BAYESIAN SEMIPARAMETRIC MODELS FOR NONIGNORABLE MISSING DATA MECHANISMS IN LOGISTIC REGRESSION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

OLCAY ÖZTÜRK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

MAY 2011

Approval of the thesis:

BAYESIAN SEMIPARAMETRIC MODELS FOR NONIGNORABLE MISSING DATA MECHANISMS IN LOGISTIC REGRESSION

submitted by **OLCAY ÖZTÜRK** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. H. Öztaş Ayhan
Head of Department, **Statistics** ________________

Assist. Prof. Dr. Zeynep Işıl Kalaylıoğlu
Supervisor, **Statistics Department, METU** ________________

Examining Committee Members:

Assoc. Prof. Dr. Meral Çetin
Statistics Department, Hacettepe University ________________

Assist. Prof. Dr. Zeynep Işıl Kalaylıoğlu
Statistics Department, METU ________________

Assoc. Prof. Dr. İnci Batmaz
Statistics Department, METU ________________

Assist. Prof. Dr. Özlem İlk
Statistics Department, METU ________________

Assist. Prof. Dr. Vilda Purutçuoğlu
Statistics Department, METU ________________

Date: ________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name:	OLCAY ÖZTÜRK
Signature	:

ABSTRACT

BAYESIAN SEMIPARAMETRIC MODELS FOR NONIGNORABLE MISSING DATA MECHANISMS IN LOGISTIC REGRESSION

Öztürk, Olcay
M.S., Department of Statistics
Supervisor : Assist. Prof. Dr. Zeynep Işıl Kalaylıoğlu

May 2011, 69 pages

In this thesis, Bayesian semiparametric models for the missing data mechanisms of nonignorably missing covariates in logistic regression are developed. In the missing data literature, fully parametric approach is used to model the nonignorable missing data mechanisms. In that approach, a probit or a logit link of the conditional probability of the covariate being missing is modeled as a linear combination of all variables including the missing covariate itself. However, nonignorably missing covariates may not be linearly related with the probit (or logit) of this conditional probability. In our study, the relationship between the probit of the probability of the covariate being missing and the missing covariate itself is modeled by using a penalized spline regression based semiparametric approach. An efficient Markov chain Monte Carlo (MCMC) sampling algorithm to estimate the parameters is established. A WinBUGS code is constructed to sample from the full conditional posterior distributions of the parameters by using Gibbs sampling. Monte Carlo simulation experiments under different true missing data mechanisms are applied to compare the bias and efficiency properties of the resulting estimators with the ones from the fully parametric approach. These simulations show that estimators for logistic regression using semiparametric missing data models maintain better bias and efficiency properties than the ones using fully parametric missing data models when the true relationship between the missingness and the missing covariate has a nonlinear form. They are comparable when this relationship has a linear form.

Keywords: Computational Bayesian statistics, emprical Bayes, Gibbs sampler, nonignorably missing covariate, penalized spline regression

ÖZ

LOJİSTİK REGRESYONDA İHMAL EDİLEMEYEN KAYIP VERİ MEKANİZMALARI
İÇİN BAYESCİ YARI-PARAMETRİK MODELLER

Öztürk, Olcay
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi : Yrd. Doç. Dr. Zeynep Işıl Kalaylıoğlu

Mayıs 2011, 69 sayfa

Bu tez çalışmasında, lojistik regresyonda ihmal edilemeyen kayıp veriye sahip ortak değişkenlerin kayıp veri mekanizmaları için Bayesci yarı-parametrik modeller geliştirilmiştir. Kayıp veri literatüründe, ihmal edilemeyen kayıp veri mekanizması tam parametrik yaklaşım ile modellenmiştir. Bu yaklaşımda, ortak değişkendeki verinin kayıp olma koşullu olasılığının probit veya logit bağlantısı, kayıp veri olan ortak değişken dâhil tüm değişkenlerin doğrusal birleşimi ile modellenir. Ancak, bu koşullu olasılığın probit (veya logit) bağlantısı ile ihmal edilemeyen kayıp veriye sahip ortak değişkenler arasındaki ilişki doğrusal olmayabilir. Bizim çalışmamızda, kayıp verili ortak değişkenin kendisi ile bu değişkende kayıp veri olma olasılığının probit bağlantısı arasındaki ilişki, yarı-parametrik bir yaklaşım kullanılarak cezalı yiv regresyonu ile modellenmiştir. Parametreleri tahmin etmek için etkili Markov zinciri Monte Carlo (MZMC) örnekleme algoritması kurulmuştur. Gibbs örnekleyicisi kullanılarak parametrelerin tam koşullu sonsal dağılımlarından örneklem çekilebilmesi için WinBUGS kodu oluşturulmuştur. Farklı gerçek kayıp veri mekanizmaları altında, önerilen tahmin edicileri yanlılık ve etkinlik özellikleri açısından tam-parametrik yaklaşımla elde edilen tahmin edicilerle karşılaştırabilmek için Monte Carlo benzetim denemeleri yapılmıştır. Bu benzetim denemeleri şu sonuçları vermektedir. Kayıp veriye sahip ortak değişken ile bu değişkendeki verinin kayıp olması arasındaki gerçek ilişkinin doğrusal olmayan formda olduğu durumlarda, yarı-parametrik kayıp veri modelleri kullanılarak elde edilen lojistik regresyon tahmin edicileri yanlılık ve etkinlik özellikleri açısından tam parametrik kayıp veri modelleri kullanılarak elde edilen tahmin edicilere göre daha iyidir. Bu ilişkinin doğrusal formda olduğu durumlarda ise tahmin edicilerin yanlılık ve etkinlik özellikleri benzerdir.

Anahtar Kelimeler: Hesaplamaya dayalı Bayesci istatistik, deneysel Bayes, Gibbs örnekleyicisi, ihmal edilemeyen kayıp verili ortak değişken, cezalı yiv regresyonu

To my beloved family and friends for their everlasting support...

ACKNOWLEDGMENTS

First of all, I would like to express my deepest appreciation to my supervisor Assist. Prof. Dr. Zeynep Işıl Kalaylıoğlu for her invaluable guidance, enthusiasm and encouragements throughout the research. She not only encouraged me to go ahead with my thesis but also answered all my questions patiently when ever I need help with something. Without her encouragement and understanding, it would have been impossible for me to finish this thesis. It has been a great honor and pleasure for me to work with her.

I gratefully acknowledge my examining committee members, Assist. Prof. Dr. Özlem İlk, Assoc. Prof. Dr. Meral Çetin, Assoc. Prof. Dr. İnci Batmaz and Assist. Prof. Dr. Vilda Purutçuoğlu for their detailed review, constructive criticism and excellent advice on my thesis.

I owe my special thanks to all members of the Department of Statistics for providing me a comfortable working environment and necessary technology to complete my thesis.

I also would like to express my warm and sincere thanks to my beloved family and friends for their everlasting support, kindly understanding and encouragement to pursue my academic goals. They always put a smile on my face and make me feel better in every circumstance. I owe my success and achievement to them.

Lastly, I wish to express my sincere thanks to all of those who supported me in any respect throughout the research.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

For any statistical analysis, the researchers generally need to use a data matrix to make an inference about the situation. In many applications such as biological, social and economical studies, some of the entries of the data matrix may not be observed because of nonresponse or technical reasons, which constitutes the missing data problem. In general, missing data can disrupt the representativeness of the sample and severely affect the conclusions drawn from the data depending on the missing percentage. The reason why the data is missing is an important question to deal with since the methods used in the analysis vary considerably depending on the type of missingness. For example, if the values are missing at random, missing observations can be simply excluded from the analysis or replaced by using the simple imputation methods in most of the cases; however if the values are missing systematically, those observations can not be simply ignored and analyst should need to make more complex analyses which are more robust to missingness. In the missing data literature, the statistical methodologists try to develop data analysis methods that produce as little bias as possible on statistical inference.

In order to handle the missing data problem more systematically, Little and Rubin [1] define mainly three different types of missingness; *missing completely at random* (MCAR), *missing at random* (MAR) and *not missing at random* (NMAR). If the reason for missingness does not depend on the observed or missing components of the data, the missing variable (response or covariate) is said to be MCAR. If it depends only on the observed components of the data, the missingness is classified as MAR. The missing variable is called NMAR if the reason for missingness depends on the missing components of the data. To account for the underlying nature of the missingness in the analysis, the mechanism leading to missingness is included in the model. The missing data mechanism is **ignorable** for the classes of MCAR and MAR but **nonignorable** for the NMAR cases. In the real life situations, the probability of the variable being missing can be highly associated with its actual value that is missing. For example, respondents in a study may be unwilling to report the annual income due to the fact that it is too high or too low. In this case the information about the pattern of missingness should be accounted in the analysis by using the missing data mechanism. On the other hand, in a biological experiment, some results can be missing because of technical reasons which are unrelated to the experimental process. Thus, analyst can directly exclude those missing observations from the analysis without losing information or introducing bias.

In this thesis, noningorable missing data mechanisms for NMAR covariates in logistic regression analysis are considered. In general, parametric approaches are used to model the missing data mechanisms in the analysis of the generalized linear models (GLMs). In such parametric approaches, the relationship between the missingness and NMAR covariates is assumed to have a linear form. However, the actual relationship might have a nonlinear form. For instance, consider a study on dietary habits evaluating the probability of person being vegetarian depending on the person's sex, age and weight. Anorectic or obese subjects can be much more likely to refuse to report any weight related information compared to the subjects with rather normal weight. That is, the probability of nonresponse is high for low or high values of weight. Since the probability of the weight information being missing is not directly proportional with the person's weight, there can exists serious biases in the estimates of the parameters if the missing data mechanism is constructed by using the linear parametric model. The semiparametric models are more flexible to capture any possible nonlinear functional relationship between the response and covariates (see Ruppert et al. [2]). Due to this fact, penalized spline regression based semiparametric approach is considered to model the missing data mechanisms. In this context, a fully Bayesian procedure is developed to efficiently estimate the parameters of interest. To compare the bias and efficiency properties of the resulting estimators with the ones from the fully parametric approach, Monte Carlo simulation experiments under different true missing data mechanisms are performed.

1.2 Literature Review

For the analysis of the GLMs with nonignorable missing data models, first maximum likelihood (ML) methods are established in the missing data literature, and then fully Bayesian procedures are developed. Ibrahim and Lipsitz [4] developed an Expectation-Maximization (EM) algorithm based ML procedure with weights, which is originated from the work of Ibrahim [3] for ignorable missing data models, for binomial regression models with nonignorably missing response and fully observed covariates. They modeled the missingness probability of the nonignorably missing response as a logistic regression. In Ibrahim et al. [5], multinomial model is considered for the missing data mechanism by writing the joint distribution of missing data indicators as a sequence of one-dimensional conditional distributions. Each one-dimensional conditional distribution for missing data indicators is modeled via logistic regression again. In their study, the GLMs with fully observed response and NMAR continuous covariates are considered, and the Monte Carlo based EM algorithm is used to estimate the parameters via the Gibbs sampler.

Huang et al. [6] developed a fully Bayesian method for estimation in GLMs with NMAR covariates, and examined the properties of the priors used for the coefficients of the multinomial missing data models under various conditions. It is shown in that paper that if the improper uniform prior is used for the parameters of the missingness mechanism, the joint posterior distribution of the parameters would be also improper. On the other hand, they also showed that using noninformative proper priors for these parameters would result in slow convergence in the posterior calculations because of poor mixing. In order to provide proper posterior inference that is insensitive to the choice of hyperparameters, empirical Bayes based priors are proposed for the parameters of missing data mechanism. This is accomplished by specifying the hyperparameters of the proper prior according to the information that is conveyed about these parameters in the observed dataset as well as the other sets of data that could have been observed from the considered model. The deviance information criterion (DIC) of Spiegelhalter et al. [7] is also extended in that paper for determining whether the missing data mechanism is ignorable or nonignorable.

The four common ways of dealing with the missing data problem in GLMs are discussed in a comparative review by Ibrahim et al. [8]. The comparative simulation study for the maximum likelihood (ML), multiple imputation (MI), fully Bayesian (FB) and weighted estimating equations (WEEs) methods is carried out. The procedures using the EM algorithm in Ibrahim [3], and Ibrahim and Lipsitz [4] are considered for ML method. The techniques for creating complete datasets by filling the missing values in Little and Rubin [1] are considered for MI method. Approaches proposed by Ibrahim et al. [5] and Huang et al. [6] are considered for FB method. The techniques developed by Robins et al. [9] are considered for WEEs method. The simulations show that the four methods give practically same results when the covariate models are specified correctly.

To avoid the effect of outliers in the GLMs with missing covariates, a robust method that downweights the influential observations is proposed by Sinha [10]. The robust method based ML estimation for fitting the GLMs when the missing data mechanism is nonignorable is demonstrated. The Monte Carlo simulation under different patterns having various percentage of outliers is also carried out to compare the proposed robust ML method with the ML method proposed by Ibrahim [3]. It is shown that robust ML method for estimating the parameters of the GLMs with noningorably missing covariates has good robustness properties under the data with outliers, and is comparable to the ML method under the data without outliers.

Recently, missing data mechanism for NMAR covariates in GLMs is modeled by using the semiparametric model in Chen and Ibrahim [11]. To determine the shape of possibly nonlinear functional relationship between the missing data indicator and the covariates, penalized spline regression based semiparametric approach is used in the missing data mechanism. In a penalized spline regression, possibly nonlinear functional relationship between the response and the covariates is modeled through a smooth function of each covariate. In their article, they used smooth function for the covariates that are suspected to be nonlinearly related with the missingness and improved the ML based parameter estimation via the EM algorithm for the proposed semiparametric missing data model. To investigate the performance of proposed semiparametric model for the missing data mechanism, Monte Carlo simulations are carried out under the various true missing data mechanism. In their simulations, they focused on modeling the fully observed covariates nonparametrically (as an unspecified smooth function which is assumption free functional formwise) while they modeled the covariates that are subject to missingness parametrically (as a strictly linear function which is a strong assumption) in the missingness model. The downside of this simulation experiment is that it may not well accommodate the situations that are encountered in real life studies. In reality it is quite difficult to determine empirically the nature of the relationship between the missingness status and the values of the covariate that is missing. Thus, unless there is a strong ground or reliable literature to safely assume linearity, one should avoid making such a strong assumption about the functional form of the relationship between them.

1.3 Objectives and Scope of the Study

As mentioned before, the missing data mechanisms for NMAR covariates in the GLMs are modeled by using parametric approaches in general. That is, the conditional probability of the covariate being missing depend on all covariates and response is modeled by using logistic or probit regression. In this approach, the relationship between the logit (or probit) of that probability and the missing covariate is assumed to be linear. However, the actual relationship might be a nonlinear type as in the study on dietary habits. If the actual relationship does not have linear form, choosing the true parametric model is not so easy in many applications. At this point, semiparametric models are more attractive than the parametric ones since they can detect the form of functional relationship between the missingness and the missing covariates. Consequently, we propose a semiparametric model to scrutinize the effect of NMAR covariates, that are nonlinearly associated with the missingness, on estimating the parameters of the main interest.

Chen and Ibrahim [11] also considered a semiparametric model for the missing data mechanism and derived ML estimation via the EM algorithm. Our work differs from their work mainly in the following aspects: i) they propose a semiparametric model in which the smooth function is used for the covariates having possibly nonlinear effect on the missingness status and the linear form is used for all the other covariates; we, on the other hand, consider smooth function for all the NMAR covariates (not only the ones that might have a nonlinear effect on the missingness status), ii) they investigated the performance of the model in which the fully observed covariates have nonlinear and the NMAR covariates have strictly linear true effect on the missingness status while we focus on the performance of the model with NMAR covariates having nonlinear true effect on the missingness status, iii) they used the EM algorithm based ML estimation and we considered the fully Bayesian approach to estimate the parameters of interest, iv) they did not consider the usage of different number of knots in the semiparametric missing data model; however, we showed the effect of various number of knots selection on the inference of main model.

Through this thesis, we made the following contributions in logistic regression analysis with nonignorably missing covariates:

i) Nonparametric modeling for all the nonignorably missing covariates (see Section 2.1): we proposed the use of spline regression in the missingness model.

ii) Knot determination in the presence of nonignorably missing values (see Section 2.1): we developed an iterative method to determine the knots in the presence of nonignorably missing covariates.

iii) Empirical Bayes based prior for σ_b^2 (see Section 2.2.1): we proposed robust prior for σ_b^2 based on empirical Bayesian where σ_b^2 is the variance of the latent random effects in the resulting mixed model.

iv) Knot analysis (see Section 3.1): we assessed sensitivity of the estimation procedure to the number of knots in the spline regression.

v) WinBUGS code (see Section 2.2.3): we provided a simple yet effective way of implementing the Bayesian estimation by Gibbs sampling in WinBUGS (Speigelhalter, Thomas, and Best [12]).

This thesis is organized as follows. Chapter 2 explains the methodology consists of model construction and Bayesian estimation. The models for each component of the missing data problem and our proposed semiparametric model for the missing data mechanism are described in Section 2.1. The Bayesian procedure to estimate the parameters including prior construction, posterior construction and Bayesian inference using Gibbs sampling via WinBUGS is described in Section 2.2. Chapter 3 gives our simulation study conducting extensive investigation on the performance of the proposed model and the accompanying estimation procedure. Chapter 4 concludes the study with some discussion.

CHAPTER 2

METHODOLOGY

In the analysis of generalized linear models (GLMs) with nonignorably missing covariates, the Expectation-Maximization (EM) algorithm based maximum likelihood (ML) procedure or fully Bayesian approach can be used to make an inference on the parameters of the main interest. Bayesian hierarchical modeling provides a flexible approach in modeling as it can easily accommodate in its theory any distributional assumption which is not necessarily Normal. It has been an attractive alternative to frequentist approaches in that sense and yet the use of it had been hindered by the difficulties in deriving the posterior distributions. Today, modern sampling methods provide an efficient computational way of obtaining the posterior densities. Models with nonignorably missing covariates are hierarchical in nature and thus can be easily handled with Bayesian approaches. Additionally, more information can be extracted from the data using the empirical Bayes based priors in addition to the model based likelihood and this is especially important in missing data problems in which the likelihood of missingness alone usually contains insufficient information about the parameters of missingness mechanism. To obtain the Bayesian estimates of the parameters, the joint posterior distribution of the parameters based on the model based likelihood and the prior distributions is required. First, each component of the hierarchical model including the proposed semiparametric missing data model is described in Section 2.1, and then prior and accompanying posterior construction for the Bayesian inference using Gibbs sampling are explained in Section 2.2.

2.1 Model

In the missing data problems, if the variable is missing completely at random (MCAR) or missing at random (MAR), missing data mechanism is not required in the model; however, it is necessary to construct a model for the missingness as well as for the data themselves if the variable is *not missing at random* (NMAR). In the missing data context, the model framework is hierarchical in nature with components *response model*, *covariate distribution* and *missing data model*. The main focus in this hierarchy is the response model as its parameters are the main parameters of interest.

Let $\{(y_i,x_i), i = 1,..,n\}$ be the set of observations obtained from the subjects selected independently for the study. Here y_i denotes the response variable and x_i denotes the vector of p correlated continuous or categorical covariates measured on the ith subject. That is $x_i = (x_{i1},...,x_{ip})$. For instance, in a study on dietary habits, y_i can be an indicator for a person being vegetarian and x_i can be the factors associated with the person's dietary habits. In such datasets, some of the covariates might be missing for some i's. We can rearrange the order of subjects so that $i = 1,...,m$ correspond to the subjects with completely observed y_i and x_i, and $i = m+1,...,n$ correspond to the subjects with at least one missing covariate value. Let s be the number of covariates that are subject to missingness and x^{miss} denote the $n \times s$ matrix of such covariates where $x_i^{miss} = (x_{i1},...,x_{is})$ is the ith row vector of this matrix. In this thesis, it is assumed that all possibly missing covariates are continuous and NMAR. Also, $p-s$ is the number of covariates that are observable for every subject in the sample and let x^{obs} denote the $n \times (p-s)$ matrix of the covariates that are observed for every subject where $x_i^{obs} = (x_{i,s+1},...,x_{ip})$ is the ith row vector of this matrix. Thus, the covariates of the ith subject can be defined as $x_i = (x_i^{miss}, x_i^{obs})$. When the data contain missing observations, missing data indicator for each missing covariate is needed to construct the missing data model that represents the underlying mechanism of missingness. We take $r_{ik} = 1$ when x_{ik} is missing and $r_{ik} = 0$ when x_{ik} is observed for $i = 1,...,n$ and $k = 1,...,s$. Consequently, the complete data can be denoted by $D_c = (r,y,x^{miss},x^{obs})$ and the complete data likelihood can be written by using the selection model approach as follows.

$$\begin{aligned} L(\phi,\beta,\alpha|D_c) &= f(r,y,x|\phi,\beta,\alpha) \\ &= f(r|y,x,\phi)f(y|x,\beta)f(x|\alpha) \qquad (2.1) \end{aligned}$$

where β, α and ϕ are the parameters of the response, covariate and missingness models respectively. By using the selection model approach, we can specify the joint distribution of each component separately. Note that, if the missingness status of each possibly missing covariate is related to only always observed covariates, not to covariates that are subject to missingness, missing data mechanism $f(r|y,x,\phi)$ is not needed to be specified at all and in this case we say that the missingness mechanism is ignorable. Each component of these probabilistic models are described below, and the last part combines these components into a hierarchical format.

