Abstract

ZHU, LIANSHENG. Analyzing Longitudinal Data with Non-Ignorable Missing. (Under the direction of Dr. Sujit Ghosh and Dr. Subhashis Ghosal.)

In longitudinal studies, data are often missing despite every attempt made to collect complete data. When the missingness is informative and hence not ignorable, it is generally difficult to analyze non-ignorable missing (NIM) data since the distributional assumptions about missing data are not easily verifiable using traditional goodness of fit tests or otherwise. Selection models and pattern-mixture models are two common approaches to analyze NIM data. Each approach has its advantages and disadvantages. Methods proposed in this thesis fall into the category of pattern-mixture models. Traditionally, patterns are determined by time to occurrence of missing. This definition often results into the problem of not all parameters being identifiable. Moreover, marginalization is commonly required and can be very tricky when outcomes are discrete. It is recognized that patterns can and need to be defined by covariates, surrogate variables and/or time to missing. We propose two approaches to model NIM data: (i) pseudo-imputation (PI) approach, in which we first obtain predictive means within each pattern, get transformed predictive means by using a suitable link function and then fit with covariates to obtain marginal estimates; (ii) joint-modeling (JM) approach, in which patterns considered as random effects are marginalized within a generalized linear mixed model framework. The JM approach is shown to be able to capture the dependence of missing indicators on missing outcomes in some degree as is the case with NIM data. Some of the main advantages of these proposed approaches include (i) the capability to handle both continuous and discrete responses, (ii) avoidance of the problem of

under-identifiability, (iii) availability of marginal estimates, and (iv) computational efficiency. When the missingness does depend on the patterns, results based on simulated data suggest that both approaches yield accurate estimates if the underlying number of patterns is specified correctly. Otherwise the PI method leads to biased results whereas the JM approach still provides reasonably accurate estimates.

Finally, we extend our approaches to a generalized additive model (GAM) replacing the GLM framework. When the underlying relationship is highly non-linear, our extended approaches with a GAM framework provide flexibility and more accurate estimates. The JM approach along with generalized additive models can provide more flexibility than the PI approach since it uses a more robust model for the missing indicator.

ANALYZING LONGITUDINAL DATA WITH NON-IGNORABLE MISSING

by

LIANSHENG ZHU

A dissertation submitted to the Graduate Faculty of
North Carolina State University
In partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy

STATISTICS

Raleigh, NC
2006

APPROVED BY

Sujit Ghosh, Co-Chair	Subhashis Ghosal, Co-Chair
Anastasios Tsiatis	Hao Zhang

To my parents, my wife and my daughter.

Biography

Liansheng Zhu was born in Qidong, a coastal town in the Eastern China. He graduated from Huilong High School at Qidong in 1994. He received a Bachelor of Sciences degree in Chemistry and Chemical Engineering from Hunan University at Changsha, China in 1998. He obtained a Master's of Sciences degree in Statistics from Kansas State University at Manhattan, Kansas in 2003. He then joined the graduate program in the Department of Statistics at North Carolina State University in August 2003. He is working as a biostatistician at PPDI starting May 2006.

Acknowledgements

First, I would like to express my deepest appreciation to my wonderful advisors Dr. Sujit Ghosh and Dr. Subhashis Ghosal for their guidance, training and support during this research. It is my honor to have them as my advisors, to learn and get inspiration from them. Also, I am grateful to the valuable suggestions and comments from Dr. Anastasios Tsiatis and Dr. Hao Zhang on this research and their presence in my committee. Many thanks also go to Dr. Jie Zhang in Novartis Pharmaceuticals from whom I got inspiration to initiate this research. Finally, I want to thank my dearest parents, sister and brother, for their endless love and support. Special thanks also go to my parents-in-law who come and help us out at this very time.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In longitudinal studies, measurements of a response variable and some explanatory variables for each subject are recorded repeatedly across time. The primary interest is often the estimation of either the change in the response variables over time or the effects of independent variables at the completion of studies. Longitudinal studies are popular in epidemiological research. Besides that, they are gaining increasing popularity in clinical trials of intervention dealing with chronic diseases. Although most studies are well designed and every attempt is made to collect data on each subject at each of the follow-up times, it is often unavoidable to have some measurements missing intermittently and/or even have subjects dropped out prior to the completion of the study. The reasons behind such missing may be study-related, for example, adverse event, competing risks, lack of effectiveness of treatments, or even early recovery. It can also be non-study related, e.g., accidental death or moving away. An example of a longitudinal study is given in Section 1.1. Section 1.2 provides notations and definitions for model description. Some statistical methods in literature dealing with non-ignorable missing are discussed in Section 1.3.

1.1 A Motivating Study

According to the National Osteoporosis Foundation, vertebral fractures are the most common osteoporotic fracture. Among postmenopausal women who sustain a vertebral

fracture, one out of five will suffer their next vertebral fracture within just one year, potentially leading to a fracture "cascade". There is a three-year clinical trial in which binary outcomes (new fracture versus no new fracture) of vertebral fracture are measured at the end of each year. Baseline measurements include patients' age, Tscore (calculated from bone mineral density), and baseline fracture status (0, 1, ≥ 2). Patients are randomly assigned to either the treatment group or the placebo group. Some patients may not have any post-baseline measurements and some patients may only have incomplete post-baseline measurements. In this specific study, it was defined that if a fracture is determined for the patient at the j^{th} year, it will be assumed that his or her post j^{th} year outcome is a fracture as well since the vertebral fracture can not be cured (once a fracture, always a fracture). Therefore, data from these patients are not counted as missing even when there are no measurements available. That means the subjects are considered as dropouts only if they do not experience fracture before missing. The primary analysis is to compare the intervention with the control on vertebral fracture based on the population excluding patients without any post-baseline data. There is reason to suspect that dropouts may depend on responses because subjects with new fractures may be more likely to quit than those with no fractures. Hence, standard statistical methods do allow the missing mechanism to depend on missing observations may yield biased estimates and hence misleading conclusions might be drawn by the clinicians.

1.2 Notations and Definitions

Let Y_{ij} be the responses of the i^{th} subject measured at the j^{th} time point, X_i denote the vector of baseline covariates, and R_{ij} be the missing indicator variable for $i = 1, \ldots, n$, and $j = 1, \ldots, m$. Let $R_{ij} = 1$ if Y_{ij} observed and $R_{ij} = 0$ otherwise. Let Y be the matrix of Y_{ij}'s and X and R be matrices of X_i's, and R_{ij}'s. The variable Y can be expressed as (Y_{obs}, Y_{miss}), where Y_{obs} denotes observed responses and Y_{miss} the missing ones.

Consider again the above example of the 3-year study. If subjects have at least one post-baseline measurement, and then they are included in the analysis. Hence, there are only three missing patterns since a monotonic missing mechanism (dropout) is observed. More generally, if subjects are included as long as they baseline measures, then there

Table 1.1: The typical data structure. (NA: missing data)

Subject	Year 1	Year 2	Year 3
1	y_{11}	y_{12}	y_{13}
2	y_{21}	y_{22}	NA
3	y_{31}	NA	NA
4	y_{41}	NA	y_{43}
5	NA	y_{52}	y_{53}
6	NA	y_{62}	NA
7	NA	NA	y_{73}
8	NA	NA	NA
.	.	.	.
.	.	.	.

can be eight possible missing patterns. A typical data structure is illustrated in Table 1, where NA indicates a missing observation. In this illustration, $R_{11} = 1$, $R_{23} = 0$ etc.

In Rubin and Little's taxonomy (1976, 1987), there are in general three missing mechanisms that can be expressed in terms of the conditional distribution of R given X and Y. Let $f(U|V)$ denote the conditional distribution of U given V.

1. Missing completely at random (MCAR): $f(R|X, Y) = f(R)$. One extension of MCAR is defined as the covariate-dependent MCAR if $f(R|X, Y) = f(R|X)$.

2. Missing at random (MAR): $f(R|X, Y) = f(R|Y_{obs}, X)$.

3. Non-ignorable missing (NIM): $f(R|X, Y) = f(R|X, Y_{obs}, Y_{miss})$.

The third mechanism (NIM) which makes no restrictive assumption on missing mechanisms is the main focus of our research.

1.3 Models and Associated Methods for NIM Data

It is generally difficult to model NIM data since the distributional assumptions about the missing data are not easily verifiable using traditional goodness of fit (GOF) tests or

otherwise. There is a large volume of literature devoted to reduce bias resulting from this type of missing. Two widely used approaches are selection models and mixture models. In mixture models, the joint density of repeated outcomes Y and missing indicator R, conditional on covariates, X is written as $f(Y, R|X) = f(Y|R, X)f(R|X)$ whereas for selection models it is written as $f(Y, R|X) = f(Y|X)f(R|Y, X)$.

1.3.1 Selection Models

Selection models, which were first introduced in the univariate setting by Heckman (1976) appearing in the econometrics literature, require one to combine a model for the full data with a selection model for missing indicators given data. Diggle and Kenward (1994) extended Heckman's work to multivariate normal data. Fizmaurice et al. (1996) proposed multivariate logistic models for incomplete binary responses while Molenberghs et al. (1997) introduced methods for analyzing longitudinal ordinal data with non-ignorable missing. In practice, one can assume a generalized linear regression model for the full-data using a suitable linear function to explain effects of covariates and a logit or probit model for the selection model.

Selection models are intuitively appealing since the inference about parameter estimates of primary interest can be obtained directly from $f(Y|X)$, and parameters in $f(R|Y, X)$ can be treated as nuisance parameters provided there are no common parameters in $f(Y|X)$ and $f(R|Y, X)$. However, some limitations do exist. For instance, for continuous responses, the common assumption of multivariate normality is not easy to verify when some responses are missing (Hogan et al., 2004). Moreover, the selection model postulated to model the dependence of dropout process on unobserved responses is not verifiable. Inference from selection models can be highly sensitive to model assumptions (Little and Rubin, 2002). Finally, computations can be very intensive due to general requirement of likelihood maximization and numerical instability occurs resulting from the possibility of a flat likelihood about the parameters of primary interest (Hogan, 1997).

1.3.2 Pattern-mixture Models

Although the terminology of pattern-mixture models that are our main focus here was first defined by Little (1993), the early application of mixture concepts can be traced back to Rubin (1977). Conventionally, marginalization of results are often in need. For example, Little (1993, 1994) and Molenberghs et al. (1998) first stratified the incomplete data by missing patterns, then fitted distinct models within each stratum, and finally summarized results over patterns. The general model can be written as

$$f(Y, R|X) = f(Y|R, X)f(R|X), \tag{1.1}$$

where the distribution of Y, conditional on R and X, is multinomial and R is treated as a covariate, and the distribution of R given X can be left unspecified or in a parametric form. Marginalization over patterns is required to obtain estimates of parameters of primary interest.

Traditionally, missing patterns are defined according to times to missing. For example, there are eight patterns for a typical data set like the one shown in Table 1.1. Without a doubt, it is often unavoidable that not all parameters are estimable because of some missing observations. In other words, pattern-mixture models are under-identified (Little 1993). Either restriction must be set for some patterns, additional assumptions be made, or information need to be borrowed from observed data. To overcome the under-identifiable problem, Wu and Bailey (1988, 1989) allow responses to depend on missing indicators through individual random effects estimated from a general linear mixed model, then average over the distribution of random effects that can be assumed to be parametric or left unspecified. The model can be written as

$$f(Y, R|X) = \int f(Y|b, X)f(R|b, X)dF(b), \tag{1.2}$$

where b is the subject-specific random parameter vector. The variables Y and R are linked through b and inference can be made based on $(Y|b, X)$ by integrating out b. This model is often called shared parameter (SP) model. Follmann and Wu (1995) extended the work of Wu and Bailey (1988, 1989) to permit generalized linear models for discrete

responses without any parametric assumption on random effects. Estimation of marginal means for binary cases was described by Birmingham and Fitzmaurice (2002).

Guo et al. (2004) proposed a random pattern-mixture model for continuous data in which they considered dropout patterns as random latent variables and determined according to many factors, including times to missing, baseline covariates and/or time-varying covariates. The key assumption of this model is that the longitudinal responses, Y, are independent of the missing indicators, R, conditional on the pattern effects. The proposed model is given by,

$$\begin{aligned} f(Y,R,b|X) &= \int f(Y|b,X,u,R)f(b|X,R,u)f(R|u,X)f(u|X)du \\ &= \int f(Y|b,X,u)f(b|X,u)f(R|u,X)f(u|X)du, \end{aligned} \tag{1.3}$$

where b is the subject-specific random parameter vector and u is the latent variable vector representing the random pattern effect. To obtain parameters of primary interest, one requires marginalization over b and u which can be computationally difficult. To avoid this task, a joint normal approach was adopted and maximum likelihood estimates were derived by EM algorithm.

As will be seen, marginalization is often required. It may not be that problematic if responses are continuous and/or the final parameter of interest is the marginal mean. However, in situations where there are a number of covariates of interest in discrete longitudinal data, methods for parameter-wise marginalized results can be very tricky. For instance, suppose that there are two covariates, treatments and patients' age, denoted as X_1 and X_2, respectively. Also, suppose that there are two missing patterns. Then, a logistic regression is usually fit for binary data conditional on each missing pattern,

$$P^{(k)} = (1 + \exp[-(\beta_0^{(k)} + X_1^{(k)}\beta_1^{(k)} + X_2^{(k)}\beta_2^{(k)})])^{-1}. \tag{1.4}$$

where k is the pattern indicator, 1 or 2. Obviously, $\beta_1^{(k)}$ is the logarithm of odds ratio between two treatments given the value of age and the pattern k. However, what we need is a marginal parameter β_1 to measure the effect of treatment irrespective of the pattern. The problem here is how to define a β_1 that has the same nice representation

as $\beta_1^{(k)}$'s. The above methods that are available in the literature do not provide such marginal parameter β_1 such that the marginalized estimates of parameters are still of meaningful representation as the conditional estimates based on patterns.

