Bayesian Compound Poisson Mixed Models for Longitudinal Semi-continuous Data with Non-ignorable Missingness

by

Qiuguang Sang

Bsc. (Chem)- Qingdao University, 2005

Ph.D. (Chem)- University of Chinese Academy of Sciences, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

In the Graduate Academic Unit of Mathematics and Statistics

Supervisor(s):	Guohua Yan, Ph.D, Statistics
	Renjun Ma, Ph.D, Statistics
Examining Board:	James Watmough, Ph.D, Mathematics,Chair
	M. Tariqul Hasan, Ph.D, Statistics
	Huajie Zhang, Ph.D, Computer Science

This thesis is accepted

Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

May, 2016

Abstract

Longitudinal semi-continuous data, which contain a fair amount of zeros along with positive values, is common in biomedical, econometric and health research. The existence of missing values in this kind of data creates additional difficulty in analysis. This thesis proposes a Compound Poisson mixed model to analyse the semi-continuous data with missing data using a selection-model approach.

In the literature, the two-part model has been used most frequently to analyse semi-continuous data. In the two-part model, the zero values and positive responses from the same subject are modeled separately (Liu et al., 2012). This separation of zero and non-zero values is likely to destroy the serial dependence structure among the repeated responses within subjects. Therefore, we modelled the zero and positive responses together of the semi-continuous data while considering multilevel random effects.

Our work is motivated by Ma et al. (2007) and Hasan et al. (2009). They have proposed a multilevel random effects zero inflated Poisson model for the clustered count data with excess zeros. Furthermore they use patten-mixture model to deal with the dropouts in the longitudinal unbalanced data. Recently, Yan et al. (2016) also analyse the semi-continuous longitudinal data with Tweedie's compound Poisson based on orthodox best linear unbiased predictor (BLUP) of the random effects given the data.

This thesis applies a Tweedie's compound Poisson mixed model for longitudinal

semi-continuous data using a Bayesian approach. Our approach can accommodate both the zero and non-zero parts of the semi-continuous responses simultaneously with multilevel random effects. The regression parameters and the random effects parameters are estimated by using the Markov chain Monte Carlo (MCMC) algorithm of the Bayesian approach.

The method is demonstrated on Fluoride intake data and BSI data. Both data sets consist of a large number of zero responses as well as a substantial amount of dropout. To account for both excessive zeros and dropout patterns, we use a selection-model with compound Poisson mixed random effects for the unbalanced longitudinal data. We also conduct a simulation study to verify the proposed model.

Dedication

This report is dedicated to my husband Jiujiang and my son Sangtian, for their support and encouragement in achieving my great potential and accomplishing my goal.

Acknowledgements

I would like to express the deepest appreciation to my supervisor Dr. Guohua Yan, for his guidance, patience, and expertise in advising me with this thesis report. His advice, encouragement, persistent help and constant instruction were much appreciated during my graduate years.

I would also like to thank my co-supervisor Dr. Renjun Ma for his constant support, motivation, helpful comments and insightful knowledge.

I am also grateful to Dr. Tariqul Hasan and Dr. Jeffery Picka for their instruction and encouragement.

I thank the Department of Mathematics and Statistics for giving me the opportunity to learn in such a friendly and exciting environment.

Finally, I am deeply indebted to my friends Qing Yu, Jiejie Wang, Jiaoxiu Li and Yuhuai Wu who always inspired me with their love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In biomedical, environmental and actuarial application, it is common that the semi-continuous longitudinal random variable combines a point mass at zero with non-zero continuous values. Modelling semi-continuous data does not have a standard distribution due to the fact that the observations measured over time are correlated and strong skewness exits in the data; appropriate statistical techniques are needed for analysing the semi-continuous longitudinal data effectively.

The two-part model is commonly used to analyse semi-continuous data in the literature. A two part model generally uses a combination of a binary mixed model and a Gaussian mixed model to analyse the zero and non-zero parts respectively, but the choice of Gaussian distribution may not be appropriate as the responses are often strongly right skewed (Anderson et al., 2010). Moreover, separating zero and non-zero responses from the same subject is likely to destroy the underlying serial dependence structure.

Both the two-part model (Duan et al., 1983) and first sample selection model (Heckman, 1976, 1979) use two equations to separately model and investigate the outcome. The sample selection model posits an underlying bivariate normal error. The model estimates an unconditional equation that describes the level subjects

would have if they all had outcomes. The two-part model estimates a conditional equation that describes only the level of outcomes for those that are positive. Therefore a new model to analyse the semi-continuous responses, without separating zero and non-zero parts, and after taking random effect into account, is required.

Tweedie's Compound Poisson model (Revfeim et al., 1984) is a popular method to model data with highly right-skewed distribution, which has probability mass at zero and non-negative support. Tweedie's model has been widely applied in various areas such as actuarial and agricultural sciences. To manage correlated data with zeros, some models have been developed which incorporate serial correlation within the subjects. These models possess subject-random effect terms to handle zero and non-zero responses. As we are interested in longitudinal data, the time-random effect should also be considered in the proposed model.

Ma et al. (2007) have proposed a zero inflated Poisson model for the clustered count data with excess zeros. Multilevel random effects are introduced in the proposed model to account for the unobserved heterogeneity which arises due to the clustering effects. Ma et al. (2009) have also used a compound Poisson distribution to model the cluster level random effects. A compound Poisson distribution is a Poisson sum of independent and identical Gamma distributions. Their proposed model is more flexible as it considers only a single distribution which is Tweedie's compound Poisson distribution. Similarly, the compound Poisson distribution can be used to model semi-continuous data with a number of zero and non-zero continuous responses. Yan et al. (2016) also analyse the semi-continuous longitudinal data with Tweedies compound Poisson based on orthodox best linear unbiased predictor (BLUP) of the random effects given the data.

Bayesian methods offer the advantage of providing knowledge in the form of prior distributions of the parameters from previous studies incorporated in the current analysis. A major difference with the classical method is model parameters are

considered as random variables in Bayesian methods. Bayesian analysis of the semi-continuous data is a recent research area. By introducing priors in the model, the Bayesian approach allows the adjustment of the uncertainty of the regression parameters. A correlation structure can easily be incorporated in the Bayesian framework by considering assumptions for the priors of random effects.

In this thesis, we investigate the compound Poisson mixed model for the semi-continuous longitudinal data in the context of the Bayesian approach. Since compound Poisson distribution has a mass at zero, we can analyse both zero and non-zero continuous responses in an integral way. We also propose subject- and time-specific distribution-free random effects to model existence of any possible over dispersion among the longitudinal responses.

The remaining part of this thesis is organized into four chapters. In chapter two, we will discuss the conventional models used to analyse semi-continuous data. The models frequently used to handle semi-continuous data are the Tobit and the Heckman's selection models. In this chapter, these models are described in brief along with their disadvantage to handle zero inflation and overdispersion. The Tobit model was the first model to deal with semi-continuous data. The sample selection model extends the Tobit model to allow different coefficients to affect the two components. Both models assume an underlying normal random variable that is censored by a random mechanism. These models are sometimes suitable for modelling a limited or censored response variable. When zeros represent actual outcome values instead of censored or missing values, the underlying normal assumption becomes dubious. An alternative strand of literature for semi-continuous data does not assume an underlying normal distribution. Duan et al. (1983) proposed a two-part model to fit data on expenditures for medical care. Jørgensen (1987) proposed a compound Poisson exponential dispersion model for semi-continuous data.

Chapter three presents the fundamental ideas of the Bayesian approach as

well as the description of compound Poisson mixed effect model which is the main model on which our research is based. Simulation plays an important role in Bayesian analysis. The difficult mathematical calculations in the Bayesian approach can easily be accomplished by simulation. Simulation is used to update estimates in Bayesian problems. Markov chain Monte Carlo (MCMC) methods, specifically the Metropolis Hastings algorithm and the Gibbs sampler, are described in chapter three. The applications of Bayesian approach are discussed along with the Compound Poisson mixed effect model.

In chapter four, we have used BSI and infant fluoride intake data with excessive zeros collected from health studies to evaluate the proposed compound Poisson mixed model using the Bayesian approach. The results of the analysis are presented along with some relevant graphs and figures which show the parameter inferences and test results. The longitudinal data often have missing observations, subsequently, the non-ignorable missing observations are challenging to analyse. There are several approaches available to handle the non-ignorable nonresponses; we used the selection-model to deal with the missing values. Finally, a brief discussion about the simulation process is presented to assess the compound Poisson mixed model.

Chapter five concludes the approach we used to show how the Bayesian framework, combined with the MCMC method to handle multilevel random effects in semi-continuous data using Tweedies compound Poisson distribution. The Bayesian approach has the advantages of ease of interpretation, incorporation of prior knowledge, and reduced small sample bias.

Chapter 2

Literature Review

2.1 Previous methods to deal with semi-continuous data

In real life continuous outcome measures together with a large number of observed values clustered at zero do not follow a standard distribution. The traditional approach is modelling the two parts of outcome measures separately. We introduce several existing modelling methods for analysing semi-continuous data in the following sections.

2.1.1 Sample selection models

2.1.1.1 Tobit model

Tobin (1958) introduces the censored regression model for analysing the household expenditure data, which is commonly known as Tobit model. He considers the normal distribution assumption to the positive responses, and then the positive responses are modelled with the appropriate covariates. Here we will discuss the Tobit model very briefly. Let $Y = (Y_1, ..., Y_n)'$ be the response. In the Tobit model, it was assumed

that Y_i^p be a normally distributed variable, which can be defined as

$$Y_i = \begin{cases} Y_i^p & \text{if } Y_i^p > 0 \\ 0 & \text{if } Y_i^p \leq 0 \end{cases}, \tag{2.1}$$

where $i = 1, 2, ..., n$ and this division of the response indicates only non zero positive responses are observed and considered in the model. By using explanatory variables, Y_i^p can be expressed as

$$Y_i^p = X_i\beta + \epsilon_i,$$

where X_i are the explanatory variables for the i^{th} response and ϵ_i be the random component, β are the fixed effect parameters and ϵ_i follows the normal distribution with mean zero and variance τ^2,

$$\tau_i \sim N(0, \tau^2).$$

In the Tobit model, probability density function is found for both zero part and non zero positive part. For the zero part of the data,

$$\begin{aligned} p(Y_i = 0) &= 1 - p(Y_i^p > 0) = 1 - p(X_i\beta + \epsilon_i > 0) \\ &= 1 - p(X_i\beta > -\epsilon_i) = 1 - p(-X_i\beta < \epsilon_i) \\ &= 1 - p(\epsilon_i > -X_i\beta) = 1 - [1 - p(\epsilon_i < -X_i\beta)] \\ &= 1 - \left[1 - p\left(\frac{\epsilon_i - 0}{\tau}\right) < -\frac{X_i\beta - 0}{\tau}\right] = 1 - \left[1 - p\left(z < -\frac{X_i\beta}{\tau}\right)\right] \\ &= 1 - \left[1 - F\left(-\frac{X_i\beta}{\tau}\right)\right] \\ &= F\left(-\frac{X_i\beta}{\tau}\right). \end{aligned} \tag{2.2}$$

where $F(\cdot)$ indicates the cumulative probability density function for the standard normal distribution $N(0,1)$. For the positive response part of the data,

$$\begin{aligned} p(Y_i = Y_i^p) &= p(Y_i = X_i\beta + \epsilon_i) \\ &= p(\epsilon_i = Y_i - X_i\beta) \\ &= p\left(\frac{\epsilon_i - 0}{\tau} = \frac{Y_i - X_i\beta}{\tau}\right) \\ &= p\left(z = \frac{Y_i - X_i\beta}{\tau}\right) \\ &= \frac{1}{\sqrt{2\pi r}} \exp\left[-\left(\frac{Y_i - X_i\beta}{\tau}\right)^2\right] \end{aligned} \quad . \tag{2.3}$$

Assuming all the observations are independent, the likelihood function looks like

$$L(\beta, \tau) = \left[\prod_{Y_i=0} F\left(-\frac{X_i\beta}{\tau}\right)\right]\left[\prod_{Y_i>0} \frac{1}{\sqrt{2\pi r}} \exp\left[-\left(\frac{Y_i - X_i\beta}{\tau}\right)^2\right]\right]. \tag{2.4}$$

The Tobit model assumes normality for the distribution of the error term, with constant variance. This is unrealistic in many applications. When the model form is correct but the distribution of ϵ_i is not normal, the ML estimators are inconsistent (Robinson 1982).

2.1.1.2 Heckman's selection model

Heckman introduces a selection model to analyse the women participation in the labour force. The model is an extension of the Tobit model to a simultaneous equation system. This model is often used to the censored dependent variables like the Tobit model. Duan et al. (1983) argues that when zeros represent nothing but the actual data then their two-part model is more clearly interpretable than Heckman's selection model. There are several ways of understanding Heckman's sample selection

model. Here we will follow the way of Min and Agresti (2002). Let $Y = (Y_1, ..., Y_n)'$ be the response. In the sample selection model, there are two basic equations, the outcome equation and the selection equation. Outcome equation can be expressed as

$$Y_i = \begin{cases} Y_i^p & \text{when } z_i^p > 0 \\ 0 & \text{when } z_i^p \leq 0 \end{cases}, \tag{2.5}$$

where z_i^p relates to the selection equation and

$$Y_i^p = X_i\beta + \epsilon_i, \tag{2.6}$$

where X are the explanatory variables, β are the fixed effect parameters and the random component ϵ_i can be defined as

$$\epsilon_i \sim N(0, \tau^2),$$

Selection equation can be defined as

$$z_i^p = W_i\phi + \psi_i, \tag{2.7}$$

where W are the explanatory variables, ϕ are the regression parameters and the random component ψ_i defined as follows

$$\psi_i \sim N(0, \eta^2),$$

We can define an indicator variable z_i, which can be expressed as

$$z_i = \begin{cases} 1 & \text{when } z_i^p > 0 \\ 0 & \text{when } z_i^p \leq 0 \end{cases}. \tag{2.8}$$

It was assumed in the model that the random components of the outcome equation and the selection equation are correlated and together follow a bivariate normal distribution, which can be expressed as

$$\begin{pmatrix} \epsilon_i \\ \psi_i \end{pmatrix} = N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Psi = \begin{pmatrix} \tau^2 & \text{cov}(\epsilon_i, \psi_i) \\ \text{cov}(\epsilon_i, \psi_i) & \eta^2 \end{pmatrix} \right].$$

By doing some small modification, it can be easily shown that the Tobit model is a special case of the selection model. If we replace z_i by Y_i then the sample selection model gives rise to the Tobit model. Sometimes the sample selection model is also termed as generalized Tobit model or type 2 Tobit model.

There are two ways of estimating the Heckman's (1974) sample selection model. They are Heckman's two-step procedure and maximum likelihood estimation technique. In the following paragraphs we will describe them very briefly. Heckman's two-step procedure is a widely used technique for estimation. In the estimation technique we need to find

$$\begin{aligned} E[Y_i^p | z_i^p > 0] &= E[X_i\beta + \epsilon_i | W_i\phi + \psi_i > 0] \\ &= X_i\beta + E[\epsilon_i | W_i\phi + \psi_i > 0] \\ &= X_i\beta + E[\epsilon_i | \psi_i > -W_i\phi]. \end{aligned} \tag{2.9}$$

As ϵ_i and ψ_i are not independent and have a bivariate normal distribution, according to Greene (2002), we can write

$$E[\epsilon_i | \psi_i > -W_i\phi] = 0 + \text{corr}(\epsilon_i, \psi_i)\tau\lambda(\alpha_\psi), \tag{2.10}$$

and

$$Var[\epsilon_i | \psi_i > -W_i\phi] = \tau^2[1 - [\text{corr}(\epsilon_i, \psi_i)]^2\delta(\alpha_\psi)], \tag{2.11}$$

where $\alpha_\psi = \frac{(-W_i\phi - 0)}{\eta}, \lambda(\alpha_\psi) = \frac{f(\alpha_\psi)}{1-F(\alpha_\psi)}, \delta(\alpha_{psi}) = \frac{\lambda(\alpha_\psi)}{\lambda(\alpha_\psi)-\alpha_\psi}$ and $f(), F()$ are the density and the cumulative distribution function of the standard normal distribution respectively and η is the square root of the variance of random component ψ_i of the selection equation.

Therefore,

$$\begin{aligned} E[Y_i^p | z_i^p > 0] &= X_i\beta + E[\epsilon_i | \psi_i > -W_i\phi] \\ &= X_i\beta + corr(\epsilon_i, \psi_i)\tau\lambda(\alpha_\psi). \end{aligned} \tag{2.12}$$

Thus,

$$\begin{aligned} Y_i^p | z_i^p > 0 &= E[Y_i^p | z_i^p > 0] + \omega_i \\ &= X_i\beta + corr(\epsilon_i, \psi_i)\tau\lambda(\alpha_\psi) + \omega_i. \end{aligned} \tag{2.13}$$

Following Heckman (1974), ω_i has mean zero and variance

$$\eta^2[1 + [\text{corr}(\epsilon_i, \psi_i)]^2(\alpha_\psi\lambda(\alpha_\psi) - [\lambda(\alpha_\psi)]^2)]. \tag{2.14}$$

In this estimation technique, ψ and η are estimated first then are used to find the estimate of α_ψ, which will help find $\lambda\alpha_\psi$. Next by using all the informations β and τ can be estimated from the following equation,

$$Y_i^p | z_i^p > 0 = X_i\beta + \text{corr}(\epsilon_i, \psi_i)\tau\lambda(\alpha_\psi) + \omega_i. \tag{2.15}$$

Maximum likelihood estimation technique can be used to estimate the parameters of the sample selection model. To do so any standard iteration technique is used to the following likelihood equation,

$$L(\beta, \phi, \Psi) = \left[\prod_{Y_i=0} p(z_i^p \leq 0)\right] \left[\prod_{Y_i>0} p(Y_i^p | z_i^p > 0) p(z_i^p > 0)\right]. \qquad (2.16)$$

Heckman's two-step procedure does not perform as well as the ML estimators. But this method is very simple and easy to implement. It is widely used and has become the standard estimation procedure for empirical micro-econometrics studies.

2.1.2 Two-part model

The Tobit model allows the same underlying stochastic process to determine whether the response is zero or positive as well as the value of a positive response. That is, the same parameters influence whether the outcome is zero or positive as well as the magnitude of the outcome, conditional on its being positive. The next two subsections discuss the two-part model which allows the two components to have different parameters.