Response Model

When the response variable takes only two possible outcomes $\{0,1\}$ (e.g.,an indicator for a person being vegetarian or not), logistic regression (sometimes called the logit model) can be used to predict the probability of the response variable being 1 depending on the covariates of interest. Logistic regression is a special case of the GLMs in which the logit link function is used. To define the relationship between y_i and x_i, the probability distribution function of y_i conditional on a given value x_i can be defined by a Bernoulli distribution in the following form.

$$f(y_i|x_i^{miss},x_i^{obs}\beta) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}\ ; \quad \mu_i = E(y_i|x_i^{miss},x_i^{obs},\beta) = P(y_i = 1|x_i^{miss},x_i^{obs},\beta) \tag{2.2}$$

in which

$$\text{logit}(\mu_i) = \log\left[\frac{\mu_i}{1-\mu_i}\right] = [1\ x_i]\ \beta = \beta_0 + \beta_1 x_{i1}^{miss} + \ldots + \beta_s x_{is}^{miss} + \beta_{s+1} x_{i,s+1}^{obs} + \ldots + \beta_p x_{ip}^{obs} \tag{2.3}$$

where logit is a link function that explains how expected response is related with a linear predictor $[1\ x_i]\ \beta$. Also, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients including the intercept, which are the parameters of the main interest. For the simulation study in Chapter 3, we consider a logistic regression model with two continuous covariates one of which is possibly missing and the other one is fully observed. Then the response model used in the simulation study is defined by

$$\text{logit}\,[E(y_i|x_i,\beta)] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \tag{2.4}$$

where we assume that response variable y_i and covariate x_{i2} are fully observed, and covariate x_{i1} is nonignorably missing.

Covariate Distribution

In a model including nonignorably missing covariates, the covariate distribution is also needed to be specified due to the random feature of the missing covariates. Additionally, if the covariates are correlated with each other, this correlation can be accounted in the covariate distribution. Let $f(x_i|\alpha)$ denote the joint distribution of $(x_{i1},...,x_{ip})$ for $i = 1,...,n$ where α denoting the set of parameters specifying this joint distribution. Then

$$\begin{aligned} f(x_i|\alpha) &= f(x_i^{miss}, x_i^{obs}|\alpha) \\ &= f(x_i^{miss}|x_i^{obs}, \alpha_{(1)}) f(x_i^{obs}|\alpha_{(2)}) \end{aligned} \tag{2.5}$$

where α_1 is the set of parameters specifying the distribution of x_i^{miss} conditional on x_i^{obs} and α_2 is the set of parameters specifying the distribution of x_i^{obs}. In this thesis, it is assumed that all fully observed covariates x_i^{obs} are fixed. Thus, the joint distribution $f(x_i^{obs}|\alpha_{(2)})$ in (2.5) can be ignored. To reduce the number of nuisance parameters in the covariate distribution, Lipsitz and Ibrahim [13] wrote the joint distribution function of x_i^{miss} conditional on x_i^{obs} as a product of piecewise conditional distributions. Let x_{ik}^{miss} denote the kth missing covariate for $k = 1,...,s$. Then the joint distribution $f(x_i^{miss}|x_i^{obs}, \alpha_{(1)})$ can be written as a product of s piecewise conditional distributions as follows:

$$\begin{aligned} f(x_i^{miss}|x_i^{obs}, \alpha_{(1)}) =& f(x_{i1}^{miss},...,x_{is}^{miss}|x_i^{obs}, \alpha_{(1)}) \\ =& f(x_{is}^{miss}|x_{i1}^{miss},...,x_{i,s-1}^{miss}, x_i^{obs}, \alpha_{(1)s}) f(x_{i,s-1}^{miss}|x_{i1}^{miss},...,x_{i,s-1}^{miss}, x_i^{obs}, \alpha_{(1)s-1}) \\ & ...f(x_{i2}^{miss}|x_{i1}^{miss}, x_i^{obs}, \alpha_{(1)2}) f(x_{i1}^{miss}|x_i^{obs}, \alpha_{(1)1}) \end{aligned} \tag{2.6}$$

where $\alpha_{(1)k}$ denotes the set of parameters specifying the conditional distribution of x_{ik}^{miss} for $k = 1,...,s$. Each conditional distribution can be written in the form of a GLM density which defines the relationship between x_{ik}^{miss} and $(x_{i1}^{miss},...,x_{i,k-1}^{miss}, x_i^{obs})$. For instance for continuous missing covariates a Normal distribution can be considered. Let the set of parameters $\alpha_{(1)k}$ consists of the coefficients $\alpha_{(1)k}^{*}$ and the variance component $\sigma_{X(k)}^2$ for $k = 1,...,s$. Then kth conditional distribution of x_{ik}^{miss} can be defined as follows:

$$x_{ik}^{miss}|\left(x_{i1}^{miss},...,x_{i,k-1}^{miss}, x_i^{obs}\right) \sim N\left[\alpha_{(1)k}^{(*,0)} + \sum_{j=1}^{k-1} \alpha_{(1)k}^{(*,j)} x_{ij}^{miss} + \sum_{j=s+1}^{p} \alpha_{(1)k}^{(*,j)} x_{ij}^{obs},\ \sigma_{X(k)}^2\right] \tag{2.7}$$

where $\alpha_{(1)k}^{(*,j)}$ is the jth component of the coefficients $\alpha_{(1)k}^{*}$ in the kth conditional distribution of x_{ik}^{miss} for $k = 1,...,s$ and $j = 0,...,k-1, s+1,...,p$. In the simulation study, since there are two continuous covariates one of which is possibly missing and the other one is fully observed in the logistic regression, covariate distribution used in the simulation study is specified as follows:

$$x_{i1} \mid x_{i2} \sim N(\alpha_0 + \alpha_1 x_{i2}, \sigma_x^2) \tag{2.8}$$

where the parameters are redefined as $(\alpha_{(1)1}^{(*,0)}, \alpha_{(1)1}^{(*,1)}, \sigma_{X(1)}^2) = (\alpha_0, \alpha_1, \sigma_x^2)$ for simplicity.

Proposed Semiparametric Missing Data Model

For each missing covariate, if the probability that x_{ik}^{miss} being missing depends on the values of observed variables (e.g. y_i or x_i^{obs}) but not on the values of possibly missing covariates, the covariate x_{ik}^{miss} is said to be missing at random (MAR) and missing data model can be ignored in the analysis. On the other hand, if the missingness on x_{ik}^{miss} is related with the values of missing covariates including itself and not necessarily with the values of the observed variables, the covariate x_{ik}^{miss} is said to be *not missing at random* (NMAR) and missing data model is nonignorable. Thus, it is necessary to model the missing data mechanism for all the missing covariates since each one is NMAR in this thesis. Let $r_i = (r_{i1}, r_{i2}, ..., r_{is})$ be the vector of missing value indicator that corresponds to x_i^{miss}. Each r_{ik} $(k = 1, ..., s)$ is a binary variable whose distribution depends on the values of x_i^{miss} that would have been observed if the missingness occurred not at random. Also the distribution of r_{ik} may possibly depend on the values of the fully observed covariates as well as the response variable. One possible way of modeling the missingness mechanism is to use a multinomial model (e.g. a joint log-linear model) for the joint distribution of r_i. Alternatively, Ibrahim et al. [5] represent the joint distribution of r_i for subject i as a product of s piecewise conditional distributions as

$$\begin{aligned} f(r_i|y_i, x_i^{miss}, x_i^{obs}, \phi) =& f(r_{i1}, ..., r_{is}|y_i, x_i^{miss}, x_i^{obs}, \phi) \\ =& f(r_{is}|(r_{i1}, ..., r_{i,s-1}), y_i, x_i^{miss}, x_i^{obs}, \phi_{(s)}) \\ & ...f(r_{i,s-1}|(r_{i1}, ..., r_{i,s-2}), y_i, x_i^{miss}, x_i^{obs}, \phi_{(s-1)}) \\ & ...f(r_{i2}|r_{i1}, y_i, x_i^{miss}, x_i^{obs}, \phi_{(2)}) f(r_{i1}|y_i, x_i^{miss}, x_i^{obs}, \phi_{(1)}) \end{aligned} \tag{2.9}$$

where $\phi = (\phi_{(1)}, ..., \phi_{(s)})$ and each $\phi_{(k)}$ is the set of parameters associated with the conditional distribution of r_{ik} for $k = 1, ..., s$. Each one-dimensional conditional distribution of r_{ik} can be modeled by logistic or probit regression. Note that each $r_{ik}$is a Bernoulli random variable with success probability $p_{ik} = P(r_{ik} = 1|r_{ik}^*, y_i, x_i^{miss}, x_i^{obs})$ where $r_{ik}^* = (r_{i1}, ..., r_{i,k-1})$ is the set of missing data indicators on the condition part. Then the probability distribution function of

r_{ik} depending on $(r^*_{ik}, y_i, x^{miss}_i, x^{obs}_i)$ can be defined by a Bernoulli distribution in the following form.

$$f(r_{ik}|r^*_{ik}, y_i, x^{miss}_i, x^{obs}_i, \phi_{(k)}) = p^{r_{ik}}_{ik}(1-p_{ik})^{1-r_{ik}} \;\; ; \quad p_{ik} = P(r_{ik} = 1|r^*_{ik}, y_i, x^{miss}_i, x^{obs}_i, \phi_{(k)}) \tag{2.10}$$

in which

$$h(p_{ik}) = \phi_{(k,0)} + r^*_{ik}\phi_{(k,1)} + y_i\phi_{(k,2)} + x^{miss}_i\phi_{(k,3)} + x^{obs}_i\phi_{(k,4)} \tag{2.11}$$

where h is a chosen link function as logit or probit. Also, $\phi_{(k,.)}$ are partitions of $\phi_{(k)}$; $\phi_{(k,0)}$ and $\phi_{(k,2)}$ are scalars, $\phi_{(k,1)}$, $\phi_{(k,3)}$ and $\phi_{(k,4)}$ are $(k-1) \times 1$, $s \times 1$ and $(p-s) \times 1$ parameter vectors associated with r^*_{ik}, x^{miss}_i and x^{obs}_i, respectively.

In the missing data literature, each missing data indicator r_{ik} is generally modeled by a logistic or probit regression where it is assumed that all covariates $(r^*_{ik}, y_i, x^{miss}_i, x^{obs}_i)$ have a linear effect on the missingness status as in (2.11). The linearity assumption in this model can be dangerous in most cases due to the fact that the data may not contain sufficient information about the true relationship between the missingness and missing covariates. For instance, let the weight of a person be the covariate in the logistic regression analysis and it is observed that some of the covariates are missing for some subjects. Also, let the true probability of this covariate being missing is high for its extreme values compared to its midrange values. Suppose that this true relationship is not known by the analyst, the linearity assumption can be constructed incorrectly by considering only the observed subjects of this missing covariate. In that case, parametric missing data model ignores the nonlinear true relationship that exists between the actual value of the covariate and its missingness status. To detect the true functional relationship (e.g. nonlinear or linear) between the missingness and the missing covariates, we modeled the covariates that are subject to missingness nonparametrically in the missing data mechanism by using an unspecified smooth function which is assumption free functional formwise. We considered separate smooth functions for each missing covariate by using the generalized additive model (GAM) approach. The possible relationship between r_{ik} and x^{miss}_{ik} that might have a nonlinear form can be modeled by using this approach. The missing data model we consider is represented as follows:

$$h(p_{ik}) = \phi_{(k,0)} + r^*_{ik}\phi_{(k,1)} + y_i\phi_{(k,2)} + \sum_{k=1}^{s} m(x^{miss}_{ik}) + x^{obs}_i\phi_{(k,4)} \tag{2.12}$$

where $m(.)$ is an unspecified smooth function. The model is *semiparametric* in the sense that $h(p_{ik})$ is made to be linearly dependent on the covariates $(r^*_{ik}, y_i, x^{obs}_i)$ and the relationship between them is expressed in fully parametric terms (i.e. in terms of the parameters $(\phi_{(k,1)}, \phi_{(k,2)}, \phi_{(k,4)}))$ whereas the interpretation of the relationship between $h(p_{ik})$ and x_i^{miss} does not depend on parameters. We can model the smooth function by using natural cubic splines, B-splines, truncated polynomials etc. In this thesis, the smooth functions $m(.)$ are approximated by low-rank thin-plate splines which is a special case of thin-plate splines. The reason for this choice is that they have good computational properties. For example, low-rank thin-plate splines have better mixing properties for the Markov chain Monte Carlo (MCMC) chains than other basis (e.g. truncated polynomials) since there exists small posterior correlation for the parameters of the thin-plate splines (e.g. see Crainiceanu et al. [14]).

Before constructing the semiparametric missing data model, let's first consider the properties of smooth function and corresponding low-rank thin-plate splines approximation. To detect the functional relationship between y_i and x_i in a flexible way ($i = 1, ..., n$), an unspecified smooth function $y_i = m(x_i)$ can be found by minimizing the following function.

$$\sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int m''(x)dx \tag{2.13}$$

where λ is the smoothing parameter and $\int m''(x)dx$ measures the "wiggliness" of the function f. The wiggliness gives the total roughness of function in the domain of x. For example, the wiggliness is zero when the function is linear; however nonlinear functions produce wiggliness value bigger than zero. To avoid the overfitting problem, the wiggliness of function which is the penalty term in (2.13) is minimized as well as the sum of residual squares. The smoothing parameter λ controls the trade-off between the goodness of fit to the data and roughness of the function estimate. Larger values of λ force $m(x)$ to be smoother whereas $m(x)$ is detecting too much detail for smaller values of λ which causes the overfitting problem. For example, the number of fixed knots in the low-rank thin-plate splines is selected according to the smoothing parameter λ. For any value of λ, the minimizer of (2.13) is a natural cubic spline which is a piecewise third-order polynomial. On the other hand, the low-rank thin-plate splines can also be used to approximate the smooth function in the following form.

$$m(x_i, \psi, u) = \psi_0 + \psi_1 x_i + \sum_{d=1}^{D} u_d |x_i - \kappa_d|^3 \tag{2.14}$$

where $\psi = (\psi_0, \psi_1)^T$ and $u = (u_1, ..., u_D)^T$ are unknown coefficients, and $\kappa_1 < \kappa_2 < ... < \kappa_D$ are fixed (known) knots suitably determined for x. Here D is the number of knots which should be chosen large enough to provide the desired flexibility, and fixed knot κ_d can be the sample quantile of x's corresponding to probability $d/(D+1)$. To estimate the unknown coefficients in the smooth function by avoiding overfitting, the following quantity is minimized as in (2.13).

$$\sum_{i=1}^{n}(y_i - m(x_i, \psi, u))^2 + \lambda\ u^T \Omega\ u \tag{2.15}$$

where Ω is a known $D \times D$ penalty matrix whose (l,k)th entry is $|\kappa_l - \kappa_k|^3$. An appropriate number of knots can be chosen so that it minimizes the penalty introduced by the coefficients of $|x_i - \kappa_d|^3$ as well as the total residual squares.

Let $Y = (y_1, y_2, ..., y_n)^T$ be a $n \times 1$ vector for the response values and define X and Z^* matrices as

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \text{and} \qquad Z^* = \begin{bmatrix} |x_1-\kappa_1|^3 & |x_1-\kappa_2|^3 & \cdots & |x_1-\kappa_D|^3 \\ |x_2-\kappa_1|^3 & |x_2-\kappa_2|^3 & \cdots & |x_2-\kappa_D|^3 \\ \vdots & \vdots & \ddots & \vdots \\ |x_n-\kappa_1|^3 & |x_n-\kappa_2|^3 & \cdots & |x_n-\kappa_D|^3 \end{bmatrix} \tag{2.16}$$

If the penalized spline fitting criterion (2.15) is divided by the error variance σ_ε^2, one can obtain the following fitting criterion.

$$\frac{1}{\sigma_\varepsilon^2}\|Y - X\psi - Z^* u\|^2 + \frac{\lambda}{\sigma_\varepsilon^2} u^T \Omega\ u \tag{2.17}$$

Let define $\sigma_u^2 = \sigma_\varepsilon^2/\lambda$ and let consider the vector u as a set of random parameters distributed as $\mathrm{MN}(0, \sigma_u^2\Omega^{-1})$ where u and ε are independent vectors. Brumback et al. [15] have shown that one can obtain an equivalent representation of the penalized spline regression in the form of linear mixed model (LMM) since the fitting criterion (2.15) is equal to the best linear unbiased predictor (BLUP) criterion in the LMM. Then the model explaining the functional relationship between y_i and x_i becomes

$$y = X\psi + Z^* u + \varepsilon \ \ ; \quad \mathrm{cov}\begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \sigma_u^2\Omega^{-1} & 0 \\ 0 & \sigma_\varepsilon^2 I_n \end{bmatrix} \tag{2.18}$$

Consider the reparametrization $b = \Omega^{1/2} u$ and define $Z = Z^*\Omega^{-1/2}$. Then the linear mixed model in 2.18 takes the form

$$y = X\psi + Zb + \varepsilon \ \ ; \quad \mathrm{cov}\begin{bmatrix} b \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \sigma_b^2 I_D & 0 \\ 0 & \sigma_\varepsilon^2 I_n \end{bmatrix} \tag{2.19}$$

where b is a $D \times 1$ vector of random effects distributed as $\mathrm{MN}(0, \sigma_b^2 I_D)$ and Z is a $n \times D$ data matrix calculated for the spline part. The details of the reparametrization used here is given in Crainiceanu et al. [14]. The parameters in the mixed model 2.19 can be estimated by using frequentist approach using Best Linear Unbiased Predictor (BLUP) or Penalized Quasi-Likelihood (PQL) estimation. We fitted this mixed model in the missing data mechanism of defined hierarchical model based on Bayesian approach. Note that the representation for the low-rank thin-plate spline regression in terms of linear mixed model is also appropriate for the binary regression model with the logit or probit link function (e.g. see Crainiceanu et al. [14]).

After all these derivations for the penalized spline regression by using low-rank thin-plate splines, the proposed semiparametric missing data mechanism defined in (2.12) takes the following mixed form.

$$h(p_{ik}) = \phi_{(k,0)} + r^*_{ik}\phi_{(k,1)} + y_i\phi_{(k,2)} + x_i^{miss}\phi_{(k,3)} + \sum_{k=1}^{s} Z_{(k)}b_{(k)} + x_i^{obs}\phi_{(k,4)} \tag{2.20}$$

where $Z_{(k)}$ is a $n \times D$ data matrix calculated for the penalized spline regression with D knots and $b_{(k)}$ is a $D \times 1$ vector of random effects distributed as $\mathrm{MN}(0, \sigma^2_{b(k)} I_D)$ corresponding to the missing covariate x_{ik}^{miss} for $k = 1, \ldots s$. Also, $Z_{(k)} = Z^*_{(k)}\Omega_{(k)}^{-1/2}$ in which $Z^*_{(k)}$ is an $n \times D$ matrix and $\Omega_{(k)}$ is a known $D \times D$ penalty matrix as follows.

$$Z^*_{(k)} = \begin{bmatrix} |x_{1k}^{miss} - \kappa_{1(k)}|^3 & |x_{1k}^{miss} - \kappa_{2(k)}|^3 & \cdots & |x_{1k}^{miss} - \kappa_{D(k)}|^3 \\ |x_{2k}^{miss} - \kappa_{1(k)}|^3 & |x_{2k}^{miss} - \kappa_{2(k)}|^3 & \cdots & |x_{2k}^{miss} - \kappa_{D(k)}|^3 \\ \vdots & \vdots & \ddots & \vdots \\ |x_{nk}^{miss} - \kappa_{1(k)}|^3 & |x_{nk}^{miss} - \kappa_{2(k)}|^3 & \cdots & |x_{nk}^{miss} - \kappa_{D(k)}|^3 \end{bmatrix} \tag{2.21}$$

and

$$\Omega_{(k)} = \begin{bmatrix} |\kappa_{1(k)} - \kappa_{1(k)}|^3 & |\kappa_{1(k)} - \kappa_{2(k)}|^3 & \cdots & |\kappa_{1(k)} - \kappa_{D(k)}|^3 \\ |\kappa_{2(k)} - \kappa_{1(k)}|^3 & |\kappa_{2(k)} - \kappa_{2(k)}|^3 & \cdots & |\kappa_{1(k)} - \kappa_{D(k)}|^3 \\ \vdots & \vdots & \ddots & \vdots \\ |\kappa_{D(k)} - \kappa_{1(k)}|^3 & |\kappa_{D(k)} - \kappa_{2(k)}|^3 & \cdots & |\kappa_{D(k)} - \kappa_{D(k)}|^3 \end{bmatrix} \tag{2.22}$$

where $\kappa_{1(k)} < \kappa_{2(k)} < \ldots < \kappa_{D(k)}$ are fixed knots suitably determined for each missing covariate x_{ik}^{miss}. Since some of the observations in x^{miss} in reality are unknown, to obtain the fixed knots, we consider the following procedure based on ML estimation. Let $x_{(1)}^{miss}$ and $x_{(2)}^{miss}$ be the missing and observed components of x^{miss} respectively and let X_{obs} denote the observed data for the covariates where $X_{obs} = (x^{miss}_{(2)}, x^{obs})$. The natural way to fill these missing observations is to sample from $f(x^{miss}|x^{obs}, \alpha_{(1)})$. To do so, we need to find the appropriate estimate of $\alpha_{(1)}$ denoted by $\hat{\alpha}_{(1)}$. Since we have m completely observed subjects, we can simply choose $\hat{\alpha}_{(1)}$ so that it maximizes the likelihood function $L(\alpha_{(1)}|X_{obs}) = f(x^{miss}_{(2)}|x^{obs}, \alpha_{(1)})$ based on the completely observed subjects. Then, we independently generate $x^{miss}_{(1),q} \sim f(x^{miss}|x^{obs}, \hat{\alpha}_{(1)})$ for $q = 1,...,Q$. By using those generated values for $x^{miss}_{(1)}$, we can obtain $D \times 1$ vectors of fixed knots $\kappa^{q}_{(k)} = (\kappa^{q}_{1(k)}, ..., \kappa^{q}_{D(k)})$ calculated by the sample quantiles of $(x^{miss}_{k(2)}, x^{miss}_{k(1),q})$ corresponding to probability $d/(D+1)$ for $d = 1,...D$ where $x^{miss}_{k(2)}$ is the observed observations for the missing covariate x^{miss}_{ik} and $x^{miss}_{k(1),q}$ is the qth imputed missing observations for the covariate x^{miss}_{ik} for $k = 1,...,s$ and $q = 1,...Q$. Then, we can obtain the fixed knots for each missing covariate x^{miss}_{ik} by using the following formula which averages the each components of the vector of fixed knots $\kappa^{q}_{(k)}$ calculated empirically by using different imputed missing observations for the covariate x^{miss}_{ik}.

$$\kappa_{d(k)} = \frac{1}{Q}\sum_{q=1}^{Q} \kappa^{q}_{d(k)} \tag{2.23}$$

where $\kappa_{d(k)}$ is the dth fixed knot for the missing covariate x^{miss}_{ik} for $d = 1,...,D$ and $k = 1,...,s$. In the simulation study, logistic regression model contains one NMAR covariate, namely x_{i1} and one fully observed covariate, namely x_{i2}. Then, we denote r_i as the missing data indicator of the covariate x_{i1} where $r_i = 1$when x_{i1} is missing and $r_i = 0$ when x_{i1} is observed. Then, the semiparametric missing data mechanism used in the simulation study by using the low-rank thin-plate splines with D knots is specified by a binary regression model in the following form.

$$f(r_i|y_i, x_{i1}, x_{i2}, Z, \phi, b) = p_i^{r_i}(1-p_i)^{1-r_i} \quad ; \quad p_i = P(r_i = 1|y_i, x_{i1}, x_{i2}, Z, \phi, b) \tag{2.24}$$

in which

$$h(p_i) = \phi_0 + \phi_1 y_i + \phi_2 x_{i1} + Zb + \phi_3 x_{i2} \tag{2.25}$$

where Z is a $n \times D$ data matrix calculated for the penalized spline regression with D knots and b is a $D \times 1$ vector of random effects distributed as $MN(0, \sigma^2_b I_D)$ corresponding to the missing covariate x_{i1}. The matrix Z are calculated by using the same procedure in (2.20). Note that parameters in (2.25) are redefined for the simplicity. The link function in (2.20) are chosen as probit in the simulation study since it has better computational properties than logit link function. The details of why the probit link function is used are given in Section 2.2.

