2011

Analysis of Multivariate Data with Random Cluster Size

Xiaoyun Li

THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

ANALYSIS OF MULTIVARIATE DATA WITH RANDOM CLUSTER SIZE

By

XIAOYUN LI

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Spring Semester, 2011

The members of the committee approve the dissertation of Xiaoyun Li defended on December 2nd, 2010.

Debajyoti Sinha
Professor Directing Dissertation

Yi Zhou
University Representative

Dan McGee
Committee Member

Stuart Lipsitz
Committee Member

Approved:

Dan McGee, Chair, Department of Statistics

Joseph Travis, Dean, College of Arts and Sciences

The Graduate School has verified and approved the above-named committee members.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this dissertation, we examine binary correlated data with present/absent component or missing data that are related to binary responses of interest. Depending on the data structure, correlated binary data can be referred as *clustered data* if sampling unit is a cluster of subjects, or it can be referred as *longitudinal data* when it involves repeated measurement of same subject over time. We propose our novel models in these two data structures and illustrate the model with real data applications.

In biomedical studies involving clustered binary responses, the cluster size can vary because some components of the cluster can be absent. When both the presence of a cluster component as well as the binary disease status of a present component are treated as responses of interest, we propose a novel two-stage random effects logistic regression framework. For the ease of interpretation of regression effects, both the marginal probability of presence/absence of a component as well as the conditional probability of disease status of a present component, preserve the approximate logistic regression forms. We present a maximum likelihood method of estimation implementable using standard statistical software. We compare our models and the physical interpretation of regression effects with competing methods from literature. We also present a simulation study to assess the robustness of our procedure to wrong specification of the random effects distribution and to compare finite sample performances of estimates with existing methods. The methodology is illustrated via analyzing a study of the periodontal health status in a diabetic Gullah population.

We extend this model in longitudinal studies with binary longitudinal response and informative missing data. In longitudinal studies, when treating each subject as a cluster, cluster size is the total number of observations for each subject. When data is informatively missing, cluster size of each subject can vary and is related to the binary response of interest

and we are also interested in the missing mechanism. This is a modified situation of the cluster binary data with present components. We modify and adopt our proposed two-stage random effects logistic regression model so that both the marginal probability of binary response and missing indicator as well as the conditional probability of binary response and missing indicator preserve logistic regression forms. We present a Bayesian framework of this model and illustrate our proposed model on an AIDS data example.

CHAPTER 1

Introduction

1.1 Motivation and Purpose

Correlated data are common in biomedical researches. In some cases the sampling unit is a cluster of subjects, such as members from same family or rats in same litter, hence observations within a cluster are usually correlated. This type of data is usually called *clustered data*. In some other cases, measurements are collected on the same subject over time. For example, in Framingham Heart Study, each participant was followed up in a long period of time to identify the factors that contribute to cardiovascular disease. Therefore, observations from same subject are correlated. This type of data is called *longitudinal data*. In this dissertation, we treat longitudinal data as a special type of clustered data where the cluster sampling unit is a subject and the modeling involves time as a factor.

For clustered data, cluster size is defined as the total number of observations for each cluster. Usually cluster size is fixed and the it is not related to the outcome of interest. For example, we are interested in a predictive modeling of periodontal disease with risk factors and examine sixteen teeth including molar and premolar teeth of each subject, then the cluster size of each subject would be a fixed number 16. An example of fixed cluster size in longitudinal studies could be a study design that measures the blood pressure of each subject right before the experiment, 10 minutes, 20 minutes, 1 hour and 3 hour after experiments and all observations are collected successfully, then each subject would have 5 observations therefore the cluster size for each subject will all be 5. However, in practice, situations are not as ideal as planned. In the dental studies, it is possible that some of the molar or premolar teeth are already removed before the patient comes to the clinic for periodontal disease examination. In the example of longitudinal study, some of the blood pressure measurements may be missing because of various reasons, such as the patient left

before the last measurement was taken or the blood pressure device failed to show the measurement, etc. If the varying cluster size is not related to the response of our interest, we can treat the cluster size as a nuisance parameter and fit the predictive model with all available information we have. However, if the varying cluster size is related to the response of interest, the inference will no longer be valid if we ignore the cluster size and fit the model with only the observations we have.

Often in toxicity experiments and biomedical researches, not only are the data within same cluster correlated, the number of cluster size can also be correlated with the outcomes of interest in the cluster [11], which is inferred as 'informative cluster size' [37] or nonignorable cluster size [18]. An example of this situation is found in dental studies. People with fewer teeth are more likely to have poor dental health. Another example is found in toxicity experiments, bigger litter size might result in decreased fetus weight since the amount of space and nutrition are limited. While in the case of longitudinal data, this is called informative missingness where the mechanic of missingness is related to the response of interest [14]. When informative cluster size exists, regular models without considering the cluster size effect would usually lead to invalid estimates. In this dissertation, we propose mixed effects models for binary responses in clustered data or longitudinal data with random cluster size that has unbiased estimates and some good statistical properties.

A brief introduction of different general model types for modeling correlated binary data without considering informative cluster size is shown throughout the rest of Chapter 1. Chapter 2 reviews the existing methods for clustered binary data with informative cluster size and longitudinal binary data with informative missing data. In Chapter 3, we propose our two-stage Bridge model for clustered binary data with informative cluster size and presents a data example. We also perform simulations to assess the performance of the model. In Chapter 4, we modify and adopt the novel Bridge model in longitudinal studies with informative missingness. Chapter 5 concludes the dissertation with conclusions and future work.

1.2 Traditional Methods for Correlated Binary Data

Ordinary regression models are the most commonly used models when the response is continuous. However, when the responses of interest are categorical variables, the normal assumption in ordinary regression no longer holds. Therefore, generalized linear models

(GLM) were developed for categorical responses in predictive modeling. In GLM, the functions of the mean would be linear to explanatory variables [2]. A generalized linear model includes three components: a random component, a systematic component and a link function.

Suppose Y is a categorical variable and $\mathbf{X}$ is a $p \times 1$ vector of covariates (exploratory variable). There are n independent observations $(y_1; \mathbf{x}_i)$,..., $(y_1; \mathbf{x}_n)$, The random component is the distribution of Y with independent observations $(y_1, ..., y_n)$ in the natural exponential family:

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp[y_i Q(\theta_i)], \tag{1.1}$$

where θ_i is the parameter of the distribution for y_i and may depend on the exploratory variables $\mathbf{X}$. Here $a(\theta_i) \geq 0$, $b(y_i)$ and $Q(\theta_i)$ are the real-value functions of θ_i [7], $i = 1, ...n$.

The systematic component uses a vector $\boldsymbol{\eta} = (\eta_1, \eta_2, ...\eta_n)$ to denote a linear relationship with the covariates $\mathbf{x} = (\mathbf{x}_1, ...\mathbf{x}_n)$:

$$\eta_i = \boldsymbol{\beta}^T \mathbf{x_i}, \quad i = 1, ...n, \tag{1.2}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients for covariates X.

The link function $g(\cdot)$ relates $\mu_i = E(Y_i)$ with systematic components η_i by

$$g(\mu_i) = \eta_i, \quad i = 1, ...n. \tag{1.3}$$

For binary responses, we usually use logit link function as the link function $g(\cdot)$:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \log(p_i) - \log(1 - p_i) = 0, \tag{1.4}$$

where $p_i = E(Y_i) = \mu_i$.

In GLM, we assume that all outcomes $Y_1, Y_2, ...Y_N$ are independent. However, observations often come in clusters, therefore the observations within clusters tend to correlate with each other. The assumption of independent responses dose not hold any more and the ordinary analysis would lead to invalid point estimates and standard errors. There are two types of methods we can use for correlated/clustered binary responses: the first one is to model the conditional effect of binary responses for each cluster given cluster-random effect. Within a cluster, when the cluster random effect is given, all binary responses are independent between each other. The most commonly used model for this type is Generalized

Linear Mixed Effect Models (GLMM). The other type of methods, instead of conditionally modeling the binary responses within clusters, we model the marginal (population-average) model for the binary responses in population, in addition to the regular GLM, it specifies a variance function and pairwise correlation parameter for the correlated binary responses within each cluster [2]. The most popular method of this type is the generalized estimating equations(GEE). We introduce these two types of methods in the following subsections. These traditional methods for correlated binary data assume cluster size being independent of the binary outcomes of interest hence do not consider the impact of informative cluster size.

Mixed Effects Model: Generalized Linear Mixed Effects Model

The generalized linear model can be extended to generalized linear mixed effects model (GLMM) by adding random effects. Let i be the index of cluster, where $i = 1, ...n$. Let j be the index of the jth subject of the ith cluster, $j = 1, ...s_i$, where s_i is the cluster size for the ith cluster. We assume that $\mathbf{u_i}$ is a $q \times 1$ vector of cluster-specific random effect on the ith cluster, let Y_{ij} denote the response for the jth subject in the ith cluster, $i = 1, ...n$ and $j = 1, ...s_i$. We also let $\mu_{ij} = E(Y_{ij}|X_{ij}, \mathbf{u}_i)$.

The linear predictor for a GLMM has the following form

$$g(\mu_{ij}) = \boldsymbol{\beta}^T \mathbf{x_{ij}} + \mathbf{u}_i^T \mathbf{z_{ij}}, \tag{1.5}$$

where $\mathbf{x_{ij}}$ is a $p \times 1$ vector of covariates for the jth subject on ith cluster, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters for fixed effects and $\mathbf{z_{ij}}$ is a $q \times 1$ vector of variables that have random effects. We assume that binary responses $(Y_{i1}, ...Y_{is_i})$ are independent given the cluster-specific effect $\mathbf{u_i}$, then the independence assumption holds in (1.5).

It is possible to attain the marginal model $E(Y_{ij}|X_{ij})$ by:

$$\begin{aligned} E(Y_{ij}|X_{ij}) &= E(Y_{ij}|X_{ij}, \mathbf{u}_i) f(\mathbf{u}_i) d\mathbf{u}_i & (1.6) \\ &= \int g^{-1}(\boldsymbol{\beta}^T \mathbf{x}_{ij} + \mathbf{u}_i^T \mathbf{z_{ij}}) f(\mathbf{u}_i) d\mathbf{u}_i & (1.7) \end{aligned}$$

where $f(\mathbf{u}_i)$ is the density function of random effect $\mathbf{u}_i$. However, this usually does not have closed form and the marginal model $E(Y_{ij}|X_{ij})$ could not maintain same link function $g(\cdot)$ such that there is $g(E(Y_{ij}|X_{ij})) = \boldsymbol{\beta}^{*T} \mathbf{x}_{ij}$.

Marginal Model

When the marginal model of the responses with respect to the covariates is our primary interest and we are only interested in cluster-level covariate effect, then marginal models are useful. The generalized estimating equations (GEE) proposed by Liang and Zeger (1986) [22] is by far the most popular marginal method for correlated binary responses.

The GEE method is a quasi-likelihood method which only specifies on the first moment and the how the variance depends on the mean [35]. Suppose in a GLM, we have link function $g(\cdot)$ and linear predictor $\eta_i = g^{-1}(\mu_i) = \boldsymbol{\beta}\mathbf{X}_i$. The quasi-likelihood would be

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} (\frac{\partial \mu_i}{\partial \eta_i})' (A_i^{1/2} R_i A_i^{1/2})^{-1} (y_i - \mu_i) X_i = 0, \tag{1.8}$$

where R_i is the marginal correlation matrix for $\mathbf{Y_i}$, termed as working correlation matrix. Matrix A_i is the $n_i \times n_i$ diagonal matrix with elements $\text{Var}(Y_{ij})$, with n_i as the cluster size for the ith cluster. The variance-covariance matrix can be written as $V_i = A_i^{1/2} R_i A_i^{1/2}$. Liang and Zeger [22] proposed a moment-based estimate for the working correlation matrix. A iterative two-stage estimation procedure is used to estimate β and the pairwise correlation.

GEE method has the advantage of easy implementation and physical marginal interpretations, which are important for data analysis. However, there are some disadvantages of GEE method. First of all, it can not specify the complete joint distribution and the likelihood function can not be attained. Since the likelihood is not attained we could not have any inference of the model. Secondly, the within-cluster correlation structure is treated as nuisance, hence we cannot estimate the within-cluster correlation and the cluster heterogeneity when using GEE method.

CHAPTER 2

Literature Review

2.1 Existing Methods for Informative Cluster Size

Discussions of 'informative cluster size' has been raised within last ten years. Researchers realized that the varying cluster size would actually influence the binary outcomes and the traditional analysis methods would give biased results. They put their efforts on adjusting or modeling the cluster size as well as the binary outcomes. Analysis of clustered binary data with informative cluster size has been proposed by Hoffman et al.(2001) [18], Williamson et al. (2003) [37], Dunson et al. (2003) [11] and Gueorguieva (2005) [17].

Within-cluster Resampling

Hoffman et al. [18] realized that traditional generalized estimating equations (GEE) with independent working correlation matrix assign more weight on the larger cluster hence the risk in larger cluster is bigger than the risk in smaller cluster. However, in the case of dental studies, people with fewer teeth (smaller sample size) tend to have poor dental health (higher risk). This is contradict with the estimation of regular GEE, which assigns lower risk to smaller cluster size (people with fewer teeth). Therefore, Hoffman et al. proposed a within-cluster resampling (WCR) method that could adjust for the informative cluster size. Suppose there are I clusters altogether. Instead of fitting the marginal model directly, he selected one observation from each cluster with replacement and analyzed these I independent observations from I clusters using the marginal models. The process is repeated a large number of Q times and the within-cluster resampling estimator is the average of the Q estimates from the resampling scheme. We let $\hat{\beta}(R; q)$ denote the qth estimate of marginal

models from the qth resampled dataset, where $q = 1, ...Q$. Then the WCR estimator $\bar{\beta}$ is

$$\bar{\beta} = Q^{-1} \sum_{q=1}^{Q} \hat{\beta}(R; q). \tag{2.1}$$

Since in each resampling process, only one observation is resampled, the I observations for fitting generalized linear model are independent. However, these Q resampled datasets are correlated and the dependency can be taken into account into the variance of $\bar{\beta}$. Hoffman also showed the asymptotic normality of the within-cluster resampling estimate (2.1) and provided a consistent variance estimator for this within-cluster resampling estimator.

Hoffman's WCR method can adjust for the incorrect weights in GEE. On the other hand, this is a marginal model and the rational of the within-cluster resampling scheme makes the within-cluster association as nuisance, hence it is impossible to estimate the subject-specific effect and within-cluster association when applying this resampling scheme.

Clustered Weighted Generalized Estimating Equation

Williamson (2003) [37] applied the idea of WCR to generalized estimating equation (GEE) and proposed the clustered weighted generalized estimating equation (CWGEE). In Williamson's paper, he used the inverse of the cluster size as the weight to adjust for the informative cluster size and used the independent working covariance structure. He also showed that WCR and CWGEE are equivalent for large sample size.