To resolve both the under-identifiability and marginalization problems, we propose two approaches in the Chapter 2. The first one will be called the pseudo imputation (PI) approach, an approximation method for estimating parameters at the marginal level. Secondly, we extend the method proposed by Guo *et al*. (2004) to general cases by jointly modeling Y and R within a Bayesian framework and will be called the joint-modeling (JM) approach. Both methods are applicable to both discrete and continuous outcomes. At this moment, it is assumed that a surrogate variable is observed and missing patterns are known. Simulation studies are conducted to illustrate the proposed methods and to compare them with some exiting methods. In Chapter 3, we extend our approaches to accommodate more general situations and relax the assumptions in Chapter 2. In Chapter 4, the proposed methods are illustrated using a real life data set from an aids study. Conclusions and future work are presented in Chapter 5.

Chapter 2

Proposed Statistical Methods

Firstly it is assumed that a surrogate variable determining missing patterns is fully observed. Then, we define Z_i, that has a known distribution, as the design matrix of missing patterns with the random effect u_i. We further assume $f(R_{ij}|Y_{ij}, X_i, Z_i, u_i) = f(R_{ij}|X_i, Z_i, u_i)$, which means that the missing mechanism is conditionally MAR given the latent variable u_i. For a given pattern, data often contain both observed responses and missed ones, the problem of under-identifiability, which is encountered by the traditional pattern-mixture models that use time to dropout as patterns, does not exist any more. The models presented in this chapter are general enough in the sense of that we do not require the assumption of monotonic or intermittent missingness pattern. The pseudo-imputation (PI) method and its extension will be described in Section 2.1, followed by the joint modeling (JM) method. In Section 2.3, simulation studies based on different scenarios are carried out. Section 2.4 presents conclusions and some general remarks.

2.1 The Pseudo-imputation (PI) Method

2.1.1 Model Description

The first step is to estimate $E(Y_{ij}|X_i, Z_i) = \mu_{ij}$ and its corresponding variance within each pattern for both observed and missing subjects. Standard methods can be used for this step here. For example, generalized linear models can be used for discrete responses.

As it is well-known in the multiple imputation technique, more than one values are simulated for each missing point based on estimated $E(Y_{ij}|Z_i, X_i)$ and its variance that are obtained at the first step. On the contrary, the pseudo-imputation method does not impute the missing values but directly models estimated $E(Y_{ij}|Z_i, X_i)$ or its transformation in linear regression framework. By doing so, estimates of marginal parameters can be obtained. In summary, the following three steps are needed to implement the PI method:

1. Given Z_i, estimate μ_{ij} by using a standard method (e.g., GLMM).

2. Transform the estimate, say $h(\hat{\mu}_{ij})$, by using a desired functional form. For example, logit for binary data and log for count data. By doing so, $h(\hat{\mu}_{ij})$ will be linear in β.

3. Fit a linear model
$$E(h(\hat{\mu}_{ij})|X_i) = X_i^T\beta \tag{2.1}$$
where the estimate of β is indeed at the marginal level.

Notice that the variance of estimates obtained at Stage 3 is under estimated since it ignores the fact that a major proportion of variation has been reduced at Stage 1. Hence, we need to derive the variance of estimates for fixed effects, denoted as $var(\hat{\beta})$. In linear models, $var(\hat{\beta}) = (X^TV^{-1}X)^{-1}$, where,

$$V = var(h(\hat{\mu}_{ij})) = var(E(h(\hat{\mu}_{ij})|Z_i)) + E(var(h(\hat{\mu}_{ij})|Z_i)) \tag{2.2}$$

If a linear mixed model is used at Stage 3 and u treated as random with mean 0 and variance Ω, then $E(var(h(\hat{\mu}_{ij})|u)) = Z^T\Omega Z + \Sigma$ where Ω and the random error Σ can be estimated at the third stage and $E(var(h(\hat{\mu}_{ij})|u))$ at the first stage. Therefore, $var(\hat{\beta})$ can be obtained without much difficulty. Explicit steps to get parameter estimates and their corresponding variance are demonstrated in the following example.

2.1.2 Example

For illustration we consider the binary data described in Section 1. The original data can often be written as $(Y_{ij}, R_{ij}, X_i, S_i)$, where S_i indicates a discrete-valued surrogate

variable. Now, we can rewrite (Y_{ij}, X_i) as

$$\begin{aligned} Y_{ij}^{(k)} &= \{Y_{ij} : S_i = k\} \\ X_i^{(k)} &= \{X_i : S_i = k\} \end{aligned} \tag{2.3}$$

Suppose the variable, baseline fracture status (BFS), is a good surrogate for responses and missing indicators and hence it links Y and R. Then, we can have three patterns corresponding to S_i=BFS$_i = k$, $k = 1, 2, 3$. Hence, the first step is to regress $Y_{ij}^{(k)}$ on $X_i^{(k)}$ and obtain $\beta^{\hat{(k)}}$ and $\hat{\mu}_{ij}^{(k)}$. In this step, a logistic regression can be used since responses are binary,

$$\begin{aligned} Y_{ijk} &\sim \text{Ber}(P_{ij}^{(k)}) \\ \mu_{ij}^{(k)} &= P_{ij}^{(k)} = [1 + \exp(-X_i^{(k)}\beta^{(k)})]^{-1}. \end{aligned} \tag{2.4}$$

For instance, $X_i^{(k)} = (1, \text{Tscore}_i^{(k)}, \text{year}_i^{(k)}, \text{trt}_i^{(k)}, \text{age}_i^{(k)})^T$. Notice that BFS is not included in the model since there is only one level of BFS in each pattern by definition.

After obtaining the estimates of $\mu_{ij}^{(k)}$, $\hat{\mu}_{ij}^{(k)}$ at the first step, the second step is to transform $\hat{\mu}_{ij}^{(k)}$ by the logit function, denoted as $h(\hat{\mu}_{ij}^{(k)})$, which is linear in β. Obviously, $h(\hat{\mu}_{ij}^{(k)}) = X_i^{(k)}\beta^{(k)}$. At the third step, the marginal estimate, $\hat{\beta}$, will be obtained by using a linear model for $\hat{\mu}_{ij}^{(k)}$ and all covariates,

$$E(h(\hat{\mu}_{ij}^{(k)})) = X_i^T\beta, \tag{2.5}$$

where $X_i^T = (1, \text{Tscore}_i, \text{year}_i, \text{trt}_i, \text{age}_i, \text{BFS}_i)^T$. The variance of $h(\hat{\mu}_{ij}^{(k)})$ is given by

$$\hat{V} = \hat{\text{Var}}(h(\hat{\mu}_{ij}^{(k)})) = \hat{\text{Var}}(E(h(\hat{\mu}_{ij}^{(k)})|S_i)) + \hat{E}(\text{Var}(h(\hat{\mu}_{ij}^{(k)})|S_i)). \tag{2.6}$$

Since the logit function is used at the second step, $h(\hat{\mu}_{ij}^{(k)}) = X_i^{(k)}\beta^{\hat{(k)}}$. Hence,

$$\begin{aligned} \hat{\text{var}}(E(h(\hat{\mu}_{ij}^{(k)})|S_i)) &= \hat{\Sigma} + S^T\hat{\Omega}S \\ \hat{E}(\text{Var}(h(\hat{\mu}_{ij}^{(k)})|S_i) &= \text{Var}(X_i^{(k)}\beta^{\hat{(k)}}) = X_i^{(k)}\text{Var}(\beta^{\hat{(k)}})(X_i^{(k)T}). \end{aligned} \tag{2.7}$$

Where S is the vector of S_i's. The estimate of variance-covariance matrix for the within-pattern variation and the between-pattern variance ($\hat{\Sigma}$ and $\hat{\Omega}$ can be obtained at the third step while $\mathrm{Var}(\hat{\beta^{(k)}})$ is available at the first step. Let V^* be the matrix of $\mathrm{Var}(\hat{\beta^{(k)}})$. Therefore, $var(\hat{\beta}) = (X^T\hat{V}^{-1}X)^{-1} = (X^T(\hat{\Sigma} + S^T\hat{\Omega}S + V^*)^{-1}X)^{-1}$.

2.1.3 Extension and Asymptotic Property

Previously, we assume there exists a desirable link function at the second stage of the proposed method. However, it is not necessary. In the following, we are going to construct the estimation equation and derive asymptotic properties in a general sense. Denote the estimates of the predictive means obtained at the first stage to be $\hat{\mu}_k$ and n_k be the number of observations for the k^{th} pattern. Notice that $\mu_k = \mu$ under the true model. Then the likelihood score equation for approximating β is

$$\sum_{k=1}^{K} n_k \left(\frac{\partial \mu_k(\beta)}{\partial \beta}\right) V_k^{-1}(\hat{\mu}_k - \mu_k(\beta)) \tag{2.8}$$

and the variance of the maximum likelihood estimator for $\hat{\beta}$ can be estimated by

$$\left[\sum_{k=1}^{K} n_k \left(\frac{\partial \mu_k(\beta)}{\partial \beta}\right) V_k^{-1} \left(\frac{\partial \mu_k(\beta)}{\partial \beta}\right)^T\right]^{-1} \tag{2.9}$$

Lemma 1. *Suppose that β_0 is the true parameter and that*
(i) $n_k^{\frac{1}{2}}(\hat{\mu}_k - \mu_k(\beta_0))$ has asymptotic distribution $N(0, V_k)$ for $k = 1, \ldots, K$,
ii) $\frac{\partial \mu_k(\beta)}{\partial \beta}$ and V_k for all k are continuous in the neighborhood of β_0 and the inverse of the limit of

$$\tfrac{1}{n}\textstyle\sum_{k=1}^{K} n_k(\frac{\partial \mu_k(\beta)}{\partial \beta})V_k^{-1}(\frac{\partial \mu_k(\beta)}{\partial \beta})^T$$

exists. Then,
(a) $\hat{\beta}$ is consistent for β_0,
(b) $n^{\frac{1}{2}}(\hat{\beta} - \beta_0)$ has an asymptotic normal distribution.

Proof. In general, consistency will hold if the expectation of the likelihood equation (1.18)

is equal to 0 and some regularity conditions are satisfied. The regularity conditions for consistency can be found in Serfling (Chapter 4, 1980), Huber (Lemma 2.1 and Theorem 2.2 in Chapter 6, 1981) , and van der Vaart (Theorems 5.7 and 5.9, 1998). Here, we only check if the expectation condition holds. Clearly, as the expectation is with respect to β_0, $E\left[(\hat{\mu}_k - \mu_k(\beta))|Z_k\right] = 0$ as the condition (i) in Lemma 1 holds. Hence, with respect to β_0,

$$E\left[\sum_{k=1}^{K} n_k(\frac{\partial \mu_k(\beta)}{\partial \beta})V_k^{-1}(\hat{\mu}_k - \mu_k(\beta))|Z_k\right] = 0.$$

By conditional expectation, it results in

$$E\{\sum_{k=1}^{K} n_k(\frac{\partial \mu_k(\beta)}{\partial \beta})V_k^{-1}(\hat{\mu}_k - \mu_k(\beta))\} = 0.$$

Therefore, $\hat{\beta}$ is the consistent estimator of β_0. To derive the large sample distributional properties of the above estimator, one can follow the usual approach for the M-estimators (See e.g., Serlfing, 1980, Chapter 7; Van der Vaart, 1998, Chapter 5). Details are omitted.

□

2.2 Joint Modeling

2.2.1 Model Description

We propose to use the following model,

$$\mu_{ij} = E(Y_{ij}|X_i, Z_i, u_i, \beta) = g_1(X_i^T\beta + Z_i u_{1i}) \tag{2.10}$$

$$\pi_{ij} = E(R_{ij} = 1|u_i, X_i, Z_i, \alpha) = g_2(X_i^T\alpha + Z_i u_{2i}) \tag{2.11}$$

$$u_i \sim F(.) \tag{2.12}$$

where g_1 and g_2 are monotone link functions, the random effects $u_i = (u_{1i}, u_{2i})$ have a bivariate distribution which can be left unspecified. X_i is a known design matrix. The parameter vectors β and α are unknown and β is of primary interest. u_{1i} and u_{2i} are correlated random effects with the design matrix Z_i for dropout patterns. Here we call Model (2.10) as the Y-model and Model (2.11) the R-model. With the key assumption of $f(Y|R, X, Z, u_1) = f(Y|X, Z, u_1)$, the joint distribution of Y and R with u for random pattern effects can be written as

$$f(Y, R, u|X, Z) = f(Y|R, X, Z, u)f(R|X, Z, u)f(u|X) = f(Y|X, Z, u)f(R|X, Z, u)f(u|X). \tag{2.13}$$

Lemma 2. *The proposed model (2.10) - (2.12) captures the dependence of R on Y.*

Proof. Notice that, under MAR, $f(R|Y, X) = f(R|Y_{obs}, X) \Leftrightarrow f(Y_{miss}|Y_{obs}, R, X) = f(Y_{miss}|Y_{obs}, X)$. Hence, it is enough to show that $f(Y|R, X)/f(Y_{obs})$ is not free of R.

