ← c(wx, wxi)
  sum.xi ← sum.xi + sum.xi0
  sum.prod.xi ← sum.prod.xi + sum.prod.xi0
  M.v ← M.v + M.vi
  q.v ← q.v + q.vi
  Q.v ← Q.v + Q.vi
  MM.v ← MM.v + MM.vi
}
beta0 ← beta0 + c(solve(M) %*% q)
beta ← cbind(beta, beta0)
var.beta ← solve(MM) %*% Q %*% solve(MM)
```

```
    std.beta ← sqrt(diag(var.beta))

    sum.wx ← sum(wx)

    sum.wx.sq ← sum(wx^2)

    mx0 ← sum.xi/sum.wx

    Dx0 ← (sum.prod.xi − sum.wx.sq * mx0 %*% t(mx0))/(sum.wx.sq−1)

    mx ← cbind(mx, mx0)

    Dx ← cbind(Dx, c(Dx0))

    tau0 ← tau0 + c(solve(M.v) %*% q.v)

    tau ← cbind(tau, tau0)

    var.tau ← solve(MM.v) %*% Q.v %*% solve(MM.v)

    std.tau ← sqrt(diag(var.tau))

  }

  list(beta=beta, beta0=beta0, std.beta=std.beta,

       tau=tau, tau0=tau0, std.tau=std.tau,

       mx=mx, mx0=mx0, Dx=Dx, Dx0=Dx0)
}


simul.fn ← function(simul=1000)
{
  beta ← NULL
  std.beta1 ←NULL
  mx ← NULL
  Dx ←NULL
  tau ← NULL
  std.tau1 ←NULL
  for(s in 1:simul)
  {
```

```
    data0 ← binary.dat(n=100, beta=c(-2,1,0.5), tau=c(-4,1,1,0),
                       mx=c(2,1), Dx=matrix(c(1,.2,.2,1),2,2),
                       eps=.05, out.x=0, out.y=0)
    fit ←rglm.exact(dat=data0,c0=999, gam=0)
    beta0 ←fit$beta0
    std.beta0 ←fit$std.beta
    beta ← rbind(beta, beta0)
    std.beta1 ← rbind(std.beta1, std.beta0)
    tau0 ←fit$tau0
    std.tau0 ←fit$std.tau
    tau ←rbind(tau,tau0)
    std.tau1 ←rbind(std.tau1, std.tau0)
    mx0 ←fit$mx0
    mx ←rbind(mx ,mx0)
    Dx0 ← fit$Dx0
    Dx ← rbind(Dx, c(Dx0))

    est1.table ← cbind(beta, std.beta1, tau, std.tau1, mx,Dx)
  }
  est1.table
}
s1 ← simul.fn(simul=1000)

write.csv(s1,``est1.table.csv")
```

Appendix B

```
library(mvtnorm)
library(MASS)

pois.dat <- function(n=200, beta=c(.5,1,.5), tau=c(-2,.5,.5,0.25),
 mx=c(.5,.5), Dx=matrix(c(.5,.1,.1,.5),2,2))
  # generating missing data with outliers
{

  xx0 <- round(rmvnorm(n, mean=mx, sigma=Dx), digits=3)

  x1 <- xx0[ ,1]
  x2 <- xx0[ ,2]

  intercept <- rep(1, n)
  xx <- cbind(intercept, x1, x2)

  eta <- c(xx %*% beta)
  mu <- round(exp(eta), digits=3)

  y <- rpois(n, mu)

  v1 <- rep(0, n)                  # x1 is considered always observed

  z <- cbind(intercept, x1, x2, y)
  zeta <- c(z %*% tau)
  pi <- round(exp(zeta)/(1+exp(zeta)), digits=3)
  v2 <- rbinom(n, 1, pi)

  data <- data.frame(y, x1, x2, mu, v1, v2, pi)
  data
}

data0<-pois.dat(200)


data0 <- pois.dat(n=200, beta=c(.5,1,.5), tau=c(-2,.5,.5,.25), mx=c(.5,.5),
 Dx=matrix(c(.5,.1,.1,.5),2,2))
```