A simple way to model semi-continuous data is to present them as a two-part mixture consisting of a continuous distribution and a mass clustered at zero. As noted by Duan et al. (1983), in the first step of the model they use a binary model for the expenditure to find whether the response is positive or not, in the contrast, the second part of the two-part model describes the conditional mean of the response given that it is positive. The two-part model usually uses a combination of a binary mixed model and a Gaussian mixed model to analyse the zero and non-zero parts separately. But sometimes the non-zero values show strongly skewed feature which causes that Gaussian mixed model may not be appropriate, so logarithmic transformation is often used in the second part of the model. Next few paragraphs show the brief description of the two-part model.

2.1.2.1 Model and notation

Let Y_i denote the ith semi-continuous response, $i = 1, 2, \ldots, n$. This response can be recoded as two variables,

$$Z_i = \begin{cases} 1 & \text{if } Y_i \neq 0 \\ 0 & \text{if } Y_i = 0 \end{cases}. \tag{2.17}$$

In the first part of the model, a logistic regression is used to deal with the dichotomous responses, which can be represented as

$$\text{logit}[p(Z_i = 0)] = \text{logit}[p(Y_i = 0)] = X_i\beta$$

$$\Rightarrow \log\left[\frac{p(Y_i = 0)}{1 - p(Y_i = 0)}\right] = X_i\beta$$

$$\Rightarrow p(Y_{ij} = 0) = \frac{\exp[X_i\beta]}{1 + \exp[X_i\beta]},$$

where X_i are the explanatory variables and β indicates the fixed effect parameters. The second part of the model shows the conditional distribution given the binary indicator,

$$V_i = \begin{cases} g(Y_i) & \text{if } Y_i \neq 0 \\ \text{irrelevant} & \text{if } Y_i = 0 \end{cases}. \tag{2.18}$$

In general, the model can be written as

$$g(Y_i|Z_i = 1) = X_i^*\gamma + \varepsilon_i, \tag{2.19}$$

where X_i^* are the explanatory variables, γ indicates the fixed effect parameters and ε_i

is the random component. The function of $g()$ is a monotone increasing function(e.r., log) that will make the transformed responses V_i to be approximately normally distributed. Maximum likelihood estimation is often used to estimate the parameters of the two-part model.

Later, a two-part random effects model is developed by Olsen and Schafer(2001) as a extension of the traditional two-part model we introduced above. One random effect is considered in the first part of the two-part model. Let Y_{ij} denote a semi-continuous response for the ith subject at the jth occasion.

For the first part,

$$Z_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \neq 0 \\ 0 & \text{if } Y_{ij} = 0. \end{cases} \tag{2.20}$$

Logistic regression can also be applied to the first part of the responses. Comparing to the aforementioned two-part model, a random effect is added in this step,

$$\text{logit}[p(Z_{ij} = 1)] = X_{ij}\beta + B_{ij}c_i, \tag{2.21}$$

where X_{ij} and B_{ij} are the explanatory variables, and β is the fixed effect parameter and c_i is the random effect parameter.

The second part of the responses can be defined as

$$V_{ij} = \begin{cases} g(Y_{ij}) & \text{if } Y_{ij} \neq 0 \\ \text{irrelevant} & \text{if } Y_{ij} = 0 \end{cases}, \tag{2.22}$$

and the model of the responses was represented as

$$g(Y_{ij}|Z_{ij} = 1) = \log(Y_{ij}|Z_{ij} = 1) = X^*_{ij}\beta^* + {B_{ij}}^*{c_i}^* + \varepsilon_i, \tag{2.23}$$

where X_{ij}^* and $B_{ij}{}^*$ are matrices of covariates, β^* and $c_i{}^*$ are parameters of fixed and random effects respectively, and ε_i is the the residual which is assumed to follow $N(0, \sigma^2)$.

In the two-part mixed model, the random coefficients from the binary and conditional parts are assumed to be jointly normal and correlated.

$$\begin{pmatrix} \Phi_1 \\ \Phi_2 \end{pmatrix} \sim N \left(0, \psi = \begin{pmatrix} \psi_{\Phi_1\Phi_1} & \psi_{\Phi_1\Phi_2} \\ \psi_{\Phi_2\Phi_1} & \psi_{\Phi_2\Phi_2} \end{pmatrix} \right).$$

2.2 Introduction to the compound Poisson distribution

In this section, I briefly review the Tweedie models based on the Tweedie exponential distribution. Tweedie distributions are a special case of exponential dispersion models.

Any exponential dispersion model can be characterized by its variance function $V()$, which describes the mean-variance relationship of the distribution when the dispersion is held constant. If Y follows an exponential dispersion model distribution that has mean μ, variance function $V()$, and dispersion ϕ, then the variance of Y can be written as

$$\text{Var}(Y) = \phi V(\mu).$$

Tweedie distributions are a special case of the exponential dispersion family for which $V(\mu) = \mu^p$ and $\text{Var}(Y) = \phi\mu^p$ (Dunn and Smyth 2005). The distribution is defined for all values of p except values in the open interval $(0, 1)$. Many important known distributions are a special case of Tweedie distributions including normal ($p = 0$), Poisson ($p = 1$), gamma ($p = 2$), and inverse Gaussian ($p = 3$). Apart from

these special cases, the probability density function of the Tweedie distribution does not have an analytical expression. For $p > 1$, it has the form

$$f(y;\mu,\phi,p) = a(y,\phi)\exp\left[\frac{1}{\phi}\left(\frac{y\mu^{1-p}}{1-p} - \kappa(\mu,p)\right)\right], \tag{2.24}$$

where

$$\kappa(\mu,p) = \mu^{2-p}/(2-p) \text{ for } p \neq 2,$$

and

$$\kappa(\mu,p) = \log(\mu) \text{ for } p = 2.$$

The function $a(y,\phi)$ does not have an analytical expression. It is usually evaluated by using the series expansion methods.

For $1 < p < 2$, the Tweedie distribution is a compound Poisson-gamma mixture distribution, which can be expressed as

$$Y = \sum_{i=1}^{N} X_i, N \sim \text{Poisson}(\lambda), X_i \sim \text{Gamma}(\alpha,\beta), N \perp X_i,$$

where N is the number of X_i, which follows a Poisson distribution with mean λ, and $X_i \sim \text{Gamma}(\alpha,\beta)$ are independently and identically distributed Gamma random variables with mean α/β and variance α/β^2. In this representation, the exponential dispersion model is denoted as compound Poisson distribution for index parameter $p \in (1,2)$. As a result, the compound Poisson distribution has a density mass at zero and a right skewed continuous positive part. This distinguished feature can be applied to analyse semi-continuous data.

The compound Poisson distribution has been applied in diverse fields, where the underlying examples can be considered as a compound Poisson process.

In an ecological study of fishery survey, Y refers to the total biomass of a

particular fish species in a certain area, N is the number of fish, and X_i is the weight of the ith fish.

In actuarial science, Y is the total claim amount of an insurance policy, N is the number of reported claims, and X_i is the individual insurance payment of the ith claim.

In the meteorological study, Y can be the totally precipitation in a given period, N is the number of rainfall events, and X_i is the precipitation of the ith rain event.

Ma et al. (2007) propose a compound Poisson random effects approach to analyse a set of medical and health care data which can characterize both the extra zeros and the remaining data with multilevel clustering. In the following section, our proposed model is discussed which is a compound Poisson mixed model based on the Bayesian approach.

2.3 Bayesian inference

Both Bayesian and frequentist inferences are widely used in statistical analysis. In some respects, Bayesian methods are older than frequentist ones, having been the basis of very early statistical reasoning as far back as the 18th century, but it was limited before the computing technology developed rapidly. The Bayesian approach has steadily gained ground, and is now recognized as an alternative to the frequentist approach due to computing advances.

This part is organized into three sections. The first gives an outline of the Bayesian method, compares the Bayesian and frequentist inferences, links their differences to fundamental differences, and argues that the Bayesian approach is more consistent and reflects better the true nature of scientific reasoning. The second section focuses on the computation methods for Bayesian approach. The final section addresses the creation of compound Poisson model using Bayesian approach.

Table 2.1: Overview of the similarities and differences between frequentist and Bayesian statistics

Topics	Frequentist statistics	Bayesian statistics
Definition of the p value	The probability of observing the same or more extreme data assuming that the null hypothesis is true in the population	The probability of the null hypothesis
Probability	Limit of empirical Frequencies	Subject belief
Inclusion of prior knowledge possible	No	Yes
Nature of the parameters in the model	Unknown but fixed	Unknown and random
Population parameter	One true value	A distribution of values reflecting uncertainty
Uncertainty is defined by	The sampling distribution based on the idea of infinite repeated sampling	Probability distribution for the population parameter
Estimation	Likelihood based(MLE) and other criteria(e.g., UMVUE)	Based on the posterior
Estimated intervals	Confidence interval: interpreted in terms of long-run behaviour of response	Credibility interval: interpreted as probability statements about the parameters

We first present the basic procedures of Bayesian inference which is a process of getting information from data. The main difference between the frequentist approach and the Bayesian approach is that the latter one considers parameters as random variables that are characterized by a prior distribution whereas the classical approach relies on the frequency interpretation of probability. In frequentist statistics, parameters are not repeatable random things but are fixed, which means that they cannot be considered as random variables. In contrast, in Bayesian statistics, the value of a parameter is uncertain, can be thought as a random variable, known specifically from prior information.

2.3.1 The ingredients of Bayesian statistics

Three ingredients underlying Bayesian statistics were firstly described by T. Bayes et al. (1763). In the following sections, these ingredients will be introduced and explained in details.

2.3.1.1 Prior knowledge

The first ingredient is the background knowledge on the parameters of the model being tested. The first ingredient refers to all information available before we start to analyse the data set, therefore, we define the first ingredient as prior distribution. We expect a prior distribution with a small variance which may reflect the uncertainty about the value of the population parameters. The smaller the variance, the more confident one is that the prior estimates reflect the population estimates.

How to define priors The data we want to analyse can moderate our priors regarding the parameters and thus lead to the updated value of priors. But how do we specify priors? The choice of a prior is based on how much information we have prior to the data collection and how accurate we believe that information to be. There are two scenarios. First, in some cases we may not be in possession of enough prior information to aid in drawing posterior inferences. From a Bayesian point of view, this lack of information is still important to consider and incorporate into our statistical specifications. Second, in some cases we may have considerable prior information regarding the value of a parameter and our sense of the accuracy around that value.

Non-informative prior distributions In many cases, we have no prior knowledge to the value that a parameter might take. We can represent prior ignorance with a non-informative prior distribution, sometimes called a diffuse distribution because such a wide range of values are considered possible. The most typical diffuse prior is a uniform probability distribution, which says that each value of the parameter is equally likely.

Informative prior distribution The other choice is an informative prior, representing cases where we have substantial prior knowledge about the value of the

parameter. For example, we might specify a Gamma distribution about our prior expectation of the parameter value. The choice of informative priors is a contentious aspect of Bayesian inference, since they are constructed from subjective opinion as opposite to previous empirical estimates. Therefore, the construction of priors is a challenge for bayesians who are looking for logical and consistent values of priors.

2.3.1.2 Posterior distribution

All conclusions from Bayesian inference are based on the posterior probability distribution of the parameters or an estimate of the posterior derived from MCMC sampling. This posterior distribution represents our prior probability distribution modified by the current data through the likelihood function. Bayesian inference is usually based on the shape of the posterior distribution, particularly the range of values over which most of the probability mass occurs. A point estimate of the parameter is usually determined from the mean of the posterior distribution.

2.3.2 Notation

We can denote a parameter or a collection of parameters by θ and the data can be denoted as Roman letter y which may be a vector of observations of a single variable or a matrix of observations of more than one variable. Bayesian inference is based on Bayes Theorem (1763) which states that the conditional probability for event A given B is related to the conditional probability of B given A as follows:

$$\begin{aligned} p(A_j|B) &= \frac{p(B, A_j)}{p(B)} \\ &= \frac{p(B|A_j)p(A_j)}{\sum_{j=1}^{J} p(B, A_j)} \\ &= \frac{p(B|A_j)p(A_j)}{\sum_{j=1}^{J} p(B|A_j)p(A_j)} \end{aligned} \quad . \tag{2.25}$$

The event of interest B was given and a collection of events A_j $(j = 1, ..., J)$ that are mutually exclusive. Where $p(B, A_j)$ indicates the joint probability when both A and B occur.

In the Bayesian approach, we use the data y to learn about the parameters θ which will be updated in the learning process. Therefore, *prior* and *posterior* information are used to represent the knowledge before and after observing the data. Bayesian approach also begins with a sampling model for the data y given a vector of unknown parameters θ.

The Bayes theorem can be used for parameter estimation replacing the terms A and B with parameter vector θ and data y to obtain,

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{p(y|\theta)\pi(\theta)}{p(y)}, \tag{2.26}$$

with

$$p(y) = \int \cdots \int \mathrm{p(y|}\theta)\pi(\theta)\,\mathrm{d}\theta. \tag{2.27}$$

The sampling model is given in the form of a probability distribution $p(y|\theta)$ which is also the likelihood if regarded a function of θ. $\pi(\theta)$ is called the prior distribution of θ which reveals the information regarding the parameter θ before observing the data y. The posterior probability distribution $p(\theta|y)$ is the distribution of θ obtained from the combinational information in both data and the prior distribution. In Bayesian inference, the main interest is to estimate the posterior distribution of θ given the data y. $p(y)$ is the integral of $\pi(\theta)p(y|\theta)$ over all possible values of θ which indicates the $p(y)$ will always be a constant for a particular data set. Hence we may leave $p(y)$ out, the posterior can be written as,

$$p(\theta|y) \propto \pi(\theta)p(y|\theta),$$

that is,

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

Over all the conceptual reviews above, a typical Bayesian analysis can be outlined in the following steps:

The first step in a Bayesian analysis is to choose a probability model for the data. This process is similar with the frequentist approach of choosing an appropriate probability distribution for the data if the parameters are fixed.

Once the probability model was chosen, a prior distribution was proposed for the unknown model parameters. The approach to choosing a prior distribution depends on the knowledge from similar data with substantive problem or the statisticians' experiences and beliefs about the unknown parameters.

After the data has been observed, the likelihood function is constructed. The likelihood is the joint probability function of the data, but viewed as a function of the parameters, treating the observed data as fixed quantities. In the Bayesian framework, all of the information about θ coming directly from the data is contained in the likelihood. Now, we can apply the Bayesian theorem to obtain the posterior distribution $p(\theta|y)$,

$$p(\theta|y) \propto \pi(\theta)p(y|\theta). \tag{2.28}$$

For the last step, point estimates of parameters and interval estimates can be summarized with an appropriate analysis after the posterior distribution has been determined.

2.3.3 Application to Bayesian models

2.3.4 Bayesian regression

A general regression model can be represented as follows:

$$y = x\beta + \varepsilon, \tag{2.29}$$

where y is the response variable and $x = (x_1, x_2, ..., x_p)'$ is a vector of explanatory variables with corresponding regression parameter $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)'$. And the residual ε is considered to be normally distributed with mean 0 and variance σ^2.

Comparing to the Bayesian model, we may consider this likelihood function of y as $f(y|\beta, \sigma)$ where β and σ are parameters. The priors for each parameters in the Bayesian model need to be defined such as β as $\pi(\beta)$ and σ as $\pi(\sigma)$. The specification of prior distributions plays a very important role in Bayesian inference since it influences the posterior directly. Finally the posterior of the model should be obtained as follows:

$$p(\beta, \sigma|x, y) \propto f(y|\beta, \sigma)\pi(\beta)\pi(\sigma). \tag{2.30}$$

2.3.4.1 Markov chain Monte Carlo methods

MCMC techniques are often applied to simulate complicated distribution which enables the estimate of posterior distribution more easily and accurately. MCMC has a great impact on the development of Bayesian theory.

What is Markov chain A Markov chain is a stochastic process whose states, discrete or continuous, are governed by a transition probability. The current state in a Markov chain only depends on the most recent previous states, the future states

are independent of past states given the present state. Or simply represented as:

$$x_t|x_{t-1}, ..., x_0 \sim P(x_t|x_{t-1}, ..., x_0) = P(x_t|x_{t-1}). \tag{2.31}$$

The Monte Carlo principle Typically a simple way of saying we can take quantities of interest of a distribution from simulated draws from the distribution.

Suppose we have a distribution $p(\theta)$ (perhaps a posterior) that we want to take quantities of interest from. To derive it analytically, we need to take integrals:

$$I = \int_{\Theta} g(\theta)p(\theta)d\theta, \tag{2.32}$$

where $g(\theta)$ some function of θ, we can approximate the integrals via Monte Carlo Integration by simulating N values from $p(\theta)$ and calculating

$$\hat{I}_N = \frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)}). \tag{2.33}$$

The Monte Carlo approximation $\hat{I}_N$ is a simulation consistent estimator of the true value $I : \hat{I}_N \to I$ as $N \to \infty$.

MCMC techniques are based on the construction of a Markov chain. To eventually reach the target posterior distribution, MCMC updates parameters iteratively which makes the estimate modified in every step depending on the previous step. After a large number of steps, known as the burn-in period, the Markov chain converges to the target posterior distribution. Implementation of Markov Chain Monte Carlo (MCMC) method has made the calculation of posterior probability much easier. Many highly complicated computations can be handled easily by using MCMC method along with the use of high technique computer and software applications. The two most popular MCMC methods are the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampling (Geman

and Geman, 1984).

Metropolis-Hastings algorithm Metropolis-Hastings (MH) algorithm is a popular MCMC method. The sampling technique is used to draw samples from a multivariate target distribution without even knowing the normalizing factor. The main idea of this method is to evolve a Markov chain so that its stationary distribution is the target distribution f. To obtain values from the target density $f(\theta)$ where θ denotes the parameters of interest, the MH algorithm employs a proposal density $q(\theta^c, \theta^p)$. This density plays an important role for two purposes: first, it provides proposal values θ^p of the parameters, given the current values θ^c, and second, it is used to decide whether this values will be accepted or not.

