University of Alberta

MODELS FOR UNIVARIATE AND MULTIVARIATE ANALYSIS OF LONGITUDINAL AND CLUSTERED DATA

by

Dandan Luo

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences

Fall 2012

Edmonton, Alberta

Abstract

Longitudinal studies of repeated observations on subjects are commonly undertaken in medical and biological sciences. The responses on a given occasion may be either univariate or multivariate. We concentrate on three topics related to longitudinal and clustered data analysis. The first topic is the development of a class of generalized linear latent variable models. The second involves the modelling of count data with excess zeros. The third is the development of a non-Gaussian linear mixed effects model for multiple outcomes.

In addressing the first problem, we propose random mean models to account for correlation among repeated measures. We extend random mean models to include mixed outcomes, renaming them random mean joint models. The difficulty in joint modelling of continuous and discrete outcomes is the lack of a natural multivariate distribution. We overcome the difficulty by introducing two cross-correlated latent processes. We apply the Monte Carlo EM (MCEM) algorithm to find the MLEs of regression coefficients and variance components, by treating the latent variables as missing data.

This thesis also proposes regression models for count data with excess zeros. We solve the problem from a perspective different from that of mixture model framework. By employing the zero truncated distribution and the zero modified distribution, we establish a broad class of distributions to model data with excess zeros. We consider the zero modified Poisson regression model and zero modified binomial regression model for cross-sectional data. We extend the zero modified regression models to models with random effects. We further extend random mean models to model zero-inflated data, and formulate the

corresponding zero modified random mean models.

A non-Gaussian linear mixed effects model for multiple outcomes is proposed to the third question. The methodology is motivated by a glaucoma study. The normality assumption for random effects may be unrealistic, raising concerns about the validity of inferences on fixed effects and random effects if it is violated. To accommodate the skewness of the responses and the associations among multiple characteristics, we propose a mixed effects model, in which non-normal random effects are assumed by the log-gamma distribution.

Acknowledgements

I would like to express my sincerely appreciation to my supervisor, Dr. Peng Zhang, for his guidance, encouragement, and support. Without his invaluable suggestions, I would never finish my dissertation.

I would like to thank all my committee members: Dr. Narasimha Prasad, Dr. Byron Schmuland, Dr. Dale Schuurmans, Dr. Grace Y. Yi and Dr. Yan Yuan, for their assistance and valuable advice for my dissertation and also for all the help they have offered to me. I also want to take this opportunity to thanks all the professors, staff and my friends at mathematical and statistical department.

I also thank my parents for their encouragement and support along the way. This thesis is dedicated to my mother Hong Luo, and my father Qian Hu.

Last, but not the least, I am deeply grateful to my husband, Nan Jiang, for his understanding and love.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Longitudinal studies of repeated observations on subjects are commonly undertaken in medical and biological sciences. They are important in the study of chronic diseases such as arthritis, nephritis, diabetes and chronic obstructive pulmonary diseases. The primary interests in the analysis of longitudinal data are the mechanisms of change over time, and the effects of covariates on the progression of the diseases. The responses on a given occasion may be either univariate or multivariate.

In this thesis, we concentrate on three topics related to longitudinal and clustered data, which together cover a broad range of types of data and applications. The first topic is the development of a flexible class of generalized linear latent variable models for univariate and bivariate longitudinal and clustered data. The second involves the modelling of count data with excess zeros in the cross-sectional and longitudinal studies. The third consists of the modelling of bivariate longitudinal data with skewed random effects.

Part I

The theories for longitudinal data are well developed. Diggle et al. (1995) [22] present a comprehensive introduction to the analysis of the longitudinal data. Molengerghs and Verbeke (2005) [32] describe recent developments in the analysis of discrete longitudinal data. The mixed effects models, the most commonly used models in longitudinal and clustered data analysis, assume that all correlations can be described by random effects. However, in some case, it is likely that other sources of correlation may be present, for example, time-series correlation. See Stiratelli et al. (1984) [89], Zeger et al. (1985) [108], Kaufmann (1987) [52], Zeger (1988) [104], Zeger and Qaqish (1988) [109], Chan and Ledolter (1995) [13], and Davis et al. (2000) [18].

Cox et al. (1981) [15] characterized two classes of models of time-dependent data: the parameter-driven model, and the observation-driven model. The parameter-driven model has been studied for count data and binary data separately, albeit not in a unified way for the members in the exponential family. In observation-driven models, also known as the transition models, the conditional distribution of the response is specified as a function of past observations. Although the observation-driven models have some advantages from a computational point of view (Harvey and Fernandes (1989) [44]), the parameter-driven model is conceptually more attractive.

In Chapter 2, we review generalized linear model, mixed effects models, and as well as the Monte Carlo sampling rules. Chapter 3 and Chapter 4 address the first topic of the thesis. In Chapter 3, we concentrate on univariate data and propose a flexible class of generalized linear latent variable models for responses with clustered and longitudinal structures. We consider a family of models, called random mean models (RMMs) (Zhang (2006) [110]), to account for correlation among repeated measures. This class is parameter-driven

(Cox et al. (1981) [15]), and encompasses all the members in the exponential family. Our objective is to model the mean of the response as a function of the covariates and an unobserved latent variable, with some specified covariance structure on the latent process. The responses are assumed to be conditionally independent given the latent variable, both over time and across subjects. The latent variable can provide an interpretation of the data generation mechanism. Linear mixed effects models and generalized linear mixed effects models are special cases of random mean models, in which the latent variables are modelled through random effects.

With regard to parameter-driven models for count data, Zeger (1988) [104], and Chan and Ledolter (1995) [13] propose two different methods for the statistical inference of regression parameters. Zeger (1988) employs a quasi-likelihood approach resembling generalized estimating equations (GEE) method developed for population-averaged methods, while Chan and Ledolter (1995) use a Markov Chain Monte Carlo algorithm. For the statistical inference in random mean models, we apply the adaptive Gaussian quadrature method in the approximation of the integration of the marginal likelihood function.

In Chapter 4, we extend random mean models to bivariate mixed outcomes, renaming them random mean joint models. A number of joint modelling strategies for mixed outcomes have been studied in the literature: those of Catalano and Ryan (1992) [12], Fitzmaurice and Laird (1995) [29], Shah et al. (1997) [86], Sammel et al. (1997) [84], Dunson (2000) [24], Gueorguieva and Agresti (2001) [41]. The difficulty in joint modelling of continuous and discrete outcomes is the lack of a natural multivariate distribution. In random mean joint models, two cross-correlated latent processes are introduced to ac-

count for the correlation between different outcomes. For longitudinal data, we propose a specific form of the cross covariance matrix of the latent process using Kronecker product. We are able to simplify the log-likelihood function of the latent variables, especially the inverse and the determinant of the high dimensional covariance matrix.

A big obstacle to the development of the parameter-driven models for the mixed outcome data is that the likelihood methods are computationally intensive, and the Monte Carlo method may be employed. For the statistical inference of random mean joint models, we apply the Monte Carlo EM (MCEM) algorithm to find the MLEs of regression coefficients and variance components, by treating the latent variables as missing data. In the implementation of the MCEM algorithm, we approximate the Q function by the importance sampling approach, and use Laplace approximation to the posterior distribution of the latent variables given the responses to find the approximate mean and covariance of the instrumental distribution. We demonstrate the methodology with two simulations and a kidney study data set.

Part II

In recent years there has been considerable interest in models for count data that allow for excess zeros in the cross-sectional and longitudinal studies. The generalized linear Poisson models for count data may encounter lack of fit due to disproportionately large frequencies of zeros. The count data with excess zeros are often overdispersed relative to Poisson distribution. This overdispersion does not arise from heterogeneity, it arises from the large frequencies of zeros. The variance-mean relationship must be correctly modelled. Addressing this issue is the second topic of this thesis.

In the literature, Mullahy (1986) [69], Heilbron (1989) (1994) [46] [47], and Lambert (1992) [55] pioneered the regression models based on the zero-inflated Poisson (ZIP) distribution. The ZIP distribution can be viewed as an extreme case of the mixture model of a Poisson distribution and the discrete distribution with point mass of one at zero. Hence, the statistical inference, and the lack of fit test of the ZIP regression models are inherited accordingly from the mixture model framework. Hall (2000) [43] adapted Lambert's methodology to an upper bounded count situation, and introduced the zero-inflated binomial (ZIB) regression models.

In addition to the cross-sectional data, zero inflation may also occur with repeated measures or longitudinal data. Many researchers have incorporated random effects into a wide variety of regression models to account for correlated responses and multiple sources of variance. Duijn and Bockenholt (1995) [23] propose a mixture model, in which the distribution of the Poisson intensity parameter is a step function and is modelled by a gamma distribution, to analyze the overdispersed repeated count data. Zero-inflated regression models for continuous data with repeated measures have also been considered by Olsen and Schafer (2001) [72], Berk and Lachenbruch (2002) [5], Tooze et al. (2002) [96], and Yau et al. (2002) [103]. Hall (2000) incorporated random effects into the ZIP and ZIB models to accommodate repeated measures, so the within-subject correlation and between-subject heterogeneity typical of repeated measures can be accommodated.

In Chapter 5, we focus on the second topic of the thesis: the construction of regression models for count data with excess zeros. We solve the problem from a perspective different from that of the mixture model framework. By employing the zero truncated distribution, and the zero modified distribution,

we establish a broad class of distributions to model data with excess zeros, including discrete distributions as well as continuous distributions.

We are mainly interested in count data in this topic. We consider the zero modified Poisson regression model and the zero modified binomial regression model for crosse-sectional data. We then extend the zero modified regression models to the ones with random effects for clustered data. Two simulations for the zero modified regression models with random intercept are conducted, and a real data example is analyzed to illustrate the new methods. We also extend the random mean models introduced in Chapter 3 to model zero-inflated data, and formulate the corresponding zero modified random mean models. A simulation is conducted to evaluate the random mean model for the temporal count data.

Part III

In Chapter 6, we present a non-Gaussian linear mixed effects model for multiple outcomes. The methodology described in this chapter is motivated by a glaucoma study. The normality assumption for random effects in the linear mixed model may be unrealistic, raising concerns about the validity of inferences on fixed effects and random effects if it is violated. For single-characteristic longitudinal data, it has been shown (Verbeke and Lesaffre (1996), (1997) [97] [98]) that deviations from the normality assumption have little impact on the estimation of the fixed effects and variance components, and much more on the empirical Bayes estimates for random effects in linear mixed models, which may still hold under multiple characteristics case.

In the preliminary study of the glaucoma data set, we found that the distribution of one of the characteristics is skewed. To accommodate the skewness of

the responses and the associations among multiple characteristics, we propose to extend the mixed effects model used in a single characteristic longitudinal study in Zhang et al. (2008) [111], to the situation where non-normal random effects are assumed by the use of the log-gamma distribution in the multiple characteristics longitudinal study. We allow the number and time of repeated measures to differ for different characteristics and units. Adapting the model from one to two characteristics makes the modelling complicated. We are able to reduce the computational complexity by introducing a linear transformation matrix and reordering random effects according to whether they show skewed pattern in the analysis of the within-subject regressions of each characteristics. Prior to application of this model, it is essential to examine the necessity of the adjustment for the skewed random effects. Having noted that the limiting distribution of the family of log-gamma distributions is normal, we propose a lack-of-fit test for comparing the log-gamma model and the Gaussian model, based on the profile likelihood function of the shape parameter.

Chapter 2

Preliminaries

2.1 Generalized Linear Model

We describe the generalized linear model as formulated by Nelder and Wedderburn (1972) [71] in Rodríguez's note (2007) [81]. We assume that Y comes from a distribution in the **exponential family** if it has probability density function

$$f(y|\ \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\phi) + c(y, \phi)\},$$

for known functions $a(\phi)$, $b(\theta)$ and $c(y, \phi)$, with parameters θ and ϕ. The parameter ϕ stands for a certain type of nuisance parameter, such as the variance σ^2 of the normal distribution. Here θ is the canonical parameter, and ϕ is the scale parameter. In all models considered in the thesis, the function $a(\phi)$ has the form

$$a(\phi) = \phi/p,$$

where p is a known prior weight, usually 1.

The parameters θ and ϕ are called location and scale parameters. It can be shown that Y has mean and variance

$$\mathrm{E}(Y) = b'(\theta)$$
$$\mathrm{Var}(Y) = b''(\theta)a(\phi),$$

where $b'(\theta)$ and $b''(\theta)$ are the first and second derivatives of $b(\theta)$. The exponential family includes normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions as special cases.

In the generalized linear model, instead of modelling the mean, a one-to-one continuous differentiable transformation $g(\mu)$ on the mean

$$\eta = g(\mu),$$

is introduced to relate the mean to the linear predictor. The function $g(\mu)$ is called the link function. Examples of link functions include the identity, log, reciprocal, logit and probit etc.

The link function relates the linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ to μ, the expectation of Y. The most commonly used link function is the canonical link

$$\theta = \eta = \mathbf{x}'\boldsymbol{\beta}.$$

2.2 Longitudinal and Clustered Data

Longitudinal studies involve repeated observations of variables obtained from a single individual at different occasions. Observations for the same individual typically exhibit positive correlation. Longitudinal data usually have a temporal order. The ordering of the repeated measures are importance in the analysis. However, many studies in the health sciences give rise to data that do not have a temporal order, but are clustered. As mentioned in Fitzmaurice et al. (2004) [31], clustered data arise in cases when intact groups are randomized to interventions or when naturally occurring groups in the population are randomly sampled. One example of the former is group-randomized trial. In a group-randomized trial, also known as a cluster-randomized trial, groups, rather than the individuals themselves, are randomized to different interventions. Examples of the latter are data arisen from random sampling of naturally occurring groups, where the sampling units could be families, households, hospital wards, medical practices, neighborhoods, or schools.