```
psi.binom ← function(y=c(0,1), mu=c(.5,.5), c0=999)
  # This program computes the psi function, its derivatives
  # and the expected values for the binomial distribtuion
{
  n ← length(y)
  r ← (y-mu)/sqrt(mu*(1-mu))
  psi.r ← ifelse(abs(r) <= c0, r, sign(r) * c0)
  deriv.psi.r ← ifelse(abs(r) <= c0, 1, 0)

  y1 ← rep(0, n)
  y2 ← rep(1, n)

  r1 ← (y1-mu)/sqrt(mu*(1-mu))
  psi.r1 ← ifelse(abs(r1) <= c0, r1, sign(r1) * c0)
  deriv.psi.r1 ← ifelse(abs(r1) <= c0, 1, 0)

  r2 ← (y2-mu)/sqrt(mu*(1-mu))
  psi.r2 ← ifelse(abs(r2) <= c0, r2, sign(r2) * c0)
  deriv.psi.r2 ← ifelse(abs(r2) <= c0, 1, 0)

  f.y1 ← dbinom(y1, 1, mu)
  f.y2 ← dbinom(y2, 1, mu)

  Epsi.r ← psi.r1 * f.y1 + psi.r2 * f.y2
  Epsi.rsq ← (psi.r1^2) * f.y1 + (psi.r2^2) * f.y2
  Er.psi.r ← r1 * psi.r1 * f.y1 + r2 * psi.r2 * f.y2

  Ederiv.psi.r ← deriv.psi.r1 * f.y1 + deriv.psi.r2 * f.y2
  Er.deriv.psi.r ← r1 * deriv.psi.r1 * f.y1 + r2 * deriv.psi.r2 * f.y2

  list(r=r, psi.r=psi.r, deriv.psi.r=deriv.psi.r,
       Epsi.r=Epsi.r, Epsi.rsq=Epsi.rsq, Ederiv.psi.r=Ederiv.psi.r,
       Er.deriv.psi.r=Er.deriv.psi.r, Er.psi.r=Er.psi.r)
}


# ----------------------------------------

psi.pois ← function(y=c(0,1), mu=c(.5,.5), c0=999)
  # This program computes the psi function, its derivatives
  # and the expected values for the Poisson distribtuion
```

```
{
  n ← length(y)
  r ← (y–mu)/sqrt(mu)
  psi.r ← ifelse(abs(r) <= c0, r, sign(r) * c0)
  deriv.psi.r ← ifelse(abs(r) <= c0, 1, 0)

  yy ← t(matrix(rep(c(0:199), n), 200, n))

  rr ← (yy–mu)/sqrt(mu)
  psi.rr ← ifelse(abs(rr) <= c0, rr, sign(rr) * c0)
  deriv.psi.rr ← ifelse(abs(rr) <= c0, 1, 0)

  f.yy ← dpois(yy, mu)

  Epsi.r ← apply(psi.rr * f.yy, 1, sum)
  Epsi.rsq ← apply((psi.rr^2) * f.yy, 1, sum)
  Er.psi.r ← apply(rr * psi.rr * f.yy, 1, sum)
  Ederiv.psi.r ← apply(deriv.psi.rr * f.yy, 1, sum)
  Er.deriv.psi.r ← apply(rr * deriv.psi.rr * f.yy, 1, sum)

  list(r=r, psi.r=psi.r, deriv.psi.r=deriv.psi.r, Epsi.r=Epsi.r,
       Epsi.rsq=Epsi.rsq, Ederiv.psi.r=Ederiv.psi.r,
       Er.deriv.psi.r=Er.deriv.psi.r, Er.psi.r=Er.psi.r)
}

# ------------------------------------------




simpson ← function(vec, a, b)
{
  # Uses Simpson's rule; function values are elements of vec
  # length of vec should be odd

  n ← length(vec)
  m ← n - 1       # m is the number of intervals
  h ← (b - a)/m  # h is the length of each interval
  d ← c(1, rep(c(4, 2), m/2 - 1), 4, 1)
  integral ← (h/3) * sum(d * vec)
  list(integral = integral)
}
```

```
rglmpois.exact ← function(dat=data0, beta0=c(.5,1,.5), tau0=c(-2,.5,.5,0.25),
 mx0=c(.5,.5), Dx0=matrix(c(.5,.1,.1,.5),2,2), c0=999, gam=0, iter=20)

