Abstract of the thesis entitled

Informative Drop-out Models for Longitudinal Binary Data

submitted by Chau, Ka Ki

for the degree of Master of Philosophy

at The University of Hong Kong in October 2003

Attrition or drop-out is a common phenomenon in longitudinal studies in which repeated observations are made on the same subject over time. Subjects always drop out prematurely, especially when the measurement process is lengthy. The problem of drop-out results in incomplete and unbalanced data which in turn results in loss of efficiency and bias in the analysed results. Regarding this problem, many modelling approaches that deal with missing data have been proposed. This variety of possible approaches differs for different types of drop-out processes and also different kinds of longitudinal data. In this thesis, we aim to develop new modelling strategies for longitudinal binary data with informative drop-out. Three different conditional AR1 models are proposed for the response and a logistic regression model for the drop-out process. In these models, both the probabilities of a positive response and the drop-out of patient in that occasion are assumed to be logit linear in some covariates and outcomes. To account for the problem of over-dispersion and accommodate population heterogeneity,

we incorporate random intercepts to one of the proposed models. We will implement the models via likelihood and Bayesian frameworks. Since the inclusion of random effects complicates the calculation considerably, we also attempt to investigate the use of Gibbs output within the Bayesian framework to carry out the Monte Carlo Approximation of the complicated likelihood function involving random effects by a classical likelihood approach. We then demonstrate these models on a methadone clinic data. Moreover, we also investigate the sensitivity of the assumption of the dropout process on the parameter estimates for the three proposed models through simulation experiments. Results show that the incorporation of the informative drop-out model helps us to understand and interpret the drop-out process across patients better.

Informative Drop-out Models for Longitudinal Binary Data

BY

CHAU, KA KI

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Philosophy
at the University of Hong Kong

Hong Kong, October 2003

DECLARATION

I declare that this thesis represents my own work, except where due acknowledge is made, and that it has not been previously included in a thesis, dissertation or report submitted to this University or to any other institution for a degree, diploma or other qualification.

Signed _____

CHAU, KA KI

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Dr Jennifer Chan for her encouragement, support and guidance throughout the years of my study. Moreover, I would like to thank all the friends and staff in the Department of Statistics and Actuarial Science. Without anyone of them, I could not have such wonderful years. Most of all, my sincere appreciation goes to my family for their constant understanding and encouragement.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background

The collection of longitudinal binary data is common in clinical trials or longitudinal studies when repeated measurements, positive or negative to certain tests, are made on the same subject over time. Since many longitudinal studies are lengthy, subjects undergoing longitudinal studies may drop-out prematurely, resulting in a large class of distinct missingness patterns.

One important issue arising from the problem of drop-out is whether the drop-out process is related to the measurement process. Drop-out processes can be classified into three types: completely random, random and informative drop-out (Rubin, 1976; Little and Rubin, 1987). Completely random drop-out (CRD) and random drop-out (RD) are often referred to be ignorable because it is not necessary to specify a model for the drop-out process in a likelihood-based analysis of the measurement process. Informative drop-out (ID), on the other hand, is said to be non-ignorable as the drop-out mechanism cannot be ignored when estimating the model parameters for the data. Special modeling strategies are therefore required for inference when the drop-out process is informative.

In this thesis, we aim to develop new modelling strategies for longitudinal binary data with informative drop-out. Three different conditional AR1 models are proposed for the outcomes or responses, and a logistic regression model for the drop-out process. In the model, both the probabilities of a positive response and the drop-out indicator of a patient in that occasion are assumed to be logit linear in some covariates and outcomes. To account for the problem of over-dispersion and accommodate population heterogeneity, we incorporate random intercepts to one of the proposed models. Since the inclusion of random effects complicates the calculation considerably, we also make contributions in improving the methodologies in GLMMs with informative drop-out using both likelihood and Bayesian approaches. We then demonstrate these models on a methadone clinic data. Moreover, we also investigate the sensitivity of the assumption of the drop-out process on the parameter estimates for the three proposed models through simulation experiments. Results show that the incorporation of the informative drop-out model helps us to understand and interpret the drop-out process across patients better.

1.2 Structure of the Thesis

This thesis consists of eight chapters. Chapter 1 introduces the background of this research. Chapter 2 briefly introduces the Generalized linear models (GLMs) and their extensions, the Generalized linear mixed models (GLMMs). As the computation of parameter estimates is complicated by the inclusion of random effects into the GLMs, so in this chapter, we will also investigate the existing methodologies, from classical to Bayesian aspects, for estimating the parameters of the GLMMs. These methods include Maximum likelihood method (ML), Simulated maximum likelihood method (SML), Newton Raphson method (NR), Monte Carlo Newton Raphson method (MCNR) and a Bayesian method using Gibbs sampler. Moreover, we will introduce our proposed method called 'Monte Carlo approximation through Gibbs output' for calculating the ML estimates of the random intercept model with an ID modelling. In chapter 3, we introduce different types of drop-out mechanisms and describe their impact on parameter estimation. We also discuss several types of models that allow for an informative drop-out process. In chapter 4, we introduce a methadone clinic data reported by Chan *et al.* (1998). In addition, we give a brief description of the three modeling strategies proposed for this data set: the conditional AR1 model, the two-group mixture model and the random intercepts model for modeling the drug-use of the

patients under a methadone maintenance program. In chapter 5, we extend the three proposed models to account for the informative drop-out process in the data set. In chapter 6, we report parameter estimates of the three models with and without an ID modeling and its interpretation in terms of the effectiveness of the methadone maintenance treatment. Moreover we also describe the calculation of AICs as measures of goodness-of-fit for the three types of models using ML or Bayesian methods. In chapter 7, we investigate the sensitivity of the assumption of drop-out process on the parameter estimates for the three models with or without an ID modelling through simulation studies. Finally, we discuss the main results of this research and propose future research direction in chapter 8.

CHAPTER 2

MODELS FOR PRIMARY RESPONSES

2.1 Introduction

One important objective of data analysis is to derive statistical models that can adequately describe the phenomenon of the data. When the data is incomplete, we will have invalid likelihood inferences especially when the missing process is non-ignorable. In fact, the validity of such inference depends on using the right data model $f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta})$. Thus the choice of data model is important.

Several decades ago, the simple linear model has often been used to study the relationship between response and explanatory variables. With $\boldsymbol{Y}$ being a vector of responses and $\boldsymbol{X}$ being a $n \times p$ matrix of the explanatory variables, the simple linear model is defined as below:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \qquad \text{where } \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters and $\boldsymbol{e}$ is a $n \times 1$ vector of error terms which follows a multivariate normal distribution with mean equals to $\boldsymbol{0}$ and variance equals to $\sigma^2\boldsymbol{I}$ where $\boldsymbol{I}$ is an identity matrix. Here the error term is assumed to be independent and identically distributed with a constant variance. This simple

linear model forms the base of analysis for the continuous data.

However, this model does not work well with time series data that are common in clinical trials and stock markets when observations are serially correlated. It also cannot deal with responses that are neither normally distributed nor linearly related to the explanatory variables. It cannot deal with categorical responses as well. So, Nelder and Wedderburn (1972) introduced the Generalized linear models (GLMs) which allow an exponential family of distributions on the data in order to describe the non-normal responses. Thereafter, the GLMs are further extended to the Generalized linear mixed models (GLMMs) with the inclusion of random effects in the GLMs to account for the serial correlation, to overcome the overdispersion problem and to accommodate population heterogeneity. As a result, these models become more applicable in many practical situations. On the other hand, since the inclusion of random effects complicates the calculation of likelihood function considerably, especially for models involving high dimensional random effects, this invokes many diversified methodologies for parameter estimation in the GLMMs.

This chapter will be presented as follows. Section 2 will first introduce the Generalized Linear Models (GLMs). Section 3 will further introduce its extension, the Generalized Linear Mixed Models (GLMMs). Then, section 4 will discuss

some model implementation methodologies of the GLMMs using both Bayesian and Classical Approaches. Finally section 5 will describe our proposed methodology of using the Monte Carlo approximation through Gibbs output.

2.2 Generalized linear models (GLMs)

Firstly introduced by Nelder and Wedderburn (1972), the Generalized linear models (GLMs) generalize the classical linear models to the exponential family of sampling distributions and have an immense impact on both theoretical and practical aspects in statistics. The model for the responses y_i defined as follows:

$$\begin{aligned} f(y_i) &= \exp\left(\frac{A_i[\theta_i y_i - b(\theta_i)]}{\phi} + c(y_i, \frac{\phi}{A_i})\right) \\ E(y_i) &= \mu_i \\ g(\mu_i) &= \eta_i = \boldsymbol{x_i}^T\boldsymbol{\beta} \end{aligned}$$

where $b(.)$ is a function of canonical parameter θ_i, $c(.)$ is a function of y_i and ϕ/A_i, A_i is a known prior weight, ϕ is the dispersion parameter, $\boldsymbol{x}_i$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of regression parameters. Note that $f(.)$ is the distribution we assign to the data, $\boldsymbol{x}_i{}^T\boldsymbol{\beta}$ is a linear function of predictors defined in the model and finally, $g(.)$ is the link function that link together the mean μ_i for the distribution of y_i and the linear function of explanatory variables or predictors $\boldsymbol{x}_i{}^T\boldsymbol{\beta}$.

This exponential family of distributions has the following properties:

1. Mean of Y is $\mu = E(Y) = b'(\theta)$.

2. Variance of Y is $Var(Y) = \phi\, b''(\theta)/A = \phi v(\mu)/A$ in which $v(\mu) = b''[(b')^{-1}(\mu)]$ is called the variance function.

There are various kinds of exponential family of distributions including Binomial distribution, Poisson distribution, Gamma distribution and Normal distribution etc. Moreover there are also different kinds of link functions for non-normal responses. For binary data, the common link functions are logit-link $g(\mu) = \log(\mu/(1-\mu))$ and probit-link $g(\mu) = \Phi^{-1}(\mu)$ for $0 < \mu < 1$. The log-link $g(\mu) = \ln(\mu)$ is usually used for Poisson count data. For continuous data, we can use identity link $g(\mu) = \mu$.

2.3 Generalized linear mixed models(GLMMs)

Although the GLMs can be used to describe the non-normal behavior of data, they cannot be used to account for the serial correlation and clustering effect in the data from longitudinal studies. The GLMs, as a result, have been improved and modified to a more general class of models, known as the Generalized linear mixed models (GLMMs) by the inclusion of random effect terms. With the

GLMMs, we can overcome the problem of overdispersion in the data and at the same time, accommodate the population heterogeneity. The main difference between the structure of GLMMs and that of GLMs is the incorporation of the random effects, $\boldsymbol{u}$, into the function of linear predictors. Thus, the model on $y_i,\ i = 1, \ldots, n$ which follows an exponential family of distributions given the random effects $\boldsymbol{u}$ becomes:

$$\begin{aligned} f(y_i|\boldsymbol{u}) &= \exp\left(\frac{A_i[\theta_i y_i - b(\theta_i)]}{\phi} + c(y_i, \frac{\phi}{A_i})\right) \\ E(y_i|\boldsymbol{u}) &= \mu_i \\ g(\mu_i) &= \boldsymbol{x_i}^T\boldsymbol{\beta} + \boldsymbol{z_i}^T\boldsymbol{u} \end{aligned}$$

where $\boldsymbol{x}_i$ is a vector of covariates, $\boldsymbol{\beta}$ is a vector of fixed effect parameters, $\boldsymbol{z}_i$ is the design matrix for the random effects and $\boldsymbol{u}$ is a vector of the random effects. A distribution $h(.)$ is assigned to the random effects such that $\boldsymbol{u} \sim h(\boldsymbol{u})$. For simplicity, researchers usually assign a multivariate normal distribution with mean $\mathbf{0}$ and variance $\sigma^2\boldsymbol{I}$ to the random effects. For robustness consideration, a general class of random effects distributions (heavy-tailed, asymmetric and nonparametric) should be adopted to widen the scope of applicability on the GLMMs. See for examples Stiratelli *et al.* (1984) who analysed a longitudinal data using GLMMs and assigned a multivariate normal distribution for the random effects to account for the heterogeneity between subjects, and Choy *et al.* (2003) who assigned a

Student-t distribution for the random effects in studying the famous Salamander Mating data set.

2.4 Model implementation methodologies

2.4.1 Introduction

The inclusion of random effects into the GLMs opens up the class of Generalized linear mixed models (GLMMs) which help to overcome the problem of over-dispersion and accommodate population heterogeneity. These models become more applicable in many practical situations. However, since the inclusion of random effects complicates the model implementation and estimation considerably, diversified methodologies in GLMMs have therefore been proposed to improve the estimation of parameters in the model. These proposed approaches are mainly classified into two types: Classical approach and Bayesian approach.

In this chapter, we will first study both classical and Bayesian approaches for model estimation in GLMMs. In the classical approach, the intractability of the likelihood function has thus led various authors to propose a host of alternative estimation methods rather than carrying out maximum likelihood estimation directly. The methodologies discussed in this thesis include the maximum like-

lihood method (ML) and its extensions such as simulated maximum likelihood (SML) method(see Geyer and Thompson, 1992 and Gelfand and Carlin, 1993 for reference), Newton-Raphson method, Monte Carlo Newton-Raphson (MCNR) method (Pentitinen 1984) , Laplace importance sampling method (Kuk 1999) and Monte Carlo relative likelihood method (Geyer and Thompson, 1992 and Geyer, 1994) for handling complicate likelihood function which involves high dimensional integral of random effects. However, methodologies like the approximate maximum likelihood and the residual maximum likelihood method (REML) (Stiratelli (1984); Schall, 1991; Drum & McCullagh 1993; McCulloch and Searle, 2001), penalized quasi-likelihood method (Green, 1990; Wolfinger, 1993; Breslow & Clayton 1993), the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) and Monte Carlo EM algorithm (Wei and Tanner, 1990), the estimating function approach (Waclawiw & Liang, 1993) and the iterative bias correction approach (Kuk, 1995) etc.. will not be pursed in this thesis. For Bayesian approach on the GLMMs, we will investigate the use of Markov chain Monte Carlo method and Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990; Zeger & Karim, 1991).

Thereafter we will propose a new model using the classical likelihood approach but applying methodologies of both Monte Carlo approximation and Gibbs sampler.

2.4.2 Bayesian and Classical Approaches

Suppose we have a n-dimensional vector of responses $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$. We are interested in the parameter estimate $\boldsymbol{\theta}$. We denote the density of y_i as $f(y_i|\boldsymbol{\theta})$. Then the joint density function is expressed as:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}).$$

In classical inference, we assign a model $f(\boldsymbol{y}|\boldsymbol{\theta})$ to an observed data and the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{y})$ represents the information on the data such that $L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. The parameter $\boldsymbol{\theta}$ of the model is treated as fixed but unknown. We differentiate the log of the likelihood function, $\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \ln L(\boldsymbol{\theta}|\boldsymbol{y})$ to obtain the parameter estimate $\widehat{\boldsymbol{\theta}}$ which maximizes the likelihood function of the observed data.

In Bayesian approach, the parameter $\boldsymbol{\theta}$ of the data model $f(\boldsymbol{y}|\boldsymbol{\theta})$ is treated as random rather than fixed as in the classical approach. So, in order to obtain the Bayesian estimates, we first need to specify a prior distribution for each parameter and then evaluate its posterior distribution from which the Bayesian estimate is given by its posterior mean. Suppose $f(\boldsymbol{\theta})$ is the prior distribution for the model parameters, then the joint posterior density of $\boldsymbol{\theta}$ is given by

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}).$$

The posterior distribution of $\boldsymbol{\theta}$ is proportional to the product of likelihood function $f(\boldsymbol{y}|\boldsymbol{\theta})$ and its prior density $f(\boldsymbol{\theta})$.

We have to collect prior information on the parameter $\boldsymbol{\theta}$ and then assign a suitable prior density $f(\boldsymbol{\theta})$ to the parameter $\boldsymbol{\theta}$ in order to construct the posterior density. More information on the Bayesian statistics can be found in Lee (1999) and Bernardo and Smith (1994).

However, there exists an argument on the specification of prior density $f(\boldsymbol{\theta})$. In some cases, conjugate prior is chosen just for convenience. In other occasions, the choice of prior can be rather subjective. In the absence of sufficient prior information, a non-informative prior with large variance is usually chosen for the model parameter $\boldsymbol{\theta}$.

In the following sections, we will investigate the commonly used methodologies in both Classical and Bayesian approaches.

2.4.3 Classical approach: Maximum Likelihood (ML) method

Maximum likelihood estimation is a prevalent classical approach for parameter estimation. In the following, we will illustrate the algorithm of this approach based on the GLMM. The ML estimation begins with the likelihood function of a model.

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ be a n-dimensional vector of responses with a joint density function $f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\theta})$ on $\Re^p$, where $\boldsymbol{z}$ is a m-dimensional vector of random effects with the density $h(\boldsymbol{u}|\boldsymbol{\theta})$, and $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ is a vector of parameters including the parameters for fixed effects $\boldsymbol{\beta}$ and the parameters for random effects $\boldsymbol{\gamma}$. The 'marginal' likelihood function $L(\boldsymbol{\theta})$ and the log-likelihood function $\ell(\boldsymbol{\theta})$ are then obtained by integrating out the unobserved random effects:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \int f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\theta})\, d\boldsymbol{u} = \int f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\beta})\, h(\boldsymbol{u}|\boldsymbol{\tau})\, d\boldsymbol{u}, \\ \ell(\boldsymbol{\theta}) &= \ln L(\boldsymbol{\theta}) = \ln \int f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\beta})\, h(\boldsymbol{u}|\boldsymbol{\tau})\, d\boldsymbol{u}. \end{aligned}$$

Then the ML estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained by maximizing $L(\boldsymbol{\theta})$ or equivalently $\ell(\boldsymbol{\theta})$. That is, we obtain $\widehat{\boldsymbol{\theta}}$ by solving

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \tag{2.1}$$

and the solution $\widehat{\boldsymbol{\theta}}$ is the ML estimate which maximizes the log likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{y})$.

Many researchers further improved this method to form various kinds of modified ML estimation method. Schall (1991) studied the use of ML approach in GLMs with random effects. But, the numerical integration method that he used is only appropriate for simple cases in which the likelihood function involves only an integral of low dimensional random effects or an integral that can be factorized into a product of low dimensional integrals.

2.4.3.1 Simulated Maximum Likelihood method (SML)

Geyer and Thompson (1992) and Gelfand and Carlin (1993) suggested the use of SML on the approximation of likelihood function. McCulloch (1997) further studied the use of this method on the GLMMs. In SML method, the likelihood function is estimated directly by simulation without considering the log-likelihood function.

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y}) &= \int f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta})\ d\boldsymbol{u} \\ &= \int f(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\theta})\ f(\boldsymbol{u}; \boldsymbol{\theta})\ d\boldsymbol{u} \\ &= \int \frac{f(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\theta})\ f(\boldsymbol{u}; \boldsymbol{\theta})}{h(\boldsymbol{u})}\ h(\boldsymbol{u})\ d\boldsymbol{u} \\ &\approx \frac{1}{M} \sum_{j=1}^{M} \frac{f(\boldsymbol{y}|\boldsymbol{u}_j; \boldsymbol{\theta})\ f(\boldsymbol{u}_j; \boldsymbol{\theta})}{h(\boldsymbol{u}_j)} \qquad \text{where } \boldsymbol{u}_j \sim h(\boldsymbol{u}) \end{aligned} \tag{2.2}$$

where M is the total number of simulated values for $\boldsymbol{u}$, $h(\boldsymbol{u})$ is the importance sampling function independent of the model parameter $\boldsymbol{\theta}$ and $\boldsymbol{u}_j$ is the j-th vector of random effects simulated from this distribution by any sampling technique. Theoretically, the estimates are independent of the choice of importance sampling function, $h(\boldsymbol{u})$. They are calculated numerically based on the likelihood function approximated by values which are simulated from the importance sampling function.

One choice of the importance sampling function $h(\boldsymbol{u})$ is the conditional distribution of $f(\boldsymbol{u}|\boldsymbol{y}; \boldsymbol{\theta}^{(k)})$ given the current parameter estimates $\boldsymbol{\theta}^{(k)}$. However

this requires simulating a new set of $\boldsymbol{u}$ whenever $\boldsymbol{\theta}$ is updated. To overcome this problem, one may set $h(\boldsymbol{u})$ to be $f(\boldsymbol{u}|\boldsymbol{y};\boldsymbol{\theta}_0)$ based on a fixed reference point $\boldsymbol{\theta}_0$. However, if such function is far away from the density of the random effects, $f(\boldsymbol{u};\boldsymbol{\theta})$, specially when the reference point $\boldsymbol{\theta}_0$ is not close to the true ML estimates $\boldsymbol{\theta}_{ML}$, the efficiency of estimates will be affected. One remedy will be to update the reference point $\boldsymbol{\theta}_0$ to the current ML estimates $\boldsymbol{\theta}_{ML}^{(k)}$ a few times.

2.4.3.2 Laplace importance sampling method

Kuk (1999) proposed the Laplace importance sampling method which is actually an application of SML combining the Laplace expansion with importance sampling method. By carrying out a second order expansion of the log-likelihood function $\ell(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}) = \ln f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\theta})$ around $\boldsymbol{u} = \hat{\boldsymbol{u}}(\boldsymbol{\theta}_0)$, where $\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0)$ is the maximizer of the joint density $f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\theta}_0) = f(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}_0)f(\boldsymbol{u};\boldsymbol{\theta}_0)$ with $\boldsymbol{y}$ fixed at the observed value and $\boldsymbol{\theta}$ fixed at a pre-specified value $\boldsymbol{\theta}_0$, the Laplace expansion is given by

$$\ell(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}) \simeq \ell(\boldsymbol{\theta}_0;\boldsymbol{y},\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0)) - \frac{1}{2}(\boldsymbol{u}-\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0))^T[-\ell''_{\boldsymbol{u}}(\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0))](\boldsymbol{u}-\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0)), \quad (2.3)$$

where $\ell''_{\boldsymbol{u}}(\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0)) = \ell''(\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0);\boldsymbol{y},\boldsymbol{\theta}_0)$ is the Hessian matrix of the second order derivatives of $\ell(\boldsymbol{u};\boldsymbol{y},\boldsymbol{\theta}_0)$ with respect to the components of $\boldsymbol{u}$ evaluated at $\boldsymbol{u} = \hat{\boldsymbol{u}}(\boldsymbol{\theta}_0)$, the maximizer of the joint density $f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\theta}_0) = f(\boldsymbol{y}|\boldsymbol{u};\boldsymbol{\theta}_0)f(\boldsymbol{u};\boldsymbol{\theta}_0)$.

Exponentiating the expansion (2.3) results in a normal density in $\boldsymbol{u}$ up to a constant multiple that does not involve $\boldsymbol{u}$. Specially,

$$f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}_0) \simeq c(\boldsymbol{y}, \boldsymbol{\theta}_0)\ \phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)), \tag{2.4}$$

where $\phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$ is the multivariate normal density function with mean vector $\boldsymbol{\mu}(\boldsymbol{\theta}_0) = \hat{\boldsymbol{u}}(\boldsymbol{\theta}_0)$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = [-\ell''_{\boldsymbol{u}}(\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0))]^{-1}$ and $c(\boldsymbol{y}, \boldsymbol{\theta}_0)$ is a constant that does not depend on $\boldsymbol{u}$. Hence, if $\phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$ in expansion (2.4) is used as an importance function for the sampling of $\boldsymbol{u}$ and the sampled $\boldsymbol{u}$ is in turn used for carrying out the Monte Carlo approximation of the likelihood function, we have

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y}) &= \int \frac{f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta})}{\phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))}\ \phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))\ d\boldsymbol{u} \\ &\approx \frac{1}{M} \sum_{j=1}^{M} \frac{f(\boldsymbol{y}, \boldsymbol{u}_j; \boldsymbol{\theta})}{\phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))} \end{aligned} \tag{2.5}$$

where

$$\boldsymbol{u}_j \sim \phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$$

and

$$\frac{f(\boldsymbol{y}, \boldsymbol{u}_j; \boldsymbol{\theta})}{\phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))} \propto c(\boldsymbol{y}, \boldsymbol{\theta}_0)$$

does not depend on $\boldsymbol{\theta}$ from (2.4). Hence the approximation as given by (2.5) is very efficient with minimum variance. Moreover since the function $\phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$ is only used as an importance sampling function, the Laplace expanision will only

affect the efficiency but not the unbiasedness of the approximation (Kuk, 1999). In fact, this is essentially a SML method with the importance sampling function $h(\boldsymbol{u})$ being replaced by $\phi(\boldsymbol{u}; \boldsymbol{\mu}(\boldsymbol{\theta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$. To solve for $\boldsymbol{\theta}$ in (2.5), NR method or its extension is often used. However, the Laplace importance sampling method may only be good when $\boldsymbol{\theta}_0$ is close to its ML estimates. So, Kuk (1999) recommended updating $\boldsymbol{\theta}_0$ to the current estimate of $\widehat{\boldsymbol{\theta}}^{(k)}$, repeating the importance sampling of $\boldsymbol{u}$ and the calculation of (2.5) a few times. However, in this way, the estimation procedure requires iteration within iterations as well as maximization to obtain $\hat{\boldsymbol{u}}(\boldsymbol{\theta}^{(k)})$ and $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}^{(k)})$ in each iteration k which is laborious.

2.4.3.3 Newton-Raphson (NR) method

As mentioned before, we can obtain the ML estimate $\widehat{\boldsymbol{\theta}}$ by solving (2.1). However, in many situations, the solution of (2.1) cannot be solved explicitly. In these cases, NR method provides a way to solve for $\widehat{\boldsymbol{\theta}}$ iteratively.

NR method is a popular iterative method for obtaining the ML estimates. With $\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \ln f(\boldsymbol{y}; \boldsymbol{\theta})$ denoting the log-likelihood function on the data $\boldsymbol{y}$ and the parameter vector $\boldsymbol{\theta}$ and $\ell'(\boldsymbol{\theta}; \boldsymbol{y})$ and $\ell''(\boldsymbol{\theta}; \boldsymbol{y})$, its first and second order derivatives, current parameter estimates in the k-th iteration $\boldsymbol{\theta}^{(k)}$ of the NR

procedures can be updated to the $(k+1)$-th iteration $\boldsymbol{\theta}^{(k+1)}$ by

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - [\ell''(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}]^{-1}\ell'(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}) \tag{2.6}$$

and the procedure continues until $\parallel \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)} \parallel$ is sufficiently small or the maximum number of iterations has been attained.