This is done with the help of the acceptance probability $\chi(\theta^c, \theta^p)$. So, χ has to be computed at each step,

$$\chi(\theta^c, \theta^p) = \min\left\{\left[\frac{f(\theta^p) \times q(\theta^p, \theta^c)}{f(\theta^c) \times q(\theta^c, \theta^p)}\right], 1\right\}. \tag{2.34}$$

Denoting by I the total number of iterations, the MH sampling proceeds in the following steps,

1. Specify the initial values of the parameters $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_m^{(0)})$.

2. Repeat for $i = 1, ..., I$

- Draw a proposal value θ^p from $q(\theta^{(i-1)},)$;
- Draw a sample $u^{(i)}$ from the uniform distribution $U(0, 1)$;
- Let

$$\theta^{(i)} = \begin{cases} \theta^{(p)} & \text{if } u^{(i)} \leq \chi(\theta^{(i-1)}, \theta^p) \\ \theta^{(i-1)} & \text{otherwise} \end{cases}. \tag{2.35}$$

3. Return the values $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_m^{(i)})$.

After a burn in period, $\theta^{(i)}$ can be considered as a sample from the target distribution $f(\theta)$. The algorithm works better if the proposal density is close to the target distribution.

Gibbs sampler The Gibbs sampling algorithm is a special case of the Metropolis-Hastings algorithm.The Gibbs sampler is a simple and powerful conditional sampling technique. In this method, the parameter vector θ is grouped into m different blocks, $\theta = (\theta_1, \theta_2, ..., \theta_m)$. Each block is then sampled separately conditioned on the remaining blocks. An attractive characteristic of Gibbs sampler is that at each iteration it uses the full conditional distributions $f(\theta_k|\theta_1, ..., \theta_{k1}, \theta_{k+1}, ..., \theta_m)$ as proposal densities for $\theta_k, k = 1, 2, ..., m$. The main advantage of using full conditionals for construction of Markov chain moves is that at any of the sampling steps no rejection is incurred. The basic algorithm of Gibbs sampler is as follows:

1. Specify the initial values of the parameters $\theta^{(0)} = (\theta_1^{(0)}, ..., \theta_m^{(0)})$.

2. Repeat for $i = i, ..., I$

- Generate $\theta_1^{(i)}$ from the full conditional $f(\cdot|\theta_2^{(i-1)}, \theta_3^{(i-1)}, ..., \theta_m^{(i-1)})$;
- Generate $\theta_2^{(i)}$ from the full conditional $f(\cdot|\theta_1^{(i)}, \theta_3^{(i-1)}, ..., \theta_m^{(i-1)})$;

$\vdots$

- Generate $\theta_m^{(i)}$ from the full conditional $f(\cdot|\theta_1^{(i)}, \theta_2^{(i)}, ..., \theta_i^{(m-1)})$.

3. Return the values $\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_m^{(i)}$.

The sequence $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_m^{(i)})$ then comprises a Markov chain with stationary distribution f. If the full conditionals are standard distributions, sampling from them is straightforward. $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_m^{(i)})$ can be considered as a sample from the target distribution f.

2.4 Missing data

2.4.1 Missing data mechanism

To decide how to handle missing data, it is helpful to know why they are missing. We review three general missingness mechanisms, moving from the simplest to the most general (Little and Rubin 1987; Little 1993).

2.4.1.1 Missing completely at random (MCAR)

If the missingness does not depends on observed as well as unobserved observations then this type of missingness is known as missing completely at random. In this type of missingness, probability of missingness is same for all the observations. For example, if answering a question depends on the show up of head after tossing a fair coin, then missingness of that answer is completely random.

2.4.1.2 Missing at random (MAR)

If the missingness of an observation only depends on observed observations then this type of missingness is known as missing at random. In MAR, probability of missingness depends only on available observations, not on unobserved observations. For example, missing information about age or income maybe depends on the other available information.

2.4.1.3 Missing not at random (MNAR)

If the missingness depends on observed as well as unobserved observations then this type of missingness is known as missing not at random. In MNAR, the probability of missingness depends on both the observed and unobserved observations. Dropouts in the medical studies can be a good example of the MNAR. A person in a study may not like the previous results and may be worried about the future results of the

study and drop out.

2.5 Methods to deal with missing data

2.5.1 Deletion methods

Many missing data approaches simplify the problem by throwing away data. We discuss below how these two simple approaches to delete missing data.

2.5.1.1 Listwise deletion-complete case analysis

A direct approach to missing data is to exclude them. Complete case analysis only considers those individuals or subjects for whom all required informations are available. In this method subjects having one or more missing information would be discarded from the analysis. This method has some advantages like, any standard statistical software can be used for the analysis and interpretations of the results will be very straightforward. If the number of missing observations are quite high in the data set then deletion of the data may loss some important features of the data and the small sample size of the data may not allow the sustainable analysis. Complete case analysis is sometimes good or consistent under MCAR mechanism but it does not work well for the MAR mechanism. One disadvantage of the method is that estimates may be biased if the units with missing values differ systematically from the completely observed cases.

2.5.1.2 Pairwise deletion-available case analysis

Another simple approach is available case analysis, which uses all available data to estimate parameters of the model. This approach is better than the complete case analysis due to considering all informations possible in the analysis comparing to the complete cases.

2.5.2 Imputation methods

Instead of deleting the subjects with missing observations, it is possible to keep all the subjects in the analysis by imputing the missing observations in the data set.

2.5.2.1 Random imputation of a single variable

Usually imputation of the missing observations is carried out by substituting the values based on the available or observed data. One of the basic advantages of this method can be specified that this method uses the complete set of data for the analysis and the missing observations are based on the available informations. Based on the way of using the observed values in imputation, imputation based techniques can be subdivided into a few categories.

Last value carried forward imputation In this method of imputation, last observed value of a subject is carried over the next missing observation. This method can be used in monotone as well as nonmonotone settings of missingness. In this technique, missing observations are substituted by the same subject's last observed information and it is assumed that the last condition would continue for the next unobserved measurements. This assumption is very strong and often does not work out. This method is sometimes used in clinical study and is found that it often fails to give the unbiased estimates. Though this method is not good but it helps to understand the pattern of the observations over time.

Imputation by related observation Sometimes related observations played a good role in case imputing the missing values. It may happen in a study that the mother's age and educational status for a child is missing then father's information can be used to fill the mother's missing information. Sometimes missing information about income can be filled by the income of another person doing the same kind of job.

Imputation by unconditional mean In this type of imputation procedure, missing value of a subject will replaced by the average of the available information of the same variable but from different subjects. So, in this technique the available observations of the subject for which the imputation is occurring will not use.

Imputation by conditional mean This approach of imputation was discussed by Buck (1960) and Little and Rubin (1987). Following Molenberghs et al. (2004) conditional mean imputation can be explained by considering a single multivariate normal sample. In this example, the mean and the covariance matrix was calculated from the complete case of the data in the first step, and then in the second step, informations from the first step were used to calculate the conditional mean from an regression of missing values of a subject conditional on the observed observations. Conditional mean from the second step was used to substitute the missing value.

Hot deck imputation Hot deck imputation procedure uses the similar kind of responding units from the sample to substitute the missing observations. This technique is one of the commonly used techniques. For example, if the information about the total number of persons in a household is missing then that information would be replaced by the similar kinds of household of that area.

Cold deck imputation In this imputation technique, missing observation will replace by a constant value from the external sources like previous survey or study. Replacing missing values by using the cold deck technique may not give the good statistical inference because the conditions of the current and the previous survey may not be the same.

2.5.2.2 Imputation of several missing variables

In the multiple imputation technique (Rubin 1978, 1987) missing values are replaced by more than one values. This technique considers the uncertainty raised due to estimating the missing values. This is a kind of modeling technique which produced the data that maintains the overall variability of the population. This technique also helps to calculate the variance of estimates. Data obtained from this technique also keep the relationship with the existing variables.

2.5.3 Model-based methods

Longitudinal data consist of time sequence of measurements on several subjects. Longitudinal data frequently involve some missing values. Subjects may withdraw from the study prematurely resulting in a dropout pattern or they may missed some occasions with an intermittent pattern. Little and Rubin (1987) introduce different mechanisms for missing values. The missing values are termed non-ignorable if the probability of being missed depends on the unobserved measurements. In this case a model is needed for both the observed and missing data for unbiased inference. In the model based technique, a model is set up for the partially missing data and the inference of the model is done on the basis of likelihood under the model. For estimating the parameters in the likelihood there exist several estimation techniques, such as maximum likelihood. Though model based techniques are not so easy to implement for all kinds of data set, this technique gives the better results than other techniques. There are quite a few ways to apply the model based approach to the missing data analysis. We will discuss the basic idea very briefly about two most frequently used model based approaches for the nonignorable missing data. Two models can be distinguished based on the factorization of the joint likelihood of response and the missing data indicator variable. One is the selection-model and the other one is the pattern mixture model. Following Little and Rubin (1987), these

two models are based on two different frameworks of the joint distribution of the response, Y and the missingness indicator variable, R and they can be expressed as

selection-model:

$$f(Y, R|X, Z, \omega_1, \omega_2) = f(Y|X, Z, \omega_1) f(R|Y, X, \omega_2),$$

pattern mixture model:

$$f(Y, R|X, Z, \psi_1, \psi_2) = f(Y|R, X, Z, \psi_1) f(R|X, \psi_2),$$

where X, Z be the matrix of covariates of the fixed and the random effects respectively, ω's and ψ's represent the parameters of the specific parts of the model and the missingness indicator R can be defined as

$$R_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ is observed} \\ 1 & \text{otherwise} \end{cases}, \tag{2.36}$$

where i represents subjects ($i = 1, 2, ..., m$) and j indicates the occasions ($j = 1, 2, ..., n_i$) of the observations.

The first term of the selection-model indicates the distribution of response given the covariates and the second term shows the missingness indicator of the response is function of responses as well as covariates. In the pattern mixture model, responses are grouped according to the missingness patterns of the data and then these groups are used for the modelling purpose. In the pattern mixture model first part shows the distribution of response given the covariates for the groups and the second part shows the missingness patterns of the response is the function of only covariates not the responses.

In the pattern mixture modelling, subjects are divided into different groups according to their patterns of dropout and then the created groups are used to

examine the effectiveness of dropout patterns to the response variable.

2.5.4 Comparisons of different approaches for handling missing data

Missing values are very common in many fields of analysis. Handling missing values is not always straightforward. We have discussed very briefly a few existing approaches for the analysis of data with missing observations. Complete case analysis and the available case analysis only works nicely at the missing completely at random (MCAR) mechanism. Almost all the statistical software can be used easily to apply these two methods in the data set. Though the interpretation from the complete case analysis is quite straightforward, this method can loss many important features of the data set. Complete case analysis sometimes suffers from the lack of reliability of interpretation because of small sample size. In the available case analysis, although it keeps subjects having missing values, it should be better than the complete case analysis. Calculating variance component may cause problems due to not having same number of observations in all the subjects. Imputation based analysis is sometimes preferable than the complete case analysis and the available case analysis because of complete data set and not losing any subjects. Like complete case analysis, imputation based analysis can be applied by any statistical software very easily. In most of the cases, this procedure requires the MCAR mechanism which is not very common. The results from the imputation based methods are quite unreliable and it is very hard to distinguish between the situations where this method works nicely and where they do not. Method, like last value carried forward, may be very unrealistic in some settings. Very often imputation based methods need specific adjustments for acceptable point estimates and sometimes these methods are not capable to give correct precision estimators (Verbeke and Molenberghs, 2000). Model based methods are flexible, and this approach can work for large data sets

and lead to large sample estimates. These methods are sometimes hard to apply but always gives the better results for interpretations. Among selection-model and pattern mixture modelling approach, pattern mixture model is convenient to apply and easy to interpret. Most of the existing statistical software can easily work with the pattern mixture approach. Although there are a few packages available for the selection-model approach, the distributional assumption of the conditional density very often creates computational hazard in applications. Little (1993) argued that the pattern mixture model is more flexible in the situation where the data are not missing completely at random and this model shows proximity to the way how the sample survey experts consider the nonresponse situation.

Chapter 3

Bayesian Mixed Model for Semi-continuous Data

3.1 Compound Poisson model

Compound Poisson Model was first used by Revfeim (1984) to model the distribution of the total rainfall in a day, assuming that each day contains a Poisson-distributed number of events and each event provides an amount of rainfall which has an exponential distribution. Thompson (1984) applied the same model to monthly total rainfalls. The compound Poisson distribution is also widely used in actuarial science and insurance to model the distribution of the total claim amount, denoted as Y, which is the sum of a random number N of independent and identically distributed claim amounts $X_1, X_2, ..., X_N$. Although the compound Poisson Model has been widely used in the field of insurance and risk management by several researchers, the idea is new in the content of zero inflated count model in health and medical science. A compound Poisson random effect approach was proposed by Ma et al. (2009) to characterize both the extra zeros and the remaining data. They used medical and health care data to evaluate their model. Their approach incorporated multilevel

clustering. Hasan et al. (2009) also used Tweedie's compound Poisson distribution to model multilevel zero inflated longitudinal count data that can unify population-averaged and cluster-specific analyses. In the next section, our proposed model is discussed which is a compound Poisson mixed model based on the Bayesian approach for the longitudinal semi-continuous data.

3.2 Compound Poisson mixed model

In our study, we adopt a Bayesian application to the compound Poisson mixed model with longitudinal semi-continuous data.

Ma et al. (2009) have proposed a compound Poisson model to handle zero inflated count data. Their model is a multi-level random effects ZIP model. They have used Tweedie's compound Poisson distribution to characterize the excess zeros, and at the same time they consider the clustering effects in the subject level. Based on their work, we use the compound Poisson model for the longitudinal semi-continuous data. The reason of using this distribution is that it can model the extra zeros and the mixed random effects in an integral way. Conditional probability given the random effects is used to allow the mixture of these zeros and non-zero components, and follows a compound Poisson distribution. We incorporate the priors in this model to obtain the posterior estimates of the parameters.

One of the main advantages of using Bayesian statistics is that Bayesian inference can provide more intuitive and meaningful inferences using all available information. On the other hand, Bayesian computation has become much easier with the availability of powerful computational tools, such as WinBUGS or OpenBUGS. In the the Bayesian approach, the main task is to implement Bayes theorem and then to derive relevant inferences or decisions from the posterior distribution. Although in any simple problems these tasks can be done algebraically, this approach does not

usually work for even moderately complex problems, hence, simulation techniques are critical in Bayesian inferences. The model is described below.

3.2.1 Model specification

Let Y_{ij} represent the response of the ith($i = 1, 2, ..., m$) individual at time j ($j = 1, 2, ..., n$). Let $U = (U_1, ..., U_i, ..., U_m)$ be the vector of the subject-specific random effects. Given the subject-specific random effects U, the time-specific random effects of second level are represented as the vector of $V = (V_{11}, ..., V_{ij}, ..., V_{mn_m})$. To formulate the model the following three assumptions are considered:

3.2.1.1 Assumption 1

Subject-specific random effect $U_1, ..., U_i, ..., U_m$ are independently and identically distributed with mean 1 and variance σ^2. That is

$$E(U_i) = 1,$$

$$Var(U_i) = \sigma^2.$$

In our data analysis, we use Gamma distribution for the subject-specific random effects.

3.2.1.2 Assumption 2

Given the subject-specific random effect U, the time-specific random effects of second level $V_{11}, ..., V_{ij}, ..., V_{mn_m}$ have a moment structure as follows:

$$E(V_{ij}|U) = U_i,$$

$$Var(V_{ij}|U) = \tau^2 U_i$$

In our data analysis, the time-specific random effect V_{ij} given the subject-random effects U follows Gamma $\left(\frac{U_i}{\tau^2}, \frac{1}{\tau^2}\right)$ distribution with mean U_i and variance $\tau^2 U_i$.

3.2.1.3 Assumption 3

Given the subject- and time-specific random effects $W = (U, V)'$, the components of Y are conditionally independent. The conditional distribution of Y_{ij}, given W, only depends on V_{ij} and follows the Tweedie compound Poisson distribution. In notation, the Tweedie compound Poisson is,

$$Y_{ij}|W \sim Tw_q(\mu_{ij}V_{ij}, \epsilon^2 V_{ij}^{1-q}) \quad \text{where} \quad 1 < q < 2. \tag{3.1}$$

And the notation of parameter μ_{ij} is

$$\mu_{ij} = \exp(X'_{ij}\beta), \tag{3.2}$$

where $X_{ij} = (X_{ij1}, X_{ij2}, ..., X_{ijp})'$ is the vector of covariates and $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ is the regression parameter vector. The zero and non-zero components are both modelled by the compound Poisson distribution with subject- and time-specific random effects.

The probability mass at zero can be written as follows,

$$P(Y_{ij} = 0|W) = \exp\left\{\frac{1}{\epsilon^2}\left(-\frac{\mu_{ij}^{2-q}V_{ij}}{2-q}\right)\right\}, \tag{3.3}$$

The conditional mean and variance of the Tweedie's compound Poisson distri-

bution is given as

$$E(Y_{ij}|W) = \mu_{ij}V_{ij}, \tag{3.4}$$

$$Var(Y_{ij}|W) = \epsilon^2\mu_{ij}^q V_{ij}. \tag{3.5}$$

We use Bayesian software OpenBUGS and/or WinBUGS to perform our analysis . But the problem is that, the compound Poisson distribution is not listed as the standard distribution list in OpenBUGS. Subsequently, we have to use a strategy to manage the analysis in OpenBUGS. We introduce a Poisson random variable as illustrated below to obtain a compound Poisson random variable. Also we add a suitable constant such as 0.02 to ensure that the the value of the Gamma random variable is positive. Here, $Y_{ij}|W$ follows a compound Poisson distribution with $Y_{ij}|W = \sum_{k=0}^{N_{ij}} G_k$(Poisson sum of Gamma), where G_k is independent Gamma(α, β) random variable and G_k is independent of N_{ij}. N_{ij} is the number of G_k, which follows the Poisson(φ) distribution.