  # fitting GLMs by rubust method
  # c0 is a constant for Huber's psi function
{
  y ← dat$y
  n ← length(y)

  x1 ← dat$x1
  x2 ← dat$x2

  v1 ← dat$v1
  v2 ← dat$v2

  xx ← cbind(x1, x2)

  a1 ← -3    # min(x)
  a2 ← 3     # min(x)
  x.mis ← seq(a1, a2, length=51)
  len1 ← length(x.mis)

  cutoff1 ← qchisq(.95, 1)
  cutoff2 ← qchisq(.95, 2)

  beta ← NULL

  mx ← NULL
  Dx ← NULL

  tau ← NULL

  for(it in 1:iter)
  {
    cat(".")

    M ← 0
    q ← 0
    Q ← 0
```

```
MM ← 0

wx ← NULL

sum.xi ← 0
sum.prod.xi ← 0

M.v ← 0
q.v ← 0
Q.v ← 0

MM.v ← 0

for (i in 1:n)
{

  x1i ← x1[i]
  v2i ← v2[i]

  if(v2i==0)

    # ------------------------------------

  {
    yi ← y[i]

    x2i ← x2[i]
    xxi ← c(x1i, x2i)
    xi ← c(1, xxi)

    etai ← c(t(xi) %*% beta0)
    mui ← exp(etai)


    disi ← c(t(xxi - mx0) %*% solve(Dx0) %*% (xxi - mx0))
    wxi ← ifelse(disi <= cutoff2, 1, (cutoff2/disi)^gam)

    std.yi ← sqrt(mui)
    wi ← std.yi * wxi

    psii ← psi.pois(y=yi, mu=mui, c0=c0)
```

```
di ← psii$psi.r - psii$Epsi.r

d1i ← sqrt(mui) * (psii$deriv.psi.r - psii$Ederiv.psi.r) +
  (1/2) * (psii$r * psii$deriv.psi.r - psii$Er.deriv.psi.r) +
  sqrt(mui) * psii$Er.psi.r

Mi ← c(wi*d1i) * (xi %*% t(xi))
qi ← c(wi*di) * xi

Qi ← qi %*% t(qi)

MMi ← - Mi

# ------------------ for density of x --------------------------

sum.xi0 ← (wxi * xxi)
sum.prod.xi0 ← wxi^2 * (xxi %*% t(xxi))

# ------------------- for tau -------------------------

zi ← c(xi, yi)
zetai ← c(t(zi) %*% tau0)
pii ← exp(zetai)/(1+exp(zetai))

std.vi ← sqrt(pii * (1-pii))

# wyi ← (psii$psi.r+.0000000000001)/(psii$r+.0000000000001) # to avoid NA
# w.vi ← std.vi * wxi * wyi

w.vi ← std.vi * wxi

psi.vi ← psi.binom(y=v2i, mu=pii, c0=c0)

d.vi ← psi.vi$psi.r - psi.vi$Epsi.r

d1.vi ← sqrt(pii * (1 - pii)) * (psi.vi$deriv.psi.r - psi.vi$Ederiv.psi.r) +
  (1/2 - pii) * (psi.vi$r * psi.vi$deriv.psi.r - psi.vi$Er.deriv.psi.r) +
  sqrt(pii * (1 - pii)) * psi.vi$Er.psi.r

M.vi ← c(w.vi*d1.vi) * (zi %*% t(zi))
q.vi ← c(w.vi*d.vi) * zi
```

```
  Q.vi <- q.vi %*% t(q.vi)