2.4.3.4 Monte Carlo Newton Raphson method (MCNR)

The NR method is a popular iterative method used to find the ML estimates. The method requires the calculation of the log-likelihood function $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ as well as its first and second order derivatives $\ell'(\boldsymbol{\theta}; \boldsymbol{y})$ and $\ell''(\boldsymbol{\theta}; \boldsymbol{y})$. However, with random effects, it is often difficult to evaluate the marginal likelihood function and hence its derivatives when the likelihood function involves high dimensional integral of random effects. One approach is the use of Monte Carlo approximation to the likelihood function using the random effects $\boldsymbol{u}$ which are simulated from a conditional function of $\boldsymbol{u}$ given the observed $\boldsymbol{y}$ and the current estimate $\boldsymbol{\theta}^{(k)}$ such as the Laplace importance sampling method. Pentitinen (1984), therefore, extended the NR method to the MCNR method. The algorithm of using the MCNR is as follows:

Algorithm:

1. Choose a starting value for $\boldsymbol{\theta}^{(0)}$.

2. Simulate $\boldsymbol{u}_j$ where $j = 1, ..., M$ with M being the total number of simulations from a conditional distribution of $f(\boldsymbol{u}|\boldsymbol{y}; \boldsymbol{\theta}^{(k)})$ based on $\boldsymbol{\theta}^{(k)}$ which is the current parameter estimates.

3. Use the simulated $\boldsymbol{u}_j$ to approximate the first and second order derivatives of a likelihood function by Monte Carlo approximation. We have

$$\ell'_M(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}) = \frac{1}{M}\sum_{j=1}^{M} \ell'(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}, \boldsymbol{u}_j) \tag{2.7}$$

$$\begin{aligned} \ell''_M(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}) &= \frac{1}{M}\sum_{j=1}^{M} \ell''(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}, \boldsymbol{u}_j) + \frac{1}{M}\sum_{j=1}^{M} \ell'(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}, \boldsymbol{u}_j)\ell'^{T}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}, \boldsymbol{u}_i) \\ &\quad - \left(\frac{1}{M}\sum_{j=1}^{M} \ell'(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}, \boldsymbol{u}_j)\right)\left(\frac{1}{M}\sum_{j=1}^{M} \ell'^{T}(\boldsymbol{\theta}^{(k)}; \boldsymbol{y}, \boldsymbol{u}_j)\right) \qquad (2.8) \\ &= -(\boldsymbol{I}_1 - \boldsymbol{I}_2) \qquad (2.9) \end{aligned}$$

where $\boldsymbol{I}_1$ in (2.9) in corresponds to the first term of (2.8). Replacing the terms $\ell'(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$ and $\ell''(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$ in the Newton Raphson iteration given in equation (2.6) by $\ell'_M(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$ and $\ell''_M(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$ respectively in equations (2.7) and (2.8), we obtain an updated estimate of $\boldsymbol{\theta}^{(k+1)}$ in each MCNR iteration.

4. Repeat steps 2-3 until convergence is achieved.

Kuk and Cheng (1997) applied MCNR method in calculating the estimates in the GLMMs and proposed some refinement and stopping criteria. Kuk and Cheng (1999) gave more comment of the method. They showed that the convergent rate for this MCNR was faster than that of Monte Carlo EM. So, it is computationally

more efficient.

2.4.3.5 Monte Carlo relative likelihood method

The MCNR deals with *pointwise* approximation of the first and second order derivatives $\ell'(\boldsymbol{\theta};\boldsymbol{y})$ and $\ell''(\boldsymbol{\theta};\boldsymbol{y})$ at the current estimate $\boldsymbol{\theta}^{(k)}$. On the other hand, Geyer and Thompson (1992) and Geyer (1994) approximated the whole likelihood function based on a relative likelihood

$$L_q(\boldsymbol{\theta};\boldsymbol{\theta}_0) = \frac{L(\boldsymbol{\theta};\boldsymbol{y})}{L(\boldsymbol{\theta}_0;\boldsymbol{y})}$$

at a reference point $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This is called a *functional* approach because the entire relative function $L_q(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ will be approximated by Monte Carlo method. For the GLMMs, the marginal relative likelihood using Monte Carlo approximation with j indexes the number of simulations used in the approximation is

$$\begin{aligned} L_q(\boldsymbol{\theta};\boldsymbol{\theta}_0) &= E\left[\frac{f(\boldsymbol{y},\boldsymbol{\mu};\boldsymbol{\theta})}{f(\boldsymbol{y},\boldsymbol{\mu};\boldsymbol{\theta}_0)} \mid \boldsymbol{y};\boldsymbol{\theta}_0\right] \\ &= \int \frac{f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\theta})}{f(\boldsymbol{y},\boldsymbol{u};\boldsymbol{\theta}_0)} f(\boldsymbol{u}|\boldsymbol{y};\boldsymbol{\theta}_0)\, d\boldsymbol{u} \\ \widehat{L}_q(\boldsymbol{\theta};\boldsymbol{\theta}_0) &= \frac{1}{M}\sum_{j=1}^{M} \frac{f(\boldsymbol{y},\boldsymbol{u}_j;\boldsymbol{\theta})}{f(\boldsymbol{y},\boldsymbol{u}_j;\boldsymbol{\theta}_0)} \qquad \text{where } \boldsymbol{u}_j \sim f(\boldsymbol{u}|\boldsymbol{y},\boldsymbol{\theta}_0). \end{aligned}$$

which is an unbias Monte Carlo approximation of the entire relative likelihood function with the Monte Carlo variance

$$\mathrm{Var}_{MC}[L_q(\boldsymbol{\theta};\boldsymbol{\theta}_0)] = \frac{1}{M}\mathrm{Var}\left[\frac{f(\boldsymbol{y},\boldsymbol{\mu};\boldsymbol{\theta})}{f(\boldsymbol{y},\boldsymbol{\mu};\boldsymbol{\theta}_0)} \mid \boldsymbol{y};\boldsymbol{\theta}_0\right].$$

Note that $\hat{\ell}_q(\boldsymbol{\theta};\boldsymbol{\theta}_0) = \ln L_q(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ approximates the relative log-likelihood function $\ell_q(\boldsymbol{\theta};\boldsymbol{\theta}_0) = \ell(\boldsymbol{\theta};\boldsymbol{y}) - \ell(\boldsymbol{\theta}_0;\boldsymbol{y})$. For $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$, the conditional variance $\mathrm{Var}_{MC}[L_q(\boldsymbol{\theta};\boldsymbol{\theta}_0)]$ is small and so $\widehat{L}_q(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ is a good approximation of $L_q(\boldsymbol{\theta};\boldsymbol{\theta}_0)$. If the initial $\boldsymbol{\theta}_0$ is far away from $\widehat{\boldsymbol{\theta}}_{ML}$, the maximiser of $\widehat{L}_q(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ may be quite different from $\hat{\boldsymbol{\theta}}_{ML}$. To solve the local nature of the approximation, Geyer and Thompson (1992) suggested updating the maximiser of $\widehat{L}_q(\boldsymbol{\theta};\boldsymbol{\theta}_0)$ as the next $\boldsymbol{\theta}_0$ to obtain an updated estimate $\widehat{\boldsymbol{\theta}}^{(k+1)}$ which is the maximiser of the Monte Carlo relative log-likelihood function

$$\hat{\ell}_q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) = \ln\left[\frac{1}{M}\sum_{j=1}^{M}\frac{f(\boldsymbol{y},\boldsymbol{\mu}_j;\boldsymbol{\theta})}{f(\boldsymbol{y},\boldsymbol{\mu}_j;\widehat{\boldsymbol{\theta}}^{(k)})}\right]; \qquad \text{where } \boldsymbol{u}_j \sim f(\boldsymbol{u}|\boldsymbol{y},\widehat{\boldsymbol{\theta}}^{(k)}).$$

The maximization of $\hat{\ell}_q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)})$ requires numerical method such as Newton Raphson iterative procedure for each k. The derivatives of $\hat{\ell}_q(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$ are

$$\hat{\ell}'_{qM}(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) = \frac{\sum_{j=1}^{M}\ell'(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}_j)L_{qj}}{\sum_{j=1}^{M}L_{qj}} \tag{2.10}$$

$$\begin{aligned}\hat{\ell}''_{qM}(\boldsymbol{\theta};\widehat{\boldsymbol{\theta}}^{(k)}) = {} & \frac{\sum_{j=1}^{M}\ell''(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}_j)L_{qj}}{\sum_{j=1}^{M}L_{qj}} + \frac{\sum_{j=1}^{M}\ell'(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}_j)\ell'^{T}(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{u}_j)L_{qj}}{\sum_{j=1}^{M}L_{qj}} \\ & - \left(\frac{\sum_{j=1}^{M}\ell'(\boldsymbol{\theta}^{(k)};\boldsymbol{y},\boldsymbol{u}_j)L_{qj}}{\sum_{j=1}^{M}L_{qj}}\right)\left(\frac{\sum_{j=1}^{M}\ell'^{T}(\boldsymbol{\theta}^{(k)};\boldsymbol{y},\boldsymbol{u}_j)L_{qj}}{\sum_{j=1}^{M}L_{qj}}\right)\end{aligned} \tag{2.11}$$

where $L_{qj} = \dfrac{f(\boldsymbol{y}, \boldsymbol{u}_j; \boldsymbol{\theta})}{f(\boldsymbol{y}, \boldsymbol{u}_j; \widehat{\boldsymbol{\theta}}^{(k)})}$ and (2.10) is a weighted average of $\ell'(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}_j)$ with weights

$$w(\boldsymbol{u}_j, \boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}^{(k)}) = L_{qj}.$$

This is different from the MCNR method when ${\ell'}_M(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$ is given by a simple average of $\ell'(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}_j)$ as in (2.7). See Kuk and Cheng (1997) & (1999) for a detail discussion.

2.4.4 Bayesian approach: Gibbs sampler

In many cases, especially when random effects are present in models, it is difficult to obtain parameter estimates by using classical approach because high-dimensional integration is involved in the computation. To avoid such tedious computation, Bayesian approach is an alternative to the classical one.

In Bayesian approach, we need to specify a prior distribution to each parameter and then obtain their posterior means or medians from their conditional distribution given the observed data. Numerical or analytic approaches to evaluate the posterior means of the complicated conditional distribution are not useful as high-dimensional integration is usually required. With regard to this point, Markov chain Monte Carlo (MCMC) methods using Gibbs sampler (Geman and Geman, 1984 and Gelfand and Smith, 1990), Metropolis Hasting algorithm (Hast-

ings, 1970 and Metropolis *et al.*, 1953) and Adaptive Rejection sampling (Ripley 1987) have been proposed. In this section, we will focus on the use of Gibbs sampling approach on the GLMMs.

The Gibbs sampler was proposed by Geman and Geman (1984) and has already been successfully applied to many diverse problems such as linear variance components models (Gelfand *et al.*, 1990), random effects GLMs (Zegar and Karim, 1991), and frailty models (Clayton, 1996).

Suppose $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are three random variables with $[\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}]$ as the joint distribution which is complicate, and $[\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}]$, $[\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}]$, and $[\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}]$ are the comparatively simpler conditional distributions of $\boldsymbol{X}$, $\mathbf{Y}$ and $\boldsymbol{Z}$ respectively. Then, given an arbitrary starting values $\boldsymbol{X}^{(0)}$, $\boldsymbol{Y}^{(0)}$ and $\boldsymbol{Z}^{(0)}$, the algorithm of Gibbs sampler proceeds as follows:

1. Draw $\boldsymbol{X}^{(1)}$ from the conditional distribution of $[\boldsymbol{X}|\boldsymbol{Y}^{(0)}, \boldsymbol{Z}^{(0)}]$.

2. Draw $\boldsymbol{Y}^{(1)}$ from the conditional distribution of $[\boldsymbol{Y}|\boldsymbol{X}^{(1)}, \boldsymbol{Z}^{(0)}]$ based on $\boldsymbol{Z}^{(0)}$ and the newly simulated $\boldsymbol{X}^{(1)}$.

3. Complete the first iteration by drawing $\boldsymbol{Z}^{(1)}$ from $[\boldsymbol{Z}|\boldsymbol{X}^{(1)}, \boldsymbol{Y}^{(1)}]$ based on the newly simulated $\boldsymbol{X}^{(1)}$ and $\boldsymbol{Y}^{(1)}$.

4. Repeat this algorithm until N iterations have completed and the simulated

values converged to the joint density function.

Suppose we have completed N iterations. We should then discard the first K iterations in the burn-in period in which the simulated samples may not be stable and use the remaining $M = N-K$ iterations to form posterior samples. The posterior sample means are thus our parameter estimates. After computing the posterior sample means, we need to check the convergence and the auto-correlation by plotting the series of simulated values and examining their auto-correlation function respectively for each parameter.

After reviewing the general idea of Gibbs sampler, we can move on to describe the procedure of applying Gibbs sampler on the GLMMs. To do so, we have to specify the joint distribution and the corresponding full conditional distribution for each variable first. Let $\boldsymbol{y}$ be the observed data having joint density function $f(\boldsymbol{y}|\ \boldsymbol{\beta}, \boldsymbol{u})$, $\boldsymbol{\beta}$ be a vector of fixed effect parameters following a b-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}_\beta$ and a variance-covariance matrix $\boldsymbol{\Sigma}_\beta$, and $\boldsymbol{u}$ be a vector of random effect parameters following a p-dimensional multivariate normal distribution with mean $\mathbf{0}$ and a variance-covariance matrix $\boldsymbol{D}$. We then have the following hierarchical model:

$$\begin{aligned} \boldsymbol{y} &\sim f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}) \\ \boldsymbol{u} &\sim MVN_p(\mathbf{0}, \boldsymbol{D}) \end{aligned}$$

$$\begin{aligned}
\boldsymbol{\beta} &\sim MVN_b(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \\
\boldsymbol{D} &\sim \text{Wishart}(\boldsymbol{R}_p, p) \\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}} &\sim \text{Wishart}(\boldsymbol{R}_b, b)
\end{aligned}$$

where $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ is a fixed vector and $\boldsymbol{R}_p$ and $\boldsymbol{R}_b$ are p and b dimensional fixed matrices for $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ respectively. The joint density is $[\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}]$. The Bayesian estimates are drawn from its conditional distribution by Gibbs sampler. For example, we sample

$$\begin{aligned}
\boldsymbol{\beta} \text{ from } [\boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \boldsymbol{D}, \boldsymbol{y}] \\
\boldsymbol{u} \text{ from } [\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \boldsymbol{D}, \boldsymbol{y}] \\
\boldsymbol{D} \text{ from } [\boldsymbol{D}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \boldsymbol{D}] \\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}} \text{ from } [\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{D}, \boldsymbol{y}]
\end{aligned}$$

Sometimes, the conditional distributions are not in standard form. So, we need to use some non-standard random variates sampling approaches such as Metropolis Hastings (see Hastings, 1970 and Metropolis *et al.*, 1953 for reference) or adaptive rejection sampling (Ripley, 1987).

2.5 Our proposed method: Monte Carlo approximation through Gibbs output

As mentioned in the previous sections, the computation of parameter estimates may be laborious when likelihood functions involve high dimensional integrals. Regarding this problem, researchers have suggested various Monte Carlo methods to approximate the likelihood functions.

McCulloch (1997) has suggested a simulated maximum likelihood (SML) method. This method requires an optimal importance sampling function $h(\boldsymbol{u})$ to draw the random effects in order to carry out the Monte Carlo approximation. However, the SML method performs poorly if the choice of importance sampling function is far away from the true distribution for the random effects. Laplace importance sampling method, an application of SML, was proposed to use a second order Laplace expansion of the log-likelihood function $\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{u}) = \ln f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta})$ around $\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0)$, the maximizer of the joint density $f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}_0)$ based on a reference point $\boldsymbol{\theta}_0$ as the importance sampling function. The resulting importance sampling function is actually a normal density with mean $\hat{\boldsymbol{u}}(\boldsymbol{\theta}_0)$ and covariance matrix $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\theta}_0)$. The approximation will only be good when the reference point $\boldsymbol{\theta}_0$ is close to the true parameter $\boldsymbol{\theta}$. (See sections 2.4.3.3 and 2.4.3.4)

Alternatively, Monte Carlo relative likelihood method proposed by Geyer and Thompson (1992) approximates a relative likelihood function based on a reference point $\boldsymbol{\theta}_0$. However, Kuk and Cheng (1999) demonstrated that the resulting maximizer may again differ substantially from the true ML estimates if the reference point is chosen improperly. Although it would then be better if we update the reference point $\boldsymbol{\theta_0}$ to the current $\widehat{\boldsymbol{\theta}}^{(k)}$ each time and then simulate a new set of $\boldsymbol{u}$ based on the new reference point, this again requires iterations within iteration that makes the procedures tedious. (See section 2.4.3.5)

With regard to this problem, Kuk *et al.* (2003) extended the Monte Carlo relative likelihood method and suggested the use of the Gibbs output in the Monte Carlo approximation. Instead of relying on a single specified reference point $\boldsymbol{\theta}_0$, they assigned a conveniently chosen prior density $h(\boldsymbol{\theta})$ to $\boldsymbol{\theta}$. Suppose the marginal likelihood based on the observed data $\boldsymbol{y}$ over the random effects $\boldsymbol{u}$ is

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \int f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta})\ d\boldsymbol{u}.$$

The likelihood function is calculated as follows:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \int\int \frac{f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta})}{f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}^*)} f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}^*) d\boldsymbol{u}\ h(\boldsymbol{\theta}^*) d\boldsymbol{\theta}^* \\ &\propto \int\int \frac{f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta})}{f(\boldsymbol{y}, \boldsymbol{u}; \boldsymbol{\theta}^*)} f(\boldsymbol{u}, \boldsymbol{\theta}^*; \boldsymbol{y}) d\boldsymbol{u}\ d\boldsymbol{\theta}^* \\ \widehat{L}(\boldsymbol{\theta}) &\propto \frac{1}{M} \sum_{i=1}^{M} \frac{f(\boldsymbol{y}, \boldsymbol{u}_i; \boldsymbol{\theta})}{f(\boldsymbol{y}, \boldsymbol{u}_i; \boldsymbol{\theta}_i)} \qquad \text{where } (\boldsymbol{u}_i, \boldsymbol{\theta}_i) \sim f(\boldsymbol{u}, \boldsymbol{\theta} | \boldsymbol{y}) \quad (2.12) \end{aligned}$$

and the proportionality constant is the marginal density $f(\boldsymbol{y})$. Then they sampled the random effects $\boldsymbol{u}_j$ and the parameters $\boldsymbol{\theta}_j$ from a joint posterior density $f(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{y})$ and substituted them into (2.12) to approximate a relative likelihood function

$$\frac{L(\boldsymbol{\theta})}{f(\boldsymbol{y})} \approx \frac{1}{M} \sum_{j=1}^{M} \frac{f(\boldsymbol{y}, \boldsymbol{u}_j|\boldsymbol{\theta})}{f(\boldsymbol{y}, \boldsymbol{u}_j|\boldsymbol{\theta}_j)}$$

using a Monte Carlo approximation. Note that unlike the usual classical likelihood method, the likelihood $L(\boldsymbol{\theta})$ or log-likelihood $\ell(\boldsymbol{\theta})$ function for the model cannot be obtained easily because we approximate a relative likelihood function instead of the likelihood function directly. This method solves the problem of choosing a proper reference point and it does not require the simulation of a new set of random effects in each iteration. Besides, if the sample size M is large, the posterior density $f(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{y})$ will concentrate around the ML estimates $\widehat{\boldsymbol{\theta}}$ and so they automatically get a good approximation of $L(\boldsymbol{\theta})$ around $\widehat{\boldsymbol{\theta}}$.

To simulate $(\boldsymbol{u}_i, \boldsymbol{\theta}_i)$ from $f(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{y})$, Kuk *et al.* (2003) suggested making use of the Gibbs output from the WinBUGs program (Bayesian inference Using Gibbs Sampler[1]). This provides a new perspective of using both classical and Bayesian

[1]WinBUGS is a piece of computer software for the Bayesian analysis of complex statistical models using MCMC method and Gibbs sampler. It could be downloaded in the website of the BUGS project by the Biostatistics Unit of the Medical Research Council, in the University of Cambridge, with URL:http://www.mrc-bsu.cam.ac/bugs.

approaches in estimation.

In this thesis, we extend the methodology in Kuk *et al.* (2003) to models incorporating a drop-out modelling so as to account for the informative drop-out process in the data.

CHAPTER 3

MODELS FOR DROP-OUT PROCESS

3.1 Introduction

In this chapter, we will discuss the models for describing the drop-out process. The chapter will be presented as follows. Section 2 introduces the common missing patterns in drop-out models. In section 3, we further introduce three types of drop-out processes when the missing pattern is monotone and explain their impacts on parameter estimation in GLMMs. Then finally in section 4, we will discuss some modelling strategies for the informative drop-out process.

3.2 Missing patterns

Missing patterns can be divided into two forms: intermittent missing and monotone missing or drop-out. While intermittent missing refers to a sequence of measurements that has one or more gaps in it, monotone missing refers to a sequence of measurements that terminates once an individual drops out. Missing pattern that consists of both types of missing is called non-monotone missing. While Laird (1988) and Troxel *et al.* (1998) considered the general non-monotone miss-

ing, many methods of analysis specifically look at data with monotone missing or drop-outs (Little, 1995 and Fitzmaurice *et al.*, 1995). In this thesis, we focus on monotone missing for long time-series data although the models we propose can also cope with non-monotone missing, because we model the probability of drop-out in each occasion instead of the total number of observable outcomes for each subject (see Alfo and Aitkin, 2000).

3.3 Drop-out process

Based on the literature of Rubin (1976) and the discussion in Little and Rubin (1987) and Laird (1988), drop-out process can be classified into three types.

1. If the probability of drop-out does not depend upon the outcomes $\boldsymbol{y}$, the data are said to be missing completely at random (MCAR) or we have a completely random drop-out (CRD).

2. If the probability of drop-out depends upon the observed outcomes $\boldsymbol{y_o}$ and possibly some covariates $\boldsymbol{z}$, but not the unobserved outcomes $\boldsymbol{y_m}$, the data are said to be missing at random (MAR) or we have a random drop-out (RD).

3. If the probability of drop-out depends upon the unobserved outcomes $\boldsymbol{y_m}$,

we have an informative drop-out (ID).

In some instances, it is difficult to distinguish between MCAR and MAR. Technically, if the probability of unobserved outcomes or non-response depends only on covariates $\boldsymbol{z}$ and possibly some covariates $\boldsymbol{x}$ of the outcome model, then the data are said to be MCAR. For example, the attrition relating to subject characteristics, such as prognosis or treatment group, is considered as MCAR. MCAR and MAR are said to be *ignorable* because it is not necessary to specify a drop-out model $f(\boldsymbol{r}|\boldsymbol{z}, \boldsymbol{\alpha})$ in order to obtain valid likelihood based inferences about the parameters $\boldsymbol{\beta}$ in the outcome or data model $f(\boldsymbol{y_o}|\boldsymbol{x}, \boldsymbol{\beta})$. For data with ignorable drop-out, 'complete case' analyses which discard any units with missing data yield valid inferences although they may mean a loss of efficiency. Likelihood based methods using standard algorithms such as scoring (Fitzmaurice *et al.*, 1995) or missing-data tools such as EM and its extension (Dempster *et al.*, 1997; Meng and Rubin, 1991, 1993) can give consistent parameter estimates $\boldsymbol{\beta}$ because the joint density of the observed data can be separated into two densities, that is

$$f(\boldsymbol{y_o}, \boldsymbol{r}; \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = f(\boldsymbol{y_o}; \boldsymbol{x}, \boldsymbol{\beta}) f(\boldsymbol{r}|\boldsymbol{y_o}; \boldsymbol{z}, \boldsymbol{\alpha})$$

and hence the estimation of $\boldsymbol{\beta}$ can be done independently of the drop-out model $f(\boldsymbol{r}|\boldsymbol{y_o}; \boldsymbol{z}, \boldsymbol{\alpha})$.

In fact, some non-likelihood based methods of analysis such as the generalized estimating equations (GEE) approach proposed by Wei and Stram (1987) and Fitzmaurice *et al.* (1995), are also possible for data that are MCAR.

When the probability of drop-out is related to the subject's unobserved outcomes, the data are said to have an ID process (Wu and Carroll, 1988). In this case, the drop-out process is non-ignorable (NI) which indicates that valid likelihood based inferences can only be made by specifying a drop-out model $f(\boldsymbol{r}|\boldsymbol{y}_o, \boldsymbol{y}_m; \boldsymbol{z}, \boldsymbol{\alpha})$. The joint density of the observed data with in ID process is

$$f(\boldsymbol{y_o}, \boldsymbol{r}|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = f(\boldsymbol{y_o}|\boldsymbol{x}, \boldsymbol{\beta}) \int f(\boldsymbol{r}|\boldsymbol{y_o}, \boldsymbol{y_m}, \boldsymbol{z}, \boldsymbol{\alpha})\ d\boldsymbol{y_m}.$$

Some researchers have investigated different modelling approaches for longitudinal continuous response with an ID process. Wu and Carroll (1988) proposed a pseudo-maximum likelihood approach for a probit drop-out model. Schluchter (1992) adopted a likelihood approach via the EM algorithm based on a log-normal survival model. Mori *et al.* (1994) used an empirical Bayes approach based on a conditional linear model.

With regard to longitudinal categorical data, Fitzmaurice *et al.* (1995) considered an ID model for binary responses using logit link and Plackett distribution. However, the likelihood approach that he considered requires solving a complicated score function. Molenberghs et al. (1995) proposed maximum likelihood

(ML) methods based on marginal multinomial models for repeated ordinal responses. Similarly, Zhao and Prentice (1990) have also used ML methods, but based on a quadratic exponential model for repeated binary responses.

The recent proposed likelihood-based methods that account for an ID process include Brown (1990), Molenberghs et al. (1995), Mori *et al.* (1994), Schluchter (1992) and Wu and Carroll (1998). These methods depend strongly on the correct specification of the underlying distribution of the longitudinal measurements and the model assumptions about the missingness probabilities.

All likelihood-based methods are sensitive to the assumed distribution of the data, but these assumptions cannot be tested if the data are non-ignorably missing. Alternative methods like non-likelihood-based methods are therefore proposed. Some examples include the Bayesian approach via Markov chain Monte Carlo methods and Gibbs' sampler (Gelfand & Smith, 1990; Gelman & Rubin, 1992), generalized estimating equations (GEE) and weighted GEE approaches (Fitzmaurice *et al.*, 1995), and the augmented inverse probability of censoring weighted (IPCW) for non-ignorable nonresponses (Rotnizky *et al.*, 1998). These methods yield consistent estimates when the drop-out process is NI and the probability of drop-out is correctly specified.