Consider the ith cluster, where $i = 1, ..., N$. Let Y_{ij} be the binary response for the jth subject in the ith cluster. He assumed that $\{n_i, Y_{i1}, ..., Y_{in_i}\}$ are independent across clusters. If a cluster I is randomly selected from the N clusters, and k_I is a realization of a uniform random variable from 1 to n_I. Then for the marginal model he considered:

$$E\mathbf{U}_I(Y_{IK_I}, \mathbf{X}_{IK_I}; \boldsymbol{\beta}) = 0.$$

Since binary response is considered, a logistic marginal model is considered and $\mathbf{U}_{ij}(\boldsymbol{\beta}) = \mathbf{U}_i(y_{ij}, \mathbf{x}_{ij}; \boldsymbol{\beta}) = \mathbf{x}_{ij}^T y_{ij} - \mu(\mathbf{x}_{ij}; \boldsymbol{\beta})$, where $\mu(\mathbf{x}_{ij}; \boldsymbol{\beta}) = \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij})(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_{ij})$. Applying the within-cluster resampling (WCR) rationale on the score function, and assuming the working correlation matrix is independence matrix, he proposed that the analytical form of the average of score function is:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{U}_{ij}(\boldsymbol{\beta}), \tag{2.2}$$

which assigns the inverse of cluster size as the weight in GEE.

However, there are some disadvantages of CWGEE. First of all, the inverse of cluster size weight that CWGEE assigns for informative cluster size data is based on the assumption of using the independence matrix as the working correlation matrix. This assumption is too strong to be true and it would lead to biased estimates [28]. Also, the CWGEE assumes the distribution for cluster sizes are the same for the population. If the distribution of cluster sizes is different in two sub-populations, even though the relationship between the response and covariates are the same conditional on cluster size, the CWGEE would not remain valid.

Dunson's Bayesian Mixed Effects Model and Its Modification

Dunson et al. (2003) proposed a Bayesian framework for joint modeling of cluster size and binary outcomes [11]. Let i be the index of cluster ($i = 1, ..., n$) and let j be the index of subject within the cluster ($j = 1, ...s_i$), and s_i is the cluster size for the ith cluster. $\mathbf{y_{ij}} = (y_{ij1}, ...y_{ijp})$ is a vector of subject outcomes. They assumed that the binary outcomes have an underlying normal latent variable such that $y_{ijk} = g_k(y^*_{ijk})$, where y^*_{ijk} is a continuous variable from normal distribution and $g_k(y^*) = I(y^* > 0)$, where $k = 1, ...p$, where p is the dimension of responses.

The vector of underlying continuous variables $\mathbf{y^*_{ij}}$ was modeled as:

$$\mathbf{y^*_{ij}} = \boldsymbol{\mu} + \boldsymbol{\alpha}\mathbf{x_i} + \boldsymbol{\Lambda}\mathbf{W_{1i}}\xi_\mathbf{i} + \boldsymbol{\Gamma}\mathbf{W_{2i}}\boldsymbol{\eta_{ij}} + \boldsymbol{\epsilon_{ij}} \tag{2.3}$$

where $\boldsymbol{\mu}$ is a vector of intercept parameters, $\boldsymbol{\alpha} = [\boldsymbol{\alpha_1}, \boldsymbol{...\alpha_p}]^T$ is a $p \times q$ matrix of regression parameters, $\mathbf{x_i}$ is a $q \times 1$ vector of cluster-specific covariates, $\boldsymbol{\Lambda} = [\boldsymbol{\lambda_1}, \boldsymbol{...\lambda_p}]^T$, where $\boldsymbol{\lambda}_i$ is the coefficient for the cluster-level random effect ξ_i. At the same time, $\boldsymbol{\Gamma} = [\boldsymbol{\gamma_1}, \boldsymbol{...\gamma_p}]$, where $\boldsymbol{\gamma}_i$ is the coefficient for the cluster-level random effect η_{ij}. Vectors $\boldsymbol{\xi_i} = (\boldsymbol{\xi_{i1}}, \boldsymbol{...\xi_{ir}})^T$ and $\boldsymbol{\eta_{ij}} = (\boldsymbol{\eta_{ij1}}, \boldsymbol{...\eta_{ijr}})^T$ are i.i.d. standard normal cluster level and subject level random effects, and $\mathbf{W_{1i}}, \mathbf{W_{2i}}$ are diagonal weighting matrices for the ith cluster.

The cluster size was modeled as a generalization of the continuation-ratio ordinal response,

$$\Pr(s_i = j | s_i \geq j, \mathbf{x_i}, \boldsymbol{\xi_i}) = F(\delta_j - \mathbf{x_{ij}}^T\boldsymbol{\beta} - \boldsymbol{\lambda_{p+1}}^T\mathbf{W_{1i}}\boldsymbol{\xi_i}), \text{ for } j = 1, ...T-1, \tag{2.4}$$

where $F(\cdot)$ is a one to one monotone function mapping from the real numbers to interval $[0, 1]$, $\boldsymbol{\delta} = (\delta_1, ..., \delta_{T-1})$ are intercept parameters for cluster size.

In Dunson et al.'s model, they incorporate the dependency between y_i and cluster size s_i by sharing the same cluster-level random effects $\mathbf{W_{1i}\xi_i}$. Bayesian framework is used and the prior distributions and posterior computations are given in their paper. Normal density is used for the random effects and both the cluster size and the binary responses share the same cluster-level random effects. However, when applied to binary responses of interest, this model can only give an explicit form for conditional model given cluster-level and subject-level random effects. Since normal random effects are used in this model, when we integrate over the random effects, there is no closed form for the integration. Therefore, we could not attain the parameters estimates and interpretation for a marginal model by integrating over the random effects.

Gueorguieva (2005) [17] commented and improved Dunson's model with correlated cluster-level random effects by using a maximum likelihood method. As mentioned in the previous section, Dunson et al. used a vector of shared cluster-level random effects for both the outcomes of interest $\mathbf{y_i}$ and cluster size s_i. However, in Gueorguieva's paper, he showed that more bias may be produced by assuming the same cluster-level random effects for both $\mathbf{y_i}$ and s_i than ignoring the cluster size effect. He suggested an improved model with correlated cluster-level random effects for $\mathbf{y_i}$ and s_i instead of using same cluster-level random effects. The model is then:

$$\mathbf{y^*_{ijk}} = \boldsymbol{\mu} + \boldsymbol{\alpha}\mathbf{x_i} + \boldsymbol{\Lambda}\mathbf{W_{1i}}\xi_{\mathbf{ik}} + \boldsymbol{\Gamma}\mathbf{W_{2i}}\boldsymbol{\eta_{ij}} + \boldsymbol{\epsilon_{ij}}, \text{ for } k = 1, ...p, \tag{2.5}$$

and

$$\Pr(s_i = j | s_i \geq j, \mathbf{x_i}, \boldsymbol{\xi_i}) = F(\delta_j - \mathbf{x_{ij}}^T\boldsymbol{\beta} - \boldsymbol{\lambda_{p+1}}^T\mathbf{W_{1i}}\boldsymbol{\xi_{i(p+1)}}), \text{ for } j = 1, ...T-1, \tag{2.6}$$

where $\xi_{i1}, ...\xi_{i,(p+1)}$ are correlated.

This model is reorganized such that the sub-unit specific random effects are combined with the random error so that the model can be fitted in PROC NLMIXED in SAS, since NLMIXED does not allow multilevel random effects.

2.2 Existing Methods for Informative Missing Data

In longitudinal data, missing-data mechanism may be related to the response of interest. When the missing-data mechanism is related to missing response and it is unknown, we refer it as "informative missing". Regular inferences that ignore the missing data would

lead to bias. Three approaches have been developed for data with non-ignorable missing dataselection approach, pattern mixture approach and shared random effect approach. In this dissertation, we only focus on the study of shared random effects approach.

Ten Have's Mixed Effects Logistic Models for Binary Longitudinal Data

Ten Have et al. [32] proposed a shared random effects logistic regression model for longitudinal binary response with informative drop-out. The definition of informative drop-out here is "drop-out is dependent on an unbservated random effect underlying the observed and unobserved binary outcomes" [**?**]. The key assumption of the shared random effect model is the conditional independence between binary response and missing-data mechanism given random effects. Let y_i be the ith binary response and z_i be the variable for drop-out status for the ith subject. Let τ be the random effect. The shared random effect model can be expressed as:

$$f(y_i, z_i) = \int f(y_i, z_i|\tau) f_\tau(\tau) d\tau = \int f_y(y_i|\tau) f_z(z_i|\tau) f_\tau(\tau) d\tau. \tag{2.7}$$

In Ten Have's model, they assumed $Y_i|\tau$ follows a Bernoulli distribution with probability of success π_{ij}, with

$$\text{logit}(\pi_{ij}) = \boldsymbol{\tau_i}^T \Sigma \mathbf{w_{ij}} + \boldsymbol{\beta}^T \mathbf{x}_{ij}, \tag{2.8}$$

where $\boldsymbol{\tau_i}$ is random effects for the ith subject, Σ is a square matrix, where $\Omega = \Sigma^T \Sigma$. $\mathbf{x_{ij}}$ is a vector of fixed covariates, $\boldsymbol{\beta}$ is the regression parameter (log odds ratio) for the corresponding covariates. They assume random effect τ_i follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix as a identity matrix.

Variable for drop-out Z_i is defined as the number of visit after which the ith subject dropped out of the study. Suppose there is a maximum of J visits. If the ith subject drops out after the jth visit, then $Z_i = j$. If the subject did not drop out of the study, then they let $Z_i = J$. S_{ij} is used for the indicator of drop-out for the jth visit of the ith subject, for example:

$$S_{ij} = \begin{cases} 1 & \text{if subject i dropped out} \\ 0 & \text{o.w.} \end{cases}. \tag{2.9}$$

They assumed $\lambda_{ij} = \Pr(Z_i = j | Z_i > j-1, \tau_i)$. Therefore, λ_{ij} can be seen as the discrete hazard rate in discrete time survival analysis. The log likelihood for z_i can be written as:

$$\log[f_z(z_i|\tau)] = \sum_{j=1}^{J-1}[s_{ij}\log\psi_{ij} + \log(1-\lambda_{ij}\sum_{j'=j}^{J} s_{ij'})], \tag{2.10}$$

where $\psi_{ij} = \frac{\lambda_{ij}}{1-\lambda_{ij}}$.

The logarithm of ψ_{ij} can be modeled as:

$$\log\psi_{ij} = \tau_i^T \Sigma^* w_{ij} + \rho^T u_{ij}, \tag{2.11}$$

where $\Sigma^* = \Sigma + \Delta$, $\mathbf{u}_{ij}$ is a vector of fixed covariates for ψ_{ij} and ρ is a vector of log odds ratio corresponding to the fixed covariates $\mathbf{u}_{ij}$. The covariates for binary longitudinal response Y_i and informative drop-out Z_i do not necessary need to be same.

They estimated the likelihood by approximating normal random effects distribution with a mixture of binomial distributions, therefore the integral over the multivariate normal distributed τ could be replaced by summation of binomial distributions.

Albert's Binary Longitudinal Model with Autoregressive Latent Processes

Albert et al. [3] extended Ten Have et al.'s shared subject-specific random effects model to a shared autoregressive latent process and model missing-data mechanism in a slightly different way. Instead of only model informative drop-out, Albert et al. also considered intermittent missing, therefore, missing-data mechanism becomes a categorical variable with

$$z_{it_j} = \begin{cases} 0 & \text{observed} \\ 1 & \text{intermittent missing} \\ 2 & \text{drop out.} \end{cases} \tag{2.12}$$

Instead of subject specific random effect, Albert et al. uses a laten process b_{it_j} for the random process where b_{it_j} and $b_{it'_j}$ are correlated with $cov(b_{it_j}, b_{it_{j'}}) = \sigma^2\exp(-\theta|t-t'|)$ and $\theta > 0$. Under the shared autoregressive latent process, they assigned a logistic regression to the binary longitudinal response Y_{it_j}:

$$\text{logit}(\Pr(E(Y_{it_j})|b_{it_j})) = \boldsymbol{\beta}^T \mathbf{x}_{it_j} + b_{it_j},$$

where $\boldsymbol{\beta}$ is a vector of regression parameters for $\mathbf{x}_{it_j}$.

The missing data mechanism Z_{it_j} ($j > 1$, Z_{it_j}) is modeled as:

$$P(Z_{it_j} = l | b_{it_j}, Z_{it_{j-1} \neq 2}) = \begin{cases} \frac{1}{1+\sum_{l=1}^{2} \exp(\boldsymbol{\nu'}_{l_i t_j} \eta_l + \gamma_l b_{it_j})}, & \text{l=0} \\ \frac{\exp(\boldsymbol{\nu}_{l_i t_j} \eta_l + \gamma_l b_{it_j})}{1+\sum_{l=1}^{2} \exp(\boldsymbol{\nu'}_{l_i t_j} \eta_l + \gamma_l b_{it_j})}, & \text{l=1,2,} \end{cases}$$

Monte Carlo EM (MCEM) was used to estimate the maximum likelihood.

CHAPTER 3

Likelihood Method for Binary Responses with Random Cluster Size

3.1 Introduction

Clustered binary response data are encountered in a multitude of biomedical settings including studies of developmental toxicology, family-based genetic traits, as well as longitudinal and spatial studies. Data analysis methods need to account for the effect of within cluster association to obtain correct estimates of the regression parameters and their corresponding standard errors [1](Aerts et al., 2002). In addition to the correlation within clusters, it is quite often that the size of cluster may be random and would also influence the outcomes of interest. This phenomenon has been earlier considered in the literature as an 'informative cluster size' scenario [37][36](Williamson et al., 2003; Williamson et al., 2008). For example, in dental studies, people often have varying number of teeth and fewer teeth may be associated with poor dental health of existing teeth.

We consider the study on Gullah speaking African-American Type-2 diabetics [12](Fernandes et al., 2009) with data collected on the periodontal health status of each tooth (component) within a patient (cluster). Here we only consider the molar and premolar teeth at 16 locations of the jaws. For each tooth (corresponding to a location in jaw), the three states of the outcome variable of primary interest were absent tooth, tooth present with disease and tooth present with no-disease. The primary objective of the study was to evaluate the influence of various patient-level as well as tooth-level explanatory variables (covariates) on these three categories of periodontal health status. This study design provides some interesting challenges for analysis. Most of the methods traditionally used for analyzing clustered binary data, including the generalized estimating equations [22](GEE; Liang and Zeger, 1986) and the mixed-effects conditional logistic regression methods (CLR; Breslow

and Day, 1980, Lin et.al., 2009)[6] [23], do not treat the cluster size as random. However, the latent factor that causes the loss of a tooth might also lead to the poor dental health of the rest of the teeth. Thus, if we omit the first level binary response (i.e., the presence/ absence of the tooth), the interpretation and evaluation of the covariate effects on the disease status become questionable.