  MM.vi <- - M.vi

}


else

  # ------------------------------------

{
  x2i <- x.mis
  xxi <- cbind(rep(x1i,len1), x2i)

  xi <- cbind(rep(1,len1), xxi)
  yi <- rep(y[i],len1)

  etai <- c(xi %*% beta0)
  mui <- exp(etai)

  disi <- (x1i - mx0[1])^2/Dx0[1,1]
  wxi <- ifelse(disi <= cutoff1, 1, (cutoff1/disi)^gam)

  std.yi <- sqrt(mui)
  wi <- std.yi * wxi

  psii <- psi.pois(y=yi, mu=mui, c0=c0)

  di <- psii$psi.r - psii$Epsi.r

  d1i <- sqrt(mui) * (psii$deriv.psi.r - psii$Ederiv.psi.r) +
    (1/2) * (psii$r * psii$deriv.psi.r - psii$Er.deriv.psi.r) +
    sqrt(mui) * psii$Er.psi.r

  f.yi <- exp(-mui) * (mui^yi)/gamma(yi+1)

  f.xi <- dmvnorm(xxi, mean=mx0, sigma=Dx0)

  zi <- cbind(xi, yi)
  zetai <- c(zi %*% tau0)
  pii <- exp(zetai)/(1+exp(zetai))
```

```
f.vi ← pii  # since v2i is always 1 here

fyi.xi ← f.yi * f.xi * f.vi

ave.fyi.xi ← simpson(fyi.xi, a1, a2)$integral
fyobsi ← ave.fyi.xi

fxmis.obsi ← fyi.xi/fyobsi
density.i ← fxmis.obsi

wwi ← c(wi*d1i)

Mi11 ← simpson(wwi * xi[,1] * xi[,1] * density.i, a1, a2)$integral
Mi12 ← simpson(wwi * xi[,1] * xi[,2] * density.i, a1, a2)$integral
Mi13 ← simpson(wwi * xi[,1] * xi[,3] * density.i, a1, a2)$integral
Mi22 ← simpson(wwi * xi[,2] * xi[,2] * density.i, a1, a2)$integral
Mi23 ← simpson(wwi * xi[,2] * xi[,3] * density.i, a1, a2)$integral
Mi33 ← simpson(wwi * xi[,3] * xi[,3] * density.i, a1, a2)$integral

Mi ← matrix(c(Mi11,Mi12,Mi13,Mi12,Mi22,Mi23,Mi13,Mi23,Mi33),3,3)

qqi ← c(wi*di) * xi

qi1 ← simpson(qqi[,1] * density.i, a1, a2)$integral
qi2 ← simpson(qqi[,2] * density.i, a1, a2)$integral
qi3 ← simpson(qqi[,3] * density.i, a1, a2)$integral

qi ← c(qi1, qi2, qi3)

Qi ← qi %*% t(qi)

# ----------------

M1.i11 ← simpson(c(wi*di) * c(yi-mui) * xi[,1] * xi[,1] * density.i, a1, a2)$integral
M1.i12 ← simpson(c(wi*di) * c(yi-mui) * xi[,1] * xi[,2] * density.i, a1, a2)$integral
M1.i13 ← simpson(c(wi*di) * c(yi-mui) * xi[,1] * xi[,3] * density.i, a1, a2)$integral
M1.i22 ← simpson(c(wi*di) * c(yi-mui) * xi[,2] * xi[,2] * density.i, a1, a2)$integral
M1.i23 ← simpson(c(wi*di) * c(yi-mui) * xi[,2] * xi[,3] * density.i, a1, a2)$integral
M1.i33 ← simpson(c(wi*di) * c(yi-mui) * xi[,3] * xi[,3] * density.i, a1, a2)$integral

M1i ← matrix(c(M1.i11,M1.i12,M1.i13,M1.i12,M1.i22,M1.i23,M1.i13,M1.i23,M1.i33),3,3)
```

```
# ----------------

M2.i1 <- simpson(c(yi-mui) * xi[,1] * density.i, a1, a2)$integral
M2.i2 <- simpson(c(yi-mui) * xi[,2] * density.i, a1, a2)$integral
M2.i3 <- simpson(c(yi-mui) * xi[,3] * density.i, a1, a2)$integral

M2.i0 <- c(M2.i1, M2.i2, M2.i3)

M2i <- qi %*% t(M2.i0)

# ----------------

MMi <- - Mi + M1i - M2i


# ------------------ for density of x ---------------------------

ave.xmisi <- simpson(x.mis * density.i, a1, a2)$integral
ave.xi <- c(x1i, ave.xmisi)

sum.xi0 <- (wxi * ave.xi)

prod.11 <- x1i^2
prod.12 <- x1i*ave.xmisi
prod.22 <- simpson(x.mis*x.mis * density.i, a1, a2)$integral

prod.mat <- matrix(c(prod.11,prod.12,prod.12,prod.22),2,2)
sum.prod.xi0 <- wxi^2 * prod.mat