In case when the drop-out mechanism is not well understood, sensitivity analyzes are suggested to assess the effect of inferences on target parameters from alternative assumptions about the drop-out process. See Laird (1988), Little (1995) and the sensitivity analysis to model assumptions of Glynn *et al.* (1986) and Baker & Laird (1988) for references. In Chapter 7, we conduct simulation studies in which simulated data sets with ID processes are fitted into three types of outcome models with and without an ID modelling. More details of the sensitivity analysis are given in that chapter.

3.4 Models for drop-out process

Correct specification of the drop-out model is crucial particularly when the drop-out process is non-ignorable. Otherwise it may lead to bias result. While GLMs or GLMMs are commonly used models for the outcomes, models for the drop-out process are more diversified. In this section, we will review some of these models.

3.4.1 Modelling the total number of occasions before drop-out

Suppose drop-out occurs in an experiment so that measurements are not observed at all T occasions for some subjects. When modelling the drop-out process, some

researchers choose to model the total number of observed outcomes for each subject in the experiment. Adopting the standard notation of Little and Rubin (1987), we introduce T random variables $r_{i1}, \ldots, r_{iT}$, where r_{it} is the drop-out indicator for subject i at time t and it equals 1 if the corresponding measurement y_{it} is observed and 0 if y_{it} is missing. For a special case of attrition when the missing data pattern is monotone, the information of such missing pattern denoted by a sequence of r_{it} can be summarized by a single random variable:

$$s_i = \sum_{t=1}^{T} r_{it};\ i = 1, \ldots, n,$$

indicating the total number of observed outcomes before drop-out for subject $i,\ i = 1, \ldots, n$ such that $s_i \leq T$.

Researchers have modelled the probability of s_i with covariates using some link functions. Diggle and Kenward (1994) and Roy and Lin (2002) regressed the conditional probability of s_i on the history of the measurement process, unobserved observations and some covariates, using a logit link function. Mori *et al.* (1994) proposed a method to estimate the rate of change of blood pressure from incomplete longitudinal data. This method involved modeling s_i by a geometric distribution with mean, a linear function of subject's rate of change.

Alfo and Altkin (2000) linked the conditional expectation of s_i to a linear function of covariates for a methadone clinic data (see Chapter 4 for detail). Let

p_{it} be the conditional probability of success ($y_{it} = 1$) at time t, $\boldsymbol{x}_{it}$ a p-dimensional vector of explanatory variables, $\boldsymbol{\beta}$ a vector of p parameters, $\boldsymbol{u}_i$, a vector of random coefficients and $\boldsymbol{z}_{it}$, a design vector for $\boldsymbol{u}_i$. Then, their proposed outcome model is

$$\theta_{it} = \log\left[\frac{p_{it}}{1-p_{it}}\right] = \boldsymbol{x}_{it}{}^T\boldsymbol{\beta} + \boldsymbol{z}_{it}{}^T\boldsymbol{u}_i + \alpha\ y_{i,t-1},$$

and the model for the informative drop-out process is

$$g[E(S_i|\boldsymbol{u}_i)] = \boldsymbol{v}_{it}{}^T\boldsymbol{\alpha}_s + \boldsymbol{u}_i{}^T\gamma_s$$

where $\boldsymbol{\alpha}_s$ and $\boldsymbol{\gamma}_s$ are vectors of parameters, $\boldsymbol{v}_{it}^T$ is a vector of covariates and g is some link functions.

Such model is defined as the *full shared parameter* model because the outcome model and the informative drop-out model are sharing random coefficient vectors $\boldsymbol{u}_i$. The parameters were estimated by means of an EM algorithm, without assuming any specific parametric distribution for the random coefficients in the outcome as well as the drop-out models. While this modelling strategy provides a simple alternate drop-out model based only on s_i, it is not suitable for modelling data with non-monotone missing pattern.

3.4.2 Modelling the probability of drop-out in each occasion

Other than modelling s_i, the total number of observations for subject i, some researchers prefer modelling directly the probability of r_{it}, a drop-out indicator at time t. This alternate type of drop-out modelling is more applicable to non-monotone missing. Fitzmaurice *et al.* (1995) and Rotnitzky *et al.* (1997) modelled the probability of r_{it} linearly on both observed and unobserved outcomes. Greenlees *et al.* (1982) and Little (1982) adopted the same modelling approach as Fitzmaurice *et al.* (1995), Molenberghs *et al.* (1994) and Rotnitzky *et al.* (1997), but used a probit link function instead. Fitzmaurice *et al.* (2001) adopted a logit link function to relate the conditional probability of r_{it} to the history of measurement process and some covariates. Wei *et al.* (1987) modelled the marginal probability of r_{it} with a quasi-likelihood function.

3.4.3 Modelling the censoring time

Instead of modelling the total number of observed outcomes s_i or the probability of drop-out indicator r_{it} in each occasion, some investigators focus on the censoring time, T_i, which is the time until a drop-out occurs for subject i. Generally speaking, T_i can be thought of as a continuous sense of s_i. One example of mod-

eling the censoring time can be found in Wu and Carroll (1988), who modelled T_i with some covariates using a probit link. Another example is that proposed by Schluchter (1992) who modelled the censoring time as trivariate-normally distributed with a subject specific random intercept and slope.

3.4.4 Modelling the drop-out mechanism implicitly

All models discussed above have incorporated a separate model for the primary responses and the drop-out process. However, some researchers do not assign a separate model for the drop-out process. They model the drop-out process implicitly in the outcome model instead. For instance, Wu and Bailey (1989) have developed a conditional linear model for the outcomes with a subject specific random intercept and slope. They further assumed that the subject specific random slope was a polynomial function of censoring time. Based only on these models, they developed estimation and testing procedures for the models without any explicit drop-out model.

CHAPTER 4

METHADONE CLINIC DATA

4.1 Introduction

In recent years, there has been a resurgence for the support of a methadone maintenance treatment (MMT) program in many countries as studies have revealed its contribution in reducing the risk of HIV infection among injecting heroin users in treatment. Associated with this expansion of methadone maintenance treatment, there is a growing research interest in trying to identify the factors that contribute to effective methadone treatment. To identify the factors that contribute to effective methadone maintenance treatment, we analyze a data set consisting of results of urine drug screens for patients who took part in a MMT at a clinic in Sydney in 1986. This research is actually motivated by this methadone clinic data set.

The chapter will be presented as follows. Section 2 introduces the data set reported by Chan *et al.* (1998) in details. Section 3 gives an overview of various modelling strategies that have been applied to this data set and describes the objective of this thesis in details.

4.2 Data set

This research is based on records of heroin users who were under methadone maintenance treatment (MMT) at a clinic in Western Sydney in 1986. Outcome measure is heroin use measured by urine testing performed once a week, on a day determined at random. Screens were recorded as positive ($y = 1$) or negative ($y = 0$) for morphine, the biological marker for heroin use. Records also contain other informations including patients' dosage of methadone d in milligram (mg) at the time (in week t) of urine test. The clinic required attendance for dosing seven days per week, with take-home doses of medication only provided in exceptional circumstances.

The analysis was performed using a restricted data set in which patients who completed less than 4 weeks of treatment were excluded. Patients with missing dose records were also excluded from our analysis. Finally, past experience showed that the treatment was most effective in the first half year of maintenance and beyond that, non-random drop-out began to occur, with patients who continued to use heroin being more likely to leave the treatment. Consequently, our study only looked at results of urine screens collected in the first 26 weeks of treatment so as to avoid the distorting effect of patients being on a withdrawal regimen, something that usually began after the first half year of maintenance.

There were 136 patients, submitting a total of 2872 urine screens with 16% of them being positive for heroin. The dosage averaged over the 2872 incidents is 64mg. Each patient submitted 4 to 26 weekly outcomes and the average number of treatment weeks per patient is 21.1 weeks. Fifty one patients dropped out before the end of 26 weeks and the rest having 26 outcomes were regarded as having completed the program. For all analyses, each urine screen result rather than each patient served as the unit of analysis. Some descriptive statistics of this data set is presented in Table 1 of Appendix A.

Since the aim of the analysis is to investigate the relationship between heroin use as detected by urine testing and various treatment factors, the response variable is the results of urine test. Regarding factors that associate with heroin use, there is a substantial body of evidence that methadone dose is important in influencing continued heroin use. Hence it is necessary to take into account the fluctuating methadone dose in assessing the influence of other treatment factors. Another factor included is the duration of treatment (called time effect) in weeks. As there is a strong belief that the time effect on heroin use levels off as time increases, such effect was transformed to $\ln t$ in all analyses where t is the duration of treatment in weeks. We have included an interaction term between the dose and time effects initially, but it was insignificant and was dropped from the model subsequently.

4.3 Modelling strategies

Many researchers have studied the methadone clinic data set. Chan *et al.* (1998) have adopted several modelling strategies within the GLM and GLMM frameworks to analyze this data set. First of all, it is clear that longitudinal binary data are not independent but serially correlated. To account for such dependency between observations, a marginal or conditional model may be used. In adopting a marginal model for binary data with a logit link function, parameters can be estimated using the Generalised Estimating Equation (GEE) approach (Liang & Zeger 1986) with a specified working correlation matrix of different auto-correlation structures. On the other hand, researchers may also consider the conditional logistic model proposed by Bonney (1987). While the marginal model is easy to interpret and the dependency structure can be explicitly modelled by a working correlation matrix incorporating different choices of correlation structures, the conditional logistic model has a more tractable likelihood function and can be extended easily to accommodate random effects. See Chan *et al.* (1998) for more details of the two approaches.

In this thesis, the conditional approach is adopted and we model the serial correlation using a first order auto-regressive (AR1) model such that the logit of the conditional probabilities $\Pr(y_t = 1|y_{t-1})$ depend linearly on the covariates

d and $\ln t$, as well as the previous outcomes y_{t-1}. The parameter estimation is carried out using the ML method.

Since results of separate fitting to each patient and score tests suggested that there was substantial between-patient variation (see Chan *et al.*, 1998 for the separate fitting and score tests). To account for the population heterogeneity and to facilitate subject-specific inference, Chan *et al.* (1998) further extended the conditional logistic model by introducing group effects. The mixture models revealed valuable information regarding the drug-taking habit of patients in different groups. By assuming that there are several groups of patients who react differently to methadone treatment, the data were fitted into two-, three- and four-group mixture models and the three-group mixture model was chosen based on Akaike information criterion (AIC). For instance, one group of patients have ceased taking heroin as a result of treatment (light-user group) while the other group of patients continued to use heroin regardless of the methadone dose received (heavy-user group). The remaining group of patients responded to the treatment in a dose-dependent fashion with reduced heroin use at a high methadone dose (medium-user group). Results confirmed that patients reacted differently to MMT.

Finally, a more direct way to account for the population heterogeneity was

to incorporate a random intercept term into the conditional logistic model. The random intercepts are assumed to follow a normal distribution with mean zero and variance σ^2. Results of fitting the conditional AR1 model, the mixture model and the random intercept model to data are similar. Both the dose and time effect of these models are significant suggesting that reduced drug use is associated with an increase in methadone dose and duration of treatment. Also, there is a strong and positive association between the present and previous outcomes y_{it} & $y_{i,t-1}$ suggesting that some patients in treatment tend to use heroin continously while others do not.

Recently, Chan and Leung (2003) have incorporated the conditional AR1 model and the mixture model with an ID modelling so as to account for the ID process of the data set. They ignored the *initial stage* problem (see Chan, 2000 for reference) in models by simply assuming that $y_{i0} = 0$ and focused on developing models that accommodate the ID process. They believed the biases in the regression coefficients caused by such assumption will be small because the time series in this data set are mostly long (see Chan, 2000).

Up tp now, few of the models on methadone clinic data did allow for the drop-out process of the data. Alfo and Aitkin (2000) have described the ID process by modeling the total number of observations before drop-out i.e. s_i of each patient

with some covariates in some link function (see section 3.4.1). They entered the initial observations y_{i1} as a covariate in the conditional outcome model for observations from $t = 2$ to $t = n_i$ and proposed two models: the first model (called AA1) has an interaction term '$y_{i1}\times$ dose' so that the dose effect becomes initial-use specific and the second model (called AA2) is a mixture model with group probabilities π_{k0} and π_{k1}, $k = 1, 2, 3$ depending on initial-use. The authors remarked that the dose effect is only significant among those non-initial users in the heavy user group (group 1) (as reported in Table 2 of Alfo and Aitkin, 2000) and the rest of group 1 together with group 2 (light users) and 3 (medium users) do not respond to the methadone treatment. They also found that after modeling the drop-out process, the parameter estimate of the dose effect changed by -0.00053 ($s.e. = 0.00019$). This result confirms the necessarity of incorporating an informative drop-out model in the analysis of the methadone clinic data.

The results of Chan & Leung (2003) and Alfo & Aitkin (2000) clearly support the incorporation of an ID model to the analysis of the methadone clinic data as the ID model allow for the selective attitude towards drop-out for heavy drug user which can otherwise lead to a false impression of reduced drug use over time and hence a false conclusion of significant time effect. On the other hand, there are still rooms for improvement in the modelling strategies for the drop-out process. *The objective of this thesis is to extend the existing conditional AR1 model and*

mixture model with an ID modelling to a random intercept model with ID that will account for the population heterogeneity in the data and compare the results between these three types of models.

CHAPTER 5

METHODOLOGY

In this chapter, we extend the three modelling strategies, namely the conditional AR1 model, the mixture model and the random intercept model by incorporating an ID model for the methadone clinic data. Tthe first two models were proposed in Chan and Leung (2003).

5.1 Conditional AR1 model with ID

Let y_{it} denote the binary outcome for patient i in week t. The vector of all possible outcomes for patient i can be separated into

$$\boldsymbol{y}_i = (y_{it})^T = (\ \underbrace{y_{i1}, ..., y_{i,n_i}}_{\text{Observed } \boldsymbol{y}_{oi}{}^T}\ ,\ \underbrace{y_{i,n_i+1}, ..., y_{i,n}}_{\text{Unobserved } \boldsymbol{y}_{mi}{}^T}\ )^T$$

where n_i denotes the number of observed y_{it} and the vector of all outcomes is denoted by $\boldsymbol{y}^T = (\boldsymbol{y}_1{}^T, \boldsymbol{y}_2{}^T, ..., \boldsymbol{y}_I{}^T)$.

Similarly, let r_{it} denote the drop-out indicator for patient i in week t such that $r_{it} = I(t > n_i)$ where $r_{it} = 1$ if y_{it} is unobserved $(t > n_i)$ and zero otherwise. Then the vector of all drop-out indicators for patient i is $\boldsymbol{r}_i = (r_{it})^T$ which is a series of n_i '0' followed by $26 - n_i$ '1'. For the outcome model, the

conditional probabilities of heroin use are logit linear in some covariates as well as the 'previous outcomes' $y_{i,t-1}$:

$$\text{logit}[\Pr(y_{it} = 1|y_{i,t-1}, \boldsymbol{\beta})] = \eta_{it} = \beta_o + \beta_d d_{it} + \beta_t \ln t + \beta_{pv}\ y_{i,t-1}. \tag{5.1}$$

For the ID model, the conditional probabilities of drop-out are logit linear in some covariates as well as the 'present outcomes' $y_{i,t}$ which signify ID:

$$\text{logit}[\Pr(r_{it} = 1|y_{it}, \boldsymbol{\alpha})] = \zeta_{it} = \alpha_o + \alpha_t \ln t + \alpha_{ps} y_{i,t} \tag{5.2}$$

for $t \leq n_i$ such that the present outcomes y_{it} are observed. At the time of drop-out when $t = n_i + 1$ and $n_i < 26$, the 'present outcome' y_{i,n_i+1} is unobserved. Then there are two possible conditional probabilities of drop-out:

$$\text{logit}[\Pr(r_{i,n_i+1} = 1|y_{i,n_i+1} = h, \boldsymbol{\alpha})] = \zeta_{i,n_i+1,h} = \alpha_o + \alpha_t \ln t + \alpha_{ps} h,\ \ h = 0, 1. \tag{5.3}$$

We model the probability of drop-out in each occasion instead of total number of occasions in the drop-out model because the latter approach can neither be extended to cases with non-monotone missing nor revealed factors that will affect the missingness pattern, although it can simplify the computation considerably.

A vector of parameters for the whole model is $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ where the p-dimensional vector of parameters in the outcome model is $\boldsymbol{\beta} = (\beta_o, \beta_d, \beta_t, \beta_{pv})^T$ and the q-dimensional vector of parameters in the drop-out model is $\boldsymbol{\alpha} = (\alpha_o, \alpha_t, \alpha_{ps})^T$.

Here we adopt a *selection model* factorization such that

$$f(\boldsymbol{y}_i, \boldsymbol{r}_i; \boldsymbol{\theta}) = f(\boldsymbol{y}_i; \boldsymbol{\theta}) f(\boldsymbol{r}_i | \boldsymbol{y}_i; \boldsymbol{\theta}). \tag{5.4}$$

See Rubin (1976) for the description of *selection models* and *pattern mixture models*.

Since y_{i,n_i+1} is unobserved for $n_i < 26$, the marginal probabilities $\Pr(r_{i,n_i+1} = 1|\boldsymbol{\theta})$ are estimated by

$$\widehat{\Pr}(r_{i,n_i+1} = 1|\boldsymbol{\theta}) = \sum_{h=0}^{1} \Pr(r_{i,n_i+1} = 1 | y_{i,ni+1} = h, \boldsymbol{\alpha}) \Pr(y_{i,ni+1} = h | y_{i,n_i}, \boldsymbol{\beta}). \tag{5.5}$$

5.1.1 Maximum Likelihood Approach

Let $p_{y,it} = \Pr(y_{it}) = \Pr(y_{it} = 1|y_{i,t-1}, \boldsymbol{\beta})$, $p_{r,ir} = \Pr(r_{it}) = \Pr(r_{it} = 1|y_{it}, \boldsymbol{\alpha})$ and $p_{yr,i,h} = \Pr(r_{i,n_i+1,h}) = \Pr(r_{i,n_i+1} = 1|y_{i,n_i+1} = h, \boldsymbol{\alpha})$, $h = 0, 1$. The 'observed data' likelihood $f(\boldsymbol{y}_o, \boldsymbol{r}|\boldsymbol{X}, \boldsymbol{\theta})$ is given by

$$\prod_{i=1}^{I} \left\{ \prod_{t=1}^{n_i} \Pr(y_{it})^{y_{it}} [1 - \Pr(y_{it})]^{(1-y_{it})} \cdot \prod_{t=2}^{n_i} [1 - \Pr(r_{it})] \cdot \left[\sum_{h=0}^{1} \Pr(r_{i,n_i+1,h}) \Pr(y_{i,n_i+1})^h [1 - \Pr(y_{i,n_i+1})]^{1-h} \right]^{I(n_i<26)} \right\}$$

where

$$\begin{aligned} L_y &= \prod_{i=1}^{I} \prod_{t=1}^{n_i} \Pr(y_{it})^{y_{it}} [1 - \Pr(y_{it})]^{(1-y_{it})} \\ L_{r0} &= \prod_{i=1}^{I} \prod_{t=2}^{n_i} [1 - \Pr(r_{it})] \end{aligned}$$

$$L_{r1} = \prod_{i=1}^{I}\left[\sum_{h=0}^{1}\Pr(r_{i,n_{i+1},h})\Pr(y_{i,n_{i+1}})^{h}[1-\Pr(y_{i,n_{i+1}})]^{1-h}\right]^{I(n_i<26)}$$

are respectively the likelihood for all y_{it}, $r_{it}=0$ and $r_{it}=1$. Then, we maximize the log-likelihood function $\ell = \ell_y + \ell_{r0} + \ell_{r1}$ for the observed data where $\ell_k = \ln L_k,\ k = y, r0, r1$ are given by

$$\begin{aligned}
\ell_y &= \sum_{i=1}^{I}\sum_{t=1}^{n_i}\ln\left(\frac{e^{y_{it}\eta_{it}}}{1+e^{\eta_{it}}}\right)\\
\ell_{r0} &= \sum_{i=1}^{I}\sum_{t=2}^{n_i}\ln\left(\frac{1}{1+e^{\zeta_{it}}}\right)\\
\ell_{r1} &= \sum_{i=1}^{I} I(n_i<26)\ln\left[\left(\frac{e^{\zeta_{i,n_i+1,1}}}{1+e^{\zeta_{i,n_i+1,1}}}\right)\left(\frac{e^{\eta_{i,n_i+1}}}{1+e^{\eta_{i,n_i+1}}}\right)+\left(\frac{e^{\zeta_{i,n_i+1,0}}}{1+e^{\zeta_{i,n_i+1,0}}}\right)\left(\frac{1}{1+e^{\eta_{i,n_i+1}}}\right)\right].
\end{aligned}$$

Newton Raphson (NR) method is used to solve for the maximum likelihood (ML) estimates of $\boldsymbol{\theta}$ from the log-likelihood function ℓ and the procedures are iterated until convergence. The first and second order derivatives of the log-likelihood function, ℓ' and ℓ'' as required in the NR procedures are given in Appendix B. The variance-covariance matrix of $\boldsymbol{\theta}$ can be obtained by inverting $-\ell''$.

5.1.2 Bayesian Approach

Let $\boldsymbol{y}$ be the observed data having joint density function $f(.)$, $\boldsymbol{\beta}$ be a vector of fixed effect parameters following a p-dimensional multivariate normal distribution with a mean $\boldsymbol{\mu_\beta}$ and a variance covariance matrix $\boldsymbol{\Sigma}_\beta$ which is a diagonal matrix with diagonal entities τ_{β_k}, and $\boldsymbol{u}_i$ be a vector of random effect parameters

following a m-dimensional multivariate normal distribution with mean $\mathbf{0}$ and a variance-covariance matrix $\boldsymbol{D}$. Very often, we set $\boldsymbol{D} = \sigma^2 \boldsymbol{I}$. Then the general framework of Bayesian hierarchy for a GLMM is

$$\begin{aligned} y_i &\sim f(y|\ \mu_i) \\ g(\mu_i) &= \boldsymbol{x_i}^T \boldsymbol{\beta} + \boldsymbol{z_i}^T \boldsymbol{u}_i \\ \boldsymbol{u}_i &\sim N(\mathbf{0}, \boldsymbol{D}),\ i = 1, \cdots, m \\ \beta_k &\sim N(\mu_{\beta_k}, \tau_{\beta_k}),\ k = 1, \cdots, p \\ \sigma^2 &\sim \text{Gamma}(a_\sigma, b_\sigma) \end{aligned}$$

where $g(\cdot)$ is a link function, μ_{β_k}, τ_{β_k}, a_σ and b_σ are fixed, and $\boldsymbol{x_i}$ and $\boldsymbol{z_i}$ are the design matrices for the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{u}_i$ respectively.

Now, $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_p, \sigma^2)^T$ which is a vector of parameters and $[\boldsymbol{y}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{\Sigma_\beta}]$ denotes the joint density. The Bayesian estimates are drawn from its conditional distribution by Gibbs sampler. For example, we sample

$$\begin{aligned} \boldsymbol{\beta} &\text{ from } [\boldsymbol{\beta}|\boldsymbol{u}, \boldsymbol{\Sigma_\beta}, \boldsymbol{D}, \boldsymbol{y}] \\ \boldsymbol{u} &\text{ from } [\boldsymbol{u}|\boldsymbol{\beta}, \boldsymbol{\Sigma_\beta}, \boldsymbol{D}, \boldsymbol{y}] \\ \boldsymbol{D} &\text{ from } [\boldsymbol{D}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{\Sigma_\beta}, \boldsymbol{D}] \\ \boldsymbol{\Sigma_\beta} &\text{ from } [\boldsymbol{\Sigma_\beta}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{D}, \boldsymbol{y}] \end{aligned}$$

In our case with binary data $\boldsymbol{y}$, the density function is $f(y_i) = {p_i}^{y_i}(1 - p_i)^{(1-y_i)}$

where p_i is the parameter of a bernoulli distribution. Usually a logit link function is used to link the linear function of covariates to the mean p_i so that $g(p_i) = \ln \frac{p_i}{1-p_i} = \sum_{k=1}^{p} \beta_k x_{ik}$. Hence the Bayesian hierarchy for the binary drug uses y_{it} as well as the drop-out indicators r_{it} with vague priors for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ is

$$
\begin{aligned}
y_{it} &\sim \text{Bernoulli}(p_{y,it}) \\
\text{logit}(p_{y,it}) &= \beta_o + \beta_d \; d_{it} + \beta_t \; \ln t + \beta_{pv} y_{i,t-1} \\
r_{it} &\sim \text{Bernoulli}(p_{it}) \\
\text{For } t \leq n_i, \qquad \text{logit}(p_{r,it}) &= \alpha_o + \alpha_t \; \ln t + \alpha_{ps} y_{it} \\
\text{For } n_i < 26, \quad \text{logit}(p_{r,i,n_i+1}) &= I_i(\alpha_o + \alpha_t \; \ln t + \alpha_{ps}) + (1 - I_i)(\alpha_o + \alpha_t \; \ln t) \\
I_i &\sim \text{Bernoulli}(p_{i,n_i+1}) \\
\text{logit}(p_{i,n_i+1}) &= \beta_o + \beta_d \; d_{i,n_i} + \beta_t \; \ln(n_i + 1) + \beta_{pv} y_{i,n_i} \\
\beta_o, \beta_d, \beta_t, \beta_{pv}, \alpha_o, \alpha_t, \alpha_{ps} &\sim N(0, 1000000).
\end{aligned}
$$

where $I_i = I(y_{i,n_i+1} = 1)$, the drug use indicator during the drop-out occasion. From the Gibbs output obtained from the Bayesian software WinBUGS, we discard the first 1000 observations in the burn-in period and take every 20 observations resulting in a sample of 500 observations. The auto-correlation functions and history plots of all posterior samples of parameters in Appendix C show that the samples have converged and are independent.