Therefore,

$$E(N_{ij}) = Var(N_{ij}) = \varphi, \tag{3.6}$$

with

$$E(G_k) = \frac{\alpha}{\beta}, \tag{3.7}$$

and

$$Var(G_k) = \frac{\alpha}{\beta^2}. \tag{3.8}$$

$$
\begin{aligned}
E(Y_{ij}|W) &= E\left(\sum_{k=0}^{N_{ij}} G_k\right) \\
&= E\left[E\left(\sum_{k=0}^{N_{ij}} G_k|N_{ij}\right)\right] \\
&= E[N_{ij}E(G_k)] \\
&= \varphi * \frac{\alpha}{\beta}
\end{aligned} \tag{3.9}
$$

$$
\begin{aligned}
Var(Y_{ij}|W) &= Var\left(\sum_{k=0}^{N_{ij}} G_k\right) \\
&= Var\left[E\left(\sum_{k=0}^{N_{ij}} G_k|N_{ij}\right)\right] + E\left[Var\left(\sum_{k=0}^{N_{ij}} G_k|N_{ij}\right)\right] \\
&= Var\left[N_{ij}E(G_k)\right] + E\left[N_{ij}Var(G_k)\right] \\
&= Var\left[N_{ij}\frac{\alpha}{\beta}\right] + E\left[N_{ij}\frac{\alpha}{\beta^2}\right] \\
&= \frac{\alpha^2}{\beta^2}Var[N_{ij}] + \frac{\alpha}{\beta^2}E[N_{ij}] \\
&= \frac{\alpha^2}{\beta^2}\varphi + \frac{\alpha}{\beta^2}\varphi
\end{aligned} \tag{3.10}
$$

Also, $P(N_{ij} = 0) = exp(-\varphi)$ which is equivalent to

$$
P(Y_{ij} = 0|W) = \exp\left\{\frac{1}{\epsilon^2}\left(-\frac{\mu_{ij}^{2-q}V_{ij}}{2-q}\right)\right\}, \tag{3.11}
$$

so

$$
\varphi = \frac{1}{\epsilon^2}\left(\frac{\mu_{ij}^{2-q}V_{ij}}{2-q}\right). \tag{3.12}
$$

Therefore, solving the following three equations, we can get the parameters

α, β, φ.

$$\mu_{ij} V_{ij} = \varphi * \frac{\alpha}{\beta}, \tag{3.13}$$

$$\epsilon^2 \mu_{ij}^q V_{ij} = \frac{\alpha^2}{\beta^2}\varphi + \frac{\alpha}{\beta^2}\varphi, \tag{3.14}$$

$$\varphi = \frac{1}{\epsilon^2}\left(\frac{\mu_{ij}^{2-q} V_{ij}}{2-q}\right). \tag{3.15}$$

The solutions are

$$\alpha = \frac{2-q}{q-1},$$

$$\beta = \frac{\mu_{ij}^{1-q}}{\epsilon^2(q-1)},$$

$$\varphi = \frac{1}{\epsilon^2}\left(\frac{\mu_{ij}^{2-q} V_{ij}}{2-q}\right).$$

3.2.2 Prior specification

To incorporate Bayesian approach in the model, the priors of the parameters have to be specified. We have considered all the regression parameters $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_k)'$ to be normally distributed with mean 0 and known variance 100. The priors for the random effects τ, σ and ε are considered as uniform distribution. The limits for the uniform priors are considered to be 0.001 and 10.

3.2.3 Posterior computation

The posterior inference is obtained from the relationship as follows,

$$\text{posterior} \propto \text{prior} \times \text{likelihood}. \tag{3.16}$$

As mentioned in section 3.1, the priors for the random terms, such as regression

parameters and the random effects parameters, are incorporated with the likelihood to obtain posterior inference. In the next chapter, the Bayesian compound Poisson mixed model is used to analyse two different semi-continuous longitudinal data sets.

3.2.4 Selection model to deal with missing data

In longitudinal studies, data are collected on a group of individuals over a period of time and these data commonly contains missing values. These missing values may be due to early withdrawal of some subjects from the study or some values for some subjects may be unavailable (intermittent setting). Assuming that this missingness follows convenient random patterns may not be realistic, so there is much interest in methods for analysing incomplete longitudinal data which allow the incorporation of more realistic assumptions about the missing data mechanism.

The selection model specifies both a model for the longitudinal outcome and a model for the missingness. Assume that there are m subjects participating in the study and the intended measurements for the ith subject are n_i. Let Y_{ij} represent the jth measurement, $j = 1, ..., n_i$, on the ith subject, $i = 1, ..., m$. Let Y_i be an $n_i \times 1$ vector containing the responses that would be obtained for the ith subject, if there were no missing values. Assume that the observed and missing components of Y_i are denoted as $Y_{i,obs}$ and $Y_{i,mis}$ respectively.

Suppose R is the missing data indicator for the response variable, then notation can be written as

$$R_{ij} = \begin{cases} 0 & y_{ij} \text{ observed} \\ 1 & y_{ij} \text{ missing} \end{cases}. \tag{3.17}$$

Let R_i be a vector of missingness indicators. For a particular realization of (Y_i, R_i), each element of R_i takes the value zero if the corresponding value of Y_i is observed and the value one if the corresponding value of Y_i is missing. The

response vector of the ith subject of jth measurement is Y_{ij}, given the subject- and time-specific random effect, the conditional distribution is assumed to follow the compound Poisson distribution with mean $\mu_{ij}V_{ij}$ and variance $\epsilon^2\mu_{ij}^q V_{ij}$.

Therefore we can model semi-continuous data with missing values as follows: set up a compound Poisson mixed model for all the data introduced in Section 1 of this chapter and then apply a selection-model approach to analyse the missing values. A key assumption underlying the selection model approach is the correct specification of the response distribution and the missing data mechanism.

Chapter 4

Data Analysis

An intervention study was conducted to identify salient parent and adolescent psychosocial factors related to somatic symptoms in adolescents in New York. Adolescents and young adults living with parents with HIV (PWH), provided an opportunity to examine risk factors for the development and maintenance of somatic symptoms. Families living with PWH often endure stresses from poverty, residential instability, substance abuse, physical deterioration, emotional distress or early parental death.The purpose of their study is to evaluate an intervention designed to reduce risk in children of PWH.

Dental fluorosis prevalence has increased in the United States, Canada, and other nations due to the widespread availability of fluoride in many forms, with fluoride ingestion during the first three years of life appearing most critical in fluorosis etiology. Repeated responses to separate series of questions about water intake, use of fluoride dentifrice, and use of fluoride supplements were collected by questionnaire as part of the longitudinal Iowa Fluoride Study and used to estimate fluoride intake. Estimated intake is reported by source and combined at different ages. Effects of subject age and other covariates such as race, gender, and family income on fluoride intake were assessed in the model.

In this chapter, we focus on the analysis of the brief symptoms inventory data and fluoride intake data using the proposed Tweedie's compound Poisson mixed model. we also analyse the data by using a model-selection approach to deal with the missing data.

4.1 Brief symptoms inventory (BSI) data

The brief symptoms inventory (BSI) data were collected from 409 adolescents (young adults or teenagers) whose parents are independently HIV infected (Bursch, 2008). These adolescents were followed from 1995 to 2002. The somatic symptoms were assessed at baseline and in each follow-up, for up to 72 months, using the average somatization subscale score of the Brief Symptom Inventory and their global severity index (GSI) scores; these scores were recorded in each visit. The BSI data are longitudinal semi-continuous as the responses GSI are recoded repeatedly over time with more than 21% of the responses being exact zeros. To illustrate the proposed methodology, for data analysis we use a subset of the brief symptoms inventory (BSI) data due to the significant number of missing values. In the subset, we considerer 57 adolescents from 57 independent parents and their measurements at 0, 3, 6, 9, 12, 15 and 18 months. At the beginning of the study, the parents and adolescents were randomly assigned either to the intervention or to the standard care. The main objective of the study is to evaluate the effectiveness of the intervention which was designed based on social learning theory, as compared to the standard care.

Let Y_{ij} be the response recorded at the jth time point ($j = 1, 2, 3, 4, 5, 6$ and 7 representing responses recorded at $0, 3, 6, 9, 12, 15$ and 18 months) of the ith adolescent ($i = 1, ..., 57$). Figure 4.1 is the histograms of the global severity index (GSI) scores at different measurement times in the study. The BSI data are longitudinal semi-continuous, as the responses are recoded repeatedly over time. The histograms

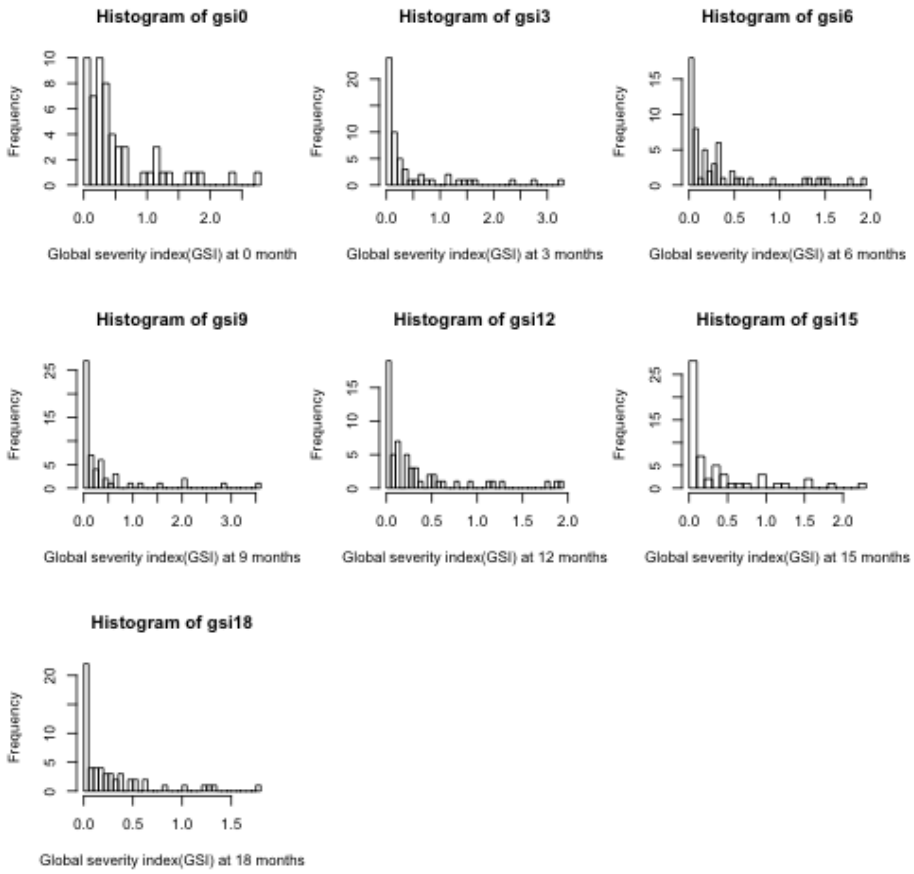

Figure 4.1: Histograms of the global severity index(GSI) scores of the brief symptoms inventory(BSI) Data

of GSI shows that the responses are right skewed with a substantial proportion of zeros. To overcome this problem we use the proposed compound Poisson mixed model for analyzing the right skewed semi-continuous data. Let U_i be adolescent-specific random effect, and let V_{ij} be the time-specific random effect as specified in assumptions 1 and 2 discussed in Chapter 3, respectively. We consider gender, age, time in study, treatment, Hispanic status, parents age and parents gender as covariates. For the purpose of analysis, the covariate gender was coded as 1 for female and 0 for male; Hispanic status was coded as 1, if adolescents is Hispanic, and 0 if adolescent is not Hispanic. The collection of regression parameters is denoted as $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)'$. The mean of the response of the ith adolescents

recorded at time j is defined as follows: $\mu_{ij} = \exp(x'_{ij}\beta)$, where

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{ Age} + \beta_2 \text{ Gender} + \beta_3 \text{ Time in study} + \beta_4 \text{ Treatment}$$

$$+\beta_5 \text{ Hispanic} + \beta_6 \text{ Parent's age} + \beta_7 \text{ Parent's gender}$$

The model is fitted using the WinBUGS software. We ran two different MCMC chains with dispersed initial values. In order to break the dependence between draws in the Markov chain, some have suggested only keeping every dth draw of the chain. This is known as thinning, which helps to get closer to i.i.d. draws and save memory since we only store a fraction of the draws. We ran two different MCMC chains with dispersed initial values and retained every 100th draw to reduce autocorrelation. To obtain the convergence, 20000 iterations have been performed.

Table 4.1: Estimates, SDs, credible intervals and MC errors for BSI data

Coefficient	Estimate	SD	2.5%	97.5%	MC error
Intercept	3.8150	1.0890	1.6980	5.9700	0.0108
Age	-0.0324	0.0075	-0.0472	-0.0176	0.0000
Gender	-0.0057	0.0510	-0.1061	0.0941	0.0004
Time in study	0.6387	0.2013	0.2443	1.0340	0.0011
Treatment	-0.3464	0.1994	-0.7381	0.0446	0.0011
Hispanic	0.3473	0.2126	-0.0709	0.7636	0.0014
Parent's age	0.0161	0.0211	-0.0252	0.0574	0.0002
Parent's gender	1.2780	0.2756	0.7281	1.8110	0.0017
τ	0.0789	0.0607	0.0031	0.2259	0.0003
σ	0.9527	0.1081	0.7619	1.1850	0.0005
ϵ	4.2170	0.2135	3.8190	4.6570	0.0010

The results presented in Table 4.1 which reports the estimates, standard deviation, 95% credible intervals and MC error of the posterior means for all parameters. The 95% credible intervals are based on the empirical 2.5% and 97.5% quantiles at index parameter p=1.5. The notion of credible interval facilitates a common-sense interpretation of statistical conclusions; a credible interval for an unknown quantity

of interest can be regarded as having a high probability of containing the unknown quantity, in contrast to a frequentist confidence interval which may be interpreted only in relation to a sequence of similar inferences that might be made in repeated practice. There is a relationship between credible intervals and significance tests. Specifically, if the effect of one covariate is significant then the 95% credible interval will not contain 0. MC error is used to evaluate the computational accuracy of the mean. Values of MC error smaller than the difference between 1% and 5% of the standard deviation of posterior estimates indicates more accurate results. In our analysis, we found all of our estimates have very small MC error which indicates the more accurate results of our estimates.

Our results presented in Table 4.1 indicates that adolescent's Age, Time in study and Parent's gender have significant effect on GSI scores of the BSI data. Compared to older adolescents, the younger adolescents tend to report more somatic symptoms. Parent's gender has positive effect, showing that youth with mother report more somatic symptoms. The positive trend of Time in study for GSI indicates that the youth tend to have more GSI scores as they become older. The covariate Treatment does not appear to have any significant effect on GSI. This indicates that the intervention is not effective enough to improve adolescents' psychiatric condition.

The estimates of dispersion and correlation parameters τ, σ, ϵ for the BSI data are 0.0789, 0.9527 and 4.2170 respectively. The estimates of the variance parameters τ, σ and ϵ indicate that there is additional variation in the responses beyond what can be characterized by the random effect. The value of τ and σ indicates that the variation of each subject and the responses at different measurement times, respectively.

Figures 4.2 and 4.3 show the trace plots of all the parameters of the model for the BSI data. Trace plots are used to verify the convergence of the parameters' inference. Since no patterns or irregularities are observed in all the tracing plots, the

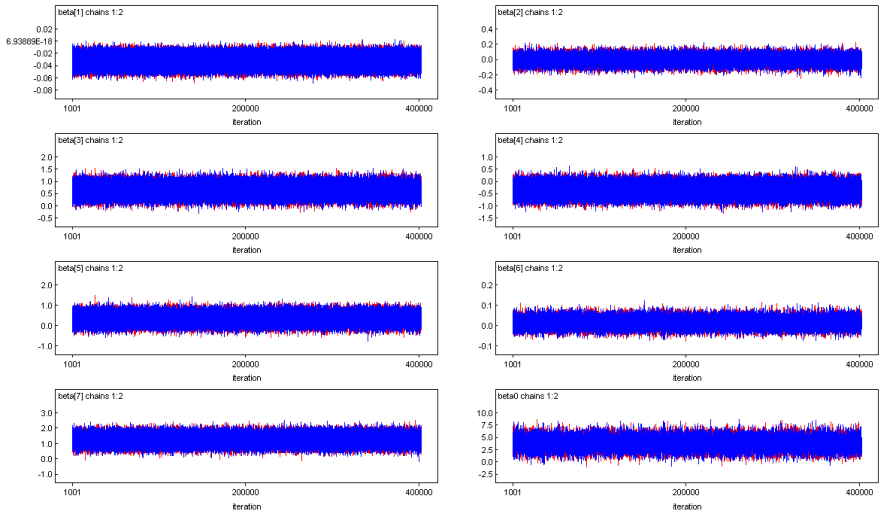

Figure 4.2: Trace plots of the parameters based on two MCMC chains for BSI data

convergence can be assumed for all of the parameters.

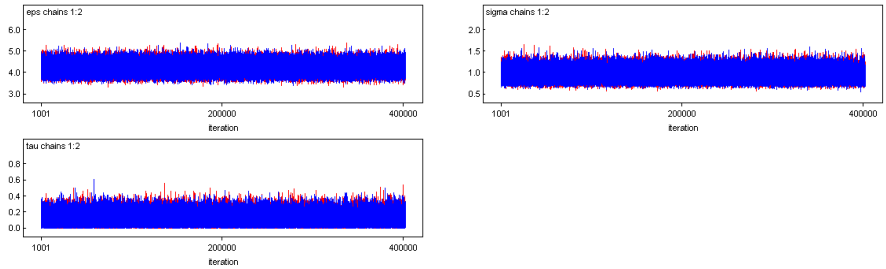

Figure 4.3: Trace plots of the parameters based on two MCMC chains for BSI data(cont.)

The autocorrelation plots for the parameters of the BSI data are organized in Figures 4.4. To minimize the autocorrelation among the responses, every 100th iteration is considered in our analysis. We observed that autocorrelation for all parameters become low only after considering lag of 20 from the original autocorrelation plots. However, the autocorrelation is comparatively higher in β_2, β_6 which are the coefficients of gender and parent's age for the compound Poisson model respectively.