Hierarchical Model

As a summary, the hierarchical model including proposed semiparametric missing data model to estimate the parameters of logistic regression with nonignorably missing covariates has the following three components:

1. $f(y_i|x_i^{miss}, x_i^{obs}\beta) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}$; $\mu_i = E(y_i|x_i^{miss}, x_i^{obs}, \beta) = P(y_i = 1|x_i^{miss}, x_i^{obs}, \beta)$

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1}^{miss} + ... + \beta_s x_{is}^{miss} + \beta_{s+1} x_{i,s+1}^{obs} + ... + \beta_p x_{ip}^{obs} \tag{2.26}$$

2.

$$x_{ik}^{miss}| \left(x_{i1}^{miss}, ..., x_{i,k-1}^{miss}, x_i^{obs}\right) \sim N\left[\alpha_{(1)k}^{(*,0)} + \sum_{j=1}^{k-1} \alpha_{(1)k}^{(*,j)} x_{ij}^{miss} + \sum_{j=s+1}^{p} \alpha_{(1)k}^{(*,j)} x_{ij}^{obs},\ \sigma_{X(k)}^2\right] \tag{2.27}$$

for $k = 1, ...s$

3.

$$f(r_{ik}|r_{ik}^*, y_i, x_i^{miss}, x_i^{obs}, Z, \phi_{(k)}, b) = p_{ik}^{r_{ik}}(1-p_{ik})^{1-r_{ik}}$$

$$h(p_{ik}) = \phi_{(k,0)} + r_{ik}^*\phi_{(k,1)} + y_i\phi_{(k,2)} + x_i^{miss}\phi_{(k,3)} + \sum_{k=1}^{s} Z_{(k)}b_{(k)} + x_i^{obs}\phi_{(k,4)}$$

$$b_{(k)}|\ \sigma_{b(k)}^2 \sim \text{MN}(0, \sigma_{b(k)}^2 I_D) \tag{2.28}$$

for $k = 1, ...s$

where $x_i^{miss} = (x_{i1}^{miss}, ..., x_{is}^{miss})$ and $x_i^{obs} = (x_{i,s+1}^{obs}, ..., x_{ip}^{obs})$ are continuous NMAR covariates and fully observed covariates respectively, and $Z = \{Z_{(k)}; k = 1, ...s\}$ and $b = \{b_{(k)}; k = 1, ..., s\}$.

Chen and Ibrahim [11] also considered a semiparametric model for the missing data mechanism in the same spirit of model 2.20. However, our work is different from their work in the following aspects: i) they considered the smooth functions for the covariates having possibly nonlinear effect on the missingness status; however we considered the smooth functions for all the missing covariates since the relationship between the missingness probability and the missing covariate is not identifiable from the observed data or the data may not contain sufficient information for the underlying missingness condition, ii) they investigated the efficiency properties of semiparametric missing data model under the assumption that the missing covariates have linear effect on the missingness probability while we focused on the model in which missing covariates have possibly nonlinear effect on the missingnees status, iii) They did not conduct a sensitivity analysis for the different number of knots while we investigated the effect of number of knots in our proposed semiparametric model on the estimation of β. In this thesis, we investigate the effect of the proposed semiparametric missing mechanism in (2.28) on estimating the β, the parameters of the main interest.

2.2 Bayesian Estimation

When we have a complex model for the data analysis (e.g. hierarchical modeling), the estimation techniques based on frequentist approach may be difficult to deal with due to the challenging derivations in the estimation procedure. For instance, the maximum likelihood (ML) estimation requires minimizing the likelihood function based on the observed data over the domain of the parameter which is sometimes a complicated procedure. However, Bayesian estimation techniques using the modern sampling methods provide many advantages for such a statistical modeling. In the past, statistical analysis based on the Bayesian approach was not so applicable because of computational incompetence. The recent developments on the computer-intensive sampling algorithms have improved the utilization of Bayesian estimation techniques since they have made possible to estimate the parameters of the complex models containing high-dimensional numerical integrations in the estimation procedure. It is well known that the Bayesian approaches are different from the frequentist approaches in philosophy. The extra information about the parameters can be incorporated to the analysis using the prior structure in Bayesian paradigm whereas frequentist methods depend on only the observed data. While estimating the parameters, the information from the likelihood based on the data and the apriori knowledge about the parameters are combined together to obtain posterior estimates. On the other hand, prior construction has an important role on the estimation of parameters in the Bayesian analysis since the estimates of the parameters may be sensitive to the choice of priors and convergence problem may exist for the sampling algorithms of posterior distribution due to the improper priors. Specifically, the posterior distribution of parameters related with the missing data mechanism are improper if improper uniform priors are used. In such situations, statistical methodologists propose the empirical Bayes based priors for the parameters. In this approach, the hyperparameters (the parameters of the prior distributions) are obtained based on the sampled data itself as well as the possible datasets that could be observed from the considered model. In addition, Bayesian estimation based on the sampling methods allows us the exact hypothesis tests on the parameters that are valid not only in large samples but also in finite samples.

In the Bayesian inference, unknown quantities such as parameters (whether fixed or random), hierarchical parameters, latent random effects and missing data are treated as random (see Gelman and Rubin [16]). The basic idea behind the Bayesian estimation based on sampling methods is simple yet requires care: after observing the data y, empirical density for the each component of parameter set $\theta = (\theta_1, ..., \theta_d)$ is obtained by drawing a set of random observations for the parameter θ_j from its full conditional posterior distribution $f(\theta_j|\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_p, y)$, then statistical inference on the parameter θ_j can be accomplished by using those empirical densities providing the distributional characteristics (e.g. moments, quantiles etc.) of the parameters. The novel Markov chain Monte Carlo (MCMC) sampling methods are very crucial tools to obtain a set of random draws for the parameter θ_j. However, the Bayesian procedure and accompanying sampling methods can be dangerous due to the inappropriate modeling, improper priors and convergence issues.

Let $\theta = (\theta_1, ..., \theta_d)$ be the parameter sets of any statistical model and $f(\theta)$ be the joint prior distribution for the parameters. Then, the joint posterior distribution which is the updated knowledge on the parameters can be simply obtained by combining the information from the prior knowledge which is put into a model format by $f(\theta)$, and the likelihood of the data $f(y|\theta)$ as follows.

$$\begin{aligned} f(\theta|y) &= f(y|\theta)f(\theta)/f(y) \\ &\propto f(y|\theta)f(\theta) \end{aligned} \tag{2.29}$$

where $f(y)$ is the marginal likelihood of the data which is a function of only observations but not a function of θ and considered as a normalizing constant for the posterior distribution $f(\theta|y)$. To obtain the marginal posterior distribution of the parameter θ_j, we need to integrate out all the other parameters:

$$f(\theta_j|y) = \int ... \int f(\theta_1, ..., \theta_d|y)\ d\theta_1 ... d\theta_{j-1} d\theta_{j+1} ... d\theta_d \tag{2.30}$$

Mostly, integration in (2.30) is very complicated (and sometimes impossible) to perform. To deal with the situation, the Gibbs sampler algorithm is constructed to obtain the random samples from the marginal posterior distribution of each parameter in the multiparameter models (see Gelfand and Smith [17]; Gilks et al. [18]; Casella and George [19]). At the tth iteration of this iterative algorithm, each parameter is updated by using the corresponding full conditional posterior distribution as follows.

1. $\theta_1^{(t)} \sim f_1(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, ..., \theta_d^{(t-1)}, y)$
2. $\theta_2^{(t)} \sim f_2(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_d^{(t-1)}, y)$

   $\vdots$

d. $\theta_d^{(t)} \sim f_d(\theta_d|\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_{d-1}^{(t)}, y)$

where each $\theta_j^{(t)}$ can be sampled from $f_j(\theta_j|\theta_1^{(t)}, ..., \theta_{j-1}^{(t)}, \theta_{j+1}^{(t-1)}, ..., \theta_d^{(t-1)}, y)$ by using the appropriate sampling algorithms (e.g. Metropolis-Hastings (M-H) algorithms, direct sampling algorithms etc.) for $t = 1, ..., T$, and $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})$ is the initial points. After obtaining a large number of T observations $\theta_j^{(1)}, \theta_j^{(2)}, ..., \theta_j^{(T)}$ for each parameter, we can simply estimate the $E\left(g(\theta_j)|y\right) = \int g(\theta_j) f(\theta_j|y) d\theta_j$ by

$$\overline{g_T} = \overline{E\left(g(\theta_j)|y\right)} = \frac{1}{T}\sum_{t=1}^{T} g(\theta_j^{(t)}) \tag{2.31}$$

where $\overline{g_T}$ is the Bayesian estimate of the parameter θ_j (i.e. the estimated expected value based on the posterior distribution) when g is the identity function. Note that full conditional posterior distribution $f_j(\theta_j|\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_d, y)$ in the Gibbs sampler algorithm can be easily derived from the joint posterior distribution of the parameters simply by extracting the terms corresponding to θ_j. The main motivation behind the Gibbs sampler algorithm is to reduce the number of iterations by using the parameter-by-parameter updating and to avoid calculating the marginal posterior distribution of the each parameter which is very complicated as in (2.30). Note also that, there is no need to use additional sums in (2.31) in terms of the observations for the other parameters $(\theta_1^{(t)}, ..., \theta_{j-1}^{(t)}, \theta_{j+1}^{(t)}, ..., \theta_d^{(t)})$ for $t = 1, ..., T$ since those parameters have already converge to their expected values in the long run due to the parameter-by-parameter updating.

In the hierarchical models (e.g. missing data models), unobserved quantities such as latent random effects and missing data can be also treated as random besides the parameters. To estimate the parameters of such models including latent random effects or missing data by using the fully Bayesian approach, Tanner and Wong [20] propose the Data Augmentation (DA) algorithm which is a special case of the Gibbs sampler and considered as the stochastic version of the EM algorithm. Let y be the observed data and y^* be the latent data such as latent random effects or missing data, and y^{com} denotes the complete data. That is $y^{com} = (y, y^*)$. Note that y^{com} is the dataset augmented by y^*. The joint conditional distribution of all the unknown quantities conditional on the observed data can be written as

$$f(\theta, y^*|y) = \frac{f(y^*, y|\theta) f(\theta)}{f(y)} \propto f(y^*, y|\theta) f(\theta) \tag{2.32}$$

where $f(y^*, y|\theta) = f(y^{com}|\theta)$ is the likelihood based on the complete data and $f(\theta)$ is the joint prior distribution of the parameters. To obtain the posterior estimates of the parameters $\theta = (\theta_1, ..., \theta_d)$, one needs to integrate out y^* from $f(\theta, y^*|y)$. Then

$$f(\theta|y) = \int f(\theta, y^*|y) dy^* = \int f(\theta|y^*, y) f(y^*|y) dy^* \quad \text{where} \quad f(y^*|y) = \int f(\theta, y^*|y) d\theta \tag{2.33}$$

Here $f(y^*|y)$ is the joint predictive density of y^* given the observed data and $f(\theta|y^*, y) = f(\theta|y^{com})$ is the joint posterior distribution of the parameters based on the complete data. Most of the time, the integrals in (2.33) are analytically difficult to obtain. However, the random draws from $f(\theta|y)$ and $f(y^*|y)$ can be obtained by using the Gibbs sampler algorithm as defined above. Then, the DA algorithm has the following two steps at the tth iteration:

1. $y^{*(t)} \sim f(y^*|\theta^{(t-1)}, y) \implies$ Imputation-Step

2. $\theta^{(t)} \sim f(\theta|y^{*(t)}, y) \implies$ Posterior-Step

where $y^{*(t)}$ and $\theta^{(t)}$ can be sampled from those conditional distributions by using the appropriate sampling algorithms (e.g. Gibbs sampler, M-H and direct sampling algorithms) for $t = 1, ..., T$. In the missing data problems, missing components can be treated as random and filled up by using the DA algorithm based on the fully Bayesian approach.

In the following sections, proposed Bayesian methodology is explained for the hierarchical modeling whose components are given in (2.26), (2.27) and (2.28) to estimate the parameters of logistic regression with nonignorably missing covariates. The prior construction for all the parameters including the fixed coefficients of each model and variance components are described in Section 2.2.1. The posterior construction including the employed Gibbs sampler algorithm for the proposed hierarchical model is explained in Section 2.2.2, and then the way of posterior calculations by using the WinBUGS is explained in Section 2.2.3.

2.2.1 Prior Construction

In the Bayesian analysis, prior knowledge on the parameters is also accounted as well as the likelihood information based on the sampled data. The effect of the prior distribution on the parameter estimation depends on the how informative the prior is and the size of the data. For example, noninformative priors lead to posterior estimates that are very close to the ML estimates since posterior knowledge about the parameters is not considerably influenced by these priors. On the other hand, informative priors have a greater influence on the estimation especially for the smaller samples. The prior selection is an important issue due to the following problems: i) the posterior estimate for the parameter may not be robust to the choice of the hyperparameters of prior, ii) improper priors can lead to improper posterior distribution, and iii) the choice of prior may result in poor mixing and hence slow convergence in MCMC framework. Thus, the choice of an appropriate prior requires care to avoid these problems. If there is no existing information about the parameters, noninformative proper priors with large variance can also be used instead of noninformative improper priors (e.g. improper uniform prior). Using conjugate proper prior is advantageous as it results in direct sampling to construct the posterior distribution in MCMC. In some circumstances, it is better to use informative priors (e.g. the empirical Bayes based priors) instead of noninformative ones to extract more information from the data since the likelihood alone may not contain sufficient information about the parameters as in missing data models. To construct the empirical Bayes based prior, the information from the data itself and the possible datasets that could be observed from the considered model is used to set the hyperparameters of this prior. For usefulness on empirical Bayesian methods one can refer to Carlin and Louis [21].

In the hierarchical modeling whose components are given in (2.26), (2.27) and (2.28), the parameter space consists of $(\beta, \alpha^*_{(1)}, \sigma^2_X, \phi, \sigma^2_b)$ where $\beta = \{\beta_j; j = 1, ..., p\}$, $\alpha^*_{(1)} = \{\alpha^{(*,j)}_{(1)k}; j = 0, ..., k-1, s+1, ..., p;\ k = 1, ..., s\}$, $\sigma^2_X = \{\sigma^2_{X(k)}; k = 1, ..., s\}$, $\phi = \{\phi_{(k,j)}; k = 1, ...s; j = 1, ...4\}$ and $\sigma^2_b = \{\sigma^2_{b(k)}; k = 1, ..., s\}$. The prior distributions for all these parameters are to be constructed for the Bayesian inference. The prior construction for the parameters is summarized in Table 2.1.

Table 2.1: The prior specification corresponding to each parameter set

Parameter Set	Prior	Explanation
β	$\beta_j \sim \mathrm{U}(-\infty,\infty)$	improper uniform prior
$\alpha^*_{(1)k}$; $k=1,...,s$	$\alpha^{(*,j)}_{(1)k} \sim \mathrm{U}(-\infty,\infty)$	improper uniform prior
$\sigma^{-2}_{X(k)}$; $k=1,...,s$	$\sigma^{-2}_{X(k)} \sim \mathrm{Gamma}(0.01,0.01)$	noninformative gamma prior
$\phi_{(k)}$; $k=1,...,s$	$\phi_{(k)} \sim \mathrm{MN}(\hat{\phi}_{(k)}, c_0\hat{\Sigma}_{\phi_{(k)}})$	empirical Bayes based multivariate normal prior
$\sigma^{-2}_{b(k)}$; $k=1,...,s$	$\sigma^{-2}_{b(k)} \sim \mathrm{Gamma}(\hat{\tau}_{\sigma_{b(k)^2}}, \hat{\varphi}_{\sigma_{b(k)^2}})$	empirical Bayes based gamma prior

Huang et al. [6] showed for the models with nonignorably missing covariates that the use of improper uniform priors for β (the parameters of the response model) and $\alpha^*_{(1)}$ (the location parameters of the missing covariate model) results in proper posterior. Therefore, we simply took an improper uniform prior for each β_j and $\alpha^{*,j}_{(1)k}$. For the inverse of each variance component $\sigma^{-2}_{X(k)}$, we used gamma prior with large variance. They also showed that the joint posterior distribution of $(\beta, \alpha^*_{(1)}, \sigma^2_X, \phi)$ is always improper whenever prior for ϕ is taken to be improper, and posterior inference on β and $\alpha^*_{(1)}$ is highly sensitive to the hyperparameters of prior for ϕ. They proposed the empirical Bayes based priors for the coefficients ϕ which provide proper joint posterior. In addition, it is also shown in that paper that these empirical Bayes based priors for ϕ result in accelerated convergence in the posterior calculations. Considering these facts, we used empirical Bayes based method to construct proper priors for ϕ.

The empirical Bayes based procedure proposed by Huang et al. [6] for constructing proper priors for their missing data model coefficients is adapted for the coefficients ϕ in our proposed semiparametric missing data model (2.28). The prior of the fixed parameters in each model for missing indicator r_{ik} is set as $\phi_{(k)} \sim \mathrm{MN}(\hat{\phi}_{(k)}, c_0\hat{\Sigma}_{\phi_{(k)}})$ where $\hat{\phi}_{(k)} = (\hat{\phi}_{(k,0)}, ..., \hat{\phi}_{(k,4)})$ is an estimate of $(p+k+1)\times 1$ mean vector and $\hat{\Sigma}_{\phi_{(k)}}$ is an estimate of $(p+k+1)\times(p+k+1)$ variance-covariance matrix, and c_0 is a constant used to account for the variation introduced by estimating the parameters. Huang et al. [6] conducted a sensitivity analysis investigating the effect of c_0 on the posterior distribution, and showed that posterior estimates of the parameters of the main interest β are not affected by the choice of c_0. Posterior estimates of

β by using our proposed semiparametric missing data model are also robust to the choice of c_0, and we set $c_0 = 10$ for the simulation study in Chapter 3. Remember that $x^{miss}_{(1)}$ and $x^{miss}_{(2)}$ are defined as the missing and observed components of x^{miss} respectively, and then $X_{obs} = (x^{miss}_{(2)}, x^{obs})$ is defined as the observed data for the covariates in Section 2.1. Also, define $D_{obs} = (r, y, x^{miss}_{(2)}, x^{obs})$ as the observed data for all variables. The following procedure can be used to obtain the hyperparameters of the multivariate normal prior.