# ------------------- for tau --------------------------

v2i <- rep(v2i, len1)
zi <- cbind(xi, yi)
zetai <- c(zi %*% tau0)
pii <- exp(zetai)/(1+exp(zetai))

std.vi <- sqrt(pii * (1-pii))

# wyi <- (psii$psi.r+.0000000000001)/(psii$r+.0000000000001) # to avoid NA
# w.vi <- std.vi * wxi * wyi
```

```
w.vi ← std.vi * wxi

psi.vi ← psi.binom(y=v2i, mu=pii, c0=c0)

d.vi ← psi.vi$psi.r - psi.vi$Epsi.r

d1.vi ← sqrt(pii * (1 - pii)) * (psi.vi$deriv.psi.r - psi.vi$Ederiv.psi.r) +
  (1/2 - pii) * (psi.vi$r * psi.vi$deriv.psi.r - psi.vi$Er.deriv.psi.r) +
  sqrt(pii * (1 - pii)) * psi.vi$Er.psi.r

ww.vi ← c(w.vi*d1.vi)

M.vi11 ← simpson(ww.vi * zi[,1] * zi[,1] * density.i, a1, a2)$integral
M.vi12 ← simpson(ww.vi * zi[,1] * zi[,2] * density.i, a1, a2)$integral
M.vi13 ← simpson(ww.vi * zi[,1] * zi[,3] * density.i, a1, a2)$integral
M.vi14 ← simpson(ww.vi * zi[,1] * zi[,4] * density.i, a1, a2)$integral
M.vi22 ← simpson(ww.vi * zi[,2] * zi[,2] * density.i, a1, a2)$integral
M.vi23 ← simpson(ww.vi * zi[,2] * zi[,3] * density.i, a1, a2)$integral
M.vi24 ← simpson(ww.vi * zi[,2] * zi[,4] * density.i, a1, a2)$integral
M.vi33 ← simpson(ww.vi * zi[,3] * zi[,3] * density.i, a1, a2)$integral
M.vi34 ← simpson(ww.vi * zi[,3] * zi[,4] * density.i, a1, a2)$integral
M.vi44 ← simpson(ww.vi * zi[,4] * zi[,4] * density.i, a1, a2)$integral

M.vi ← matrix(c(M.vi11, M.vi12, M.vi13, M.vi14,
                M.vi12, M.vi22, M.vi23, M.vi24,
                M.vi13, M.vi23, M.vi33, M.vi34,
                M.vi14, M.vi24, M.vi34, M.vi44),4,4)

qq.vi ← c(w.vi*d.vi) * zi

q.vi1 ← simpson(qq.vi[,1] * density.i, a1, a2)$integral
q.vi2 ← simpson(qq.vi[,2] * density.i, a1, a2)$integral
q.vi3 ← simpson(qq.vi[,3] * density.i, a1, a2)$integral
q.vi4 ← simpson(qq.vi[,4] * density.i, a1, a2)$integral

q.vi ← c(q.vi1, q.vi2, q.vi3, q.vi4)

Q.vi ← q.vi %*% t(q.vi)

# ---------------
```

```
M1.vi11 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,1] * zi[,1] * density.i, a1, a2)$integral
M1.vi12 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,1] * zi[,2] * density.i, a1, a2)$integral
M1.vi13 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,1] * zi[,3] * density.i, a1, a2)$integral
M1.vi14 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,1] * zi[,4] * density.i, a1, a2)$integral

M1.vi22 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,2] * zi[,2] * density.i, a1, a2)$integral
M1.vi23 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,2] * zi[,3] * density.i, a1, a2)$integral
M1.vi24 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,2] * zi[,4] * density.i, a1, a2)$integral

M1.vi33 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,3] * zi[,3] * density.i, a1, a2)$integral
M1.vi34 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,3] * zi[,4] * density.i, a1, a2)$integral

M1.vi44 <-simpson(c(w.vi*d.vi) * c(v2i-pii) * zi[,4] * zi[,4] * density.i, a1, a2)$integral