In clustered data, measurements within a cluster are expected to be more similar than the measurements in different clusters. The correlation or association is usually used to measure the degree of clustering among the measurements within the same cluster. Many standard statistical techniques require independence assumption of the data, hence special models are proposed for clustered data, which explicitly describe and account for the correlation or association. As longitudinal data are special case for clustered data, albeit with a natural ordering of the measurements within a cluster, a general method for clustered and longitudinal data analysis is reviewed in this section.

The last few years have seen remarkable advances in methods for analyzing

longitudinal and clustered data. In particular, there now exists a broad and flexible class of models for correlated data based on a regression paradigm. Regression models have been developed for correlated continuous, binary responses and count responses. Mixed effects models for repeated measures data have become popular because their flexible covariance structure allows for non-constant correlation among the observations from the same individual. Much work has been done to extend the linear regression model and the generalized linear models to repeated measures (Laird and Ware (1982) [54]; Stiratelli et al. (1984) [89]; Liang and Zeger (1986) [57]; Zeger et al. (1988) [107]; Lindstrom and Bates (1990) [59]). We first review the linear mixed effects models for continuous response, and then the generalized linear mixed effects model.

2.2.1 Linear Mixed Effects Models

The most common approach for analyzing continuous clustered or longitudinal data is linear mixed effects models (LMMs). The underlying premise of linear mixed effects models is that the subset of the regression parameters vary randomly from one individual to another, thereby accounting for the sources of natural heterogeneity in the population. Individuals in the population are assumed to have their own subject specific mean response trajectories over time and a subset of the regression parameters are now regarded as being random.

We assume there are N individuals. For the ith individual, we have collected n_i repeated observations, with the response variable Y_{ij} measured at time t_{ij}, $i = 1, \ldots, N, j = 1, \ldots, n_i$. The linear mixed effects model can be expressed as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i,$$

where $\boldsymbol{\beta}$ is a $k \times 1$ vector of fixed effects; $\mathbf{b}_i$ is a $q \times 1$ vector of random effects and usually $\mathbf{b}_i \sim MVN(\mathbf{0}, G_i(\boldsymbol{\sigma_1^2}))$; $\mathbf{X}_i$ is a $n_i \times k$ matrix of covariates; $\mathbf{Z}_i$ is a $n_i \times q$ matrix of covariates (usually $q < k$); $\mathbf{e}_i$ is a $n_i \times 1$ vector of errors and $\mathbf{e}_i \sim MVN(\mathbf{0}, R_i(\boldsymbol{\sigma^2}))$.

Conditional on the random effects $\mathbf{b}_i$ and covariates $\mathbf{X}_i$, the responses in $\mathbf{Y}_i$ from the same individuals are assumed to be independent. The marginal or population-averaged mean and covariance can be calculated as

$$E(\mathbf{Y}_i) = E\{E(\mathbf{Y}_i|\mathbf{b}_i)\} = E(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i) = \mathbf{X}_i\boldsymbol{\beta},$$

and

$$\begin{aligned}
\text{Cov}(\mathbf{Y}_i) &= \text{Cov}\{E(\mathbf{Y}_i|\mathbf{b}_i)\} + E\{\text{Cov}(\mathbf{Y}_i|\mathbf{b}_i)\} \\
&= \text{Cov}(\mathbf{Z}_i\mathbf{b}_i) + \text{Cov}(\mathbf{e}_i) \\
&= \mathbf{Z}_i\text{Cov}(\mathbf{b}_i)\mathbf{Z}_i' + \text{Cov}(\mathbf{e}_i) \\
&= \mathbf{Z}_i G_i \mathbf{Z}_i' + R_i.
\end{aligned}$$

The marginal covariance matrix is not diagonal, thereby accounting for the correlation among the repeated observations on the same subjects in a longitudinal study.

2.2.2 Generalized Linear Mixed Effects Models

The Generalized Linear Mixed Effects Models (GLMMs) (Schall (1991) [85], Zeger and Karim (1991) [105], Breslow and Clayton (1993) [8]) are the most frequently used random effects models in the context of discrete repeated measurements, and are a popular way to model such type of data arising in clinical

trials and epidemiological studies of cancer and other diseases. They are an extension of the class of generalized linear models in which random effects are added to the linear predictor. This modification extends the broad class of generalized linear models to accommodate correlation via random effects, while retaining the ability to model nonnormal distributions and allowing nonlinear models of specific form. The class of GLMMs includes the special cases of linear mixed models, random coefficient models, random effects logistic regression, and random effects Poisson regression, and etc.

The incorporation of random effects is a natural way to model or accommodate correlation in the context of a nonlinear model for nonnormal data. It generates a rich class of correlated data models that would be difficult to specify directly. Readily available, flexible, multivariate distributions analogous to the multivariate normal distribution do not exist for most nonnormally distributed data.

Inferences for these models can be of the usual variety, that is, modeling the effect of predictors on the mean, in which case the random effects and correlation are "nuisance" features of the model. In other situations, however, both estimation and testing of the variances of the random effects, as well as prediction of the realized values of the random effects, may be of interest.

As before, Y_{ij}, is the jth outcome for the cluster i, $i = 1, \ldots, N; j = 1, \ldots, n_i$ and $\mathbf{Y}_i$ is the vector of all measurements for the cluster i. We formulate the generalized linear mixed effects models using a two-step specification.

First-step: Assume that the conditional distribution of each Y_{ij}, given the random effects $\mathbf{b}_i$, belongs to the exponential family with conditional mean

connecting to the linear predictor through the link function g

$$g(\mathrm{E}(Y_{ij}|\mathbf{b}_i)) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i.$$

Second-step: The $\mathbf{b}'_i s$ are assumed to vary independently from one individual to another and $\mathbf{b}_i \sim MVN(\mathbf{0}, G_i)$.

Maximum likelihood (ML) is the standard method of estimation for generalized linear models. Evaluation of the likelihood and hence likelihood inference with GLMMs is computationally difficult. However, the random effects on which the likelihood is conditioned must be integrated out of the distribution prior to maximization as a function of the fixed effects. Although several useful computational methods currently exist, the development of new methods for GLMMs continues to be an active research area.

2.3 Background Material on Monte Carlo

We review the basics of Monte Carlo methods for the remainder of the chapter. In statistics, we are often tasked with computing the expected value of a function $f(\mathbf{x})$ with respect to a probability density function $p(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$, especially when n is not small

$$I = \int_{\mathbf{x}} f(\mathbf{x})p(\mathbf{x})d\mathbf{x}.$$

If a cumulative distribution is non-decreasing and easily invertible then we can draw samples from its distribution by using *inverse sampling*. However, many distributions are difficult or impossible to invert, and in some cases a closed-form representation might not exist or be computationally intractable

to obtain. This is a problem since finding expected values of functions is often a step in a lot of statistical problems. We outline two methods, importance sampling, and rejection sampling, that are useful when direct simulation from p is difficult or impossible but direct simulation from another distribution q is possible. We refer to the distribution similar to p as the *instrumental* distribution, and label it q.

2.3.1 Importance Sampling

The basic idea of importance sampling is to draw from a similar distribution other than $p(\mathbf{x})$, say $q(\mathbf{x})$, and then correct for the bias introduced by sampling from the wrong distribution. Suppose, we sample $\mathbf{x}^{(i)}, i = 1, \ldots, N$, independently from the distribution $q(\mathbf{x})$, then estimate the expectation of $f(\mathbf{x})$ with respect to $p(\mathbf{x})$ by

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} \frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})} f(\mathbf{x}^{(i)}). \tag{2.1}$$

In (2.1), we can see the bias correction, or the importance weight $p(\mathbf{x}^{(i)})/q(\mathbf{x}^{(i)})$ can be determined exactly for a given sampling point $\mathbf{x}^{(i)}$.

In practice, the actual $p(\mathbf{x})$ or $q(\mathbf{x})$ will often be unnormalized. The general form of the approximation accounting for the unnormalized $\tilde{p}$ or $\tilde{q}$ can be expressed as

$$\hat{I}_1 = \sum_{i=1}^{N} \omega_i f(\mathbf{x}^{(i)}),$$

where

$$\omega_i = \frac{\tilde{p}(\mathbf{x}^{(i)})/\tilde{q}(\mathbf{x}^{(i)})}{\sum_{k=1}^{N} \tilde{p}(\mathbf{x}^{(k)})/\tilde{q}(\mathbf{x}^{(k)})}.$$

It is easy to see that $\mathrm{E}(\hat{I}) = I$. The asymptotic variance σ_q^2 is defined as

$$\sigma_q^2 = \int \left[\frac{f(\mathbf{x})p(\mathbf{x})}{q(\mathbf{x})}\right]^2 q(\mathbf{x})\ d\mathbf{x} - I^2, \tag{2.2}$$

and

$$\mathrm{Var}(\hat{I}) = \frac{\sigma_q^2}{N}.$$

As the number of samples is increased, the variance of the estimate $\hat{I}$ will decrease. The density q^* that minimizes this asymptotic variance (2.2) is known to be proportional to $|f(\mathbf{x})|p(\mathbf{x})$ (Kahn and Marshall (1953) [51]). The selection of $q(\mathbf{x})$ will have a huge impact on the accuracy of our estimation. In fact, one of the biggest problems with using the importance sampling method is that a poor selection of the sampling distribution will lead to a high-variance estimate $\hat{I}$, that yields the wrong answer without any indication.

2.3.2 Laplace Approximation

The Laplace approximation (Tierney and Kadane (1986) [95]) is very useful for Monte Carlo as it may be used to construct accurate instrumental density, q. The Laplace approximation is an analytic approximation to the expectation with respect to a distribution p. We assume $l(\mathbf{x}) = \log p(\mathbf{x})$ admits a second-order Taylor expansion about the mode of $p(\mathbf{x})$. Let $\hat{\mathbf{x}}$ denote the maximizer

of $l(\mathbf{x})$ satisfying the equation $l^{(1)}(\mathbf{x}) = 0$. The Laplace method can be applied to approximate integrals of the form

$$\begin{aligned}\int e^{l(\mathbf{x})} d\mathbf{x} &\approx \int \exp\left\{ l(\hat{\mathbf{x}}) + \frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})' l^{(2)}(\hat{\mathbf{x}})(\mathbf{x}-\hat{\mathbf{x}}) \right\} d\mathbf{x} \\ &= \left| -2\pi l^{(2)}(\hat{\mathbf{x}})^{-1} \right|^{\frac{1}{2}} e^{l(\hat{\mathbf{x}})}.\end{aligned}$$

Similarly, we have

$$\int \mathbf{x}\ e^{l(\mathbf{x})}\ d\mathbf{x} \approx \left| -2\pi l^{(2)}(\hat{\mathbf{x}})^{-1} \right|^{\frac{1}{2}} e^{l(\hat{\mathbf{x}})}\ \hat{\mathbf{x}},$$

and

$$\int \mathbf{x}\mathbf{x}'\ e^{l(\mathbf{x})}\ d\mathbf{x} \approx \left| -2\pi l^{(2)}(\hat{\mathbf{x}})^{-1} \right|^{\frac{1}{2}} e^{l(\hat{\mathbf{x}})} (\hat{\mathbf{x}}\hat{\mathbf{x}}' - l^{(2)}(\hat{\mathbf{x}})^{-1}).$$

It suggests that p is approximately normally distributed with mean $\hat{\mathbf{x}}$, and variance $-l^{(2)}(\hat{\mathbf{x}})^{-1}$. The Laplace approximation will be exact when $l(\mathbf{x})$ is a quadratic function of $\mathbf{x}$. The integrand $f(\mathbf{x})$ can be unnormalized posterior distributions of the random effects. Tierney and Kadane reviewed the use of Laplace's approximation for moments of the posterior distribution in Bayesian problems, and discussed modifications that result in higher-order accuracy. In chapter 4, we will use the Laplace approximation to estimate the mean and variance of p, and use these to construct an accurate instrumental distribution function by shifting and scaling a heavy tailed t distribution by the approximate mean and standard deviation respectively.

2.3.3 Rejection Sampling

Rejection sampling is a way to generate an i.i.d. sequence from the target distribution p by thinning out an i.i.d sequence from q

$$p(\mathbf{x}) = af(\mathbf{x})q(\mathbf{x}), \tag{2.3}$$

where a is the normalizing constant. This type of density function in (2.3) is quite common in the random effects model and the random mean model, in which f is the conditional density function of the response given the random effects or the latent variable, and q is the density function of the random effects or the latent variable.

A random sample from p can be selected as follows by multivariate rejection sampling.

Step 1: Generate $\mathbf{x}$ from q and sample u from the uniform $(0, 1)$ distribution.

Step 2: Accept $\mathbf{x}$, and let $\mathbf{z} = \mathbf{x}$, if $u \leq f(\mathbf{x})/\tau$ where $\tau = \sup\{f(\mathbf{x})\}$. Otherwise, go to step 1.

This simple method of simulating from p is often very fast even if the acceptance rate is quite low provided that it is easy to simulate from the assumed random effects density.

Theorem 2.1: *The random variables generated from the rejection sampling algorithm are i.i.d. with distribution p.*

Proof:

$$\begin{aligned} P(Z \leq z) &= P(X \leq z | X \ \textit{accepted}) \\ &= \frac{P(X \leq z, U \leq f(X)/\tau)}{P(U \leq f(X)/\tau)} \\ &= \frac{\int_{-\infty}^{z} q(x) f(x)/\tau \ dx}{\int_{-\infty}^{\infty} q(x) f(x)/\tau \ dx} \\ &= \int_{-\infty}^{z} \frac{q(x) f(x)}{\int_{-\infty}^{\infty} q(x) f(x) \ dx} dx. \end{aligned}$$

The derivation is easily extended for the multivariate case.