An appropriate GEE based method [37](Williamson et al., 2003) should accommodate random (informative) cluster-sizes when we are interested in estimating the effects of covariates on a randomly selected tooth from a population of 'exchangeable' teeth. These extensions of GEE methods assume that all teeth are 'at risk' of disease, except, we may have a biased sample because the disease status for some teeth are 'missing from observation'. For estimating effects of covariates on the risk of disease for a typical tooth of a randomly selected patient, Hoffman et al. [18] introduced the 'within-cluster resampling' (WCR) technique and Williamson et al.(2003)[37] proposed the simpler and asymptotically equivalent 'clustered weighted GEE' (CWGEE) method. The CWGEE and WCR aim to estimate the effect on a typical cluster member from a random cluster via properly weighing the observations based on their cluster sizes. The model assumes 'independence' working correlation, which is unlikely to hold in practice, and consequently may lead to inefficient estimates (Panageas et al., 2007) [28]. At the same time, the rate of periodontal decay is expected to vary with tooth location and modeling the cluster size as a whole using the above methods may not be appropriate. Under a Bayesian paradigm, Dunson et al. (2003) [11] used a continuation ratio probit model for the informative cluster size and underlying normal linear mixed models for the categorical and continuous sub-unit responses. However, this method does not address the expression and estimation of the interpretable marginal effects of covariates on the risk of a tooth being present and on the risk of disease. Here, we focus on estimating the effects of both cluster- and tooth-specific covariates on whether a particular tooth (at a location) from a randomly selected patient is present, and the effects of these covariates on the risk of disease when that particular tooth is found to be present. In Section 3.2, we propose a two-stage model with useful practical and physical interpretations of the marginal as well as conditional (given cluster-specific random effects) regression functions to express the covariate effects on the probability of a present tooth and on the conditional probability of disease given the tooth being present. We use logit link with the convenient log-odds interpretation (Agresti, 2006) [2] for the conditional regression function for the binary response in each of two stages.

In the usual binary mixed model with logit link and Gaussian random effects, the logistic structure is no longer preserved marginally after integrating out the random effects. We use a bivariate extension of the bridge density of Wang and Louis (2003) [34] instead of the traditional Gaussian density for the random effects in both stages to facilitate the marginal probability of each binary response to approximately preserve the logistic structure. In Section3.3, we apply a maximum likelihood (ML) estimation tool, readily amenable to available statistical softwares (viz. SAS) using routine nonadaptive importance sampling, to analyze the motivating data example. Section 3.4 provides simulation studies to compare our method's performances with competing methods, and to evaluate the robustness of our estimates when our modeling assumptions are not valid. Finally, necessary discussions and concluding comments are in Section 5.

3.2 Proposed Model

3.2.1 Bridge Distribution

One drawback of the regular generalized linear mixed effects model we introduced in Chapter 1.2 is that even though the cluster-specific model given random effects has a logistic form, the marginal model integrating over the random effect will usually not keep the logistic form. Wang and Louis (2003) [34] introduced a bridge distribution that can match the logistic functional shape of both the conditional and marginal for binary response with random cluster effects.

Wang and Louis [34] derived the Bridge distribution in the following way: $G(b)$ is denoted as a distribution for the random effects such that the marginal functional shape would remain the same as the conditional functional shape, i.e.

$$\int H(b+\alpha_s^T X)dG(b) = H(k+\phi\alpha_s^T X), \tag{3.1}$$

where H is the cumulative distribution function and b is the random effects, and α_s is the conditional regression effect. X is the covariates and ϕ is a re-scaling parameter and it is between 0 and 1. If we let $\nu = \alpha_s^T X$ and differentiate the equation above respect to ν then we will have

$$h * g_{-b}(\nu) = \phi h(k+\phi\nu), \tag{3.2}$$

where $*$ denotes the convolution.

By taking Fourier transforms of both sides of the equation and organizing and using the Fourier Inversion Theorem, we would eventually get (in mild conditions):

$$g_\phi(x) = \frac{1}{2x}\int e^{i(k/\phi - x)\xi}\frac{\mathcal{F}h(\xi/\phi)}{\mathcal{F}h(\xi)}d\xi. \tag{3.3}$$

In the logit link, the function $H(\cdot)$ would be $H(\nu) = e^\nu(1+e^\nu)^{-1}$ and $h(\nu) = 3^\nu(1+e^\nu)^{-2}$. By plugging in the Fourier transform we could get the the probability density function (pdf) for bridge distribution:

$$g_\phi(x) = \frac{1}{2\pi}\frac{\sin(\phi\pi)}{\cosh(\phi x) + \cos(\phi\pi)}(0 < \phi < 1, -\infty < x < \infty), \tag{3.4}$$

where $\cosh(x) = \frac{e^x + e^{-x}}{2}$. The mean of this bridge distribution is $\mu = 0$ and the variance is $\sigma_b^2 = \pi^2(\frac{1}{\phi^2} - 1)/3)$. Comparing to the normal distribution and logistic distribution, the bridge distribution has slightly heavier tails than the Normal distribution and lighter tails than logistic [34]. A graph from Wang and Louis paper (2003) [34] is shown in figure 3.2.1.

The ϕ is a cluster-heterogeneity parameter (attenuation of marginal regression effect) such that

$$\alpha_p = \alpha_s * \phi, \tag{3.5}$$

where α_p is the marginal regression effect vector. Therefore the smaller ϕ is, the bigger cluster-heterogeneity we have. Here $\phi = 1$ means there is no heterogeneity within cluster, $\phi = 0$ means maximum heterogeneity within cluster.

The bridge distribution also has a closed form cumulative density function (c.d.f.), expected as

$$G_\phi(x) = 1 - \frac{1}{\pi\phi}\left[\frac{\pi}{2} - \arctan\{\frac{e^{\phi x} + \cos(\phi\pi)}{\sin(\phi\pi)}\}\right], \tag{3.6}$$

and inverse of the cumulative density function as

$$G_\phi^{-1}(x) = \frac{1}{\phi}\log\left[\frac{\sin(\phi\pi x)}{\sin\{\phi\pi(1-x)\}}\right]. \tag{3.7}$$

The bridge distribution has its advantage of keeping both the conditional form and the marginal form as logistic form. However, this property cannot hold under linear combination. We will discuss this in Section 3.2.3.

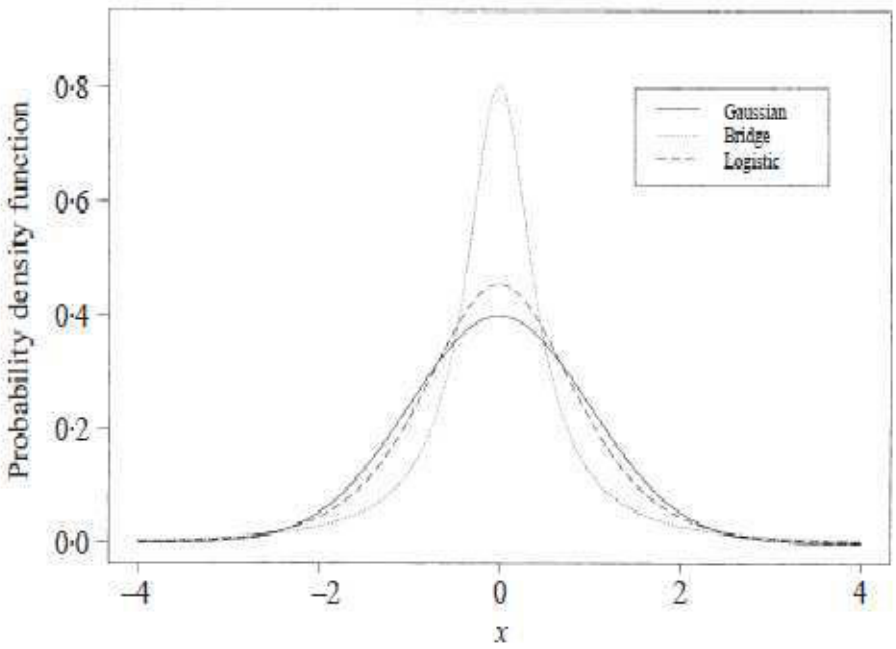

Fig. 1. Probability density functions of the Gaussian, logistic and bridge, for logistic, distributions each with zero mean and unit variance.

Figure 3.1: Excerpt from Wang and Louis (2003) p. 768

3.2.2 Copula Model

Copula models are referred to the 'functions that join the multivariate distribution functions to their one-dimension marginal distributions' [27]. A copula is a multivariate joint cumulative distribution function defined on n-dimensional unit cube $[0, 1]^n$, where each dimension is uniform distributed on the interval $[0, 1]$ [31]. There are many different types of copulas. Here we introduce the Gaussian Copula. We will use this Gaussian Copula to get the joint distribution of the random effects in section 3.2.3 .

Let U, V be uniformly distributed on the interval $[0, 1]$. We have the Gaussian Copula:

$$C_\rho(u, v) = \Phi_{X,Y,\rho}(\Phi^{-1}(u), \Phi^{-1}(v)), \tag{3.8}$$

where $\Phi_{X,Y,\rho}$ is the c.d.f. of a bivariate normal distribution with mean $\mathbf{0}$ variance 1 and correlation ρ, and Φ^{-1} is the inverse cdf of standard normal distribution.

If we differential both sides, we can get the joint density function of u and v with correlation ρ:

$$c_\rho(u, v) = \frac{f_{X,Y,\rho}(\Phi^{-1}(u), \Phi^{-1}(v))}{f_N(\Phi^{-1}(u)) f_N(\Phi^{-1}(v))}, \tag{3.9}$$

where

$$f_{X,Y,\rho}(x, y) = \frac{1}{2\phi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right).$$

3.2.3 Model and Notations

Let Y_{1ij} be the binary response of the presence of the tooth at location $j = 1, ...K$ (sub-unit) for the patient $i = 1, ...n$ (cluster), given by

$$Y_{1ij} = \begin{cases} 1 & \text{when tooth at location } j \text{ of patient } i \text{ is present,} \\ 0 & \text{when tooth at location } j \text{ of patient } i \text{ is absent,} \end{cases}$$

where $Y_{1i} = \sum_{j=1}^{K} Y_{1ij}$ is the total number of teeth present in patient i (cluster-size) and can take values from 0 to K, where K is the number of locations. Also, let Y_{2ij} be the binary response as to whether the tooth at location j for patient i is healthy or diseased given $Y_{1ij} = 1$, i.e., given the tooth is present. In particular, Y_{2ij} is given by

$$Y_{2ij} = \begin{cases} 1 & \text{the tooth at location } j \text{ for patient } i \text{ is healthy (not diseased) given } Y_{1ij} = 1, \\ 0 & \text{the tooth at location } j \text{ for patient } i \text{ is diseased given } Y_{1ij} = 1. \end{cases}$$

Additionally, each tooth has a $d \times 1$ covariate vector $\mathbf{X_{ij}}$, which can be both patient-specific (like age, gender, body mass index, etc), or location-specific (like tooth position indicator with respect to upper/lower jaw). We assume that the distribution of the Stage 1 clustered binary response Y_{1ij} (the indicator of tooth presence), given the patient-specific random effects $\mathbf{B_i} = (B_{1i}, B_{2i})$ to be Bernoulli with success probability p_{1ij}, given by

$$\text{logit}[E(Y_{1ij}|\mathbf{B_i}, \mathbf{X_{ij}})] = \text{logit}(p_{1ij}) = \beta_0 + \beta_1^T \mathbf{X_{ij}} + B_{1i}. \tag{3.10}$$

We assume that the first-stage random cluster-effects B_{1i} follows a bridge distribution (Wang and Louis, 2003) with unknown parameter $0 < \phi_1 < 1$ and density

$$f(B_{1i}|\phi_1) = \frac{1}{2\pi} \frac{\sin(\phi_1 \pi)}{\cosh(\phi_1 B_{1i}) + \cos(\phi_1 \pi)}, \tag{3.11}$$

where $\cosh(x) = \frac{1}{2}(e^x + e^{-x})$.

We assume the parameter ϕ_1 to be common for all clusters to assure the exchangeability of cluster (random) effects on the binary response of the presence of each tooth. Using the Bridge density of (4.4) instead of, say, the normal density for cluster-effects B_{1i}, allows the marginal probability of the tooth being present at each location j to have logit link with regression parameter being proportional to the conditional regression parameter β_1, i.e.,

$$\text{logit}[E(Y_{1ij}|\mathbf{X_{ij}})] = \text{logit}(p^*_{1ij}) = \phi_1 \beta_0 + \phi_1 \beta_1^T \mathbf{X_{ij}}, \tag{3.12}$$

where $0 < \phi_1 < 1$ measures the attenuation of the marginal regression effect $\phi_1 \beta_1$ due to heterogeneity of clusters (patients).

When $Y_{1ij} = 1$ (tooth is present), we also consider a random effects model for the indicator of the present tooth being healthy. We assume that the conditional distribution Y_{2ij} given $Y_{1ij} = 1$ and patient specific $\mathbf{B_i} = (B_{1i}, B_{2i})$ to be Bernoulli with success probability p_{2ij}, given by:

$$\text{logit}[E(Y_{2ij}|Y_{1ij} = 1, \mathbf{X_{ij}}, \mathbf{B_i})] = \text{logit}(p_{2ij}) = \alpha_0 + \alpha_1^T \mathbf{X_{ij}} + B_{2i}, \tag{3.13}$$

where the second-stage random effect B_{2i} also follows a bridge density $f(B_{2i}|\phi_2)$ of (4.4) with unknown parameter $0 < \phi_2 < 1$. Clearly, when $Y_{1ij} = 0$, the random variable Y_{2ij} is not available with probability 1.

Adopting a bivariate density for (B_{1i}, B_{2i}) with bridge density for each marginal is the key to preserve the marginal logistic regression structures for Y_{1ij} in (3.12) as well as for Y_{2ij}. This bivariate bridge distribution is modeled using an inverse probability integral transformation

$$B_{ki} = F_{bk}^{-1}(\Phi(Z_{ki})), k = 1, 2, \tag{3.14}$$

where (Z_{1i}, Z_{2i}) has a bivariate standard normal distribution with mean 0, variance 1 and correlation ρ, $\Phi(\cdot)$ is the cumulative distribution function of the univariate standard normal density and $F_{bk}^{-1}(\cdot), k = 1, 2$ is the inverse cumulative distribution

$$F_{bk}^{-1}(u) = \frac{1}{\phi_k} \log \left\{ \frac{\sin(\phi_k \pi u)}{\sin[\phi_k \pi (1-u)]} \right\}$$

of the marginal bridge distribution for $0 < u < 1$. The transformation of (4.10) assures that the marginal densities of B_{1i} and B_{2i} are Bridge with c.d.f given by

$$F_{bk}(B_{ki}) = 1 - \frac{1}{\pi\phi_k}\left\{\frac{\pi}{2} - \arctan\left[\frac{\exp(\phi_k B_{ki}) + \cos(\phi_k\pi)}{\sin(\phi_k\pi)}\right]\right\} .$$

The correlation parameter ρ induces the within patient association between B_{1i} and B_{2i}. When $\rho = 0$ (independence of B_{1i} and B_{2i}), the corresponding marginal model of Y_{2ij} given $Y_{1ij} = 1$ (integrating over the subject-specific random effect B_{2i}) is given by

$$\text{logit}[E(Y_{2ij}|\mathbf{X_{ij}}, Y_{1ij} = 1)] = \text{logit}(p^*_{2ij}) = \phi_2\alpha_0 + \phi_2\alpha_1\mathbf{X_{ij}}, \tag{3.15}$$

where $\alpha^*_0 = \phi_2\alpha_0$ and $\alpha^*_1 = \phi_2\alpha_1$.