Notice that,

$$\begin{aligned} f(Y|R,X) &= \int f(Y,u|R,X)du \\ &= \int f(Y|u,R,X)f(u|X,R)du \\ &= \int f(Y|u,X)f(u|X,R)du \\ &= \int f(Y|u,X)\frac{f(R|u,X)f(u|X)}{f(R|X)}du \\ &= \frac{\int f(Y|u,X)f(R|u,X)f(u|X)du}{\int f(R,u|X)du} \end{aligned} \tag{2.14}$$

and

$$f(Y_{obs}|R,X) = \frac{\int\int f(Y_{obs},Y_{miss}|u,X)f(R|u,X)f(u|X)dudY_{miss}}{\int f(R,u|X)du} \tag{2.15}$$

Hence,

$$\frac{f(Y|R,X)}{f(Y_{obs}|R,X)} = \frac{\int f(Y_{obs},Y_{miss}|u,X)f(R|u,X)f(u|X)du}{\int\int f(Y_{obs},Y_{miss}|u,X)f(R|u,X)f(u|X)dudY_{miss}} \tag{2.16}$$

which can not be free of R in general. Therefore, our proposed model does capture the dependency of R on Y which is the key assumption of NIM. □

For the purpose of illustration, we investigate the dependence of R on Y through a numerical example by using Monte Carlo approximation. Suppose the univariate responses, $Y_i \sim Ber(1,p_i)$ and $R_i \sim Ber(1,\pi_i)$, where $\text{logit}(p_i) = X_i^T\beta$, $\beta^T = (\beta_0,\ldots,\beta_5) = (-6.9,-0.5,0.7,0.04,1,1.8)$ are true values, $X_i^T = (1,\text{Tscore}_i,\text{trt}_i,\text{age}_i,\text{BFS}_i)^T$, $\text{logit}(\pi_i) = -1.5 - 0.5x_i + 0.5Y_i$ and $i = 1\ldots100$. Values of covariates are generated by the same way as in Section 2.3. By the data construction, R does depend on Y. Now, we would like to see if our formulation captures the dependency. Equivalently, if $\frac{f(Y|R=1,X)}{f(Y_{obs}|R=1,X)} = \frac{f(Y|R=0,X)}{f(Y_{obs}|R=0,X)}$, then the ratio is free of R. Otherwise, it is not. Given values of parameters, Equation (2.14) can be approximated by

$$\frac{f(Y|R,X)}{f(Y_{obs}|R,X)} \approx \frac{N^{-1}\sum_{l=1}^{N} f(Y_{obs},Y_{miss}|u_l,X)f(R|u_l,X)}{N^{-1}\sum_{l=1}^{N}\int f(Y_{obs},Y_{miss}|u_l,X)f(R|u_l,X)dY_{miss}} \tag{2.17}$$

where u_l has the bivariate standard normal distributions with correlation coefficient $\rho = 0.6$. For each given set of values of u_l, terms inside of summand can be evaluated. By repeating it $N = 5000$ times, ratios that can be obtained with respect to R are 0.0099 and 0.0027. Clearly, they are different. Hence, our proposed method captures dependency of R on Y.

Let $D_{obs} = (Y_{obs}, X, Z, R)$ denote the observed data. Write $D = (Y, X, Z, R)$ and thus D is not fully observed. For simplicity, let us assume u has the bivariate normal distribution with mean 0 and variance-covariance matrix Σ_u. We consider independent priors for β, α, and Σ_u, so that $\pi(\beta, \alpha, \Sigma_u) = \pi(\beta)\pi(\alpha)\pi(\Sigma_u)$. Then, the joint posterior of $\theta = (\beta, \alpha, \Sigma_u)$ is

$$\pi(\beta, \alpha, \Sigma_u | D_{obs}) = \left[\prod f(Y_{obs,ij} | X_i, Z_{ij}, u_i) f(R_{ij} | X_i, Z_{ij}, u_i) f(u_i | X_i) \right] \pi(\beta, \alpha, \Sigma_u). \tag{2.18}$$

In the above equation, β is the parameter of primary interest and α, Σ_u are considered as nuisance parameters.

2.2.2 Prior Selection and Posterior Computation

The choice of prior distributions reflects information about unknown parameters. It is desirable to pick mathematically manageable functions as prior distributions which results in computationally convenient posterior distributions. Hence, conjugate prior distributions are often desired. If responses are discrete as in our case, finding conjugate priors can be hardly possible, particularly for parameters β and α in the proposed model.

Moreover, the data provide little information about the additional variance components (Σ_u) for the latent variable - missing patterns. Therefore, the specification of a proper prior for Σ_u might be a crucial element in this missing-data problem. This prior distribution should carry some degree of information, capturing a reasonable amount of variation between dropout patterns and association between $f(Y|X, Z, u)$ and $f(R|X, Z, u)$. Intuitively, a Wishart distribution with the scale matrix Q and the degrees of freedom η can be a good candidate as the prior for the inverse of Σ_u. For priors of β and α, we simply take the normal distribution with mean 0 and precision τI. A small value of τ is

chosen resulting in nearly a non-informative prior.

To sample from the joint posterior distribution $\pi(\beta, \alpha, \Sigma_u | D_{obs})$, a standard Gibbs sampling algorithm obtains samples from the following conditional distributions: 1) $[\beta|\alpha, \Sigma_u, D_{obs}]$, 2) $[\alpha|\beta, \Sigma_u, D_{obs}]$, and 3) $[\Sigma_u|\beta, \alpha, D_{obs}]$. Closed forms of the above three full conditional posterior distributions are not available. However, if these full conditional distributions are log concave functions, we can use the adaptive rejection sampling (Gilks, 1992). By construction, the above condition holds. Therefore, direct sampling or derivative-free adaptive rejection sampling algorithm (Gilks, 1992) can be adopted to draw samples from each of the above three conditional posterior distributions.

Convergence of MCMC samples is often an issue for Bayesian computing. Fortunately, it can be easily checked in WinBUGS by looking at trace plots with running multiple chains. In our cases two chains mix quite well after a few thousand samples from each chain. Initial samples can be discarded as burn-in samples and then remaining iterations can be retained to obtain approximate posterior estimates of the desired parameters. Density plots can be obtained to summarize posterior distributions. Literature on how to deal with these issues in MCMC samples includes Brooks (1998), Cowles and Carlin (1996), and Gilks et al. (1996).

2.2.3 Model Selection

The joint-modeling approach requires determination of the number of patterns. Initially, we assume the underlying number of patters are known. In reality, it remains unknown and ultimately needs to be determined. It can be determined arbitrarily or in a purely data-driven way. Hence, it is critical to assess model adequacy that is the matter of model selection. Here, we consider a Bayesian model selection criterion, namely, a deviance information criterion (DIC) proposed by Spiegellhalter et al. (2002). It can be used with informative, uninformative, and improper priors.

DIC is defined as, $DIC = \bar{D}(\theta) + p_D$, where θ is the vector of all parameters involved in the model, p_D is the estimated effective number of parameters in the posterior distribution given by $p_D = \bar{D}(\theta) - D(\bar{\theta})$, and $\bar{D}(\theta)$ is the posterior mean of the deviance and $D(\bar{\theta})$ is the deviance evaluated at the posterior mean of parameters. Mathematically, $D(\theta) =$

$-2\log f(y|\theta)$ where $f(y|\theta)$ is the likelihood. DIC can be easily obtained in WinBUGS.

It had been observed that this definition proposed by Spiegellhalter et al. (2002) can hardly be utilized due to that the R2WinBUGS program is not able to call the deviance function because of WinBUGS called from the outside of R2WinBUGS (Sturtz et al., 2005). Actually, it is available in R2WinBUGS program now. Instead, Sturtz et al. (2005) adopted the definition of p_D introduced by Gelman et al. (2003, Section 6.7) which is defined as the half of the posterior variance of the Bayesian deviance. In other words, $DIC = \bar{D}(\theta) + \frac{1}{2}\hat{\text{Var}}(D(\theta))$. An additional advantage of choosing this definition is that $p_D > 0$ even when $\bar{D}(\theta) < D(\bar{\theta})$.

Gelman et al. (2003) noted that both definitions of p_D can be derived from the asymptotic χ^2 distribution of the deviance relative to its minimum. A model with the smallest DIC will be chosen. Since both DIC's can be easily obtained by using WinBUGS, we will compare their performance in later simulation studies.

2.3 Simulation Study

To study the performance of proposed methods, we conduct a series of simulation studies covering different missing mechanisms. Data are simulated mimicking the study described in Section 1.1. The full data are generated from the following model:

$$\begin{aligned} Y_{ij} &\sim Ber(p_{ij}) \\ p_{ij} &= \left[1 + \exp\{-X_i^T\beta\}\right]^{-1} \end{aligned} \tag{2.19}$$

where $\beta^T = (\beta_0, \ldots, \beta_6) = (-7.5, -0.47, 0.5, 0.7, 0.04, 1.05, 1.82)$ are true values of parameters for covariates $X_i^T = (1, \text{Tscore}_i, \text{year}_i, \text{trt}_i, \text{age}_i, \text{BFS}_i)^T$, $i = 1, \ldots, 600$ index for subjects, $j = 1, 2, 3$ the time points. The values of Tscore and age for each individual are generated from $\mathcal{N}(-2.7, 0.5^2)$ and $\mathcal{N}(73, 5.4^2)$, respectively. The values of $\text{year}_i = 1, 2$, or 3, $\text{trt}_i = 1$ or 0, and $\text{BFS}_i = 1, 2$, or 3 are generated with equal probabilities. Then, a surrogate variable, say $W_{ij} \sim N(S_i^T\beta^*, 1)$, where $\beta^{*T} = (-5.25, -0.4, 0.35, 0.02, 0.63, 1.27, 0)$, and $S_i^T = (\text{intercept}_i, \text{Tscore}_i, \text{trt}_i, \text{age}_i, \text{bfs1}_i I(\text{BFS}_i = 1), \text{bfs2}_i I(\text{BFS}_i = 2), \text{year}_i)^T$. Then, according to the order of W_{ij}'s, partition them into 3 groups, say Z_{ij} . For

example, subjects with the values of W_{ij}'s within the first 1/3 of the ordered W_{ij}'s are allocated into the first group, or $Z_{ij} = 1$.

After data are simulated, we fit 9 models to each dataset. These model include the full data analysis (FA), last observation carried forward (LOCF), pseudo-imputation (PI) methods with 3 patterns (PI3), PI with 6 patterns (PI6), PI and with 12 patterns (PI12), the joint modeling approach with 3 patterns (JM3), JM with 6 patterns (JM6), and JM with 12 patterns (JM12), the shared-parameter model using the subject-specific random effect as the link between Y-model and R-model. For the joint modeling approach, u is modeled as to have the bivariate normal distribution with a mean vector of 0's and a variance-covariance Σ_u. The prior distribution for both β and α is assumed to be $N(0, 10^{-4})$ and for the inverse of Σ_u we use the Wishart distribution with the 2-dimensional identity matrix as the scale matrix and the degree of freedom of 2. For each scenario considered, 500 data sets are generated. Models for generating missing indicators are described in the following sections.

2.3.1 Results based on Scenario 1

In the first scenario, the missing indicator R_{ij} are generated using the following model:

$$\begin{aligned} R_{ij} &\sim \text{Ber}(\pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \alpha_0 + \alpha_1 * \text{trt}_i + \alpha_2 * \text{year}_i + Z_{ij}u_i \end{aligned} \tag{2.20}$$

where trt_i, a subvector of X_i, denotes the treatment assignment, $(\alpha_0, \alpha_1, \alpha_2) = (-3.3, -0.5, 0.1)$, and $u_i \sim \mathcal{N}(0, 1.75^2)$. On average, 16% of data were found to be missing.

A portion of one simulated data set based on the first scenario is given in Table 2.1. As shown, some subjects fail to comply after the first year, some after the second year, and some complete the study. Figure 2.1 summarizes the results for the fixed effects in the Y-model from the FA, LOCF, PI3 and JM3 approaches. The dash lines indicate true values of parameters. The LOCF method produces substantially biased estimates for almost every fixed effect. On the other hand, the pseudo-imputation method and the joint-modeling approach provide remarkably accurate estimates similar to the the ideal

Table 2.1: Proportion of data simulated in the first scenario

Subject	Trt	Tscore	age	year	BFS	u	y	r	y_{new}
110	1	−2.9	61.1	1	3	0.50	0	1	0
110	1	−2.9	61.1	2	3	0.50	0	0	NA
110	1	−2.9	61.1	3	3	0.50	0	0	NA
228	2	−1.6	61.6	1	1	−0.12	0	1	0
228	2	−1.6	61.6	2	1	−0.12	0	1	0
228	2	−1.6	61.6	3	1	−0.12	0	0	NA
173	2	−2.2	59.7	1	2	−0.50	1	1	1
173	2	−2.2	59.7	2	2	−0.50	1	1	1
173	2	−2.2	59.7	3	2	−0.50	1	1	1
471	1	−2.3	60.1	1	1	−0.12	0	1	0
471	1	−2.3	60.1	2	1	−0.12	0	1	0
471	1	−2.3	60.1	3	1	−0.12	1	1	1

full-data analysis, when the number of patterns is correctly specified. Also, estimates obtained from the pseudo-imputation method generally have larger variations than those from the joint-modeling approach.

Table 2.2 shows biases and SE's of estimates from the PI method with different number of missing patterns. The biases and SE's appear to be larger as the number of patterns increases except that the estimate of the treatment effect has the least bias when the number of patterns assumed to be 6. It implies that the PI method might be very sensitive to the number of patterns chosen. When the number of patterns is specified incorrectly, results can be very misleading, even if a right surrogate variable is used and the missingness indeed depends on the random pattern variable controlled by that surrogate variable.