Chapter 3

Random Mean Models

3.1 Introduction

In longitudinal and panel studies, random effects models are popular choices for modelling unobserved heterogeneity. In these models, the outcomes are modelled as independent variables conditionally on the subject-specific random effects, and the covariance of the marginal distribution of the responses can be expressed as functions of time, but also a function of the subset of covariates. The induced correlation structure from the random effects seems somewhat awkward and may be unrealistic in most cases. Moreover, it is likely that other sources of correlation, such as time series correlation, may be present. A better model would take into account the possible serial dependence within subject-specific measurements.

Considerable effort has been devoted to the development of methods. Zeger (1988) [104], Campbell (1994) [10], Brannas and Johansson (1994) [7], Chan and Ledolter (1995) [13], Davis et al. (2000) [18], and Hay and Pettitt (2001) [45] discussed models for regression analysis with a time series of counts. Cor-

relation is assumed to arise from an unobservable process added to the linear predictor in a log linear model. The mean function is specified by a linear predictor modified by a latent process. An alternative model is Poisson counts mixed by gamma random effects to give negative binomial marginal. Henderson and Shimakura (2003) [48] considered a Poisson-gamma model to account for between-subjects heterogeneity and within-subjects serial correlation. Thall (1988) [91] proposed a family of Poisson regression models incorporating a mixed random multiplicative component in the rate function of each subject. Duijn and Böckenholt (1995) [23] considered a mixture model with two gamma distributions for the Poisson parameter.

In this chapter, we consider a family of flexible models, random mean models (RMMs) in Zhang (2006) [110], to characterize the correlation among the repeated measures in the longitudinal and clustered data. In RMMs, a latent variable is assumed to be associated with the mean function of the response for each individual. The introduction of the latent variables is to account for the correlation of the repeated responses, but they are not really of intrinsic interest. The linear mixed effects models and the generalized linear mixed effects models are special cases of random mean models, in which the latent variables are modelled through the random effects.

The epileptic seizures data set analyzed by Thall and Vail (1990) [92], and by Breslow and Clayton (1993) [8], is a motivation data set for the new model. The scientific question concerned the effectiveness of the drug progabide to reduce the rate of epileptic seizures. Seizures data were collected from a placebo-controlled clinical trial of 59 epileptics, which aimed to examine the effectiveness of the drug progabide in treating epileptic seizures. Patients suffering from partial seizures were enrolled in the study, and were randomly

assigned to progabide or a placebo, in addition to a standard chemotherapy. For each patient, the number of baseline epileptic seizures before the study was recorded in the preceding eight weeks. The number of epileptic seizures was then reported during four consecutive two-week periods after the randomization.

Breslow and Clayton (1993) considered subject level and unit level random variation in their Model III. The set of independent random effects associated at unit level with each visit was included in the log-linear model to represent nonspecific overdispersion beyond that introduced by the subject-to-subject variation. However, the random mean model, would be better suited to explaining the possible serial dependence within subject-specific measurements than their Model III, in the computational consideration and in the modelling itself.

The formulation and discussion of the random mean models are provided in Section 3.2 and Section 3.3. The discussion particularly focuses on the count data and continuous data, whose marginal moments can be explicitly expressed by the parameters in the probability function of the latent variable. Section 3.4 is about the statistical inference of the random mean model. We apply the adaptive Gaussian quadrature method in the approximation of the integration of the marginal likelihood function. We revisit and analyze the epileptic seizures data set in Section 3.6. Concluding remarks are presented in Section 3.7.

3.2 Model Formulation

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ be the observed data vector, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$ is the vector of responses for the ith individual, $i = 1, \ldots, N$, and $\mathbf{x}_{ij}$ be vector of covariates associated with the jth response, $j = 1, \ldots, n_i$. Conditional on the latent variable u_{ij}, the response y_{ij} is assumed to be arisen from a distribution in the exponential family

$$f(y_{ij}|u_{ij}, \boldsymbol{\beta}, \phi) = \exp\{(y_{ij}\theta_{ij} - b(\theta_{ij}))/(a_{ij}(\phi)) + c(y_{ij}, \phi)\},$$

with linear predictor, $\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_{ij}$, where $\boldsymbol{\beta}$ is the vector of regression coefficients, and the conditional mean $\mu_{ij} = E(Y_{ij}|U_{ij} = u_{ij})$ satisfies $g(\mu_{ij}) = \eta_{ij}$, for some link function g. The transformed conditional mean linearly depends on both fixed effects $\boldsymbol{\beta}$ and the latent variable U_{ij}. In addition, given the random variable $\mathbf{U}_i = (U_{i1}, \ldots, U_{in_i})$, the responses Y_{ij}'s are independent of one another. At the second level of the hierarchy, it is assumed that $\mathbf{U}_i$ is a n_i-variate random variable from a parametric distribution $q_i(\mathbf{u}_i; \boldsymbol{\sigma}_1^2)$ with variance components $\boldsymbol{\sigma}_1^2$. A common assumption is that $\mathbf{U}_i$ comes from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma_i = \Sigma_i(\boldsymbol{\sigma}_1^2)$. The $\mathbf{U}_i$'s are assumed to be independent from one individual to another. The conditional mean, conditional variance and canonical parameter are related through the equation $\mu_{ij} = b'(\theta_{ij})$ and $\mathrm{Var}(Y_{ij}|u_{ij}) = b''(\theta_{ij})a_{ij}(\phi)$ (McCullagh and Nelder (1989) [63]).

We assume that the latent variable $\mathbf{U}_i = (U_{i1}, \ldots, U_{in_i})$, $i = 1, \ldots, N$, is realization from a stationary process, whose variance and covariance functions are not arbitrary but follow some pattern. For example, the covariance

structure of the latent process could be compound symmetry, toeplitz, autoregressive, exponential and other patterns. The empirical observations about the nature of the correlation among repeated measures in longitudinal studies indicate that: (i) the repeated measures are positively correlated, (ii) the correlations often decrease with increasing time separation, i.e., measures taken closer together in time are expected to be more highly correlated than measures further apart in time. By selecting a covariance structure for the latent variable $\mathbf{U}_i$, the random mean models can better accommodate the dependence among data.

For example, when the number of measurement occasions is relatively small, and all individuals are measured at the same set of occasions, we may choose an unstructured covariance matrix

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix}. \tag{3.1}$$

Other covariance patterns could be compound symmetry, with constant variance across occasions, and constant correlation coefficients

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{pmatrix}; \tag{3.2}$$

The Toeplitz covariance patterns for any pair of responses that are equally separated in time have the same correlation, and constant variance across occasions

$$\sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \dots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \dots & 1 \end{pmatrix}; \tag{3.3}$$

or the autoregressive model for equally spaced or approximately equally spaced data

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho^2 & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}. \tag{3.4}$$

3.3 Longitudinal Count and Continuous Data

In this section, we will discuss how the marginal moments of the response are related to the regression coefficients and parameters in the probability density function of the latent variable. Firstly, we express the marginal covariance function of the responses in a general form under the random mean model.

Lemma 3.1: *The marginal covariance between any pair of responses Y_{ij} and*

Y_{ik}, for $j \neq k$, in the random mean model can be expressed as

$$
\begin{aligned}
& Cov(Y_{ij}, Y_{ik}) \\
&= E[Cov\{Y_{ij}, Y_{ik}|\ \sigma(U_{ij}, U_{ik})\}] + Cov[E\{Y_{ij}|\ \sigma(U_{ij}, U_{ik})\}, E\{Y_{ik}|\ \sigma(U_{ij}, U_{ik})\}] \\
&= Cov\{E(Y_{ij}|U_{ij}), E(Y_{ik}|U_{ik})\} \\
&= Cov\{g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta} + U_{ij}), g^{-1}(\mathbf{x}'_{ik}\boldsymbol{\beta} + U_{ik})\}, \qquad (3.5)
\end{aligned}
$$

where g is the link function. The second equality is from the conditional independence assumption.

3.3.1 Count Data

In the case of count data, we are able to express the marginal covariance function of the responses explicitly in terms of the parameters in the distribution of the latent process (see Zeger (1988) [104], and Davis et al. (2000) [18]).

We now derive how the latent process introduces the autocorrelation into repeated measures. Suppose that, given the latent variable U_{ij}, the response Y_{ij} is from Poisson distribution with conditional mean $\mathrm{E}(Y_{ij}|U_{ij}) = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + U_{ij})$. For each individual, conditional on the latent variable $\mathbf{U}_i$, $\mathbf{Y}_i$ is a sequence of independent counts. We now show the marginal moments of the response can be expressed as a function of the log linear regression coefficients and the variance components of the distribution function of the latent variable $\mathbf{U}_i$.

Suppose that U_{ij} is normally distributed with mean 0 and variance σ^2, and the latent variable $\mathbf{U}_i$ has an autoregressive covariance structure

$$\mathrm{Cov}(U_{ij}, U_{ik}) = \sigma^2 \rho^{|k-j|},\ 0 \leq \rho \leq 1.$$

By treating $e^{U_{ij}}$ is from the lognormal distribution, the marginal mean of the response can be calculated as

$$\mathrm{E}(Y_{ij}) = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma^2/2). \tag{3.6}$$

From (3.6), the unconditional mean of the response depends on the moment of the latent variable, which seems undesirable in the regression content. Particularly, in order to have $\mathrm{E}(Y_{ij}) = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})$, it is required that U_{ij} has mean $-\sigma^2/2$ under the normality assumption. However, as long as $\mathbf{U}_i$ is mean stationary, all regression coefficients except the intercept are invariant under changing assumptions about $\mathrm{E}(U_{ij})$.

By Lemma 3.1, the marginal covariance can be calculated as

$$\begin{aligned}
\mathrm{Cov}(Y_{ij}, Y_{ik}) &= \mathrm{Cov}\{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + U_{ij}), \exp(\mathbf{x}'_{ik}\boldsymbol{\beta} + U_{ik})\} \\
&= \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{x}'_{ik}\boldsymbol{\beta})\mathrm{Cov}(e^{U_{ij}}, e^{U_{ik}}) \\
&= \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{x}'_{ik}\boldsymbol{\beta})\{e^{\sigma^2(1+\rho^{|k-j|})} - e^{\sigma^2}\} \\
&= \mathrm{E}(Y_{ij})\mathrm{E}(Y_{ik})(e^{\sigma^2\rho^{|k-j|}} - 1), \text{ for } j \neq k.
\end{aligned} \tag{3.7}$$

The marginal variance can be calculated as

$$\begin{aligned}
\mathrm{Var}(Y_{ij}) &= \mathrm{E}\{\mathrm{Var}(Y_{ij}|U_{ij})\} + \mathrm{Var}\{\mathrm{E}(Y_{ij}|U_{ij})\} \\
&= \mathrm{E}\{\mathrm{Var}(Y_{ij}|U_{ij})\} + \mathrm{Var}(\mu_{ij}) \\
&= \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma^2/2) + \exp(2\mathbf{x}'_{ij}\boldsymbol{\beta})(e^{2\sigma^2} - e^{\sigma^2}) \\
&= \mathrm{E}(Y_{ij}) + E(Y_{ij})^2(e^{\sigma^2} - 1).
\end{aligned} \tag{3.8}$$

By (3.7) and (3.8), the normalized covariance function can be calculated

as

$$\begin{aligned}\rho_Y(j,k) &= \frac{E(Y_{ij})E(Y_{ik})(e^{\sigma^2\rho^{|k-j|}}-1)}{[\{E(Y_{ij})+E(Y_{ij})^2(e^{\sigma^2}-1)\}\{E(Y_{ik})+E(Y_{ik})^2(e^{\sigma^2}-1)\}]^{\frac{1}{2}}} \\ &= \frac{e^{\sigma^2\rho^{|k-j|}}-1}{[\{E(Y_{ij})^{-1}+e^{\sigma^2}-1\}\{E(Y_{ik})^{-1}+e^{\sigma^2}-1\}]^{\frac{1}{2}}}. \end{aligned} \tag{3.9}$$

From the above calculations, we see how the latent process introduces both overdispersion and autocorrelation into the responses. For the binomial distribution, we are unable to obtain closed form of the marginal mean and marginal covariance function of the responses, but the numerical evaluation is possible.

3.3.2 Continuous Data

In the case of continuous responses, the conditional mean of the response Y_{ij}, given the latent variable U_{ij} is expressed as

$$E(Y_{ij}|U_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + U_{ij} + e_{ij}, \tag{3.10}$$

where e_{ij} is the measurement error and independent of the latent variable. The marginal covariance function of the responses is

$$\text{Cov}(Y_{ij}, Y_{ik}) = \text{Cov}(U_{ij}, U_{ik}),$$

for $j \neq k$.