5.2 Mixture model with ID

Chan *et al.* (1998) showed that there are substantial group effects in the methadone clinic data and extended the outcome model to accommodate the group effects. They assigned a multinominal distribution for $\boldsymbol{\beta}$ such that $\boldsymbol{\beta} = \boldsymbol{\beta}_m = (\beta_{mo}, \beta_{md}, \beta_{mt}, \beta_{mpv})^T$ at a probability π_m. Although Chan *et al.* (1998) have fitted 2-, 3- and 4-group mixture models to the methadone clinic data and selected a 3-group mixture model by AIC, Chan and Leung (2003) considered a 2-group mixture model ($m = 2$) with an ID modelling for simplicity. The model can easily be extended to 3 or more groups and AIC can be used in the model selection. Moreover, Chan *et al.* (1998) proposed a group specific intercept and dose coefficients while the time in treatment and previous outcome effects are fixed across groups in the outcome model as

$$\text{logit}[\Pr(y_{it} = 1 | y_{i,t-1}, \boldsymbol{\beta}_m)] = \eta_{itm} = \beta_{mo} + \beta_{md}\, d_{it} + \beta_t\ \ln t + \beta_{pv} y_{i,t-1}$$

and it occurs at probability π_m ($m = 1, 2$ and $\pi_1 + \pi_2 = 1$). For the drop-out model, equations (5.2) and (5.3) follow. A vector of parameters for the whole model is $\boldsymbol{\theta} = ({\boldsymbol{\beta}_1}^T, {\boldsymbol{\beta}'}^T, {\boldsymbol{\beta}_2}^T, \boldsymbol{\alpha}^T, \pi_1)^T$ where $\boldsymbol{\beta}_m = (\beta_{mo}, \beta_{md})^T,\ m = 1, 2$ and $\boldsymbol{\beta}' = (\beta_t, \beta_{pv})^T$.

Since y_{i,n_i+1} as well as the latent group membership for patient i are unob-

served, the marginal probabilities $\Pr(r_{i,n_i+1}=1|\boldsymbol{\theta})$ are estimated by

$$\widehat{\Pr}(r_{i,n_i+1}=1|\boldsymbol{\theta}) =$$

$$\sum_{m=1}^{2}\sum_{h=0}^{1}\Pr(r_{i,n_i+1}=1|y_{i,ni+1}=h,\boldsymbol{\beta}=\boldsymbol{\beta}_m,\boldsymbol{\alpha})\Pr(y_{i,ni+1}=h|y_{i,n_i},\boldsymbol{\beta}=\boldsymbol{\beta}_m)\ \pi_m.$$

5.2.1 Maximum Likelihood Approach

The 'observed' data likelihood $f(\boldsymbol{y}_o,\boldsymbol{r}|\boldsymbol{\theta})$ is given by

$$\prod_{i=1}^{I}\left\{\sum_{m=1}^{2}\pi_m\left[\prod_{t=1}^{n_i}\Pr(y_{itm})^{y_{it}}[1-\Pr(y_{itm})]^{(1-y_{it})}\right]\cdot\prod_{t=2}^{n_i}[1-\Pr(r_{it})]\cdot\right.$$
$$\left.\left[\sum_{m=1}^{2}\sum_{h=0}^{1}\Pr(r_{i,n_i+1,hm})\Pr(y_{i,n_i+1,m})^h[1-\Pr(y_{i,n_i+1,m})]^{1-h}\ \pi_m\right]^{I(n_i<26)}\right\}$$

where

$$\begin{aligned}
L_y &= \prod_{i=1}^{I}\left\{\sum_{m=1}^{2}\pi_m\left[\prod_{t=1}^{n_i}\Pr(y_{itm})^{y_{it}}[1-\Pr(y_{itm})]^{(1-y_{it})}\right]\right\},\\
L_{r0} &= \prod_{i=1}^{I}\prod_{t=2}^{n_i}[1-\Pr(r_{it})],\\
L_{r1} &= \prod_{i=1}^{I}\left[\sum_{h=0}^{1}\sum_{m=1}^{2}\Pr(r_{i,n_i+1,hm})\Pr(y_{i,n_i+1,m})^h[1-\Pr(y_{i,n_i+1,m})]^{1-h}\ \pi_m\right]^{I(n_i<26)}.
\end{aligned}$$

Then we maximize the log-likelihood function $\ell=\ell_y+\ell_{r0}+\ell_{r1}$ for the observed data where $\ell_v=\ln L_v,\ v=y,r0,r1$ are given by

$$\begin{aligned}
\ell_y &= \sum_{i=1}^{I}\ln\left[\sum_{m=1}^{2}\pi_m\left(\prod_{t=1}^{n_i}\frac{e^{y_{it}\eta_{itm}}}{1+e^{\eta_{itm}}}\right)\right]\\
\ell_{r0} &= \sum_{i=1}^{I}\sum_{t=2}^{n_i}\ln\left(\frac{1}{1+e^{\zeta_{it}}}\right)
\end{aligned}$$

$$\ell_{r1} = \sum_{i=1}^{I} I(n_i < 26)$$
$$\ln\left\{\sum_{m=1}^{2}\left[\left(\frac{e^{\zeta_{i,n_i+1,1}}}{1+e^{\zeta_{i,n_i+1,1}}}\right)\left(\frac{e^{\eta_{i,n_i+1,m}}}{1+e^{\eta_{i,n_i+1,m}}}\right)\pi_m + \left(\frac{e^{\zeta_{i,n_i+1,0}}}{1+e^{\zeta_{i,n_i+1,0}}}\right)\left(\frac{1}{1+e^{\eta_{i,n_i+1,m}}}\right)\pi_m\right]\right\}.$$

Again, NR method is used to solve for $\boldsymbol{\theta}$ from the log-likelihood function ℓ and the procedures are iterated until convergence. The first and second order derivatives, ℓ' and ℓ'' as required in the NR procedures are given in Appendix B. However, Chan and Leung (2003) remarked that the computation required in the NR procedures is complicate as this model involves many parameters and there is also a problem of convergence for the ML estimates. To solve this problem, one may resort to adjusting the Hessian matrix $-\ell''$ according to (5.8) as given in the coming section of (5.3.1).

5.2.2 Bayesian Approach

Bayesian hierarchy with vague priors for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}'$, $\boldsymbol{\beta}_2$, $\boldsymbol{\alpha}$ and π_1 are

$$\begin{aligned} y_{it} &\sim \text{Bernoulli}(p_{y,it}) \\ \text{logit}(p_{y,it}) &= I_{y,i}(\beta_{1o} + \beta_{1d}\ d_{it} + \beta_t\ \ln t + \beta_{pv} y_{i,t-1}) + \\ &\quad (1 - I_{y,i})(\beta_{2o} + \beta_{2d}\ d_{it} + \beta_t\ \ln t + \beta_{pv} y_{i,t-1}) \\ I_{y,i} &\sim \text{Bernoulli}(\pi_1) \end{aligned}$$

For $t \le n_i$, $\quad \text{logit}(p_{r,it}) = \alpha_o + \alpha_t\ \ln t + \alpha_{ps} y_{it}$

$$
\begin{aligned}
\text{For } n_i < 26, \quad \text{logit}(p_{r,i,n_i+1}) &= I_{r,i}[\alpha_o + \alpha_t \ \ln(n_i + 1) + \alpha_{ps}] + \\
&\quad (1 - I_{r,i})[\alpha_o + \alpha_t \ \ln(n_1 + 1)] \\
I_{r,i} &\sim \text{Bernoulli}(p_{y,i,n_i+1}) \\
\text{logit}(p_{y,i,n_i+1}) &= I_{y,i}[\beta_{1o} + \beta_{1d} \ d_{i,n_i} + \beta_t \ \ln(n_i + 1) + \beta_{pv} y_{i,n_i}] + \\
&\quad (1 - I_{y,i})[\beta_{2o} + \beta_{2d} \ d_{i,n_i} + \beta_t \ln(n_i + 1) + \beta_{pv} y_{i,n_i}] \\
\beta_{1o}, \beta_{1d}, \beta_t, \beta_{pv}, \beta_{2o}, \beta_{2d}, \alpha_0, \alpha_t, \alpha_{ps} &\sim N(0, 1000000) \\
\pi_1 &\sim U(0, 1)
\end{aligned}
$$

where $I_{y,i} = I(\boldsymbol{\beta} = {\boldsymbol{\beta}_1}^T, {\boldsymbol{\beta}'}^T)$, the group-1 indicator for patient i and $I_{r,i} = I(y_{i,n_i+1} = 1)$, the drug use indicator during the drop-out occasion. From the Gibbs output, we discard the first 1000 observations in the burn-in period and take every 38 observations resulting in a sample of 500 observations. Again, the auto-correlation functions and history plots of all posterior samples in Appendix C show that the samples have converged and are independent.

The results of the analysis for the conditional AR1 and 2-group mixture models with an ID modelling using both ML and Bayesian approaches are reported in Chan and Leung (2003) and are included in Tables 2 and 3 respectively in Appendix A for reference.

5.3 Random intercept model with ID

Believing the regression coefficients are patient-specific, Chan *et al.* (1998) incorporated a random intercept into the conditional AR1 model and assign a normal distribution to the random intercept $u_i,\ i = 1, \cdots, I$ such that $u_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. To allow for an ID process in the data, we further extend the model by incorporating an ID model. Hence the outcome model is

$$\text{logit}[\Pr(y_{it} = 1|y_{i,t-1}, \boldsymbol{\beta})] = \eta_{it} = u_i + \beta_o + \beta_d d_{it} + \beta_t \ln t + \beta_{pv} y_{i,t-1}$$

and the drop-out model follows equations (5.2) and (5.3). A vector of parameters for the whole model is $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \sigma^2)^T$.

Since y_{i,n_i+1} and the random intercept u_i is unobserved, the marginal probabilities $\Pr(r_{i,n_i+1} = 1|\boldsymbol{\theta})$ are estimated by

$$\widehat{\Pr}(r_{i,n_i+1} = 1|\boldsymbol{\theta}) =$$

$$\sum_{h=0}^{1} \Pr(r_{i,n_i+1} = 1|y_{i,n_i+1} = h, \boldsymbol{\alpha}) \int_{-\infty}^{\infty} \Pr(y_{i,n_i+1} = h|y_{i,n_i}, \boldsymbol{\beta}, u_i)\ \phi(u_i|0, \sigma^2)\ du_i$$

where $\phi(u_i|\ 0, \sigma^2)$ is the density function of the normal distribution $N(0, \sigma^2)$.

5.3.1 Maximum Likelihood Approach

The 'observed data' likelihood function $L(\boldsymbol{\theta}) = f(\boldsymbol{y}, \boldsymbol{r}|\boldsymbol{x}, \boldsymbol{\theta})$ is

$$\prod_{i=1}^{I} \left\{ \prod_{t=2}^{n_i} [1 - \Pr(r_{it})] \int_{-\infty}^{\infty} \prod_{t=1}^{n_i} \Pr(y_{it})^{y_{it}} [1 - \Pr(y_{it})]^{(1-y_{it})} \right.$$

$$\left[\sum_{h=0}^{1} \Pr(r_{i,n_i+1,h}) \Pr(y_{i,n_i+1})^h [1 - \Pr(y_{i,n_i+1})]^{1-h}\right]^{I(n_i<26)} \phi(u_i|\ 0, \sigma^2)\ du_i \Bigg\} \tag{5.6}$$

Using the Laplace Importance sampling method with the importance function $\phi(\boldsymbol{u}|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ where $\boldsymbol{\mu}^* = (u_1^*, \ldots, u_I^*)$ is the ML estimate of $\boldsymbol{u}$ with respect to the log of the 'complete data' likelihood $\ell(\boldsymbol{u}) = \ln f(\boldsymbol{y}, \boldsymbol{r}, \boldsymbol{u}|\boldsymbol{\theta}_0)$ when $\boldsymbol{\theta}$ is set to a fixed value $\boldsymbol{\theta}_0$ in (5.6) and $\boldsymbol{\Sigma}^* = \left[\dfrac{\partial^2 \ell(\boldsymbol{u})}{\partial \boldsymbol{u}^2}\right]^{-1}$ is the information matrix (see section 2.4.3.4 and McCulloch (1997)), the 'observed data' likelihood function $L(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\theta}_0) = f(\boldsymbol{y}, \boldsymbol{r}; \boldsymbol{\theta}, \boldsymbol{\theta}_0)$ becomes

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{I} \left[\prod_{t=2}^{n_i} \left(\frac{1}{1+e^{\zeta_{it}}}\right)\right] \left[\prod_{t=1}^{n_i} \left(\frac{e^{y_{it}\eta_{it}}}{1+e^{\eta_{it}}}\right)\right] \left[\left(\frac{e^{\zeta_{i,n_i+1,1}}}{1+e^{\zeta_{i,n_i+1,1}}}\right)\left(\frac{e^{\eta_{i,n_i+1}}}{1+e^{\eta_{i,n_i+1}}}\right) + \left(\frac{e^{\zeta_{i,n_i+1,0}}}{1+e^{\zeta_{i,n_i+1,0}}}\right)\left(\frac{1}{1+e^{\eta_{i,n_i+1}}}\right)\right]^{I(n_i<26)} \right\} \frac{\prod_{i=1}^{I} \phi(u_i|0,\sigma^2)}{\phi_I(\boldsymbol{u}|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)}\ \phi_I(\boldsymbol{u}|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)\ d\boldsymbol{u}$$

and the log of its Monte Carlo approximation, $\ell_M(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{u})$ is

$$\ln \left\{ \frac{1}{M} \sum_{j=1}^{M} \left\{ \prod_{i=1}^{I} \left[\prod_{t=2}^{n_i} \left(\frac{1}{1+e^{\zeta_{it}}}\right)\right] \left[\prod_{t=1}^{n_i} \left(\frac{e^{y_{it}\eta_{itj}}}{1+e^{\eta_{itj}}}\right)\right] \left[\left(\frac{e^{\zeta_{i,n_i+1,1}}}{1+e^{\zeta_{i,n_i+1,1}}}\right) \left(\frac{e^{\eta_{i,n_i+1,j}}}{1+e^{\eta_{i,n_i+1,j}}}\right) + \left(\frac{e^{\zeta_{i,n_i+1,0}}}{1+e^{\zeta_{i,n_i+1,0}}}\right)\left(\frac{1}{1+e^{\eta_{i,n_i+1,j}}}\right)\right]^{I(n_i<26)} \frac{\phi(u_{ij}|0,\sigma^2)}{\phi_I(\boldsymbol{u}_j|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)} \right\} \right\}, \tag{5.7}$$

where the random effects $\boldsymbol{u}_j = (u_{1j}, \ldots, u_{Ij})$ are drawn from $N_I(\boldsymbol{u}|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ and M is the number of simulations. Then we obtain its first and second order derivatives, namely $\ell'_M(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{u})$ and $\ell''_M(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{u})$ and use the Monte Carlo

Newton Raphson method to find the ML estimate of $\boldsymbol{\theta}$. To solve the local nature of the approximation based on the reference point $\boldsymbol{\theta}_0$, the ML estimate $\widehat{\boldsymbol{\theta}}^{(k)}$ at iteration k, say is set as the next $\boldsymbol{\theta}_0$ to obtain an updated ML estimate $\widehat{\boldsymbol{\theta}}^{(k+1)}$ using an updated set of random effects $\boldsymbol{u}^{(k)}$ (based on the updated reference point $\widehat{\boldsymbol{\theta}}^{(k)}$) in (5.7) and the procedures are iterated until convergence. However the estimation becomes laborious as it requires iterations within iterations.

Moreover, we have encountered convergence problem possibly because the importance sampling function $N_I(\boldsymbol{u}|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$ for $\boldsymbol{u}$ deviates a lot from the true posterior and hence the Monte Carlo approximationtain in (5.7) is not good enough. Furthermore, the MCNR procedure is quite erratic due to the use of Monte Carlo gradient vector $\ell'_M(\boldsymbol{\theta}^{(v)}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{u}^{(k)})$ and Hessian matrix $-\ell''_M(\boldsymbol{\theta}^{(v)}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{u}^{(k)})$ at iteration v within iteration k which is sometimes not a positive definite. Kuk and Cheng (1997) thus suggested halving $\boldsymbol{I}_2$ s times until

$$\boldsymbol{I}_1 - \frac{1}{2^s}\boldsymbol{I}_2 \tag{5.8}$$

is positive definite and use it for $-\ell''_M(\boldsymbol{\theta}^{(v)}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\theta}^{(k)})$. Note that $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ are given by (2.9) with $\ell'_M(\boldsymbol{\theta}^{(v)}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\theta}^{(k)})$ and $\ell''_M(\boldsymbol{\theta}^{(v)}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{\theta}^{(k)})$ corresponds to $\ell'_M(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$ and $\ell''_M(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$ in (2.8).

Instead of adopting this method, we adopt another method, the Monte Carlo approximation through Gibbs output as discussed in section 2.5 to approximate

a relative likelihood function $\frac{L(\boldsymbol{\theta})}{f(\boldsymbol{y})}$ in our analysis. We sample both the random effects $\boldsymbol{u}_i$ as well as the parameters $\boldsymbol{\theta}_i$ from a joint posterior density as in the Bayesian analysis and use them to evaluate a relative likelihood function by Monte Carlo approximation. Since this methodology applies the Bayesian results in the evaluation of a classical likelihood function, it integrates both the Bayesian and classical approaches together.

Referring to (2.12), the approximated likelihood function for the observed data is

$$L(\boldsymbol{\theta}) \propto \frac{1}{M} \sum_{j=1}^{M} \frac{f(\boldsymbol{y}, \boldsymbol{u}_j|\boldsymbol{\theta})}{f(\boldsymbol{y}, \boldsymbol{u}_j|\boldsymbol{\theta}_j)} = \frac{1}{M} \sum_{j=1}^{M} L_{qj}, \qquad (\boldsymbol{u}_j, \boldsymbol{\theta}_j) \sim f(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{y})$$

where

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{u}_j|\boldsymbol{\theta}) &= \prod_{i=1}^{I} \left\{ \left[\prod_{t=2}^{n_i} \left(\frac{1}{1+e^{\zeta_{it}}} \right) \right] \left[\prod_{t=1}^{n_i} \left(\frac{e^{y_{it}\eta_{itj}}}{1+e^{\eta_{itj}}} \right) \right] \left[\left(\frac{e^{\zeta_{i,n_i+1,1}}}{1+e^{\zeta_{i,n_i+1,1}}} \right) \left(\frac{e^{\eta_{i,n_i+1,j}}}{1+e^{\eta_{i,n_i+1,j}}} \right) \right. \right. \\ &\quad \left. \left. + \left(\frac{e^{\zeta_{i,n_i+1,0}}}{1+e^{\zeta_{i,n_i+1,0}}} \right) \left(\frac{1}{1+e^{\eta_{i,n_i+1,j}}} \right) \right]^{I(n_i<26)} \phi(u_{ij}|0,\sigma^2) \right\} \end{aligned}$$

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{u}_j|\boldsymbol{\theta}_j) &= \prod_{i=1}^{I} \left\{ \left[\prod_{t=2}^{n_i} \left(\frac{1}{1+e^{\zeta^*_{itj}}} \right) \right] \left[\prod_{t=1}^{n_i} \left(\frac{e^{y_{it}\eta^*_{itj}}}{1+e^{\eta^*_{itj}}} \right) \right] \left[\left(\frac{e^{\zeta^*_{i,n_i+1,1,j}}}{1+e^{\zeta^*_{i,n_i+1,1,j}}} \right) \left(\frac{e^{\eta^*_{i,n_i+1,j}}}{1+e^{\eta^*_{i,n_i+1,j}}} \right) \right. \right. \\ &\quad \left. \left. + \left(\frac{e^{\zeta^*_{i,n_i+1,0,j}}}{1+e^{\zeta^*_{i,n_i+1,0,j}}} \right) \left(\frac{1}{1+e^{\eta^*_{i,n_i+1,j}}} \right) \right]^{I(n_i<26)} \phi(u^*_{ij}|0,\sigma^{*2}_j) \right\}, \end{aligned}$$

$\phi(\cdot|\mu, \sigma^2)$ is a Normal density with mean μ and variance σ^2, the random effects as well as parameters $(\boldsymbol{u}_j, \boldsymbol{\theta}_j) =$

$(u^*_{1j}, u^*_{2j}, \ldots, u^*_{Ij}, \beta^*_{oj}, \beta^*_{dj}, \beta^*_{tj}, \beta^*_{pvj}, \alpha^*_{oj}, \alpha^*_{tj}, \alpha^*_{ps,j}, \sigma^{*2}_j)'$ are drawn from $f(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{y})$,

$$\begin{aligned}
\eta_{itj} &= u^*_{ij} + \beta_o + \beta_d d_{it} + \beta_t \ln t + \beta_{pv}\ y_{i,t-1},\ t = 1, \ldots, n_i + 1, \quad \text{is a linear} \\
&\quad \text{function of parameters } \boldsymbol{\beta} \text{ \& sample value } u^*_{ij}; \\
\zeta_{it} &= \alpha_o + \alpha_t \ln t + \alpha_{ps}\ y_{it},\ t = 2, \ldots, n_i, \quad \text{is a linear function of parameters } \boldsymbol{\alpha}; \\
\eta^*_{itj} &= u^*_{ij} + \beta^*_{oj} + \beta^*_{dj} d_{it} + \beta^*_{tj} \ln t + \beta^*_{pv,j}\ y_{i,t-1},\ t = 1, \ldots, n_i + 1, \quad \text{is a linear} \\
&\quad \text{function of sample values;} \\
\zeta^*_{itj} &= \alpha^*_{oj} + \alpha^*_{tj} \ln t + \alpha^*_{ps,j}\ y_{it},\ t = 2, \ldots, n_i, \quad \text{is a linear function of sample values;} \\
\zeta^*_{i,n_i+1,h,j} &= \alpha^*_{oj} + \alpha^*_{tj} \ln t + \alpha^*_{ps,j}\ h,\ h = 0, 1 \ \text{ is a linear function of sample values.}
\end{aligned}$$

To approximate the relative likelihood function $\dfrac{L(\boldsymbol{\theta})}{f(\boldsymbol{y})}$ closely using the Monte Carlo approximation, the number of simulation M should be large. Here we use $M = 20,000$ sets of simulation for $(\boldsymbol{u}_j, \boldsymbol{\theta}_j)$.

The full conditional density $f(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{y})$ is not in a standard form for $\boldsymbol{u}$ and hence non-standard sampling methods such as Metropolis Hastings or Adaptive Rejection sampling can be used. Our Gibbs output are again obtained from WinBUGS, adopting vague prior for $\boldsymbol{\theta}$. To obtain the ML estimates using the Monte Carlo Newton Raphson method (see section 2.4.3.4), we calculate the first and second order derivatives of the log-likelihood function denoted by $\ell'(\boldsymbol{\theta}; \boldsymbol{y})$ and $\ell''(\boldsymbol{\theta}; \boldsymbol{y})$ respectively. See Appendix B for details. Then ML estimates are

updated iteratively until converge by

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - [\ell''(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})]^{-1}\ \ell'(\boldsymbol{\theta}^{(k)}; \boldsymbol{y})$$

where k is the number of iterations in the Newton Raphson method. Bayesian estimates are set to be the initial values. However the standard error of σ^2, is found to be negative. One possible reason may be due to the constant variability of the random intercepts u_i across patients denoted by σ^2, leading to a low standard error for σ^2 and hence resulting in a very low or even a negative standard error of σ^2. The other reason may be due to the non-positive-definite for the Hessian matrix $-\ell''_M(\boldsymbol{\theta}^{(v)}; \boldsymbol{y}, \boldsymbol{r}, \boldsymbol{u}^{(k)})$.

5.3.2 Bayesian Approach

The Bayesian hierarchy with vague priors for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, σ^2 and $\boldsymbol{u}$ is

$$\begin{aligned}
y_{it} &\sim \text{Bernoulli}(p_{y,it}) \\
\text{logit}(p_{y,it}) &= u_i + \beta_o + \beta_d\ d_{it} + \beta_t\ \ln t + \beta_{pv} y_{i,t-1} \\
r_{it} &\sim \text{Bernoulli}(p_{r,it}) \\
\text{For } t \leq n_i, \qquad \text{logit}(p_{r,it}) &= \alpha_o + \alpha_t\ \ln t + \alpha_{ps} y_{it} \\
\text{For } n_i < 26, \quad \text{logit}(p_{i,n_i+1}) &= I_i * (\alpha_o + \alpha_t\ \ln t + \alpha_{ps}) + (1 - I_i) * (\alpha_0 + \alpha_t\ \ln t) \\
I_i &\sim \text{Bernoulli}(p_{y,i,n_i+1})
\end{aligned}$$

$$
\begin{aligned}
\text{logit}(p_{y,i,n_i+1}) &= u_i + \beta_o + \beta_d \; d_{i,n_i} + \beta_t \; \ln(n_i + 1) + \beta_{pv} y_{i,n_i} \\
u_i &\sim N(0, \sigma^2) \\
\beta_o, \beta_d, \beta_t, \beta_{pv}, \alpha_o, \alpha_t, \alpha_{ps} &\sim N(0, 1000000) \\
\sigma^2 &\sim IG(0, 1000000)
\end{aligned}
$$

where $r_{it} = 0$ for $t \leq n_i$, $r_{i,n_i+1} = 1$ for $n_i < 26$ and $I_i = I(y_{i,n_i+1} = 1)$, the drug use indicator during drop-out. From the Gibbs output, we discard the first 1000 observations in the burn-in period and take every 35 observations resulting in a sample of 525 observations. The auto-correlation and history plots in appendix C show that the sample is satisfactory.

CHAPTER 6

RESULT

6.1 Goodness of fit

6.1.1 Introduction

In order to gain an insight of the goodness of fit of each model, we have computed the Akaike Information Criterion (AIC) defined as

$$AIC = -2\ell + 2p \tag{6.1}$$

where ℓ is the log-likelihood of the model and p is the number of parameters in the model. We compute the AIC rather than the second order Akaike Information Criterion (AIC_c), a second order variant of AIC (see Sakamoto *et al.*,1986 for reference), defined as

$$AIC = -2\ell + 2p\left(\frac{n}{n-p-1}\right),$$

because our sample size n is large with respect to the number of parameters p. Since we have adopted both the likelihood and Bayesian methods to obtain the parameter estimates of models, we thus have to use different methods to approximate ℓ so as to obtain AIC. For Bayesian inference, the computation of

AIC requires ℓ which cannot be obtained from the estimation procedures as in the classical likelihood base inference, we thus 'approximate' ℓ using some special procedures that will be described in the following section.

On the other hand, it should be noted that we cannot directly compare the AICs between models of different types, with and without an ID modelling, or under ML and Bayesian approaches because these models are not in the same hierarchy and parameter estimations are not done using the same method. In addition, as we adopt different methods to approximate ℓ in order to obtain AIC, the AICs reported serve only a reference for the goodness-of-fit of each model therefore.