Comparison of the joint-modeling approach with a different number of patterns and the SP model is given in Figure 2.2. The estimates appear to be approximately unbiased despite the fact that JM model is based on the wrong number of patterns. The SP model provides acceptable results, compared with the JM model. The estimates based on the SP model are generally more biased than the PI approach and the JM approach

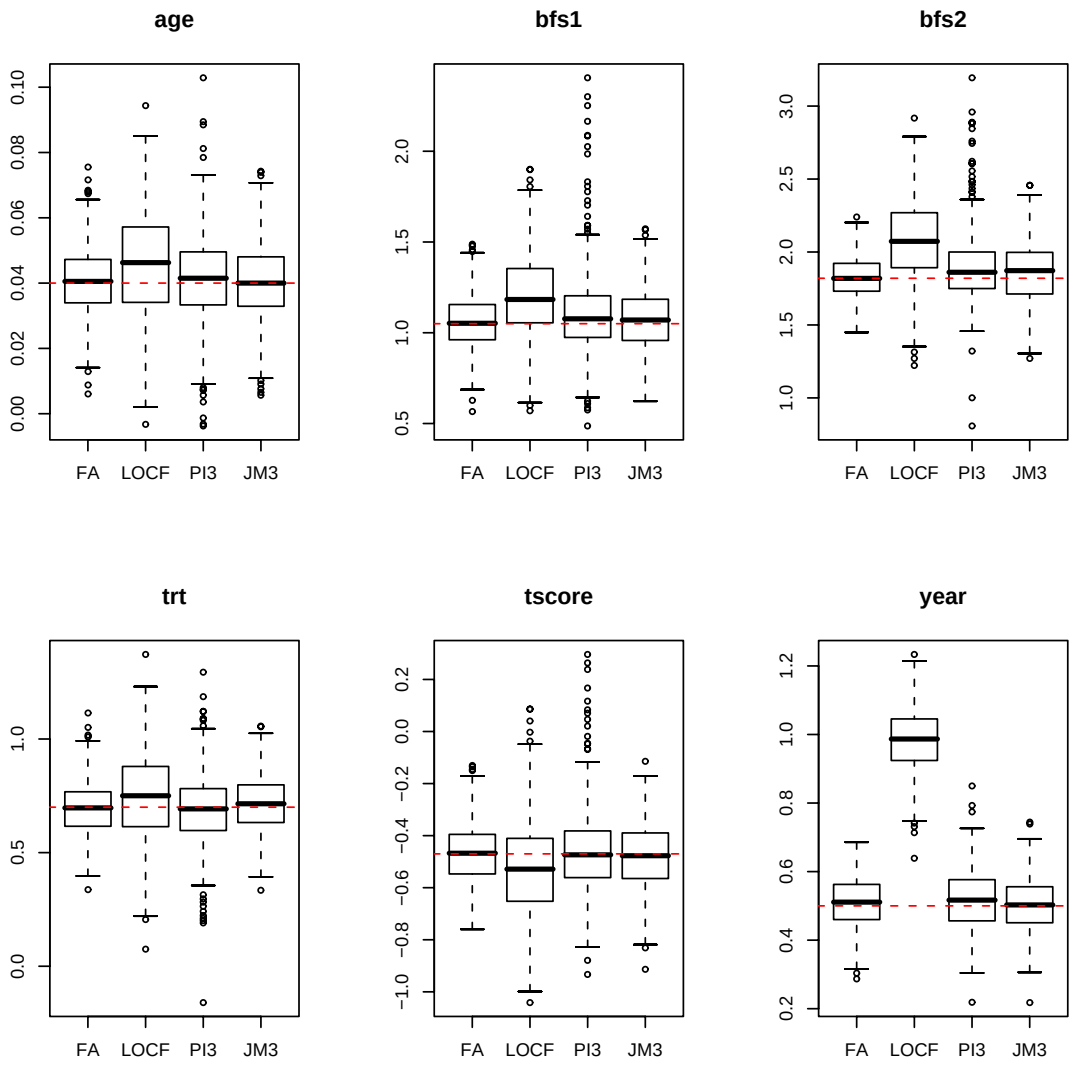

Figure 2.1: Parameter estimates from 4 methods under Scenario 1. FA, the full-data analysis. LOCF, the last observation carried forward method. PI3, the pseudo-imputation method with 3 patterns. JM3, the joint-modeling approach with 3 patterns.

Table 2.2: Comparison of Bias's and SE's of estimates for the scenario 1. FA: the full-data analysis. PI3, PI6, and PI12: the pseudo-imputation methods with 3, 6, and 12 patterns, respectively.

	FA	PI3	PI6	PI12
Parameter	Bias(SE)	Bias(SE)	Bias(SE)	Bias(SE)
age	0.001(0.01)	0.001(0.01)	0.002(0.02)	0.011(0.06)
bfs1	0.01(0.15)	0.06(0.23)	0.14(0.50)	0.52(2.18)
bfs2	0.01(0.14)	0.07(0.24)	0.17(0.46)	0.63(2.40)
year	0.01(0.07)	0.02(0.09)	0.03(0.09)	0.06(0.44)
trt	0.006(0.12)	0.011(0.16)	0.002(0.033)	0.078(1.01)
tscore	-4.1E-05(0.11)	-6.5E-03(0.16)	-9.5E-03(0.25)	-4.6E-02(0.72)

but the biases are not substantial. It is expected as the SP model can be considered as a special case of the JM approach in a way that each subject is a pattern. We now turn to model selection, using two definitions of DIC as described in Section 2.2.3. Results are presented in Table 2.3. Both versions of DIC's pick the 3-pattern model most frequently. Certainly, DIC's perform quite well in picking up the true model with high precision (about 90%). The SP model is never preferred by DIC, mainly because it does not improve the precision of estimates but involves more random parameters than needed to be estimated. In addition, the joint-modeling approach is able to accurately estimate fixed effects in the R-model, which other models do not provide. Results from the JM approach are presented in Table 2.4. Obviously, the joint-model approach provides accurate estimates regardless how many number of patterns are specified in the model and moreover the DIC picks up the true model, subject to about 10% type I error.

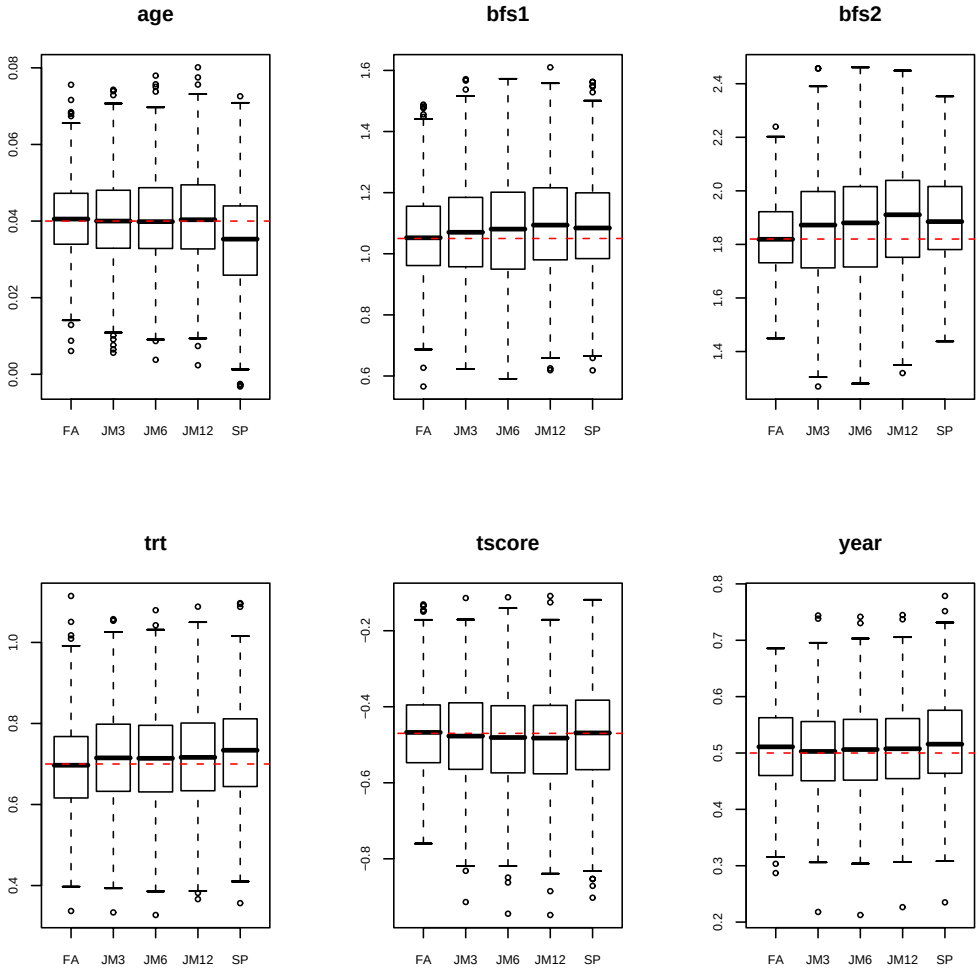

Figure 2.2: Parameter estimates of fixed effects in the Y-model from joint-modeling approach with different patterns under Scenario 1.

Table 2.3: Comparison of two versions of DIC's. $DIC1$ (Spiegellhalter et al., 2002) and $DIC2$ (Gelman et al., 2003).

	JM3	JM6	JM12	SP
$DIC1$	89.4%	9.2%	1.4%	0%
$DIC2$	89.4%	9.0%	1.6%	0%

Table 2.4: Comparison of estimates and their SE's for fixed effects in the R-model from joint-modeling approach based on the scenario 1

	JM3	JM6	JM12
Parameter (Truth)	Estimate(SE)	Estimate(SE)	Estimate(SE)
intercept(-3.30)	$-3.28(1.13)$	$-3.27(1.08)$	$-3.35(1.08)$
year (0.10)	0.10(0.13)	0.10(0.13)	0.10(0.13)
trt (0.50)	0.51(0.22)	0.51(0.22)	0.51(0.22)

2.3.2 Results Based on Scenario 2

In the second scenario, R_{ij}'s are allowed to depend on responses and treatments and are generated by the following model,

$$\begin{aligned} R_{ij} &\sim Ber(\pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \alpha_0 + \alpha_1 \text{trt}_i + \alpha_2 Y_{ij} + \alpha_3 \text{year}_i \end{aligned} \tag{2.21}$$

where $(\alpha_0, \alpha_1, \alpha_2) = (-3, 0.8, 1.2, 0.1)$. Again on the average about 16% of data are missing. Notice that the true model generating the data in this scenario does not belong to the class of models considered by JM and PI approaches. Moreover, unlike in the first scenario, the true number of missing patterns is no longer known.

The simulation results based on Scenario 2 are shown in Figure 2.3, in which 4 fixed effects are presented and there are no big difference among all models for the rest of fixed effects. As expected, the LOCF approach produces highly biased estimates. Comparing the PI approach and the JM method, we are able to see the same trend as in first scenario. Estimates of the PI approach fluctuate as the number of patterns varies. The JM approach provides stable estimates regardless of the number of patterns posited in the model. Overall, estimates obtained from the PI method have more variations than those from the JM approach. The SP model has the similar performance as the JM model except that its parameter estimate for the variable"age" is more biased. As expected, estimates of the treatment effects are moderately biased. However, the JM approach produces estimates with less biases than the PI method. On average, the JM

approach reduces the bias by about 31% as compared to the estimate based on the full-data analysis. Except the treatment effect, the JM approach appears to provide unbiased estimates for other fixed effects despite the fact that the assumed model is not correct. Hence, the JM approach is preferred over the PI method. With respect to model selection, again both definitions of DIC's have the similar preference. On average, the JM approach with 3 patterns are selected 71.6%, 18.8% for 6 patterns, and 9.6% for 12 patterns. Therefore, the JM approach with 3 patterns might be the best model for this type of data set.

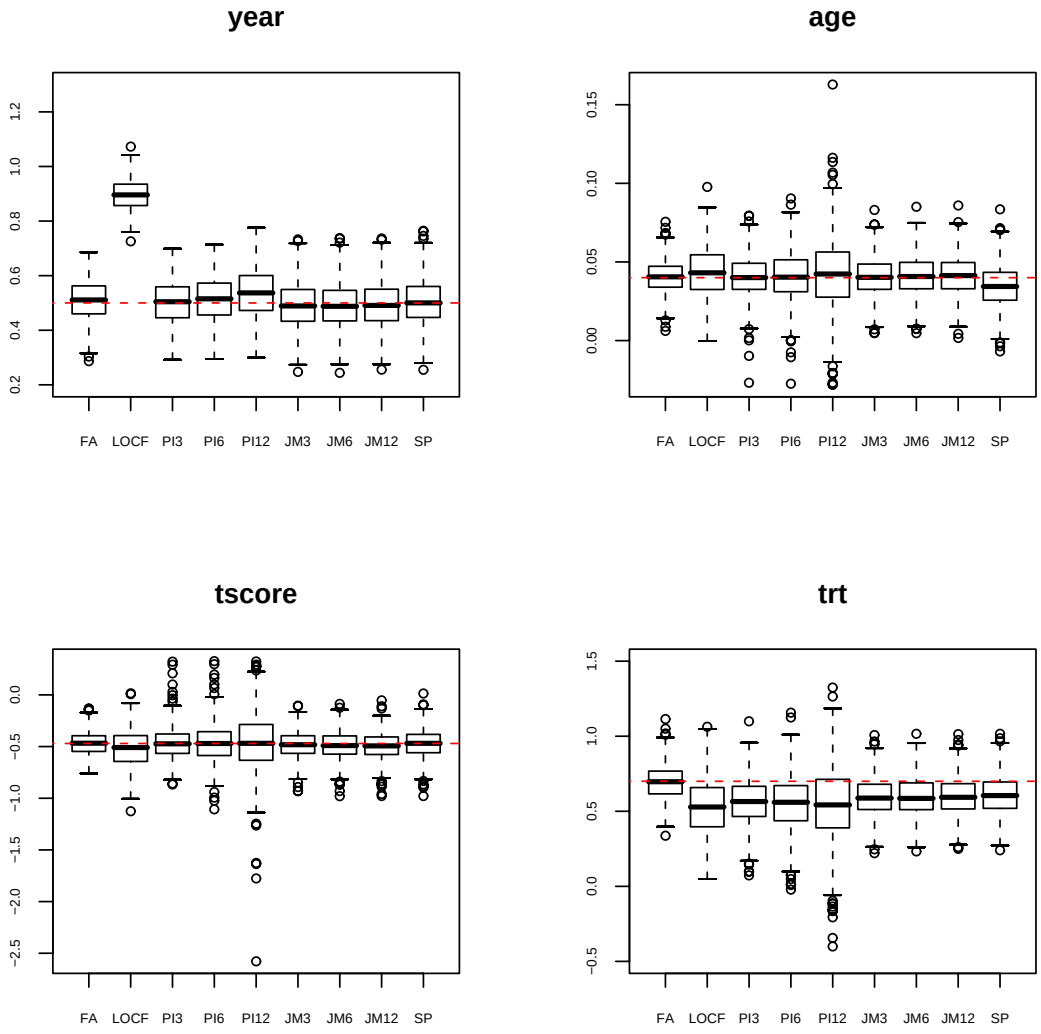

Figure 2.3: Partial parameter estimates under Scenario 2.