    M1.vi <- matrix(c(M1.vi11, M1.vi12, M1.vi13, M1.vi14,
                      M1.vi12, M1.vi22, M1.vi23, M1.vi24,
                      M1.vi13, M1.vi23, M1.vi33, M1.vi34,
                      M1.vi14, M1.vi24, M1.vi34, M1.vi44),4,4)


    # ----------------

    M2.vi1 <- simpson(c(v2i-pii) * zi[,1] * density.i, a1, a2)$integral
    M2.vi2 <- simpson(c(v2i-pii) * zi[,2] * density.i, a1, a2)$integral
    M2.vi3 <- simpson(c(v2i-pii) * zi[,3] * density.i, a1, a2)$integral
    M2.vi4 <- simpson(c(v2i-pii) * zi[,4] * density.i, a1, a2)$integral

    M2.vi0 <- c(M2.vi1, M2.vi2, M2.vi3, M2.vi4)

    M2.vi <- q.vi %*% t(M2.vi0)

    # ----------------

    MM.vi <- - M.vi + M1.vi - M2.vi

  }


  M <- M + Mi
  q <- q + qi

  Q <- Q + Qi
```

```
    MM <- MM + MMi

    wx <- c(wx, wxi)

    sum.xi <- sum.xi + sum.xi0
    sum.prod.xi <- sum.prod.xi + sum.prod.xi0

    M.v <- M.v + M.vi
    q.v <- q.v + q.vi

    Q.v <- Q.v + Q.vi

    MM.v <- MM.v + MM.vi
  }

  beta0 <- beta0 + c(solve(M) %*% q)
  beta <- cbind(beta, beta0)

  var.beta <- solve(MM) %*% Q %*% solve(MM)
  std.beta <- sqrt(diag(var.beta))

  sum.wx <- sum(wx)
  sum.wx.sq <- sum(wx^2)

  mx0 <- sum.xi/sum.wx
  Dx0 <- (sum.prod.xi - sum.wx.sq * mx0 %*% t(mx0))/(sum.wx.sq-1)

  mx <- cbind(mx, mx0)
  Dx <- cbind(Dx, c(Dx0))

  tau0 <- tau0 + c(solve(M.v) %*% q.v)
  tau <- cbind(tau, tau0)

  var.tau <- solve(MM.v) %*% Q.v %*% solve(MM.v)
  std.tau <- sqrt(diag(var.tau))

}

list(beta=beta, beta0=beta0, std.beta=std.beta,
     tau=tau, tau0=tau0, std.tau=std.tau,
     mx=mx, mx0=mx0, Dx=Dx, Dx0=Dx0)
```

```
}


fit1 <- rglmpois.exact(dat=data0, beta0=c(.5,1,.5), tau0=c(-2,.5,.5,.25),
 mx0=c(.5,.5), Dx0=matrix(c(.5,.1,.1,.5),2,2), c0=999, gam=0, iter=20)


simul.fn <- function(simul=1000)
{
  beta <- NULL
  std.beta1<-NULL
  mx <- NULL
  Dx <-NULL
  tau <- NULL
  std.tau1<-NULL

  for(s in 1:simul)
  {

    data0 <- pois.dat(n=200, beta=c(0.5,1,0.5), tau=c(-2,0.5,0.5,0.25),
                      mx=c(0.5,0.5), Dx=matrix(c(0.5,.1,.1,0.5),2,2))
    fit<-rglmpois.exact(dat=data0,c0=999, gam=0 )
    beta0 <-fit$beta0
    std.beta0<-fit$std.beta
    beta <- rbind(beta, beta0)
    std.beta1 <- rbind(std.beta1, std.beta0)
    tau0<-fit$tau0
    std.tau0<-fit$std.tau
    tau<-rbind(tau,tau0)
    std.tau1<-rbind(std.tau1, std.tau0)
    mx0<-fit$mx0
    mx<-rbind(mx ,mx0)
    Dx0 <- fit$Dx0
    Dx <- rbind(Dx, c(Dx0))

    est2.table <- cbind(beta, std.beta1, tau, std.tau1, mx, Dx)
  }
  est2.table
}
s1 <- simul.fn(simul=1000)

write.csv(s1,``est2.table.csv")
```