Figures 4.5 represents the quantile plots of all the parameters of the model for the BSI data. Quantile plot is a plot of mean and 95% credible interval against

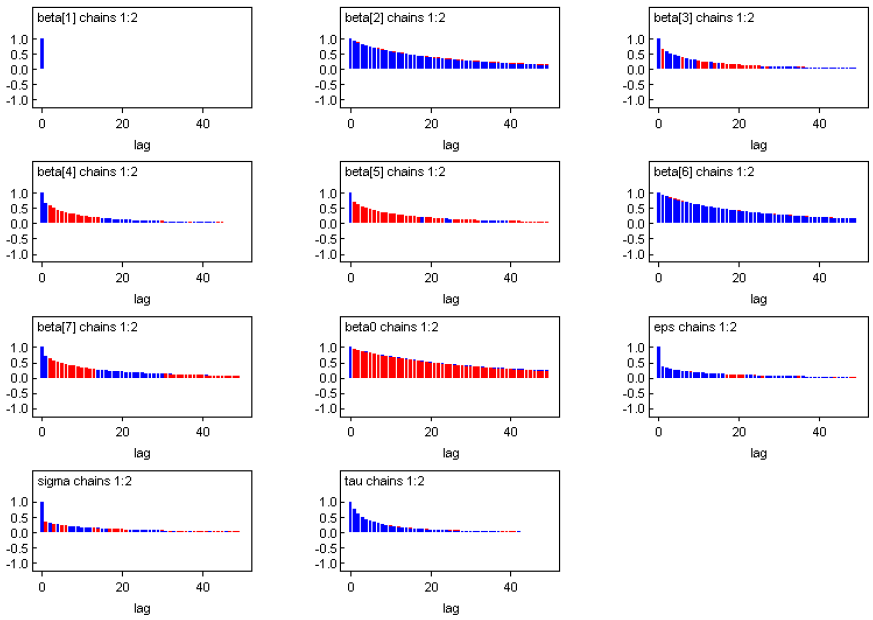

Figure 4.4: Auto-correlation plots of the parameters based on two MCMC chains for BSI data

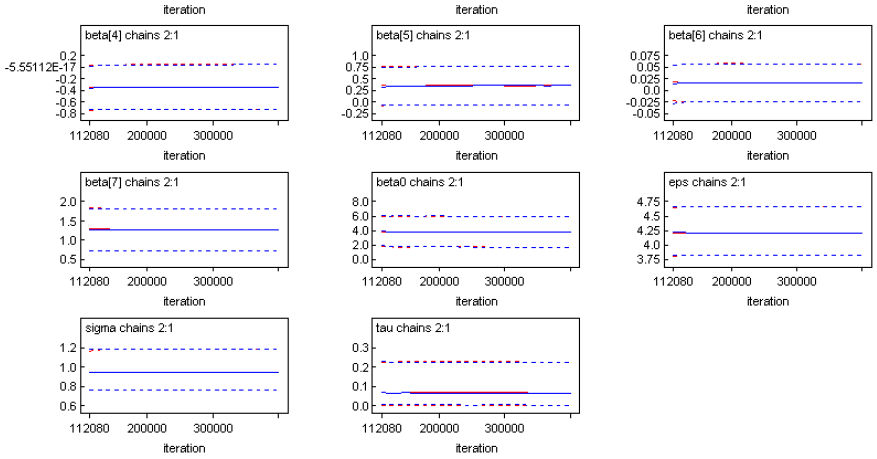

Figure 4.5: Quantile plots of the parameters based on two MCMC chains for BSI data

iteration number. A quantile plot is also used to identify the convergence of the parameters. Since no patterns or irregularities are observed in any of the plots, the convergence can be assumed for all of the parameters.

Figures 4.6 shows the Gelman Rubin statistic plot for the convergence check.

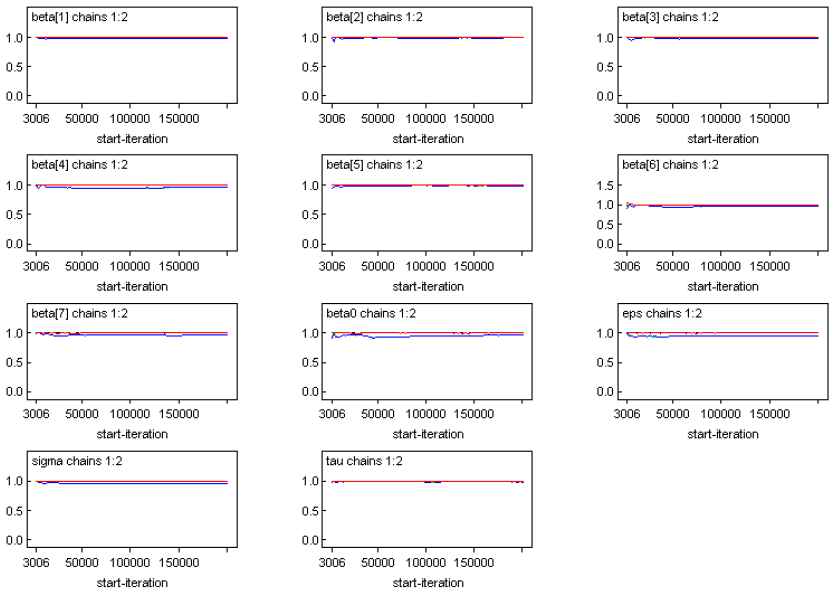

Figure 4.6: Gelman Rubin statistics plots of the parameters based on two MCMC chains for BSI data

The Gelman Rubin statistic compares the variance within and between multiple chains. As we know, this plot requires multiple parallel chains starting with overdispersed initial values; we ran two chains with different initial values. The multivariate potential scale reduction factor is plotted against the number of iterations to simultaneously assess the convergence of all model unknowns. In particular, the Gelman Rubin plot shows the following elements versus the number of iterations:

- Width of central 80% interval constructed from pooled runs (plotted in green)
- Average width of 80% intervals constructed from each run (plotted in blue)

  $\text{ratioR} = \frac{\text{pooled}}{\text{within}}$. (in red)

In our analysis, we found all the parameters in the BSI data shows a steady convergence pattern after iterations as the lines are close to 1.

4.2 Fluoride intake data

4.2.1 Complete case

In this section, we analyse longitudinal semi-continuous fluoride intake data (Davis, 2002) using the proposed compound Poisson mixed model. In a longitudinal study of fluoride intake, infants were enrolled at birth and followed over time. During the first nine months of the study, total fluoride intake (mg/kg body weight) was assessed at months 1.5, 3, 6, and 9, together with their covariates information, such as sex, race, mother's age and education level, household income, and the number of children in the family. The data set contains 4151 observations from 1363 infants. In some cases, the information of fluoride was missing or not recorded. Firstly, we just consider 596 infants in the data set which was recorded completely for all four measurement times. Gamma distribution is used to model the subject- and time-specific random effects. The responses given two level random effects follow a Compound Poisson distribution. In other words, the conditional probability are the Poisson sum of Gamma distributions, the values of the index parameter of Tweedie's compound Poisson distribution for this situation should lie between $1 < p < 2$. In our analysis, we have considered the values of p as 1.5. The models were fitted using the WinBUGS (Spiegelhalter et al., 2003) software. We ran two different MCMC chains with dispersed initial values. We also retained every 20th draw to reduce autocorrelation.

Let Y_{ij} be the jth ($j = 1, 2, 3, 4$ for the response recorded at $1.5, 3, 6, 9$ months) observation of the ith infant ($i = 1, ..., 596$) since the infant is 1.5 months as the baseline age. Let U_i be infant-specific random effect, and V_{ij} be the time-specific random effect as specified in Assumptions 1 and 2, respectively. Figure 4.7 shows that the Fluoride intake data are longitudinal semi-continuous data as the responses are recoded repeatedly over time and about 15.5% of the responses are exact zeros. The

figure also indicates the distribution of the semi-continuous data are right skewed, and the responses should not be Poisson distribution since the large amount of zeros. To overcome this problem, we use the proposed compound Poisson mixed model approach for analyzing fluoride intake data.

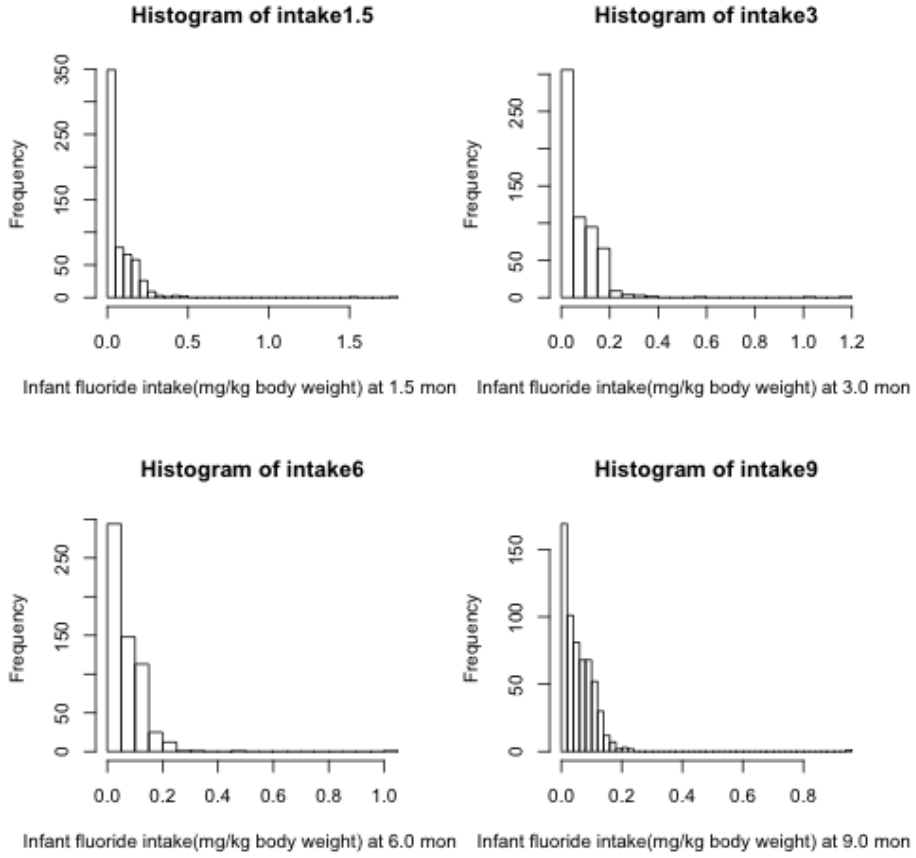

Figure 4.7: Histograms of infant fluoride intake data (original data)

In the final model, we included age, age^2 (quadratic evaluation of age), sex (1 for male, 0 for female), race (1 for white, 0 nonwhite), twokids (1 for two kids at home, 0 otherwise), threekids (1 for three or more kids at home, 0 otherwise), income (1 for annual household income of $\$30,000$ or more, 0 otherwise) as covariates. Following assumption 3 regarding conditioning on the subject and time-specific random effects, we cam assume that Y_{ij} is compound Poisson distribution with $\mu_{ij} = \exp(x'_{ij}\beta)$, where

$$\log(\mu_{ij}) = \beta_0 + \beta_1 \text{ Age} + \beta_2 \ Age^2 + \beta_3 \text{ Sex} + \beta_4 \text{ Race}$$

$$+\beta_5 \text{ Twokids} + \beta_6 \text{ Threekids} + \beta_7 \text{ Income}$$

The analysis results are presented in Table 4.2. The predictors Two kids, Three kids and Income have significant effect on the fluoride intake. The indicators for two kids and for three or more kids in a family are both negative as compared to one or no kid which indicate that infants living with more than one kid in the family tend to have a lower fluoride intake level. The predictor Income also has a negative effect: the infants whose families have more income will have a lower fluoride intake level.

The estimates of dispersion and correlation parameters τ, σ, ϵ for the Fluoride Intake data are 0.8247, 0.0338 and 1.2650, respectively. The estimates of the variance parameters τ, σ and ϵ indicate that there is additional variation in the responses beyond what can be characterized by the temporal random effects.

Table 4.2: Estimates, SDs, credible intervals and MC errors for the fluoride intake complete data

Coefficient	Estimate	SD	2.5%	97.5%	MC error
Intercept	2.1590	0.2116	1.7695	2.5960	0.0075
Age	0.0118	0.0286	-0.0441	0.0671	0.0005
Age^2	-0.0040	0.0027	-0.0091	0.0013	0.0000
Sex	-0.0277	0.0644	-0.1549	0.0969	0.0020
Race	-0.1347	0.2052	-0.5662	0.2374	0.0075
Twokids	-0.1877	0.0729	-0.3281	-0.0430	0.0023
Threekids	-0.2549	0.0818	-0.4219	-0.0979	0.0025
Income	-0.1977	0.0681	-0.3333	-0.0643	0.0020
σ	0.8247	0.0285	0.7702	0.8818	0.0005
τ	0.0338	0.0244	0.0026	0.0920	0.0008
ϵ	1.2650	0.0237	1.2190	1.3120	0.0005

Figure 4.8 and 4.9 show the trace plots of all the parameters of the model for the Fluoride Intake data. Trace plots are used to verify the convergence of the parameters' inference. Since no patterns or irregularities are observed in all the

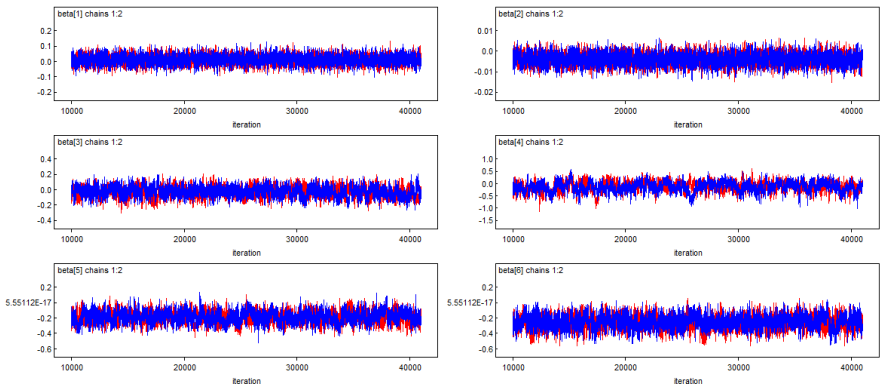

Figure 4.8: Trace plots of the parameters based on two MCMC chains for fluoride intake complete data

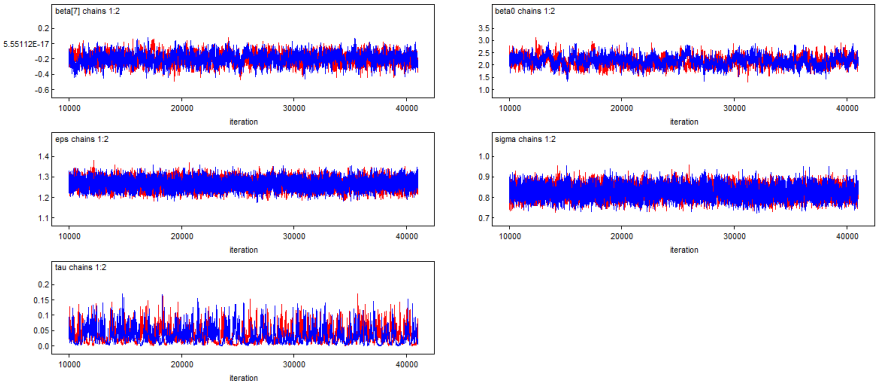

Figure 4.9: Trace plots of the parameters based on two MCMC chains for fluoride intake complete data (Continued)

tracing plots, the convergence can be assumed for all of the parameters.

The autocorrelation plots for the parameters of the fluoride intake complete data are organized in Figures 4.10. To minimize the autocorrelation among the responses, every 50th iteration is considered in our analysis. We observed that autocorrelation for all parameters did not become low even considering a lag of 50 iterations. The plots here shows that the autocorrelation of covariances such as Sex, Twokids, Threekids become ignorable after considering to thin at 50. However, the

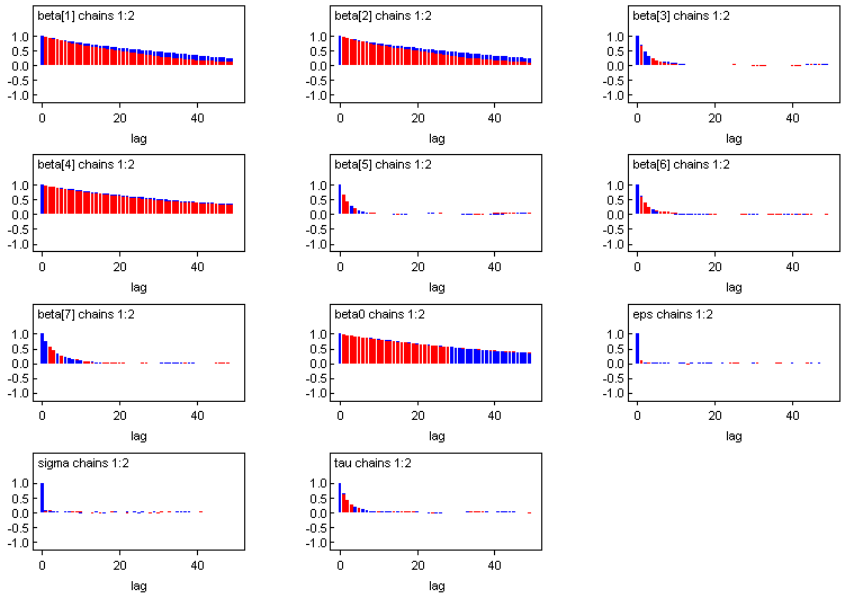

Figure 4.10: Auto-correlation plots of the parameters based on two MCMC chains for fluoride intake complete data

autocorrelation is comparatively higher in $\beta_1, \beta_2, \beta_4$ which are the coefficients of Age, Age^2 and Race for the compound Poisson model respectively.