1. Obtain $\hat{\alpha}_{(1)}$ which maximizes the likelihood function $L(\alpha_{(1)}|X_{obs}) = f(x^{miss}_{(2)}|x^{obs}, \alpha_{(1)})$ based on the m completely observed subjects.

2. Generate Q independent samples $x^{miss}_{(1),q} \sim f(x^{miss}|x^{obs}, \hat{\alpha}_{(1)})$ to obtain the imputed datasets $D^{imp}_q = (r, y, x^{miss}_{(1),q}, x^{miss}_{(2)}, x^{obs})$ for $q = 1, ..., Q$.

3. Obtain $\hat{\phi}^q_{(k)}$ which maximizes the likelihood function $L^q_{(k)}(\phi_{(k)}|D^{imp}_q) = f(r_{(k)}|r^*_{(k)}, y, x^{miss}_{(1),q}, x^{miss}_{(2)}, \phi_{(k)})$ corresponding to each missing indicator $r_{(k)}$. Note that each $r_{(k)}$ is fitted by the model $f(r_{(k)}|r^*_{(k)}, y, x^{miss}_{(1),q}, x^{miss}_{(2)}, \phi_{(k)}) = p^{r_{(k)}}_{(k)}(1-p_{(k)})^{1-r_{(k)}}$ in which $h(p_{(k)}) = \phi_{(k,0)} + r^*_{(k)}\phi_{(k,1)} + y\phi_{(k,2)} + x^{miss}\phi_{(k,3)} + x^{obs}\phi_{(k,4)}$ for $k = 1, ..., s$ and $q = 1, ..., Q$.

4. Denote $L^q_{(k)}(\phi_{(k)}|D^{imp}_q)$ as the imputed likelihood function for $\phi_{(k)}$. Then compute the information matrix $I^q_{(k)}(\hat{\phi}^q_{(k)})$ for each $\hat{\phi}^q_{(k)}$ which maximizes the likelihood $L^q_{(k)}(\phi_{(k)}|D^{imp}_q)$ where the information matrix can be calculated as $I(\phi) = \frac{-\partial^2 lnL(\phi)}{\partial\phi\partial\phi^T}$.

Then $\hat{\phi}_{(k)}$ and $\hat{\Sigma}_{\phi_{(k)}}$ which specify the hyperparameters of the empirical Bayes based prior of $\phi_{(k)}$ are calculated as

$$\hat{\phi}_{(k)} = \frac{1}{Q}\sum_{q=1}^{Q}\hat{\phi}^q_{(k)} \quad \text{and} \quad \hat{\Sigma}_{\phi_{(k)}} = \frac{1}{Q}\sum_{q=1}^{Q}\left[I^q_{(k)}(\hat{\phi}^q_{(k)})\right]^{-1} \tag{2.34}$$

In our proposed missing data model (2.28), $D \times 1$ vector of random effects $b_{(k)}$ is distributed as $\text{MN}(0, \sigma^2_{b(k)} I_D)$. We, at first, considered a noninformative gamma prior for the inverse of each variance component $\sigma^{-2}_{b(k)}$ for $k = 1, ..., s$. For the missing data model in the simulation study, we observed that posterior estimate of $\sigma^2_{b(k)}$ was quite sensitive to the choice of the hyperparameters of this gamma prior. We, for instance, used different gamma priors for $\sigma^{-2}_{b(k)}$ such as Gamma(0.1,0.1), Gamma(0.01,0.01) and Gamma(0.001,0.001). Alternatively, we used noninformative uniform prior distributions for the standard deviations $\sigma_{b(k)}$ as recommended by Gelman [22]. However, constructing a noninformative uniform prior such as $\text{U}(0.1, 100)$

resulted in severe numerical overflow in Gibbs sampling. On the other hand, constructing a more informative uniform prior with narrower range such as $\mathrm{U}(0.1, 10)$ would introduce too much informativeness on the parameters $\sigma^2_{b(k)}$ even there was no any existing prior knowledge about those parameters. Also one should note that $\sigma^2_{b(k)}$ is induced in the model by transformation of semiparametric models into mixed models and thus they are not part of the primary model. Therefore, the data lacks sufficient information about them. Accordingly, we proposed an empirical Bayes based gamma priors for each $\sigma^{-2}_{b(k)}$. In the simulation study, we also observed that using empirical Bayes based gamma priors for $\sigma^{-2}_{b(k)}$ results in accelerated convergence in the posterior calculations of $b_{(k)}$ and $\sigma^{-2}_{b(k)}$ (see the corresponding Brooks-Gelman-Rubin's convergence diagnostics and autocorrelation plots in Appendix D.1). To calculate the hyperparameters of the empirical based Gamma$(\hat{\tau}_{\sigma_{b(k)^{-2}}}, \hat{\varphi}_{\sigma_{b(k)^{-2}}})$ prior for each $\sigma^{-2}_{b(k)}$, we proposed the use of following procedure.

1. Obtain $\hat{\alpha}_{(1)}$ which maximizes the likelihood function $L(\alpha_{(1)}|X_{obs}) = f(x^{miss}_{(2)}|x^{obs}, \alpha_{(1)})$ based on the m completely observed subjects.

2. Generate Q independent samples $x^{miss}_{(1),q} \sim f(x^{miss}|x^{obs}, \hat{\alpha}_{(1)})$ to obtain the imputed datasets $F^{imp}_q = (r, Z_q)$ where $Z_q = \{Z_{(k),q}; k = 1, ..., s\}$ and $Z_{(k),q}$'s are $n \times D$ data matrices for the nonparametric part of missing data model calculated from $(x^{miss}_{(1),q}, x^{miss}_{(2)})$ as in (2.28) for $k = 1, .., s$ and $q = 1, ..., Q$.

3. Obtain $\hat{b}^q_{(k)}$ which maximizes the likelihood function $L^q_{(k)}(b_{(k)}|F^{imp}_q) = f(r_{(k)}|Z, b)$ corresponding to each missing indicator $r_{(k)}$. Note that each $r_{(k)}$ is fitted by the model $f(r_{(k)}|Z, b) = p^{r_{(k)}}_{(k)}(1 - p_{(k)})^{1-r_{(k)}}$ in which $h(p_{(k)}) = c_{(k),q} + \sum_{k=1}^{s} Z_{(k),q} b^q_{(k)}$ where $c_{(k),q}$ is the intercept coefficients for $k = 1, ..., s$ and $q = 1, ..., Q$.

4. Denote $L^q_{(k)}(b_{(k)}|F^{imp}_q)$ as the imputed likelihood function for $b_{(k)}$. Then compute the information matrix $I^q_{(k)}(\hat{b}^q_{(k)})$ for each $\hat{b}^q_{(k)}$ which maximizes the likelihood $L^q_{(k)}(b_{(k)}|F^{imp}_q)$ where the information matrix can be calculated as $I(b) = \frac{-\partial^2 lnL(b)}{\partial b \partial b^T}$.

5. Calculate the estimated variance-covariance matrix for each $D \times 1$ parameter vector $b^q_{(k)}$ where D is the number of knots by using $\hat{\Sigma}^q_{b_{(k)}} = \left[I^q_{(k)}(\hat{b}^q_{(k)})\right]^{-1}$.

6. Then we determined the shape $\hat{\tau}_{\sigma_{b(k)^2}}$ and scale $\hat{\varphi}_{\sigma_{b(k)^2}}$ parameters of the gamma prior for each $\sigma^{-2}_{b(k)}$ so that the mean of this prior is equal to the averages of $\left[\mathrm{diag}\left(\frac{1}{Q}\sum_{q=1}^{Q} \hat{\Sigma}^q_{b_{(k)}}\right)\right]^{\cdot -1}$ and variance is equal to 1. Here, $[]^{\cdot -1}$ denotes the elementwise inverse.

2.2.2 Posterior Construction

Let $D_{obs} = (r, y, x_{(2)}^{miss}, x^{obs})$ denote the observed dataset where $x_{(1)}^{miss}$ and $x_{(2)}^{miss}$ denote missing and observed components of x^{miss} respectively. Note that $n \times D$ data matrices $Z = \{Z_{(k)}; k = 1, ..., s\}$ for the nonparametric part of the missing data model are not included in D_{obs} since $Z_{(k)}$'s are the function of $x_{(k)}^{miss}$. Let also $D_{com} = (D_{obs}, x_{(1)}^{miss}, b)$ denotes the complete dataset which is augmented by the unobserved covariates $x_{(1)}^{miss}$ and the latent random effects b. Then the joint conditional distribution of all the unknown quantities based on the observed data D_{obs} is represented as

$$\begin{aligned} f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2, x_{(1)}^{miss}, b|D_{obs}) &= \frac{f(x_{(1)}^{miss}, b, D_{obs}|\ \beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2) f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2)}{f(D_{obs})} \\ &\propto f(D_{com}|\ \beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2) f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2) \quad (2.35) \end{aligned}$$

where $f(D_{com}\ |\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2)$ and $f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2)$ are the complete likelihood and the joint prior distribution of the parameters respectively. To conduct a posterior inference on the parameters, the desired joint posterior distribution $f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2|D_{obs})$ and the joint predictive density $f(x_{(1)}^{miss}, b|D_{obs})$ are defined as

$$\begin{aligned} f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2|D_{obs}) &= \int\int f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2, x_{(1)}^{miss}, b|D_{obs})\ dx_{(1)}^{miss} db \\ &= \int\int f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2\ |D_{com}) f(x_{(1)}^{miss}, b|D_{obs})\ dx_{(1)}^{miss} db \end{aligned} \quad (2.36)$$

and

$$f(x_{(1)}^{miss}, b|D_{obs}) = \int f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2, x_{(1)}^{miss}, b|D_{obs})\ d\theta \quad (2.37)$$

where $\theta = (\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2)$ is the set of all parameters. Note that the multiple integrals in (2.36) and (2.37) are analytically difficult to obtain. Therefore, the data augmentation (DA) algorithm based on the Gibbs sampling approach can be performed to sample random draws from the desired joint posterior distribution $f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^2|D_{obs})$ and the joint predictive density $f(x_{(1)}^{miss}, b|D_{obs})$ based on the observed data. Accordingly, the unobserved quantities

$x_{(1)}^{miss}$ and b are to be sampled from the joint conditional distribution $f(x_{(1)}^{miss}, b|\ \beta, \alpha_{(1)}^{*}, \sigma_X^2, \phi, \sigma_b^2, D_{obs})$, and the unknown parameters $(\beta, \alpha_{(1)}^{*}, \sigma_X^2, \phi, \sigma_b^2)$ are to be sampled from the joint posterior distribution $f(\beta, \alpha_{(1)}^{*}, \sigma_X^2, \phi, \sigma_b^2|\ x_{(1)}^{miss}, b, D_{obs})$ based on the complete data at each iteration of DA algorithm. By using the Gibbs sampling algorithm within the DA algorithm, each of the unknown quantities β, $\alpha_{(1)}^{*}$, σ_X^2, ϕ, σ_b^2, $x_{(1)}^{miss}$ and b can be sampled from the corresponding full conditional distributions to obtain the random draws from the joint posterior distribution $f(\beta, \alpha_{(1)}^{*}, \sigma_X^2, \phi, \sigma_b^2|D_{obs})$ based on the observed sample. After obtaining the random draws from this joint posterior distribution, we can easily conduct a posterior inference on all of the parameters. Thus, at the tth iteration of this iterative algorithm, the missing components $x_{(1)}^{miss}$, the latent random effects b and the parameters are sampled from the following full conditional distributions.

$$\begin{aligned}
x_{(1)(t)}^{miss} &\sim f(x_{(1)}^{miss}|\ \beta_{(t-1)}, \alpha_{(1)(t-1)}^{*}, \sigma_{X(t-1)}^2, \phi_{(t-1)}, b_{(t-1)}, D_{obs}) \\
b_{(t)} &\sim f(b|\ \phi_{(t-1)}, \sigma_{b(t-1)}^2, x_{(1)(t)}^{miss}, D_{obs}) \\
\beta_{(t)} &\sim f(\beta|\ x_{(1)(t)}^{miss}, D_{obs}) \\
\alpha_{(1)(t)}^{*} &\sim f(\alpha_{(1)}^{*}|\ \sigma_{X(t-1)}^2, x_{(1)(t)}^{miss}, D_{obs}) \\
\sigma_{X(t)}^2 &\sim f(\sigma_X^2|\ \alpha_{(1)(t)}^{*}, x_{(1)(t)}^{miss}, D_{obs}) \\
\phi_{(t)} &\sim f(\phi|\ x_{(1)(t)}^{miss}, b_{(t)}, D_{obs}) \\
\sigma_{b(t)}^2 &\sim f(\sigma_b^2|\ b_{(t)})
\end{aligned}$$

The functional forms of each full conditional distribution of parameters and latent variables can be easily derived from (2.35) by extracting the terms corresponding to the associated parameter. The joint prior distribution of the parameters can be assumed independent apriori and the complete data likelihood can be written as

$$\begin{aligned}
f(D_{com}|\ \beta, \alpha_{(1)}^{*}, \sigma_X^2, \phi, \sigma_b^2) =& f(r, y, x_{(1)}^{miss}, x_{(2)}^{miss}, x^{obs}, b|\ \beta, \alpha_{(1)}^{*}, \sigma_X^2, \phi, \sigma_b^2) \\
=& f(r|\ y, x_{(1)}^{miss}, x_{(2)}^{miss}, x^{obs}, \phi, b) f(y|\ x_{(1)}^{miss}, x_{(2)}^{miss}, x^{obs}, \beta) \times \\
& f(x_{(1)}^{miss}, x_{(2)}^{miss}|\ x^{obs}, \alpha_{(1)}^{*}, \sigma_X^2) f(b|\ \sigma_b^2) \qquad (2.38)
\end{aligned}$$

The appropriate MCMC sampling algorithms (e.g. M-H algorithm) or direct sampling algorithms can be used within the Gibbs sampling algorithm to sample from these full conditional distributions. In this thesis, WinBUGS(Speigelhalter, Thomas, and Best [12]), a flexible and user-friendly software for the Bayesian analysis using MCMC methods, is used to carry out the sampling from the full conditional distributions and ultimately attain the joint posterior distribution in (2.36).

We, at first, considered the logit link function in the proposed semiparametric missing data model (2.28). However, the tedious `trap` messages were often appeared in WinBUGS due to the computational difficulties in sampling from the full conditional distribution of b. Note that the full conditional distribution of b is specified by $f(r|\ y, x_{(1)}^{miss}, x_{(2)}^{miss}, x^{obs}, b) f(b|\ \sigma_b^2)$. When the logit link function is used, the kernels of this full conditional distribution belong to different distributions which lead to inconvenient sampling for the b's. As a remedy for this ill condition, we considered the probit link function for the semiparametric missing data model. For the analysis of binary response data, Albert and Chib [23] proposed the use of probit link and augmenting the data by suitably constructed latent variables. We adopted this approach and latent W_{ik}'s are introduced for each missing data indicator r_{ik} for $k = 1, ..., s$ and $i = 1, ..., n$ such that

$$r_{ik} = \left\{ \begin{array}{ccc} 1 & ; & W_{ik} \geq 0 \\ 0 & ; & W_{ik} < 0 \end{array} \right\} \tag{2.39}$$

$$W_{ik} \sim N(\mu_{W_{ik}}, 1) \tag{2.40}$$

$$\mu_{W_{ik}} = \phi_{(k,0)} + r_{ik}^* \phi_{(k,1)} + y_i \phi_{(k,2)} + x_i^{miss} \phi_{(k,3)} + \sum_{k=1}^{s} Z_{(k)} b_{(k)} + x_i^{obs} \phi_{(k,4)} \tag{2.41}$$

Then

$$f(r_{ik}|W_{ik}) = \Phi(W_{ik})^{r_{ik}} (1 - \Phi(W_{ik}))^{(1-W_{ik})} \quad \text{and} \quad W_{ik}|\ (r_{ik}^*, y_i, x_i^{miss}, x_i^{obs}, Z, \phi_{(k)}, b) \sim N(\mu_{W_{ik}}, 1) \tag{2.42}$$

where Φ is the cumulative distribution function of standard normal distribution. Let $W_i = (W_{i1}, ..., W_{is})^T$ be the vector of latent W_{ik}'s and $\mu_{W_i} = (\mu_{W_{i1}}, ..., \mu_{W_{is}})^T$ be the vector of $\mu_{W_{ik}}$'s for $i = 1, ..., n$. Then $W_i \sim \mathrm{MN}(\mu_{W_i}, I_s)$. To apply this approach, we used the complete dataset denoted by $D_{com*} = (D_{obs}, x_{(1)}^{miss}, b, W)$. To obtain the joint posterior distribution based on the observed data, W also needs to be integrated out in (2.36). Accordingly, the parameters, missing components $x_{(1)}^{miss}$, and latent variables b and W are sampled from the following full conditional distributions at the tth iteration of the Gibbs sampling algorithm.

$$\begin{aligned}
x^{miss}_{(1)(t)} &\sim f(x^{miss}_{(1)}|\ \beta_{(t-1)},\alpha^*_{(1)(t-1)},\sigma^2_{X(t-1)},\phi_{(t-1)},b_{(t-1)},W_{(t-1)},D_{obs})\\
b_{(t)} &\sim f(b|\ \phi_{(t-1)},\sigma^2_{b(t-1)},x^{miss}_{(1)(t)},W_{(t-1)},D_{obs})\\
W_{(t)} &\sim f(\phi_{(t)},x^{miss}_{(1)(t)},b_{(t)},D_{obs})\\
\beta_{(t)} &\sim f(\beta|\ x^{miss}_{(1)(t)},D_{obs})\\
\alpha^*_{(1)(t)} &\sim f(\alpha^*_{(1)}|\ \sigma^2_{X(t-1)},x^{miss}_{(1)(t)},D_{obs})\\
\sigma^2_{X(t)} &\sim f(\sigma^2_X|\ \alpha^*_{(1)(t)},x^{miss}_{(1)(t)},D_{obs})\\
\phi_{(t)} &\sim f(\phi|\ x^{miss}_{(1)(t)},b_{(t)},W_{(t)},D_{obs})\\
\sigma^2_{b(t)} &\sim f(\sigma^2_b|\ b_{(t)})
\end{aligned}$$

where the likelihood based on the complete data in (2.38) becomes

$$\begin{aligned}
f(D_{com*}|\ \beta,\alpha^*_{(1)},\sigma^2_X,\phi,\sigma^2_b) =& f(r,y,x^{miss}_{(1)},x^{miss}_{(2)},x^{obs},b,W|\ \beta,\alpha^*_{(1)},\sigma^2_X,\phi,\sigma^2_b)\\
=& f(r|\ W)f(W|\ y,x^{miss}_{(1)},x^{miss}_{(2)},x^{obs},\phi,b)f(y|\ x^{miss}_{(1)},x^{miss}_{(2)},x^{obs},\beta)\times\\
& f(x^{miss}_{(1)},x^{miss}_{(2)}|\ x^{obs},\alpha^*_{(1)},\sigma^2_X)f(b|\ \sigma^2_b) \qquad (2.43)
\end{aligned}$$

Notice that the full conditional distribution for b is specified by the product of the two distributions, namely $f(W|\ y,x^{miss}_{(1)},x^{miss}_{(2)},x^{obs},\phi,b)$ and $f(b|\ \sigma^2_b)$ when the probit link used in the missing data model. Since both distributions have normal kernels, the full conditional distribution for b is also normal which is very easy to sample from. As a result, the `trap` messages vanished and sampling was accomplished. For the specified hierarchical model in the simulation study, the functional forms of the full conditional distributions to be used in the Gibbs sampling algorithm are given in Appendix A.

2.2.3 Bayesian Inference using Gibbs Sampling via WinBUGS

In the previous sections, the general methodology for the Bayesian analysis of logistic regression with NMAR covariates is explained. As described previously, the model framework consists of three components, namely response model, covariate distribution and missing data model, and we proposed a semiparametric approach to model the underlying missing data mechanism for each missing indicator r_{ik} to provide a flexible way of modeling the true relationship between the missingness and the missing covariates, linear or nonlinear. In Chapter 3, the simulation study is designed to investigate the effect of the proposed semiparametric missing data model under the different true missing data mechanism on the parameters of main interest. Specifically, in these simulations, a logistic regression response model with two continuous explanatory variables $x_i = (x_{i1}, x_{i2})$ is considered where x_{i1} is possibly nonignorably missing and x_{i2} is fully observed. Thus, the hierarchical model with the following three components is considered to estimate the parameters of logistic regression in the simulation study.