When B_{1i} and B_{2i} are not independent, (i.e., $\rho \neq 0$), $E(Y_{2ij}|Y_{1ij} = 1, \mathbf{X_{ij}})$ integrating over $\mathbf{B_i}$ does not have closed form. We show the detailed derivation of $E(Y_{2ij}|Y_{1ij} = 1, \mathbf{X_{ij}})$ in the Appendix that why this marginal regression function has the logistic form of (3.15) only when B_{1i} and B_{2i} are dependent. However, we can show numerically that $E(Y_{2ij}|Y_{1ij} = 1, \mathbf{X_{ij}})$ still has an approximately logistic form given by $\text{logit}[E(Y_{2ij}|Y_{1ij} = 1, \mathbf{X_{ij}})] \simeq \alpha^*_0 + \alpha^{*T}_1\mathbf{X_{ij}}$ for a wide range of values of ϕ_1 and ϕ_2, and moderate within cluster association ρ. Moreover, when the cluster-heterogeneities for B_{1i} and B_{2i} are moderate (i.e., $\phi_1 \geq 0.6$, $\phi_2 \geq 0.6$), this linear relationship is approximately the same as that in (3.15) with $\alpha^*_1 \simeq \phi_2\alpha_1$. We use Monte Carlo simulations to evaluate the linear form of $\text{logit}[P(Y_2 = 1|Y_1 = 1; x)]$ for $\alpha_0 = \beta_0 = 1$, $\alpha_1 = \beta_1 = -1$ and different values of ϕ_1, ϕ_2, ρ. To simplify the simulations, we assume $\phi_1 = \phi_2 = \phi$. We choose different combinations of ρ and ϕ for different simulations, where $\rho \in \{0.2, 0.4, 0.6, 0.8\}$ and $\phi \in \{0.4, 0.6, 0.8\}$. We consider only one cluster-level covariate X ranging from -3 to 3 (corresponding to ± 3 standard deviation range for a standardized continuous covariate) with increments of 0.1. We use 1000 simulated clusters for each X value and the total number of components within each cluster is 16. Gaussian copula transformation of (4.10) is used to simulate random effects (B_1,B_2) with marginal bridge density. The responses (Y_1, Y_2) for each value of $X = x \in [-3, 3]$ are generated using simulated (B_1, B_2) and simulations from binary distributions with $\text{logit}[E(Y_1|B_1, B_2; x)] = 1 - x + B_1$ and $\text{logit}[E(Y_2|Y_1 = 1, B_1, B_2; x)] = 1 - x + B_2$ with $\alpha_0 = \beta_0 = 1$ and $\alpha_1 = \beta_1 = -1$.

We would like to explore whether $\text{logit}[E(Y_2|Y_1 = 1; x)]$ is linear in x, that is, $\text{logit}[P(Y_2 = 1|Y_1 = 1; x)] - \text{logit}[P(Y_2 = 1|Y_1 = 1; x = 0)] = \alpha^*_1 x$; We would also like to compare the

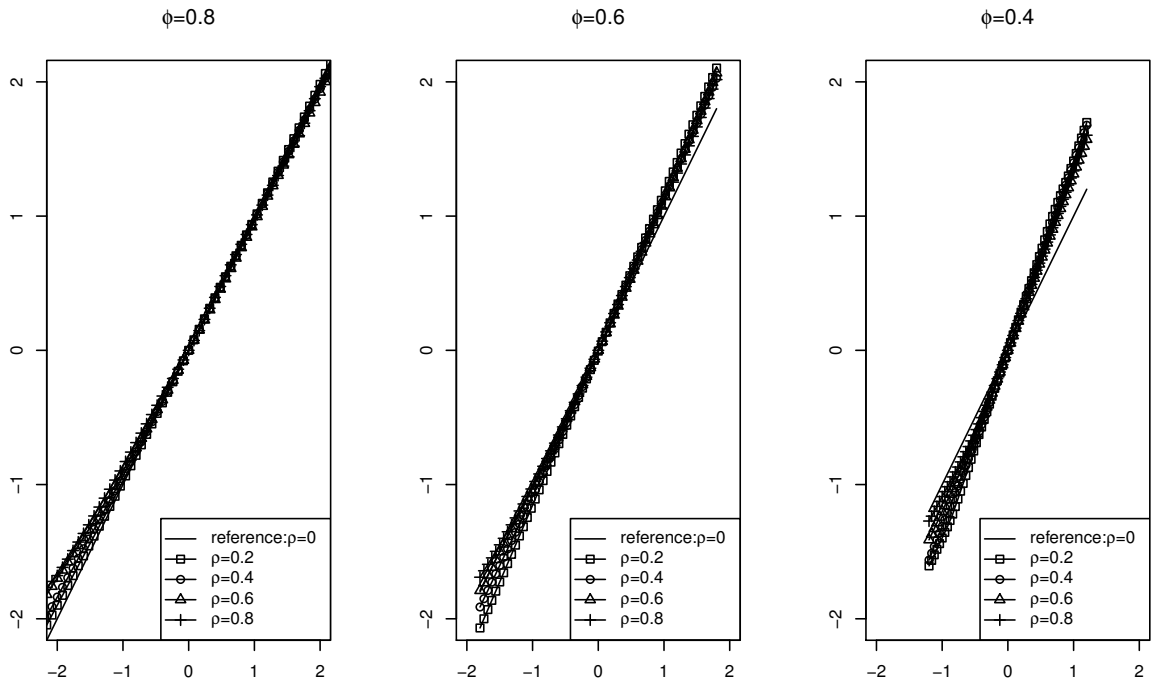

Figure 3.2: Plot of the Monte Carlo approximation of $P[Y_2 = 1|Y_1 = 1, X] - P[Y_2|Y_1 = 1, X = 0]$ (vertical axis) versus $\phi\alpha_1 X$ (horizontal axis) for different values of ϕ and ρ

real attenuation α_1^*/α_1 with estimated ϕ_2 from equation (3.15). We plot the Monte Carlo approximation of $\text{logit}[E(Y_2|Y_1 = 1, X = x)] - \text{logit}[E(Y_2|Y_1, X = 0)]$ versus $\phi_2\alpha_1 x$ for different values of $\phi = \phi_1 = \phi_2$ and ρ in Figure 3.2.

The solid straight line corresponds to the equation in (3.15) when B_1 and B_2 are independent (i.e., $\rho = 0$). From Figure 3.2, we can see that when the heterogeneity level ϕ is moderate ($\phi \geq 0.6$), the marginal model $\text{logit}[P(Y_2 = 1|Y_1 = 1; x)]$ is still approximately linear in x for different values of correlation ρ. Also, $\text{logit}[P(Y_2 = 1|Y_1 = 1; x)]$ will be well approximated by equation (3.15) in this case. When the cluster heterogeneity level is high,corresponding to $\phi \leq 0.4$ in Figure 3.2, we can still approximate the marginal relationship $\text{logit}(P[Y_2 = 1|Y_1 = 1, x]) \simeq \alpha_0^* + \alpha_1^{*T} x$, preserving the logistic form with desirable physical interpretation, except that in this case $\alpha_1^* \neq \phi_2\alpha_1$ and the actual attenuation α_1^*/α_1 is more than ϕ_2, the attenuation expected from (3.15).

The contribution to the likelihood for the patient i (cluster) is given by

$$\int_{\mathbf{B_i}} \left[\prod_{j=1}^{K} \{ \overbrace{P(Y_{1ij} = y_{1ij}|\mathbf{B_i}, \mathbf{X_{ij}})}^{a} [\overbrace{P(Y_{2ij} = y_{2ij}|Y_{1ij} = 1, \mathbf{B_i}, \mathbf{X_{ij}})}^{b}]^{y_{1ij}} \} \right] f_b(\mathbf{B_i}) d\mathbf{B_i}, \qquad (3.16)$$

where $f_b(\mathbf{B_i})$ is the joint density of (B_{1i}, B_{2i}) given as

$$f_Z\{\Phi^{-1}[F_{b1}(b_{1i})], \Phi^{-1}[F_{b2}(b_{2i})]\} \frac{f_{b1}(b_{1i}) \times f_{b2}(b_{2i})}{f_N\{\Phi^{-1}[F_{b1}(b_{1i})]\} \times f_N\{\Phi^{-1}[F_{b2}(b_{2i})]\}} \qquad (3.17)$$

with $f_Z(\cdot, \cdot)$ is the bivariate normal density with mean 0, variance 1 and correlation ρ, $f_{bj}(\cdot)$ is the density of B_j for $j = 1, 2$ and $f_N(\cdot)$ is the standard normal density function. Note that (b) contributes to the likelihood in (3.16) only when $Y_{1ij} = 1$, i.e. the tooth j of patient i is present. The integral in (3.16) does not have a closed form, so ML estimation can be implemented using numerical approximation through routine nonadaptive importance sampling techniques available in PROC NLMIXED of SAS (V9.1).

3.3 Application: Periodontal Data

In this section, we apply our methods to analyze data on the periodontal health study of Gullah-speaking African-American Type-2 diabetic patients [12] (Fernandes et al., 2009), where the health status for each tooth (excluding the 3rd molars) was assessed by hygienists using a periodontal probe [9] (Darby and Walsh, 1995) for six pre-specified sites. We considered only the molars and pre-molars in this analysis which gives us a maximum count of 16 teeth (8 in each jaw) within each patient. The current data analysis uses 113 patients with full covariate information and with at least one molar/pre-molar teeth present.

In our model, two binary responses are (1) presence of a tooth and (2) binary health status when the tooth is present. The health-status of a present tooth is based on whether the mean clinical attachment loss (mean over the 6 sites) of the tooth is $\geq$ 3mm. This is an indicator of moderate to severe periodontitis (Armitage, 1999)[5]. Our patient-specific covariates for both stages 1 and 2 include age (in years), gender (1 = female, 0 = male) , body mass index (BMI), binary glycemic level (1 = high, 0 = controlled), binary poverty indicator (1 = poor, 0 = otherwise), binary brush-floss indicator (1 = if the patient brushed twice and flossed once every day, 0 = otherwise), smoker (1 = present or past smoker, 0 = never a smoker) and the only tooth-specific covariate is the molar location (1 = molar/pre-molar

Table 3.1: Estimates of regression parameters (conditional on random effects) using our proposed ML method for the periodontal data. OR indicates 'odds-ratio' and SE denotes the corresponding standard errors of the estimate.

Parameter	log OR (SE)	P-value
Presence of tooth: Y_1 (Stage 1 response)		
Age	-0.010 (0.008)	0.258
Gender	0.451 (0.214)	**0.037**
Smoker	-0.059 (0.221)	0.790
Molar location	0.0569 (0.102)	0.577
Health status for a present tooth: Y_2 (Stage 2 response)		
Age	0.005 (0.010)	0.619
Gender	0.743 (0.269)	**0.007**
Smoker	0.154 (0.270)	0.569
Molar location	-1.187 (0.158)	**<.0001**
Other Parameters		
ϕ_1	0.857 (0.024)	**<.0001**
ϕ_2	0.852 (0.030)	**<.0001**
ρ	0.218 (0.175)	0.215

tooth in upper jaw, 0 = otherwise). Our objective is to assess the effects of both patient- and tooth-specific covariates on the binary responses defined in equation (3.10) and (3.13). We fit the proposed two-stage model as described in Section 2 and estimate both conditional as well as marginal regression parameters with bivariate Bridge random effects. The estimated ϕ_1 and ϕ_2 also allows us to estimate the marginal regression effect $\phi_1\beta_1$ of (3.12) and the approximate marginal regression effect $\alpha_1^* = \phi_2\alpha_1$ of $\text{logit}[P(Y_{2ij} = 1|Y_{1ij} = 1, \mathbf{X_{ij}})]$ in (3.15).

Table 3.1 provides the ML point estimate of regression parameters (conditional on random effects) in terms of log odds-ratio (OR), corresponding standard error and p-values for our model. We omit the covariates BMI, binary glycemic level, poverty and brush-floss indicators because they have been found to be statistically very insignificant for both responses. For the Stage 1 regression of equation (3.10), gender is the only covariate found to have a significant effect ($p = 0.037$), indicating that females are likely to have more molars/pre-molars (log OR = 0.451, 95% confidence interval (C.I.)= $[0.027, 0.875]$) present as compared

Table 3.2: Estimates of the marginal regression parameters using ML and CWGEE methods for the periodontal data. Estimates in the proposed ML model corresponds to Stage 2 regression. OR indicates 'odds-ratio' and SE denotes the corresponding standard errors of the estimate.

	Proposed ML model		CWGEE	
	log OR (SE)	P-value	log OR (SE)	P-value
Age	0.004 (0.009)	0.6194	0.003(0.009)	0.739
Gender	0.633 (0.223)	**0.007**	0.510(0.217)	**0.019**
Smoker	0.132 (0.230)	0.569	0.167(0.223)	0.454
Molar location	-1.012 (0.132)	**<.0001**	-0.936(0.152)	**<.0001**

to the males. For the Stage 2 regression of equation (3.13), the effect of gender is found to be statistically significant, indicating that under the same latent risk group (cluster), after adjusting for other covariates, the odds of a healthy present molar/pre-molar for females is about 2.10(= exp(0.743)) times higher than the odds for males. The binary indicator 'molar location' is also significant revealing that a molar/pre-molar present in the lower jaw is more likely (log OR $= -1.187$, 95% C.I.$= [-1.501, -0.874]$) to be healthy in comparison to that present in the upper jaw. The estimates (standard error) of the attenuation parameters ϕ_1 and ϕ_2 are 0.857(0.02) and 0.852(0.03) respectively. The estimated value of ϕ_1 confirms that heterogeneity induced by cluster effect causes moderate attenuation of the marginal regression parameter $\phi_1\beta_1$ of (3.12). The estimated value of ϕ_2 also imply that we can use the approximation of $\text{logit}[P(Y_{2ij} = 1|Y_{1ij} = 1, \mathbf{X_{ij}})]$ with $\alpha_1^* = \phi_2\alpha_1$ in (3.15) to estimate the effects of covariates on the risk of disease for a present tooth.

Table 3.2 compares our ML estimates of fixed-effects parameters in terms of log odds-ratio (OR), corresponding standard errors and p-values for the approximate marginal regression parameter $\alpha_1^* = \phi_2\alpha_1$ of (3.15) with estimates of comparable quantities from the CWGEE method of Williamson et al. (2003). Since the WCR method is asymptotically equivalent to the CWGEE, we didn't include it in Table 3.2. The marginal effect (after integrating out cluster effects) of gender and molar location on 'risk of a tooth being healthy' is statistically significant in both models. We like to emphasize that for CWGEE, the 'risk of tooth being healthy' means the probability of being healthy for a 'typical tooth' of a random person. On the contrary, our model's regression functions pertains to the conditional probability of

tooth being healthy given that tooth is present, as expressed in (3.15). We can see from the Table 3.2 that our proposed model have slightly larger estimated log OR for the significant covariates than the CWGEE method, however, these are estimates of log OR of two different events.

To evaluate model diagnostics, we calculate the Pearson residuals of both response variables in our proposed model. The Pearson residuals r_{1ij} for the first response variable (the presence/absence of the tooth), are calculated using the formula: $r_{1ij} = \frac{Y_{1ij}-\widehat{p}_{1ij}}{\sqrt{\widehat{p}_{1ij}(1-\widehat{p}_{1ij})}}$, and plotted against $\widehat{\beta}_0 + \widehat{\beta}_1^T \mathbf{X_{ij}}$ in Figure 3.3. Similarly, we calculate the Pearson residuals $r_{2ij} = \frac{Y_{2ij}-\widehat{p}_{2ij}}{\sqrt{\widehat{p}_{2ij}(1-\widehat{p}_{2ij})}}$ for the tooth health status when the tooth is present. Figure 3.4 gives the Pearson residual plot of r_{2ij} versus $\widehat{\alpha}_0 + \widehat{\alpha}_1^T \mathbf{X_{ij}}$. In both figures, LOWESS smoothing curves for residuals are indistinguishable to the line of zero residuals (parallel to the x-axis), indicating no trend and residual bias over different values of fitted regression effects for both the responses. This suggests that no further (non-linear) transformations of covariates are necessary.