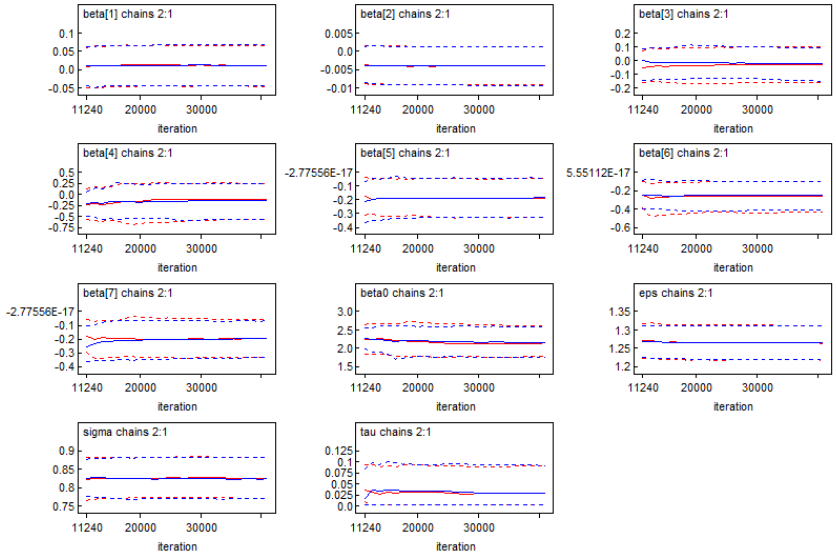

Figure 4.11: Quantile plots of the parameters based on two MCMC chains for fluoride intake complete data

Figures 4.11 represents the quantile plots of all the parameters of the model

for the fluoride intake data. Since no patterns or irregularities are observed in any of the plots, the convergence can be assumed for all of the parameters.

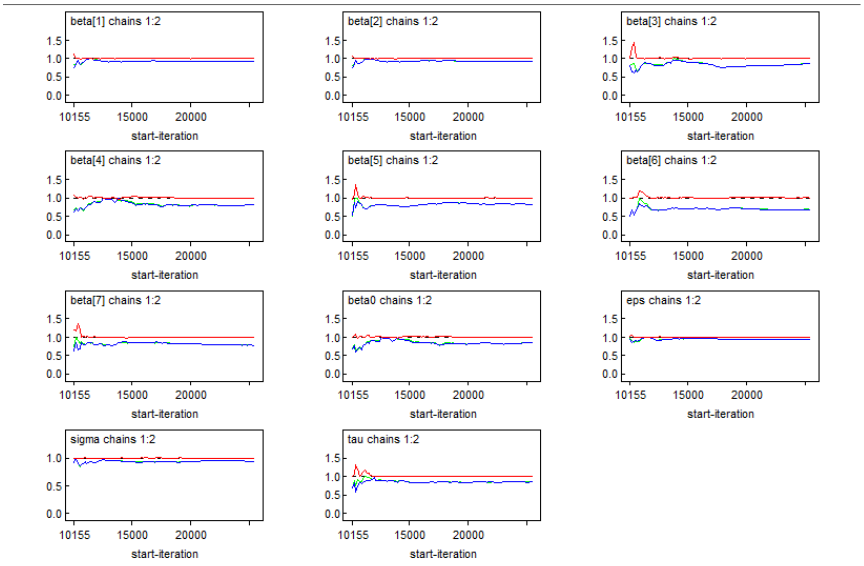

Figure 4.12: Gelman Rubin statistics plots of the parameters based on two MCMC chains for fluoride intake complete data

Figures 4.12 shows the Gelman Rubin statistic plot for the convergence check. In our analysis, we found all the parameters in the Fluoride Intake data shows a steady convergence pattern after iterations as the lines are close to 1.

4.2.2 Selection model for the full data

The data set contains 4151 observations from 1363 infants and fluoride intake is missing at least one time in 56% of the infants, resulting in 596 (44%) complete cases. The complete case was analysed in the last section. Furthermore, we are interested in the data with missing values for the response variable fluoride intake. Since the covariance race and annual household income include missing values, we disposed the data set and eliminated the observations where race and income are missing. The records of some subjects are incomplete, which were also eliminated. A data set, including 2512 observations from 628 subjects, was used to analyse; there

are 382 zero responses and 41 missing values in this data set. Age, race, sex, the number of kids and income were also recorded for each individual at the beginning of the study; these were used as covariates. For the purpose of analysis, the covariate sex was coded as 1 for male and 0 for female; race was coded as 1 for white and 1 for otherwise; and annual household income was coded as 1 for \$30,000 or more and 0 for otherwise.

Table 4.3: Variables included in the data set from the longitudinal study of fluoride intake

Col	Variable	Commets
1	ID	subject identifier
2	age	months; values are 1.5,3,6,9
3	sex	1=male, 0=female
4	race	1=white, 0= non-white, .= missing
5	mother's age	1= 30+ years, 0= less than 30 years
6	mother's education 1	1 = high school graduate (but not college graduate) 0 = otherwise
7	mother's education 2	1 = college graduate, 0 = otherwise
8	first baby in the family	1= yes, 0 = no
9	number of children in the family	1= two children at home, 0= otherwise
10	number of children in the family	1= three or more, 0= otherwise
11	annual household income	1= \$30,000, 0 = less than \$30,000, .=missing
12	total fluoride intake	mg/kg (. = missing)

Table 4.3 shows that one predictor income includes missing values, we may eliminate the subject whose family income is missing in order to guarantee that only response variable includs the missing values. Therefore, the full data set we used to analyse includes 2512 observations, of which 124 response fluoride intake were missing.

Figure 4.7 indicates that the distribution of the responses should not be Gaussian as commonly used for the continuous part of the two parts model. To overcome this problem, we use the proposed compound Poisson mixed model approach for analyzing fluoride intake data.

Firstly, we ignore the missing data and impute the missing values as mean of the response; we consider the full model with all the covariates, and then refine the model by dropping covariates simultaneously as some covariates are insignificant; The result shows that the covariates twokids, threekids and income are significant, therefore, I only keep these three covariates in the model. Since the time-specific

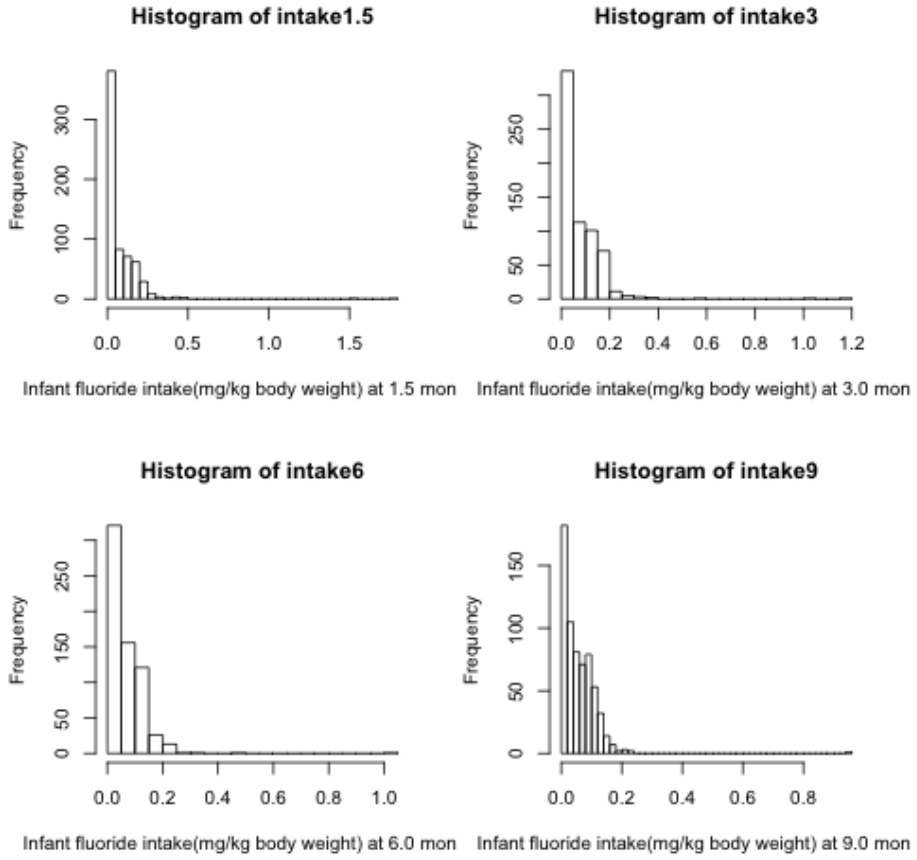

Figure 4.13: Histograms of infant fluoride intake data with missing

random effect can be ignorable, I eliminated the second assumption in the compound Poisson mixed model.

The model for this data set excluding missing values is considered as follows:

$$\log(\mu_{ij}) = {\beta_0}^* + {\beta_1}^* \text{ twokids} + {\beta_2}^* \text{ threekids} + {\beta_3}^* \text{ income}$$

The aim of analysing this data set is to identify the significant effects of the birth order and annual income on Fluoride intake amount after the infants were born, from 1.5 months to 9 months, considering the random effect at subject level.

Due to the existence of missing value in the response, one more part should be considered in the model. I added one covariate 'missing' which was coded as 0

for missing and 1 for not missing responses in the full data set. This covariate is a binary variable, therefore, the logistic regression model was used to analyse the binary response missing with covariates fluoride intake.

The logit model for this data set is considered as follows:

$$logit(p_{[i,j]}) = {\theta_0}^{**} + {\theta_1}^{**} \text{ fluoride}$$

where $p[i, j]$ is the probability that a response value is missing, and the covariate missing follows Bernoulli distribution with parameter p.

In the second part of the model, I analysed whether the covariates fluoride intake , twokids, threekids and income effects the missing value significantly.

4.2.3 Results from full data with missing values

In total, 2512 observations were observed for 618 infants at 1.5, 3, 6, 9 months. Tweedies compound Poisson distribution is used to model the subject-specific random effects. Given the random effects, the responses follow the Poisson sum of Gamma distribution and the values of the index parameter of Tweedies Compound Poisson distribution for this situation should lie between $1 < p < 2$. In our analysis, we have considered the values of p as 1.1, 1.5 and 1.9, in order to evaluate the estimates for different values of p. The models were fitted using the OpenBUGS software. We ran two different MCMC chains with dispersed initial values. We also retained every 50th draw to reduce autocorrelation.

*: the coefficients for the compound Poisson mixed model

**: the coefficients for the logistic model.

Table 4.5 reports the Estimates, SD, the 95% credible intervals and the MC error of the posterior means for all parameters. The 95% credible intervals are based on the empirical 2.5% and 97.5% quantiles at index parameter p=1.5.

The estimates of the parameters and their standard deviation for different

Table 4.4: Posterior estimates and standard deviations at different values of the index parameter for the fluoride intake data

	p=1.1		p=1.5		p=1.9	
Coefficient	Estimates	SE	Estimates	SE	Estimates	SE
Intercept*	2.2760	0.1275	2.3340	0.0855	2.0690	0.1719
Twokids*	-0.1012	0.1244	-0.1914	0.0886	-0.1415	0.1086
Threekids*	-0.0210	0.2644	-0.2078	0.1026	-0.0531	0.1532
Income*	-0.3282	0.0551	-0.0958	0.0855	-0.1082	0.0731
Intercept**	4.7290	0.5669	4.7430	0.5778	4.5990	0.6417
Fluoride**	0.0384	0.0364	0.0419	0.0586	0.0822	0.0909
σ	0.9881	0.0095	0.9862	0.0120	0.7806	0.0568
ϵ	0.9997	0.0004	0.9986	0.0014	0.8869	0.0809

Table 4.5: Estimates, SDs, credible intervals and MC errors for the fluoride intake full data with missing

Coefficient	Estimates	SD	2.5%	97.5%	MC error
Intercept*	2.0100	0.0770	1.8580	2.1610	0.0024
Twokids*	-0.1530	0.0808	-0.3111	0.0041	0.0021
Threekids*	-0.2010	0.0934	-0.3811	-0.0146	0.0025
Income*	-0.1576	0.0768	-0.3090	-0.0092	-0.0024
Intercept**	3.6510	0.3025	3.0800	4.2530	0.0055
Fluoride**	0.1151	0.0888	-0.0080	0.3264	0.0018
σ	0.9519	0.0340	0.8882	1.0220	0.0003
ϵ	1.3060	0.0257	1.2560	1.3570	0.0007

values of the index parameter p are presented in Table 4.3. It is notable that the changes of the values of p do not have any significant effect on the estimates of the parameters. There is a slight change in the values of estimates of the parameters. The result shows that threekids and income have a negative impact on the fluoride intake. The fluoride intake is not significant on the probability of missing. The amount of fluoride intake decreases with the increase of birth order and income. Babies of people with higher household income tend to have less fluoride intake than the babies of people with lower household income. The negative trend of birth order for fluoride intake indicates that the babies tend to take less fluoride than older siblings.

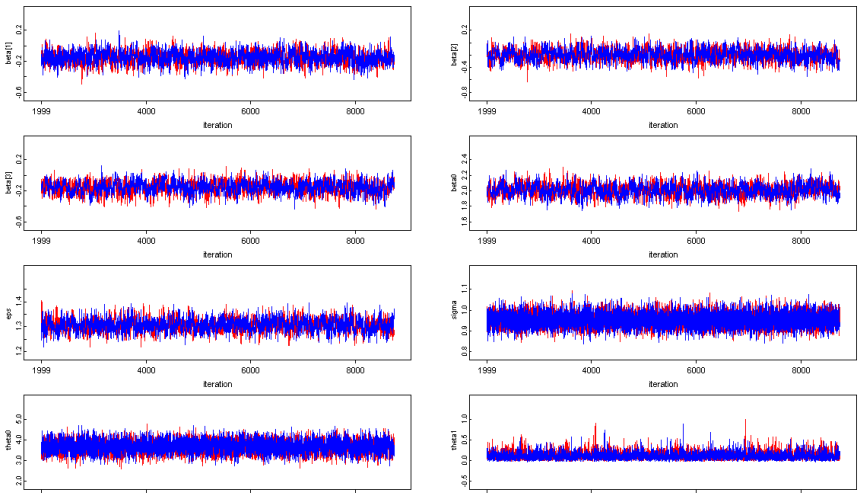

Figure 4.14: Trace plots of the parameters based on two MCMC chains for fluoride intake full data

Figures 4.14 shows the trace plots of all the parameters of the model for the fluoride intake data. Any irregular patten in a trace plot indicates lack of convergence. Since no patterns or irregularities are observed in any of the plots, the convergence can be achieved for all of the parameters.

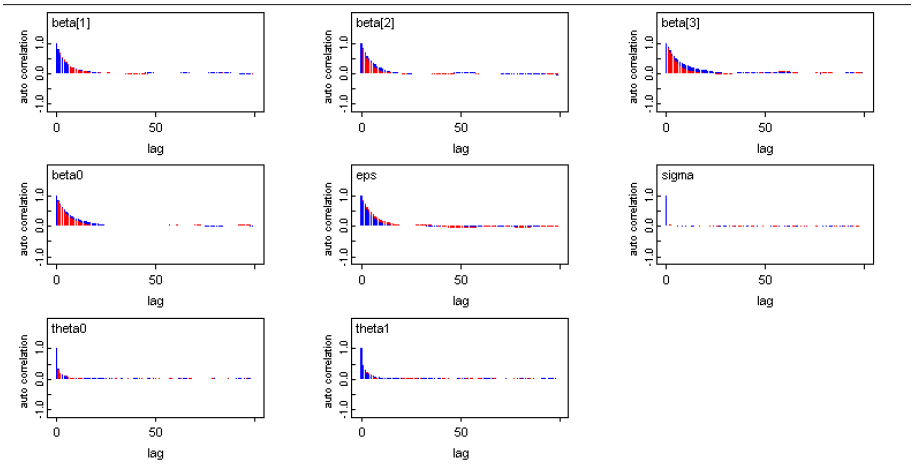

Figure 4.15: Auto-correlation plots of the parameters based on two MCMC chains for fluoride intake full data

The autocorrelation plots for the parameters of the fluoride intake full data are organized in Figures 4.15. We observe that autocorrelation for all parameters

become low only after considering a lag of 50 iterations from the original autocorrelation plots. To minimize the autocorrelation among the responses, every 50th iteration is considered in our analysis. The plots here show that the autocorrelation become ignorable after considering to thin at 50. However, the autocorrelation is comparatively higher in $\beta_1, \beta_2, \beta_3, \theta_1$ which are the coefficients of twokids, threekids and income for the compound Poisson model, and the coefficient of fluoride intake for logistic model respectively.

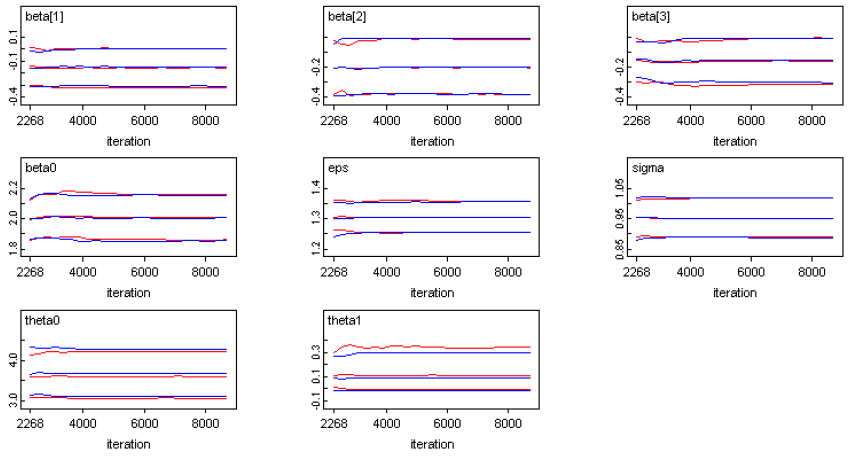

Figure 4.16: Quantile plots of the parameters based on two MCMC chains for fluoride intake full data

Figures 4.16 represents the quantile plots of all the parameters of the model for the fluoride intake data. Since no patterns or irregularities are observed in any of the plots, the convergence can be assumed for all of the parameters.

Figures 4.17 shows the Gelman Rubin statistic plot for the convergence check. The Gelman Rubin statistic compares the variance within and between multiple chains. In our analysis, we found all the parameters in the fluoride intake data shows a steady convergence pattern after iterations as the lines are close to 1.