Table 2.5: Results of simulation for the joint modeling with the misspecification of the distribution for random patterns.

Parameter	Bias	SE	MSE
age	0.001	0.011	0.02
bfs1	0.047	0.176	0.03
bfs2	0.083	0.203	0.05
year	0.008	0.081	0.01
trt	0.014	0.127	0.02
tscore	-0.011	0.127	0.02

2.3.3 Sensitivity Analysis for the JM Approach

Notice that it is assumed the random pattern u has the bivariate normal distribution in the JM approach and $f(R|X,u)$ follows a logistic-regression model. In this section, sensitivity of the model will be assessed on these two respects.

First, we use the same model as the one in Scenario 1 with three patterns but u_i is generated by a mixture of two normal distributions given by $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ with weights 2/3 and 1/3, respectively, where $\mu_1 = (-1.5, 0)^T$, $\mu_2 = (1.5, 0)^T$, $\Sigma_1 = 1$, and $\Sigma_2 = 3.06$

For this scenario about 11.5% of data are missing. After data are generated, we fit the JM with 3 patterns. The biases, standard errors (SE's), and MSE's of parameter estimates are reported in Table 2.5. None of these biases is substantially large ($\leq 5\%$). Both SE's and MSE's are fairly small. Hence, the JM approach provides consistent estimates of fixed effects in the Y-model even though the distribution of random effects is misspecified.

Second, we use the data simulated in Scenarios 1 with 3 patterns and fit the same Y-model but a probit regression for the R-model instead of the logistical regression used for data generation. In other words, the assumed R-model is not correct. Here, we only consider 3 patterns in the simulation. Results are summarized in Figure 2.4. Apparently, both models provide almost identical estimates although variations of estimates provided

by the JM with a probit regression are slightly larger. Therefore, the JM generally delivers robust estimates even when the model for the missing indicators are not correctly postulated.

2.4 Conclusions

We proposed two approaches, the pseudo-imputation method and the joint-modeling method, that are general enough in the sense that both methods can be applied to continuous and discrete outcomes. Here, we assumed that a surrogate variable can be correctly identified. Based on our simulation studies, we see that the LOCF approach provides biased estimates even when the missingness is not informative. The PI method performed well only if the missingness depends on missing patterns and the underlying number of patterns is correctly specified. When the number of patterns is incorrectly defined, the PI method produces biased estimates. On the other hand, the JM method gives approximately unbiased estimates as long as the the missing depends on patterns even when the number of patterns is not correctly specified. The SP model performs similarly in overall as the JM model although it seems to produce comparatively more based estimates. In cases of the missingness depending on the unobserved responses themselves, all methods provide biased estimates but the JM based method reduces the bias substantially for the estimate of the treatment effect. In terms of model selection, the DIC criteria picks the right model about 90% for the joint-modeling approach. Furthermore, the JM approach is not sensitive to mild model violations. Overall, the joint model approach is preferred.

In the cases of absence of a surrogate variable, patterns may be defined according to available covariates but extreme care must be given since misspecification of patterns may introduce serious bias and produce misleading conclusions. This issue is explained in the next chapter.

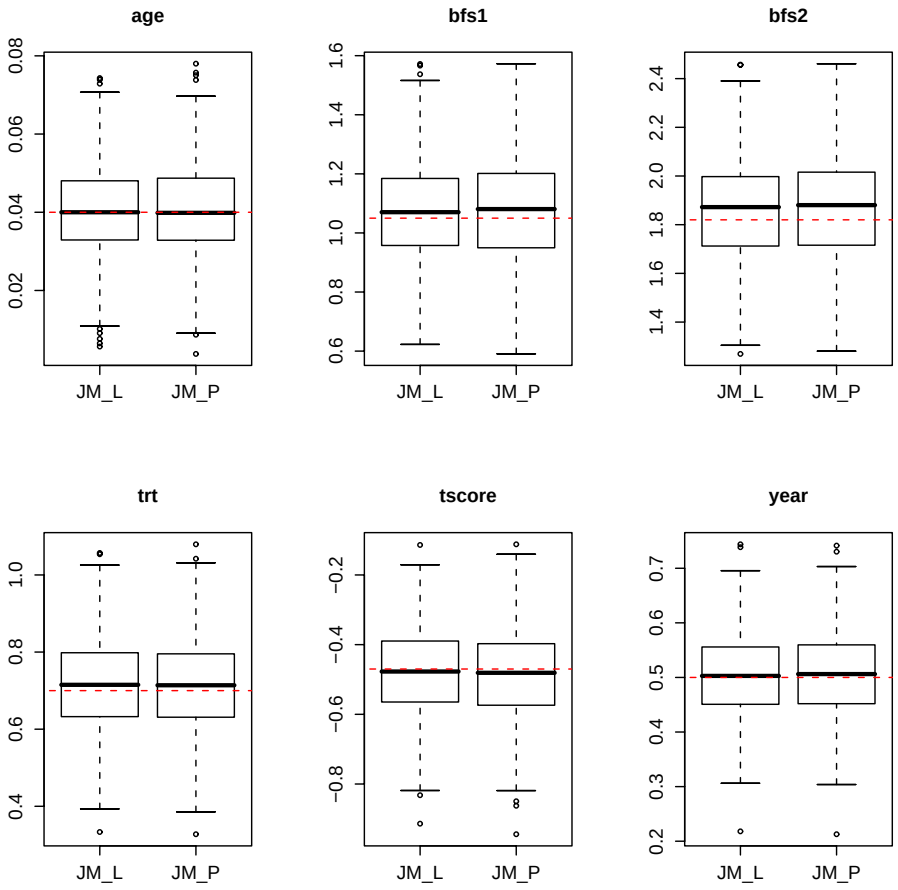

Figure 2.4: Comparison of results between JM_L and JM_P with 3 patterns under Scenario 1. JM_L, the joint model with the use of the logistic regression in the R-model; JM_P, the use of the probit regression in the R-model

Chapter 3

Modeling NIM Data using GAM

A parametric generalized linear model may not be appropriate especially when some measurements are non-ignorable missing. Also a specific parametric structure can not be validated easily. Such parametric structures may lead to possible bias when the underlying model for the full data is nonlinear. Particularly, the generalized linear model used in the first step for the PI approach may not produce reliable estimates of the predictive means that will be used in latter steps. Therefore, final estimates might be biased and inference will be misleading. Hence, instead of the generalized linear model, we consider the use of the generalized additive model (GAM) proposed by Hastie and Tibshirani (1990). As it is known, parameters in GAM can be difficult to interpret. However, parameters need not to be interpretable in the first step of the PI approach and hence we can still apply GAM to obtain the predictive means for latter use. Moreover, in situations that only the treatment effect or effects of other discrete covariates may be of primary interest, estimations in terms of odds ratios can be obtained without much trouble. A brief description about generalized additive models and model fitting techniques is given in Section 3.1. Section 3.2 provides simulation results comparing a generalized linear model and a generalized additive model within the PI method. Possibilities of fitting the GAM within the JM approach will be illustrated in Section 3.3.

3.1 Model Description

Many nonparametric methods do not perform well when there is a large number of independent variables in the model. The sparseness of data in this setting inflates the variance of the estimates. The problem of rapidly increasing variance for increasing dimensionality is sometimes referred to as the curse of dimensionality . To overcome this difficulty, Stone (1985) proposed additive models. These models estimate an additive approximation to the multivariate regression function. The benefits of an additive approximation are at least twofold. First, since each of the individual additive terms is estimated using a univariate smoother, the curse of dimensionality is avoided, at the cost of not being able to approximate universally. Second, estimates of the individual terms explain how the dependent variable changes with the corresponding independent variables.

Hastie and Tibshirani (1990) proposed generalized additive models that extend the traditional additive models (Stone, 1985) to accommodate more distribution functions. These models assume that the mean of the dependent variable depends on an additive predictor through a nonlinear link function. Generalized additive models permit the response probability distribution to be any member of the exponential family of distributions. Many widely used statistical models belong to this general class, including additive models for Gaussian data, nonparametric logistic models for binary data, and nonparametric log-linear models for Poisson data.

The generalized additive models can be expressed as

$$E(Y_i|X_i) = g(s_0 + \sum_{i=1}^{p} s_i(X_i)) \tag{3.1}$$

where $g(\cdot)$ is a known link function, $s_i(\cdot)$'s are smooth functions, and p is the number of the predictor variables. The GAM can also be used to incorporate partial linear models (Hastie and Tibshirani, 1986). This is achieved by replacing some smooth functions with linear functions. Thus, parameters in linear functions in the GAM can be interpretable, which makes it possible to draw inferences for some covariates if they are of primary interest.

The estimating procedure for generalized additive models consists of two loops. The local scoring algorithm is used in the outer loop and a weighted backfitting algorithm for the inner loop. The local scoring algorithm uses a modified form of adjusted dependent variable regression, as described for generalized linear models in McCullagh and Nelder (1989), with the additive predictor taking the role of the linear predictor (Hastie and Tibshirani, 1986). The backfitting algorithm (Friedman and Stuetzle, 1981; Hastie and Tibshirani, 1986) is a general algorithm for an additive model using any regression-type fitting mechanisms. This algorithm always converges for Gaussian distribution (Hastie and Tibshirani, 1986). However, for other distributions, numerical instabilities with weights may cause convergence problems. Even when the algorithm converges, the individual functions need not be unique, since dependence among the covariates can lead to more than one representation for the same fitted surface. A weighted backfitting algorithm has the same form as for the un-weighted case, except that the smoothers are weighted according to distribution functions. Both PROC GAM in SAS/STAT and GAM in R package of MGCV can be used to fit these models.

3.2 Simulation Studies

The objective of the first simulation is to compare the performance of different models used at the first step in the PI method. The model used in the Chapter 2 is a generalized linear model (GLM) and we also use a generalized additive model (GAM) in this simulation. For simplicity, these two strategies are referred to as the PI-GLM method and the PI-GAM method. At the second stage, we will still use a generalized linear model as described in Chapter 2. Data are generated by using the same models described in Chapter 2 under Scenario 1. The underlying Y-model follows a logistic regression and the missing indicator depends on missing patterns. One thousand Monte Carlo samples are simulated and the sample size for each sample is 1800.

Simulation results are summarized in Figures 3.1 − 3.3. When the configuration of patterns is correctly known, both models yield precise estimates. When patterns are misspecified, the PI-GAM method provides slightly more biased estimates than the PI-GLM method for most covariates, but the estimates are not significantly more biased

than estimates from the PI-GLM method. Thus, the methods appear to be approximately unbiased. Also from the boxplots in Figure 3.1, we see that the sampling distributions of the parameter estimates are very similar. Overall, the PI-GLM method and the PI-GAM method have the similar performance in terms of parameter estimation. Hence, there is no significant loss of efficiency to fit a GAM at the first step of the PI method even when the underlying Y-model is a GLM.

In the second simulation study, the full data are generated from the following model

$$\begin{aligned} Y_i &\sim Ber(p_i) \\ \text{logit}(p_i) &= \beta_0 + \beta_1 \text{trt}_i + f(\text{age}_i) \end{aligned} \tag{3.2}$$

where $\beta^T = (\beta_0, \beta_1) = (-1.5, 0.3)$ are true values of parameters and $f(\text{age}) = 0.001(\text{age}-20)(\text{age}-30)(\text{age}-40)I(\text{age} < 40) + 0.1(\text{age}-40)I(\text{age} > 40)$. The values of age for each individual are generated from $Uniform(18, 58)$. The surrogate variable $W_i = 0.7 \times \text{logit}(p_i) + e_i$, where $e_i \sim \mathcal{N}(0, 0.8^2)$. Patterns are determined in the same way as in Chapter 2.

Notice that the model used to generate data in this simulation is determined according to the real data from the AIDS study which we will analyze in the last chapter. To simplify the simulation, we only include two independent variables, treatment (trt) and age in the model. The treatment variable is binary and the age variable is continuous. The curve of the spline function versus the corresponding covariate is given in Figure 3.4. Obviously, the relation with age is non-linear.

Then, the missing indicator R_i is generated using the following model:

$$\begin{aligned} R_{ij} &\sim \text{Ber}(\pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \alpha_0 + \alpha_1 * \text{trt}_i + Z_i u_i \end{aligned} \tag{3.3}$$

where trt_i, a subvector of X_i, denotes the treatment assignment, $(\alpha_0, \alpha_1) = (-3, 1.7)$, and $u_i \sim \mathcal{N}(0, 0.75^2)$. On average, 15% of data were found to be missing.

As in the first simulation, a total of one thousand Monte Carlo samples are simulated with 1800 subjects in each sample. After data generated, we use both the PI-GLM

method and the PI-GAM method to fit to each dataset. Difference from the first simulation is that we also fit the GLM and the GAM accordingly at the second stage. A summary of results is presented in Figure 3.5. It is clear from Figure 3.5 that the PI-GAM method provides more accurate estimates compared to the PI-GLM method.

In summary, the PI method with the GAM has similar performance even when the underlying Y-model follows a generalized linear model. Moreover, it yields precise estimates when the underlying Y-model is nonlinear. Hence, the PI-GAM method is preferred universally. In practice, the underlying distribution is unknown and missing data adds more uncertainty. A good strategy would be to obtain the predictive mean by fitting a GAM at the first stage of the PI method. Then plot the predictive mean versus covariates to discover the potential data structure and fit the model accordingly. When the treatment effect is of primary interest, we can fit the GAM within the PI method at both stages.