1. $$f(y_i|x_{i1}, x_{i2}, \beta) = \mu^{y_i}(1-\mu_i)^{1-y_i}\ ; \quad \mu_i = E(y_i|x_{i1}, x_{i2}, \beta) = P(y_i = 1|x_{i1}, x_{i2}, \beta)$$

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \tag{2.44}$$

2. $$x_{i1}\,|\,x_{i2} \sim N(\alpha_0 + \alpha_1 x_{i2},\ \sigma_x^2) \tag{2.45}$$

3. $$f(r_i|W_i) = \Phi(W_i)^{r_i}(1-\Phi(W_i))^{(1-r_i)}$$

$$W_i\ |(y_i, x_{i1}, x_{i2}, Z, \phi, b) \sim N(\mu_{W_i}, 1) \quad \text{and} \quad \mu_{W_i} = \phi_0 + \phi_1 y_i + \phi_2 x_{i1} + Zb + \phi_3 x_{i2}$$

$$b \sim \text{MN}(0, \sigma_b^2 I_D) \tag{2.46}$$

where all parameters and variables are as explained in Section 2.1 and Section 2.2.2. In this section, we use this hierarchical model to illustrate how to create a WinBUGS code to carry out Gibbs sampling for the proposed semiparametric approach. We also shed light onto how WinBUGS update the Markov chain in each iteration of Gibbs algorithm for our model. To computationally obtain the posterior distributions of the all parameters via Gibbs sampling, the latent variables and the parameters are to be sampled from the corresponding full conditional distributions. All the components of the hierarchical model and the prior distributions of the parameters are needed to be defined in the WinBUGS interface, and then WinBUGS constructs automatically all the full conditional distributions corresponding to the all unknown quantities (e.g. the parameters, the missing components $x_{(1)}^{miss}$ and the latent variables b and W in our model) and determine the appropriate sampling algorithm to sample from these full conditional distributions. However, it is a great advantage for the user to obtain the functional forms of these full conditional distributions by hand for such complex analyses in order to construct appropriate priors that result in Gibbs samplings performed efficiently. In WinBUGS interface, the tedious `trap` messages were observed while running the simulation study due to the inappropriate prior construction. The functional forms of these full conditional distributions which are derived in Appendix A have very important role on the appropriate prior construction and the appropriate link function choice for the missing data model. To sample from the full conditional distribution of each parameter and latent variable, WinBUGS choose the appropriate sampling algorithm which are defined in Appendix B. In Table B.1, the kernels of the full conditional distributions, and the corresponding sampling algorithms within the Gibbs sampling are explained.

Table 2.2: The kernels of the full conditional distributions and corresponding sampling algorithms

Latent/ Parameter	Full Conditional Kernels	Sampling Method	Explanation
$x_{(1)}^{miss}$	Normal, Logistic	Metropolis Hastings method with normal proposal	for real non-linear function
b	Normal	Direct sampling	
w	Truncated Normal	Derivative-Free Adaptive Rejection sampling	for log-concave function
β	Logistic	Slice sampling	for logistic regression
α	Normal	Direct sampling	
ϕ	Multivariate Normal	Direct sampling	
σ_x^{-2} and σ_b^{-2}	Normal, Gamma	Direct sampling from Gamma	conjugate gamma prior

In WinBUGS, the initial points for each latent variables and parameters are needed to be determined to start the corresponding Markov chain. The initial points are arbitrarily chosen for the parameters of response model and covariate distribution, namely $(\beta_0, \beta_1, \beta_2)$, (α_0, α_1) and σ_x^2. However, arbitrarily chosen initials for the parameters $(\phi_0, \phi_1, \phi_2, \phi_3)$ and σ_b^2 lead to the `trap` messages in WinBUGS since the possibly inappropriate initials do not allow to start sampling from the corresponding full posterior distributions. The calculated means for the empirical Bayes based multivariate normal priors of ϕ proposed in Section 2.2.1 are used for the initial points of these parameters. Also, initial points for the missing components $x_{1(1)}^{miss}$ are chosen from the simulation based imputed datasets described in the procedure for the empirical Bayes based priors. The starting points for the random effects b are chosen as the averages of the ML estimations $\hat{b}^q$ from the imputed datasets (see the procedure for calculating the hyperparameters of the empirical Bayes based gamma priors of $\sigma^{-2}_{b,(k)}$ in Section 2.2.1). The initial points are chosen arbitrarily for the latent variables W.

The following WinBUGS code is created for posterior calculations based on the proposed hierarchical model in Section 2.1. This code can be specifically used for logistic regression model with two continuous covariates one of which is possibly missing(NMAR) and the other one is fully observed. However, it can be used for different GLMs having NMAR covariates with just minor adjustments. The likelihood part is specified in WinBUGS as follows:

```
for(i in 1:N)

   { # model for covariate subject to missingness(X1)
   X[i,1] ~ dnorm(mu[i],invsigma2x)
   mu[i]<-alpha[1]+alpha[2]*X[i,2]

   # model for response
   Y[i] ~ dbern(piY[i])
   logit(piY[i])<-beta[1]+beta[2]*X[i,1]+beta[3]*X[i,2]

   # model for missing data model
   for(j in 1:k)
   { znknots[i,j]<-pow(X[i,1]-knots[j],3) }

   for(j in 1:k)
   { z[i,j]<-inprod(znknots[i,],invsqrtomega[,j]) }
   W[i] ~ dnorm(wmean[i],1) I(lb[R[i]+1],ub[R[i]+1])

   wmean[i]<-phi[1]+phi[2]*Y[i]+phi[3]*X[i,2]+mfixed[i]+mrand[i]
   mfixed[i]<-phi[4]*X[i,1]
   mrand[i]<-inprod(b[],z[i,]) }
```

In WinBUGS, user should load the data: total sample size `N`, $n \times 1$ response vector `Y`, $n \times 1$ missing indicator vector `R`, $n \times 2$ design matrix `X[,]`, number of knots `k`, $k \times 1$ fixed knots vector `knots` for x_1 and $n \times k$ matrix `invsqrtomega` for calculating the random effect design matrix z. WinBUGS is called from within MATLAB using `mat2bugs.m` function to run the simulations in Chapter 3. These fixed values are calculated by using MATLAB, and sent to WinBUGS via `mat2bugs.m` function. Note that random effect design matrix z is calculated in WinBUGS since we want to update this matrix at every iteration of Gibbs sampling, which means that updated missing $x^{miss}_{(1)}$ values at each iteration are considered in the random effect desing matrix z. Probit link in the code is specified by defining truncated normal distribution with lower bound `lb[R[i]+1]` and upper bound `ub[R[i]+1]` where `lb=(-50,0)` and `ub=(0,50)` are fixed constants with arbitrarily chosen large values in magnitude. The priors of parameters and the distribution of the random coefficients b are specified in WinBUGS as follows:

```
# Priors for Parameters of Response and Covariate Models
for(j in 1:2){ alpha[j] ~ dflat() }
for(j in 1:3){ beta[j] ~ dflat() }
invsigma2x ~ dgamma(0.01,0.01)

# Distribution of the random coefficients
for(j in 1:k)
{ b[j] ~ dnorm(0,invsigma2b) }

# Priors for Missingness Part
phi[1:4] ~ dmnorm(phipriormean[1:4],phipriorcov[1:4,1:4])
invsigma2b ~ dgamma(invsigma2b1,invsigma2b2)
```

The parameters of the priors for ϕ and σ_b^{-2} are determined in MATLAB using the empirical Bayesian method described in Section 2.2.1 and sent to WinBUGS via `mat2bugs.m` function again. In WinBUGS code, `phipriormean` and `phipriorcov` are empirical Bayesian based mean vector and variance-covariance matrix of multivariate normal prior of ϕ. Also, `invsigma2b1` and `invsigma2b2` are emprical Bayesian based shape and scale parameters of gamma prior of σ_b^{-2}.

CHAPTER 3

SIMULATION STUDY

To assess the performance of the proposed hierarchical model and the accompanying estimation procedure for estimating the parameters of the main interest β, we designed Monte Carlo simulations under the various true missing data mechanisms. In the previous chapters, we explained the general methodology to estimate the parameters of the logistic regression with the proposed semiparametric missing data model in order to capture any functional relationship between the missingness and the missing covariates more flexibly. To investigate the bias and efficiency properties of the resulting estimators, a logistic regression with two continuous covariates is considered as a specific example of this estimation procedure in all simulations, and we assume that one of the covariates namely x_{i1} is *not missing at random* (NMAR) and the other one namely x_{i2} is completely observed. We considered mainly three different true missingness mechanisms where each of them constitutes 15% and 35% missingness for the missing covariate x_{i1}. Under all true missingness mechanisms, we fitted three hierarchical models one of which has our proposed semiparametric missing data model and the others have the fully parametric missing data model. The all simulated datasets and the fixed values (e.g. fixed knots vector `knots`) used by WinBUGS are constructed in MATLAB, and WinBUGS is called from within MATLAB using `mat2bugs.m` function to fit each hierarchical model for the analysis of each simulated dataset. The sample `simulation.m` function to run the simulation study for the specified true missingness mechanism and the missing percentage (e.g. for the true missingness mechanism II and 15% missing percentage) is given in Appendix C. The data generation procedures for the simulations are explained in Section 3.1. The fitted hierarchical models for the analysis of the simulated datasets are explained and the MCMC diagnostic checks for these fitted models are carried out in Section 3.2. The results obtained by the simulations are given and discussed in Section 3.3.

3.1 Simulated Dataset

For all subsequent simulations, we use $N = 100$ replicates including $n = 250$ response variable y_i and covariates (x_{i1}, x_{i2}). The binary response variables are independently generated from the following logistic regression model

$$f(y_i|x_{i1}, x_{i2}, \beta) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}\ ; \quad \mu_i = E(y_i|x_{i1}, x_{i2}, \beta)$$

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \tag{3.1}$$

where we assume that response variable y_i and covariate x_{i2} are always completely observed, and covariate x_{i2} is NMAR. Our main goal is estimating β_1, the parameter representing the association between the response and the NMAR covariate while the full observed covariate is an adjusting factor. The values $(\beta_0, \beta_1, \beta_2) = (2, 1, -1)$ are chosen as logistic regression coefficients which gives approximately equal number of cases $(y_i = 1)$ and controls $(y_i = 0)$ for each simulated dataset if the covariates have the following properties. For the simulated dataset, the covariates (x_{i1}, x_{i2}) are independently generated from the following normal distributions

$$x_{i2} \sim N(1,1) \quad \text{and} \quad x_{i1} \sim N(\alpha_0 + \alpha_1 x_{i2}, \sigma_x^2) \tag{3.2}$$

where $(\alpha_0, \alpha_1, \sigma_x^2) = (-1.5, 0.5, 0.75^2)$. Since we assume that covariate x_{i2} is fully observed and covariate x_{i1} is subject to missingness, we consider the following missingness construction for the simulated possibly missing covariate x_{i1}. Let r_i be the missing value indicator corresponding to x_{i1} and is 1 for missing x_{i1}. For the missing data generation process, we consider the missingness models of the following form:

$$f(r_i|\ y_i, x_{i1}, x_{i2}, \phi) = p_i^{r_i}(1-p_i)^{1-r_i}$$

$$h(p_i) = \phi_0 + \phi_1 y_i + \phi_2 x_{i2} + \phi_3 x_{i1} + \phi_3 m(x_{i1}) \tag{3.3}$$

where h is a probit link function and m is a smooth nonlinear function of x_{i1}. Specifically we consider the true missingness mechanisms (TMMs) below and use these to generate r_i's:

$$\text{TMM 0} : h(p_i) = \phi_0 + \phi_1 y_i + \phi_2 x_{i2} + \phi_3 x_{i1} \tag{3.4}$$

$$\text{TMM I} : h(p_i) = \phi_0 + \phi_1 y_i + \phi_2 x_{i2} + \phi_3 x_{i1} + \phi_4 x_{i1}^2 \tag{3.5}$$

$$\text{TMM II} : h(p_i) = \phi_0 + \phi_1 y_i + \phi_2 x_{i2} + \phi_3 x_{i1} + \phi_4 \left(0.5 - \frac{1}{1+(x_{i1}+1)^4}\right) \tag{3.6}$$

With TMM 0, we want to determine the risk of using the proposed semiparametric missing data model when in fact the underlying true missingness mechanism indicates exact linear relationship between the NMAR covariate and its missingness probability. With TMM I and TMM II, we want to see how effective the proposed semiparametric missing data model when the NMAR covariate and its missingness probability is indeed nonlinearly related. We used $(\phi_1, \phi_2, \phi_3) = (1, 1, 1)$ and $(\phi_1, \phi_2, \phi_3, \phi_4) = (1, 1, 1, 3)$ as parameters of TMM 0 and TMM I respectively. For TMM II, the parameters $(\phi_1, \phi_2, \phi_3, \phi_4) = (1, 1, 0, 5)$ are chosen so that it represents a common real life situation in which the nonresponse (hence the missing information on the covariate) is more likely to occur if the true measurements are extreme. For instance, anorectic or obese people can be much more likely to refuse to report any weight related information compared to the people with rather normal weight. The functional relationship between the missing covariate x_{i1} and its missingness probability p_i for TMM II is given in Figure 3.1. For all the true missingness mechanisms considered, ϕ_0 is chosen to have 15% or 35% missing x_{i1}. The sample `Phi0.m` function to calculate ϕ_0 for the specified true missingness mechanism and the missing percentage (e.g. for the true missingness mechanism II and 15% missing percentage) is given in Appendix C.

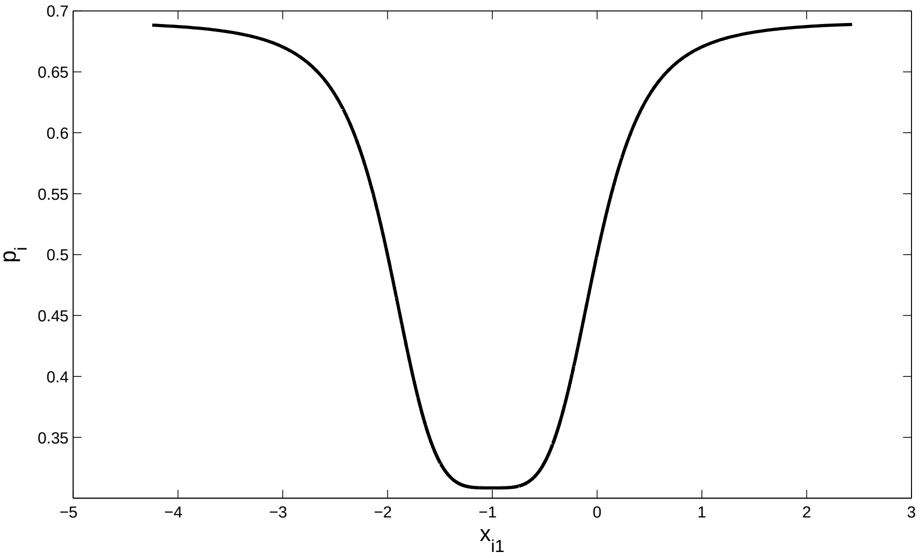

Figure 3.1: The functional relationship between the missing covariate x_{i1} and its missingness probability p_i for TMM II

3.2 Model Fitting

To investigate the performance of the proposed semiparametric missingness model, we consider the following three hierarchical models for the analysis of each simulated dataset.

$$\text{Model I}: \left\{ \begin{array}{c} \operatorname{logit}(P(y_i = 1|x_{i1}, x_{i2}, \beta)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\ x_{i1}|x_{i2} \sim N(\alpha_0 + \alpha_1 x_{i2}, \sigma_x^2) \\ h(P(r_i = 1|y_i, x_{i1}, x_{i2}, \phi)) = \phi_0 + \phi_1 y_i + \phi_2 x_{i2} + \phi_3 x_{i1} \end{array} \right\}$$

$$\text{Model II}: \left\{ \begin{array}{c} \operatorname{logit}(P(y_i = 1|x_{i1}, x_{i2}, \beta)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\ x_{i1}|x_{i2} \sim N(\alpha_0 + \alpha_1 x_{i2}, \sigma_x^2) \\ h(P(r_i = 1|y_i, x_{i1}, x_{i2}, \phi)) = \phi_0 + \phi_1 y_i + \phi_2 x_{i2} + \phi_3 x_{i1} + \phi_4 x_{i1} x_{i2} \end{array} \right\}$$

$$\text{Model III}: \left\{ \begin{array}{c} \operatorname{logit}(P(y_i = 1|x_{i1}, x_{i2}, \beta)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \\ x_{i1}|x_{i2} \sim N(\alpha_0 + \alpha_1 x_{i2}, \sigma_x^2) \\ h(P(r_i = 1|y_i, x_{i1}, x_{i2}, \phi)) = \phi_0 + \phi_1 y_i + \phi_2 x_{i2} + \phi_3 x_{i1} + \sum_{j=1}^{D} b_j Z_{ij} \end{array} \right\}$$

where h is probit link function, b_j is the jth unobserved random effect distributed as $N(0, \sigma_b^2)$, and z_{ij} is the (i, j)th element of the random effect design matrix Z calculated based on x_{i1} as in (2.28) for $j = 1, ..., D$. For Model III, we consider the number of knots (D) from 2 to 8 to assess the effect of different number of knots selection on the inference of main model. To calculate the jth fixed knot ($j = 1, ..., D$) from the missing covariate x_{i1}, we consider our proposed procedure based on ML estimation given in Section 2.1 to fill missing components of x_{i1} and attain the fixed knots based on the incomplete covariate x_{i1}. The imputation process in this procedure is repeated $Q = 100$ times.

For Models I&II, we also use improper uniform priors for αs and βs, noninformative gamma prior for σ_x^{-2} and the empirical Bayes based prior for ϕ with $c_0 = 10$ as in Model III (see the prior construction for our proposed hierarchical model in Section 2.2.1). In order to determine the burn-in point and the size of the Markov chain for the inferential purposes, we created three chains starting from three dispersed sets of initial points and used the convergence diagnostic tools available in WinBUGS. Accordingly, Brooks-Gelman-Rubin's convergence diagnostics indicate that about first 1000 iterations for Models I&II and about 6000 iterations for Model III should be burnt (i.e. left out of the Bayesian inference). The Brooks-Gelman-

Rubin's convergence diagnostics for the proposed hierarchical model (Model III) are given in Appendix D.2. Also, the size of the Markov chain is set to be 4000 iterations for Models I&II, and 12000 iterations for Model III so that Monte Carlo standard errors of the means delivered by WinBUGS are all about the same magnitude. To alleviate the autocorrelation observed in some parameters, we selected every 3rd and 5th iterations for Models I&II and Model III respectively. The computation time for Model III (e.g. with $D = 5$) was approximately 145 seconds while estimation for Models I&II took approximately 15 seconds on a PC (2.0 GB RAM, 2.80 GHz Pentium Dual-Core GPU).

3.3 Results

To evaluate the efficiency of our proposed semiparametric missing data model in terms of estimating the parameter of our main interest namely β_1, simulation based bias and mean squared error (MSE) estimates of each parameter in the logistic regression are calculated. Besides the bias and MSE we also include simulation based standard error (SE) estimates of the parameters in the following tables as it is useful to see how MSE is decomposed between bias and SE.

Table 3.1: Simulation Results for TMM 0

			Bias (SE) MSE		
Missing Perc.(ϕ_0)	Model	# of knots (D)	$\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = -1$
15%(-2.55)	I	-	-0.105 (0.430) 0.196	-0.062 (0.223) 0.054	0.081 (0.195) 0.044
	II	-	-0.114 (0.428) 0.196	-0.068 (0.220) 0.053	0.083 (0.197) 0.045
	III	2	-0.108 (0.413) 0.198	-0.064 (0.222) 0.053	0.076 (0.197) 0.044
		3	-0.124 (0.427) 0.198	-0.074 (0.219) 0.053	0.076 (0.194) 0.043
		4	-0.131 (0.426) 0.199	-0.079 (0.219) 0.054	0.078 (0.195) 0.044
		5	-0.136 (0.418) 0.193	-0.082 (0.215) 0.053	0.081 (0.190) 0.043
		6	-0.139 (0.421) 0.197	-0.084 (0.217) 0.054	0.079 (0.196) 0.044
		7	-0.136 (0.422) 1.196	-0.084 (0.216) 0.053	0.078 (0.194) 0.044
		8	-0.139 (0.422) 1.197	-0.084 (0.217) 0.054	0.079 (0.196) 0.047
35%(-1.25)	I	-	-0.154 (0.495) 0.269	-0.134 (0.283) 0.098	0.153 (0.221) 0.072
	II	-	-0.175 (0.505) 0.285	-0.149 (0.289) 0.105	0.158 (0.221) 0.074
	III	2	-0.223 (0.506) 0.305	-0.179 (0.305) 0.125	0.193 (0.215) 0.084
		3	-0.303 (0.527) 0.369	-0.230 (0.308) 0.148	0.224 (0.222) 0.099
		4	-0.243 (0.535) 0.346	-0.192 (0.306) 0.131	0.187 (0.243) 0.094
		5	-0.292 (0.518) 0.353	-0.222 (0.297) 0.137	0.210 (0.241) 0.102
		6	-0.293 (0.463) 0.301	-0.221 (0.273) 0.124	0.211 (0.227) 0.096
		7	-0.325 (0.495) 0.351	-0.247 (0.281) 0.140	0.230 (0.228) 0.105
		8	-0.306 (0.520) 0.365	-0.232 (0.307) 0.148	0.216 (0.230) 0.099

The bias estimate for the jth component of β is calculated by $\text{Bias}_j = \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_j^{(i)} - \beta_j$ where $\hat{\beta}_j^{(i)}$ is the estimate of the jth β parameter based on the ith simulated sample. Simulation based SE estimate of the β_j is then calculated by $\text{SE}_j = \left(\frac{1}{N-1}\sum_{i=1}^{N}(\hat{\beta}_j^{(i)} - \bar{\beta}_{.j})^2\right)^{1/2}$ where $\bar{\beta}_{.j}$ is the average of the estimates of the $\hat{\beta}_j$s over N simulations. Then the MSE is defined as $\text{MSE}_j = \text{Bias}_j^2 + \text{SE}_j^2$. The sample `Result.m` function to calculate the simulation based Bias, SE and MSE estimates of the parameters is given in Appendix C.