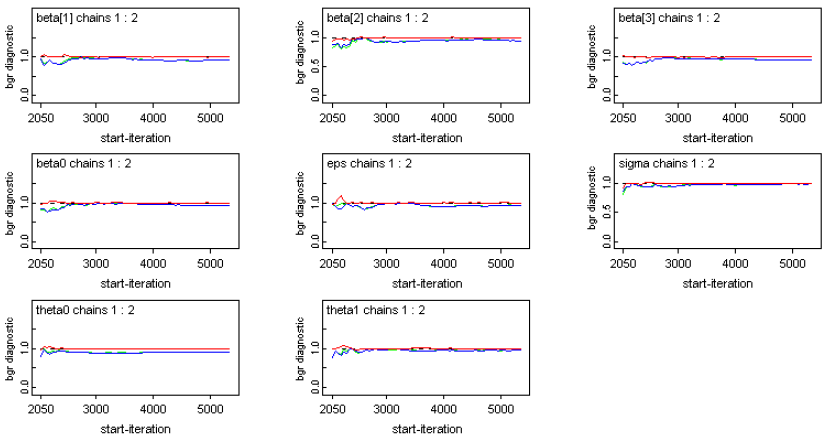

Figure 4.17: Gelman Rubin statistics plots of the parameters based on two MCMC chains for fluoride intake full data

4.3 Simulation

In this section, we examine the performance of the compound Poisson mixed model with multilevel random effects through simulation studies for semi-continuous data set. To do that we carried out simulation runs for the semi-continuous responses with missing.

4.3.1 Simulation procedure

The generation procedure is discussed below:

• We first generate 596 variates, $(u_1, ..., u_{596})$ following Gamma distribution with mean 1 and variance σ^2 .

• In the second step, we generate four variates $(v_{i1}, v_{i2}, v_{i3}, v_{i4})$ following Gamma distribution with mean u_i and variance $\tau^2 u_i$ for each u_i.

• Next step, we generate $(Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})$ for each u_i and v_j using

$$\text{Tw}_q(\mu_{ij}V_{ij}, \epsilon^2 V_{ij}^{1-q}) \quad \text{where } 1 < q < 2, \tag{4.1}$$

The conditional mean and variance of the Tweedie's compound Poisson distribution

is given as

$$E(Y_{ij}|W) = \mu_{ij}V_{ij}, \tag{4.2}$$

$$Var(Y_{ij}|W) = \epsilon^2\mu_{ij}^q V_{ij}, \tag{4.3}$$

and

$$\mu_{ij} = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2). \tag{4.4}$$

- Finally, we generate $(Z_{i1}, ...Z_{ij}, ..., Z_{i4})$ using Bern$(1, p_{[i,j]})$, where

$$\text{logit}(p_{[i,j]}) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 Y. \tag{4.5}$$

The covariates X_1, X_2 are described in the following section. We used $\beta_0 = 1, \beta_1 = 0, \beta_2 = 0, \theta_0 = -1, \theta_1 = -0.1, \theta_2 = 0.1, \theta_3 = -1, \sigma = 0.1, \tau = 0.1$ as the true parameter values.

4.3.2 Summary statistics

The simulation results given in Table 4.6 indicate that the derived result can identify estimating procedure good behaviour.

Table 4.6: Estimates, True values, SDs, credible intervals and MC errors for the generated data

Coefficient	True value	Estimates	SD	2.5%	97.5%	MC error
Intercept*	1.0000	0.5960	0.0274	0.5432	0.6501	0.0006
x1*	0.0000	-0.0272	0.0233	-0.0730	0.0183	0.0004
x2*	0.0000	-0.0313	0.0220	-0.0745	0.0126	0.0003
Intercept**	-1.0000	-0.7330	0.1678	-1.1020	-0.4344	0.0037
x1**	-0.1000	-0.0762	0.0712	-0.2129	0.0673	0.0010
x2**	0.1000	0.0520	0.0696	-0.0888	0.1873	0.0011
y**	-1.0000	-1.3420	0.3513	-2.0830	-0.7135	0.0078
σ	0.1000	0.1447	0.0335	0.1017	0.2215	0.0007
τ	0.1000	0.2656	0.1117	0.1067	0.5004	0.0053
ϵ	1.0000	1.2710	0.0461	1.1670	1.3460	0.0019

In our analysis, we found all of our estimates have a very small MC error which indicates the accuracy of our estimates. Extensive simulation results indicate that the estimates are reasonably close to true values.

4.3.2.1 Convergence issues

For data analysis using the Bayesian approach, it is imperative that we ensure convergence of the Markov Chains used before we take results from them. WinBUGS or OpenBUGS have several different methods to check for convergence. Here, we describe several of those methods. We first need to estimate the length of the burn-in period before we can take a sample from the converged chain. To estimate the burn-in, we may look at the trace plots of the chains to see whether they have mixed well. As the result of the generated data, we may check the compound Poisson mixed model for semi-continuous data and logistic model for part of missing data. All graphs come from the output of OpenBUGS. Figure 4.14 and Figure 4.15 show a segment of the trace history of parameter for iterations from 22000. As we can see, the chains appear to have mixed well by this stage.

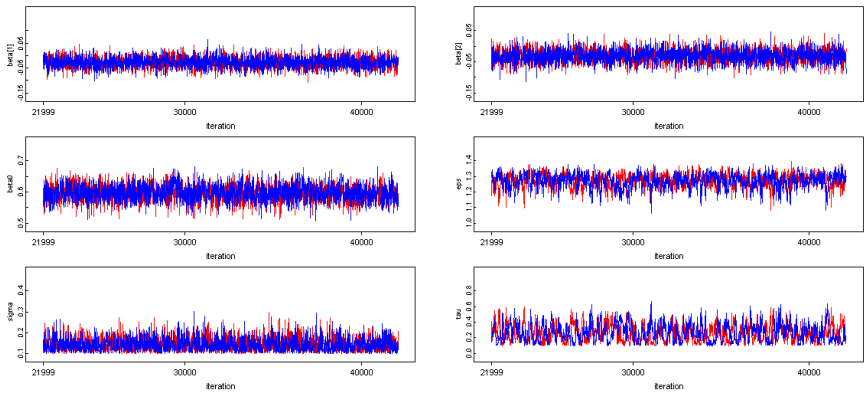

Figure 4.18: Trace plots of the parameters based on two MCMC chains for generated data, after 22000 burn in iterations

A more precise method to verify convergence is to look at the Gelman-Rubin statistic (described in last section). To perform this test, we need to run two or more

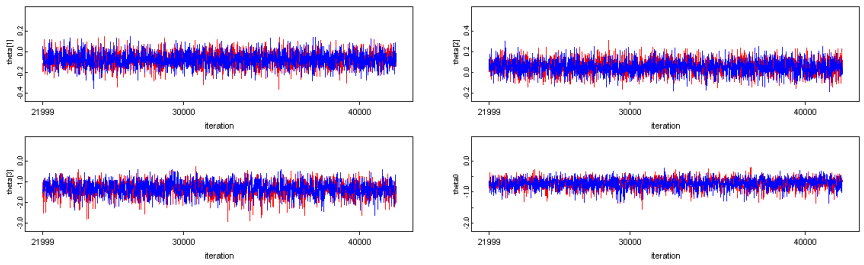

Figure 4.19: Trace plots of the parameters based on two MCMC chains for generated data (cont.)

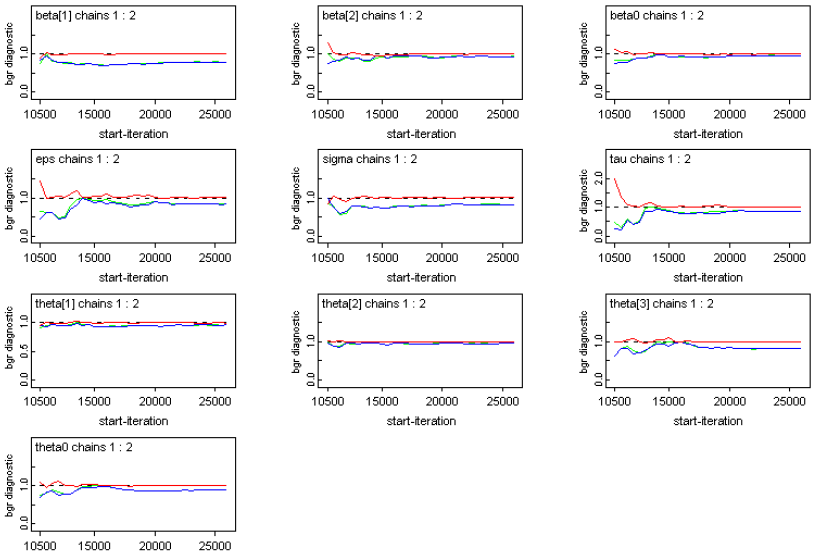

Figure 4.20: Gelman Rubin statistics plots of the parameters based on two MCMC chains for generated data

chains in parallel, with initial values over-dispersed relative to the true posterior. The BGR test compares the variances within and between the chains. In the plots of the BGR statistic in Figure 4.20, the blue and green lines represent within and between chain variations, respectively, and the red line is the ratio of the between and within chain variations. When the lower two lines stabilise (in these diagrams at approximately 15000 iterations) and the red line converges to 1, then we accept that the chain has converged.

To obtain an independent sample, we may wish to thin the converged chain, so

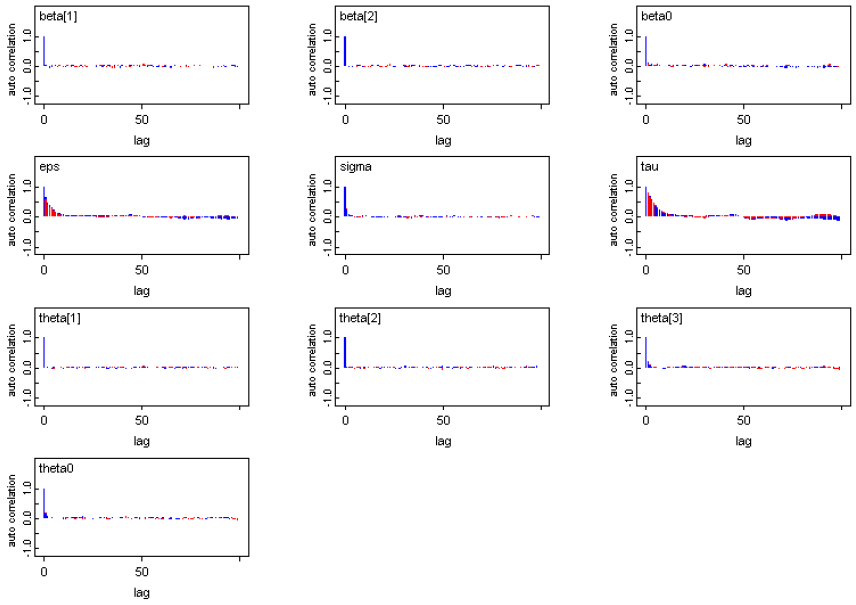

Figure 4.21: Auto-correlation plots of the parameters based on two MCMC chains for generated data

that the correlation between successive states is small. The auto-correlation functions for all the parameters are shown in Figure 4.21. To minimize the autocorrelation among the responses, every 20th iteration is considered in our analysis. The plots here show that the autocorrelation becomes ignorable after considering to thin at 20. In order to obtain a sample of a given size, we need to perform a considerably larger number of iterations which may be costly in terms of computing time.

Finally, I introduce a straightforward way to check for convergence of looking at the Monte Carlo standard error of the posterior mean, which can be obtained from OpenBUGS. We want the values of standard error to be less than 5% of the sample standard deviation. In all our examples, they are all qualified.

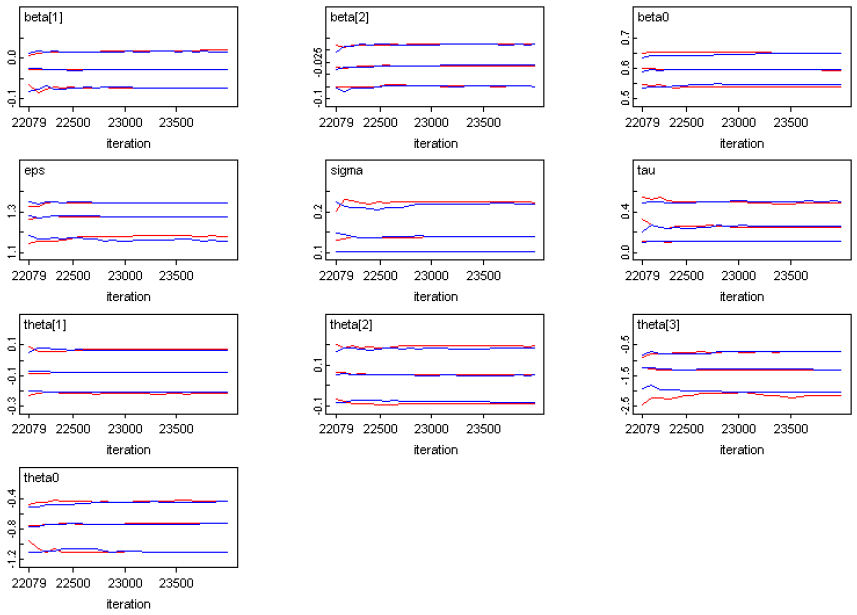

Figure 4.22: Quantile plots of the parameters based on two MCMC chains for generated data

Chapter 5

Discussion

Conclusion

In medical, environmental, econometric and social sciences, it is common to encounter longitudinal data. Longitudinal data may be modelled in different ways depending on the background and the objective of study.

Longitudinal data presents multiple methodological challenges in data analyses. The primary problem is the correlation among the repeated responses of the same subject. Classic models of data analysis, such as multiple linear and logistic regressions, based on the independence of observations may not apply to longitudinal data.

Another issue in longitudinal studies is missing values. Since subjects are followed up for a period of time in longitudinal studies, it is common that some subjects will drop out of the studies. One naive approach would be to delete the subjects with any missing value during the study period; as it does not utilize all available data, such an approach is not efficient. However, sometimes the remarkable bias in the estimate from deletions may yield misleading or even wrong conclusions.

Very often researchers may encounter longitudinal continuous outcome measures with a large number of observed values clustered at zero, called a semi-continuous

longitudinal data.

In this thesis, we have proposed a compound Poisson mixed model for longitudinal semi-continuous data, combined with a logistic regression model to analyse the missing values in the response variable. The advantages of our approach are to analyse zero and non-zero parts of the semi-continuous data in an integral way.

We have set out to show how the Bayesian framework, combined with MCMC method to handle multilevel random effects in semi-continuous data using Tweedie's compound Poisson distribution. We have considered two level random effects in our model. The Bayesian approach has the advantages of ease of interpretation, incorporation of prior knowledge, and reduced small sample bias. The methods used in this project are general, and we may implement them for any similar set of data.

Two semi-continuous longitudinal data sets, the brief symptoms inventory data and fluoride intake data, are analysed by the proposed Tweedie's compound Poisson mixed model using Bayesian approach. We also analyse the data by using a model selection to deal with the missing data.

Further study

There are many areas we could investigate in more details.

Observations of each individual are likely to be serially correlated for longitudinal data since subjects are followed over a period of time. It is of interest to investigate the serially correlated random effects in the model.

The correlation $\rho_{(j,j')}$ between two measurement times for same subject can be

expressed as a flexible structure,

$$\Re = [\rho_{(j,j')}]_{n_m \times n_m} = \begin{bmatrix} 1 & \rho_{(1,2)} & \rho_{(1,3)} & \cdots & \rho_{(1,n_m-1)} \\ \rho_{(2,1)} & 1 & \rho_{(2,3)} & \cdots & \rho_{(2,n_m-1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{(n_m-1,1)} & \rho_{(n_m-1,2)} & \rho_{(n_m-1,3)} & \cdots & 1 \end{bmatrix}. \tag{5.1}$$

To be specific, for exchangeable correlation structure, we can use $\rho_{(j,j')} = \rho$ in (5.1), where the correlation matrix can be expressed as:

$$\Re = [\rho_{(j,j')}]_{n_m \times n_m} = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}. \tag{5.2}$$

Autoregressive of order 1 (AR(1)) structure can be accommodated by considering $\rho_{(j,j')} = \rho^{|j-j'|}$ for any $j \neq j'$. To estimate ρ under AR(1) structure, it would be sufficient to estimate lag 1 ($\rho^1 = \rho$) correlation only, which can be obtained from (5.1) as

$$\Re = [\rho_{(j,j')}]_{n_m \times n_m} = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n_m-1} \\ \rho & 1 & \rho & \cdots & \rho^{n_m-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n_m-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n_m-1} & \rho^{n_m-2} & \rho^{n_m-3} & \cdots & 1 \end{bmatrix}. \tag{5.3}$$

For the Toeplitz structure, assumes $\text{Corr}(Y_{ij}, Y_{i,j+k}) = \rho_k$ for one subject at different

measurements.

$$\Re = [\rho_{(j,j')}]_{n_m \times n_m} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho^{n-m-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho^{n-m-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho^{n-m-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n-m-1} & \rho^{n-m-2} & \rho^{n-m-3} & \cdots & 1 \end{bmatrix}. \tag{5.4}$$

In following work, we may add these correlation structures into the composed model.

Despite the many conceptual advantages of Bayesian statistics, the models are too complex and the Bayesian computation is technically challenging. Sometimes, a model works well for one data set but not for another data set, or a model may work in one package but not in another, therefore, to resolve these issues we may need to code everything from scratch.

It may be possible that the semi-continuous longitudinal data contain a large amount of zero values such that our proposed compound Poisson mixed model becomes inadequate. We may address this extra zero problem by adding a zero-inflation component to the compound Poisson distribution. In a Bayesian context, this should be straightforward in principle. Again, the computational challenges need to be tackled.

we may explore the practical performance of DIC as a model selection criterion for comparing various models. DIC is a Bayesian version of the classical deviance for model assessment. Therefore the model accuracy and model comparison will be explored in the future study.