6.1.2 Procedures

6.1.2.1 Bayesian approach

Since we use the WinBUGS package to implement the models when adopting a Bayesian approach, we have already obtained the Gibbs outputs of each models. As the Gibbs output of each model consists of posterior sample of all iterations after the burn-in period, we first calculate the log-likelihood function of the posterior sample for each iteration. For the random intercept model with an ID

modelling, suppose

$$(u_{1j}, \ldots, u_{Ij}, \beta_{oj}, \beta_{dj}, \beta_{tj}, \beta_{pv,j}, \alpha_{oj}, \alpha_{tj}, \alpha_{ps,j}, \sigma_j^2)'$$

are the posterior sample in the $j-th$ iteration obtained from WinBUGS. Then, we calculate the 'observed data' log-likelihood function of this sample condition on the random effects as

$$\begin{aligned} \ell_{oj} &= \ln\left\{\prod_{i=1}^{I}\left[\prod_{t=2}^{n_i}\left(\frac{1}{1+e^{\zeta_{itj}}}\right)\right]\left[\prod_{t=1}^{n_i}\left(\frac{e^{y_{it}\eta_{itj}}}{1+e^{\eta_{itj}}}\right)\right]\left[\left(\frac{e^{\zeta_{i,n_i+1,1,j}}}{1+e^{\zeta_{i,n_i+1,1,j}}}\right)\right. \right. \\ &\left.\left.\left(\frac{e^{\eta_{i,n_i+1,j}}}{1+e^{\eta_{i,n_i+1,j}}}\right)+\left(\frac{e^{\zeta_{i,n_i+1,0,j}}}{1+e^{\zeta_{i,n_i+1,0,j}}}\right)\left(\frac{1}{1+e^{\eta_{i,n_i+1,j}}}\right)\right]^{I(n_i<26)}\right\} \quad (6.2) \end{aligned}$$

where

$$\eta_{itj} = u_{ij} + \beta_{oj} + \beta_{dj}\ d_{it} + \beta_{tj}\ln t + \beta_{pv,j}\ y_{i,t-1}, \quad (6.3)$$

$$\zeta_{itj} = \alpha_{oj} + \alpha_{tj}\ln t + \alpha_{ps,j}\ y_{i,t-1},, \quad (6.4)$$

$$\zeta_{i,n_i+1,h,j} = \alpha_{oj} + \alpha_{tj}\ln t + \alpha_{ps,j}\ h,\ h = 0, 1. \quad (6.5)$$

and repeat the calculation for each iteration $j,\ j = 1, \ldots, M$ $(M = 525)$. After getting the 'observed data' conditional log-likelihood functions ℓ_{oj} for all iterations, we approximate ℓ as required in the calculation of AIC in (6.1) by the minimum of these ℓ_{oj}, i.e. $\ell \approx \min_j\{\ell_{oj}\}$ (see Wood, 2001 for the procedures in the calculation of BIC for reference). The computation of AIC for the conditional AR1 model and 2-group mixture model is similar to that for the random intercept model.

For the conditional AR1 model, the 'observed data' log-likelihood is the same as (6.2) with (6.4) and (6.5) remain unchanged except

$$\eta_{itj} = \beta_{oj} + \beta_{dj}\ d_{it} + \beta_{tj} \ln t + \beta_{pv,j}\ y_{i,t-1}$$

and the posterior sample in the $j-th$ iteration is $(\beta_{oj}, \beta_{dj}, \beta_{tj}, \beta_{pv,j}, \alpha_{oj}, \alpha_{tj}, \alpha_{ps,j})'$. Regarding the 2-group mixture model, the 'observed data' log-likelihood is defined as

$$\begin{aligned}\ell_{oj} \quad = \quad & \prod_{i=1}^{I} \left\{ \sum_{m=1}^{2} \pi_m \left[\prod_{t=1}^{n_i} \frac{e^{y_{it}\eta_{itkj}}}{1+e^{\eta_{itkj}}} \right] \cdot \prod_{t=2}^{n_i} \left[\frac{1}{1+e^{\zeta_{itj}}} \right] \cdot \left[\sum_{m=1}^{2} \left[\left(\frac{e^{\zeta_{i,n_i+1,1,j}}}{1+e^{\zeta_{i,n_i+1,1,j}}} \right) \right.\right.\right. \\ & \left.\left.\left. \left(\frac{e^{\eta_{i,n_i+1,k,j}}}{1+e^{\eta_{i,n_i+1,m,j}}} \right) \pi_{kj} + \left(\frac{e^{\zeta_{i,n_i+1,0,j}}}{1+e^{\zeta_{i,n_i+1,0,j}}} \right) \left(\frac{1}{1+e^{\eta_{i,n_i+1,m,j}}} \right) \pi_{mj} \right] \right]^{I(n_i<26)} \Bigg\}\end{aligned}$$

where

$$\eta_{itmj} \quad = \quad \beta_{omj} + \beta_{dmj}\ d_{it} + \beta_{tj} \ln t + \beta_{pv,j}\ y_{i,t-1}, \qquad m = 1, 2,$$

with (6.4) and (6.5) remain unchanged, the posterior sample in the $j-th$ iteration is

$$(\beta_{o1j}, \beta_{d1j}, \beta_{o2j}, \beta_{d2j}, \beta_{tj}, \beta_{pv,j}, \alpha_{oj}, \alpha_{tj}, \alpha_{ps,j}, \pi_{1j})'$$

and $\pi_{2j} = 1 - \pi_{1j}$.

6.1.2.2 Likelihood approach

For the likelihood base inference, the calculation of AICs for the conditional AR1 and 2-group mixture models with ID modelling are easy because ℓ can be obtained

from the estimation procedures. However, for the random intercept model with an ID modelling, such ℓ cannot be easily obtained from the estimation procedures because we only approximate the relative likelihood function $\dfrac{L(\boldsymbol{\theta})}{f(\boldsymbol{y})}$ instead of $L(\boldsymbol{\theta})$ directly using the Monte Carlo approximation (see sections 2.5 and 5.4.1). As a result, we resort to another method to approximate the log-likelihood function ℓ. Again, the 'observed data' log-likelihood function is given by (5.6) and its Monte Carlo approximation using the idea of SML method is

$$\begin{aligned}\ell_o &= \ln\left\{\frac{1}{M}\sum_{j=1}^{M}\left\{\prod_{i=1}^{I}\left[\prod_{t=2}^{n_i}\left(\frac{1}{1+e^{\widehat{\zeta}_{it}}}\right)\right]\left[\prod_{t=1}^{n_i}\left(\frac{e^{y_{it}\widehat{\eta}_{itj}}}{1+e^{\widehat{\eta}_{itj}}}\right)\right]\right.\right.\\ &\qquad\left.\left.\left[\left(\frac{e^{\widehat{\zeta}_{i,n_i+1,1}}}{1+e^{\widehat{\zeta}_{i,n_i+1,1}}}\right)\left(\frac{e^{\widehat{\eta}_{i,n_i+1,j}}}{1+e^{\widehat{\eta}_{i,n_i+1,j}}}\right)+\left(\frac{e^{\widehat{\zeta}_{i,n_i+1,0}}}{1+e^{\widehat{\zeta}_{i,n_i+1,0}}}\right)\left(\frac{1}{1+e^{\widehat{\eta}_{i,n_i+1,j}}}\right)\right]^{I(n_i<26)}\right\}\right\}\end{aligned}$$

where

$$\begin{aligned}\widehat{\eta}_{itj} &= u_j+\widehat{\beta}_o+\widehat{\beta}_d d_{it}+\widehat{\beta}_t\ln t+\widehat{\beta}_{pv}\ y_{i,t-1},\\ \widehat{\zeta}_{it} &= \widehat{\alpha}_o+\widehat{\alpha}_t\ln t+\widehat{\alpha}_{ps}\ y_{i,t-1},\\ \widehat{\zeta}_{i,n_i+1,h} &= \widehat{\alpha}_o+\widehat{\alpha}_t\ln t+\widehat{\alpha}_{ps}\ h,\ h=0,1,\end{aligned}$$

$(\widehat{\beta}_o,\widehat{\beta}_d,\widehat{\beta}_t,\widehat{\beta}_{pv},\widehat{\alpha}_o,\widehat{\alpha}_t,\widehat{\alpha}_{ps},\widehat{\sigma}^2)'$ are the maximum likelihood (ML) estimates and $\boldsymbol{u}_j=(u_{1j},\ldots,u_{Ij})$ are drawn from $N(0,\widehat{\sigma}^2)$. Then we use ℓ_o to approximate ℓ as required in the calculation of AIC in (6.1).

6.2 Interpretation

We report results of parameter estimates and AIC values for the three types of models using ML or Bayesian approaches with and without ID modelling, in Tables 2-4 in Appendix A. The maximum likelihood (ML) estimates and Bayesian estimates for the outcome and drop-out models *across the three types of models* with or without an ID modelling are qualitatively the same. The results of ML and Bayesian estimates are also comparable. Since the strength and direction of effects for covariates in both outcome and drop-out models are consistent across models, we will only describe the general results in the following sections.

6.2.1 On drug use

The significant dose and time effect indicate that a decrease in drug use is associated with an increase in methadone dosage and an increase in duration of treatment. The strong and positive association between the previous and present outcomes suggests that patients who use drug in their previous occasion are more likely to take drugs in the following occasion. This implies some patients tend to use drug continuously (called heavy drug-users) while others (called light drug-users) do not. This again justifies why we propose a mixture model to account for the group effects among patients.

6.2.2 Comparison between models with ID and models without ID

Comparing the results of the three types of models with and without an ID modeling, it is found that the parameter estimates in the outcome models are quantitatively similar. In spite of these consistent results, the incorporation of an ID model do give us a better understanding of the effect of drop-out process on the treatment outcome. In the drop-out models, the positive and significant time effect tells us that drug users staying longer in the methadone maintenance treatment (MMT) program are more likely to drop out from the program. The significant present outcome effect indicates that patients currently using heroin are more likely to drop out from the program. This also signifies the presence of informative drop-out in the data set and in turn suggests the suitability of modelling the data with an ID model. The low values of intercept in the drop-out models, moreover, indicate that the probability of drop-out for patients is low in general. This finding can be evidenced by the fact that only 51 (38%) patients drop out of the MMT program before 26 weeks.

On the other hand, while most of the parameter estimates in the outcome models with or without an ID modeling are quantitatively similar, the time effect in models with an ID modeling is smaller than those without an ID modelling

by 17% in magnitude or 7% in odds on the average. This finding eliminates our worry that the time effect of the treatment may be primarily due to the drop-out of heavy drug users which in turn leads to a false impression that if patients stay longer in the treatment, they will reduce drug use. Now, we still find an significiant time effect although it is weaker after accounting for the drop-out process of the data. This suggests patients staying longer in the program are more likely to reduce drug use and hence serves as a support to the policy of encouraging patients to stay longer in the MMT program.

6.2.3 Identification of patients

The significance of the variance of the random intercepts σ^2 suggests the presence of patient heterogeneity although its values differ in magnitude between the ML and Bayesian estimates in model with an ID modelling. To gain more knowledge on this, we classify the patients into 2 groups based on the Gibbs output of the 2-group mixture model with an ID modelling. In this classification, if the mean of the posterior group indicators of a patient is greater than 0.5, he or she will be classified into group 1, otherwise, group 2. There are 92 (68%) patients in group 1 returning only 89 positive heroin screens out of a total of 1951screens. Since the dose effect is significant in this group, patients are thus mainly light to medium (here called light) heroin users who respond to treatment in a dose-dependent

fashion with reduced heroin use at high methadone dose. Patients in group 2 are heavy users returning 380 positive heroin screens out of 921 screens. This group has insignificant dose effect indicating that patients respond poor to the treatment, with continuous heroin use regardless of the methadone dose received. This classification can indeed help doctors know more about the patients' habits of drug taking, so that they can modify the MMT program to suit individual needs.

On the other hand, if we classify patients into light heroin user group when the mean of his/her posterior random intecepts in the Gibbs output for the random intercept model with an ID modelling is less than or equal to 0.523, we will obtain the same classification except patient 28. Adopting the former classification using the 2-group mixture model, the average of the random intercepts for those patients in the heavy user group is 0.4445 whereas it is -0.4488 for the light user group. Results of the classification are summarised in Table 5 and group indicators and random intercepts of all patients are listed in Table 6 in Appendix A .

CHAPTER 7

SIMULATION STUDY

A simulation study is carried out to study the sensitivity or robustness of the parameters in the outcome model when data actually contains an ID process are fitted to the conditional AR1 model, 2-group mixture model and random intercept model with or without an ID modelling .

7.1 Procedure of simulation

We simulate a total of 60 data sets, each consisting of I= 300 heroin users from each of the three types of models with an ID modelling and set the maximum number n_i of outcomes per patient to be 10 or 20. The total number of outcomes without drop-out should be $N = 300 \times 10 = 3,000$ and 6000 for the maximum $n_i =$ 10 and 20 respectively. That means, we simulate 30 data sets when the maximum $n_i = 10$ and another 30 data sets when $n_i = 20$. We adjust the values of the true parameters for each type of models in order to achieve a desirable percentage of missing outcomes as well as convergent parameter estimates. Besides, we also set a higher percentage of missing outcomes and drop-out in order to obtain a more detectable result.

True parameter estimates of the conditional AR1 model with an ID modelling are set to be $\boldsymbol{\beta} = (-1.5, -0.5, -0.5, 3.5)$ for the outcome model and $\boldsymbol{\alpha} = (-2.0, 0.1, 0.5)$ for the drop-out model when the maximum $n_i = 10$. About 45% of outcomes out of totally $N = 3000$ possible outcomes are missing and 77% of patients drop out when the maximum $n_i = 10$. When the maximum $n_i = 20$, about 60% of outcomes out of totally $N = 6000$ possible outcomes are missing and 95% of patients drop out when the maximum $n_i = 20$.

For the 2-group mixture model with an ID modelling, the true parameter estimates are set to be $(\beta_{ko}, \beta_{kd}) = (-2.5, -0.5)$ and (-.5,0.008) respectively for group 1 and 2 and $(\beta_t, \beta_p) = (-0.37, 2.5)$ for the outcome model and $\boldsymbol{\alpha} = (-5.2, 0.7, 2.2)$ for the drop-out model. About 60% of outcomes are missing and 32% of patients drop out when the maximum $n_i = 10$. When the maximum $n_i = 20$, about 33% of outcomes are missing and 62% of patients drop out.

For the random intercept model with ID, the true parameter estimates are $\boldsymbol{\beta} = (-2.3, -0.83, -0.16, 3.6)$ for the outcome model, $\boldsymbol{\alpha} = (-4.4, 0.6687, 1.987)$ for the drop-out model and $\sigma^2 = 0.05$. There are about 20% of missing outcomes and 50% of drop-out patients when the maximum n_i is 10 while about 46% of missing outcomes and 87% of drop-out patients when the maximum n_i is 20. It is obvious that the percentages of missing outcomes and drop-outs of all models

when n_i is 20 are more than those when n_i is 10 because patients are more likely to drop-out as they stay longer in treatment.

The stimulated data with an ID process are then fitted into models with or without incorporating the drop-out model using WinBUGS package. The first 1000 observations of the Gibbs output of all models are treated as the burn-in period. The resulting sample size of the conditional AR1 model is about 1000 while the sample sizes of the 2-group mixture and random intercept models are both around 500. The sample sizes for the latter two models are smaller because we set a larger sampling intervals between iterations from their Gibbs output in order to reduce their higher autocorrelations between iterations.

7.2 Results of simulation

Results of simulation study for the three models are summarised in Tables 7-9 in Appendix A. For each type of model with and without an ID modelling, the mean of 30 Bayesian estimates, the average of 30 standard errors (ASE) and the mean squared error (MSE) are reported for each parameter.

7.3 Interpretation

Regarding the conditional AR1 model, all the parameter estimates in the outcome model, $\boldsymbol{\beta}$ are significant for model with or without an ID modelling. When n_i is 10, three parameters namely β_o, β_d and β_t show an improvement in MSE (smaller MSE) in model with an ID modelling when compared with model without an ID modelling. When $n_i = 20$, two parameters namely β_o and β_t show similar improvement in MSE. Such improvement implies that when data with an ID process are fitted to a model without an ID modelling, the bias will be more serious. Moreover, the improvement in MSE for β_o and β_t in model with an ID modelling is greater for $n_i = 20$ than $n_i = 10$. This indicates that for data with a higher percentage of missing, it is more necessary to fit it into a model with an ID modelling.

From the mixture and random intercept models, there are more significant parameters ($\beta_{o1}, \beta_{d1}, \beta_t$ & β_{pv} in the mixture models and $\beta_o, \beta_d, \beta_t$ & β_{pv} in the random intercept models) when $n_i = 20$ than when $n_i = 10$. Although some parameters of $\boldsymbol{\beta}$ show improvement in MSE for models with an ID modelling, we cannot observe a general trend of effect. Moreover there is also no general trend of greater improvement for data with a higher percentage of missing.

The lack of general trend of effects for the mixture and random intercept

models may be due to the complexity of modelling and the small number of simulated data set. But, due to the time constraint, the number of simulated data set cannot be further increased. On the other hand, this simulation study considers data with an ID process only and fits these data into models with or without an ID modelling. Alternatively, one may repeat the study, but using data with no ID process to evaluate the sensitivity or robustness of the parameters in the outcome model.

CHAPTER 8

DISCUSSION

After comparing the parameter estimates for the three types of models with or without informative drop-out (ID) modeling, we found that they all reveal similar information regarding the significant treatment factors that will reduce the drug use of patients in the methadone clinic data. However since there are substantial evidence that ID is presence, the incorporation of an ID model with the outcome model is necessary. Models with ID modelling allow for the selective attitude towards drop-out for the heavy drug users and avoid the illusion of 'reduced drug use over time' as they gradually drop out. Moreover, it also helps us to identify the relationship between the drop-out process and the drug taking habit of patients.

If we compare the three types of ID models namely the conditional AR1 model, the mixture model and the random intercept model according to the extent of information they carried, it seems that the mixture model as well as the random intercept model are more preferable because they allow for the population clustering and heterogeneity even though they convert similar informations on drug taking habits and drop out behaviors of patients. Both the mixture model and the

random intercept model help us to classify patients into heavy and light drup-user groups, members of which respond differently to methadone maintenance treatment (MMT). Only patients in the light drug-user group respond to MMT in a dose dependent way. Such identification of patients reveals valuable informations to doctors and help them to explore more effective and patient specific procedures in reducing the drug use of patients under MMT.

For the simulation study, we found that generally the bias in parameter estimates is more serious for models without ID modeling when the data sets actually contain an ID process. This suggests the necessarity of incorporating an ID model when the data contains an ID process. However, there is no significant reduction of bias for the other two types of ID models, namely the mixture and random intercept models. This may be because the studied models are complicate in nature with large number of parameters. Besides, although it is easy to execute WinBUGS for fitting models, it is also time-consuming to run it. As a result, we cannot simulate a large number of data sets for model fitting due to time constraint. This may also be a possible reason for obtaining the unfavourable results in simulation. For future improvement, we recommend writing a fortran program to execute the model fitting rather than relying on WinBUGS entirely.

For future research direction, we suggest fitting the models to data with non-

monotone missing so as to demonstrate their power of allowing for different types of missing pattern. On the other hand, further extension of the random intercept model with ID can be done by adopting a more general distribution for the random effects such as those heavy-tail distributions including Student-t and other leptokurtic and platykurtic alternatives. This will greatly increase the applicability of the models to data with different types of random effects.

Moreover, the methadone clinic data, in fact, contains two other drug uses namely the benzodiazepine and amphetamines. Chan *et al.* (1997) modelled simultaneously the logit of each type of drug use and their log odds ratio linearly in some covariates without ID as a bivariate conditional AR1 model. The conclusion is that while methadone maintenance is effective in reducing heroin use, it does not suppress non-opioid drug use. A challenging extension will be to incorporate an ID modelling to this multiple responses model to facilitate the study of multiple drug use while controlling their possible interaction as well as an ID process in the data.

REFERENCES

[1] Alfo, M. and Aitkin, M. (2000), "Random coefficient models for binary longitudinal responses with attrition", *Statistics and Computing*, **10**, No. 279-287.

[2] Baker, S. G. and Laird, N. M. (1988), "Regression analysis for Categorical Variables with Outcome Subject to Nonignorable Nonresponse", *Journal of the American Statistical Association*, **83**, No. 401, 62-69.

[3] Bernardo, J. M. and Smith, A. F. M. (1994), *Bayesian Statistics*, Wiley.

[4] Bonney, G. E. (1987), "Logistic regression for dependent binary observations", *Biometrics*, **43**, 951-973.

[5] Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models", *Journal of American Statistical Association*, **88**, 9-25.

[6] Brown, C. H. (1990), "Protecting Against Nonrandomly Missing Data in Longitudinal Studies", *Biometrics*, **46**, 143-155.

[7] Chan, J.S.K. (2000), " Initial stage problem in autoregressive binary regression", *The Statistican*, **49**, 495-502.

[8] Chan, J. S. K. and Kuk, A. Y. C. (1997), "Maximum Likelihood Estimation for Probit-Linear Mixed Models with Correlated Random Effects",

Biometrics, **53**, 86-97.

[9] Chan, J.S.K., Kuk, A.Y.C. and Bell, J. (1997), "A Likelihood approach to analysing longitudinal bivariate binary data", *Biometrical Journal*, **39**, No.4, 409-421.

[10] Chan, J.S.K., Kuk, A.Y.C., Bell, J. and McGilchrist, C. (1998), "The analysis of methadone clinic data using marginal and conditional logistic models with mixture or random effects", *The Australian & New Zealand Journal of Statistics*, **40**, No.1, 1–10.

[11] Chan, J.S.K. and Leung, D.Y.P. (2003), "Informative Drop-out Models for Longitudinal Binary Data using Likelihood and Bayesian Approaches", submitted for publication.

[12] Choy, B.S.T., Chan, J.S.K. and Yam, H.K. (2002), "Robust analysis on salamander data, Generalized Linear model with random effects", to be appeared in *Bayesian Statistics*, **7**.

[13] Clayton, D. G. (1996), "Generalized linear mixed models", in *Markov chain Monte Carlo in practice*, eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter, London: Chapman & Hall, 275-302.

[14] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1997), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal*

Statistical Society, B, **59**, 1-38.

[15] Diggle, P. and Kenward, M. G. (1994), "Informative drop-out in longitudinal data analysis", *Appl. Statist*, **43**, 49–93.

[16] Drum, M. L. and McCullagh, P. (1993), "REML estimation with exact covariance in the logistic mixed model", *Biometrics*, **49**, 677-689.

[17] Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G. and Ibrahim, J. G. (2001), "Bias in Estimating Association Paremeters for Longitudinal Binary Responses with Drop-Outs", *Biometrics*, **57**, 15-21.

[18] Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S, R. (1995), "Regression Models for Longitudinal Binary Responses with Informative Drop-outs", *J. R. Statist. Soc. B*, **57**, No. 4, 691-704.

[19] Gelfand, A. and Carlin, B. (1993), "Maximum-likelihood estimation for constrained- or missing data models", *Canadian Journal of Statistics*, **21**, 303-311.

[20] Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities", *Journal of the American Statistical Association*, **85**, 398-409.

[21] Gelman, A. and Rubin, D. B. (1992), "Inference on iterative simulation using multiple sequences", *Statistical Science*, **7**, no. 4, 457-511.

[22] Geman, S., and Geman, D. (1984), "Stochastic relaxation, Gibbs Distributions and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

[23] Geyer, C. J. (1994), "On the convergence of Monte Carlo maximumlikelihood calculations", *Journal of the Royal Statistical Society, B*, **56**, 261-274.

[24] Geyer, C. J. and Thompson, E. A. (1992), "Constrained Monte Carlo maximum likelihood for dependent data", *Journal of the Royal Statistical Society, B*, **54**, 657-699.

[25] Glynn, R. J., Laird, N. M. and Rubin, D. B. (1986), "Selection Modelling versus mixture modelling with nonignorable nonresponse. ", in *Drawing Inference from Self Selected Samples* (ed. H. Wainer), 115-142, New York: Springer.

[26] Green, P. J. (1990), "On use of EM algorithm for penalized likelihood estimation", *Journal of the Royal Statistical Society, B*, **52**, 443-452.

[27] Greenlees, J. S., Reece, W. S. and Zieschang, K. D. (1982), "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed", *Journal of the American Statistical Association*, **77**, Number 378, 251-261.

[28] Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov

Chains and their applications", *Biometrika*, **57**, 97-109.

[29] Kuk, A. Y. C. (1995), "Asymptotically unbiased estimation in generalized linear models with random effects", *Journal of Royal Statistical Society, B*, **57**, 395-407.

[30] Kuk, A. Y. C. (1999), "Laplace importance sampling for generalized linear mixed models", *Journal of Statistical Computing Simulation*, **63**, 143-158.

[31] Kuk, A. Y. C. and Cheng, Y. W. (1997), "The Monte Carlo Newton-Raphson Algorithm", *Journal of Statistical Computing Simulation*, **59**, 233-250.

[32] Kuk, A. Y. C. and Cheng, Y. W. (1999), "Pointwise and functional approximations in Monte Carlo maximum likelihood estimation", *Statistics and Computing*, **9**, 91-99.

[33] Kuk, A.Y.C., Chan, J.S.K. and Yam, H.K. (2003), "Monte Carlo Approximation through Gibbs output in Generalized linear mixed models", submitted for publication.

[34] Laird, N. M. (1988), "Missing data in longitudinal studies", *Statist. Med.*, **7**, 305-315.

[35] Liang, K. Y. and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models", *Biometrika*, **73**, 13-22.

[36] Little, R. J. A. (1982), "Models for Nonresponse in Sample Surveys", *Journal of the American Statistical Association*, **77**, Number 378, 237-250.

[37] Little, R. J. A. (1995), "Modeling the Drop-out Mechanism in Repeated-Measures Studies", *Journal of the American Statistical Association*, **90**, No. 431, 1112-1121.

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

[39] McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models", *Journal of the American Statistical Association*, **92**, Issue 437, 162-170.

[40] McCulloch, C. E. and Searle, S. R. (2001), *Generalized, linear, and mixed models*, New York: John Wiley & Sons, Inc.

[41] McGilchrist, C. A. (1994), "Estimation in generalized mixed models", *J. R. Statist. Soc. B*, **56**, 61-69.

[42] Meng, X. L. and Rubin, D. B. (1991), "Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm", *J. Am. Statist. Ass.*, **86**, 899-909.

[43] Meng, X. L. and Rubin, D. B. (1993), "Maximum likelihood estimation via the ECM algorithm: a general framework", *Biometrika*, **80**, 267-278.

[44] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N. and Teller, A. H. (1953), "Equations of State calculations by fast computing machines", *Journal of Chemical Physics*, **21**, 1087-1091.

[45] Molenberghs, G., Kenward, M. G. and Lesaffre, E. (1995), "Regression models for longitudinal binary responses with informative dropout ", *Journal of the Royal Statistical Society, B*, **57**, 691-704.