3.4 Statistical Inference and Estimation

The joint probability for the response $\mathbf{y}_i$ and the latent variable $\mathbf{u}_i$ can be expressed as

$$f(\mathbf{y}_i, \mathbf{u}_i) = f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta}, \phi)\ q_i(\mathbf{u}_i; \boldsymbol{\sigma}_1^2),$$

where

$$f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta}, \phi) = \prod_{j=1}^{n_i} f(y_{ij}|u_{ij}),$$

under the conditional independence assumption. Since the latent variable $\mathbf{u}_i$ is unobserved, the inference about the parameters $\boldsymbol{\beta}, \phi$ and $\boldsymbol{\sigma}_1^2$ is based on the marginal likelihood function of the observed data

$$L(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}_1^2; \mathbf{y}) = \prod_{i=1}^{N} \int f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta}, \phi)\ q_i(\mathbf{u}_i; \boldsymbol{\sigma}_1^2)\ d\mathbf{u}_i. \tag{3.11}$$

The maximum likelihood estimates of $\boldsymbol{\beta}, \phi$ and $\boldsymbol{\sigma}_1^2$ are simply those values of $\boldsymbol{\beta}, \phi$ and $\boldsymbol{\sigma}_1^2$ that maximize this likelihood function. However, as the case of the generalized linear mixed effects models, the introduction of the latent variables in the random mean models produces a greater degree of analytic complexity. The marginal likelihood function of the random mean model, obtained after integration over the latent variables, nearly always involves intractable integrals.

In general, no analytic expressions are available for the integrals in (3.11), and numerical approximations are needed. The numerical approximations can be subdivided into two categories. One is based on the approximation of the

integral, and the other is based on an approximation of the data. The Gaussian quadrature method, simply approximates the integral in the marginal likelihood function as a weighted sum of integrand evaluated over selected set of abscissas. The adaptive gaussian quadrature rules (Liu and Pierce (1994) [60], Pinheiro and Bates (1995) [73]) are the numerical integration centered at the mode of the integrand, and rescaled according to the curvature of the log function of the integrand. The penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) are examples of the approximation of the data, in which data are decomposed into the mean and error term, with a Taylor expansion of the mean (Goldstein (1991) [36]; Schall (1991) [85]; Breslow and Clayton (1993) [8]; McGilchrist (1994) [65]). The difference among different methods is the order of the Taylor expansion and the point at which the approximation is expanded. Unfortunately, PQL approximate estimation procedure exhibits many numerical problems and it is not so uncommon that it fails to converge in practical applications.

When the dimension of the integral is small, the numerical approximation is preferable over the Monte Carlo methods. We will adopt the adaptive Gaussian quadrature method in the evaluation of the marginal likelihood function of random mean models.

3.5 Adaptive Gaussian Quadrature Approximations

In this section, we describe the adaptive Gaussian quadrature approximation to evaluate the integral of the log-likelihood function of the random mean

models. The ordinary Gaussian quadrature is used to approximate integrals with respect to a given kernel by a weighted average of the integrand evaluated at predetermined abscissas. The weights and abscissas used in Gaussian quadrature rules for the most common kernels can be obtained from the tables of Abramowitz and Stegun (1964) [1] or by using an algorithm proposed by Golub and Welsch (1969) [38], and Golub (1973) [37].

A natural candidate for the kernel function for the quadrature rule is the distribution of the latent variable, $\mathcal{N}(\mathbf{0}, \Sigma_i)$. The Gaussian quadrature rule can be viewed as a deterministic version of a Monte Carlo integration algorithm, in which random samples of the latent variable are generated from the $\mathcal{N}(\mathbf{0}, \Sigma_i)$ distribution. The samples and the weights in the Gaussian quadrature rule are fixed, while in the Monte Carlo integration algorithms they are left to be randomly chosen.

However, several authors (Albert and Follmann (2000) [2], Lesaffre and Spiessens (2001) [56]) have pointed out that ordinary Gaussian quadrature can perform poorly for too few quadrature points even in quite simple models. Essentially, the problem with ordinary Gaussian quadrature is that the integrand is evaluated on a fixed grid of points, regardless of its behavior over the range of integration. Regions in which the integrand behaves badly may be underrepresented or even completely missed in the ordinary Gaussian quadrature method. In such cases, it is advantageous to customize the quadrature to the shape of the integrand, by concentrating quadrature points in the regions of the bad behavior. This is the idea behind adaptive Gaussian quadrature.

In adaptive Gaussian quadrature (Liu and Pierce (1994); Pinheiro and Bates (1995)), the grid of abscissas is centered at the conditional modes of the integrand, rather than at 0 as in the ordinary Gaussian quadrature, and

rescaled according to the curvature of the log function of the integrand. The requirement for effective results with adaptive Gaussian quadrature is that the ratio of the integrand to some Gaussian curve be a moderately smooth function. This arises frequently, when the integrand is a likelihood function, such as the product of a likelihood function and a Gaussian density function, and the product of several likelihood functions, etc (Liu and Pierce (1994)).

There exists a close relationship among the Laplace approximation, importance sampling, and adaptive Gaussian quadrature rule (Pinheiro and Bates (1995)). Moreover, the importance sampling tends to be much more efficient than the Monte Carlo integration (Geweke (1989) [33]).

We now derive the adaptive Gaussian quadrature rule for the marginal likelihood function of the random mean models. Let $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \phi, \boldsymbol{\sigma}_1^{2'})$ denote the vector of all unknown parameters. The log-likelihood of the marginal distribution function can be written as

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{N} \log \int f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta}, \phi)\ q_i(\mathbf{u}_i; \Sigma_i(\boldsymbol{\sigma}_1^2))\ \mathrm{d}\mathbf{u}_i, \tag{3.12}$$

where $f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta})$ is the conditional probability function of responses $\mathbf{y}_i$ given the latent variable $\mathbf{u}_i$.

Let $f(\mathbf{y}_i, \mathbf{u}_i)$ denote the joint density function of the response and the latent variable. The first and second derivatives of the log function of $f(\mathbf{y}_i, \mathbf{u}_i)$ with respect to $\mathbf{u}_i$ are calculated as

$$\frac{\partial \log f(\mathbf{y}_i, \mathbf{u}_i)}{\partial \mathbf{u}_i} = \sum_{j=1}^{n_i} I_j(y_{ij} - \mu_{ij}) - \Sigma_i^{-1}\mathbf{u}_i \tag{3.13}$$

$$\frac{\partial^2 \log f(\mathbf{y}_i, \mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T} = -\mathbf{W}_i - \Sigma_i^{-1}, \tag{3.14}$$

where I_j is n_i dimensional unit vector with 1 at the jth element, μ_{ij} is the conditional mean $\mathrm{E}(Y_{ij}|u_{ij})$, and the diagonal matrix $\mathbf{W}_i$ has the conditional variance $\mathrm{Var}(Y_{ij}|u_{ij})$ as the diagonal element.

It follows from (3.14) that $\frac{\partial^2 \log f(\mathbf{y}_i,\mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T}$ is negative-definite and, as a result, $\log f(\mathbf{y}_i, \mathbf{u}_i)$ is a strictly concave function of $\mathbf{u}_i$. Therefore, there is a unique point of maximum $\hat{\mathbf{u}}_i$ corresponding to $\frac{\partial \log f(\mathbf{y}_i,\mathbf{u}_i)}{\partial \mathbf{u}_i} = 0$. By taking a second-order Taylor expansion of $\log f(\mathbf{y}_i, \mathbf{u}_i)$ around $\hat{\mathbf{u}}_i$, the integrand in (3.12) is approximately $\mathcal{N}(\hat{\mathbf{u}}_i, -\frac{\partial^2 \log f(\mathbf{y}_i,\mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T}^{-1}|_{\mathbf{u}_i=\hat{\mathbf{u}}_i})$ up to a normalizing constant.

The critical step for the success of importance sampling is the choice of an importance distribution that approximates the integrand. For RMMs, the integrand $f(\mathbf{y}_i, \mathbf{u}_i)$ is approximated by $\mathrm{N}(\hat{\mathbf{u}}_i, -\frac{\partial^2 \log f(\mathbf{y}_i,\mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T}^{-1}|_{\mathbf{u}_i=\hat{\mathbf{u}}_i})$ density, after accounting for some constant coefficient. This is the importance distribution used in the adaptive Gaussian quadrature rule, so that the grid of abscissas is centered around the conditional modes $\hat{\mathbf{u}}_i$ and $\sqrt{2}(-\frac{\partial^2 \log f(\mathbf{y}_i,\mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T}|_{\mathbf{u}_i=\hat{\mathbf{u}}_i})^{-\frac{1}{2}}$ is used for scaling.

Define $\mathbf{u}_k = (u_{k_1}, \ldots, u_{k_{n_i}})'$, where u_{k_l} and ω_{k_l}, $k_l = 1, \ldots, N_{GQ}$, denote, respectively, the abscissas and the weights for the one-dimensional Gaussian quadrature rule based on the $\mathcal{N}(0, 1)$ kernel. Centering and scaling the abscissas $\mathbf{u}_k$ according to

$$\tilde{\mathbf{u}}_{ik} = \hat{\mathbf{u}}_i + \sqrt{2}\left(-\frac{\partial^2 \log f^T(\mathbf{y}_i, \mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T}\bigg|_{\mathbf{u}_i=\hat{\mathbf{u}}_i}\right)^{-\frac{1}{2}} \mathbf{u}_k, \ i = 1, \ldots, N.$$

Gaussian quadrature rules for multiple integrals are complex. However, using the structure of the integrand in the random mean models, we can break the problem into successive applications of simple one-dimensional Gaussian quadrature rules. Each component of $\tilde{\mathbf{u}}_{ik}$ follows the one dimensional adaptive

Gaussian quadrature rule. The adaptive Gaussian quadrature approximation to the log-likelihood function is

$$l_{AGQ}(\boldsymbol{\theta};\mathbf{y}) = \sum_{i=1}^{N} \frac{n_i}{2}\log 2 - \log\left(-\frac{\partial^2 \log f(\mathbf{y}_i, \mathbf{u}_i)}{\partial \mathbf{u}_i \partial \mathbf{u}_i^T}\bigg|_{\mathbf{u}_i=\hat{\mathbf{u}}_i} \right)^{\frac{1}{2}} + \log \sum_{k}^{N_{GQ}} f(\mathbf{y}_i, \tilde{\mathbf{u}}_{ik}) W_k,$$

where $W_k = \exp(\|\mathbf{u}_k\|^2) \prod_{l=1}^{n_i} \omega_{k_l}$.

3.6 Data Example: Epileptic Seizures Data

We revisit the epileptic seizures data set analyzed by Thall and Vail (1990) [92], and Breslow and Clayton (1993) [8]. Seizures data are from a placebo-controlled clinical trial of 59 epileptics. For each patient, the number of baseline epileptic seizures before the study was recorded in the preceding eight weeks. The number of epileptic seizures was then reported during four consecutive two-week periods after the randomization.

We model the seizure data through random mean model. Specifically, the conditional mean of the response is assumed to linearly depend on the covariates and latent variable on the log scale

$$\log(\mu_{ij}) = \text{base}_i + \beta_0 + \beta_1 \text{ trt}_i + \beta_2 \text{ logage}_i + \beta_3 \text{ visit}_j + u_{ij},$$
$$i = 1, \ldots 59, j = 1, \ldots, 4,$$

with some covariance pattern structure imposed on the latent process. In this data set, the dependence arises from the potential serial correlation between years for each patient. We consider the autoregressive covariance structure, and compound symmetry structure for the latent process. Model III in Breslow

Parameter	Estimate	Standard Error	P-value
intcpt	−1.0838	1.0736	0.3171
trt	−0.3179	0.1419	0.0291
logage	0.3580	0.3175	0.2645
visit	−0.0503	0.0402	0.2160
σ^2	0.3459	0.0598	<.0001
ρ	0.6896	0.0767	<.0001

Table 3.1: Estimates of the fixed effects and variance components of RMM with AR(1) covariance structure

Parameter	Estimate	Standard Error	P-value
intcpt	−0.9469	1.1336	0.4072
trt	−0.3081	0.1498	0.0445
logage	0.3166	0.3361	0.3503
visit	−0.0511	0.0330	0.1272
σ^2	0.3578	0.0647	<.0001
ρ	0.6416	0.0866	<.0001

Table 3.2: Estimates of the fixed effects and variance components of RMM with compound symmetry covariance structure

and Clayton (1993) corresponds to a random mean model with compound symmetry covariance structure. We also model the data through GLMM with random intercept for the purpose of model comparison.

Table 3.1 lists the estimates and standard errors of of the fixed effects and variance components of RMM with AR(1) covariance structure for the latent variables. Table 3.2 lists the estimates and standard errors of the fixed

Parameter	Estimate	Standard Error	P-value
intcpt	−0.8837	1.1427	0.4424
trt	−0.3132	0.1513	0.0429
logage	0.3184	0.3395	0.3522
visit	−0.0591	0.0203	0.0051
σ^2	0.2692	0.0618	<.0001

Table 3.3: Estimates of the fixed effects and variance components of GLMM with random intercept

Model	AIC	BIC	-2loglike
AR(1)	1267.9	1280.3	1255.9
Compound Symmetry	1265	1277.4	1253
Random Intercept	1346.9	1357.3	1336.9

Table 3.4: Model Comparison

Mean	Std Dev	Min	Max
0.04043	0.49727	-1.27969	1.63323

Table 3.5: Summaries of the empirical estimates of the latent variables

effects and variance components of RMM with compound symmetry covariance structure of the latent variables. Table 3.3 lists the estimates and standard errors of the fixed effects and variance components of GLMM with random intercept.

The model selection criteria (both Akaike information criterion (AIC) and Bayesian information criterion (BIC)) suggest that compound symmetry covariance structure is preferred over AR(1) covariance structure. From Table 3.2, we see that treatment effect is statistically significant and the data provide evidence that the treatment is helping to lessen the disease symptoms.

The AIC and BIC of the two random mean models with AR(1) and compound symmetry covariance structures are pretty close to each other. Sometimes, if the correlation of the response is of real interest, we can choose the model with AR(1) covariance structure, and the correlation function is calculated by formula (3.9). We take the patient with id number 104 for example, and plot the estimated correlation function over time separation in Figure 3.1. The correlation of changes over time separation decreases rapidly as the time separation increases.