3.4 Simulation Studies

To assess the performance of our proposed ML estimators based on the two-stage model, we conduct two simulation studies to (a) compare the performance with the CWGEE and the ML method for single-stage generalized linear mixed model, and (b) investigate the robustness of our ML estimates when the true distribution of the random effects is different from bridge density.

Each simulated data in our studies has $N = 50$ clusters with one binary cluster-specific covariate X_i indicating cluster exposure status ($X_i = 1$ for exposed, $X_i = 0$ for unexposed). We choose equal number of clusters in both the exposed and unexposed clusters. The maximum possible cluster size (total number of components) for each cluster is 9. We generated two binary responses: (i) Y_{1ij} being the presence of the component j in cluster i and (ii) Y_{2ij} being the health status of component j only when $Y_{1ij} = 1$, using

$$\text{logit}[\Pr(Y_{1ij} = 1|\mathbf{X_{ij}}, B_{1i}, B_{2i})] = 1 - 2x_i + B_{1i} \quad \text{and} \tag{3.18}$$

$$\text{logit}[\Pr(Y_{2ij} = 1|Y_{1ij} = 1, \mathbf{X_{ij}}, B_{1i}, B_{2i})] = 1 - 2x_i + B_{2i}, \tag{3.19}$$

with $\alpha_0 = \beta_0 = 1$ and $\alpha_1 = \beta_1 = -2$. Using equation (4.10), the random effects $\mathbf{B_i} = (B_{1i}, B_{2i})$ in (3.18) and (3.19) are simulated via standard bivariate normal variables

(Z_{1i}, Z_{2i}) with correlation ρ. We set $\phi_1 = \phi_2 = 0.8$ for the Bridge random effects (B_{1i}, B_{2i}), to induce moderate variability within clusters in the simulated data.

Simulation 1:

For each of four different values of $\rho = 0$, 0.3, 0.5 and 0.8, we simulate 100 replicates of the dataset. We fit our proposed two-stage model as in Section 2 using ML method with $\text{logit}[\Pr(Y_{1ij} = 1|\mathbf{X_{ij}}, \mathbf{B_i})] = \beta_0 + \beta_1 \mathbf{X_{ij}} + B_{1i}$ and $\text{logit}[\Pr(Y_{2ij} = 1|Y_{1ij} = 1, \mathbf{X_{ij}}, \mathbf{B_i})] = \alpha_0 + \alpha_1 \mathbf{X_{ij}} + B_{2i}$, with $\mathbf{B_i}$ as distributed in (3.17). For comparison, we also obtain (i) the CWGEE estimate for the marginal logistic regression model $\text{logit}[\Pr(Y_{2ij} = 1|\mathbf{X_{ij}})] = \widetilde{\alpha}_0 + \widetilde{\alpha}_1 x_{ij}$, and (ii) the one-stage ML estimate using $\text{logit}[\Pr(Y_{2ij} = 1|Y_{1ij} = 1, \mathbf{X_{ij}}, B_{2i})] = \alpha_0 + \alpha_1 \mathbf{X_{ij}} + B_{2i}$ based only on the responses Y_{2ij} for cases when $Y_{1ij} = 1$. Note that $\mathbf{X_{ij}} = X_i$, the binary exposure status.

Table 3.3 provides a comparison of our proposed two-stage ML method with (i) and (ii) as described above, for the regression function of Y_2. The one-stage ML method treats the cluster-size $(\sum_{j=1}^{K} Y_{1ij})$ as non-random (given) and uses a single bridge random cluster effect B_{2i} (instead of bivariate $\mathbf{B_i}$) with equation (3.13) for the likelihood contribution. The comparisons are based on the estimate with respect to the average of regression parameter estimates, bias and coverage probability of the 95% confidence interval (CI) over the replicates. When the two random effects are independent, (i.e., $\rho = 0$), the one-stage ML method performs as well as two-stage ML method. However, when ρ is even moderately different from 0, both one-stage ML and CWGEE give biased estimates and the bias increase as $|\rho|$ increases. Our two-stage ML method produces robust estimates for all ρ as far as the relative bias and 95% coverage probabilities are concerned.

Simulation 2:

Here, we investigate the robustness of our ML estimators when the true distribution of the random effects (B_{1i}, B_{2i}) has multiple modes unlike the unimodal shape of the normal density and bridge density of (4.4). The simulated data were generated using the same scheme as that in Simulation 1, except that (B_{1i}, B_{2i}) are generated from a mixture of Gaussian densities $(\frac{1}{2}N(-1, 1) + \frac{1}{2}N(1, 1))$ with two modes, however with finite mean zero. Simulations are conducted for various choices of clusters ($N = 50, 100$ and 500). Table 3.4 shows that when total number of clusters is $N = 50$, the ML regression estimates based on the wrongly specified random effects density of (4.4) are still quite close to the true value

Table 3.3: Results of the simulation study to compare two-stage ML method vs. the CWGEE method: 100 replicated datasets with $N = 50$ clusters and maximum cluster-size $J = 9$ were simulated from the two-stage model of equation (3.10) and (3.13) with bridge density as in equation (3.17); $\phi_1 = \phi_2 = 0.8$, $\beta_1 = \alpha_1 = -2$ and $\rho = 0, 0.3, 0.5, 0.8$.

		True value	Mean estimate	Relative Bias	Coverage probability
			$\rho = 0$		
Proposed model	α_1	-2	-1.992	0.004	92%
One-stage ML	α_1		-2.004	-0.002	93%
CWGEE	$\widetilde{\alpha}_1$		-1.612	0.194	86%
			$\rho = 0.3$		
Proposed model	α_1	-2	-1.960	0.020	95%
One-stage ML	α_1		-1.866	0.068	93%
CWGEE	$\widetilde{\alpha}_1$		-1.546	-0.227	80%
			$\rho = 0.5$		
Proposed model	α_1	-2	-1.985	0.008	97%
One-stage ML	α_1		-1.858	0.710	93%
CWGEE	$\widetilde{\alpha}_1$		-1.528	0.236	80%
			$\rho = 0.8$		
Proposed model	α_1	-2	-1.960	0.02	95%
One-stage ML	α_1		-1.784	0.108	86%
CWGEE	$\widetilde{\alpha}_1$		-1.541	0.230	80%

with a small relative bias (-0.003 for β_1 and 0.065 for α_1). The coverage probabilities of 95% interval estimates of ML method are 0.88 and 0.91 respectively for α_1 and β_1. The coverage probabilities are a little low, which suggests that the estimated standard errors from the inverse of the information matrix are too small. In this case, limited simulations we have performed show that a robust sandwich variance estimate (White, 1982) tends to give less biased estimates of the standard errors, and thus coverage probabilities closer to the nominal 95% coverage probability. As the total number of clusters N increases, the relative biases remain relatively small (less than 10%). Thus, we can see that even when the underlying (true) distribution of random effects is very different from unimodal bridge density, our ML method performs with satisfactory robustness as far as the estimation of regression parameters are concerned.

Table 3.4: Results of simulation study of misspecification of our proposed model using 100 replicated datasets with maximum cluster-size $J = 9$ random effects simulated from a mixture of Normal densities: $\frac{1}{2}N(-1,1) + \frac{1}{2}N(1,1)$.

	True value	Mean estimate	Relative Bias	Coverage probability
		N=50		
β_1	-2	-1.994	0.003	88%
α_1	-2	-1.870	0.065	91%
		N=100		
β_1	-2	-1.918	0.041	91%
α_1	-2	-1.836	0.082	86%
		N=500		
β_1	-2	-1.897	0.0514	84%
α_1	-2	-1.865	0.068	80%

3.5 Conclusions

Motivated by a periodontal study where a patient's tooth at a particular location of the jaw can be either absent or diseased or healthy, we develop a two-stage random effects regression model for clustered binary responses with random cluster sizes. In our model, the Stage 1 logistic regression function expresses the effects of covariates and random cluster effects on the presence of a component (tooth) within a cluster (patient). The Stage 2 logistic regression function explains the effects of covariates and cluster effects on the binary disease status of a tooth given that the tooth is present. The effects of clustering is efficiently handled through cluster-specific random effects at both stages and allowing a within cluster association among these two random effects. We also show numerically that when the correlation between two random effects is not high ($\rho \leq 0.8$) and the cluster-heterogeneity level is at least moderate, the marginal regression parameter α_1^* for the second-stage response is approximately proportional to the conditional regression parameter α_1, as in (3.15).

The existing methods like GEE, CWGEE or Bayesian methods either estimate the marginal or the conditional (given random effects) regression parameters of the response at Stage 2. The attenuation of the marginal regression effect (after integrating random patient-effect) is not captured in these methods. Instead, our method provides conditional (on the

random intercepts) as well as marginal inference for model parameters with approximate log-odds interpretation of the covariate effects at both stages. Appropriately, our method is useful when the cluster size is also considered as an important response by itself, as in the dental study. Unlike the CWGEE method, our method evaluates the probability for a present tooth and the conditional probability that the tooth is healthy given being present. For some studies, particularly when the maximum number of components are fixed and the components are not exactly exchangeable (e.g., teeth from different locations within the mouth, patients from different geographical areas, etc), it may not make sense to talk about the risk of a 'typical' cluster member. Our approach which evaluates the risk on a cluster component when it is present is more appropriate. Because the event of interest of our approach is very different from the event of interest of a CWGEE or some other missing data approach, there is no direct relationship between the regression parameters of these approaches. The comparison of regression estimates of the simulation study of Section 4 to some extent demonstrates how much these regression parameters from different models can differ from each other, rather than relative merit of one estimator over the other.

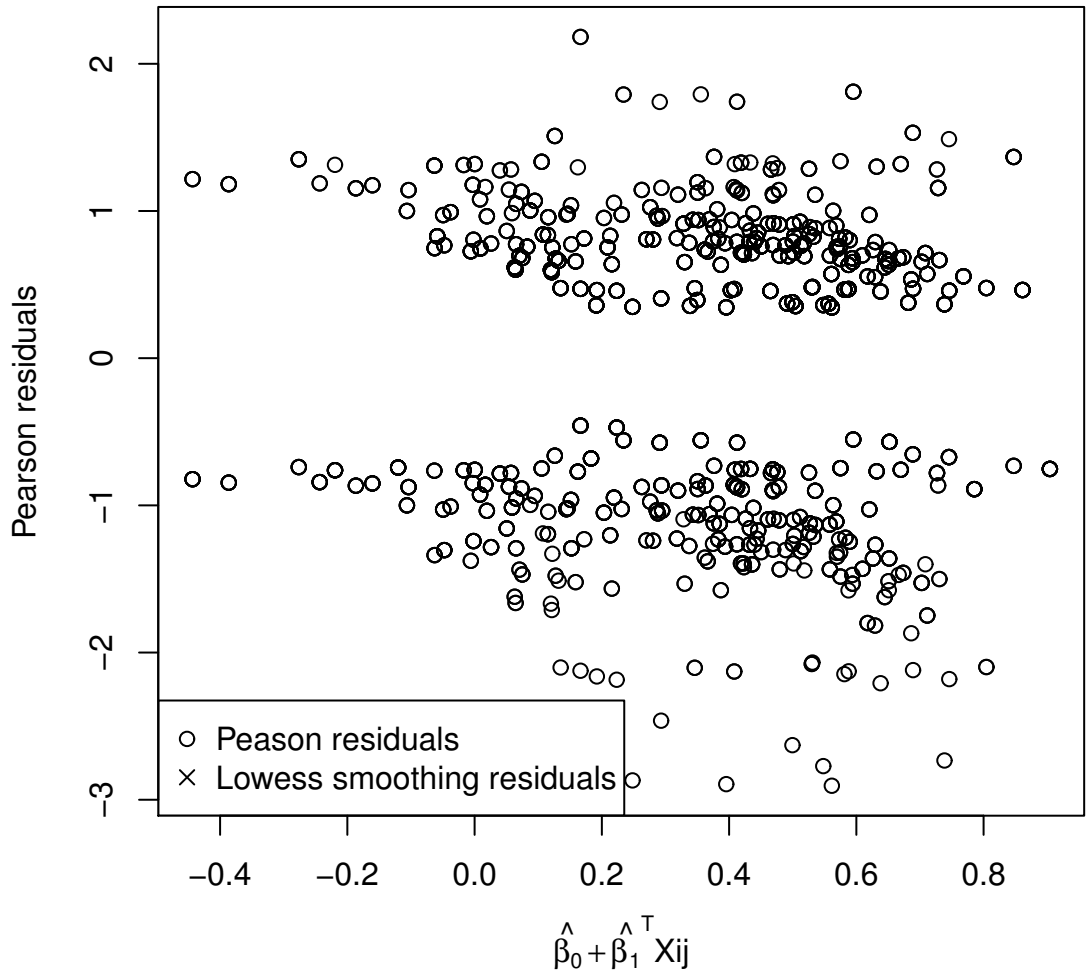

Figure 3.3: Pearson residual plot for Stage-1 regression (presence/absence status of a tooth) versus the estimated linear regression $\widehat{\beta}_0 + \widehat{\beta}_1^T \mathbf{X_{ij}}$.

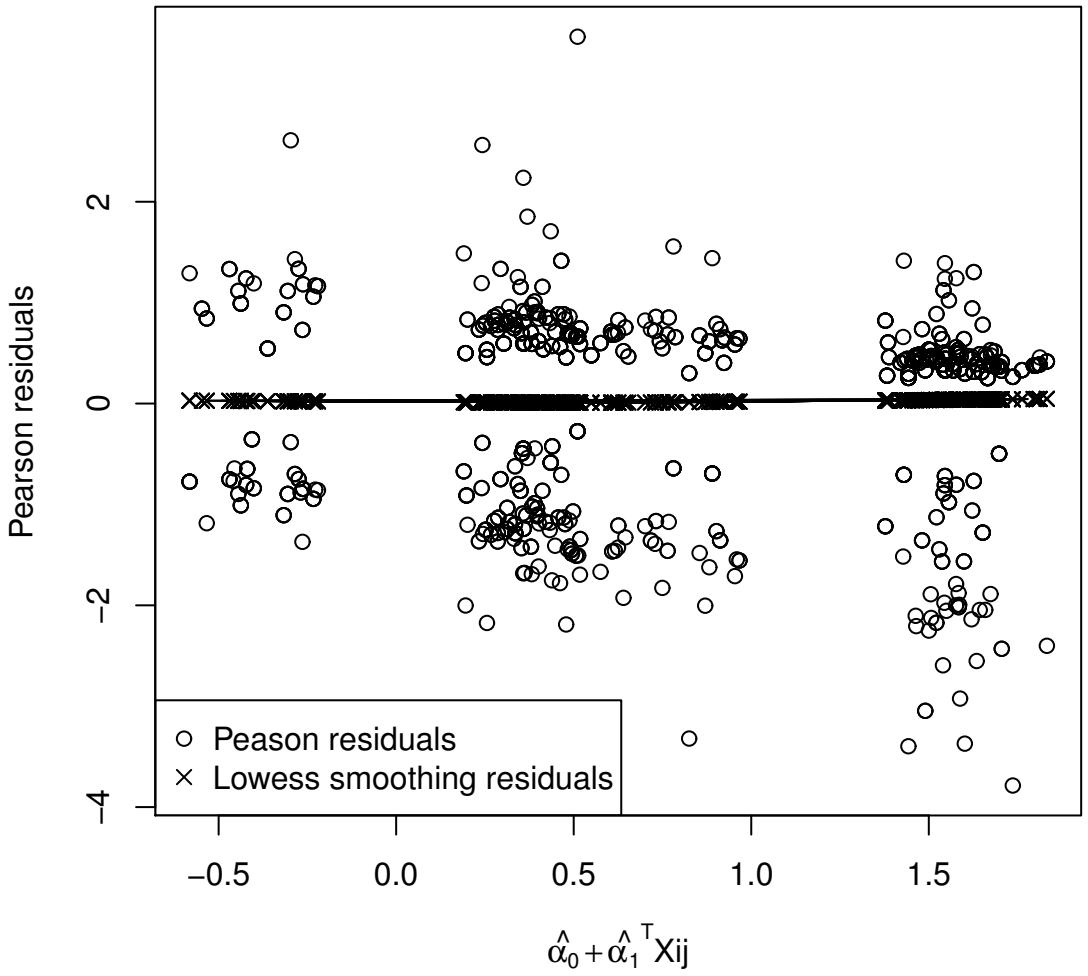

Figure 3.4: Pearson residual plot for Stage-2 regression (health status of present tooth) versus the estimated linear regression $\widehat{\alpha_0} + \widehat{\alpha}_1^T \mathbf{X_{ij}}$.