Based on Table 3.1 we can evaluate how risky it is to use our proposed semiparametric model for missingness when the true missingness construction has linear missing covariate effect. For 15% missingness, Model III especially with small number of knots seems to be competing well with Models I&II in terms of bias and variation. If we increase the missingness amount to 35%, biases of the estimates of the parameters β_0, β_1 and β_2 increase for Model III. We think that this increase is due to the fact that true missingness mechanism is indeed linear and our proposed semiparametric model does not fit as well as parametric model(linear) when the missing percentage is high. If there is strong belief about a linearity on the NMAR covariates in the missingness mechanism, it is better to use linear missingness model for severe missing percentages, however our proposed semiparametric model can be equally preferable when the missingness amount is around 15%.

Table 3.2: Simulation Results for TMM I

Missing Perc.(ϕ_0)	Model	# of knots (D)	Bias (SE) MSE $\beta_0 = 2$	 $\beta_1 = 1$	 $\beta_2 = -1$
15%(-10.45)	I	-	-0.133 (0.412) 0.188	0.102 (0.266) 0.081	0.062 (0.189) 0.039
	II	-	-0.091 (0.435) 0.198	0.112 (0.269) 0.085	0.040 (0.196) 0.040
	III	2	-0.045 (0.421) 0.179	0.121 (0.258) 0.081	0.025 (0.192) 0.037
		3	-0.034 (0.426) 0.183	0.112 (0.253) 0.076	0.010 (0.198) 0.039
		4	-0.036 (0.428) 0.185	0.113 (0.252) 0.076	0.015 (0.197) 0.039
		5	-0.033 (0.425) 0.182	0.105 (0.248) 0.072	0.016 (0.198) 0.039
		6	-0.040 (0.421) 0.178	0.098 (0.240) 0.067	0.021 (0.196) 0.039
		7	-0.034 (0.426) 1.182	0.090 (0.242) 0.067	0.020 (0.197) 0.039
		8	-0.038 (0.420) 1.178	0.091 (0.243) 0.067	0.022 (0.196) 0.039
35%(-5.90)	I	-	-0.358 (0.456) 0.336	0.212 (0.390) 0.197	0.142 (0.202) 0.061
	II	-	-0.253 (0.470) 0.285	0.248 (0.386) 0.211	0.076 (0.211) 0.050
	III	2	-0.238 (0.474) 0.282	0.249 (0.385) 0.211	0.097 (0.210) 0.053
		3	-0.090 (0.469) 0.228	0.242 (0.359) 0.187	0.050 (0.205) 0.044
		4	-0.017 (0.493) 0.244	0.260 (0.369) 0.204	-0.003 (0.221) 0.049
		5	-0.014 (0.486) 0.237	0.238 (0.371) 0.194	-0.012 (0.219) 0.048
		6	-0.022 (0.496) 0.247	0.238 (0.357) 0.184	-0.034 (0.226) 0.052
		7	-0.017 (0.509) 0.260	0.220 (0.361) 0.179	-0.038 (0.232) 0.055
		8	-0.026 (0.495) 0.246	0.211 (0.357) 0.172	-0.045 (0.228) 0.054

In Table 3.2 and Table 3.3, we can see the performance of our proposed model when the true missingness model is non-linear. We define two non-linear true missingness schemes (TMM I&II). We observe almost similar results in both tables regarding bias, simulation based SE, and MSE. For 15% missingness, the larger the number of knots used for approximating the true smooth function, the smaller the bias Model III has. In terms of the SE, our proposed semiparametric model gives slightly better results for all parameters. When we merge biases and SE estimates, the resulting MSE for all parameters based on Model III are smaller than those for Models I&II, which means that our proposed semiparametric model (with number of knots from 5 to 8) is more efficient than the parametric models. When we increase the missing percentage to 35%, biases of the estimates of all parameters (especially parameter of our main interest β_1) in Model III with larger number of knots are smaller than those for Models I&II. In addition, SE of $\hat{\beta}_1$ decreases for Model III. In terms of MSE of all parameters, Model III (with number of knots from 6 to 8) gives better results than Models I&II for 35% missingness. This implies that one is better of fitting Model III containing higher number of knots when the percentage of missingness is relatively large.

Table 3.3: Simulation Results for TMM II

Missing Perc.(ϕ_0)	Model	# of knots (D)	Bias (SE) MSE $\beta_0 = 2$	$\beta_1 = 1$	$\beta_2 = -1$
15%(-2.75)	I	-	-0.003 (0.491) 0.241	0.040 (0.281) 0.080	0.068 (0.209) 0.048
	II	-	0.047 (0.492) 0.244	0.061 (0.282) 0.083	0.029 (0.217) 0.048
	III	2	0.026 (0.505) 0.256	0.055 (0.289) 0.087	0.041 (0.220) 0.050
		3	0.024 (0.486) 0.237	0.051 (0.280) 0.081	0.018 (0.223) 0.050
		4	0.016 (0.481) 0.231	0.033 (0.280) 0.080	0.016 (0.219) 0.048
		5	0.003 (0.479) 0.230	0.022 (0.278) 0.078	0.018 (0.222) 0.049
		6	-0.003 (0.477) 0.228	0.016 (0.278) 0.077	0.016 (0.220) 0.048
		7	-0.001 (0.477) 0.227	0.017 (0.276) 0.076	0.014 (0.220) 0.049
		8	-0.013 (0.480) 0.230	0.013 (0.277) 0.077	0.015 (0.223) 0.050
35%(-0.95)	I	-	0.013 (0.600) 0.360	0.150 (0.416) 0.195	0.175 (0.206) 0.073
	II	-	0.218 (0.621) 0.434	0.250 (0.412) 0.233	0.046 (0.233) 0.056
	III	2	0.070 (0.615) 0.384	0.181 (0.421) 0.210	0.129 (0.218) 0.064
		3	0.175 (0.614) 0.408	0.238 (0.402) 0.218	0.027 (0.239) 0.058
		4	0.217 (0.607) 0.415	0.187 (0.383) 0.182	-0.046 (0.252) 0.066
		5	0.190 (0.607) 0.404	0.162 (0.379) 0.170	-0.053 (0.256) 0.068
		6	0.174 (0.606) 0.398	0.146 (0.377) 0.164	-0.058 (0.261) 0.071
		7	0.161 (0.601) 0.387	0.137 (0.376) 0.160	-0.062 (0.257) 0.070
		8	0.166 (0.586) 0.383	0.140 (0.371) 0.157	-0.065 (0.257) 0.070

CHAPTER 4

CONCLUSION

To account for the underlying nature of the missingness, the missing data mechanism of the *not missing at random* (NMAR) variables is needed to be considered in the analysis of missing data problems in order to produce as little bias as possible on statistical inference. As a common practical approach, the missing data mechanism for NMAR covariates in the generalized linear models (GLMs) is modeled by parametric approaches (e.g. the missingness probabilities of the nonignorably missing response or covariates are modeled by logistic or probit regression) in the missing data literature. However, the probability of covariate being missing may not be linearly related with missing covariate itself. In this study, we modeled the missing data mechanism using generalized additive model (GAM) approach in which we included an unspecified smooth function representing the effect of the missing covariate. We used penalized spline regression to approximate the unspecified smooth function and conveniently write the likelihood function. The accompanying semiparametric model for the missingness mechanism leads to capturing any possible nonlinear functional relationship between the missingness probability of the covariate and the missing covariate itself in a more flexible way.

In Section 2.1, we provided the usage of low-rank thin-plate splines to approximate the unspecified smooth functions in the proposed semiparametric missing data mechanism. Spline regression is based on regressing the response on the covariate piecewise. The pieces are determined by what is called knots. The knots can be obtained as the quantiles of the covariate which is nonlinearly associated with the response. We developed an iterative algorithm to suitably determine the knots in the presence of missing data. The algorithm is based on multiple imputation of the missing covariate. We also proposed the use of empirical Bayes based priors for the parameters of the missing data model (see Section 2.2.1). The empirical

Bayes based priors result in accelerated convergence in the corresponding posterior calculations. In Section 2.2.2, the steps of Gibbs sampling algorithm are constructed to efficiently estimate the parameters by using fully Bayesian approach. The WinBUGS code, provided in Section 2.2.3, is also useful for computationally generating the posterior inference based on the proposed hierarchical model discussed.

In Chapter 3, we provided a detailed analysis on the performance of penalized spline regression for the missingness models in logistic regression with NMAR covariates. The analyses implied that if there is strong belief about a linearity on the NMAR covariates in the missing data model and the missingness percentage is high, one may use the parametric approach with fully linear model. However, if missingness percentage is low, one can equally use a fully linear model or the proposed penalized spline regression for the missingness model. On the other hand if there is belief about nonlinearity on NMAR covariates in the missingness mechanism, the proposed semiparametric approach for the missing data model seems to provide a more effective method for estimating the association between the response and the NMAR covariates regardless of the missing percentage. Overall, simulation results imply that the method with semiparametric missing data model provides more reliable β estimates compared to the method with parametric missing data model if there is belief about a nonlinearity on the NMAR covariates in the missingness mechanism.

In the simulation study, we also observed that the performance of penalized spline regression highly depends on the number of knots especially if there is nonlinear relationship between the missingness probability and the missing covariate. When the small number of knots is used in the spline regression, the true nonlinear functional relationship may not be detected by the unspecified smooth functions. However, there may exist the issue of identifiability for modeling the missingness status when the number of knots is large since it is well known that the parameters of the missing data model can be unestimable due to the insufficient information contained in the dataset regarding the missingness model parameters. The number of knots, therefore, should be determined with great care so as not to end up with identifiability problems. As a future research topic, a specific knot determination method may be developed for the proposed semiparametric missing data model. Alternatively, one can tune up the knot number by trying out several different ones before settling on the final one. The deviance information criterion (DIC) of Spiegelhalter et al. [7] may be helpful for making a choice among penalized spline regression models with different knots.

In this thesis, it is assumed that the response variable is fully observed and NMAR covariates are continuous in the logistic regression. However, the proposed semiparametric missing data model can also be adopted for the generalized linear models with NMAR response as a future study. In addition, the semiparametric approach for modeling the missingness can be extended to other complex model settings, including generalized mixed models, longitudinal data models and survival models. Also, it can be extended to missing data that occur in designed experiments.

REFERENCES

[1] Little, R. J. A and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

[2] Ruppert, D., Wand, M. P., Carroll R. J. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.

[3] Ibrahim, J. G. (1990) Incomplete data in generalized linear models. *Journal of the American Statistical Association*, **85**(411), 765–769.

[4] Ibrahim, J. G. and Lipsitz, S. R. (1996) Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics*, **52**, 1071–1078.

[5] Ibrahim, J. G., Lipsitz, S. R. and Chen M. H. (1999) Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society B*, **61**(1), 173–190.

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

[7] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, **64**(4), 583–639.

[8] Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring, A. H. (2005) Missing-Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association*, **100**(469), 469, 32–346.

[9] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994) Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, **89**(427), 849–866.

[10] Sinha, S. K. (2008) Robust methods for generalized linear models with nonignorable missing covariates. *The Canadian Journal of Statistics*, **36**(2), 277–299.

[11] Chen, Q., Ibrahim, J. G. (2006) Semiparametric Models for Missing Covariate and Response Data in Regression Models. *Biometrics*, **62**, 177–184.

[12] Spiegelhalter, D., Thomas, A., Best, N. (2003) WinBUGS Version 1.4 User Manual. *Medical Research Council Biostatistics Unit*, Cambridge, UK.

[13] Lipsitz, S. R. and Ibrahim, J. G. (1996) A conditional model for incomplete covariates in parametric regression models. *Biometrika*, **83**(4), 916–922.

[14] Crainiceanu, C. M., Ruppert, D. and Wand, M.P. (2005) Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, **14**(14).

[15] Brumback, B., Ruppert, D. and Wand, M. P. (1999) Comment on Variable Selection and Function Estimation in Additive Nonparametric Regression Using Data-based Prior by Shively, Kohn and Wood. *Journal of the American Statistical Association*, **94**(447), 794–797.

[16] Gelman, A. and Rubin, D. (1996) Markov chain Monte Carlo methods in biostatistics. *Statistical Methods in Medical Research*, **5**(4), 339–355.

[17] Gelfand, A. and Smith, A. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**(410), 398–409.

[18] Gilks, W., Richardson, S. and Spiegelhalter, D. (1996a) *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

[19] Casella, G. and George, E. (1992) Explaining the Gibbs sampler. *American Statistician*, **46**(3), 167–174.

[20] Taner, M. A. and Wong, W. H. (1987) The calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**(398), 528–540.

[21] Carlin, B. P. and Louis, T. A. (1997) *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman and Hall.

[22] Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**(3), 515–533.

[23] Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**(422), 669–679.

[24] Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000) WinBUGS-A Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

[25] Gilks, W. R. and Wild P. (1992) Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, **41**(2), 337–348.

[26] Gilks, W. R. (1992) Derivative-free Adaptive Rejection Sampling for Gibbs Sampling. *Bayesian Statistics 4* (eds. Bernardo, J., Berger, J., Dawid, A. P. and Smith, A. F. M.), Oxford University Press, 641–649.

[27] Neal, R. M. (2003) Slice Sampling. *The Annals of Statistics*, **31**(3), 705–741.

[28] Metropolis, N., Rosenbluth, A. W., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

[29] Hastings, W. K. (1970) Monte Carlo sampling-based methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

[30] Brooks, S. P. and Gelman, A. (1998) General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **7**(4), 434–455.

APPENDIX A

The Functional Forms of Full Conditional Distributions

The full conditional distribution of each unknown quantity, namely β, α, σ_x^{-2}, ϕ, σ_b^{-2}, $x_{(1)}^{miss}$, b and W for the following special case of proposed hierarchical model is derived in this appendix.

1. $$f(y_i|x_{i1},x_{i2},\beta) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} \ ; \quad \mu_i = E(y_i|x_{i1},x_{i2},\beta) = P(y_i = 1|x_{i1},x_{i2},\beta)$$

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \tag{A.1}$$

2. $$x_{i1}|\, x_{i2} \sim N(\alpha_0 + \alpha_1 x_{i2},\ \sigma_x^2)$$

$$\tag{A.2}$$

3. $$f(r_i|W_i) = [\Phi(W_i)]^{r_i}\,[1-\Phi(W_i)]^{1-r_i}$$

$$W_i\ |(y_i,x_{i1},x_{i2},Z,\phi,b) \sim N(\mu_{W_i},1) \quad \text{and} \quad \mu_{W_i} = \phi_0 + \phi_1 y_i + \phi_2 x_{i1} + Zb + \phi_3 x_{i2}$$

$$b|\ \sigma_{b(k)}^2 \sim \text{MN}(0,\sigma_b^2 I_D) \tag{A.3}$$

where x_{i1} is a covariate subject to missingness and x_{i2} is a fully observed covariate, and y_i is a fully observed binary response for $i = 1,...n$. The variable r_i is a missing data indicator for the covariate x_{i1}. The variable $b = (b_1,...,b_D)^T$ is a $D \times 1$ vector of random effects where D is the number of knots. The matrix Z is a $n \times D$ design matrix for the random effects where $Z_i = (Z_{i1},...,Z_{iD})$. The cumulative distribution of standard normal distribution for the probit link function is denoted by Φ. Let x_{i1}^{miss} and x_{i1}^{obs} denote missing and observed components of x_{i1}, and let first m subjects are missing and remaining $n-m$ subjects are observed for x_{i1}.

Then the joint conditional distribution of all the unknown quantities conditional on the observed data $D_{obs} = (r, y, x_1^{obs}, x_2)$ is written as

$$\begin{aligned} f(\beta, \alpha, \sigma_x^{-2}, \phi, \sigma_b^{-2}, x_1^{miss}, b, W | r, y, x_1^{obs}, x_2) =& \frac{f(D_{com} |\ \beta, \alpha, \sigma_x^{-2}, \phi, \sigma_b^{-2}) f(\beta, \alpha, \sigma_x^{-2}, \phi, \sigma_b^{-2})}{f(D_{obs})} \\ \propto& f(D_{com} |\ \beta, \alpha_{(1)}^*, \sigma_X^{-2}, \phi, \sigma_b^{-2}) \times \\ & f(\beta, \alpha_{(1)}^*, \sigma_X^{-2}, \phi, \sigma_b^{-2}) \end{aligned} \tag{A.4}$$

where $D_{com} = (r, y, x_1^{miss}, x_1^{obs}, x_2, b, w)$ is the complete data set and $f(D_{com} |\ \beta, \alpha_{(1)}^*, \sigma_x^{-2}, \phi, \sigma_b^{-2})$ is the likelihood based on the complete data. The complete data likelihood for the special case of the proposed hierarchical model can be written as

$$\begin{aligned} f(D_{com} |\ \beta, \alpha, \sigma_x^{-2}, \phi, \sigma_b^{-2}) =& f(r, y, x_1^{miss}, x_1^{obs}, x_2, b, W |\ \beta, \alpha, \sigma_x^{-2}, \phi, \sigma_b^{-2}) \\ =& f(r |\ W) f(W |\ y, x_1^{miss}, x_1^{obs}, x_2, \phi, b) f(y |\ x_1^{miss}, x_1^{obs}, x_2, \beta) \times \\ & f(x_1^{miss}, x_1^{obs} |\ x_2, \alpha, \sigma_x^{-2}) f(b |\ \sigma_b^{-2}) \end{aligned} \tag{A.5}$$

Note that the design matrix Z for the random effects are not included in D_{obs} since Z is the function of (x_1^{miss}, x_1^{obs}). Then the joint conditional distribution in (A.4) becomes

$$\begin{aligned} f(\beta, \alpha, \sigma_x^{-2}, \phi, \sigma_b^{-2}, x_1^{miss}, b, W | r, y, x_1^{obs}, x_2) \propto& f(r |\ W) f(W |\ y, x_1^{miss}, x_1^{obs}, x_2, \phi, b) \times \\ & f(y |\ x_1^{miss}, x_1^{obs}, x_2, \beta) f(x_1^{miss}, x_1^{obs} |\ x_2, \alpha, \sigma_x^{-2}) \times \\ & f(b |\ \sigma_b^{-2}) f(\beta, \alpha_{(1)}^*, \sigma_X^2, \phi, \sigma_b^{-2}) \end{aligned} \tag{A.6}$$

Let's first construct each component of (A.6) separately as follows

1.

$$f(r|W) = \prod_{i=1}^{n} f(r_i|w_i) = \prod_{i=1}^{n} [\Phi(w_i)]^{r_i} [1 - \Phi(w_i)]^{1-r_i} \tag{A.7}$$

2.

$$\begin{aligned} f(W |\ y, x_1^{miss}, x_1^{obs}, x_2, \phi, b) =& \prod_{i=1}^{m} f(W_i |\ y_i, x_{i1}^{miss}, x_{i2}, \phi, b) \prod_{i=m+1}^{n} f(W_i |\ y_i, x_{i1}^{obs}, x_{i2}, \phi, b) \\ =& \prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi}} e^{\left\{ -\frac{1}{2} \left(W_i - \phi_0 - \phi_1 y_i - \phi_2 x_{i1}^{miss} - Z_i b - \phi_3 x_{i2} \right)^2 \right\}} \right] \times \\ & \prod_{i=m+1}^{n} \left[\frac{1}{\sqrt{2\pi}} e^{\left\{ -\frac{1}{2} \left(W_i - \phi_0 - \phi_1 y_i - \phi_2 x_{i1}^{obs} - Z_i b - \phi_3 x_{i2} \right)^2 \right\}} \right] \end{aligned} \tag{A.8}$$

3.

$$
\begin{aligned}
f(y|\, x_1^{miss}, x_1^{obs}, x_2, \beta) &= \prod_{i=1}^{m} f(y_i|\, x_{i1}^{miss}, x_{i2}, \beta) \prod_{i=m+1}^{n} f(y_i|\, x_{i1}^{obs}, x_{i2}, \beta) \\
&= \prod_{i=1}^{m} \mu_{*i}^{y_i} (1-\mu_{*i})^{1-y_i} \prod_{i=m+1}^{n} \mu_i^{y_i} (1-\mu_i)^{1-y_i}
\end{aligned}
$$

$$
\text{where} \quad \mu_{*i} = \frac{1}{1+e^{-\left\{\beta_0+\beta_1 x_{i1}^{miss}+\beta_2 x_{i2}\right\}}} \quad \text{and} \quad \mu_i = \frac{1}{1+e^{-\left\{\beta_0+\beta_1 x_{i1}^{obs}+\beta_2 x_{i2}\right\}}} \tag{A.9}
$$

4.

$$
\begin{aligned}
f(x_1^{miss}, x_1^{obs}|\, x_2, \alpha, \sigma_x^{-2}) &= \prod_{i=1}^{m} f(x_{i1}^{miss}|\, x_{i2}, \alpha, \sigma_x^{-2}) \prod_{i=m+1}^{n} f(x_{i1}^{obs}|\, x_{i2}, \alpha, \sigma_x^{-2}) \\
&= \prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi}\sigma_x} e^{\left\{ -\frac{1}{2\sigma_x^2} \left(x_{i1}^{miss} - \alpha_0 - \alpha_1 x_{i2} \right)^2 \right\}} \right] \times \\
&\quad \prod_{i=m+1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma_x} e^{\left\{ -\frac{1}{2\sigma_x^2} \left(x_{i1}^{obs} - \alpha_0 - \alpha_1 x_{i2} \right)^2 \right\}} \right]
\end{aligned} \tag{A.10}
$$