3.3 The JM Approach with GAM

It has been shown via simulation studies presented in Chapter 2 that the JM approach performs better than other approaches. As the treatment effect is of primary interest, one may prefer to adopt the JM approach with the GAM in the Y-model (Equation 2.8) over the PI method. However, it is generally difficult, at least for practical statisticians, to fit smooth functions within Bayesian framework since there is no existing program or procedure available to resolve this task in ready-made software.

Speed (1991) developed connection between nonparametric regression and mixed model. It allows one with some programming experience to code smoothing functions in the Bayesian way. Some earlier work made connections of this kind but did not mention the mixed models (Wahba, 1978; Wecker and Ansley, 1983; and Green, 1985) Many types of splines can be obtained with different choices of knots and penalty functions. For the purpose of simplifying programming, we primarily focus on penalized splines, also called P-splines, a terminology introduced by Eilers and Marx (1996) and Marx and Eilers (1998), which also coined with the names of pseudosplines and low-rank smoothers (Hastie, 1996; Ruppert and Carroll, 2000). We will use P-splines throughout afterwards.

Penalized splines use relatively small and fixed number of knots but do not sacrifice

much degradation in the quality of the fit and the specification of knots is very much a minor detail(Wand, 2003). Ruppert (2002) also noted that the number of knots is not a crucial parameter and typically 5-20 is large enough. A surge of research on P-splines has appeared, including contribution on mixed-model representations of P-splines (Brumback et al., 1999) and other work (Berry et al., 2002; Durban and Currie, 2003; Lang and Brezger, 2004). Crainiceanu et al. (2005) suggested to use P-splines which tend to have good numerical properties and provide much smaller posterior correlation of parameters than other basis.

For illustration purpose, we consider the data following Model 3.2. Then, the Y-model in the JM approach can be written as

$$g_1^{-1}(\mu_i) = \beta \text{trt}_i + f(\text{age}_i) + Z_i u_i \tag{3.4}$$

Then, the P-spline representation of $f(\text{age}_i)$ is

$$f(\text{age}_i) = \tau_0 + \tau_1 \text{age}_i + \sum_{k=1}^{K} v_k |\text{age}_i - \kappa_k|_+^3 \tag{3.5}$$

where $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ are fixed knots based on percentiles of the age variable.

We use the equivalence between penalized splines (P-splines) and mixed models for semiparametric modeling presented by Ruppert et al. (2003). Specially, the P-spline is equal to the best linear predictor and then Equation 3.4 can be written as

$$g_1^{-1}(\mu) = X\Upsilon + S_K v + Zu \tag{3.6}$$

where $\Upsilon = (\beta, \tau_0, \tau_1)^T$, $v = (v_1, \ldots, v_K)^T$, and S_K be the matrix with i^{th} row $S_{Ki} = (|\text{age}_i - \kappa_1|^3, \ldots, |\text{age}_i - \kappa_K|^3)$. $E(v) = 0$ and $\text{var}(v) = \lambda \Sigma_K^{-1}$, where the $(l,k)^{th}$ entry of Σ_K is $|\kappa_l - \kappa_k|^3$. Thus, the GAM becomes the generalized linear mixed model that can be easily fitted in Bayesian inferential perspective by placing priors on the model parameters.

We apply the approach described above to data in the second simulation in Section 3.2. We choose 15 knots according to percentiles of the age variable. From the Figure

3.6, the JM approach with GAM yields accurate estimate. Hence, the proposed modeling approach performs well and can be useful in application.

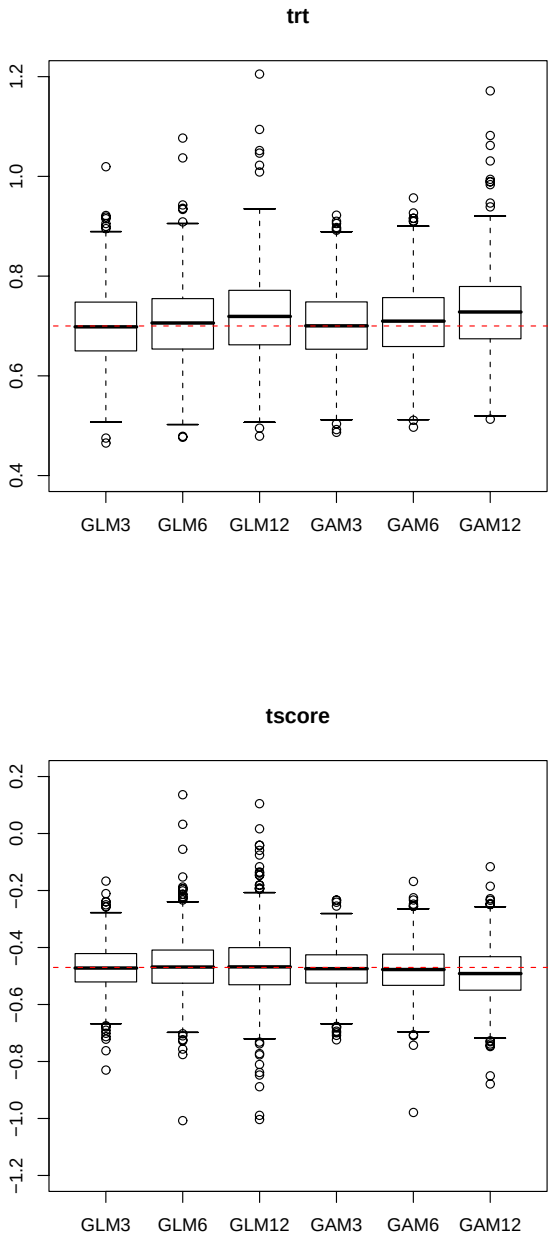

Figure 3.1: Parameter estimates for covariates "trt" (top) and "tscore" (bottom) in the first simulation. GLM: the generalized linear model and GAM: the generalized additive model at the first step of the PI method. 3, 6 and 12 indicated the number of patterns specified for both methods. The dashed horizonal line indicates the true value of β.

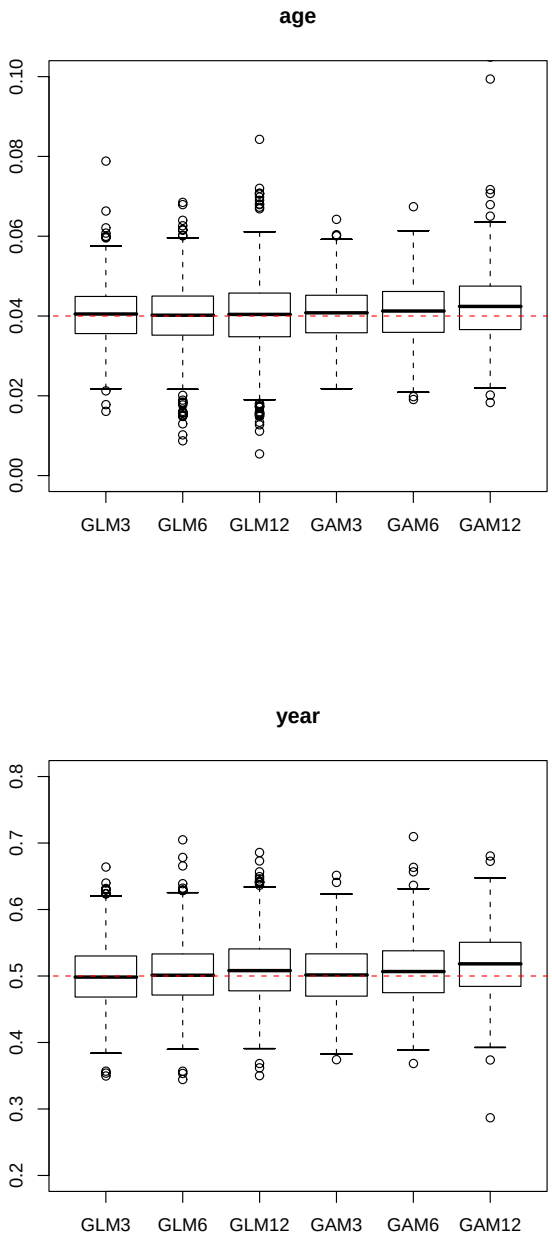

Figure 3.2: Parameter estimates for covariates "age" (top) and "year" (bottom) in the first simulation. GLM: the generalized linear model and GAM: the generalized additive model at the first step of the PI method. 3, 6 and 12 indicated the number of patterns specified for both methods. The dashed horizonal line indicates the true value of β.

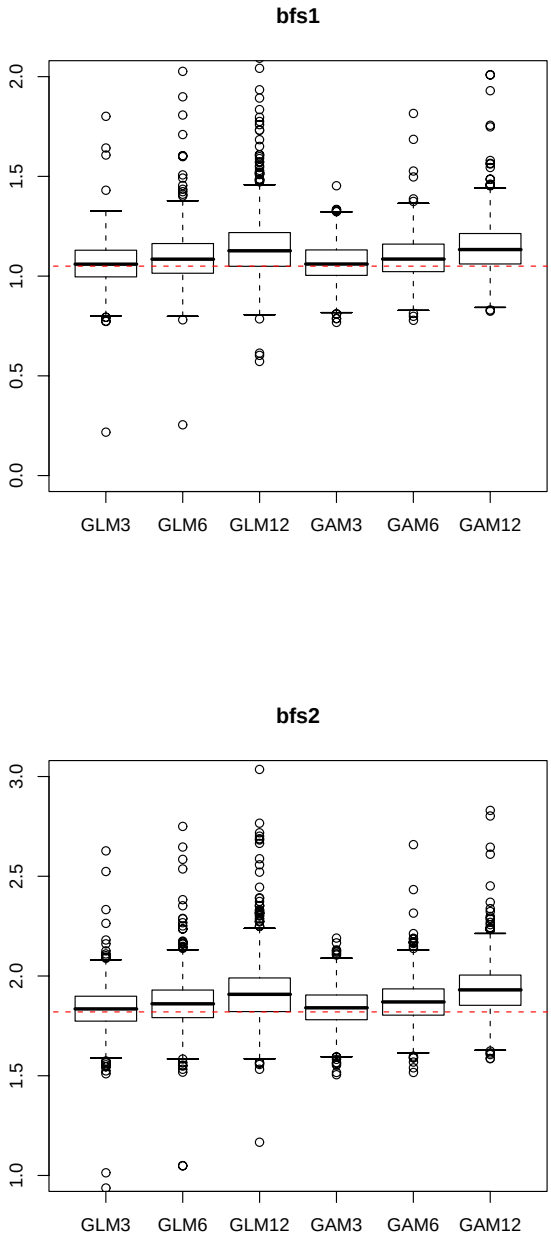

Figure 3.3: Parameter estimates for covariates "bfs1" (top) and "bfs2" (bottom) in the first simulation. GLM: the generalized linear model and GAM: the generalized additive model at the first step of the PI method. 3, 6 and 12 indicated the number of patterns specified for both methods. The dashed horizonal line indicates the true value of β.