[46] Mori, M., Woolson, R. F. and Woodworth, G. G. (1994), "Slope Estimation in the Presence of Informative Right Censoring: Modeling the Number of Observations as a Geometric Random Variable", *Biometrics*, **50**, 39-50.

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

[48] Penttinen, A. (1984), "Modelling interaction in spatial point patterns: parameter estimation by the maximum likelihood methods", *Jy. Stud. Comput. Sci. Econ. Statist.,*, **7**, 1-105.

[49] Ripley, B. (1987), *Stochastic Simulation*, New York: John Wiley.

[50] Rotnitzky, A. and Robins, J. M. (1997), "Analysis of semi-parametric regression models with non-ignorable non-response", *Statistics in Medicine*, **16**, 81-102.

[51] Rotnitzky, A. and Robins, J. M. abd Scharfstein, D. O. (1998), "Semipara-

metric regression for repeated outcomes with nonignorable nonresponse", *J. Am. Statist. Ass.*, **93**, 1321-1339.

[52] Roy, J and Lin, Xihong (2002), "Analysis of Multivariate Longitudinal Outcomes With Nonignorable Dropouts and Missing Covariates: Changes in Methadone Treatment Practices", *Journal of the American Statistical Association*, **97**, No. 457, 40-52.

[53] Rubin, D. B. (1976), "Inference and missing data", *Biometrika*, **63**, 3, 581-592.

[54] Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986), *Akaike Information Criterion Statistics*, Reidel, Dordrecht.

[55] Schall, R. (1991), "Estimation in generalized linear models with random effects", *Biometrika*, **78**, 719-727.

[56] Schluchter, M. D. (1992), "Methods for the analysis of informatively censored longitudinal data", *Statistics in Medicine*, **11**, 1861-1870.

[57] Stiratelli, R., Laird, N. and Ware, J. H. (1984), "Random-effects Models for Serial Observations with Binary Response", *Biometrics*, **40**, 961-971.

[58] Troxel, A. B. (1998), "Analysis of longitudinal data with non-ignorable non-monotone missing values", *Appl. Statist*, **47**, 425-438.

[59] Waclawiw, M. A. and Liang, K. Y. (1993), "Prediction of random effects in

the generalized linear model", *J. Am. Statist. Ass.*, **88**, 171-178.

[60] Wei, L. J. and Stram, D. O. (1987), "Analysing repeated meaurements with possibly missing observations by modelling marginal distributions", *Statistics in Medicine*, **7**, 139-148.

[61] Wei G. C. G. and Tanner M. A. (1990), "A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms," *Journal of the American Statistical Association*, **85**, 699-704.

[62] Wolfinger, R. W. (1993), "Laplace's approximation for nonlinear mixed models", *Biometrika*, **80**, 791-795.

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

[64] Wu, M. C. and Bailey, K. R. (1989), "Estimation and Comparison of Changes in the Presence of Informative Right Censoring: Conditional Linear Model", *Biometrics*, **45**, 939-955.

[65] Zeger, S. L. and Karim, M. R. (1991), "Generalized linear models with random effects: A Gibbs sampling approach", *Journal of the American Statistical Association*, **86**, 79-86.

[66] Zhao, L. P. and Prentice, R. L. (1990), "Correlated binary regression using

a quadratic exponential model", *Biometrika*, **77**, 642-648.

APPENDIX A

TABLES AND PLOTS

Table 1. Basic information for the methadone clinic data set.

Number of observation (N)	2872
Number of patient (I)	136
Number of drop-out patient ($n_i < 26$) (ND)	51
Average number of treatment week per patient	21.1 (4 to 26)
Average dosage of methadone	64mg.
Percentage of positive heroin test	16.3%

Table 2. Parameter estimates and *s.e. (in italic)* for conditional AR1 models

Model	Method	Intercept	Dose	Time	Previous	Present	AIC
with ID β	MLE	-0.865 *0.220*	-0.00908 *0.00282*	-0.367 *0.067*	2.388 *0.120*		3016.30
α		-11.329 *0.957*		2.936 *0.311*		0.895 *0.412*	
with ID β	Bayes	-0.848 *0.212*	-0.00913 *0.002818*	-0.369 *0.062*	2.397 *0.122*		2594.06
α		-6.49 *0.978*		0.692 *0.243*		2.087 *0.864*	
without ID β	MLE	-0.842 *0.219*	-0.00884 *0.00282*	-0.405 *0.063*	2.396 *0.120*		2090.02
without ID β	Bayes	-0.834 *0.227*	-0.00894 *0.00293*	-0.409 *0.062*	2.398 *0.115*		2090.22

Table 3. Parameter estimates and *s.e. (in italic)* for Mixture models

Model	Method	Intercept	Dose	Time	Previous	Present	π	AIC
with ID β_1	Bayes	-1.315 *0.614*	-0.0147 *0.0082*	-0.373 *0.0729*	1.556 *0.1367*		0.647	2717.85
β_2		-0.233 *0.3109*	0.0013 *0.0043*	-0.373* *0.0729**	1.556* *0.1367**			
α		-6.539 *0.927*		0.704 *0.267*		2.148 *0.705*		
without ID β_1	MLE	-1.173 *0.517*	-0.0153 *0.0069*	-0.432 *0.0686*	1.561 *0.1371*		0.669	1965.53
β_2		-0.132 *0.310*	0.0010 *0.0042*	-0.432* *0.0686**	1.561* *0.1371**			
without ID β_1	Bayes	-1.323 *0.577*	-0.0139 *0.0077*	-0.424 *0.0687*	1.565 *0.1401*		0.659	2671.20
β_2		-0.1679 *0.2964*	0.0011 *0.0041*	-0.424* *0.0687**	1.565* *0.1401**			

*Set to be the same across groups.

Table 4. Parameter estimates and *s.e. (in italic)* for Random intercepts models

Model	Method	Intercept	Dose	Time	Previous	Present	Sigma	AIC
with ID β	MLE	-0.671 *0.033*	-0.0144 *0.000001*	-0.265 *0.001*	1.297 *0.003*		0.500 N.A.	2769.99
α		-8.768 *0.655*		1.341 *0.204*		4.235 *0.315*		
with ID β	Bayes	-0.640 *0.386*	-0.0164 *0.006*	-0.351 *0.077*	1.410 *0.143*		1.839 *0.439*	2711.20
α		-6.459 *0.986*		0.695 *0.265*		2.039 *0.763*		
without ID β	MLE	-0.507 *0.177*	-0.0169 *0.000072*	-0.386 *0.053*	1.357 *0.503*		1.837 N.A.	2242.72
without ID β	Bayes	-0.643 *0.404*	-0.0156 *0.006*	-0.421 *0.072*	1.43 *0.140*		1.837 *0.414*	2189.88

Table 5. Classification of patients in MMT

	Heroin use	Heavy Gp. 2	Light Gp. 1	Total
Dropout	Number of patients	28	23	51
($n_i < 26$)	Average dose	63.5	67.6	65.3
	Weeks in treatment	13.4	12.5	13.0
	% of positive test	42.4	4.2	25.8
	average random intercepts	0.255	-0.09	0.161
Not dropout	Number of patients	21	64	85
($n_i = 26$)	Average dose	57.9	66.1	64.1
	Weeks in treatment	26	26	26
	% of positive test	38.6	4.6	13.0
	average random intercepts	0.190	-0.356	-0.166
Total	Number of patients	49	87	136
	Average dose	60.2	66.3	64.4
	Weeks in treatment	18.8	22.4	21.1
	% of positive test	40.2	4.6	16.0
	% of drop-out	57.1	26.4	37.5
	average random intercepts	0.445	-0.449	-0.04

Table 6. Information of each patient

Patient	Average dose	No. of obs.	No. of +ve test	Group indicator	Random intercept
1	76.5	26	7	0.036	1.053
2	50	26	1	0.996	-0.8836
3	48	10	3	0.306	0.5778
4	48.5	10	0	0.97	-0.991
5	80	26	2	0.982	-0.07834
6	40	15	8	0.004	1.245
7	51.7	6	1	0.486	0.3312
8	60	4	4	0.002	2.118
9	43.7	26	5	0.624	0.2582
10	53.8	26	18	0	2.072
11	66.2	26	3	0.96	0.02427
12	51.9	26	0	1	-1.493
13	40	26	0	0.994	-1.576
14	82.5	14	10	0	2.299
15	79.2	26	0	1	-1.32
16	80	4	0	0.81	-0.3786
17	105.7	21	1	0.976	-0.1123
18	40	5	0	0.838	-0.5215
19	64.1	11	0	0.97	-0.9212
20	61.3	26	2	0.982	-0.3101
21	58.1	15	7	0.002	1.416
22	40	9	4	0.118	0.9635
23	65	22	10	0	1.486
24	55.9	11	4	0.164	0.8974
25	61.3	8	2	0.394	0.5104

Patient	Average dose	No. of obs.	No. of +ve test	Group indicator	Random intercept
26	79.2	26	0	1	-1.28
27	56.9	26	0	1	-1.492
28	30.6	26	6	0.206	0.369
29	62.4	26	0	1	-1.432
30	48.3	6	3	0.166	1.073
31	46.6	26	1	0.996	-0.905
32	29.4	9	4	0.156	0.799
33	45	12	1	0.904	-0.386
34	59	26	1	0.996	-0.765
35	40	26	0	0.998	-1.597
36	27	6	2	0.534	0.343
37	77.7	26	9	0	1.334
38	57.8	21	1	0.98	-0.594
39	40	26	4	0.832	-0.069
40	41.2	26	10	0.018	0.95
41	79.9	25	0	0.998	-1.131
42	69.8	12	1	0.842	-0.009
43	69.2	26	2	0.99	-0.221
44	64.8	26	0	1	-1.378
45	59.3	15	8	0	1.517
46	62.2	26	5	0.45	0.521
47	63.5	26	14	0	1.748
48	60.8	26	15	0	1.865
49	59.9	14	7	0.006	1.444
50	96.5	24	8	0	1.64

Patient	Average dose	No. of obs.	No. of +ve test	Group indicator	Random intercept
51	79.2	25	5	0.18	0.854
52	23.5	13	0	0.962	-1.291
53	62.8	26	0	1	-1.391
54	69.6	26	0	0.998	-1.294
55	44	26	0	1	-1.532
56	54.4	26	1	0.992	-0.767
57	81.5	26	2	0.984	-0.026
58	37.7	26	8	0.104	0.701
59	72.5	26	0	1	-1.317
60	51	26	0	0.998	-1.585
61	56.5	10	3	0.174	0.83
62	43.1	26	11	0.004	1.101
63	120	12	1	0.874	0.32
64	100	18	0	0.996	-0.858
65	53.1	26	2	0.988	-0.383
66	93.5	26	11	0	1.82
67	60	10	0	0.958	-0.874
68	45.5	26	0	1	-1.548
69	100	26	1	1	-0.314
70	69.6	26	0	1	-1.342
71	70.4	12	6	0.004	1.475
72	77.7	26	5	0.388	0.742
73	80	26	1	0.998	-0.531
74	67.3	26	0	1	-1.361
75	55.8	26	7	0.16	0.68

Patient	Average dose	No. of obs.	No. of +ve test	Group indicator	Random intercept
76	83.1	26	3	0.91	0.244
77	78.5	26	1	0.998	-0.611
78	100	26	1	0.996	-0.313
79	50	26	3	0.94	-0.116
80	120	7	3	0.014	1.8
81	53.1	13	5	0.066	1.052
82	30	26	5	0.72	-0.003
83	51.9	26	5	0.596	0.358
84	70	5	0	0.846	-0.574
85	96.9	26	1	1	-0.402
86	40	6	0	0.88	-0.735
87	36.2	17	4	0.462	0.391
88	73.7	26	2	0.984	-0.128
89	43	15	8	0.016	1.308
90	75	26	1	1	-0.531
91	65	26	0	1	-1.402
92	100	17	11	0	2.431
93	61.3	26	0	1	-1.428
94	38.3	26	12	0.006	1.162
95	40	26	0	1	-1.569
96	58.5	26	1	0.996	-0.803
97	59.8	26	0	1	-1.454
98	81.3	26	1	0.988	-0.504
99	70	26	0	1	-1.364
100	80.2	26	1	0.994	-0.649

Patient	Average dose	No. of obs.	No. of +ve test	Group indicator	Random intercept
101	50.8	26	2	0.968	-0.396
102	90	8	0	0.938	-0.554
103	66.5	26	1	0.998	-0.77
104	30.8	26	9	0.12	0.626
105	50.4	10	5	0.024	1.252
106	77.7	26	2	0.976	-0.087
107	76.7	26	0	0.998	-1.24
108	120	4	0	0.782	-0.183
109	74.4	26	5	0.398	0.676
110	75.8	26	1	0.998	-0.605
111	40	10	3	0.384	0.523
112	99	26	0	1	-1.135
113	25.1	18	3	0.706	-0.095
114	50.4	26	6	0.334	0.518
115	53.5	13	2	0.714	0.081
116	81	26	3	0.9	0.321
117	65.6	18	0	0.998	-1.096
118	60	26	3	0.948	-0.002
119	90	26	0	1	-1.174
120	83.7	26	2	0.974	-0.007
121	61.2	26	13	0	1.649
122	38.3	26	3	0.93	-0.293
123	42.3	26	17	0	1.872
124	89.6	26	21	0	3.005
125	70.6	16	10	0	2.002

Patient	Average dose	No. of obs.	No. of +ve test	Group indicator	Random intercept
126	48.1	26	0	1	-1.539
127	66.9	26	1	0.998	-0.667
128	68.8	26	0	1	-1.309
129	90	24	0	0.998	-1.02
130	90	26	2	0.966	0.024
131	80	17	9	0	1.847
132	28.2	11	0	0.964	-1.105
133	60	26	4	0.796	0.213
134	58.3	24	4	0.404	0.534
135	99.6	26	0	0.998	-1.096
136	54.6	26	12	0	1.378

Table 7. Parameter estimates, true parameter values, *ASE* and *MSE* for parameters in the conditional AR1 data model with and without ID

	With ID model				Without ID model			
	β_0	β_d	β_t	β_{pv}	β_0	β_d	β_t	β_{pv}
	Max. $n_i = 10$							
True	-1.5	-0.5	-0.5	3.5	-1.5	-0.5	-0.5	3.5
Est.	*-1.496*	*-0.528*	*-0.526*	*3.572*	*-1.509*	*-0.529*	*-0.577*	*3.569*
ASE	*0.156*	*0.166*	*0.119*	*0.196*	*0.158*	*0.165*	*0.119*	*0.198*
MSE	5.883E-07*	2.787E-05*	2.321E-05*	1.811E-04	2.856E-06	2.979E-05	2.068E-04	1.659E-04
	Max. $n_i = 20$							
True	-1.5	-0.5	-0.5	3.5	-1.5	-0.5	-0.5	3.5
Est.	*-1.496*	*-0.545*	*-0.512*	*3.513*	*-1.524*	*-0.543*	*-0.545*	*3.510*
ASE	*0.150*	*0.157*	*0.099*	*0.181*	*0.151*	*0.158*	*0.097*	*0.183*
MSE	5.797E-07*†	6.940E-05	5.007E-06*†	5.562E-06	1.970E-05	6.458E-05	6.964E-05	3.778E-06

Note: ASE: average standard error.
MSE: mean squared error.
Parameter estimates in italic are significant.
* improvement of MSE in model with ID compared with model without ID.
† greater improvement of MSE in model with ID for data with higher % of missing.

Table 8. Parameter estimates, true parameter values, *ASE* and *MSE* for parameters in the 2-group mixture data model with and without ID

	With ID model						Without ID model					
	β_{01}	β_{d1}	β_{02}	β_{d2}	β_t	β_{pv}	β_{01}	β_{d1}	β_{02}	β_{d2}	β_t	β_{pv}
	Max. $n_i = 10$											
True	-2.5	-0.5	-0.5	0.008	-0.37	2.5	-2.5	-0.5	-0.5	0.008	-0.37	2.5
Est.	*-2.521*	-0.407	-0.509	0.029	*-0.357*	*2.440*	*-2.460*	-0.320	-0.502	-0.045	*-0.471*	*2.340*
ASE	*0.287*	*0.268*	*0.352*	*0.259*	*0.104*	*0.189*	*0.290*	*0.274*	*0.339*	*0.265*	*0.106*	*0.203*
MSE	1.51E-05*	2.96E-04	3.09E-06	1.55E-05	6.08E-06*	1.23E-04*†	5.50E-05	1.12E-03	1.95E-07	9.57E-05	3.54E-04	8.81E-04
	Max. $n_i = 20$											
True	-2.5	-0.5	-0.5	0.008	-0.37	2.5	-2.5	-0.5	-0.5	0.008	-0.37	2.5
Est.	*-2.598*	*-0.537*	-0.557	0.026	*-0.359*	*2.511*	*-2.443*	*-0.405*	-0.526	0.006	*-0.505*	*2.514*
ASE	*0.237*	*0.194*	*0.306*	*0.193*	*0.075*	*0.144*	*0.236*	*0.205*	*0.297*	*0.204*	*0.072*	*0.149*
MSE	3.30E-04	4.60E-05*	1.13E-04	1.07E-05	4.377E-06*†	4.05E-06*	1.12E-04	3.09E-04	2.30E-05	9.91E-08	6.24E-04	7.15E-06

Note: ASE: average standard error.
MSE: mean squared error.
Parameter estimates in italic are significant.
* improvement of MSE in model with ID compared with model without ID.
† greater improvement of MSE in model with ID for data with higher % of missing.

Table 9. Parameter estimates, true parameter values, *ASE* and *MSE* for parameters in the random intercept data model with and without ID

	With ID model					Without ID model				
	β_0	β_d	β_t	β_{pv}	σ^2	β_0	β_d	β_t	β_{pv}	σ^2
	Max. $n_i = 10$									
True	-2.3	-0.83	-0.16	3.6	0.05	-2.3	-0.83	-0.16	3.6	0.05
Est.	*-2.359*	*-0.855*	0.193	*3.582*	0.038	*-2.360*	*-0.851*	0.055	*3.587*	0.049
ASE	*0.195*	*0.170*	*0.126*	*0.201*	*0.059*	*0.195*	*0.172*	*0.120*	*0.202*	*0.081*
MSE	1.188E-04*	2.103E-05	4.305E-03	1.176E-05	5.315E-06	1.234E-04	1.498E-05	1.590E-03	5.798E-06	1.156E-08
	Max. $n_i = 20$									
True	-2.3	-0.83	-0.16	3.6	0.05	-2.3	-0.83	-0.16	3.6	0.05
Est.	*-2.370*	*-0.891*	*0.223*	*3.573*	0.024	*-2.363*	*-0.897*	0.074	*3.580*	0.015
ASE	*0.175*	*0.146*	*0.094*	*0.165*	*0.036*	*0.173*	*0.150*	*0.090*	*0.168*	*0.025*
MSE	1.667E-04	1.303E-04*	5.051E-03	2.489E-05	2.344E-05*	1.348E-04	1.558E-04	1.882E-03	1.449E-05	4.160E-05

Note: ASE: average standard error.
MSE: mean squared error.
Parameter estimates in italic are significant.
* improvement of MSE in model with ID compared with model without ID.

APPENDIX B

FIRST AND SECOND ORDER DERIVATIVES

1. For the conditional AR1 model with ID, the first order derivatives are:

$$\begin{aligned}
\ell'_{\beta_k} &= \sum_{i=1}^{I}(p'_{\beta_k,y,i} + p_{yr,i}^{-1}\ I_i\ p'_{\beta_k,yr,i}) \\
\ell'_{\alpha_k} &= \sum_{i=1}^{I}(p'_{\alpha_k,r,i} + p_{yr,i}^{-1}\ I_i\ p'_{\alpha_k,yr,i})
\end{aligned}$$

and the second order derivatives are:

$$\begin{aligned}
\ell''_{\beta_{k_1}\beta_{k_2}} &= \sum_{i=1}^{I}[p''_{\beta_{k_1},\beta_{k_2},y,i} + {p_{yr,i}}^{-2}\ I_i(p''_{\beta_{k_1},\beta_{k_2},yr,i}\ p_{yr,i} - p'_{\beta_{k_1},yr,i}\ p'_{\beta_{k_2},yr,i})] \\
\ell''_{\alpha_{k_1}\alpha_{k_2}} &= \sum_{i=1}^{I}[p''_{\alpha_{k_1},\alpha_{k_2},r,i} + {p_{yr,i}}^{-2}\ I_i(p''_{\alpha_{k_1},\alpha_{k_2},yr,i}\ p_{yr,i} - p'_{\alpha_{k_1},yr,i}\ p'_{\alpha_{k_2},yr,i})] \\
\ell''_{\beta_{k_1}\alpha_{k_2}} &= \sum_{i=1}^{I} {p_{yr,i}}^{-2}\ I_i(p''_{\beta_{k_1},\alpha_{k_2},yr,i}\ p_{yr,i} - p'_{\beta_{k_1},r,i}\ p'_{\alpha_{k_2},yr,i})
\end{aligned}$$

where for $i = 1, \ldots I$ ($I = 136$); $t = 1, \ldots, n_i$ and $k, k_1, k_2 = 1, \ldots, p$ ($p = 4$ for the outcome model) or $k, k_1, k_2 = 1, \ldots, q$ ($q = 3$ for the drop-out model);

the design matrix for the outcome model is $(\boldsymbol{x}'_{11}, \boldsymbol{x}'_{12}, \ldots, \boldsymbol{x}'_{I,n_I})'$, $\boldsymbol{x}_{it} = (1, d_{it}, \ln t, y_{i,t-1})$;

the design matrix for the dropout model is $(\boldsymbol{z}'_{11}, \boldsymbol{z}'_{12}, \ldots, \boldsymbol{z}'_{I,n_I})'$, $\boldsymbol{z}_{it} = (1, \ln t, y_{it})$;

$I_i = I(n_i < 26)$,

$z_{i,n_i+1,k1} = 1$, $z_{i,n_i+1,k0} = 0$ for $k = 3$ and $z_{i,n_i+1,j1} = z_{i,n_i+1,j0} = z_{i,n_i,j}$ otherwise,

$$\mu_{it} = \frac{e^{\eta_{it}}}{1+e^{\eta_{it}}}, \mu'_{it} = \mu_{it}(1-\mu_{it}) = \frac{e^{\eta_{it}}}{(1+e^{\eta_{it}})^2}, \mu''_{it} = \mu_{it}(1-\mu_{it})(1-2\mu_{it}) = \frac{e^{\eta_{it}}(1-e^{\eta_{it}})}{(1+e^{\eta_{it}})^3};$$

$$\nu_{it} = \frac{e^{\zeta_{it}}}{1+e^{\zeta_{it}}}, \nu'_{it} = \nu_{it}(1-\nu_{it}) = \frac{e^{\zeta_{it}}}{(1+e^{\zeta_{it}})^2}, \nu''_{it} = \nu_{it}(1-\nu_{it})(1-2\nu_{it}) = \frac{e^{\zeta_{it}}(1-e^{\zeta_{it}})}{(1+e^{\zeta_{it}})^3};$$

$\nu_{i,n_i+1,h},\ \nu'_{i,n_i+1,h}$ & $\nu''_{i,n_i+1,h}$ are similarly defined based on $\zeta_{i,n_i+1,1}$ & $\zeta_{i,n_i+1,0},\ h = 0, 1;$

$$p'_{\beta_k,y,i} = \sum_{t=1}^{n_i} x_{itk}(y_{it} - \mu_{it});\ p''_{\beta_{k_1},\beta_{k_2},y,i} = -\sum_{t=1}^{n_i} x_{itk_1}\ x_{itk_2}\ \mu'_{it};$$

$$p'_{\alpha_k,r,i} = -\sum_{t=2}^{n_i} z_{itk}\ \nu_{it};\ p''_{\alpha_{k_1},\alpha_{k_2},r,i} = -\sum_{t=2}^{n_i} z_{itk_1}\ z_{itk_2}\ \nu'_{it};$$

$$p_{yr,i} = \mu_{i,n_i+1}\nu_{i,n_i+1,1} + (1-\mu_{i,n_i+1})\nu_{i,n_i+1,0};$$

$$p'_{\beta_k,yr,i} = x_{i,n_i+1,k}\ \mu'_{i,n_i+1}(\nu_{i,n_i+1,1} - \nu_{i,n_i+1,0});$$

$$p'_{\alpha_k,yr,i} = \mu_{i,n_i+1}\ z_{i,n_i+1,k1}\ \nu'_{i,n_i+1,1} + (1-\mu_{i,n_i+1})\ z_{i,n_i+1,k0}\ \nu'_{i,n_i+1,0};$$

$$p''_{\beta_{k_1},\beta_{k_2},yr,i} = x_{i,n_i+1,k_1}\ x_{i,n_i+1,k_2}\ \mu''_{i,n_i+1}\ (\nu_{i,n_i+1,1} - \nu_{i,n_i+1,0});$$

$$p''_{\alpha_{k_1},\alpha_{k_2},yr,i} = \mu_{i,n_i+1}\ z_{i,n_i+1,k_1,1}\ z_{i,n_i+1,k_2,1}\ \nu''_{i,n_i+1,1} +$$

$$(1-\mu_{i,n_i+1})\ z_{i,n_i+1,k_1,0}\ z_{i,n_i+1,k_2,0}\ \nu''_{i,n_i+1,0};$$

$$p''_{\beta_{k_1},\alpha_{k_2},yr,i} = x_{i,n_i+1,k_1}\mu'_{i,n_i+1}\ (z_{i,n_i+1,k_2,1}\ \nu'_{i,n_i+1,1} - z_{i,n_i+1,k_2,0}\ \nu'_{i,n_i+1,0}).$$