5.

$$
f(b|\, \sigma_b^{-2}) = \prod_{j=1}^{D} f(b_j|\sigma_b^{-2}) = \prod_{j=1}^{D} \left[\frac{1}{\sqrt{2\pi}\sigma_b} e^{\left\{ -\frac{1}{2\sigma_b^2} b_j^2 \right\}} \right] \tag{A.11}
$$

6.

$$
\begin{aligned}
f(\beta, \alpha, \sigma_x^{-2}, \phi, \sigma_b^{-2}) =& f(\beta) f(\alpha) f(\sigma_x^2) f(\phi) f(\sigma_b^{-2}) \\
f(\beta) =& 1 \quad -\infty < \beta < \infty \\
f(\alpha) =& 1 \quad -\infty < \alpha < \infty \\
f(\sigma_x^{-2}) =& \frac{\mu^r}{\Gamma(r)} (\sigma_x^{-2})^{r-1} e^{-\mu \sigma_x^{-2}} \quad \text{where} \quad r = 0.01, \mu = 0.01, \sigma_x^{-2} > 0 \\
f(\phi) =& \frac{1}{(2\pi)^{4/2} |\hat{\Sigma}_\phi|^{1/2}} e^{-\frac{1}{2}(\phi-\hat{\mu}_\phi)' \hat{\Sigma}_\phi^{-1} (\phi-\hat{\mu}_\phi)} \quad -\infty < \phi < \infty \\
f(\sigma_b^{-2}) =& \frac{\mu^r}{\Gamma(r)} (\sigma_b^{-2})^{r-1} e^{-\mu \sigma_b^{-2}} \quad \text{where} \quad r = \hat{\tau}_{\sigma_{b^2}}, \mu = \hat{\varphi}_{\sigma_{b^2}}, \sigma_b^{-2} > 0
\end{aligned} \tag{A.12}
$$

Then the functional form of full conditional distribution of each unknown quantity (e.g parameters and latent variables) can be obtained by extracting the terms corresponding to the parameters/ latent variable of specific interest:

1.

$$f(x_{i1}^{miss}|\,\beta,\alpha,\sigma_x^{-2},\phi,b,W,x_{i2}) \propto \left[\frac{1}{\sqrt{2\pi}}e^{\left\{-\frac{1}{2}\left(W_i-\phi_0-\phi_1 y_i-\phi_2 x_{i1}^{miss}-Z_i b-\phi_3 x_{i2}\right)^2\right\}}\right]\times$$
$$\left[\mu_{*i}^{y_i}(1-\mu_{*i})^{1-y_i}\right]\left[\frac{1}{\sqrt{2\pi}\sigma_x}e^{\left\{-\frac{1}{2\sigma_x^2}\left(x_{i1}^{miss}-\alpha_0-\alpha_1 x_{i2}\right)^2\right\}}\right] \tag{A.13}$$

which is the nonlinear function of x_{i1}^{miss} consisting of normal and logistic kernels. The Metropolis Hastings algorithm (with normal proposal distribution) is performed to obtain the random draws from $f(x_{i1}^{miss}|\,\beta,\alpha,\sigma_x^{-2},\phi,b,W,x_{i2})$ for $i=1,...,m$. Note that the components x_{i1}^{miss} and x_{i1}^{obs} are combined as x_{i1} for the other full conditional distributions for simplicity.

2.

$$f(b_j|\,\phi,\sigma_b^{-2},W,y,x_1,x_2) \propto \left[\frac{1}{\sqrt{2\pi}\sigma_b}e^{\left\{-\frac{1}{2\sigma_b^2}b_j^2\right\}}\right]\times$$
$$\prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}}e^{\left\{-\frac{1}{2}\left(W_i-\phi_0-\phi_1 y_i-\phi_2 x_{i1}-\sum_{k=1}^{D}Z_{ik}b_k-\phi_3 x_{i2}\right)^2\right\}}\right] \tag{A.14}$$

which consists of only normal kernels. The direct sampling algorithm (from normal distribution) is performed to obtain the random draws from $f(b_j|\,\phi,\sigma_b^{-2},W,y,x_1,x_2)$ for $j=1,...,D$.

3.

$$f(W_i|\,\phi,b,r_i,y_i,x_{1i},x_{2i}) \propto [\Phi(W_i)]^{r_i}\,[1-\Phi(W_i)]^{1-r_i}\left[\frac{1}{\sqrt{2\pi}}e^{\left\{-\frac{1}{2}(W_i-\phi_0-\phi_1 y_i-\phi_2 x_{i1}-Z_i b-\phi_3 x_{i2})^2\right\}}\right] \tag{A.15}$$

In Winbugs, W_i's are defined by truncated normal distributions by the data augmentation approach proposed by [23] where

$$\begin{aligned} W_i|\;(\phi,b,r_i,y_i,x_{1i},x_{2i}) \sim N\,[\phi_0+\phi_1 y_i+\phi_2 x_{i1}+Z_i b+\phi_3 x_{i2},\ 1]& \\ \text{truncated at the left by 0}\quad \text{if } r_i = 1& \\ \text{truncated at the right by 0}\quad \text{if } r_i = 0& \end{aligned} \tag{A.16}$$

which consists of only truncated normal kernel. The Derivative Free Adaptive Rejection sampling algorithm is performed to obtain the random draws from $f(W_i|\,\phi,b,r_i,y_i,x_{1i},x_{2i})$ for $i=1,...,n$.

4.

$$f(\beta_j|\, y,x_1,x_2) \propto \prod_{i=1}^{n} \mu_i^{y_i}(1-\mu_i)^{1-y_i} \text{ where } \mu_i = \frac{1}{1+e^{-\{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}\}}} \tag{A.17}$$

which consists of only logistic kernels. The Slice sampling algorithm is performed to obtain the random draws from $f(\beta_j|\, y,x_1,x_2)$ for $j = 0,1,2$.

5.

$$f(\alpha_j|\, \sigma_x^{-2},x_1,x_2) \propto \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma_x} e^{\left\{-\frac{1}{2\sigma_x^2}(x_{i1}-\alpha_0-\alpha_1 x_{i2})^2\right\}} \right] \tag{A.18}$$

which consists of only normal kernels. The direct sampling algorithm (from normal distribution) is performed to obtain the random draws from $f(\alpha_j|\, \sigma_x^{-2},x_1,x_2)$ for $j = 1,2$.

6.

$$f(\phi|\, b,W,y,x_1,x_2) \propto \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}} e^{\left\{-\frac{1}{2}(W_i-\phi_0-\phi_1 y_i-\phi_2 x_{i1}-Z_i b-\phi_3 x_{i2})^2\right\}} \right] \times$$
$$\frac{1}{(2\pi)^{4/2}|\hat{\Sigma}_\phi|^{1/2}} e^{-\frac{1}{2}(\phi-\hat{\mu}_\phi)'\hat{\Sigma}_\phi^{-1}(\phi-\hat{\mu}_\phi)} \tag{A.19}$$

which consists of only multivariate normal kernel. The direct sampling algorithm (from multivariate normal distribution) is performed to obtain the random draws from $f(\phi|\, b,W,y,x_1,x_2)$.

7.

$$f(\sigma_x^{-2}|\, \alpha,x_1,x_2) \propto \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma_x} e^{\left\{-\frac{1}{2\sigma_x^2}(x_{i1}-\alpha_0-\alpha_1 x_{i2})^2\right\}} \right] \left[\frac{\mu^r}{\Gamma(r)} (\sigma_x^{-2})^{r-1} e^{-\mu\sigma_x^{-2}} \right] \tag{A.20}$$

which consists of only normal and gamma kernels. Since the gamma distribution is conjugate prior for the precision parameter σ_x^{-2}, the direct sampling algorithm (from gamma distribution) is performed to obtain the random draws from $f(\sigma_x^{-2}|\, \alpha,x_1,x_2)$.

8.

$$f(\sigma_b^{-2}|\, b) \propto \prod_{j=1}^{D} \left[\frac{1}{\sqrt{2\pi}\sigma_b} e^{\left\{-\frac{1}{2\sigma_b^2} b_j^2\right\}} \right] \left[\frac{\mu^r}{\Gamma(r)} (\sigma_b^{-2})^{r-1} e^{-\mu\sigma_b^{-2}} \right] \tag{A.21}$$

which consists of only normal and gamma kernels. Since the gamma distribution is conjugate prior for the precision parameter σ_b^{-2}, the direct sampling algorithm (from gamma distribution) is performed to obtain the random draws from $f(\sigma_b^{-2}|\, b)$.

APPENDIX B

MCMC Sampling Algorithms Used by WinBUGS

In WinBUGS, the appropriate sampling method for the specified full conditional distribution is chosen by the hierarchy given in Table B.1. For instance, if the the full conditional distribution has a closed form (e.g. if the conjugate prior is used for the parameter) then direct sampling algorithm is used to obtain the random draws, otherwise one of the following sampling methods is used in the following hierarchy (see Lunn et al.[24]).

Table B.1: The hierarchy used by WinBUGS for the sampling methods

Full Conditional Distribution	Sampling Method
Closed form (conjugate)	Direct sampling using standard algorithms
Log-concave	Derivative-free adaptive rejection sampling
Non-Log-concave (restricted range)	Slice sampling
Non-Log-concave (unrestricted range)	Metropolis-Hastings method

B.1 Derivative-Free Adaptive Rejection Sampling Method

The standard adaptive rejection sampling algorithm proposed by Gilks and Wild [25] can be used to sample from the log-concave target distribution $f(x)$ where $f(x)$ is log-concave function if $\frac{d^2}{dx^2}logf(x)$ is negative for all x in the domain. In this algorithm, the derivative of $logf(x)$ is needed to be calculated at each iteration which is computationally difficult for some cases. The derivative-free adaptive rejection method is then proposed by Gilks [26] to avoid these derivative calculations. Let $h(x) = logf(x)$ such that $h(x)$ is concave everywhere in the domain. To obtain a random number from $f(x)$, the following procedure can be considered.

Initialization Step

1. Choose at least three points $S = \{x_1 \leq x_2 \leq ... \leq x_{k-1} \leq x_k\}$ such that at least one of them lies to each side of the mode of $f(x)$. The points are on either side of the mode if chord $P_{x_1}P_{x_2}$ has a positive slope and chord $P_{x_{k-1}}P_{x_k}$ has a negative slope.

2. Construct the piece-wise linear lower bound $u(x)$ to $h(x)$ from the chords $P_0P_{x_1}$ and $P_{x_k}P_1$.

3. Construct the piece-wise linear upper bound $l(x)$ to $h(x)$ from the chords $P_{x_1}P_{x_2}$ and $P_{x_{k-1}}P_{x_k}$.

Sampling Step

4. Construct the envelope function $e(x) = e^{u(x)}$.

5. Sample X from $e(x)$.

6. Sample U from $U(0,1)$.

Squeezing Step

7. Construct the squeezing function $s(x) = e^{l(x)}$.

8. If $U \leq \frac{s(X)}{e(X)}$ then accept X as a sample from $f(x)$.

9. else go to the rejection step.

Rejection Step

10. If $U \leq \frac{f(X)}{e(X)}$ then accept X as a sample from $f(x)$.

11. else go to the updating step.

Updating Step

12. Add rejected point X to the set of points S to make use of the information gained in rejection step, and go back to the initialization step.

Repeat this algorithm until one X is accepted.

B.2 Slice Sampling Method

The slice sampling method proposed by Neal [27] is a type of MCMC algorithm to obtain the random draws from a target distribution $f(x)$. The advantage of this method is that it does not require any proposal distribution. Let $A = \{(x,u) : 0 \leq u \leq f(x)\}$ be the area under the graph $f(x)$, and let (x_0, u_0) be the starting point in A. Then at the jth iteration of this algorithm, the following procedure is considered to obtain the sequences of random numbers from $f(x)$ until we reach the desired sample size.

1. Sample u_j from $U(0, f(x_{j-1}))$.
2. Sample x_j from $U(\{x : f(x) \geq u_j\})$.

B.3 Metropolis-Hastings Sampling Method

The Metropolis-Hastings sampling method proposed by Metropolis et al. [28] and Hastings [29] is a type of MCMC algorithm to obtain a sequences of random samples from a target distribution $f(x)$. At the $(t+1)$th iteration, state of the chain X_{t+1} is obtained by sampling a candidate point Y from the proposal distribution $q(.|X_t)$ where the proposal distribution q should be easy to sample from and the point Y depends on the previous state X_t. The candidate point is accepted as X_{t+1} with the following probability; otherwise the state X_{t+1} remains same as X_t.

$$\alpha(X_t, Y) = \min\left(1, \frac{f(Y)q(X_t|Y)}{f(X_t)q(Y|X_t)}\right) \tag{B.1}$$

Then the following procedure is considered at the $(t+1)$the iteration to obtain the sequences of random numbers from $f(x)$ until we reach the desired sample size.

1. Sample a candidate point Y from the proposal distribution $q(.|X_t)$.
2. Sample U from $U(0,1)$.
3. If $U \leq \alpha(X_t, Y)$ then accept Y as X_{t+1} else set $X_{t+1} = X_t$.