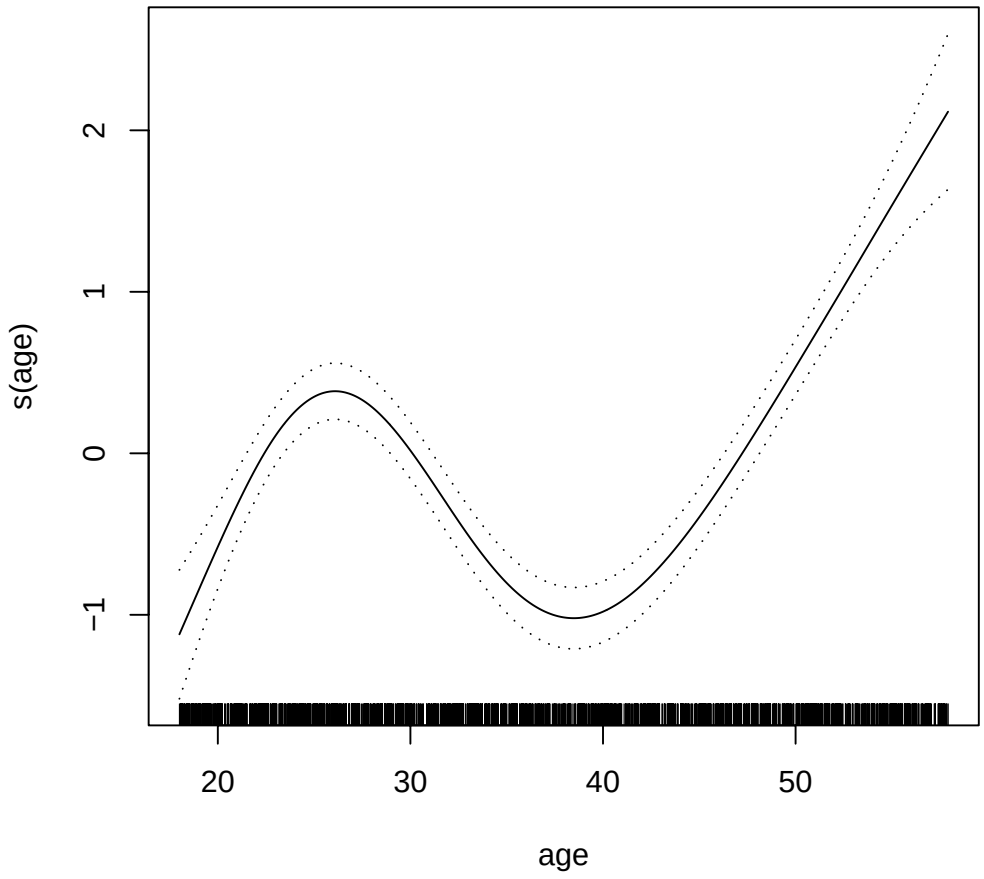

Figure 3.4: Smooth estimate of age

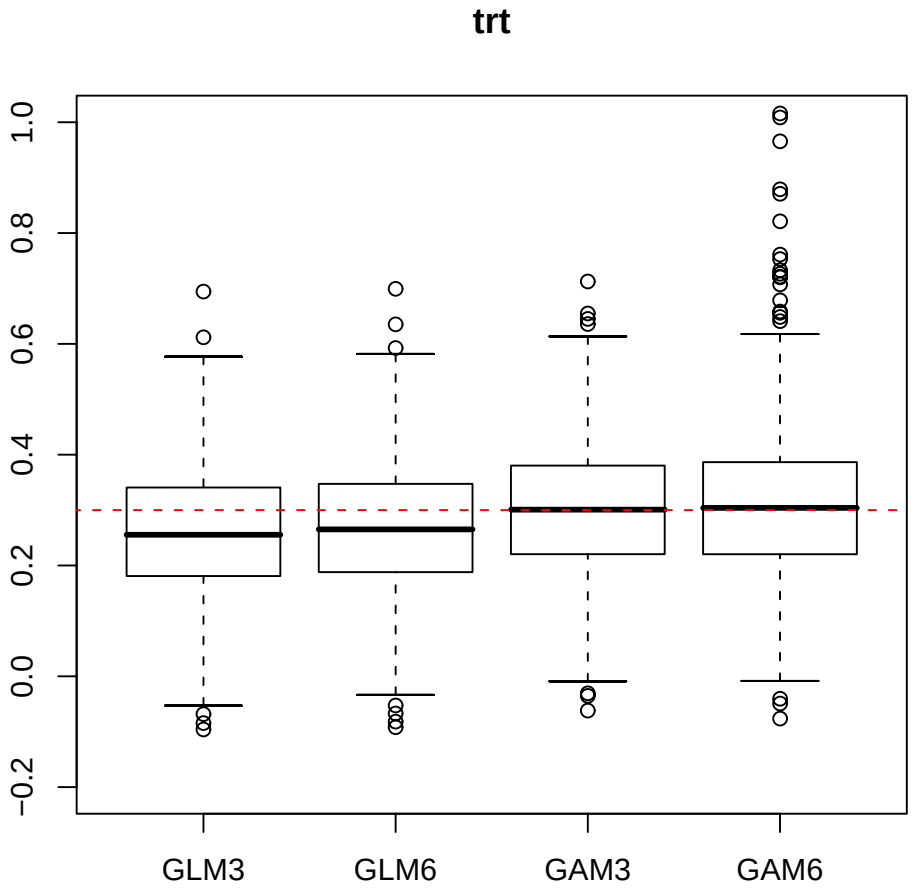

Figure 3.5: Parameter estimates of treatment effects from the PI approach with different modeling techniques. GLM: the generalized linear model and GAM: the generalized additive model at the first step of the PI method. 3 and 6 indicated the number of patterns specified for both methods. The data generation model is a GAM under Scenario 1. The dashed horizonal line indicates the true value of β.

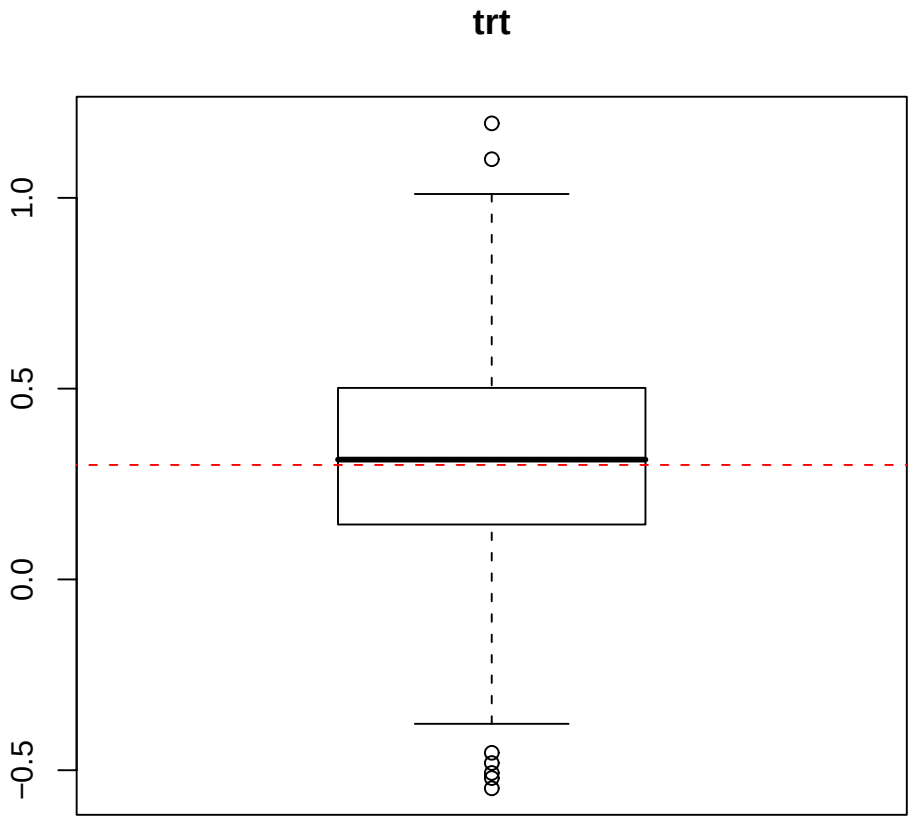

Figure 3.6: Parameter estimate of the treatment effect from the JM approach with GAM when data generated by a GAM under Scenario 1. The dashed horizonal line indicates the true value of β.

Chapter 4

Application: CPCRA Aids Study

The study described in the Section 1.1 is still ongoing and data could not be obtained during the research work. Alternatively, we illustrate our methods using is a dataset, called CPCRA Aids Study, whose data structure is very similar to that described in Section 1.1. We fit all the methods developed in Chapters 2 and 3 to this dataset.

4.1 Study Description

The CPCRA 010 Clinical Trial is a randomized, prospective, double-blind study comparing fluconazole vs. placebo (trt) for primary and secondary prophylaxis of mucosal candidiasis in HIV-infected women with 323 patients enrolled (Schuman et al., 1997). Binary indicators for a positive vaginal culture for the candidiasis fungus are used as the responses measured every 3 months. In total, there are 5 visit times including the first visit for baseline measure. Many baseline variables have been measured, including age, antiretroviral use (aruse), diagnosis of vaginal candidasis at baseline (canvbl), diagnosis of vaginal oral candidasis at baseline (canobl), CD4 counts at baseline (cd4bl), intravenous drug use (idu), progression of disease at baseline (podbl), and race (white/non-white). The covariate vector in the analysis is $X = (\text{age}, \text{aruse}, \text{canvbl}, \text{canobl}, \text{cd4bl}, \text{idu}, \text{podbl}, \text{race}, \text{trt})^T$. The primary interest is to compare effects of two treatment arms on aids occurrences. Data structure of this study is very similar to the one we described in Section 1.1.

4.2 Data Analysis Assuming a GLM

Just like other longitudinal studies, a large proportion of data were missing in this study as well. Briefly, the percentage of missing outcomes at the each visit time is $0.31\%, 19.20\%, 26.32\%$, 33.75%, and 39.63%. They are missed both intermittently and monotonic. It is well known that CD4 counts often serve as the surrogate variable in AIDS trial. At first, patients are partitioned into 4 groups according to their CD4 counts at the baseline (0-100, 101-200, 201-300, and > 300). Group frequencies are 28.2%, 26.6%, 26.9% and 18.3%. To check whether the groups can serve as the random pattern variable, we need to assess the dependence of both responses and missing indicators on them. It is commonly believed that the likelihood of HIV occurrence is highly associated with CD4 counts. Thus, the first dependence holds. Furthermore, the percentages of missing data for groups 1-4 are 34.9%, 21.86%, 17.70%, and 18.64%, respectively. In other words, subjects with lower cd4 counts appear to be more vulnerable to be missing than those with higher cd4 counts. Hence, missingness seems to depend on CD4 counts. Therefore, groups based on the CD4 counts can be treated as the random pattern variable in our proposed methods which functions as the bridge between Y-model and R-model.

For jointly modeling Y and R, all covariates can be included in the Y-model. To determine variables included in the R-model, a multivariate logistic regression can be employed to perform the variable selection. We found that canobl, cd4bl, idu, trt, and time have relatively strong effect on the missingness. Hence, the vector of covariates included in the Y-model is (age, aruse, canvbl, canobl, cd4bl, idu, podbl, race, trt, time)^T and (canobl, cd4bl, idu, trt, time)^T for the R-model. Besides the joint model, we also consider three other competing models. They are a multivariate regression model by assuming that the mechanism of missing is completely at random (MCAR), the pseudo-imputation (PI) method, a pattern-mixture model with the multiple imputation technique (PMMI), and a shared parameter (SP) model. Both PI and PMMI are fitting the same model at the first step to obtain a predictive probability for each subject at each visit time. Then, the predictive probability is treated as the response to fit a model at the second step in PI whereas an imputation method is used to fill in the missing value in PMMI. After imputation, a complete dataset is formed and then re-fitted with

a multivariate logistic regression model to obtain marginal estimates of parameters.

Results are summarized in Table 4.1. Estimates from MCAR are far away from those obtained from other three methods. In terms of parameter estimates, the PMMI, PI, JM, and SP approaches perform similarly but the SP approach results into the largest standard error whereas the PMMI and JM approaches had the smallest standard errors. The variable, cd4bl, is significant under the PMMI and PI approaches but not using the JM approach. It is reasonable for values of cd4bl within patterns are more similar than between patterns. All methods detect the significant difference between treatments, which confirms the original finding in the primary analysis conducted by Schuman et al. (1997), although they did a survival analysis on time until occurrence of the first disease event. As the number of patterns increases, estimates from the PI approach are slightly varying and their standard errors increase. The overall parameter estimates (and standard errors) from the JM approach seem to be stable as the number of patterns varies. It is the same phenomena that we observed in the simulation studies presented in Chapter 2. When the DIC criterion is computed, the JM approach with 8 patterns (DIC = 3089) is preferred over the JM approach with 4 patterns (DIC = 3102) but the overall difference between corresponding parameter estimates is not substantial.

4.3 Data Analysis Assuming a GAM

In this section, methods discussed in Chapter 3 are applied: the PI method with GAM's at both steps (PI-GAM) and the JM approach with a GAM (JM-GAM). For the JM-GAM method, we only report results based on 5 knots for the smooth function. Other number of knots were tried but results were similar. Here, the age variable is modeled using the smoothing function and other variables are fitted using a linear form. To avoid repetitions, we assume 4 missing patterns in this section. After applying these two methods on the data set, it is found that there are not significant difference among estimates obtained from methods assuming generalized linear models or generalized additive models. Hence, only estimates of the treatment effect, the primary interest, are presented in Table 5.2. Clearly, the estimates are very similar.

Although methods with GAMs did not show advantage over methods with GLMs at

Table 4.1: Parameter estimates and their standard errors from analyzing aids data by utilizing various methods. * indicates the fixed effect is significant at the significance level $\alpha = 0.05$

	MCAR	PMMI4	PI4	PI8	JM4	JM8	SP
age	-0.01	-0.02*	-0.02*	-0.03*	-0.02*	-0.02*	-0.04*
(*s.e.*)	*0.01*	*0.01*	*0.01*	*0.01*	*0.01*	*0.01*	*0.01*
aruse	0.51	-0.05	-0.05	-0.02	-0.03	-0.04	-0.05
(*s.e.*)	*0.25*	*0.15*	*0.17*	*0.22*	*0.16*	*0.15*	*0.25*
canvbl	0.03	0.01	0.01	-0.01	0.01	-0.01	0.01
(*s.e.*)	*0.20*	*0.12*	*0.13*	*0.19*	*0.12*	*0.12*	*0.22*
canobl	-0.62	-0.256	-0.265*	-0.34	-0.253	-0.34*	-0.23
(*s.e.*)	*0.21*	*0.13*	*0.14*	*0.20*	*0.14*	*0.13*	*0.21*
cd4bl	-0.002	-0.001*	-0.001	-0.001	-0.0016	-0.0025	-0.002*
(*s.e.*)	*0.0008*	*0.0005*	*0.0007*	*0.0011*	*0.0011*	*0.0017*	*0.0001*
idu	-0.13	-0.22	-0.24	-0.26	-0.24	-0.28*	-0.36
(*s.e.*)	*0.21*	*0.12*	*0.13*	*0.19*	*0.13*	*0.13*	*0.21*
podbl	0.44	0.48*	0.51*	0.58*	0.51*	0.44*	0.55*
(*s.e.*)	*0.27*	*0.15*	*0.16*	*0.22*	*0.16*	*0.16*	*0.25*
race	1.43	0.52*	0.56*	0.61*	0.50*	0.52*	0.45
(*s.e.*)	*0.39*	*0.18*	*0.18*	*0.27*	*0.19*	*0.19*	*0.30*
time	0.021	0.017	0.02	0.02	0.02	0.02	0.01
(*s.e.*)	*0.028*	*0.014*	*0.014*	*0.016*	*0.015*	*0.014*	*0.02*
trt	0.87	0.55*	0.59*	0.58*	0.56*	0.55*	0.64*
(*s.e.*)	*0.20*	*0.12*	*0.13*	*0.18*	*0.13*	*0.12*	*0.21*

Table 4.2: Parameter estimates and their standard errors from analyzing aids data by utilizing various methods.

	PI-GLM4	JM-GLM4	PI-GAM4	JM-GAM4
trt	0.59	0.60	0.56	0.57
$(s.e.)$	*0.13*	*0.13*	*0.13*	*0.12*

this illustrative example, one must try both methods while fitting the models. Then on the reasons of parsimony one may suggest the use of the simpler model. Hence, for this dataset a GLM seems to be adequate.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

It is generally difficult to model data with non-ignorable missing. Pattern-mixture models and selections models are commonly used to fit data of these kind. We propose two approaches, the pseudo-imputation (PI) approach and the joint-modeling (JM) approach, that fall into the category of pattern-mixture models.