Appendix A

R Code

A.1 Arrangement of BSI data for compound Poisson mixed model

```
# Load in BSI data.
setwd("/Users/qiuguangsang/Downloads/Qiuguang_study/Thesis/data/bsi")
getwd()
bsi<-read.table("bs1.txt", head=T, sep="\t")
head(bsi)

attach(bsi)
length(unique(I)) #269 parents.
table(J)
# table(I,J)   # J indexes kids within a family (ranges from 1 to 5).
# K is time point ranging from 1 to 17.
detach(bsi)

# identify study_month patterns of kids.
```

```
patterns<-NULL
for (i in 1:269){
  tmpi<-bsi[bsi$I==i,]
  for (j in 1:max(tmpi$J)){
    kid<-as.character(i*1000+j)
    month<-tmpi$study_month[tmpi$J==j]
    patternij<-paste(as.character(month), collapse=",")
    patternijdiff<-paste(as.character(diff(month)), collapse=",")
    maxij<-ifelse(length(month)==1, 0, max(table(diff(month))))
    patterns<-rbind(patterns, c(kid, patternij, patternijdiff,
    maxij))
  } # end for j.
} # end for i.

# keep only kids starting with some equal time periods.
length(unique(patterns[,3])) # 315 patterns!
dim(patterns) # [1] 409   4


patterns<-patterns[patterns[,4]>=6,]
patterns<-patterns[substr(patterns[,2],1,1)=="0",]
# > dim(patterns)
# [1] 258   4

catchphrase<-"0,3,6,9,12,15,18"#,24"
tmp<-which(substr(patterns[,2], 1, nchar(catchphrase))
==catchphrase)
```

```
length(tmp)

# Keep the 74 kids with "0,3,6,9,12,15,18".
patterns<-patterns[tmp,]
length(unique(round(as.integer(patterns[,1])/1000))) # 57.

# Keep the "first" kid in each family with "0,3,6,9,12,15,18".
sub_bsi<-NULL
for (i in unique(as.integer(substr(patterns[,1], 1,

nchar(patterns[,1])-3))){
  tmp<-patterns[as.integer(substr(patterns[,1],1,

  nchar(patterns[,1])-3))==i,1]
  as.integer(tmp)->tmp; j<-min(tmp%%1000)
  bsi_ij<-bsi[bsi$I==i & bsi$J==j,]
  sub_bsi<-rbind(sub_bsi, bsi_ij[1:7,])
}

# sub_bsi would be the data used.#
```

A.2 Arrangement of fluoride intake data for compound Poisson mixed model

```
## "Clean" data.
```

```
# load in "fluoride_full_data.txt".
dt<-read.table("fluoride_full_data.txt", sep="", head=T,
na.strings=".")
dt<-dt[!is.na(dt$income),]
table(dt$id)->tmp
ttmp<-tmp[tmp==4]
ddt<-dt[dt$id%in%names(ttmp),]

y<-matrix(ddt$flouride, byrow=T, ncol=4)*100
z<-ifelse(is.na(y), 0, 1)
sex<-matrix(ddt$sex, byrow=T, ncol=4); sex<-sex[,1]
race<-matrix(ddt$race, byrow=T, ncol=4); race<-race[,1]
momage<-matrix(ddt$momage, byrow=T, ncol=4); momage<-momage[,1]
momedu1<-matrix(ddt$momedu1, byrow=T, ncol=4); momedu1<-momedu1[,1]
momedu2<-matrix(ddt$momedu2, byrow=T, ncol=4); momedu2<-momedu2[,1]
first<-matrix(ddt$first, byrow=T, ncol=4); first<-first[,1]
twokids<-matrix(ddt$twokids, byrow=T, ncol=4); twokids<-twokids[,1]
threekids<-matrix(ddt$threekids, byrow=T, ncol=4);
threekids<-threekids[,1]
income<-matrix(ddt$income, byrow=T, ncol=4);income<-income[,1]
age<-c(1.5, 3.0, 6.0, 9)
y[!is.na(y) & y==0]<-.03
ytransposevec<-c(t(y)); ztransposevec<-c(t(z))

## Write out data for BUGS.
cat("list(m=", m, ",",
"n=", n, ",",
```

```
"y = structure(.Data=c(",paste(ytransposevec, collapse=","), "),
.Dim = c(628, 4)),",
"z = structure(.Data=c(",paste(ztransposevec, collapse=","), "),
 .Dim = c(628, 4)),",
"twokids=c(", paste(twokids, collapse=","), "),",
"threekids=c(", paste(threekids, collapse=","), "),",
"income=c(", paste(income, collapse=","), "))", sep="",
file="datatmp.txt")
```

A.3 R code for generating semi-continuous data with missing

```
# Simulate longitudinal semi-continuous data with missing


q<-1.5
sigma<-0.1; tau<-0.1; #eps<-0.1
eps<-1
u.pres<-sigma^(-2); v.pres<-tau^(-2); y.pres<-eps^(-2)
beta0<-dnorm(0,0.01)


for (j in 1:3){
  theta[j]<-dnorm(0, 0.01)}
theta0<--1
for (j in 1:2){
  beta[j]<-dnorm(0, 0.01)}
```

```
m<-596; n<-4

set.seed(1)

# Level 1:
u<-rgamma(m, 1/sigma^2, 1/sigma^2)

# Level 2:
v<-matrix(NA, m,n)
for (i in 1:m){
  v[i,]<-rgamma(n, u[i]/tau^2, 1/tau^2)
}

# Level 3:
x1<-matrix(rnorm(m*n), m,n)
x2<-matrix(rnorm(m*n), m,n)
mu<-exp(beta0+beta[1]*x1+beta[2]*x2)
phi<-mu^(2-q)*v/(2-q)/eps^2
shape<-(2-q)/(q-1)
rate<-mu^(1-q)/eps^2/(q-1)

N<-matrix(NA, m,n)->y
for (i in 1:m){
  for (j in 1:n){
    N[i,j]<-rpois(1,phi[i,j])
    y[i,j]<-rgamma(1, N[i,j]*shape, rate[i,j])
```

```
  }
}
mean(y); sd(y); mean(y==0)

# checking missing or not
z<-matrix(NA, m,n)
linpred<-theta0+theta[1]*x1+theta[2]*x2+theta[3]*y
prob <- exp(linpred)/(1 + exp(linpred))

for (i in 1:m){
  for (j in 1:n){
    z[i,j]<-rbinom(1,1,prob[i,j])
    if (z[i,j]==1) {y[i,j]<-NA}
  }
}
x1<-round(x1,digits=2)
x2<-round(x2,digits=2)
y<-round(y,digits=2)
X1<-as.vector(x1)
X2<-as.vector(x2)
Y<-as.vector(y)
Z<-as.vector(z)
genedata<-cbind(X1,X2,Y,Z)

Y[!is.na(Y) & Y==0]<-.001
ytransposevec<-c(t(Y)); ztransposevec<-c(t(Z))
x1transposevec<-c(t(X1))
```

```
x2transposevec<-c(t(X2))
cat("list(m=", 596, ",",
    "n=", 4, ",",
    "y = structure(.Data=c(",paste(ytransposevec,

    collapse=","), "), .Dim = c(596, 4)),",
    "z = structure(.Data=c(",paste(ztransposevec,

    collapse=","), "), .Dim = c(596, 4)),",
    "x1= structure(.Data=c(",paste(x1transposevec,

    collapse=","), "), .Dim = c(596, 4)),",
    "x2=structure(.Data=c(",paste(x2transposevec,

    collapse=","), "), .Dim = c(596, 4)))",
    sep="", file="data_gene_2.txt")
```

Appendix B

BUGS Code

B.1 WinBUGS code for testing the significance of covariance of BSI data

```
model{
    # fixed index parameter:
    q<-1.5
    shape<-(2-q)/(q-1)
    u.pres<-pow(sigma,-2)  # need prior for sigma.
    v.pres<-pow(tau,-2)   # need prior for tau.
    y.pres<-pow(eps,-2) # need priors for eps, beta0, beta.

    # following are compound poisson model:

    for (i in 1:m){
    # Level 1:
    u[i]~dgamma(u.pres, u.pres)
```

```
# Level 2:
v.shape[i]<-u[i]*v.pres
for (j in 1:n){
v[i,j]~dgamma(v.shape[i], v.pres)
# Level 3:
log(mu[i,j])<-beta0+beta[1]*study_month[j]+beta[2]*age[i]
+beta[3]*gender[i]+beta[4]*treatment[i]+
beta[5]*hispanic[i]+beta[6]*parent_age[i]+beta[7]*parent_gender[i]
phi[i,j]<-pow(mu[i,j],2-q)*v[i,j]/(2-q)*y.pres
N[i,j]~dpois(phi[i,j]) #T(,100)
y.shape[i,j]<-(N[i,j]+0.2)*shape
y.rate[i,j]<-pow(mu[i,j],1-q)*y.pres/(q-1)
y[i,j]~dgamma(y.shape[i,j], y.rate[i,j])

} # end of j loop.
} # end of i loop.

# Priors.
sigma~dunif(0.01,10)
tau~dunif(0.0000001,10)
eps~dunif(0.001,10)
# u.pres~dgamma(0.1,0.1)
# v.pres~dgamma(0.1,0.1)
# y.pres~dgamma(0.1,0.1)
beta0~dnorm(0,0.01)
for (j in 1:7){
beta[j]~dnorm(0, 0.01)
```

```
    }
    } # end of model
```

B.2 OpenBUGS code for testing the significance of covariance of fluoride intake full data

```
model{
  # fixed index parameter:
  q<-1.5
  shape<-(2-q)/(q-1)
  u.pres<-pow(sigma,-2)  # need prior for sigma.
  ## v.pres<-pow(tau,-2)   # need prior for tau.
  y.pres<-pow(eps,-2) # need priors for eps, beta0, beta.

  ## following are compound poisson model:

  for (i in 1:m){
    ## Level 1:
    u[i]~dgamma(u.pres, u.pres)
    ## Level 2:
    ## v.shape[i]<-u[i]*v.pres
    for (j in 1:n){
     ## v[i,j]~dgamma(v.shape[i], v.pres)
      v[i,j]<-u[i]
```

```
      # Level 3:
      log(mu[i,j])<-beta0+beta[1]*twokids[i]
      +beta[2]*threekids[i]+beta[3]*income[i]
      phi[i,j]<-pow(mu[i,j],2-q)*v[i,j]/(2-q)*y.pres
      N[i,j]~dpois(phi[i,j]) #T(,100)
      y.shape[i,j]<-(N[i,j]+0.2)*shape
      y.rate[i,j]<-pow(mu[i,j],1-q)*y.pres/(q-1)
      y[i,j]~dgamma(y.shape[i,j], y.rate[i,j])
      # missing mechanism
                  ##  Linear regression on logit
      logit(p[i,j]) <- theta0 + theta1*y[i,j]
     ##  Likelihood function for each data point
      z[i,j]   ~ dbern(p[i,j])
    } ## end of j loop.
  } ## end of i loop.

  ## Priors.
  sigma~dunif(0.1,1)
  eps~dunif(0.1,1)
  theta0~dnorm(0,0.01)   #  Prior for intercept
 # theta1~dnorm(0,0.01)

  beta0~dnorm(0,0.01)
  for (j in 1:3){
    beta[j]~dnorm(0, 0.01)
  }
} ## end of model
```

B.3 WinBUGS code for testing the significance of covariance of fluoride intake complete data

```
## Model file.
cat("
model{
  # fixed index parameter:
  q<-1.5
  shape<-(2-q)/(q-1)
  u.pres<-pow(sigma,-2)  # need prior for sigma.
  v.pres<-pow(tau,-2)   # need prior for tau.
  y.pres<-pow(eps,-2) # need priors for eps, beta0, beta.

  # following are compound poisson model:

  for (i in 1:m){
    # Level 1:
    u[i]~dgamma(u.pres, u.pres)
    # Level 2:
    v.shape[i]<-u[i]*v.pres
    for (j in 1:n){
      v[i,j]~dgamma(v.shape[i], v.pres)
      # Level 3:
      log(mu[i,j])<-beta0+beta[1]*age[j]+beta[2]*age[j]*age[j]
       +beta[3]*sex[i]+beta[4]*race[i]+beta[5]*twokids[i]
       +beta[6]*threekids[i]+beta[7]*income[i]
      phi[i,j]<-pow(mu[i,j],2-q)*v[i,j]/(2-q)*y.pres
```

```
      N[i,j]~dpois(phi[i,j]) #T(,100)
      y.shape[i,j]<-(N[i,j]+0.2)*shape
      y.rate[i,j]<-pow(mu[i,j],1-q)*y.pres/(q-1)
      y[i,j]~dgamma(y.shape[i,j], y.rate[i,j])

    } # end of j loop.
  } # end of i loop.

  # Priors.
  sigma~dunif(0.1,1)
  tau~dunif(0.1,1)
  eps~dunif(0.1,1)
  beta0~dnorm(0,0.01)
  for (j in 1:7){
    beta[j]~dnorm(0, 0.01)
  }
} # end of model
", file="fluoridebug.txt")
```

B.4 OpenBUGS code for testing the compound Poisson mixed model with generated data

```
model{

    # fixed index parameter:
    q<-1.5
    shape<-(2-q)/(q-1)
```

```
   u.pres<-pow(sigma,-2)  # need prior for sigma.

   v.pres<-pow(tau,-2)   # need prior for tau.

   y.pres<-pow(eps,-2) # need priors for eps, beta0, beta.

# following are compound poisson model:
   # Level 1:

   for (i in 1:m){
   u[i]~dgamma(u.pres, u.pres)
   }



   # Level 2:
   for (i in 1:m){
   v.shape[i]<-u[i]*v.pres
   for (j in 1:n){
   v[i,j]~dgamma(v.shape[i], v.pres)
   }
   }
   # v.pres<-pow(tau,-2)   # need prior for tau.

   # Level 3:
   for (i in 1:m){
   for (j in 1:n){
   log(mu[i,j])<-beta0+beta[1]*x1[i,j]+beta[2]*x2[i,j]
```

```
phi[i,j]<-pow(mu[i,j],2-q)*v[i,j]/(2-q)*y.pres
N[i,j]~dpois(phi[i,j])#T(,100)

y.shape[i,j]<-(N[i,j]+0.2)*shape
y.rate[i,j]<-pow(mu[i,j],1-q)*y.pres/(q-1)
y[i,j]~dgamma(y.shape[i,j], y.rate[i,j])

#  Linear regression on logit
logit(p[i,j]) <- theta0 + theta[1]*x1[i,j] + theta[2]*x2[i,j]
+theta[3]*y[i,j]
#  Likelihood function for each data point
z[i,j]   ~ dbern(p[i,j])

}
}

# Priors.
sigma~dunif(0.1,10)
tau~dunif(0.1,10)
eps~dunif(0.1,10)
theta0~dnorm(0,0.01)   #  Prior for intercept
for (j in 1:3){
theta[j]~dnorm(0, 0.01)}
beta0~dnorm(0,0.01)
for (j in 1:2){
beta[j]~dnorm(0, 0.01)}
}
```

Bibliography

[1] Anderson, C. J., Verkuilen, J. and Johnson, T. (2010). *Applied generalized linear mixed models: continuous and discrete data*, Springer, New York.

[2] Bayes, T and Price, R. (1763). An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London* 53 (0): 370-418

[3] Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, Series B 22, 302-306.

[4] Bursch, B., Lester, P., Jiang, L., Rotheram-Borus, M. J. and Weiss, R. (2008) Psychosocial predictors of somatic symptoms in adolescents of parents with HIV: a six-year longitudinal study. *AIDS Care* 20, 667-676.

[5] Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*, Springer, New York.

[6] Duan, N., Manning, W. G., Morris, C. N., Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Economic and Business Statistics*, 1, 115-126.

[7] Dunn, Peter K. and Smyth, Gordon K. (2005). Series evaluation of Tweedie exponential dispersion model densities, *Statistics and Computing*, 15(4). 267-280.

[8] Geman, S. and D. Geman (1984) Stochastic relaxation, Gibbs distribu- tion and Bayesian restoration of images, *IEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721-741.

[9] Greene, W. H. (2002). *Econometric Analysis*. New York University.

[10] Hasan, M. T., Sneddon, G., and Ma, R. (2009). Pattern-Mixture Zero- Inflated Mixed Models for Longitudinal Unbalanced Count Data with Excessive Zeros,*Biometrical Journal*, 51, pp. 946-960.

[11] Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, 57, 97-109

[12] Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economics and Social Measurement* 5, 475-492.

[13] Heckman J. (1979). Sample Selection Bias as a Specification Error. *Econometrica* 47, 153-161.

[14] Jørgensen B. (1989). Exponential Dispersion Models, *Journal of Royal Statistical Society, Series B (methodological)*, 49(2), pp.127-162.

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

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

[17] Liu, L., Strawderman, R. L., Johnson, B. A. and O'Quigley, J. M. (2012). Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study, *Statistical Methods in Medical Research* Published online on 2 April 2012, 1-20. DOI: 10.1177/0962280212443324.

[18] Ma R. and Jørgensen B. (2007). *Nested generalized linear models: an orthodox best linear unbiased predictor approach.*, Journal of Royal Statistical Society, B 69, pp.625-641.

[19] Ma R., Jørgensen B and Willms J.D. (2009). Clustered binary data with random cluster sizes : a dual poisson modelling approach, *Statistical Modelling*, 9(2), pp.137-150.

[20] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, 21, 1087-1091.

[21] Min, Y. and Agresti, A. (2002). Modeling Nonnegative Data with Clumping at Zero: A Survey. *Journal Of The Iranian Statistical Society* 1, 7-33.

[22] Molenberghs, G., Thijs, H., Jansen. I., Beunckens. C., Kenward. MG., Mallinckrodt. C. and Carroll. RJ. (2004). Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 5, 445-464.

[23] Olsen, M. K. and Schafer, J. L. (2011). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association* 96, 730-745.

[24] Revfeim, K.J.A. (1984). An initial model of the relationship between rainfall events and daily rainfalls, *Journal of Hydrology*, 75, 357-364.

[25] Robinson, P. (1982). On the Asymptotic Properties of Estimators of Models Containing Limited Dependent Variables, *Econometrica*, 50(1), 27-41.

[26] Thompson, C.S. (1984). Homogeneity analysis of a rainfall series: an application of the use of a realistic rainfall model, *Journal of Climatology*, 4, 609 - 619.

[27] Tobin J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica* 26, 24-36.

[28] Verbeke G. and Molenberghs G. (2000). Linear Mixed Models for Longitudinal Data. *Springer-Verlag*, New York.

[29] Yan G., Ma R. and Hasan, M. T. (to be submitted). Compound Poisson mixed models for semi-continuous longitudinal data with application to health data.

Vita

Candidate's full name: Qiuguang Sang

Universities Attended:

Master of Science, April 2016, University of New Brunswick,
Fredericton, NB, Canada

Ph.D. of Science, June 2012, University of Chinese Academy of Sciences,
Beijing, China

Bachelor of Science, June 2005, Qingdao University,
Qingdao, Shandong, China

Conference Presentations:

Sang, Qiuguang, Yan, Guohua and Ma, Renjun (2014). Is the survival of a rat essentially independent of its blood pressure history in this dose-response experiment? The 2nd Workshop on Statistical Modelling of Complexly Correlated Data with Applications of the AARMS Correlative Research Group.