Table 3.5 summarizes the empirical estimates of the latent variables. We locate the patient 227 and patient 206, who have the maximum and minimum

empirical estimates of the latent variable. Patient 206 has observations 11, 0, 0, 5, with predicted latent variables -0.46853, -1.27969, -1.26316, -0.82414. Patient 227 has observations 18, 24, 76, 25, with predicted latent variables 0.33872, 0.57229, 1.63329, 0.68067. We also identify patients 225 and 112 as having the highest overall count levels based on the empirical estimates. Our findings are similar to those from Breslow and Clayton.

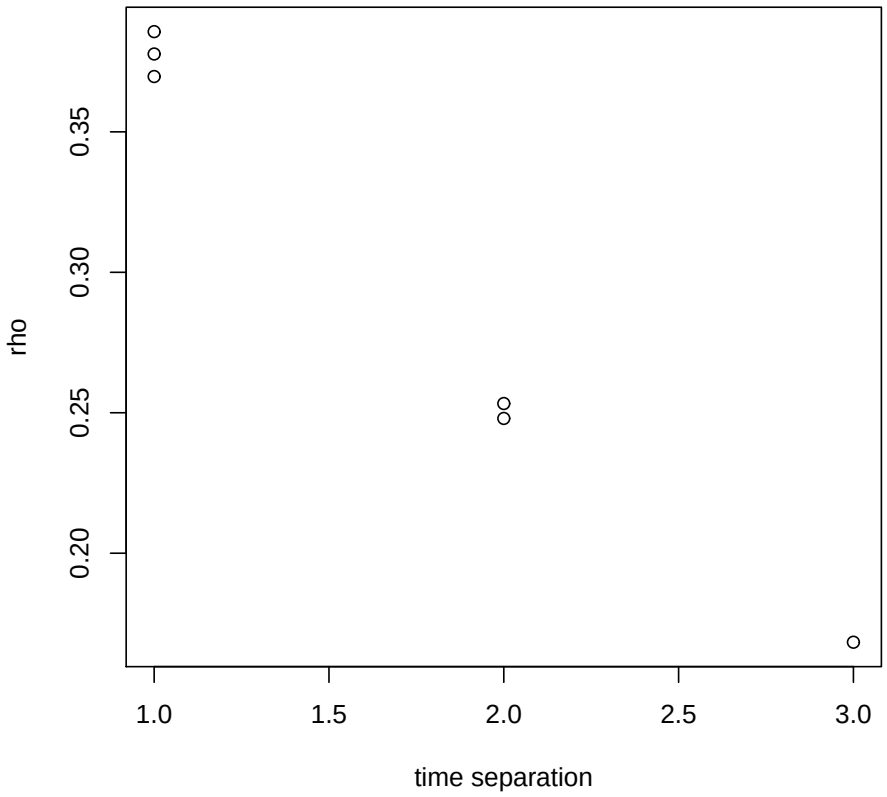

Figure 3.1: Plot of the estimated correlation function

The residual plots in Figure 3.2 show that the mean of the residuals is approximately zero, with the increase of variability with the mean. This is expected for Poisson count data, since the variance is supposed to increase

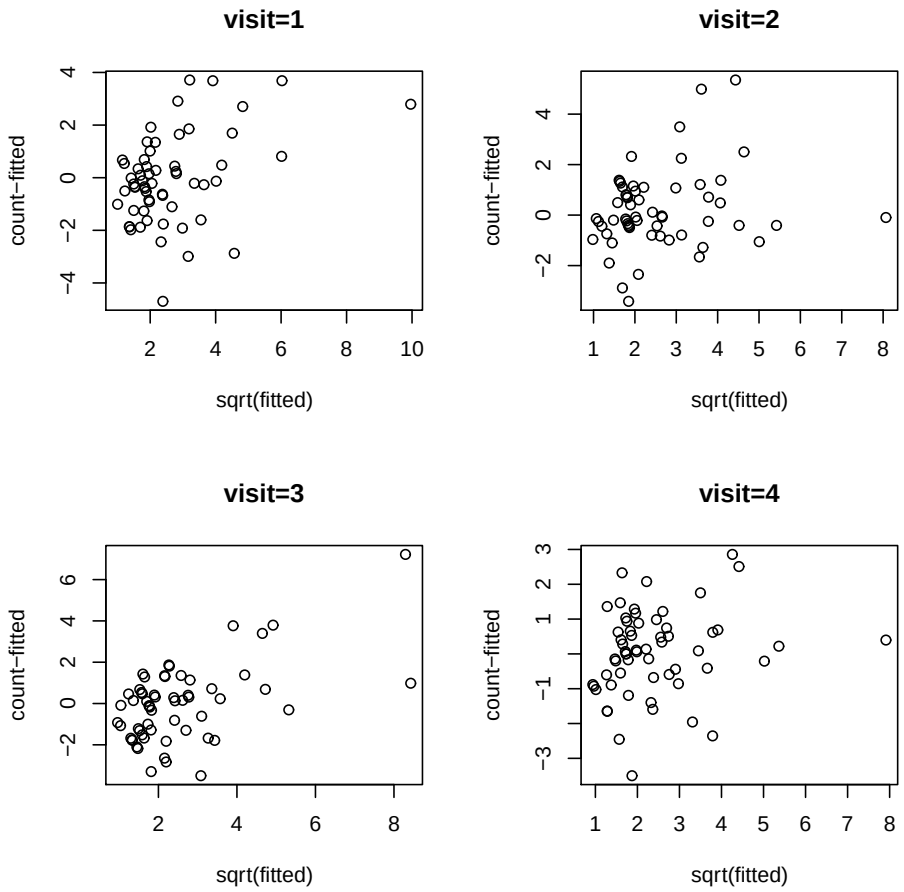

Figure 3.2: Plots of the residuals verses the square root of the fitted values

linearly with the mean. The largest residuals are 225, 225, 227, 135 at visits $1, 2, 3, 4$, separately.

3.7 Conclusion

In summary, random mean models have many attracting implications. Firstly, RMMs provide a more reasonable description of longitudinal and clustered structures of the data than do LMMs and GLMMs. The LMMs and GLMMs impose a strong assumption on the mean model of the responses, where the overall effect of the unobserved factors linearly depends on the subject-specific effects. By contrast, the latent variables introduced to model the overall effects of unobserved factors in RMMs, make the modelling of the correlation in the responses much easier to interpret. The latent processes can also reflect the natural heterogeneity due to many unmeasured factors.

The second advantage of RMMs is from the computational side. Except for the normal distribution, there is no closed form of the marginal distribution function of the responses. The preferred model should be computationally efficient. In GLMMs, if many subject effects and additional random error terms that represent nonspecific overdispersion beyond those introduced by the subject effects are specified, it will increase the computational complexity of the marginal likelihood function. See, for example, Model III of the seizure data set in Breslow and Clayton (1993) [8]. However, the dimension of the integration over latent variables in RMMs is the number of observations for each individual, which is usually small in practice. This is because the unobserved latent variable is usually composed of a small number of replications, and a large number of repeated measures.

Thirdly, RMMs can accommodate the modelling of longitudinal and clustered data in a broader sense. It is able to describe all the distributions in the exponential family. We showed how the marginal moments of the repeated measures depend on the parameters in the distribution function of the latent variable for count and continuous data. Conceptually, we can extend the idea to the other distributions in the exponential family. Specifically, an unobservable process is added to the linear predictor, and the linear predictor is related to the conditional mean through the link function. Besides the autoregressive correlation structure, other covariance structures for the latent process can be selected.

Chapter 4

Random Mean Joint Models

4.1 Introduction

The modelling of various forms of clustered data, such as repeated measurements in a longitudinal study, has received much attention in recent years. Most research has concentrated on a single outcome variable, but many studies in health and medicine application produce mixed outcomes for each subject. In this chapter, we extend the random mean models to clustered continuous and discrete outcomes.

Joint models are potentially advantageous in several statistical and practical respects. Jointly modelling different outcomes can address the questions of interest - the overall relationships between those outcomes and the joint influence of factors on them. These questions can not be answered directly by analyzing different outcomes separately. Further, joint analysis avoids multiple testing and naturally leads to global tests, thus resulting in increased power and better control of Type I error rates. See Pocock et al. (1987) [75], Fitzmaurice and Laird (1997) [30].

Models for multivariate clustered data require accounting for two types of correlation: correlation among different outcomes and correlation among repeated measures of the same outcome over time. A number of joint modelling strategies for mixed outcomes have been studied in the literature: Catalano and Ryan (1992) [12], Fitzmaurice and Laird (1995) [29], Shah et al. (1997) [86], Sammel et al. (1997) [84], Dunson (2000) [24], Gueorguieva and Agresti (2001) [41]. The general approach first specifies a model for the joint distribution of mixed outcomes, then fits the model to data at hand, and finally uses the model to make an inference. A difficulty in joint modelling of continuous and discrete outcomes is the lack of a natural multivariate distribution.

One of the approaches directly specifies the joint distribution by factorizing it into a conditional distribution of one type of outcomes and a marginal distribution of the other type of outcomes. For the mixed clustered data, Catalano and Ryan (1992) [12] parameterized the model such that the joint distribution is factorized as the product of the marginal distribution of the continuous response and the conditional distribution of the discrete response given the continuous response. They assume that the binary response has a corresponding unobserved continuous latent variable, and the latent variable and the continuous response have a joint Gaussian distribution. The marginal distribution of the continuous response is related to covariates through a linear link function, and the conditional distribution of the binary response is related to covariates through a probit link function. The lack of a marginal interpretation and the lack of robustness to misspecification may be considered unattractive features of their approach. Fitzmaurice and Laird (1995) [29] factorize the joint distribution as the product of a marginal Bernoulli distribution for the discrete response, and a conditional Gaussian distribution for

the continuous response given the discrete one. They describe an extension of this model to model clustered data, using generalized estimating equations method, (see Liang and Zeger (1986) [57], and Zeger and Liang (1986) [106]). Cox and Wermuth (1992) [16] compare a number of models for the joint distribution of continuous and binary response variables.

Another approach directly formulates a joint model for both types of outcomes. It involves the introduction of the correlated random effects to incorporate correlations among mixed outcomes. Daniels and Normand (2006) [17] adopt a Bayesian approach to jointly modelling multilevel multidimensional continuous and discrete outcomes with serial dependence. Fieuws and Verbeke (2006) [28] propose a pairwise modelling strategy, in which all possible pairs are modelled separately based on a mixed model. The inference is based on pseudo-likelihood principles. Faes et al. (2008) [25] extend the approach of Fieuws and Verbeke (2006) to different types of outcomes. An alternative strategy uses the copulas to account for the correlations (see Song et al. (2009) [88], and DeLeon and Wu (2011) [20]).

In this chapter, we extend random mean models to mixed characteristics and discuss random mean joint models for clustered continuous and discrete outcomes. Specifically, two correlated latent processes are introduced into the modelling. One is for the continuous outcomes, and the other is for the discrete outcomes. The correlation among two different outcomes is accounted through the correlation of two latent processes. For the longitudinal data, we propose a specific form of the cross-covariance matrix of the latent variables using Kronecker product. The expression of the log-likelihood function of the joint latent variables, especially the inverse and the determinant of the high dimensional covariance matrix is hence simplified.

For the statistical inference, we apply the Monte Carlo EM (MCEM) algorithm to find the MLEs of regression coefficients and variance components of the generalized linear latent variable model, by treating the latent variables as missing data. In the implementation of the MCEM algorithm, we approximate the Q function by the importance sampling approach, and use the Laplace approximation to the posterior distribution of the latent variables given the responses to give the approximate mean and covariance of the instrumental distribution. We demonstrate the methodology with simulations and a kidney study data set.

The remaining chapter is organized as follows. Section 4.2 and Section 4.3 are model formulation and statistical inference of random mean joint models. Simulation studies of count and continuous data with different cluster sizes are presented in Section 4.4. The kidney data is analyzed in Section 4.5. The conclusion remarks and the appendix for the derivative of the log-likelihood function of the latent variable are given in Section 4.6 and Section A.

4.2 Model Formulation

For modelling the observations, let Y_{kij} denote the kth response variable at time j of the ith subject, and $\mathbf{x}_{kij}$ be vector of covariates associated with the response y_{kij}, $k = 1, 2$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. Let $\mathbf{Y}_{1i} = (Y_{1i1}, Y_{1i2}, \ldots, Y_{1in_i})$ and $\mathbf{Y}_{2i} = (Y_{2i1}, Y_{2i2}, \ldots, Y_{2in_i})$ denote the sequences of outcomes of continuous and discrete responses, respectively, for the ith subject, $i = 1, \ldots, N$.

Joint analysis of mixed outcomes, requires either direct or indirect specification of the joint density $f(\mathbf{y}_{1i}, \mathbf{y}_{2i})$, $i = 1, \ldots, N$, and incorporates the association between two different types of outcomes at each time point, as

well as the association from repeated measurements of the same type of outcome.