For both approaches, we make a key assumption that there is an observable surrogate variable used to determine missing patterns and then the missing indicator will be independent of missing responses conditional on the given missing pattern. Simulation studies in Chapter 2 show that both approaches yield precise point estimates when this assumption holds. The JM approach provides stable estimates with less variations than the PI approach even when the number of missing patterns is misspecified. When the assumption is violated, estimates obtained from the JM approach are much less biased than those from the PI approach. Looking at the models constructed to conduct simulations at the second scenario, we can see that the odds ratio of $Y = 1$ to $Y = 0$ being missing is very high. It is reasonable to believe that the JM approach would provide acceptable estimates while the magnitude of odds ratio is relatively small.

In Chapter 3, we apply our approaches to cases where a generalized linear regression may not be appropriate. Hence, fitting data in the generalized additive model framework may be desired, particularly when the primary interest is the treatment difference. In

simulation studies, it has been shown that parameter estimates for both approaches with a generalized additive model are nearly equally accurate in general even when data are generated from the generalized linear model. When there exists a non-linear relationship, these two approaches can take advantage of their flexibility, capture the nonlinearity existing in data, and yield precise estimates. The JM approach along with generalized additive models can provide more flexibility than the PI approach since it allows to fit a generalized additive model for the missing indicator model. It can be very much desired since no interpretation is required for parameters in the missing indicator model, no matter what the underling model is.

5.2 Future Work

5.2.1 Pattern Finder

Previously, we assumed that the surrogate variable solely determining the group membership of subjects can be observed. In this section, we discuss a few ideas to relax this assumption and extend our methods to more general settings. For example, there is no surrogate variable on record but only covariates measured at the baseline, or along with covariates, we have full measurements of time-dependent surrogate variable. Thus, before we use the proposed models as in Chapter 2, we might need to use some form of classification or clustering to create the pattern variable. Comparison of clustering methods is out of scope of this dissertation. We only provide a brief review on this subject in the following sections. Once the clustering method is determined and missing patterns are defined, one can apply our proposed approaches to address the missing data problem.

There is a rich volume of literature devoted into the field of classification and clustering. Some historical information about the development of theories of classification is discussed by numerous authors (Cain 1962; Reyment et al., 1984; Sutcliffe, 1994). Generally, classification can be referred to as either supervised learning or unsupervised learning. In supervised learning, a subject is assumed to belong to one of the known number of groups whose characteristics have been defined using a training set and the objective is to determine the class for subjects who miss their group labels. On the other

hand, the unsupervised learning is to find groups in data without the help of a response variable and the number of classes is to be determined. Supervised learning is of little interest in our cases and some discussions of it can be found in Krzanowski and Marriott (1995), Ripley (1996) and Fayyad et al. (1996). Classification methods reviewed in the followings fall into the category of the unsupervised learning.

Graphic aids can be of immediate attentions, such as scatter plots, Andew's plot (Andrew, 1972) and Chernoff faces (Chernoff, 1973). In addition, formal classification methods are either partitional or hierarchical. Classification tree is the most commonly used hierarchical methods (Bobisud and Bobisud, 1972; McMorris et al., 1983). Clustering criteria include single link (Sneath, 1957), complete link (McQuitty, 1960), centroid (Gower, 1967b), and others (Podani, 1989). The most common partition method is the K-mean clustering (Friedman and Rubin, 1967; MacQueen, 1967; Babfield and Bassill, 1977; Linde et al., 1980; Ismail and Kamel, 1989). There are also abundant of research on Bayesian clustering. Chang and Afifi (1974) proposed a Bayes procedure for classifying observations with one binary variable and c continuous variables. Vlachonikolis (1990) extended this approach to more general cases in which observations consist of b binary variable and c continuous variables. Stephens (2000) considered the problem of label switching in mixture models. Kang (2005) applied Dirichlet mixtures to clustering techniques to estimate nonparametric regression.

One possibility that allows flexibility is a way to produce reasonable configurations as we go along a markov chain. We start with a configuration based on the best of our current knowledge or a known surrogate variable. Then we move stochastically an observation or observations from the present configuration to a new one depending on some criteria, possibly defined in terms of some goodness of fit and penalty for the size of a given configuration. Each time movement is one of the following elementary move: (a) Join two group, (b) Split one group, or (c) Move one observation from present group to a different one.

5.2.2 Testing the MAR Assumption

In both approaches we proposed, there is a key assumption that, conditional on the missing pattern, the cause of the missingness is not to do with missing responses. In other word, the MAR can be assumed with a given pattern. This was an assumption and can not be verified. There are some likelihood ratio tests introduced to test the missing completely at random hypothesis (Fuchs, 1982; Little, 1988). A non-parametric test has been proposed by Diggle (1989). Chen and Little (1999) generalized the test statistics in Little (1988) to the generalized estimating equation setting to avoid distributional assumptions. Even though the above methods are not for testing the MAR assumption, we can still us them as preliminary screening tools. If the MCAR assumption holds, then one can definitely apply our methods. In other hand, we might be able to determine the magnitude of violation.

Bibliography

[1] Berry S. M., Carroll, R. J., and Ruppert, D. (2002), Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association* 97: 160-169.

[2] Brooks, S. (1998), Markov Chain Monte Carlo and its Applications. *The Statistician*, 47:69-100.

[3] Cowles, M.K., and Carlin, B.P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91: 883-904.

[4] Crainiceanu, C. M., Rupper, D., and Wand, M. P. (2005), Bayesian analysis fpr penalized spline regression using WinBUGS. *Journal of Statistical Software* 14 (14):1-18.

[5] Demirtas, H. (2005), Bayesian Analysis of Hierarchical Pattern-mixture Models for Clinical Trials Data with Attrition and Comparisons to commonly Used Ad-hoc and Model-based Approaches. *J. Biopharmaceutical Stat.* 15: 383-402.

[6] Daniels, M. J. and Hogan, J. W. (2000), Reparameterizing the Pattern Mixture Model for Sensitivity Analyses under Informative Dropout. *Biometrics* 56: 1241-1248.

[7] Diggle , P. J. and Kenward, M. G. (1994), Informative Dropouts in Longitudinal Data Analysis (with discussion), *Applied Statistics* 43: 49-93.

[8] Durban, M. and Currie, I. D. (2003), A note on P-spline additive models with correlated errors. *Computational Statistics* 18: 251-262.

[9] Eilers, P. H. C. and Marx, B. D. (1996), Flexible smoothing with B-splines and penalties. *Statistical Science* 11:89-121.

[10] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems . *The Annals of Statistics* 1: 209-230.

[11] Fitzmaurice, G. M. (2003), Methods for Handling Dropouts in Longitudinal Clinical Trials. *Statistica Neerlandica* 57, 1:75-99.

[12] Fitzmaurice, G. M. Heath, A. F. and Clifford, P. (1996), Logistic Regression Models for Binary Panel Data with Attrition. *Journal of the Royal Statistics Society, Series A*, 159: 249-263.

[13] Fitzmaurice, G. M. and N. M. Laird (2000), Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies, *Biostatistics* 1: 141-156.

[14] Fitzmaurice, G. M., Laird, N. M., and Shneyer, L. (2001), An Alternative Parameterization of the General Linear Mixture Model for Longitudinal Data with Nonignorable Dropouts. *Statistics in Medicine* 20: 1009-1021.

[15] Follman, D. and M. Wu (1995), An approximate generalized linear model with random effects for informative missing data, *Biometrics* 51: 151-168.

[16] Friedman, J.H. and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association* 76, 817 - 823.

[17] Gelfand, A. E. and Smith, A. F. M. (1990), Sampling-based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85: 398-409.

[18] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis, 2nd edition*, Boca Raton, FL: Chapman and Hall/CRC Press.

[19] GILKS, W. R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. *In Bayesian Statistics (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.)* 641-649. Oxford Univ. Press.

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

[21] Green, P. J. (1985), Linear models for field trials, smoothing and cross validation. *Biometrika* 72:523-537.

[22] Guo, W., Ratcliffe, S.J. and Ten Have, T.T. (2004), "A Random Pattern-Mixture Model for Longitudinal Data with Dropouts", *Journal of the American Statistical Association*, 99:929-937.

[23] Hastie, T.J. and Tibshirani, R.J. (1986), "Generalized Additive Models (with discussion)," *Statistical Science*, 1, 297 - 318.

[24] Hastie, T.J. and Tibshirani, R.J. (1990), Generalized Additive Models, New York: Chapman and Hall.

[25] Hastie, T.J. (1996), Pseudosplines. *Journal of the Royal Statistics Society, Series B* 58: 379-396.

[26] Heckman, J. (1976), The Common Structrure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models. *Annals of Economical and Social Measures* 5: 475-492.

[27] Hogan, J. W. and Laird, N.M. (1997), Mixture Models for the Joint Distribution of Repeated Measures and Event Times. *Statistics in Medicine* 16: 239-257.

[28] Hogan, J. W., Roy, J., and Korkonzelou, C. (2004), Tutorial in Biostatistics: Handling Dropout in Longitudinal Studies. *Statistics in Medicine* 23: 1455-1497.

[29] Hogan, J.W., Lin, X. and Herman, B. (2004), "Mixtures of Varing Coefficient Models for Longitudinal Data with Discrete or Continuous Nonignorable Dropout", *Biometrics*, 60: 854-864.

[30] Huber, P.J. (1981) *Robust Statistics*. New York: John Wiley and Sons.

[31] Kang, C. (2005), *Regression via clustering using Dirichlet mixtures*. PhD Thesis, North Carolina State University.

[32] Lang, S. and Brezger, A. (2004), Bayesian P-splines. *Journal of Computational and Graphical Statistics* 13: 183-212.

[33] Lavori P. W., Dawson, R., Shera, D. (1995), A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data. *Statistics in Medicine* 14: 1913-1925.

[34] Liang, K. Y. and Zeger, S. L. (1986), Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 73: 13-22.

[35] Little, R. J. A. (1988a), A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* 82: 1198-1202.

[36] Little, R. J. A. (1988b), Missing-data Adjustments in Large Surveys. *Journal of Business and Economical Statistics* 6: 287-296.

[37] Little, R. J. A. (1993), Pattern-mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association* 88: 125-134.

[38] Little, R. J. A. (1994), A Class of Pattern-mixture Models for Normal Incomplete data. *Biometrika*, 81: 471-483.

[39] Little, R. J. A. (1995), Modeling the Missing Mechanism in Repeated-Measures Studies. *Journal of the American Statistical Association* 90: 1112-1121.

[40] Little, R .J. A. and Rubin, D. B. (1987, 2002), *Statistical Analysis with missing data*, Wiley, New York.

[41] Marx, B. D. and Eilers, P. H. C. (1998), Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis* 28: 193-208.

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

[43] Minini, P. and Chavance, M. (2004), Sensitivity analysis of longitudinal normal data with drop-outs. *Statistics in Medicine* 23: 1039-1054.

[44] Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1997), The Analysis of Longitudinal Ordinal Data with Nonrandom Drop-out. *Biometrika* 84: 33-44.

[45] Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, M. G. (1998), Missing Data Mechanism and Pattern-mixture Models. *Statistica Neerlandica* 52: 153-161.

[46] Molenberghs, G., Thijs, H., Kenward, M. G., and Verbeke, G. (2003), Sensitivity Analysis of Continuous Incomplete Longitudinal Outcomes. *Statistica Neerlandica* 57(1): 112- 135.

[47] Paik, M. C. (1997), The Generalized Estimating Equation Approach when Data are not Missing Completely at Random. *Journal of the American Statistical Association* 92: 1320-1329.

[48] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995), Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association* 90: 106-121.

[49] Rosenbaum, P. R. and Rubin, D. B. (1983), Assessing Sensitivity to an Unobserved Binary Covariates in an Observational Study with Binary Outcome. *Journal of the Royal Statistics Society, Series B* 45: 212-218.

[50] Rubin, D. B. (1976), Inference and Missing Data. *Biometrika* 63: 581-592.

[51] Rubin, D. B. (1977), Formalizing Subjective Notions about the Effect of Nonresponse in Sample Surveys. *Journal of the American Statistical Association* 72: 538-543.

[52] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, Ney York.

[53] Ruppert, D. (2002), Selecting the number of knots for penalized splines. Journal of *Computational and Graphical statistics* 11: 735-757.

[54] Ruppert, D., and Carroll, R. J. (2000), Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* 42: 205-224.

[55] Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Wiley, Ney York.

[56] Serfing, R.J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.

[57] Speed, T. (1991), Comment on paper by Robinson. *Statistical Science* 6: 42-44.

[58] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002), Bayesian measures of model complexity and fit. *Journal of the Royal Statistics Society, Series B* 64: 583-640.

[59] Stone, C.J. (1985), "Additive Regression and Other Nonparametric Models," *Annals of Statistics*, 13, 689 - 705.

[60] Sturtz, S., Ligges, U., and Gelman, A. (2005), R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, 12(3):1:16

[61] Tang, G., Little, R. J. A. and Raghunathan, T. E. (2003), Analysis of Multivariate Missing Data with Nonigorable Nonresponse. *Biometrika* 90(4):747-764.

[62] Vach, W. and M. Blettner (1995), Logistic regression with incompletely observed categorical covariates: investigating the sensitivity against violations of the missing at random assumption, *Statistics in Medicine* 12: 1315-1330.

[63] van der Vaart, A.W. (1998) *Asymptotic Statistics*. Cambridge: Cambridge University Press.

[64] Wahba, G. (1978), Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistics Society, Series B* 40: 364-372.

[65] Wand, M. P. (2003), Smoothing and mixed models. *Computational Statistics* 18: 223-249.

[66] Wecker, W. E. and Ansley, C. F. (1983), The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the american statistical association* 78:81-89.

[67] Wu, M. C. and K. R. Bailey (1988), Analyzing changes in the presence of informative right censoring caused by death and withdrawal, *Statistics in Medicine* 7: 337-346.

[68] Wu, M. C. and K. R. Bailey (1989), Estimation and comparison of changes in the presence of informative right censoring: conditional linear model, *Biometrics* 45: 939-955.

[69] Wu, M. C. and Carroll, R. J. (1988), Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics* 44: 175-188.