CHAPTER 4

Analysis of Longitudinal Data with Informative Missing

4.1 Introduction

Prospective longitudinal studies involve observing information on the same individual over a period of time. Some of the individuals may fail to show up during the follow up visits or fail to provide information at one or more of the scheduled visits. At the same time, individuals may drop out of the study for various reasons. Using the terminology of [26], if the missing data is missing completely at random (MCAR) or missing at random (MAR), where the missingness is unrelated to the responses at all or is dependent to the missing responses given observed responses and other covariates, valid inference can still be drawn using likelihood or Bayesian approaches if all covariates contributing to the missing data process are included in the model [21]. However, in some cases, the probability of non-response is related to the missing outcome values themselves, which is termed non-random missing. Ignoring the mechanism then invalidates inferences.

We consider a longitudinal study of HIV-infected patients, where likely a non-random mechanism applies, although definitively establishing the nature of the missing-data mechanism is fundamentally non possible. The data come from two randomized phase III double-blinded studies designed to compare two therapeutic treatments, zidovudine (AZT) and didanosine (DDI) [19, 15, 13]. The response of interest is dichotomized CD4 counts (> 200 versus ≤ 200 cells per cubic millimeter), which is a reasonable predictor for development of opportunistic infections. The cutoff 200 has been used as a standard threshold for clinicians [29]. CD4 count below 200 is indicative for high risk of opportunistic infections. In our analysis, there are 431 patients at baseline (week 0), and they are followed up every week up to at most 5 weeks. Each patient has measurement at baseline, but may miss one or more

visits during the 5 weeks of follow up. Out of these 431 patients, only 202 have measurements at all 5 follow-up visits. It is prudent to allow for non-random missingness.

An appropriate model for non-random missingness should address the dependency between primary responses and missingness. The full density function of the responses and missingness process can be specified within a selection model framework [30, 26] or a pattern-mixture model framework [24, 25]. That said, in this paper, we focus on the third framework: shared random effects. This framework accommodates non-random missingness by introducing random effects shared between the outcome and missingness models [4]. TenHave [**?**] proposed such models for binary longitudinal responses; these authors made use of the logit link. Albert [4] considered longitudinal data in a long time sequence and extended Ten Have's model by incorporating Gaussian auto-regressive latent processes instead of subject-specific random effects. However, neither of these two methods yields interpretable marginal effects of covariates on the responses, which nevertheless is the main interest in many longitudinal studies. We therefore focus on estimating the effects of both time-stationary and time-varying covariates on both binary response of interest and missingness. Our proposed model is novel in that the expression for, at the same time, the conditional and marginal effects of covariates on the binary longitudinal response have closed forms and in addition allow a log-odds interpretation. A class of so-called *Bridge distributions* is used [34] instead of the traditional Gaussian densities for the random effects, to facilitate the marginal probability of each binary response preserve the logistic structure.

A Bayesian inferential framework is used in this article, which has a number of advantages. First, we can attain exact posterior distributions of the parameters. Unlike maximum likelihood methods, Bayesian estimates attained from the posterior distributions by Markov Chain Monte Carlo (MCMC) methods, such as means and quantiles, do not depend on asymptotic normality assumptions [10]. At the same time, Bayesian methods allow us to incorporate informative prior information of parameters from historical control or pilot studies, thus efficiently using all available information.

This chapter proceeds as follows. In Section 4.2, we formulate the model. A Bayesian statistical framework is presented in Section 4.3. The model is applied to the HIV data in Section 4.4. Section 4.5 concludes. .

4.2 Model Formulation

Suppose that there are $t = 1, \ldots, T$ intended visits for each subject $i = 1, \ldots, n$. Let Y_{it} denotes the binary longitudinal response for the ith subject at time t. Let M_{it} be the missing indicator for the ith subject at time t, i.e.,

$$M_{it} = \begin{cases} 0 & i\text{th subject is observed at time } t, \\ 1 & i\text{th subject is missing at time } t. \end{cases} \tag{4.1}$$

We assume that each subject i has a $d \times 1$ covariate vector $\mathbf{x}_{it}$ at time t for the outcome and another one, a $p \times 1$ covariate vector $\mathbf{w}_{it}$, for missingness. They can have some but not all covariates in common; they can include both time stationary (such as gender, baseline age) and time-varying (such as time or weight) covariates.

Non-random missingness is induced by a subject-specific random effect $\mathbf{B}_i = (B_{1i}, B_{2i})$, where B_{1i} is the random effect for Y_{it} and B_{2i} is the random effect for M_{it}. Given $\mathbf{B}_i$, we assume that Y_{it} and M_{it} are independent. Then, the marginal distribution of (Y_{it}, M_{it}) for the ith subject at visit time t given by $f(y_{it}, m_{it}|\mathbf{x}_{it}, \mathbf{w}_{it})$ can be computed as

$$f(y_{it}, m_{it}|\mathbf{x}_{it}, \mathbf{w}_{it}) = \int_{\mathbf{B}_i} f(y_{it}|\mathbf{x}_{it}, B_{1i}) f(m_{it}|\mathbf{w}_{it}, B_{2i}) f_{\mathbf{B}}(\mathbf{b}) d\mathbf{b},$$

where $f(y_{it}|\mathbf{x}_{it}, B_{1i})$ is the conditional model for Y_{it} given $\mathbf{B}_i$ and $f(m_{it}|\mathbf{w}_{it}, B_{2i})$ is the conditional model for M_{it} given random effect $\mathbf{B}_i$. Finally, $f_{\mathbf{B}_i}(\mathbf{b})$ is the density function for the random effect $\mathbf{B}_i$.

The relationship between B_{1i} and B_{2i} is the key to the non-random missingness. There are different ways to relate B_{1i} and B_{2i}. In TenHave's [**?**] and Albert's [4] papers, the same random effect $B_i = B_{1i} = B_{2i}$ was assigned to both Y_{it} and M_{it}, but with unknown coefficient parameter for $B_{2i} = B_i$. In the following, we consider this as well as another way to construct B_{1i} and B_{2i}, and then discuss the relative merits.

4.2.1 Model 1

In this model, we assume the same random effect $B_i = B_{1i} = B_{2i}$ for both Y_{it} and M_{it}, $t = 1, \ldots, T$. The distribution of Y_{it} given B_i is taken to be Bernoulli with success probability p_{1it}, identified by

$$\text{logit}[P(Y_{it} = 1|B_i, \mathbf{x}_{it})] = \boldsymbol{\alpha}^T \mathbf{x}_{it} + B_i. \tag{4.2}$$

where $\boldsymbol{\alpha}$ is a $d \times 1$ vector regression parameters of $\mathbf{x}_{it}$. Likewise, we let M_{it} given B_i follow a Bernoulli with probability of success p_{2it}:

$$\text{logit}P(M_{it} = 1|B_i, \mathbf{w}_{it}) = \boldsymbol{\beta}^T \mathbf{w}_{it} + \gamma B_i, \tag{4.3}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector regression parameters of $\mathbf{w}_{it}$ and γ is the coefficient for the random effect B_i. The random effect B_i is the link between Y_{it} and M_{it}. If $\gamma = 0$, then missingness is independent of B_i and therefore does not relate to Y_{it}. If $\gamma \neq 0$ then, given the model would be correctly formulated, this could be seen as some evidence for non-random missingness. We assume that the random intercept B_i for Y_{it} and M_{it} follows a so-called Bridge distribution [34] with unknown parameter $0 < \eta < 1$ and density

$$f(u|\eta) = \frac{1}{2\pi} \frac{\sin(\eta\pi)}{\cosh(\eta u) + \cos(\eta\pi)}, \tag{4.4}$$

where $\cosh(x) = \frac{1}{2}(e^x + e^{-x})$.

To ensure exchangeability of subject random effects on the binary longitudinal response, we assume η is the same across all subjects. Using the Bridge density (4.4) instead of, say, the normal density for random effect B_i, allows the marginal probability of the binary longitudinal response at each visit to have logit link with regression parameter being proportional to the conditional regression parameter $\boldsymbol{\alpha}$, i.e.,

$$\text{logit}[P(Y_{it} = 1|\mathbf{x}_{it})] = \eta \boldsymbol{\alpha}^T \mathbf{x}_{it}, \tag{4.5}$$

where $0 < \eta < 1$ measures the attenuation of the marginal regression effect $\eta\boldsymbol{\alpha}$ due to heterogeneity among subjects. However, Bridge distribution is not closed under linear combination [34]. Therefore, the marginal model of M_{it} would not maintain the logistic form. In our proposed second model, this drawback will be overcome.

4.2.2 Model 2

In the second model, we assume B_{1i} and B_{2i} to be generated from same normal random generator Z_i but with different Bridge parameters $0 < \eta_1 < 1$ and $0 < \eta_2 < 1$. Since η_1 and η_2 account for the variation of B_{1i} and B_{2i}, there is no need to include coefficients for either random effects in their models.

We assume the distribution of our intended binary longitudinal response Y_{it}, given the random intercept $\mathbf{B}_i$ to be Bernoulli with success probability p_{1it}, given by

$$\text{logit}[P(Y_{it} = 1|\mathbf{B}_i, \mathbf{x}_{it})] = \boldsymbol{\alpha}^T \mathbf{x}_{it} + B_{1i}. \tag{4.6}$$

We assume that the random intercept B_{1i} for Y_{it} follows a Bridge distribution [34] with unknown parameter $0 < \eta_1 < 1$ and density as in (4.4) with $\eta = \eta_1$. As before, to guarantee that exchangeability holds, we keep η_1 constant across all subjects. Also here, the Bridge density (4.4) ensures the marginal probability of the binary longitudinal response at each visit to have logit link with regression parameter proportional to the conditional regression parameter $\boldsymbol{\alpha}$, i.e.,

$$\text{logit}[P(Y_{it} = 1|\mathbf{x}_{it})] = \eta_1 \boldsymbol{\alpha}^T \mathbf{x}_{it}, \tag{4.7}$$

where $0 < \eta_1 < 1$ measures the attenuation of the marginal regression effect $\eta_1 \boldsymbol{\alpha}$ due to heterogeneity among subjects.

We also let the distribution of missing indicator M_{it} given random effect $\mathbf{B}_i$ to be Bernoulli with probability of success p_{2it}, given by:

$$\text{logit}P(M_{it} = 1|\mathbf{B}_i, \mathbf{w}_{it}) = \boldsymbol{\beta}^T \mathbf{w}_{it} + B_{2i}, \tag{4.8}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters of $\mathbf{w}_{it}$, and the random intercept B_{2i} follows a Bridge distribution with unknown parameter $0 < \eta_2 < 1$ and density as in (4.4) with $\eta = \eta_2$. Again, we assume η_2 is the same across all subjects. The marginal model of M_{it} can now be expressed as:

$$\text{logit}[P(M_{it} = 1|\mathbf{w}_{it})] = \eta_2 \boldsymbol{\beta}^T \mathbf{w}_{it}. \tag{4.9}$$

where $0 < \eta_2 < 1$ measures the attenuation of the marginal regression effect $\eta_2 \boldsymbol{\beta}$ due to heterogeneity among subjects.

Suppose that B follows a Bridge distribution with parameter η. Then B has mean 0 and variance $\sigma_b^2 = \pi^2(\eta^{-2} - 1)/3$. Therefore, when $\eta = 1$, the variable B degenerates to a constant 0. In this model, when $\eta_1 = 1$ or $\eta_2 = 1$, Y_{it} or M_{it} would not depend on $\mathbf{B}_i$, which under the specified model corresponds to the absence of non-random missingness.

The advantage of this model is that both marginal models of Y_{it} and M_{it} preserve logistic form, while in the first model only the marginal model of Y_{it} does. However, since B_{1i} and B_{2i}

are generated from same normal random effects with different variation scale, the correlation between B_{1i} and B_{2i} could not be negative. Therefore, only when the association between Y_{it} and M_{it} is positive, can this model be fitted appropriately.

To link Y_{it} and M_{it} using the random effects, we generate B_{1i} and B_{2i} from the same standard normal generator but with potentially different attenuation degrees. Random effects B_{1i} and B_{2i} can be modeled using the inverse probability integral transformation:

$$B_{ki} = F_{\eta_k}^{-1}[\Phi(Z_i)], \qquad k = 1, 2, \tag{4.10}$$

where Z_i has a $N(0,1)$, $\Phi(\cdot)$ is the cumulative distribution function of the univariate standard normal distribution and $F_{\eta_k}^{-1}(\cdot)$, k=1,2 is the inverse cumulative distribution

$$F_{\eta_k}^{-1}(u) = \frac{1}{\eta_k}\log\left\{\frac{\sin(\eta_k\pi u)}{\sin[\eta_k\pi(1-u)]}\right\} \tag{4.11}$$

of the marginal Bridge distribution for $0 < u < 1$. The transformation of (4.10) ensures that the marginal densities of B_{1i} and B_{2i} are Bridge with c.d.f. given by

$$F_{\eta_k}(B_{ki}) = 1 - \frac{1}{\pi\eta_k}\left\{\frac{\pi}{2} - \arctan\left[\frac{\exp(\eta_k B_{ki}) + \cos(\eta_k\pi)}{\sin(\eta_k\pi)}\right]\right\} .$$

4.3 Bayesian Estimation

4.3.1 Bayesian Estimation for Model 1

In this section, prior distributions are selected for the parameters. The likelihood and posterior distribution will be derived. Practically, we use the MCMC algorithm available in the `OpenBUGS` [33] package to obtain the posteriors. Parameters in Model 1 include $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \eta)$ where $\boldsymbol{\theta_1} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma)$. Assuming all parameters are independent, we assign priors:

$$\pi(\boldsymbol{\theta}) = \pi_1(\boldsymbol{\theta_1})\pi_2(\eta), \tag{4.12}$$

where π_1 has a multivariate normal density $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and π_2 is $U(0,1)$.