Similar to the case of single type of outcomes, we introduce latent variables to account for the correlation among the repeated measurements for mixed outcomes. Specifically, conditional on the latent variable U_{kij}, the response Y_{kij} is assumed to be arisen from a distribution in the exponential family

$$f(y_{kij}|u_{kij}, \boldsymbol{\beta}_k, \phi_k) = \exp\{(y_{kij}\theta_{kij} - b_k(\theta_{kij}))/a_{kij}(\phi_k) + c_k(y_{kij}, \phi_k)\}, \quad (4.1)$$

with linear predictor, $\eta_{kij} = \mathbf{x}'_{kij}\boldsymbol{\beta}_k + u_{kij}$, where $\boldsymbol{\beta}_k$ is the vector of regression coefficients, and the conditional mean $\mu_{kij} = E(Y_{kij}|U_{kij} = u_{kij})$ satisfies $g_k(\mu_{kij}) = \eta_{kij}$, for some link function g_k, $k = 1, 2$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. The transformed mean linearly depends on both the fixed effects $\boldsymbol{\beta}_k$ and latent variable u_{kij}. At the second level of the hierarchy, it is assumed that $\mathbf{U}_i = (\mathbf{U}'_{1i}, \mathbf{U}'_{2i})'$ is a $2 \times n_i$-variate random variable from a parametric distribution $q_i(\mathbf{u}_i; \boldsymbol{\sigma}_1^2)$ with variance components $\boldsymbol{\sigma}_1^2$. A common assumption is that $\mathbf{u}_i$ comes from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma_i = \Sigma_i(\boldsymbol{\sigma}_1^2)$. The $\mathbf{U}_i$'s are assumed to be independent from one individual to another. The conditional mean, conditional variance and canonical parameter are related through the equations $\mu_{kij} = b'_k(\theta_{kij})$ and $\mathrm{Var}(Y_{kij}|u_{kij}) = b''(\theta_{kij})a_{kij}(\phi_k)$(McCullagh and Nelder (1989) [63]).

To complete the model specification, we need to introduce the association between two latent processes.

Definition: Let (U_{1_t}, U_{2_t}) represent a pair of stochastic processes that are jointly wide sense stationary. Then the cross-covariance function is defined

as

$$\gamma_{U_1 U_2}(\tau) = \mathrm{E}[(U_{1_{t+\tau}} - \mu_1)(U_{2_t} - \mu_2)],$$

for all time t, where μ_1 and μ_2 are the means of U_{1_t} and U_{2_t} respectively.

The cross-covariance function of weakly stationary processes is a function of time separation or lag τ. The cross-correlation function between two time series is described by the normalized cross-covariance function.

Definition: The cross-correlation function is defined as the normalized cross-covariance function

$$\rho_{U_1 U_2}(\tau) = \frac{\gamma_{U_1 U_2}(\tau)}{\sqrt{\gamma_{U_1}(0)\gamma_{U_2}(0)}}, \tag{4.2}$$

where $\gamma_{U_1}(\cdot)$ and $\gamma_{U_2}(\cdot)$ are the autocovariance functions of processes U_{1_t} and U_{2_t}, respectively.

When two series, U_{1_t} and U_{2_t}, satisfy the equation

$$U_{2_t} = AU_{1_t} + \omega_t, \tag{4.3}$$

where A is a known constant, a simple form of the expression of cross-covariance function can be derived. Equation (4.3) just says that one series is predictable from the other series plus an additional error term. We assume that, for convenience, U_{1_t} and U_{2_t} have zero means, and the noise ω_t is uncorrelated with U_{1_t}. For such two processes, it can be shown that the cross-covariance function

can be calculated as

$$\begin{aligned}\gamma_{U_2U_1}(\tau) &= \mathrm{E}(U_{2_{t+\tau}}U_{1_t}) \\ &= A\ \mathrm{E}(U_{1_{t+\tau}}U_{1_t}) + \mathrm{E}(\omega_{t+\tau}U_{1_t}) \\ &= A\ \gamma_{U_1}(\tau),\end{aligned} \tag{4.4}$$

and

$$\gamma_{U_1U_2}(\tau) = A\ \gamma_{U_1}(\tau). \tag{4.5}$$

By (4.2), (4.4) and (4.5), as long as the autocorrelation function of the process U_{1_t} is known, the cross-correlation function $\rho_{U_1U_2}(\tau)$ is determined.

Returning to random mean joint models, we model the correlation between different types of outcomes by the cross-correlation function of two latent processes. Specifically, assuming that the latent processes $\mathbf{U}_{1i}$ and $\mathbf{U}_{2i}$, satisfying equation (4.3), are AR(1) processes with the same autocorrelation coefficient ρ, the resulting covariance matrix of the latent variable $\mathbf{U}_i = (\mathbf{U}'_{1i}, \mathbf{U}'_{2i})'$ is

$$\Sigma_i = \begin{pmatrix} \sigma_1^2 & A\sigma_1^2 \\ A\sigma_1^2 & \sigma_2^2 \end{pmatrix} \otimes \begin{pmatrix} 1 & \rho & \cdots & \rho^{n_i-1} \\ \rho & 1 & \cdots & \rho^{n_i-2} \\ \cdots & \cdots & \cdots & \cdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \cdots & 1 \end{pmatrix} \tag{4.6}$$

$$= R \otimes T_i, \tag{4.7}$$

where σ_1^2 and σ_2^2 represent the variances of two series $\mathbf{U}_{1i}$ and $\mathbf{U}_{2i}$ separately. In (4.7), T_i has an autoregressive covariance structure. In practice, other patterns could be chosen, which depends on the feature of the particular data

set.

The Kronecker product of two matrices possesses very nice properties in terms of the operations of inverse and determinant. It is known that $\Sigma_i^{-1} = R^{-1} \otimes T_i^{-1}$ and $|\Sigma_i| = |R|^{n_i}|T_i|^2$, which can simplify the computation of the inverse and determinant of a high dimensional matrix significantly.

By (4.7), the density function of the latent variable $\mathbf{U}_i$ can be written as

$$(2\pi)^{-n_i}(|R|^{n_i}|T_i|^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{u}_i'(R^{-1}\otimes T_i^{-1})\mathbf{u}_i\right\}, \tag{4.8}$$

and the log-likelihood function of $\mathbf{U}_i$ is

$$-n_i\log(2\pi) - \frac{n_i}{2}\log|R| - \log|T_i| - \frac{1}{2}\mathbf{u}_i'(R^{-1}\otimes T_i^{-1})\mathbf{u}_i. \tag{4.9}$$

The parameters in the covariance matrix, Σ_i, are either appearing in R, or in T_i. The derivatives of Σ_i with respect to each parameter are still the Kronecker products of two matrices, with R or T_i replaced by the corresponding derivative matrix. In the appendix section, the first and second derivatives of (4.9), the log-likelihood function of the latent variable, with respect to each parameter are provided.

Some discussions about the properties of the cross-correlation function for the two latent processes are given below. By (3.9) and (4.3), the cross-correlation function is calculated as

$$\begin{aligned}\rho_{U_1U_2}(0) &= \frac{A\sigma_{U_1}^2}{\sqrt{A^2\sigma_{U_1}^2+\sigma_\omega^2}\sqrt{\sigma_{U_1}^2}} \\ &= \frac{A}{\sqrt{A^2+\frac{\sigma_\omega^2}{\sigma_{U_1}^2}}},\end{aligned}$$

since

$$\sigma_{U_2}^2 = A^2\sigma_{U_1}^2 + \sigma_\omega^2. \tag{4.10}$$

We know that $\rho_{U_1U_2}$ is bounded by -1 and 1. It is not surprising to see that as $|A|$ approaches infinity, $|\rho_{U_1U_2}(0)|$ has a limit 1. The cross-correlation function can measure the predictability of another series from a given series.

When closely examining the equation (4.3), we observed that there is a natural constraint on the range of the coefficient A. By (4.10), we have

$$A^2 < \frac{\sigma_{U_2}^2}{\sigma_{U_1}^2},$$

which implies that R in (4.7) is definite positive.

4.3 Statistical Inference and Estimation

The joint probability function of $\mathbf{Y}_i = (\mathbf{Y}_{1i}', \mathbf{Y}_{2i}')'$ and $\mathbf{U}_i = (\mathbf{U}_{1i}', \mathbf{U}_{2i}')'$, $i = 1, \ldots, N$, can be expressed as

$$f(\mathbf{y}_i, \mathbf{u}_i) = f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta}, \phi)\ q_i(\mathbf{u}_i; \boldsymbol{\sigma}_1^2),$$

where

$$f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta}, \phi) = \prod_{k=1}^{2}\prod_{j=1}^{n_i} f(y_{kij}|u_{kij}; \boldsymbol{\beta}_k, \phi_k),$$

under the conditional independence assumption. Since the latent variables $\mathbf{u}_i$'s are unobserved, inference about the parameters $\boldsymbol{\beta}_k, \phi_k$ and $\boldsymbol{\sigma}_1^2$ is based on

the marginal likelihood function of the observed data

$$L(\boldsymbol{\beta}, \phi, \boldsymbol{\sigma}_1^2; \mathbf{y}) = \prod_{i=1}^{N} \int f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta}, \phi)\ q_i(\mathbf{u}_i; \boldsymbol{\sigma}_1^2)\ d\mathbf{u}_i. \tag{4.11}$$

The maximum likelihood estimates of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')', \phi = (\phi_1, \phi_2)$ and $\boldsymbol{\sigma}_1^2$ are simply those values of $\boldsymbol{\beta}, \phi$ and $\boldsymbol{\sigma}_1^2$ that maximize this likelihood function. However, the integration (4.11) always involves intractable integrals. Much work has been focused on approximate techniques that seek to avoid the integration.

Besides the numerical approximation methods mentioned in Chapter 3, alternative attempts to carry out the integrations are via fully Bayesian analysis using the Markov chain Monte Carlo (MCMC) techniques (Zeger and Karim, (1991) [105]) or using the Monte Carlo EM algorithm to implement "exact" likelihood analysis (Tanner (1993) [90]; McCulloch (1997) [64]; Booth and Hobert (1999) [6]). McCulloch (1997) suggested obtaining a sample via Markov chain Monte Carlo techniques, in particular the Metropolis-Hastings algorithms. Estimates obtained from the MCMC are presented without reliable standard errors. The reason is possibly that calculating the standard error of an estimate is often not trivial in the MCMC methods. Indeed, establishing the existence of a central limit theorem for a Monte Carlo estimate based on a Markov chain can be difficult (Chan and Geyer (1994) [14]; Meyn and Tweedie (1993) [68]; Tierney (1994) [94]). The usual way of establishing the existence of central limit theorems is to show that the Markov chain itself is geometrically ergodic. Although many Markov chains in the MCMC algorithms have been shown to be geometrically ergodic, myriad complex MCMC algorithms are currently in use to which these results do not apply (see (Hobert

and Geyer (1998) [49]; Mengersen and Tweedie (1996) [67]; and Roberts and Rosenthal (1999) [80])). Furthermore, even though a CLT exists, estimating the asymptotic variance may not be easy (see Geyer (1992) [34] and Mykland et al. (1995) [70]).

The Monte Carlo EM (MCEM) algorithm, introduced by Wei and Tanner (1990) [100], and Tanner (1996) [90], is an extension of the EM algorithm that estimates the expectation in the E-step with a Monte Carlo approximation. Booth and Hobert (1999) [6] proposed to use rejection sampling and multivariate t importance sampling to generate independent samples to construct Monte Carlo approximations at the E-step. Because of the hierarchical structure of the random mean model, we can apply the Monte Carlo EM algorithm for the inference of the random mean joint models.

4.3.1 EM and Monte Carlo EM Algorithms

The EM algorithm is based on the idea of replacing one difficult likelihood maximization with a sequence of easier maximization whose limit is the solution to the original problem. It is particularly suited to "missing" data problems, as the missing data can sometimes make calculations cumbersome. However, filling in the "missing data" will often make the calculation become more smoothly (Casella (2001) [11]).

In general, if $\mathbf{y} = (y_1, \ldots, y_n)$ denotes the incomplete data, and $\mathbf{u} = (u_1, \ldots, u_n)$ denotes the augmented data, making $(\mathbf{y}, \mathbf{u})$ the complete data. Let $L(\boldsymbol{\theta}|\mathbf{y})$ be the incomplete-data likelihood, and $L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u})$ be the complete-data likelihood. When $L(\boldsymbol{\theta}|\mathbf{y})$ is difficult to work with, it will sometimes be the case that the complete-data likelihood will be easier to work with. From

an initial value $\boldsymbol{\theta}^{(0)}$, we create a sequence $\boldsymbol{\theta}^{(r)}$ according to the value that maximizes the Q function

$$Q(\boldsymbol{\theta},\ \boldsymbol{\theta}^{(r)}) = \mathrm{E}[\log\ L(\boldsymbol{\theta}|\mathbf{y},\mathbf{u})|\ \mathbf{y};\boldsymbol{\theta}^{(r)}], \tag{4.12}$$

The E step imputes the unobserved log-likelihood of the complete data, consisting of the observed data and the missing data, by the conditional expectation of the complete data log-likelihood given the observed data.

A Monte Carlo version of EM was introduced by Wei and Tanner (1990) [100], and Mcculloch (1997) [64]. In this section, we apply the Monte Carlo EM algorithm to the random mean joint models, in which the Q function in the E-step of the EM algorithm is replaced by a Monte Carlo approximation. It also involves Monte Carlo approximations of the gradient vectors and hessian matrices by using independent sampling. Since it is independent samples, rather than Markov chain sampling, it is easier to assess standard error and faster to converge.