2. For the mixture model with ID, we adopt a $m = 2$ group mixture with a group specific intercept and dose coefficients while the time in treatment and previous outcome effects are fixed across groups. The first order derivatives are:

$$\begin{aligned}
\ell'_{\beta_{m_1,k}} &= \sum_{i=1}^{I} {p_{y,i}}^{-1}\left(\sum_{m=1}^{2} f_{m_1,k,m}\ \pi_m\ p_{y,im}\ p'_{\beta_k,y,im}\right) + \sum_{i=1}^{I} I_i\ {p_{yr,i}}^{-1}\left(\sum_{m=1}^{2} f_{m_1,k,m}\ \pi_m\ p'_{\beta_k,yr,im}\right) \\
\ell'_{\alpha_k} &= \sum_{i=1}^{I} p'_{\alpha_k,r,i} + \sum_{i=1}^{I} I_i\ {p_{yr,i}}^{-1}\left(\sum_{m=1}^{2} \pi_m\ \ p'_{\alpha_k,yr,im}\right)
\end{aligned}$$

$$\ell'_{\pi_1} = \sum_{i=1}^{I} {p_{y,i}}^{-1} \left(\sum_{m=1}^{2} h_m \; p_{y,im} \right) + \sum_{i=1}^{I} I_i \; {p_{yr,i}}^{-1} \left(\sum_{m=1}^{2} h_m \; p_{yr,i} \right)$$

and the second order derivatives are:

$$\begin{aligned}
\ell''_{\beta_{m_1,k_1}\beta_{m_2,k_2}} &= \sum_{i=1}^{I} \left\{ {p_{y,i}}^{-2} \left[p_{y,i} \left(\sum_{m=1}^{2} f_{m_1,k_1,m} f_{m_2,k_2,m} \pi_m p_{y,im} (p''_{\beta_{k_1},\beta_{k_2},y,im} + p'_{\beta_{k_1},y,im} \; p'_{\beta_{k_2},y,im}) \right) \right.\right. \\
&\quad \left.\left. - \left(\sum_{m=1}^{2} f_{m_1,k_1,m} \; \pi_m \; p_{y,im} \; p'_{\beta_{k_1},y,im} \right) \left(\sum_{m=1}^{2} f_{m_2,k_2,m} \; \pi_m \; p_{y,im} \; p'_{\beta_{k_2},y,im} \right) \right] \right\} \\
&\quad + \sum_{i=1}^{I} \left\{ I_i \; {p_{yr,i}}^{-2} \left[p_{yr,i} \left(\sum_{m=1}^{2} f_{m_1,k_1,m} \; f_{m_2,k_2,m} \; \pi_m \; p''_{\beta_{k_1},\beta_{k_2},yr,im} \right) - \right. \right. \\
&\quad \left. \left. \left(\sum_{m=1}^{2} f_{m_1,k_1,m} \; \pi_m \; p'_{\beta_{k_1},yr,im} \right) \left(\sum_{m=1}^{2} f_{m_2,k_2,m} \; \pi_m \; p'_{\beta_{k_2},yr,im} \right) \right] \right\} \\
\ell''_{\alpha_{k_1}\alpha_{k_2}} &= \sum_{i=1}^{I} p''_{\alpha_{k_1},\alpha_{k_2},r,i} + \sum_{i=1}^{I} \left\{ I_i \; {p_{yr,i}}^{-2} \right. \\
&\quad \left. \left[p_{yr,i} \left(\sum_{m=1}^{2} \pi_m \; p''_{\alpha_{k_1},\alpha_{k_2},yr,im} \right) - \left(\sum_{m=1}^{2} \pi_m \; p'_{\alpha_{k_1},yr,im} \right) \left(\sum_{m=1}^{2} \pi_m \; p'_{\alpha_{k_2},yr,im} \right) \right] \right\} \\
\ell''_{\beta_{m_1,k_1}\alpha_{k_2}} &= \sum_{i=1}^{I} \left\{ I_i \; {p_{yr,i}}^{-2} \left[p_{yr,i} \left(\sum_{m=1}^{2} f_{m_1,k_1,m} \; \pi_m \; p''_{\beta_{k_1}\alpha_{k_2},yr,im} \right) - \right. \right. \\
&\quad \left. \left. \left(\sum_{m=1}^{2} f_{m_1,k_1,m} \; \pi_m \; p'_{\beta_{k_1},yr,im} \right) \left(\sum_{m=1}^{2} \pi_m \; p'_{\alpha_{k_2},yr,im} \right) \right] \right\} \\
\ell''_{\beta_{m_1,k}\pi_1} &= \sum_{i=1}^{I} \left\{ {p_{y,i}}^{-2} \left[p_{y,i} \left(\sum_{m=1}^{2} f_{m_1,k,m} \; h_m \; p_{y,im} \; p'_{\beta_k,y,im} \right) - \left(\sum_{m=1}^{2} f_{m_1,k_1,m} \; \pi_m \; p_{y,im} \; p'_{\beta_k,y,im} \right) \right. \right. \\
&\quad \left. \left. \left(\sum_{m=1}^{2} h_m \; p_{y,im} \right) \right] \right\} + \sum_{i=1}^{I} \left\{ I_i \; {p_{yr,i}}^{-2} \left[p_{yr,i} \left(\sum_{m=1}^{2} f_{m_1,k,m} \; h_m \; \mu'_{i,n_i+1,m} \right) - \right. \right. \\
&\quad \left. \left. \left(\sum_{m=1}^{2} f_{m_1,k,m} \; h_m \; \mu'_{i,n_i+1,m} \right) \left(\sum_{m=1}^{2} h_m \; p_{yr,im} \right) \right] \right\} \\
\ell''_{\alpha_k,\pi_1} &= \sum_{i=1}^{I} \left\{ I_i \; {p_{yr,im}}^{-2} \left[p_{yr,i} \left(\sum_{m=1}^{2} h_m \; p'_{\alpha_k,yr,im} \right) - \left(\sum_{m=1}^{2} \pi_m \; p'_{\alpha_k,yr,im} \right) \left(\sum_{m=1}^{2} h_m \; p_{yr,im} \right) \right] \right\} \\
\ell''_{\pi_1\pi_1} &= -\sum_{i=1}^{I} {p_{y,i}}^{-2} \left(\sum_{m=1}^{2} h_m \; p_{y,im} \right)^2 - \sum_{i=1}^{I} I_i \; {p_{yr,i}}^{-2} \left(\sum_{m=1}^{2} h_m \; p_{yr,im} \right)^2
\end{aligned}$$

where $i = 1, \ldots I;\ \ t = 1, \ldots, n_i;\ \ m, m_1, m_2 = 1, 2;$

$f_{m_1,k,m} = I(m = m_1 \text{ or } k = 3, 4);\ h_1 = 1,\ h_2 = -1;$

$\mu_{itm},\ \mu'_{itm},\ \mu''_{itm}$ are similarly defined as in conditional AR1 model for group m;

$\nu_{it},\ \nu'_{it},\ \nu''_{it}$ and $z_{i,n_i+1,kh},\ \nu_{i,n_i+1,h},\ \nu'_{i,n_i+1,h},\ \nu''_{i,n_i+1,h},\ h = 0, 1$ are also similarly defined;

$$p_{y,i} = \sum_{m=1}^{2} \pi_m \left[\prod_{t=1}^{n_i} {\mu_{itm}}^{y_{it}} (1 - \mu_{itm})^{1-y_{it}} \right] = \sum_{m=1}^{2} \pi_m \ p_{y,im};$$

$$p'_{\beta_k,y,im} = \sum_{t=1}^{n_i} x_{itk} \ (y_{it} - \mu_{itm});$$

$$p''_{\beta_{k_1},\beta_{k_2},y,im} = -\sum_{t=1}^{n_i} x_{itk_1} \ x_{itk_2} \ \mu_{itm}(1 - \mu_{itm}).$$

$$p'_{\alpha_k,r,i} = -\sum_{t=2}^{n_i} z_{itk} \ \nu_{it};\ \ p''_{\alpha_{k_1},\alpha_{k_2},r,i} = -\sum_{t=2}^{n_i} z_{itk_1} \ z_{itk_2} \ \nu'_{it};$$

$$p_{yr,i} = \sum_{m=1}^{2} \pi_m \ [\mu_{i,n_i+1,m} \ \nu_{i,n_i+1,1} + (1 - \mu_{i,n_i+1,m}) \ \nu_{i,n_i+1,0}] = \sum_{m=1}^{2} \pi_m \ p_{yr,im},$$

$$p'_{\beta_k,yr,im} = x_{i,n_i+1,k} \ \mu'_{i,n_i+1,m}(\nu_{i,n_i+1,1} - \nu_{i,n_i+1,0});$$

$$p'_{\alpha_k,yr,im} = \mu_{i,n_i+1,m} \ z_{i,n_i+1,k1} \ \nu'_{i,n_i+1,1} + (1 - \mu_{i,n_i+1,m}) \ z_{i,n_i+1,k0} \ \nu'_{i,n_i+1,0};$$

$$p''_{\beta_{k_1},\beta_{k_2},yr,im} = x_{i,n_i+1,k_1} \ x_{i,n_i+1,k_2} \ \mu''_{i,n_i+1,m} \ (\nu_{i,n_i+1,1} - \nu_{i,n_i+1,0});$$

$$p''_{\alpha_{k_1},\alpha_{k_2},yr,im} = \mu_{i,n_i+1,m} \ z_{i,n_i+1,k_1,1} \ z_{i,n_i+1,k_2,1} \ \nu''_{i,n_i+1,1} +$$

$$(1 - \mu_{i,n_i+1,m}) \ z_{i,n_i+1,k_1,0} \ z_{i,n_i+1,k_2,0} \ \nu''_{i,n_i+1,0};$$

$$p''_{\beta_{k_1},\alpha_{k_2},yr,im} = x_{i,n_i+1,k_1} \ \mu'_{i,n_i+1,m} \ (z_{i,n_i+1,k_2,1} \ \nu'_{i,n_i+1,1} - z_{i,n_i+1,k_2,0} \ \nu'_{i,n_i+1,0}).$$

3. For the random intercept model, we assign a normal distribution to the random intercept term of each patient. The first order derivatives are:

$$\ell'_{\beta_k} = (L_q^{-1})\frac{1}{M}\sum_{j=1}^{M} L_{q,j}\left[\sum_{i=1}^{I}(p'_{\beta_k,y,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\beta_k,yr,i,j})\right]$$

$$\ell'_{\alpha_k} = (L_q^{-1})\frac{1}{M}\sum_{j=1}^{M} L_{q,j}\left[\sum_{i=1}^{I}(p'_{\alpha_k,r,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\alpha_k,yr,i,j})\right]$$

$$\ell'_{\ln\sigma^2} = (L_q^{-1})\frac{1}{M}\sum_{j=1}^{M} L_{q,j}\left[\sum_{i=1}^{I}\phi'_{\ln\sigma^2,ij}\right]$$

and the second derivatives are:

$$\begin{aligned}
\ell''_{\beta_{k_1}\beta_{k_2}} = L_q^{-2}\Bigg\{ L_q\Bigg\{\frac{1}{M}\sum_{j=1}^{M}\Bigg\{ L_{q,j}\Bigg[\Big[\sum_{i=1}^{I}[p''_{\beta_{k_1},\beta_{k_2},y,i,j} + \\
{p_{yr,i,j}}^{-2}\ I_i(p''_{\beta_{k_1},\beta_{k_2},yr,i,j}\ p_{yr,i,j} - p'_{\beta_{k_1},yr,i,j}\ p'_{\beta_{k_2},yr,i,j})]\Big] + \\
\Big[\sum_{i=1}^{I}(p'_{\beta_{k_1},y,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\beta_{k_1},yr,i,j})\Big]\Big[\sum_{i=1}^{I}(p'_{\beta_{k_2},y,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\beta_{k_2},yr,i,j})\Big]\Bigg\} - \\
\Bigg\{\frac{1}{M}\sum_{j=1}^{M} L_{q,j}\Bigg[\sum_{i=1}^{I}(p'_{\beta_{k_1},y,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\beta_{k_1},yr,i,j})\Bigg]\Bigg\} \\
\Bigg\{\frac{1}{M}\sum_{j=1}^{M} L_{q,j}\Bigg[\sum_{i=1}^{I}(p'_{\beta_{k_2},y,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\beta_{k_2},yr,i,j})\Bigg]\Bigg\}\Bigg\}
\end{aligned}$$

$$\begin{aligned}
\ell''_{\alpha_{k_1}\alpha_{k_2}} = L_q^{-2}\Bigg\{ L_q\Bigg\{\frac{1}{M}\sum_{j=1}^{M}\Bigg\{ L_{q,j}\Bigg[\Big[\sum_{i=1}^{I}[p''_{\alpha_{k_1},\alpha_{k_2},y,i,j} + \\
{p_{yr,i,j}}^{-2}\ I_i(p''_{\alpha_{k_1},\alpha_{k_2},yr,i,j}\ p_{yr,i,j} - p'_{\alpha_{k_1},yr,i,j}\ p'_{\alpha_{k_2},yr,i,j})]\Big] + \\
\Big[\sum_{i=1}^{I}(p'_{\alpha_{k_1},r,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\alpha_{k_1},yr,i,j})\Big]\Big[\sum_{i=1}^{I}(p'_{\alpha_{k_2},r,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\alpha_{k_2},yr,i,j})\Big]\Bigg\} - \\
\Bigg\{\frac{1}{M}\sum_{j=1}^{M} L_{q,j}\Bigg[\sum_{i=1}^{I}(p'_{\alpha_{k_1},r,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\alpha_{k_1},yr,i,j})\Bigg]\Bigg\} \\
\Bigg\{\frac{1}{M}\sum_{j=1}^{M} L_{q,j}\Bigg[\sum_{i=1}^{I}(p'_{\alpha_{k_2},r,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\alpha_{k_2},yr,i,j})\Bigg]\Bigg\}\Bigg\}
\end{aligned}$$

$$\ell''_{\beta_{k_1}\alpha_{k_2}} = \sum_{i=1}^{I} {p_{yr,i}}^{-2}\ I_i(p''_{\beta_{k_1},\alpha_{k_2},yr,i,j}\ p_{yr,i,j} - p'_{\beta_{k_1},r,i,j}\ p'_{\alpha_{k_2},yr,i,j})$$

$$\ell''_{\alpha_{k_1}\beta_{k_2}} = L_q^{-2}\left\{L_q\left\{\frac{1}{M}\sum_{j=1}^{M}\left\{L_{q,j}\left[\left[\sum_{i=1}^{I}{p_{yr,i,j}}^{-2}\ I_i\right.\right.\right.\right.$$

$$\left.(p''_{\beta_{k_1},\alpha_{k_2},yr,i,j}\ p_{yr,i,j} - p'_{\beta_{k_1},r,i,j}\ p'_{\alpha_{k_2},yr,i,j})\right] +$$

$$\left.\left.\left.\left[\sum_{i=1}^{I}(p'_{\alpha_{k_1},r,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\alpha_{k_1},yr,i,j})\right]\left[\sum_{i=1}^{I}(p'_{\beta_{k_2},y,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\beta_{k_2},yr,i,j})\right]\right]\right\}\right\} -$$

$$\left\{\frac{1}{M}\sum_{j=1}^{M}L_{q,j}\left[\sum_{i=1}^{I}(p'_{\alpha_{k_1},r,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\alpha_{k_1},yr,i,j})\right]\right\}$$

$$\left.\left\{\frac{1}{M}\sum_{j=1}^{M}L_{q,j}\left[\sum_{i=1}^{I}(p'_{\beta_{k_2},y,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\beta_{k_2},yr,i,j})\right]\right\}\right\}$$

$$\ell''_{(\ln\sigma^2)^2} = L_q^{-2}\left\{L_q\left\{\frac{1}{M}\sum_{j=1}^{M}\left\{L_{q,j}\left[\left[\sum_{i=1}^{I}\phi''_{(\ln\sigma^2)^2,ij}\right] + \left[\sum_{i=1}^{I}\phi'_{\ln\sigma^2,ij}\right]^2\right]\right\} -\right.\right.$$

$$\left.\left\{\frac{1}{M}\sum_{j=1}^{M}L_{q,j}\left[\sum_{i=1}^{I}\phi'_{\ln\sigma^2,ij}\right]\right\}^2\right\}$$

$$\ell''_{\ln\sigma^2\beta_k} = L_q^{-2}\left\{L_q\left\{\frac{1}{M}\sum_{j=1}^{M}\left\{L_{q,j}\left[\sum_{i=1}^{I}\phi'_{\ln\sigma^2,ij}\right]\left[\sum_{i=1}^{I}(p'_{\beta_k,y,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\beta_k,yr,i,j})\right]\right\}\right\} -\right.$$

$$\left.\left\{\frac{1}{M}\sum_{j=1}^{M}L_{q,j}\left[\sum_{i=1}^{I}\phi'_{\ln\sigma^2,ij}\right]\right\}\left\{\frac{1}{M}\sum_{j=1}^{M}L_{q,j}\left[\sum_{i=1}^{I}(p'_{\beta_k,y,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\beta_k,yr,i,j})\right]\right\}\right\}$$

$$\ell''_{\ln\sigma^2\alpha_k} = L_q^{-2}\left\{L_q\left\{\frac{1}{M}\sum_{j=1}^{M}\left\{L_{q,j}\left[\sum_{i=1}^{I}\phi'_{\ln\sigma^2,ij}\right]\left[\sum_{i=1}^{I}(p'_{\alpha_k,r,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\alpha_k,yr,i,j})\right]\right\}\right\} -\right.$$

$$\left.\left\{\frac{1}{M}\sum_{j=1}^{M}L_{q,j}\left[\sum_{i=1}^{I}\phi'_{\ln\sigma^2,ij}\right]\right\}\left\{\frac{1}{M}\sum_{j=1}^{M}L_{q,j}\left[\sum_{i=1}^{I}(p'_{\alpha_k,r,i,j} + {p_{yr,i,j}}^{-1}\ I_i\ p'_{\alpha_k,yr,i,j})\right]\right\}\right\}$$

where $i = 1,\ldots I,\quad t = 1,\ldots,n_i, j = 1,\ldots M$ and $k, k_1, k_2 = 1,\ldots,p$ ($p = 4$ for the outcome model) or $k, k_1, k_2 = 1,\ldots,q$ ($q = 3$ for the drop-out model) respectively,

$(u^*_{1j}, u^*_{2j},\ldots,u^*_{Ij}, \beta^*_{oj}, \beta^*_{dj}, \beta^*_{tj}, \beta^*_{pvj}, \alpha^*_{oj}, \alpha^*_{tj}, \alpha^*_{ps,j}, \sigma^{*2}_j)'$ are drawn from $f(\boldsymbol{u}, \boldsymbol{\theta}|\boldsymbol{y})$

for each simulation j;

$$\mu_{itj} = \frac{e^{\eta_{itj}}}{1+e^{\eta_{itj}}};\ \eta_{itj} = u^*_{ij} + \beta_o + \beta_d d_{it} + \beta_t \ln t + \beta_{pv}\ y_{i,t-1};$$

$$\mu^*_{itj} = \frac{e^{\eta^*_{itj}}}{1+e^{\eta^*_{itj}}};\ \eta^*_{itj} = u^*_{ij} + \beta^*_{oj} + \beta^*_{dj} d_{it} + \beta^*_{tj} \ln t + \beta^*_{pv,j}\ y_{i,t-1};$$

$\mu'_{itj},\ \mu''_{itj}$ are similarly defined as in conditional AR1 model;

$$\nu_{it} = \frac{e^{\zeta_{it}}}{1+e^{\zeta_{it}}};\ \zeta_{it} = \alpha_o + \alpha_t \ln t + \alpha_{ps}\ y_{it};$$

$$\nu^*_{itj} = \frac{e^{\zeta^*_{itj}}}{1+e^{\zeta^*_{itj}}};\ \zeta^*_{itj} = \alpha^*_{oj} + \alpha^*_{tj} \ln t + \alpha^*_{ps,j}\ y_{it};$$

$\nu'_{it},\ \nu''_{it}$ and $z_{i,n_i+1,kh}, \nu'_{i,n_i+1,h},\ \nu''_{i,n_i+1,h},\ h = 1, 0$ are also similarly defined;

$$p_{y,i,j} = \prod_{t=1}^{n_i} {\mu_{itj}}^{y_{it}} (1-\mu_{itj})^{1-y_{it}};$$

$$p_{r,i} = \prod_{t=2}^{n_i} (1-\nu_{it});$$

$$p_{yr,i,j} = \mu_{i,n_i+1,j}\ \nu_{i,n_i+1,1,j} + (1-\mu_{i,n_i+1,j})\ \nu_{i,n_i+1,0,j};$$

$$\phi_{ij} = (2\pi e^{\ln\sigma^2})^{-\frac{1}{2}}\ \exp\left(-\frac{{u^*_{ij}}^2}{2e^{\ln\sigma^2}}\right);$$

$$p'_{\beta_k,y,i,j} = \frac{\partial \ln p_{y,i,j}}{\partial \beta_k} = \sum_{t=1}^{n_i} x_{itk}(y_{it} - \mu_{itj});$$

$$p''_{\beta_{k_1}\beta_{k_2},y,i,j} = \sum_{t=1}^{n_i} (-x_{itk_1} x_{itk_2} \mu'_{itj});$$

$$p'_{\alpha_k,r,i,j} = \frac{\partial \ln p_{r,i,j}}{\partial \alpha_k} = \sum_{t=2}^{n_i} (-z_{itk}\ \nu_{itj});$$

$$p''_{\alpha_{k_1}\alpha_{k_2},r,i,j} = \sum_{t=2}^{n_i} (-z_{itk_1} z_{itk_2}\ \nu'_{itj});$$

$$p'_{\beta_k,yr,i,j} = x_{i,n_i+1,k}\ \mu'_{i,n_i+1,j}(\nu_{i,n_i+1,1,j} - \nu_{i,n_i+1,0,j});$$

$$p'_{\alpha_k,yr,i,j} = \mu_{i,n_i+1,j}\ z_{i,n_i+1,k,1}\ \nu'_{i,n_i+1,1,j} + (1-\mu_{i,n_i+1,j})\ z_{i,n_i+1,k,0}\ \nu'_{i,n_i+1,0,j};$$

$$p''_{\beta_{k_1},\beta_{k_2},yr,i,j} = x_{i,n_i+1,k_1}\ x_{i,n_i+1,k_2}\ \mu''_{i,n_i+1,j}\ (\nu_{i,n_i+1,1,j} - \nu_{i,n_i+1,0,j});$$

$$p''_{\alpha_{k_1},\alpha_{k_2},yr,i,j} = \mu_{i,n_i+1,j}\ z_{i,n_i+1,k_1,1}\ z_{i,n_i+1,k_2,1}\ \nu''_{i,n_i+1,1,j} +$$

$$(1-\mu_{i,n_i+1,j})\ z_{i,n_i+1,k_1,0}\ z_{i,n_i+1,k_2,0}\ \nu''_{i,n_i+1,0,j};$$

$$p''_{\beta_{j_1},\alpha_{j_2},yr,ik} = x_{i,n_i+1,k_1}\ \mu'_{i,n_i+1,j}\ (z_{i,n_i+1,k_2,1}\ \nu'_{i,n_i+1,1,j} - z_{i,n_i+1,k_2,0}\ \nu'_{i,n_i+1,0,j});$$

$$\phi'_{\ln\sigma^2,ij} = \frac{\partial \ln \phi_{\ln\sigma^2,ij}}{\partial \ln \sigma^2} = \frac{1}{2}\left(\frac{{u^*_{ij}}^2}{\sigma^2} - 1\right);$$

$$\phi''_{(\ln\sigma^2)^2,ij} = -\frac{1}{2}\frac{{u^*_{ij}}^2}{\sigma^2};$$

$$L_{q,j} = \prod_{i=1}^{I} \frac{p_{y,i,j}\ p_{r,i}\ p_{yr,i,j}\ \phi_{ij}}{{p_{y,i,j}}^*\ {p_{r,i,j}}^*\ {p_{yr,i,j}}^*\ {\phi_{ij}}^*};$$

$$L_q = \frac{1}{M}\sum_{j=1}^{M}\left[\prod_{i=1}^{I} \frac{p_{y,i,j}\ p_{r,i}\ p_{yr,i,j}\ \phi_{ij}}{{p_{y,i,j}}^*\ {p_{r,i,j}}^*\ {p_{yr,i,j}}^*\ {\phi_{ij}}^*}\right];$$

${p_{y,i,j}}^*,\ {p_{r,i,j}}^*$ are defined similarly as $p_{y,i,j},\ p_{r,i}$ using μ^*_{itj} and ν^*_{itj} respectively;

${p_{yr,i,j}}^*$ are defined similarly as $p_{yr,i,j}$ using $\mu^*_{i,n_i+1,j}$ and $\nu^*_{i,n_i+1,j}$;

$${\phi_{ij}}^* = (2\pi e^{\ln\sigma^{*2}})^{-\frac{1}{2}}\ \exp\left(-\frac{{u^*_{ij}}^2}{2e^{\ln\sigma^{*2}}}\right);$$

APPENDIX C

WINBUGS PROGRAMS

Conditional AR1 model with ID:

```
model
        { for (i in 1:N) {
          y[i] ~ dbern(py[i])
          logit(py[i]) <- beta0 + betad*dose[i] + betat*lnt[i] + betap*prev[i]
          lnt[i] <- log(time[i])
                        }

     beta0 ~ dnorm(0.0, 0.000001)
     betad ~ dnorm(0.0, 0.000001)
     betat ~ dnorm(0.0, 0.000001)
     betap ~ dnorm(0.0, 0.000001)
     alpha0 ~ dnorm(0.0, 0.000001)
     alphat ~ dnorm(0.0, 0.000001)
     alphap ~ dnorm(0.0, 0.000001)

      for (i in 1:NR) {
          d0[i] <- 0
          d0[i] ~ dbern(pd0[i])
          logit(pd0[i]) <- alpha0 + alphat*lntd[i] + alphap*yd[i]
          lntd[i] <- log(timed[i])
                      }

     for (i in 1:R) {
      d1[i] <- 1
      d1[i] ~ dbern(pd1[i])
      logit(pd1[i]) <- indy1[i]*xalpha1[i] + (1-indy1[i])*xalpha0[i]
      xalpha0[i] <- alpha0 + alphat*lnty1[i]
      xalpha1[i] <- alpha0 + alphat*lnty1[i] + alphap
      lnty1[i] <- log(timey1[i]+1)

     indy1[i] ~ dbern(py1[i])
     logit(py1[i]) <- beta0 + betad*dosey1[i] + betat*lnty1[i] + betap*y1[i]
                   }
      }
```

```
2-group mixture model with ID
model
        { for (i in 1:N) {
          y[i] ~ dbern(py[i])
          logit(py[i]) <- indg[pat[i]]*xbeta1[i] +  (1-indg[pat[i]])*xbeta2[i]
          xbeta1[i] <- beta01 + betad1*dose[i] + betat*lnt[i] + betap*prev[i]
          xbeta2[i] <- beta02 + betad2*dose[i] + betat*lnt[i] + betap*prev[i]
          lnt[i] <- log(time[i])
                        }

      beta01 ~ dnorm(0.0, 0.000001)
      beta02 ~ dnorm(0.0, 0.000001)
      betad1 ~ dnorm(0.0, 0.000001)
      betad2 ~ dnorm(0.0, 0.000001)
      betat ~ dnorm(0.0, 0.000001)
      betap ~ dnorm(0.0, 0.000001)
      alpha0 ~ dnorm(0.0, 0.000001)
      alphat ~ dnorm(0.0, 0.000001)
      alphap ~ dnorm(0.0, 0.000001)

      pg ~ dunif(0,1)

      for (i in 1:T) {
      indg[i] ~ dbern(pg)
                  }

       for (i in 1:NR) {
           d0[i] <- 0
           d0[i] ~ dbern(pd0[i])
           logit(pd0[i]) <- alpha0 + alphat*lntd[i] + alphap*yd[i]
           lntd[i] <- log(timed[i])
                     }

      for (i in 1:R) {
       d1[i] <- 1
       d1[i] ~ dbern(pd1[i])
       logit(pd1[i]) <-  indy1[i]*xalpha1[i] + (1-indy1[i])*xalpha0[i]
       xalpha0[i] <- alpha0 + alphat*lnty1[i]
       xalpha1[i] <- alpha0 + alphat*lnty1[i] + alphap
       lnty1[i] <- log(timey1[i]+1)

      indy1[i] ~ dbern(py1[i])
      logit(py1[i]) <- indg[paty1[i]]*xbetay11[i] +  (1-indg[paty1[i]])*xbetay12[i]
      xbetay11[i] <- beta01 + betad1*dosey1[i] + betat*lnty1[i] + betap*y1[i]
      xbetay12[i] <- beta02 + betad2*dosey1[i] + betat*lnty1[i] + betap*y1[i]
                   }
       }
```