Making use of the conditional independence between outcome and missingness, the likelihood contribution for the ith subject can be written as:

$$\begin{aligned} L(\boldsymbol{\theta}, Z_i|\mathbf{Y_i}, \mathbf{M_i}) &= \prod_{t\in\Omega_i} f_{Y_{it}|B_i}(y_{it}|b_i)\prod_{t=1}^{T} f_{M_{it}|B_i}(m_{it}|b_i) \\ &= \prod_{t\in\Omega_i}\frac{\exp\{(\boldsymbol{\alpha}^T\mathbf{X_{it}} + b_i)y_{it}\}}{1+\exp(\boldsymbol{\alpha}^T\mathbf{X_{it}} + b_i)}\prod_{t=1}^{T}\frac{\exp\{(\boldsymbol{\beta}^T\mathbf{w}_{it} + \gamma b_i)m_{it}\}}{1+\exp(\boldsymbol{\beta}^T\mathbf{w}_{it} + \gamma b_i)}, \end{aligned} \tag{4.13}$$

where Ω_i is the set of measurement occasions that patient i is observed, $b_i = F_\eta^{-1}\Phi(z_i)$.

Observations across subjects are independent. Therefore, letting $\mathbf{Y} = (\mathbf{Y_1}, \ldots, \mathbf{Y_n})$ and $\mathbf{M} = (\mathbf{M_1}, \ldots, \mathbf{M_n})$, the full likelihood of $\boldsymbol{\theta}$ based on ($\mathbf{Y}$, $\mathbf{M}$) is:

$$L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{M}) = \prod_{i=1}^{n} \int_{Z_i} L(\boldsymbol{\theta}, Z_i|\mathbf{Y_i}, \mathbf{M_i}) f_{Z_i}(z_i) dZ_i, \tag{4.14}$$

where $f_{Z_i}(\cdot)$ is the standard normal density. Letting $\mathbf{Z} = (Z_1, \ldots, Z_n)$, the joint posterior distribution of $(\boldsymbol{\theta}, \mathbf{Z})$ can be written as:

$$P(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y}, \mathbf{M}) \propto \prod_{i=1}^{n} L(\boldsymbol{\theta}, Z_i|\mathbf{Y_i}, \mathbf{M_i}) f_{Z_i}(z_i) \pi(\boldsymbol{\theta}). \tag{4.15}$$

We can draw posterior samples, using MCMC, from the posterior distribution $P(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{Y}, \mathbf{M})$.

4.3.2 Bayesian Estimation for Model 2

We proceed in a way similar to what has been spelled out for Model 1. Parameters in Model 2 include $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \eta_1, \eta_2)$ where $\tilde{\boldsymbol{\theta}}_1 = (\boldsymbol{\alpha}, \boldsymbol{\beta})$. The prior now takes the form, assuming independence:

$$\pi(\tilde{\boldsymbol{\theta}}) = \pi_1(\tilde{\boldsymbol{\theta}}_1)\pi_2(\eta_1)\pi_3(\eta_2), \tag{4.16}$$

where π_1 has a multivariate normal density $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ and further π_2 and π_3 are independent $U(0, 1)$. Using the conditional independence properties, the likelihood contribution for subject i equals:

$$\begin{aligned} L(\tilde{\boldsymbol{\theta}}, Z_i|\mathbf{Y_i}, \mathbf{M_i}) &= \prod_{j \in \Omega_i} f_{Y_{it}|B_{1i}}(y_{it}|b_{1i}) \prod_{j=1}^{l} f_{M_{it}|B_{2i}}(m_{it}|b_{2i}) \\ &= \prod_{j \in \Omega_i} \frac{\exp\{(\boldsymbol{\alpha}^T \mathbf{x}_{it} + b_{1i}) y_{it}\}}{1 + \exp(\boldsymbol{\alpha}^T \mathbf{x}_{it} + b_{1i})} \prod_{j=1}^{l} \frac{\exp\{(\boldsymbol{\beta}^T \mathbf{w}_{it} + b_{2i}) m_{it}\}}{1 + \exp(\boldsymbol{\beta}^T \mathbf{w}_{it} + b_{2i})}, \end{aligned} \tag{4.17}$$

with notation identical to what has been introduced for Model 1 and further $B_{2i} = F_{\eta_2}^{-1}\Phi(Z_i)$. The full likelihood takes similar expression as (4.14) except that $\boldsymbol{\theta}$ has been replaced by $\tilde{\boldsymbol{\theta}}$. Let $\mathbf{Z} = (Z_1, ..., Z_n)$ as earlier, the joint posterior distribution of $(\tilde{\boldsymbol{\theta}}, \mathbf{Z})$ can be written as:

$$P(\tilde{\boldsymbol{\theta}}, \mathbf{Z}|\mathbf{Y}, \mathbf{M}) \propto \prod_{i=1}^{n} L(\tilde{\boldsymbol{\theta}}, Z_i|\mathbf{Y_i}, \mathbf{M_i}) f_{Z_i}(z_i) \pi(\tilde{\boldsymbol{\theta}}), \tag{4.18}$$

where $L(\tilde{\boldsymbol{\theta}}, Z_i|\mathbf{Y_i}, \mathbf{M_i})$ is given in (4.17).

4.4 Data Analysis

4.4.1 Analysis using Model 1

We will analyze the data introduced in Section 1, which concerns 431 patients with AIDS or AIDS complex. Recall that the outcome is dichotomized CD4 count, with the dichotomization rule described in Section 1. Furthermore, patients were randomized into two treatments, AZT and DDI. To this effect, define an indicator variable AZT_i that takes the value 1 for patients on AZT and 0 for those on DDI. Measurements were taken at weeks $t = 0, 1, \ldots, 5$ with $t = 0$ referring to baseline. Also dichotomized age is considered a predictor with $\text{AGE}_i = 1$ if the patient's baseline age is greater than or equal to 35 and 0 otherwise.

We first fit Model 1 from Section 2, with the predictor:

$$\text{logit}\Pr(Y_{it} = 1|\mathbf{X}_{it}, B_i) = \alpha_0 + \alpha_1 * \text{AGE}_i + \alpha_2 * t + \alpha_3 * \text{AZT}_i * t + B_i. \tag{4.19}$$

Before baseline, all patients received treatment DDI. After baseline, each patient is assigned to AZT and DDI randomly. Therefore, there should be no AZT effect at baseline. Because we have added the interaction between AZT and time to our model, the interpretation of the AZT main effect is the drug's effect at baseline, whence we exclude it.

The missing-indicator model is:

$$\text{logit}\Pr(M_{it} = 1|\mathbf{X}_{it}, B_i) = \beta_0 + \beta_1 * \text{AGE}_i + \beta_2 * t + \beta_3 * \text{AZT}_i + \gamma B_i, \tag{4.20}$$

where γ is the parameter allowing for non-random missingness under the purported model.

`OpenBUGS` is used to fit the model and generate samples from the joint posterior distribution and random effects, using Gibbs sampling method. Three chains with varied initial values were ran and the convergence checked using Gelman and Rubin's convergence diagnostics $\hat{R}$ [8, 16]. Upon convergence, another 3000 iterations were ran to compute the posterior estimates. The estimate $\widehat{\gamma} = -0.453$ with 95% credible interval $(-0.57, -0.35)$, suggesting that the binary response and the missingness indicator are negatively associated. Given that Model 2 only allows for positive association, we redefine the binary response as $Y^*_{it} = 1 - Y_{it}$. Model 1 is re-fitted using Y^* and M and the posterior summaries of the parameters given in Table 4.5. Summary statistics in Table 4.5 imply that, generally, CD4 counts decreases over time, indicating disease progression. There is increasing missingness

over time as well, and patients randomized to AZT tend to have slightly higher missingness overall. $\widehat{\gamma}$ is positive, showing the new response Y^* is positively associated with missingness. The posterior mean of η is 0.435 with 95% credible interval $(0.37, 0.50)$, indicating moderate to strong heterogeneity among patients.

4.4.2 Analysis using Model 2

Let us turn to Model 2. Also here, we use Y^* to ensure positive association between our binary responses. The logits now slightly change to:

$$\begin{aligned}
\text{logit}\,\Pr(Y_{it}^* = 1|\mathbf{X}_{it}, B_i) &= \alpha_0 + \alpha_1 * \text{AGE}_i + \alpha_2 * t + \alpha_3 * \text{AZT}_i * t + B_{1i}, \\
\text{logit}\,\Pr(M_{it} = 1|\mathbf{X}_{it}, B_i) &= \beta_0 + \beta_1 * \text{AGE}_i + \beta_2 * t + \beta_3 * \text{AZT}_i + B_{2i}.
\end{aligned}$$

Recall that the marginal models are $\text{logit}[P(Y_{it}^* = 1|\mathbf{x}_{it})] = \eta_1 \boldsymbol{\alpha}^T \mathbf{x}_{it}$ and $\text{logit}[P(M_{it} = 1|\mathbf{w}_{it})] = \eta_2 \boldsymbol{\beta}^T \mathbf{w}_{it}$. In line with our approach to Model 1, we used OpenBUGS to fit the model; convergence and posterior computations were performed in a similar fashion. Posterior summaries are provided in Table 4.5. We infer that, in general, CD4 counts decrease over time, indicating disease progression. Missingness increased over the follow-up period, while patients randomized to AZT tend to have a slightly higher chance of missing follow-up visits. The posterior mean of η_1 is 0.458 and the 95% credible interval is $(0.38; 0.56)$, indicating strong heterogeneity of CD4 counts among patients. Posterior mean of η_2 is 0.704 with 95% credible interval $(0.65, 0.76)$, showing moderate heterogeneity of missing-data mechanism among patients.

The tables show that there is strong similarity between both models. We employ the Bayes factor for model selection. Let M_ℓ denotes Model ℓ, $(\ell = 1, 2)$. The Bayes factor [20] is:

$$\text{BF}(M_1, M_2) = \frac{\Pr(\mathbf{y}_{obs}|M_1)}{\Pr(\mathbf{y}_{obs}|M_2)} = \frac{\int \Pr(\mathbf{y}_{obs}|\boldsymbol{\theta}_1, M_1)\pi(\boldsymbol{\theta}_1|M_1)d\boldsymbol{\theta}_2}{\int \Pr(\mathbf{y}_{obs}|\boldsymbol{\theta}_2, M_2)\pi(\boldsymbol{\theta}_2|M_2)d\boldsymbol{\theta}_2},$$

where $\boldsymbol{\theta}_\ell$ is a vector of all parameters in M_ℓ and $\pi(\boldsymbol{\theta}_\ell|M_\ell)$ are prior densities. The Bayes factor $\text{BF}(M_1, M_2)$ is very close to 0, providing strong evidence that Model 2 is better than Model 1.

4.5 Conclusions

In this article, we proposed a shared random-effects framework for jointly modeling binary longitudinal data and missingness, based on shared Bridge random variables. This framework

builds upon earlier work by Albert *et al.* and Ten Have *et al.* These authors only estimate the conditional (given random effects) regression parameters of the binary longitudinal response. The marginal (integrating over random effects) regression parameters and the attenuation of the marginal regression effect are not explicit in these methods. Instead, our model provides both the conditional and marginal regression parameter estimates with convenient log-odds interpretation of the covariate effects for both the binary longitudinal response. Exchangeable correlation is assumed between visits for each subject. The framework is not restricted to the models considered here. Indeed, we can extend the approach by incorporating a time-varying structure among observations over time (for instance, using banded correlation), when the number of visits per subject is large. Given that, in our data set, each person only has 6 visits, there is not enough information to apply the time-varying structure model. A Bayesian framework was used to allow flexibly incorporating prior information from historical studies and conveniently make inferences on posterior distributions.

Table 4.1: Posterior summaries of the parameters for Model 1

	Parameter	Mean	Median	Standard deviation	95% Cr.I.[a]
Binary response:[b]	AGE	-0.634	-0.650	0.399	(-1.37, 0.13)
	TIME	0.321	0.319	0.083	(0.15,0.49)
	AZT*TIME	-0.160	-0.161	0.119	(-0.39,0.06)
Missing indicator:	AGE	-0.120	-0.108	0.192	(-0.25,0.52)
	TIME	0.598	0.596	0.043	(0.51,0.68)
	AZT	0.428	0.433	0.175	(0.08,0.79)
	γ	0.453	0.449	0.056	(0.35,0.57)
Others:	η	0.435	0.434	0.036	(0.37, 0.50)

[a] 'Cr. I.' is an abbreviation of Credible Interval.

[b] To make results of Model 1 and Model 2 comparable, we let binary response be 0 if CD4> 200 such that the association between binary response and missingness is positive and therefore it is possible to fit with Model 2.

Table 4.2: Posterior summaries of the parameters for Model 2

	Parameter	Mean	Median	Standard deviation	95% Cr. I. [a]
Binary response:[b]	AGE	-0.402	-0.386	0.402	(-1.18,0.36)
	TIME	0.332	0.329	0.082	(0.18,0.50)
	AZT*TIME	-0.149	-0.150	0.114	(-0.37,0.08)
Missing indicator:	AGE	0.170	0.181	0.201	(-0.25,0.54)
	TIME	0.606	0.605	0.044	(0.52,0.69)
	AZT	0.424	0.425	0.181	(0.07,0.77)
Others:	η_1	0.458	0.455	0.044	(0.38, 0.56)
	η_2	0.704	0.703	0.028	(0.65,0.76)

[a] 'Cr. I.' is an abbreviation of Credible Interval.

[b] To make results of Model 1 and Model 2 comparable, we let binary response be 0 if CD4> 200 such that the association between binary response and missingness is positive and therefore it is possible to fit with Model 2.

CHAPTER 5

Conclusion and Future Work

In this dissertation, we have developed a two-stage random effects logistic regression models for clustered binary responses with random cluster sizes and modified it into a subject-specific random effects model as well as a correlated random effects regression model for longitudinal binary response with informative missingness. Our models are attractive because both the conditional model given random effects and the marginal model integrating over random effects preserve logistic forms. At the same time, our models can efficiently handel binary data with informative cluster size and longitudinal data with informative missingness as well as attain the estimates of heterogeneity among clusters (subjects).

In Chapter 3, we proposed a two-stage random effects logistic model that using bivariate correlated random effects for both the presence/absence of a component and disease/health status. In this two-stage model, we only model disease status when the component is present, which have the accurate probability meaning. At the same time, we utilized the bivariate extension of Bridge distribution so that both the marginal as well as the conditional model of interest can preserve logistic form and the covariates effects for both marginal and conditional model have closed form estimation and log-odds expression. Maximum Likelihood (ML) was proposed and the model can be easily fitted in available statistical software such as SAS. Simulations have been conducted in Chapter 4 and have showed that our proposed model is unbiased and efficient. The simulation of robustness of random effect missspecification also showed that our proposed model is robust even when the random effects are severely differed from Bridge distribution. We have illustrated this model via a periodontal disease data.