In the EM algorithm, the E step imputes the log-likelihood of the complete data, consisting of the observed data and the latent variables, by the conditional expectation of the complete data log-likelihood given the observed data. An important property of the EM algorithm is that the likelihood of the observed data always increases along an EM sequence. For the MCEM algorithm, this property does not hold. But it is shown that (Chan and Ledolter (1995) [13]), under suitable regularity conditions, the MCEM sequence will, with high probability , converge to the maximum likelihood estimate. In the Monte Carlo EM algorithm, the conditional expectation of the log-likelihood of the complete data is estimated by averaging the conditional log-likelihoods

of simulated variates.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \phi', \boldsymbol{\sigma}_1^2)'$ denote the vector of unknown parameters. To set up the MCEM algorithm in the context of the random mean joint models, we consider the latent variable, $\mathbf{u}$, to be the missing data. Let $f(\mathbf{y}, \mathbf{u}; \boldsymbol{\theta})$ represent the joint density of the complete data, $(\mathbf{y}', \mathbf{u}')'$. At the E-step of the MCEM algorithm, the interest lies in using Monte Carlo averages of simulated variables to estimate

$$Q(\boldsymbol{\theta},\ \boldsymbol{\theta}^{(r)}) = E[\log\ f(\mathbf{y}, \mathbf{u}; \boldsymbol{\theta})|\ \mathbf{y}; \boldsymbol{\theta}^{(r)}], \tag{4.13}$$

where the expectation of the log-likelihood function of the complete data is with respect to $h(\mathbf{u}|\ \mathbf{y}; \boldsymbol{\theta}^{(r)})$, the conditional distribution of the latent variable $\mathbf{u}$ given the response $\mathbf{y}$ with parameter value $\boldsymbol{\theta}^{(r)}$. Generally, we have

$$h(\mathbf{u}|\ \mathbf{y}; \boldsymbol{\theta}) \propto f(\mathbf{y}|\ \mathbf{u}; \boldsymbol{\beta}, \phi)\ q(\mathbf{u}; \boldsymbol{\sigma}_1^2),$$

where the normalizing constant is the marginal likelihood function of $\mathbf{y}$. By drawing a random sample, $\mathbf{u}^1, \ldots, \mathbf{u}^L$, from $h(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}^{(r)})$, the Monte Carlo approximation of $Q(\boldsymbol{\theta},\ \boldsymbol{\theta}^{(r)})$ is given by

$$\hat{Q}_{r+1}(\boldsymbol{\theta},\ \boldsymbol{\theta}^{(r)}) = \frac{1}{L} \sum_{l=1}^{L}\ \log\ f(\mathbf{y}, \mathbf{u}^l; \boldsymbol{\theta}). \tag{4.14}$$

In the implementation of the MCEM, due to independence among subjects, one may sample from $h(\mathbf{u}_i|\ \mathbf{y}_i, \boldsymbol{\theta}^{(r)})$ for the ith individual, $i = 1, \ldots, N$. Because of the introduction of Monte Carlo error at the E-step, the incomplete data log-likelihood (4.14) is not guaranteed to increase at every iteration. However, the Monte Carlo EM algorithm still converges to the maximum like-

lihood estimate under suitable regularity conditions Chan and Ledolter (1995) [13].

The M-step maximizes the approximate Q function (4.14) obtained in the Monte Carlo E-step, with respect to $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{(r+1)}$. The MCEM algorithm iterates between the "approximate" E-step and the M-step, drawing a sample of the unobserved variables at each iteration from the conditional distribution given the observed data at the updated parameter value; and maximizing the approximate Q function obtained from the new sample to update the estimate of the parameter. As McCulloch (1997) [64] has pointed out, the Monte Carlo M-step is usually relatively simple in the generalized linear mixed model context. This is still true under the random mean model setting. The reason is that $\hat{Q}_{r+1}(\boldsymbol{\theta},\ \boldsymbol{\theta}^{(r)})$ is the sum of log-likelihood functions from two generalized linear models. The first term involves $\boldsymbol{\beta}$ and ϕ, and the second one involves only $\boldsymbol{\sigma}_1^2$. The first term depends on the distribution $f(\mathbf{y}|\ \mathbf{u}; \boldsymbol{\beta}, \phi)$, so it can be maximized via iteratively reweighted least squares. The maximizer of the second term can sometimes be written in closed form.

In summary, the choice of missing data has two advantages. First, on knowing the $\mathbf{u}$, the $\mathbf{y}'s$ are independent. Second, the M-step of the EM algorithm maximizes (4.14) with respect to $\boldsymbol{\beta}, \phi$ and $\boldsymbol{\sigma}_1^2$ could be separated into two parts. The M-step with respect to $\boldsymbol{\beta}$ and ϕ only needs $f(\mathbf{y}|\mathbf{u})$, so it becomes to a standard generalized linear model problem, with the values of $\mathbf{u}$ treated as known. The maximizer of the second term can sometimes be written in a closed form. The MCEM algorithm for the random mean models is as follows

1. Choose starting values $\boldsymbol{\theta}^{(0)}$, and initial sample size L.

2. Generate L random samples, $\mathbf{u}_r^1, \ldots, \mathbf{u}_r^L$ from $h(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}^{(r)})$ using rejection

or importance sampling methods.

3. Using the approximation (4.14) to obtain $\boldsymbol{\theta}^{(r+1)}$ by maximizing $\hat{Q}_{r+1}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)})$.

4. If convergence is achieved, then declare $\boldsymbol{\theta}^{(r+1)}$ to be the maximum likelihood estimate of $\boldsymbol{\theta}$; otherwise, return to Step 2.

4.3.2 Implementation

The implementation of the Monte Carlo E-step involves sampling the unobserved $\mathbf{u}$ from the conditional distribution $h(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}^{(r)})$. This requires us to choose an "appropriate" Monte Carlo sampler that simulates $\mathbf{u}$ from a distribution that is as close as possible to the target distribution $h(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}^{(r)})$. The choice could be rejection sampling, importance sampling, or dependent samples from an invariant target distribution based on the Markov chain Monte Carlo methods. Rejection sampling is more efficient when sample sizes are small, whereas importance sampling is better with larger sample sizes. Both of them are useful when direct simulation from h is difficult or impossible but direct simulation from another distribution similar to h is possible. When the acceptance rate for the rejection sampler is very low, it may be more efficient to use importance sampling.

We have discussed the rejection sampling and importance sampling schemes in the preliminary section. It is straight forward to implement the rejection sampling scheme in the random mean joint model. However, the application of the importance sampling is not easy, hence we mainly discuss the importance sampling in this section. Specifically, we apply the Laplace importance sampling method to the approximation of the Q function. The approximation of (4.13) based on the importance sampling is

$$\hat{Q}_{r+1}(\boldsymbol{\theta},\ \boldsymbol{\theta}^{(r)}) = \sum_{l=1}^{L}\ \omega_l\ \log\ f(\mathbf{y}, \mathbf{u}^l; \boldsymbol{\theta}), \tag{4.15}$$

where

$$\omega_l = \frac{\exp\{\log\ f(\mathbf{y}, \mathbf{u}^l; \boldsymbol{\theta})\}/h^*(\mathbf{u}^l)}{\sum_{k=1}^{L} \exp\{\log\ f(\mathbf{y}, \mathbf{u}^k; \boldsymbol{\theta})\}/h^*(\mathbf{u}^k)}, \tag{4.16}$$

and $\mathbf{u}^l$ are random samples from the importance density h^*.

We consider the Laplace approximation to suggest the proposal distribution. Booth and Hobert (1999) [6] proposed the multivariate Student t importance density whose mean and variance match the mode and curvature of h. More specifically, we write $h(\mathbf{u}_i|\mathbf{y}_i, \boldsymbol{\theta}) = a_i\ \exp\{l(\mathbf{u}_i)\}$, where a_i is the unknown normalizing constant. By using the notation introduced in section 4.2, $l(\mathbf{u}_i)$ can be expressed as

$$\sum_{k=1}^{2}\sum_{j=1}^{n_i} \log\{f(y_{kij}|\ \mathbf{u}_i; \boldsymbol{\beta}_k, \phi_k)\} - \frac{1}{2}\log|2\pi\Sigma_i| - \frac{1}{2}\mathbf{u}_i'\Sigma_i^{-1}\mathbf{u}_i.$$

Let $l^{(1)}(\mathbf{u}_i)$ denote the vector of the first derivatives of $l(\mathbf{u}_i)$, and $l^{(2)}(\mathbf{u}_i)$ denote the hessian matrix of the second derivatives of $l(\mathbf{u}_i)$ with respect to $\mathbf{u}_i$

$$l^{(1)}(\mathbf{u}_i) = \mathrm{vec}\left\{\mathrm{vec}\left\{\frac{y_{kij} - \mu_{kij}}{a_{kij}(\phi)b_k''(\theta_{kij})g'(\mu_{kij})}\right\}_j\right\}_k - \Sigma_i^{-1}\mathbf{u}_i,$$

and

$$l^{(2)}(\mathbf{u}_i) = -W_i - \Sigma_i^{-1},$$

where W_i is the diagonal matrix of iterative weights

$$\text{diag}\left\{\text{diag}\left\{1/\{a_{kij}(\phi_k)b_k''(\theta_{kij})g'(\mu_{kij})^2\right\}_j\right\}_k,$$

for $j = 1, \ldots, n_i$ and $k = 1, 2$. Suppose that $\tilde{\mathbf{u}}_i$ is the maximizer of $l(\mathbf{u}_i)$ satisfying the equation $l^{(1)}(\mathbf{u}_i) = 0$. The Laplace approximation of the mean and variance are $\tilde{\mathbf{u}}_i$ and $-l^{(2)}(\tilde{\mathbf{u}}_i)^{-1}$ respectively. Of course, we can also choose multivariate normal distribution with mean and variance are $\tilde{\mathbf{u}}_i$ and $-l^{(2)}(\tilde{\mathbf{u}}_i)^{-1}$ as the importance function.

The approximations to the conditional mean and variance of $\mathbf{u}_i$ are

$$E(\mathbf{u}_i|\mathbf{y}_i) \approx \tilde{\mathbf{u}}_i \tag{4.17}$$

$$\text{Var}(\mathbf{u}_i|\mathbf{y}_i) \approx -l^{(2)}(\tilde{\mathbf{u}}_i)^{-1}. \tag{4.18}$$

One important consideration in implementing the Monte Carlo EM is the specification of L. It is inefficient to start with a large value of L when $\boldsymbol{\theta}^{(r)}$ is far from the mode. Rather, one may increase L as the current approximation moves closer to the true value of the maximizer. We implement the Monte Carlo EM algorithm through R package.

4.3.3 Information Matrix

Denote the MLE from the MCEM algorithm by $\hat{\boldsymbol{\theta}}$. Louis (1982) [61] showed that the observed information matrix is given by

$$-\text{E}\left\{\frac{\partial^2 l(\boldsymbol{\theta}|\ \mathbf{y}, \mathbf{u})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \text{Var}\left\{\frac{\partial l(\boldsymbol{\theta}|\ \mathbf{y}, \mathbf{u})}{\partial\boldsymbol{\theta}}\right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

where the expectation and variance are with respect to $h(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\theta}})$. Within the context of the Monte Carlo EM, the observed information matrix of the observed data at the posterior mode $\hat{\boldsymbol{\theta}}$ can be estimated via

$$-\sum_{l=1}^{L}\omega_l \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log\, f(\mathbf{y},\mathbf{u}^l;\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}} - \sum_{l=1}^{L}\left(\omega_l \frac{\partial}{\partial\boldsymbol{\theta}}\log\, f(\mathbf{y},\mathbf{u}^l;\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}}\right)\left(\omega_l \frac{\partial}{\partial\boldsymbol{\theta}}\log\, f(\mathbf{y},\mathbf{u}^l;\boldsymbol{\theta})\Big|_{\hat{\boldsymbol{\theta}}}\right)'.$$

4.4 Simulation Study

We conducted simulation studies to investigate the performance of the proposed method. We report on cases of longitudinal continuous and count mixed responses with different cluster sizes.

4.4.1 Simulation Study I

In each of 100 Monte Carlo data sets, observations y_{1ij} and y_{2ik}, $j = 1, \ldots, n_i = 5, k = 1, \ldots, n_i = 5$, were generated for the ith individual, $i = 1, \ldots, N = 150$. The components in $\mathbf{y}_i = (\mathbf{y}'_{1i}, \mathbf{y}'_{2i})'$, are assumed to be conditionally independent given the latent variable $\mathbf{u}_i = (\mathbf{u}'_{1i}, \mathbf{u}'_{2i})'$.

The conditional distributions of the continuous and count responses are $y_{1ij}|u_{1ij} \sim N(\mu_{1ij}, \sigma^2)$, where

$$\mu_{1ij} = \beta_{10} + x_{1ij}\beta_{11} + u_{1ij},$$

and $y_{2ik}|u_{2ik} \sim$ Poisson (μ_{2ik}), where

$$\log(\mu_{2ik}) = \beta_{20} + x_{2ik}\beta_{21} + u_{2ik}.$$

We choose the autoregressive covariance structure (4.6) for the latent variable $\mathbf{u}_i$. The data were simulated according to the above model with $\beta_{10} = 6.71$, $\beta_{11} = 1.35$, $\beta_{20} = 1$, $\beta_{21} = 1.6$, $\sigma^2 = 0.09$, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 1$, $\rho = 0.7$, $A = 1.2$, $x_{1ij} = j/15$, and $x_{2ik} = k/15$.

In the implementation of the MCEM algorithm, we started at the above values, and used the importance sampling scheme $L \leftarrow L + L/m$ with $m = 10$ and initial $L = 50$. The multivariate Student t distribution with 40 degrees of freedom was chosen as the instrumental distribution. The mean and variance of the t distribution are calculated by (4.17) and (4.18) at each iteration. The algorithm is stopped when the relative change in the parameter values from successive iterations is small

$$\max|\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}| < \delta,$$

where δ is predetermined as 0.001.

In table 4.1, the average of the estimates over 100 data sets, the average of estimated standard errors, and the empirical mean square error for the estimates are listed from the third column to the fifth column. All the estimates of parameters in the regression model and the variance components are unbiased. By comparing the average of the standard error estimates and the standard deviations of the parameter estimates, the Monte Carlo error can be judged. From Figure 4.1 to Figure 4.9, we plot the estimates of parameters over the iteration times in one simulation data set. After about 50 iterations, all estimates are stabilized.