Random intercept model with ID

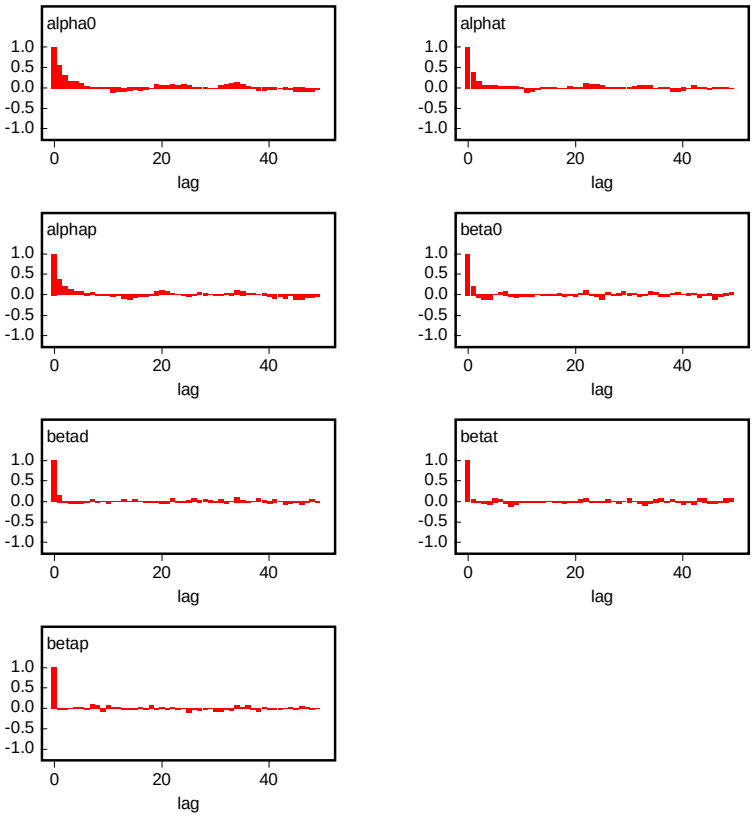
```
model
        { for (i in 1:N) {
          y[i] ~ dbern(p[i])
          logit(p[i]) <- beta0 + betad*dose[i] + betat*lnt[i] + betap*prev[i] + beta0r[pat[i]]
          lnt[i] <- log(time[i])
                        }

     beta0 ~ dnorm(0.0,0.000001)
     betad ~ dnorm(0.0, 0.000001)
     betat ~ dnorm(0.0,0.000001)
     betap ~ dnorm(0.0, 0.000001)

     alpha0 ~ dnorm(0.0, 0.000001)
     alphat ~ dnorm(0.0, 0.000001)
     alphap ~ dnorm(0.0, 0.000001)

     tau0r ~ dgamma(0.001,0.001)
     sig0r <- 1/tau0r

     for (i in 1:T) {
     beta0r[i] ~ dnorm(0.0,tau0r)
                  }

     for (i in 1:NR) {
          d0[i] <- 0
          d0[i] ~ dbern(pd0[i])
          logit(pd0[i]) <- alpha0 + alphat*lntd[i] + alphap*yd[i]
          lntd[i] <- log(timed[i])
                     }

     for (i in 1:R) {
      d1[i] <- 1
      d1[i] ~ dbern(pd1[i])
      logit(pd1[i]) <- (1-indy1[i])*xalpha0[i] + indy1[i]*xalpha1[i]
      xalpha0[i] <- alpha0 + alphat*lnty1[i]
      xalpha1[i] <- alpha0 + alphat*lnty1[i] + alphap
      lnty1[i] <- log(timey1[i]+1)

     indy1[i] ~ dbern(py1[i])
     logit(py1[i]) <- beta0 + betad*dosey1[i] + betat*lnty1[i] + betap*y1[i] + beta0r[paty1[i]]
                  }
       }
```

APPENDIX D

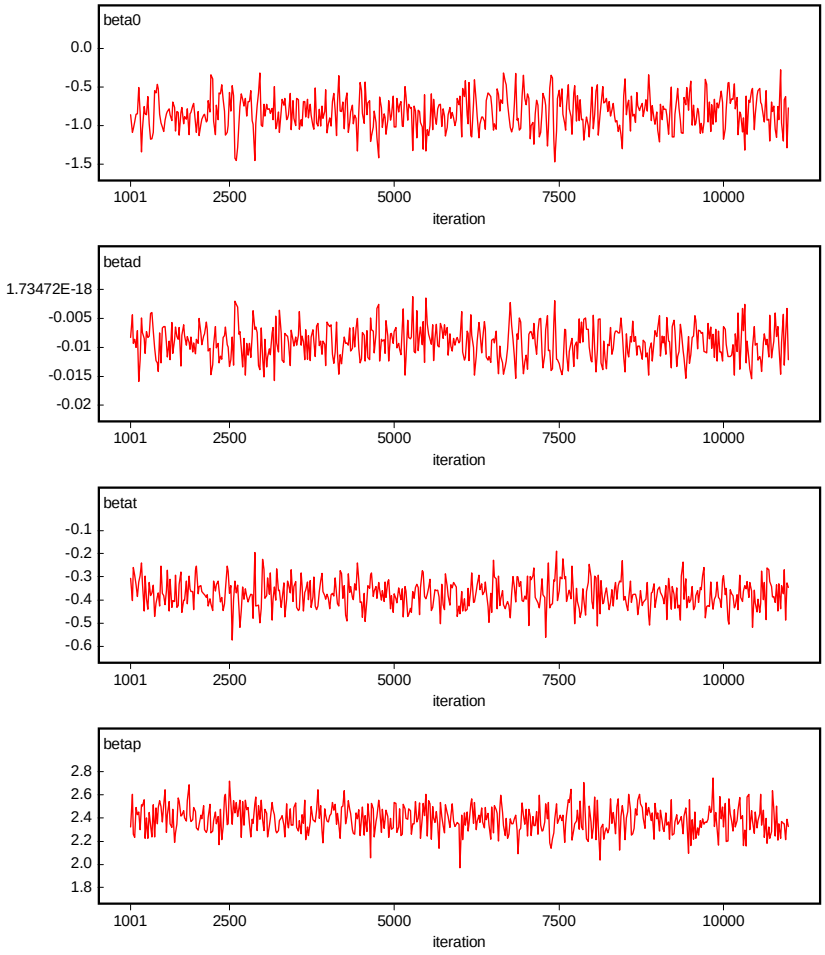
AUTO-CORRELATION FUNCTIONS AND HISTORY PLOTS

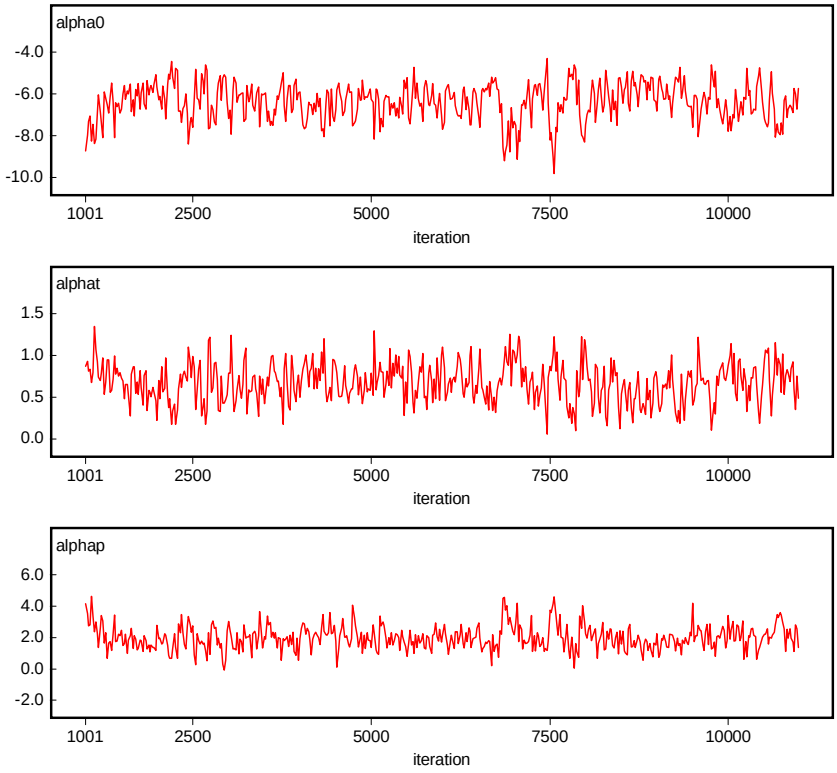
Conditional AR1 model with ID

Autocorrelation function



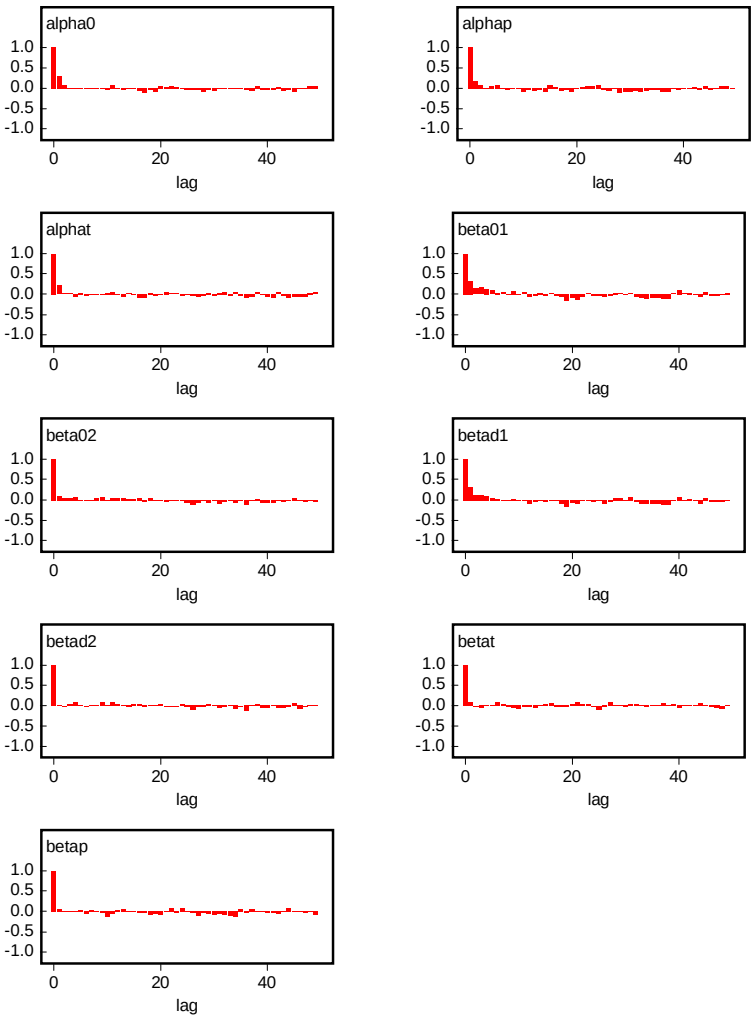
History plots



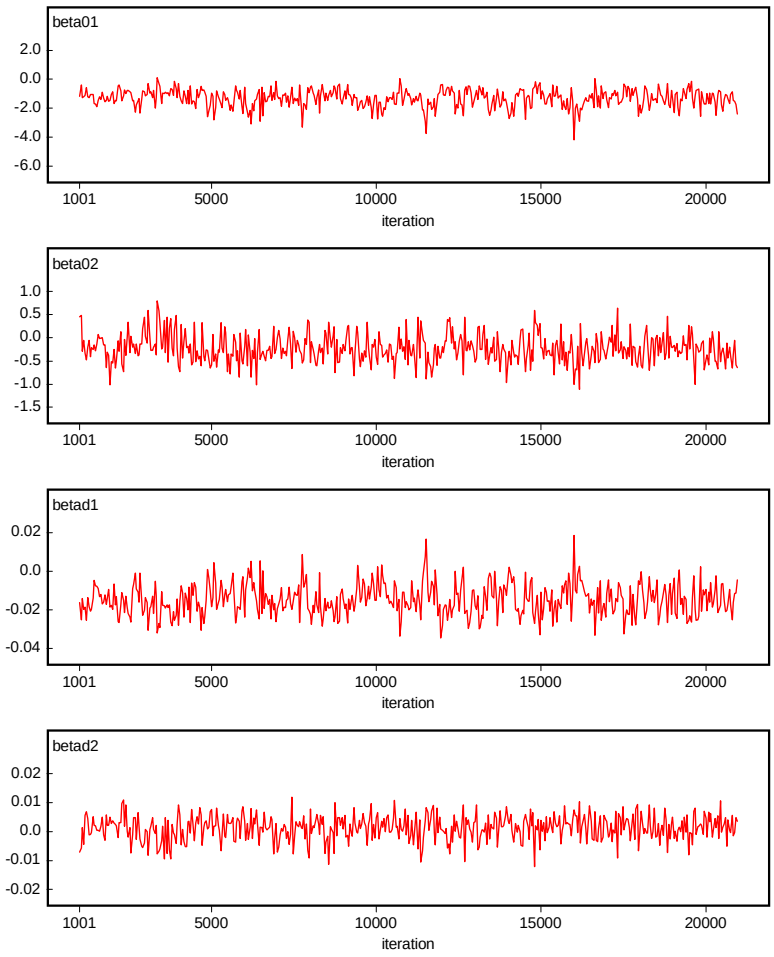

alpha0
-4.0
-6.0
-8.0
-10.0
1001
2500
5000
7500
10000
iteration
alphat
1.5
1.0
0.5
0.0
1001
2500
5000
7500
10000
iteration
alphap
6.0
4.0
2.0
0.0
-2.0
1001
2500
5000
7500
10000
iteration

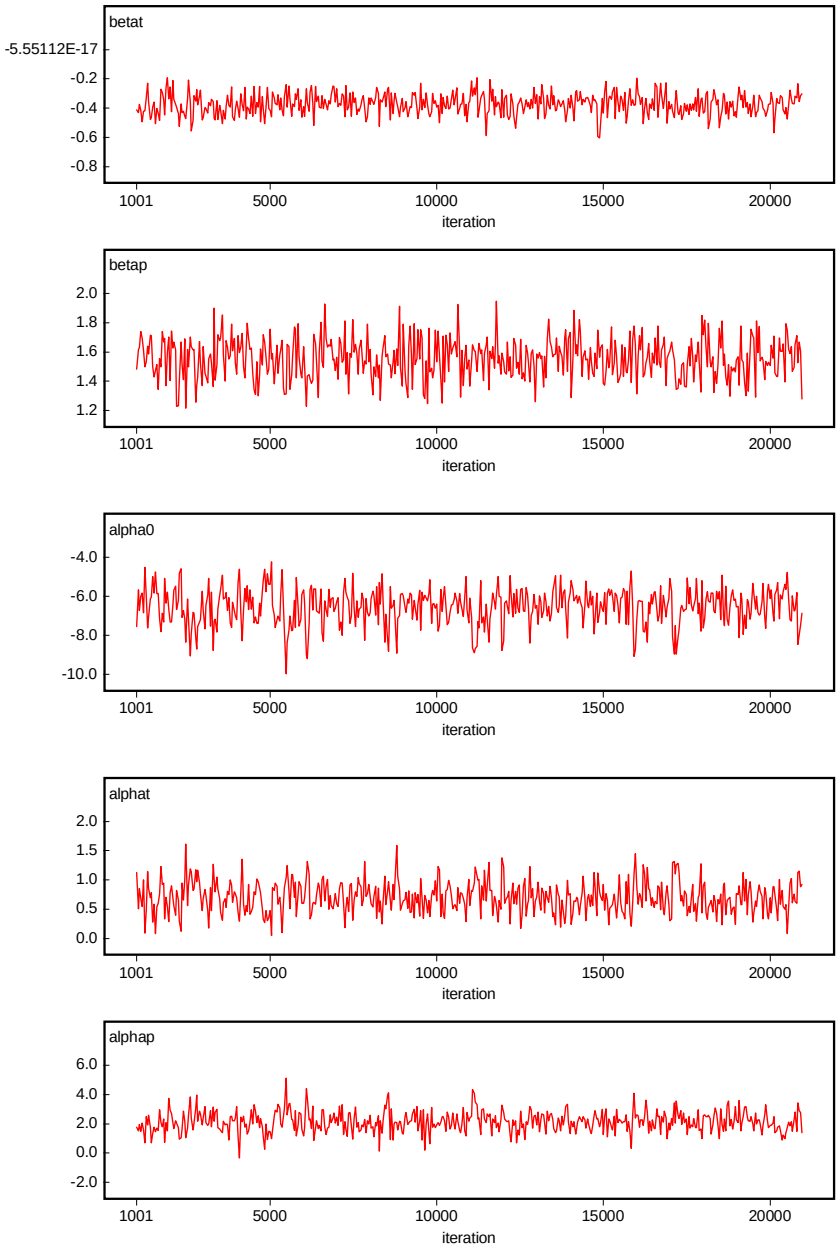
2-group mixture model with ID

Autocorrelation function



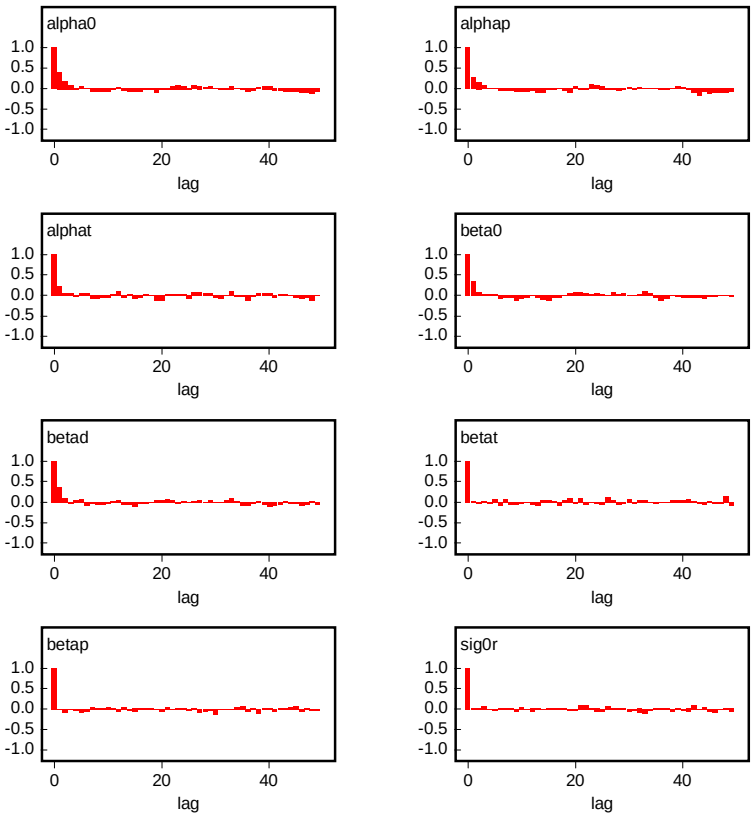
History plots



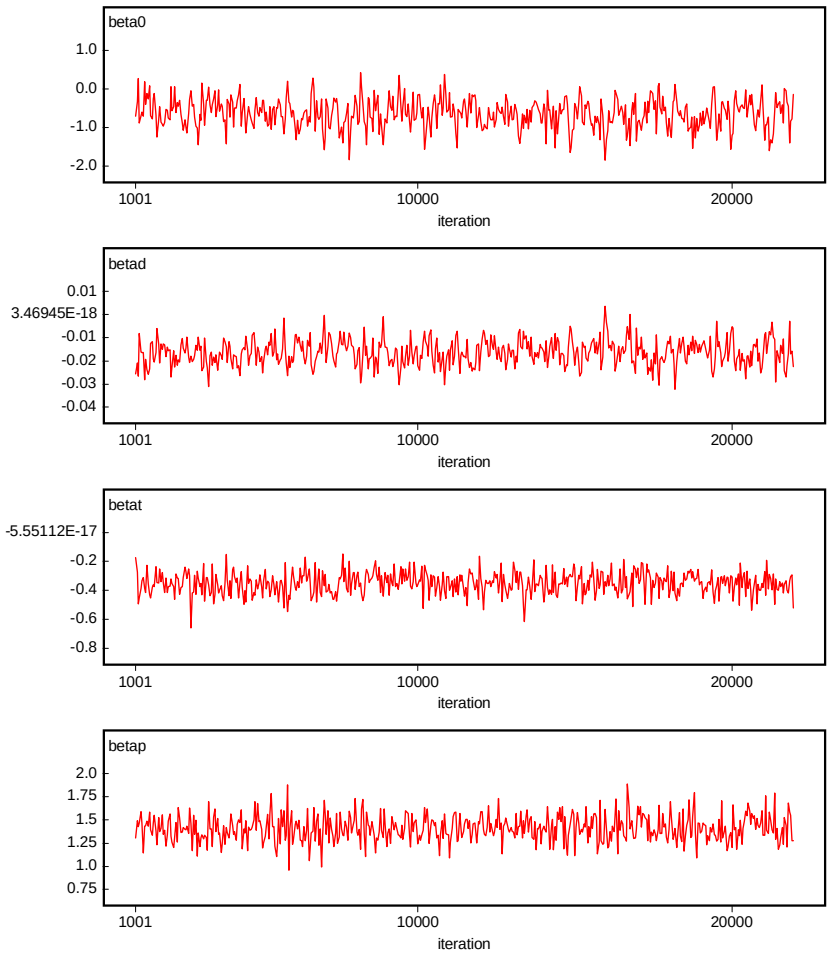

betat
-5.55112E-17
-0.2
-0.4
-0.6
-0.8
1001
5000
10000
15000
20000
iteration
betap
2.0
1.8
1.6
1.4
1.2
1001
5000
10000
15000
20000
iteration
alpha0
-4.0
-6.0
-8.0
-10.0
1001
5000
10000
15000
20000
iteration
alphat
2.0
1.5
1.0
0.5
0.0
1001
5000
10000
15000
20000
iteration
alphap
6.0
4.0
2.0
0.0
-2.0
1001
5000
10000
15000
20000
iteration

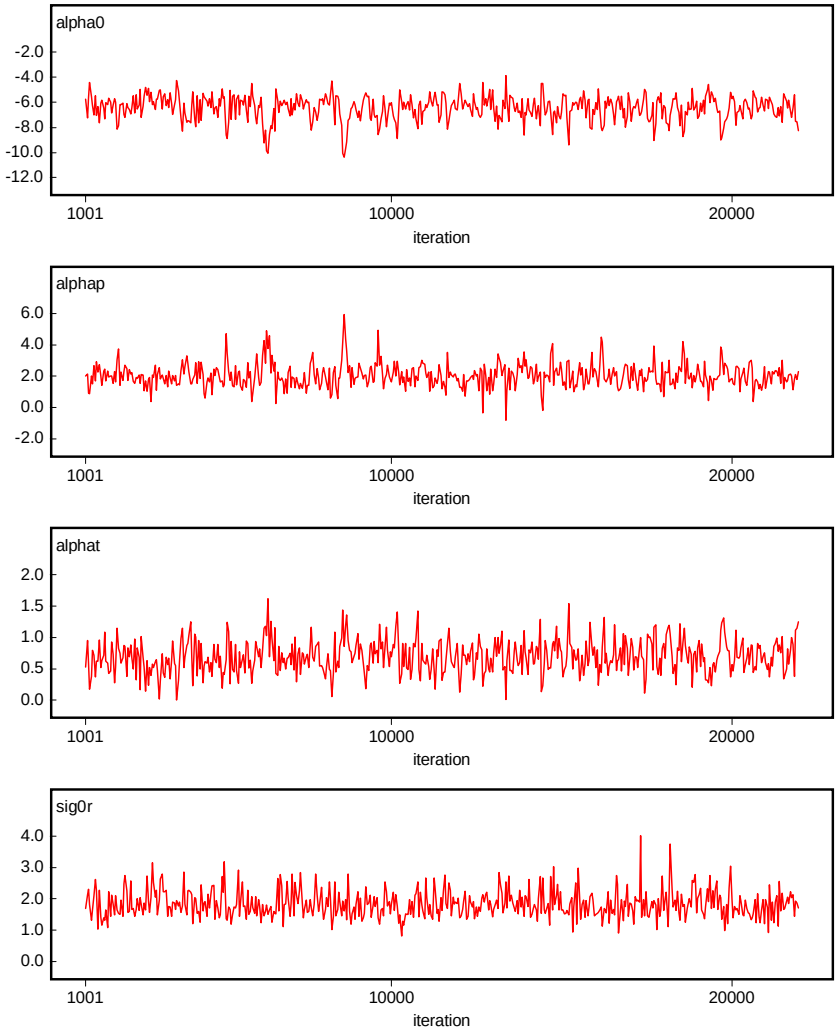
Random intercept model with ID

Autocorrelation function



History plots




alpha0
-2.0
-4.0
-6.0
-8.0
-10.0
-12.0
1001
10000
20000
iteration
alphap
6.0
4.0
2.0
0.0
-2.0
1001
10000
20000
iteration
alphat
2.0
1.5
1.0
0.5
0.0
1001
10000
20000
iteration
sig0r
4.0
3.0
2.0
1.0
0.0
1001
10000
20000
iteration