In Chapter 4, we proposed one shared random effect logistic model and one correlated random effects model for both binary longitudinal responses and missing-data mechanism. We also discussed the advantages and disadvantages of both models in the sense of model

specification. The shared random effect logistic model could only preserve the logistic form for binary longitudinal responses but not for the missing-indicator, which may decrease the attraction of the model's novelty. Meanwhile, the second model that includes two correlated Bridge random effects that are both generated from same normal generator restricts the association between binary longitudinal response and missing-data indicator has to be positive. Even though this condition can be easily validated by switching the binary responses categories, we still need to make sure we are aware of this why fitting the real life data example using this model. When the association is actually negative and no modification has done before fitting this second model, the model would have trouble converge. Bayesian analysis framework has been used in these two models to potentially incorporate prior information. Bayes factor has been used as the criteria of model selection to seek the best model for a longitudinal study of CD4 counts from HIV-infected patients.

In the future, we would like to further explore the models for binary longitudinal responses with informative missing data and seek for a possibly better model that would not have restrictions we discussed above while can still maintain the logistic form for both conditional model and marginal model. We would also like to incorporate time-varying latent variables instead of subject-specific random variables when we will have data from longitudinal studies in a long span of time and with many visits.

APPENDIX A

Derivation of $E[Y_{2ij}|Y_{1ij}=1,X_{ij}]$

In the following, we derive the expression of $E[Y_{2ij}|Y_{1ij}=1,\mathbf{X_{ij}}]$ and show that $E[Y_{2ij}|Y_{1ij}=1,\mathbf{X_{ij}}]$ has logistic form of (3.15) only when B_{1i} and B_{2i} are independent.

$$\begin{aligned} E(Y_{2ij}|Y_{1ij}=1,\mathbf{X_{ij}}) &= E[E(Y_{2ij}|Y_{1ij}=1,\mathbf{X_{ij}},B_{2i})] \\ &= \int_{B_{2i}} E(Y_{2ij}|Y_{1ij}=1,\mathbf{X_{ij}},B_{2i}) f(B_{2i}|Y_{1ij}=1,\mathbf{X_{ij}})dB_{2i}, \text{ (using (3.13))}, \end{aligned}$$

From Wang 2003 [34], $E(Y_{2ij}|Y_{1ij}=1,\mathbf{X_{ij}})$ has logistic form only if $f(B_{2i}|Y_{1ij}=1,\mathbf{X_{ij}}) = f(B_{2i})$.

We have

$$f(B_{2i}|Y_{1ij}=1,\mathbf{X_{ij}}) = \frac{\int_{B_{1i}} f(B_{1i},B_{2i})P(Y_{1ij}=1|B_{1i},\mathbf{X_{ij}})dB_{1i}}{P(Y_{1ij}=1)} \quad \text{(A.1)}$$

$$= f(B_{2i})\frac{\int_{B_{1i}} f(B_{1i}|B_{2i})P(Y_{1ij}=1|B_{1i},\mathbf{X_{ij}})dB_{1i}}{P(Y_{1ij}=1|\mathbf{X_{ij}})}. \quad \text{(A.2)}$$

Hence $f(B_{2i}|Y_{1ij}=1,\mathbf{X_{ij}}) = f(B_{2i})$ if and only if

$$\frac{\int_{B_{1i}} f(B_{1i}|B_{2i})P(Y_{1ij}=1|B_{1i},\mathbf{X_{ij}})dB_{1i}}{P(Y_{1ij}=1|\mathbf{X_{ij}})} = 1,$$

So that we have

$$\int_{B_{1i}} f(B_{1i}|B_{2i})P(Y_{1ij}=1|B_{1i},\mathbf{X_{ij}})dB_{1i} = P(Y_{1ij}=1|\mathbf{X_{ij}}), \text{ for all } B_{2i}\in(-\infty,+\infty). \quad \text{(A.3)}$$

Equation A.3 holds if and only if $f(B_{1i}|B_{2i}) = f(B_{1i})$. Therefore, $E(Y_{2ij}|Y_{1ij}=1,\mathbf{X_{ij}})$ has logistic form only if $f(B_{2i}|Y_{1ij}=1,\mathbf{X_{ij}}) = f(B_{2i})$.

APPENDIX B

SAS codes for model with informative cluster size

```
/* Joint Model for Y_{1ij} and Y_{2ij}*/

data Dental.joint_glim_est_full;
set Dental.glim_est_full_part Dental.glim_est_full_pard;
run;



proc nlmixed data=Dental.Caldata METHOD=ISAMP noad qpoints=400
tech=NRRIDG seed=6929000;

parms/ data = Dental.joint_glim_est_full;
*bounds 0 < rho_bri12 < 1;

pi = constant('pi');

/* binary teeth existence */

uni1 = probnorm(r1);
phi1 = 1.0/sqrt(1 + 3/pi/pi*std_bri1*std_bri1);
B1 = 1.0/phi1*log(sin(pi*uni1*phi1)/sin( phi1*pi*(1-uni1) )   );

xb_bin1= B1 + t_0 + t_age*age + t_sex*sex + t_bmi*bmi+t_smoker*smoker +
```

```
t_hba1c*hba1c+ t_maxilla*maxilla + t_pov*pov + t_bfl*bfl;


p1  = exp(xb_bin1)/(1 + exp(xb_bin1) );

llik1 =   teeth_cal*log( p1 ) + (1 - teeth_cal)*log(1 - p1 )  ;

/* binary disease */

uni2 = probnorm(r2);

phi2 = 1.0/sqrt(1 + 3/pi/pi*std_bri2*std_bri2);
B2 = 1.0/phi2*log(sin(pi*uni2*phi2)/sin( phi2*pi*(1-uni2) )   );

xb_bin2= B2 + d_0 + d_age*age + d_sex*sex + d_bmi*bmi+d_smoker*smoker +
d_hba1c*hba1c + d_maxilla*maxilla + d_pov*pov + d_bfl*bfl;

p2  = exp(xb_bin2)/(1 + exp(xb_bin2) );

llik2 =   teeth_cal*(d_cal*log(1-p2)+(1 - d_cal)*log(p2));

if d_cal =. then ll = llik1 ;
else ll = llik1 + llik2;

model zzz ~ general(ll);

random r1 r2 ~ normal( [0,0], [1,rho_bri12 , 1] ) subject=id;

ESTIMATE 't_0m' t_0*phi1;
  ESTIMATE 't_agem' t_age*phi1;
  ESTIMATE 't_sexm' t_sex*phi1;
  ESTIMATE 't_bmim' t_bmi*phi1;
```

```
  ESTIMATE 't_smokerm' t_smoker*phi1;
  ESTIMATE 't_hba1cm' t_hba1c*phi1;
  ESTIMATE 't_maxillam' t_maxilla*phi1;
  ESTIMATE 't_povm' t_pov*phi1;
  ESTIMATE 't_bflm' t_bfl*phi1;


 ESTIMATE 'd_0m' d_0*phi2;
  ESTIMATE 'd_agem' d_age*phi2;
  ESTIMATE 'd_sexm' d_sex*phi2;
  ESTIMATE 'd_bmim' d_bmi*phi2;
  ESTIMATE 'd_smokerm' d_smoker*phi2;
  ESTIMATE 'd_hba1cm' d_hba1c*phi2;
  ESTIMATE 'd_maxillam' d_maxilla*phi2;
  ESTIMATE 'd_povm' d_pov*phi2;
  ESTIMATE 'd_bflm' d_bfl*phi2;
estimate 'phi1' phi1;
estimate 'phi2' phi2;
estimate 'rho_bri12' rho_bri12;
run;
```

APPENDIX C

OpenBUGS codes for model 2 with Informative Missingness

```
model
{
   zmu<-0
   vara<-1


for (i in 1:M)
  {
    z[i] ~ dnorm(zmu, vara)
    v[i]<-phi(z[i])
    B1[i]<-1/ka*log(sin(ka*3.1416*v[i])/sin(ka*3.1416*(1-v[i])))
    B2[i]<-1/kb*log(sin(kb*3.1416*v[i])/sin(kb*3.1416*(1-v[i])))


  }


  for (j in 1:N)
{
# ones trick
ones[j]<-1
```

```
p1.bound[j]<-max(0,min(p1[j], 1))
logit(p1[j])<-beta0+ beta1*AGE35[j]+ beta2*time_lin[j]+beta3*AZT_lin[j]+B1[id[j]]

p2.bound[j]<-max(0,min(p2[j],1))
logit(p2[j])<-gamma0+gamma1*AGE35[j]+gamma2*time_lin[j]+gamma3*AZT[j]+B2[id[j]]



l[j]<-pow(p1.bound[j], y[j]*(1-m[j]))*pow(1-p1.bound[j], (1-y[j])*(1-m[j]))*pow(p2.bound

ones[j]~dbern(l[j])

}

ka~dunif(0.001,0.999)
kb~dunif(0.001,0.999)
beta0~dnorm(-1.5,1)
beta1~dnorm(0,1)
beta2~dnorm(0,1)
beta3~dnorm(0,1)
gamma0~dnorm(-1,1)
gamma1~dnorm(0,1)
gamma2~dnorm(0,1)
gamma3~dnorm(0,1)


}
```

REFERENCES

[1] M. Aerts, G. Geys, H.and Molenberghs, and L. M. Ryan. *Topics in Modelling of Clustered Data*. Chapman and Hall, 2002. 3.1

[2] A. Agresti. *Categorical Data Analysis. 2nd Edition*. Wiley, 2002. 1.2, 1.2, 3.1

[3] P. S. Albert, D.A. Follmann, S. A. Wang, and E. B. Suh. A latent autoregressive model for longitudinal binary data subject to informative missingness. *Biometrics*, 58:631–642, August 2002. 2.2

[4] P.S. Albert, D. A. Follmann, S. A. Wang, and E. B. Sub. A latent autoregressive model for longitudinal binary data subject to informative missingness. *Biometrics*, 58:631–642, 2002. 4.1, 4.2

[5] G. C. Armitage. Development of a classification system for periodontal diseases and conditions. *Annals of Periodontology*, 4:1–6, 1999. 3.3

[6] N.E. Breslow and N.E. Day. Lyon: International Agency for Research on Cancer, 1980. 3.1

[7] G. Casella and R. L. Berger. *Statistical Inference*. Thomson Learning, 2002. 1.2

[8] M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of Amercian Statistical Assocation*, 91:883–904, 1996. 4.4.1

[9] M. L. Darby and M. M. Walsh. *Dental Hygiene: Theory and Practice*. 1st edition, W. B. Saunders Company, USA., 1995. 3.3

[10] D. B. Dunson. Bayesian laten variable models for clustered outcomes. *Journal of Royal Statistical Society*, 62:355–366, 2000. 4.1

[11] D. B. Dunson, Z. Chen, and J. Harry. A bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*, 59:342–351, 2003. 1.1, 2.1, 2.1, 3.1

[12] J. K. Fernandes, E. Wiegandyan, C. F. Salinas, S. G. Grossi, J. J. Sanders, M. F. Lopes-Virella, and E.H. Slate. Periodontal disease status in gullah african americans with type 2 diabetics living in south carolina. *Journal of Periodontology*, 80:1062–1068, 2009. 3.1, 3.3

[13] D. M. Finkelstein, P. L. Williams, G. Molenberghs, J. Feinberg, W. Powderly, J. Kahn, R. Dolins, and D. Cotton. Patterns of opportunistic infections in patients with hiv infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 12:38–45, 1996. 4.1

[14] D. Follmann and M. Wu. An appropximate generalized linear model with random effects for informative missing data. *Biometrics*, 51:151–168, 1995. 1.1

[15] J. E. Gallant, R. D. Moore, D. D. Richman, J. Keruly, and R. E. Chaisson. Incidence and natual history of cytomegalovirus disease in patients with advanced human immunodeficiency virus disease treated with zidovudine. *Journal of Infectious Diseases*, 166:1223–1227, 1992. 4.1

[16] A Gelman and D. B. Rubin. Inference from interative simulation using multiple sequences (with discussion). *Statistical Science*, 7:457–511, 1992. 4.4.1

[17] R. V. Gueorguieva. Comments about joint modeling of cluster size and binary and continuous subunit-specific outcomes. *Biometrics*, pages 862–867, 2005. 2.1, 2.1

[18] E. B. Hoffman, P. K. Sen, and C. R. Weinberg. Within-cluster resampling. *Biometrika*, 88(4):1121–1134, 2001. 1.1, 2.1, 2.1, 3.1

[19] J. O. Kahn, S. W. Lagakos, and D. D. Richman. A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. *New England Journal of Medicine*, 327:581–587, 1992. 4.1

[20] R. E. Kass and A. E. Raftery. Bayes factor. *Journal of the American Statistical Association*, 90:773–795, 1995. 4.4.2

[21] N. M. Laird. Missing data in longitudinal studies. *Statistics in Medcine*, 7:305–315, 1988. 4.1

[22] K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986. 1.2, 1.2, 3.1

[23] L. Lin, D. Bandyopadhyay, S. R. Lipsitz, and D. Sinha. Association models for clustered data with binary and continuous responses. *Biometrics*, 66:287–293, 2009. 3.1

[24] R. J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of American Statistical Association*, 88:125–134, 1993. 4.1

[25] R. J. A. Little. A class of pattern-mixtrue models for normal incomplete data. *Biometrika*, 81:471–483, 1994. 4.1

[26] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 1987. 4.1

[27] R. B. Nelsen. *An Introduction to Copulas*. Springer, New York, 1999. 3.2.2

[28] K. S. Panageas, D. Schrag, A. R. Localio, E.S. Venkatraman, and C. B. Begg. Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Statistics in Medicine*, 26:20117–2035. 2.1, 3.1

[29] J. Phair, A. Munoz, R. Detels, R. Kaslow, C. Rinaldo, and A. Saah. The risk of pneumocystis carinii pneumonia among men infection with human immunodeficiency virus type 1. *New England Journal of Medicine*, 332:161–165, 1990. 4.1

[30] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976. 4.1

[31] A. Sklar. Fonctions de rpartition a n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Universit de Paris*, 8:229–231, 1959. 3.2.2

[32] T. R. Ten Have, A. R. Kunselman, E. P. Pulkstenis, and J. R. Landis. Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics*, 54(1):367–383. 2.2

[33] A. Thomas, B. OHara, U. Ligges, and S. Sturtz. Making bugs open. *R news*, 6(1):12–17, 2006. 4.3.1

[34] Z. Wang and T. A. Louis. Matching conditional and marginal shapes in binary mixed-effects models using a bridge distribution. *Biometrika*, pages 765–775, 2003. 3.1, 3.2.1, 3.2.1, 4.1, 4.2.1, 4.2.1, 4.2.2, A

[35] R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika*, 61:439–447, 1974. 1.2

[36] H-Y Williamson, J.M.and Kim, A. Manatunga, and D. G. Addiss. Modeling survival data with informative cluster size. *Statistics in Medicine*, 27:543=555, 2008. 3.1

[37] J. M. Williamson, S. Datta, and G. A. Satten. Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59:36–42, 2003. 1.1, 2.1, 2.1, 3.1

BIOGRAPHICAL SKETCH

Xiaoyun Li

Xiaoyun Li was born in Guangzhou, China. She completed her B.S. degree Mathmatical Statistics from Peking University, Beijing in 2006. In August 2006, she started her graduate study at the Statistics Department at Florida State University and gained her M.S. degree in April 2008. In the Fall of 2008, she started her doctoral dissertation under Dr. Debajyoti Sinha and she defended her dissertation in December 2010.

Xiaoyun Li's current research interests include longitudinal data analysis with informative missing, Bayesian analysis, modeling clustered data with informative cluster size and clinical trials.