APPENDIX C

MATLAB Codes Used in the Simulation Study

Sample *Simulation.m* Function

```
% The sample Simulation.m function runs the simulation study for the true missingness
% mechanism II and 15% missing percentage
function Simulation(mcsize,n,choose,take_data)

% Input Arguments:
% mcsize    :number of Monte Carlo iterations
% n         :sample size
% choose    :
%             0 ---> run all models
%             1 ---> run model I(parametric missing data model without interaction)
%             2 ---> run model II(parametric missing model model with interaction)
%             3 ---> run model III(semiparametric missing data model with 2 knots)
%             4 ---> run model III(semiparametric missing data model with 3 knots)
%             5 ---> run model III(semiparametric missing data model with 4 knots)
%             6 ---> run model III(semiparametric missing data model with 5 knots)
%             7 ---> run model III(semiparametric missing data model with 6 knots)
%             8 ---> run model III(semiparametric missing data model with 7 knots)
%             9 ---> run model III(semiparametric missing data model with 8 knots)
% take_data    :
%             0 ---> write data for one iteration to bugs_data_1_i.txt, not continue MC
%             1 ---> do not take data, continue MC

% parameters that are fixed throughout the Monte Carlo simulation study:
% true values of the coefficients in the model for covariate subject to missingness(x1)
alpha = [-1.5 0.5];
sigmax = 0.75;

% true values of the coefficients in the response model
beta0 = 2;    % gives 50-50 cases& controls
beta1 = 1;
beta2 =-1;
```

```
% true values of the coefficients in the missing data mechanism
phi0 = -2.75;    % depends on missing percentage of x1
phi1 = 1;
phi2 = 1;
phi3 = 1;
phi4 = 5;

% preliminaries for the output files:
fname1 = ’Model_I_Results.txt’;
fname2 = ’Model_II_Results.txt’;
fname3 = ’Model_III(2)_Results.txt’;
fname4 = ’Model_III(3)_Results.txt’;
fname5 = ’Model_III(4)_Results.txt’;
fname6 = ’Model_III(5)_Results.txt’;
fname7 = ’Model_III(6)_Results.txt’;
fname8 = ’Model_III(7)_Results.txt’;
fname9 = ’Model_III(8)_Results.txt’;
fname10 = ’Case_Miss_Percentage.txt’;


fid1 = fopen(fname1, ’a’);
fid2 = fopen(fname2, ’a’);
fid3 = fopen(fname3, ’a’);
fid4 = fopen(fname4, ’a’);
fid5 = fopen(fname5, ’a’);
fid6 = fopen(fname6, ’a’);
fid7 = fopen(fname7, ’a’);
fid8 = fopen(fname8, ’a’);
fid9 = fopen(fname9, ’a’);
fid10 = fopen(fname10, ’a’);

% Monte Carlo Simulation for mcsize trials:
for mc = 1:mcsize
% Data Generation Process

   % generate x1 and x2
   x2 = normrnd(1,1,n,1);
   x1 = normrnd(alpha(1) + alpha(2)*x2,sigmax);

   % sort x1&x2 by x1 in ascending order
   [x1,ix] = sort(x1);
   x2=x2(ix);

   % generate y, fully observed binary response
   numpy = exp( beta0+ beta1*x1 + beta2*x2 );
   py = numpy./(1+numpy);
   y = binornd(1,py);
   obs_perc_case(mc) = mean(y);
   % generate r, binary missing indicator
   r = zeros(n,1);
   numpr = phi0+phi1*y+phi2*x2+phi4*(0.5-1./(1+(x1+1).^ 4));
   w = normrnd(numpr,1);
   r(find(w>=0)) = 1;
   r(find(w<0)) = 0;
   obs_miss(mc) = mean(r);

   % indexes of missing and complete observations in x1
   miss_id = find(r==1);    full_id = find(r==0);

   % assigning NaN to the missing value in x1
   x1(miss_id) = NaN;

   % covariates of subjects whose x1 are observed
   x1_fully_obs = x1(full_id);
   x2_fully_obs = x2(full_id);

% To be used for determining the hyperparameters of prior

   % obtaining mle of parameters of x1|x2 using the fitted x1|x2 model based on
   % fully observed subjects
   [alpha_mle,dev,stats_alpha] = glmfit(x2_fully_obs,x1_fully_obs,’normal’);

   % regression imputation for missing x1 by using the fitted model x1|x2
   s=100;
   Imputed_x1=zeros(n,s); % nonmissing observation has zero value in this matrix
   mu_reg=zeros(n,1); % nonmissing observation has zero value in this vector

   for j=1:s

      mu_reg(miss_id) = alpha_mle(1)+alpha_mle(2)*x2(miss_id);
      tau=1; % selected by tuning
      Imputed_x1(miss_id,j) = normrnd(mu_reg(miss_id),tau);

   end

   updated_x1 = zeros(n,s);
   updated_x1(miss_id,:)  = Imputed_x1(miss_id,:);
   updated_x1(full_id,:)  = x1(full_id)*ones(1,s);

   x1_NA_initial = zeros(n,1);
   x1_NA_initial(miss_id,:)  = mean(updated_x1(miss_id,:),2);
   x1_NA_initial(full_id,:)  = NaN;
```

```
% Modeling Response Y with NMAR Covariate X1 and Always Observed X2:
% For Model I:
if choose == 1||choose == 0

   % Constructing a prior distribution for the parameters of r|y,x1,x2 using
   % the fully observed and imputed subjects (empirical bayes based prior for phi)

   % we find s mle estimators of phi parameter for different imputed x1 subjects
   phi_mle_j = zeros(4,s);
   phi_cov_j = zeros(4,4,s);

   for j=1:s

      x1_taken = updated_x1(:,j);
      [B,dev,stats] = glmfit([y x2 x1_taken],[r ones(n,1)],...
      'binomial','link','probit');
      phi_mle_j(:,j) = B;
      phi_cov_j(:,:,j) = stats.covb;

   end

   phi_prior_mean_1 = mean(phi_mle_j,2);
   c = 10;
   phi_prior_cov_1 = (1/c).*inv(mean(phi_cov_j,3));

   % creating initial values
   alpha_1 = [0 1]';
   beta_1 = [0 0 0]';
   phi_1 = mean(phi_mle_j,2);
   invsigma2x = 1;
   w = zeros(n,1);
   w(find(r==1)) = 1;
   w(find(r==0)) = -1;

   init0_1 = struct('alpha',alpha_1,'beta',beta_1,'phi',phi_1,...
   'X1',x1_NA_initial,'invsigma2x',invsigma2x,'W',w);

   % create data structure
   dataStruct_1 = struct('N',n,'X1',x1,'X2',x2,'Y',y,'R',r,...
   'phipriormean',phi_prior_mean_1,'phipriorcovinv',phi_prior_cov_1);

   if take_data = 1

      mat2bugs('bugs_data_2_1.txt','N',n,'X1',x1,'X2',x2,'Y',y,'R',r,...
      'phipriormean',phi_prior_mean_1,'phipriorcovinv',phi_prior_cov_1);
      mat2bugs('bugs_init_2_1.txt','alpha',alpha_1,'beta',beta_1,'phi',phi_1,...
      'X1',x1_NA_initial,'invsigma2x',invsigma2x,'W',w);
      return

   end
   % Bayesian Analysis
   [samples, stats, structArray] = matbugs_v2(dataStruct_1, ...
   fullfile(pwd, ’modelI.txt’), ...
   ’init’, init0_1, ...
   ’nChains’, 1, ...
   ’view’, 0, ’nburnin’, 1000, ’nsamples’, 4000, ...
   ’thin’, 3, ’overrelax’, 1, ’DICstatus’, 0, ...
   ’monitorParams’, ’alpha’, ’beta’, ’phi’, ...
   ’Bugdir’, ’C:/Program Files/WinBUGS14’);

   z1 = [stats.mean.beta ];
   fprintf(fid1,’%g %g %g \n’,z1);

end % end for Model I

% For Model II:
if choose == 2||choose == 0

   % Constructing a prior distribution for the parameters of r|y,x1,x2 using
   % the fully observed and imputed subjects (empirical bayes based prior for phi)

   % we find s mle estimators of phi parameter for different imputed x1 subjects
   phi_mle_j = zeros(5,s);
   phi_cov_j = zeros(5,5,s);

   for j=1:s

      x1_taken = updated_x1(:,j);
      [B,dev,stats] = glmfit([y x2 x1_taken x1_taken.*x2],[r ones(n,1)],...
      ’binomial’,’link’,’probit’);
      phi_mle_j(:,j) = B;
      phi_cov_j(:,:,j) = stats.covb;

   end

   phi_prior_mean_2 = mean(phi_mle_j,2);
   c = 10;
   phi_prior_cov_2 = (1/c).*inv(mean(phi_cov_j,3));

   % creating initial values
   alpha_2 = [0 1]’;
   beta_2 = [0 0 0]’;
   phi_2 = mean(phi_mle_j,2);
   invsigma2x = 1;
   w = zeros(n,1);
   w(find(r==1)) = 1;
   w(find(r==0)) = -1;

   init0_2 = struct(’alpha’,alpha_2,’beta’,beta_2,’phi’,phi_2,...
   ’X1’,x1_NA_initial,’invsigma2x’,invsigma2x,’W’,w);
   % create data structure
   dataStruct_2 = struct('N',n,'X1',x1,'X2',x2,'Y',y,'R',r,...
   'phipriormean',phi_prior_mean_2,'phipriorcovinv',phi_prior_cov_2);

   if take_data = 1

      mat2bugs('bugs_data_2_2.txt','N',n,'X1',x1,'X2',x2,'Y',y,'R',r,...
      'phipriormean',phi_prior_mean_2,'phipriorcovinv',phi_prior_cov_2);
      mat2bugs('bugs_init_2_2.txt','alpha',alpha_2,'beta',beta_2,'phi',phi_2,...
      'X1',x1_NA_initial,'invsigma2x',invsigma2x,'W',w);
      return

   end

   % Bayesian Analysis
   [samples, stats, structArray] = matbugs_v2(dataStruct_2, ...
   fullfile(pwd, 'modelII.txt'), ...
   'init', init0_2, ...
   'nChains', 1, ...
   'view', 0, 'nburnin', 1000, 'nsamples', 4000, ...
   'thin', 3, 'overrelax', 1, 'DICstatus', 0, ...
   'monitorParams', 'alpha', 'beta', 'phi', ...
   'Bugdir', 'C:/Program Files/WinBUGS14');

   z2 = [stats.mean.beta stats.std.beta];
   fprintf(fid2,'%g %g %g \n',z2);

end % end for Model II

% For Model III (with 2 knots):
if choose == 3||choose == 0

   % finding the knots by using fully observed and imputed subjects
   nknots = 2; % number of knots
   p=(1:nknots)./(nknots+1);
   S knotss = zeros(s,nknots);
   for j=1:s

      x1_taken = updated_x1(:,j);
      knotss(j,:)=(quantile(x1_taken,p))';

   end
   knots = mean(knotss,1)'; % average value of knots for s different updated x1
   omega_nknots = zeros(nknots,nknots);
   for i=1:nknots   for j=1:nknots

      omega_nknots(i,j) = (abs(knots(i)-knots(j)))^ 3;

   end   end
   invsqrtomega_knots = (omega_nknots)^ -1/2;
```

```
% constructing a prior distribution for the parameters of r|y,x2,x1,z
% using the fully and imputed subjects:

% phi:  coefficients of intercept,y,x2,x1
% brand:  random coefficients of z (Not used for priors, just used for initials and
% prior of the parameter invsigma2b)

% we find s mle estimators for phi and brand parameters
% for different imputed x1 subjects

% emprical bayes based prior for phi
phi_mle_j = zeros(4,s);
phi_cov_j = zeros(4,4,s);

for j=1:s

   x1_taken = updated_x1(:,j); [B,dev,stats] = glmfit([y x2 x1_taken],[r ones(n,1)],...
   ’binomial’,’link’,’probit’);
   phi_mle_j(:,j) = B;
   phi_cov_j(:,:,j) = stats.covb;

end

phi_prior_mean_3 = mean(phi_mle_j(1:4,:),2);
c=10;
phi_prior_covinv_3 = (1/c).*inv(mean(phi_cov_j(1:4,1:4,:),3));

% emprical bayes based prior for invsigma2b
brand_mle_j = zeros(nknots+1,s);
brand_cov_j = zeros(nknots+1,nknots+1,s);

for j=1:s

   x1_taken = updated_x1(:,j);
   z_nknots = zeros(n,nknots);
   for i=1:n   for zi=1:nknots

      z_nknots(i,zi) = (abs(x1_taken(i)-knots(zi)))^ 3;

   end   end
   z = z_nknots*invsqrtomega_knots;
   [B,dev,stats] = glmfit([z],[r ones(n,1)],’binomial’,’link’,’probit’);
   brand_mle_j(:,j) = B;
   brand_cov_j(:,:,j) = stats.covb;

end

brand_mean = mean(brand_mle_j(2:1+nknots,:),2);
brand_cov = mean(brand_cov_j(2:1+nknots,2:1+nknots,:),3);
invsigma2b_prior_mean = 1/mean(diag(brand_cov));
```

```
invsigma2b_prior_var = 1;
invsigma2b_prior_par1 = (invsigma2b_prior_mean^ 2)/invsigma2b_prior_var;
invsigma2b_prior_par2 = invsigma2b_prior_mean/invsigma2b_prior_var;

% creating initial values
alpha_3 = [0 1]’;
beta_3 = [0 0 0]’;
phi_3 = phi_prior_mean_3;
brand = brand_mean;
invsigma2x = 1;
invsigma2b = invsigma2b_prior_mean;
w = zeros(n,1);
w(find(r==1)) = 1;
w(find(r==0)) = -1;

init0_3 = struct(’alpha’,alpha_3,’beta’,beta_3,’phi’,phi_3,...
’X1’,x1_NA_initial,’brand’,brand,’invsigma2x’,invsigma2x,...
’invsigma2b’,invsigma2b,’W’,w);

% creating data structure
dataStruct_3 = struct(’N’,n,’X1’,x1,’X2’,x2,’Y’,y,’R’,r,...
’nknots’,nknots,’knots’,knots,’invsqrtomega’,invsqrtomega_knots,...
’phipriormean’, phi_prior_mean_3,’phipriorcovinv’,phi_prior_covinv_3,...
’invsigma2bpriorpar1’,invsigma2b_prior_par1,’invsigma2bpriorpar2’...
,invsigma2b_prior_par2);

if take_data == 1

   mat2bugs(’bugs_data_2_3.txt’,’N’,n,’X1’,x1,’X2’,x2,’Y’,y,’R’,r,...
   ’nknots’,nknots,’knots’,knots,’invsqrtomega’,invsqrtomega_knots,...
   ’phipriormean’, phi_prior_mean_3,’phipriorcovinv’,phi_prior_covinv_3,...
   ’invsigma2bpriorpar1’,invsigma2b_prior_par1,’invsigma2bpriorpar2’...
   ,invsigma2b_prior_par2);
   mat2bugs(’bugs_init_2_3.txt’,’alpha’,alpha_3,’beta’,beta_3,’phi’,phi_3,...
   ’X1’,x1_NA_initial,’brand’,brand,’invsigma2x’,invsigma2x,’invsigma2b’...
   ,invsigma2b,’W’,w);
   return

end

% Bayesian Analysis
[samples, stats, structArray] = matbugs_v2(dataStruct_3, ...
fullfile(pwd, ’modelIII(2).txt’), ...
’init’, init0_3, ...
’nChains’, 1, ...
’view’, 0, ’nburnin’, 6000, ’nsamples’, 12000, ...
’thin’, 5, ’overrelax’, 1, ’DICstatus’, 0, ...
’monitorParams’, ’alpha’, ’beta’, ’phi’, ...
’Bugdir’, ’C:/Program Files/WinBUGS14’);
```

```
   z3 = [stats.mean.beta];
   fprintf(fid3,’%g %g %g \n’,z3);

   end % end for Model III (with 2 knots)


% For Model III (with 3 knots):
if choose == 4||choose == 0
⋮
end % end for Model III (with 3 knots)

% For Model III (with 4 knots):
if choose == 5||choose == 0
⋮
end % end for Model III (with 4 knots)

% For Model III (with 4 knots):
if choose == 5||choose == 0
⋮
end % end for Model III (with 4 knots)

% For Model III (with 5 knots):
if choose == 6||choose == 0
⋮
end % end for Model III (with 5 knots)

% For Model III (with 6 knots):
if choose == 7||choose == 0
⋮
end % end for Model III (with 6 knots)

% For Model III (with 7 knots):
if choose == 8||choose == 0
⋮
end % end for Model III (with 7 knots)

% For Model III (with 8 knots):
if choose == 9||choose == 0
⋮
end % end for Model III (with 8 knots)

end % end for Monte Carlo simulation

fclose(fid1);
fclose(fid2);
fclose(fid3);
fclose(fid4);
fclose(fid5);
fclose(fid6);
fclose(fid7);
fclose(fid8);
fclose(fid9);
fclose(fid10);

% average observed missing percentage
avrg_obs_miss_perc = mean(obs_miss)

end % end of Simulation.m function
```

Sample *Phi0.m* Function

```
% The sample Phi0.m function calculates the intercept parameter phi0
% for the true missingness mechanism and 15% missing percentage
function Phi0

% True parameters:
beta0 = 2;
beta =[1 -1];
alpha = [-1.5 0.5];
sigma2 = 0.75;
phi =[1 1 1 5];
n = 10000;
missing_percentage = 0.15

% generate x1 and x2
x2 = normrnd(1,1,n,1);
x1 = normrnd(alpha(1) + alpha(2) * x2,sigma2);

% generate y, fully observed binary response
numpy = exp( beta0+ beta(1)*x1 + beta(2)*x2 );
py = numpy./(1+numpy);
y = binornd(1,py);

% choosing phi0 by missing percentage
phi0_init = -50:0.05:50;
miss_prob = zeros(size(phi0_init));
r = zeros(n,1);
w = zeros(n,1);

for k=1:length(phi0_init)

      numpr = phi0_init(k)+phi(1)*y+phi(2)*x2+phi(4)*(0.5-1./(1+(x1+1).^ 4));
      w = normrnd(numpr,1);
      r(find(w>=0)) = 1;
      r(find(w<0)) = 0;
      miss_prob(k) = mean(r);

end
```

```
A = [phi0_init' miss_prob'];
miss_prob_diff = abs(miss_prob-missing_percentage);
[Y I] = min(miss_prob_diff);
phi0 = phi0_init(I)

end % end of Phi0.m function
```

Sample *Results.m* Function

```
% The sample Results.m function calculates the biases and
% mean squared errors (MSEs) of the beta parameters
function Results(text_name,n)

% Input Arguments:
% text_name = 'Model_III(2)_Results.txt'
% n :  number of samples

true_beta = [2 1 -1];

fid_text = fopen(text_name,'r')
data = fscanf(fid_text,'%f',[3 n]);
data = data';

T_mean_average = zeros(1,3);
T_mean_std = zeros(1,3);
T_bias = zeros(1,3);
T_MSE = zeros(1,3);

for j=1:3

      T_mean_average(j) = mean(data(:,j));
      T_mean_std(j) = std(data(:,j));
      T_bias(j) = T_mean_average(j)-true_beta(j);
      T_MSE(j) = T_bias(j)^ 2 + var(data(:,j));

end

T_mean_average
T_bias
T_MSE

fclose(fid_text);

end % end of Result.m function
```

APPENDIX D

Convergence Diagnostics for Gibbs Sampling

D.1 Comparison of Noninformative and Empirical Bayesian Based Gamma Prior for σ_b^{-2}

To monitor the convergence of Markov chains within the Gibbs sampling, we used the Brooks-Gelman-Rubin's convergence diagnostics by creating three chains starting from three dispersed sets of initial points in WinBUGS. In the following figures for the Brooks-Gelman-Rubin's convergence diagnostics, red, green and blue lines indicate the Gelman-Rubin statistic R, the between-chain variance B and the within-chain variance W respectively over iteration-time. The Gelman-Rubin statistic R is the ratio calculated by

$$R = \frac{B}{W} \tag{D.1}$$

The ratio R tends to 1 as convergence is approached. Accordingly, we can determine the burn-in point and the size of the Markov chain by monitoring the Gelman-Rubin statistic R over iteration-time. For the details of the Brooks-Gelman-Rubin's convergence diagnostics, one can refer to Brooks and Gelman [30]. We also observed the autocorrelation plots of the Markov chains to detect the slow convergence due to the poor mixing.

The following figures are obtained by using the Model III (with 2 knots) in Section 3.2 via WinBUGS. If we consider the noninformative $\text{Gamma}(0.01, 0.01)$ prior for the inverse variance component σ_b^{-2}, slow convergence is appeared for the latent random effects b and σ_b^{-2} due to poor mixing in the corresponding Markov chains as shown in Figure D.1 and Figure D.3. However, the proposed empirical Bayes based gamma priors for σ_b^{-2} leads to better mixing and accelerated convergence for the b and σ_b^{-2} as shown in Figure D.2 and Figure D.4.

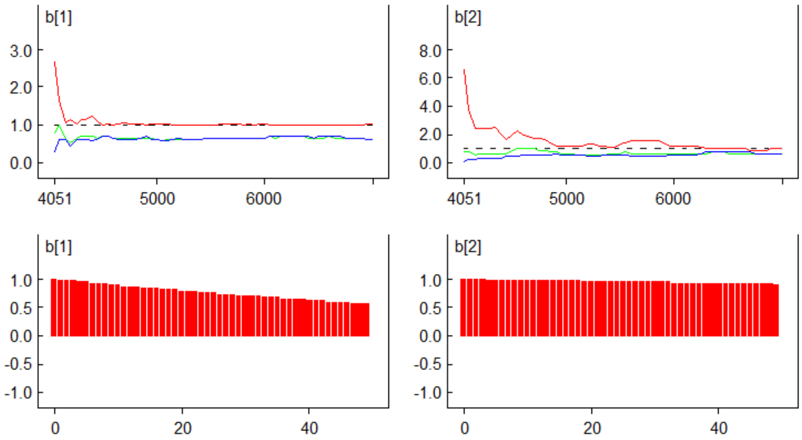

Figure D.1: The Brooks-Gelman-Rubin's Convergence Diagnostics and the Autocorrelation Plots for the random effects b when the noninformative gamma prior is used for σ_b^{-2}

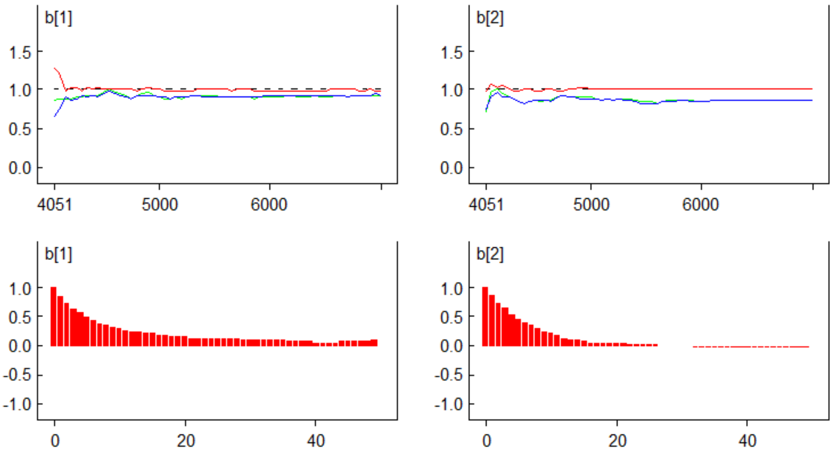

Figure D.2: The Brooks-Gelman-Rubin's Convergence Diagnostics and the Autocorrelation Plots for the random effects b when the empirical Bayes based gamma prior is used for σ_b^{-2}

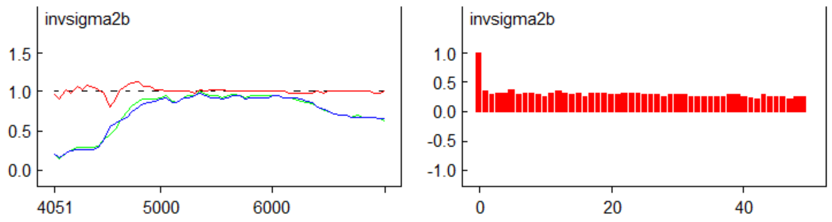

Figure D.3: The Brooks-Gelman-Rubin's Convergence Diagnostics and the Autocorrelation Plots for the parameter σ_b^{-2} when the noninformative gamma prior is used for σ_b^{-2}

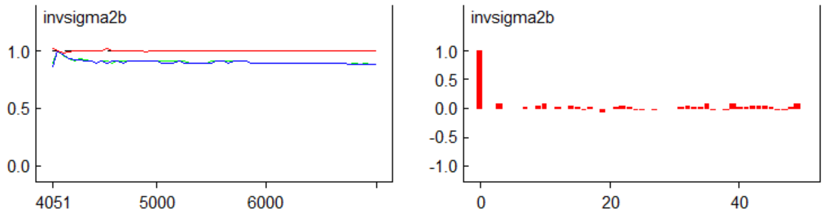

Figure D.4: The Brooks-Gelman-Rubin's Convergence Diagnostics and the Autocorrelation Plots for the parameter σ_b^{-2} when the empirical Bayes based gamma prior is used for σ_b^{-2}

D.2 Brooks-Gelman-Rubin's Convergence Diagnostics for the Parameters of the Proposed Hierarchical Model

The following figures show the Brooks-Gelman-Rubin's convergence diagnostics (see Appendix D.1) for the parameters of Model III (with 2 knots) in Section 3.2. The burn-in point and the size of the Markov chain are determined by using these figures for the proposed hierarchical model.

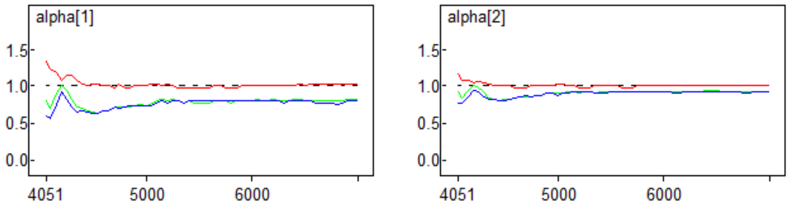

Figure D.5: The Brooks-Gelman-Rubin's Convergence Diagnostics for the parameter α

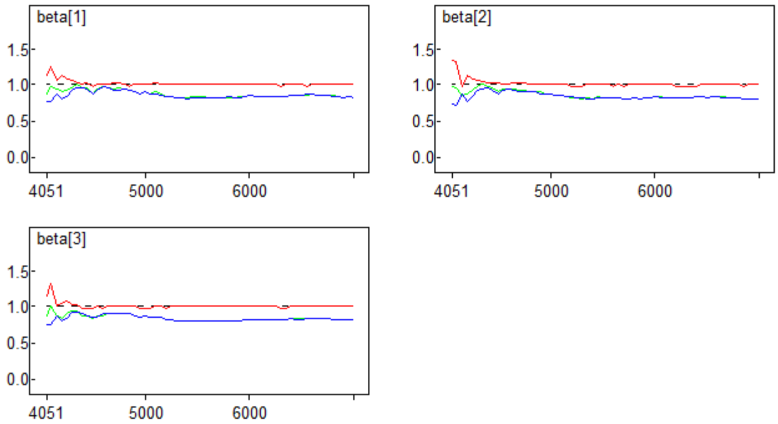

Figure D.6: The Brooks-Gelman-Rubin's Convergence Diagnostics for the parameter β

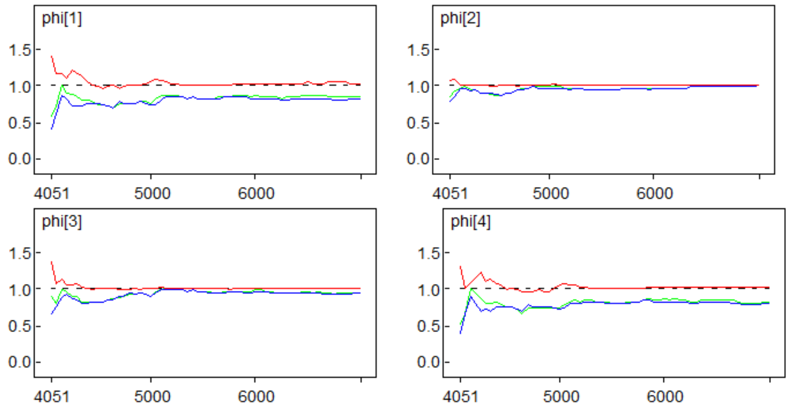

Figure D.7: The Brooks-Gelman-Rubin's Convergence Diagnostics for the parameter ϕ

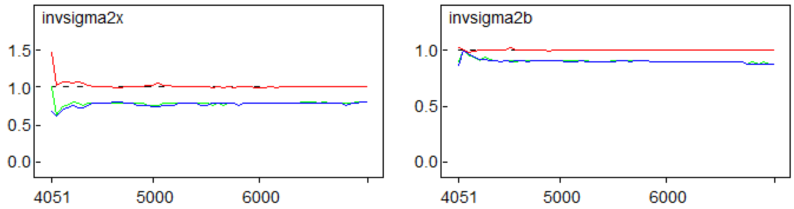

Figure D.8: The Brooks-Gelman-Rubin's Convergence Diagnostics for the parameters σ_x^{-2} and σ_b^{-2}