Model	Parameter(True)	Average estimate	Estimated SE	Empirical SE
Linear	$\beta_{10}(6.71)$	6.70589	0.05697	0.05877
	$\beta_{11}(1.35)$	1.37775	0.25765	0.25079
Log-linear	$\beta_{20}(1)$	1.00137	0.08168	0.08361
	$\beta_{21}(1.6)$	1.61480	0.34933	0.36108
Variance	σ (0.3)	0.29741	0.01030	0.02657
Component	σ_1^2 (0.5)	0.49681	0.06699	0.04059
	σ_2^2 (1)	0.99098	0.13365	0.07586
	ρ (0.7)	0.69613	0.03207	0.03006
	A (1.2)	1.19985	0.06088	0.06321

Table 4.1: Simulation results for 100 data sets from normal and Poisson random mean joint model

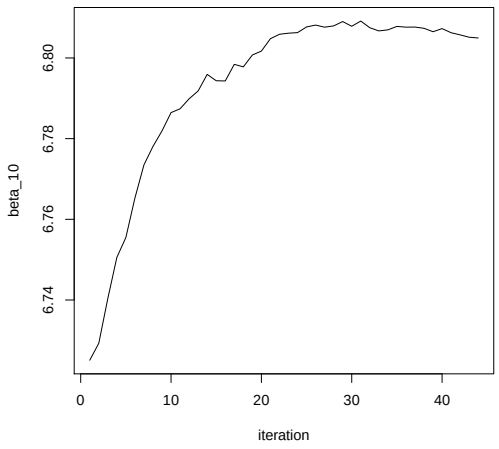

Figure 4.1: Changes for coefficient estimate β_{10} of one simulated data set

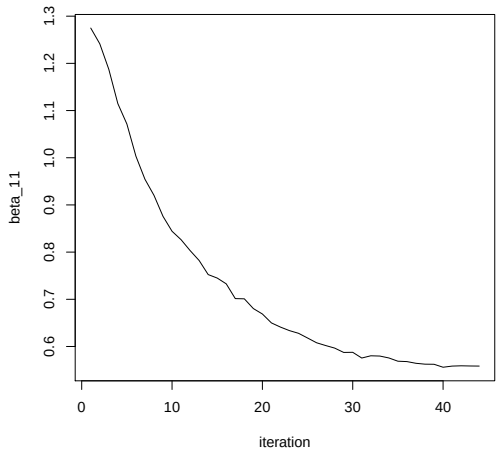

Figure 4.2: Changes for coefficient estimate β_{11} of one simulated data set

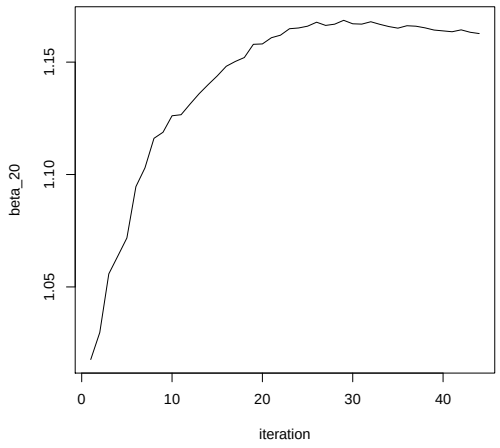

Figure 4.3: Changes for coefficient estimate β_{20} of one simulated data set

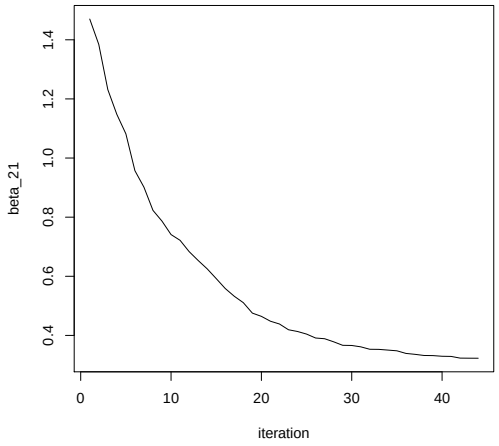

Figure 4.4: Changes for coefficient estimate β_{21}
of one simulated data set

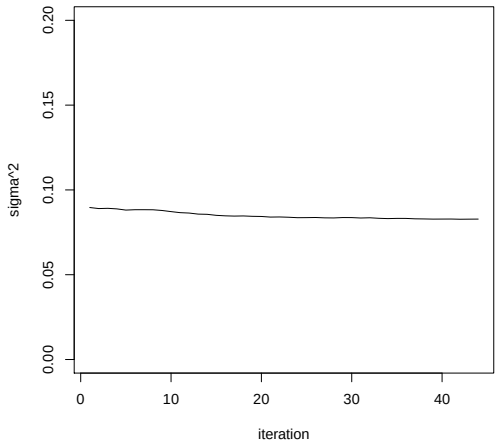

Figure 4.5: Changes for variance component estimate σ
of one simulated data set

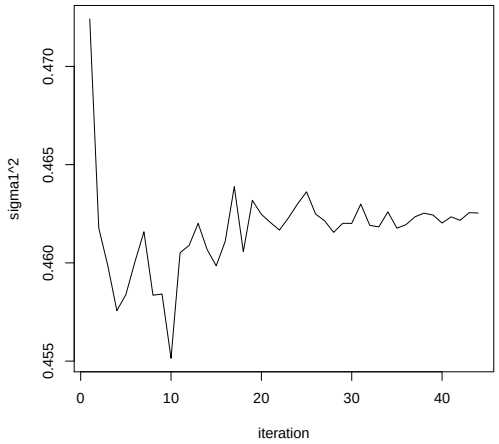

Figure 4.6: Changes for variance component estimate σ_1^2 of one simulated data set

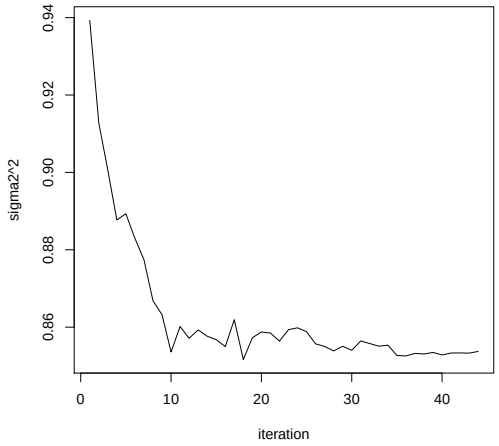

Figure 4.7: Changes for variance component estimate σ_2^2 of one simulated data set

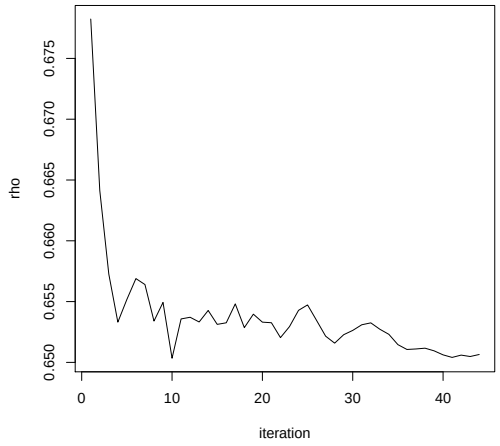

Figure 4.8: Changes for variance component estimate ρ of one simulated data set

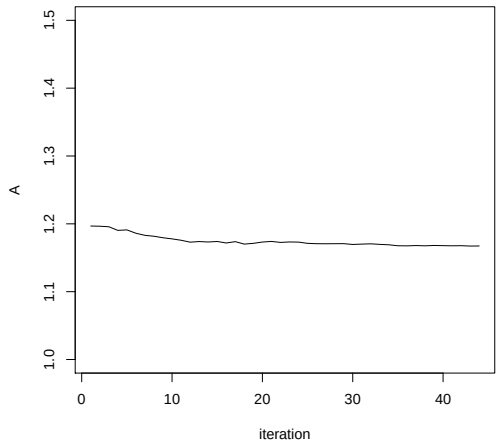

Figure 4.9: Changes for variance component estimate A of one simulated data set

4.4.2 Simulation Study II

In the second simulation study, we reduce the cluster size of repeated measurements from 5 to 3. In each of 100 Monte Carlo data sets, observations y_{1ij} and y_{2ik}, $j = 1, \ldots, n_i = 3, k = 1, \ldots, n_i = 3$, were generated for the ith individual, $i = 1, \ldots, N = 150$. The components in $\mathbf{y}_i = (\mathbf{y}'_{1i}, \mathbf{y}'_{2i})'$, are assumed to be conditionally independent given the latent variable $\mathbf{u}_i = (\mathbf{u}'_{1i}, \mathbf{u}'_{2i})'$.

The conditional distributions of the continuous and count responses are $y_{1ij}|u_{1ij} \sim N(\mu_{1ij}, \sigma^2)$, where

$$\mu_{1ij} = \beta_{10} + x_{1ij}\beta_{11} + u_{1ij},$$

and $y_{2ik}|u_{2ik} \sim$ Poisson (μ_{2ik}), where

$$\log(\mu_{2ik}) = \beta_{20} + x_{2ik}\beta_{21} + u_{2ik}.$$

We choose the autoregressive covariance structure (4.6) for the latent variable $\mathbf{u}_i$. The data were simulated according to the above model with $\beta_{10} = 6.71$, $\beta_{11} = 0.35$, $\beta_{20} = 1$, $\beta_{21} = 0.6$, $\sigma^2 = 0.09$, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 1$, $\rho = 0.7$, $A = 1.2$, $x_{1ij} = j/15$, and $x_{2ik} = k/15$.

In the implementation of the MCEM algorithm, we started at the above values, and used the importance sampling scheme $L \leftarrow L + L/m$ with $m = 10$ and initial $L = 50$. The multivariate Student t distribution with 40 degrees of freedom was chosen as the instrumental distribution. The mean and variance of the t distribution are calculated by (4.17) and (4.18) at each iteration. The

algorithm is stopped when

$$\max|\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}| < \delta,$$

where δ is predetermined as 0.001.

In table 4.2, the average of the estimates over 100 data sets, the average of the estimated standard errors, and the empirical mean square error for the estimates are provided from the third to the fifth columns. By comparing the average of the standard error estimates and the standard deviations of the parameter estimates, the Monte Carlo error can be judged. However, the difference between the two may also reflect inadequacy of asymptotic standard error estimates for the cluster sizes used. The result of Simulation II, with smaller cluster size, exhibits larger standard errors of the estimates and empirical standard errors.

Model	Parameter(True)	Average estimate	Estimated SE	Empirical SE
Linear	$\beta_{10}(6.71)$	6.70955	0.06406	0.07527
	$\beta_{11}(0.35)$	0.31594	0.44483	0.50411
Log-linear	$\beta_{20}(1)$	1.00415	0.09986	0.11104
	$\beta_{21}(0.6)$	0.49150	0.68730	0.73820
Variance	σ^2 (0.3)	0.29647	0.01026	0.03066
Component	σ_1^2 (0.5)	0.50485	0.06656	0.05161
	σ_2^2 (1)	1.00144	0.13212	0.10134
	ρ (0.7)	0.69969	0.03465	0.03441
	A (1.2)	1.19789	0.06049	0.07580

Table 4.2: Simulation results for 100 data sets from normal and Poisson random mean joint model

4.5 Example

Data on 120 patients who received a renal transplant from year 1998 to 2003 were extracted from Organ Procurement and Transplant Network (OPTN) and United Network for Organ Sharing (UNOS), which direct the transplant community to reduce disparity in access to transplant, allocate organs over as wide of a geographic area as possible, and ensure organs to be allocated on the basis of medical necessity. The most common measure of kidney function is considered to be estimated glomerular filtration rate (eGFR), which was repeatedly measured after the first transplant in the study, and calculated by the formula "4-variable MDRD" (serum creatinine, age, race, and gender). Kidney function is considered to be normal, when eGFR is larger than 90. On the other hand, when eGFR is less than 15, the kidney function is treated as a failure by the nephrologist. Patients need to have another kidney transplants or back to dialysis. We model the kidney function status after the first kidney transplant among the following-up approximately equally spaced periods. Specifically, we treat eGFR as the continuous variable and retransplant status as the binary variable. We then combine them together as mixed outcomes to jointly describe how kidney function changes and the disease progresses over time after the first transplant. We model this data set through the random mean joint model to account for the association between mixed outcomes over time following the first kidney transplant, and include covariates, such as, age, gender and time, in the regression model.

Let y_{1ij} and y_{2ik} denote eGFR and the retransplant status, where $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, and $k = 1, \ldots, n_i$. The binary response y_{2ik} takes 1, if the retransplant occurs, otherwise 0. We assume y_{1ij} and y_{2ik} are conditionally

independent given the latent variable $\mathbf{u}_i = (\mathbf{u}'_{1i}, \mathbf{u}'_{2i})'$ with the conditional distributions $y_{1ij}|u_{1ij} \sim N(\mu_{1ij}, \sigma^2)$, where

$$\mu_{1ij} = \beta_{10} + \beta_{11}\text{time} + \beta_{12}\text{age} + \beta_{13}\text{gender} + u_{1ij},$$

and $y_{2ik}|u_{2ik} \sim$ Binary (μ_{2ik}), where

$$\text{logit}(\mu_{2ik}) = \beta_{20} + \beta_{21}\text{time} + \beta_{22}\text{age} + \beta_{23}\text{gender} + u_{2ik}.$$

We assume the distribution of the latent variable $\mathbf{u}_i = (\mathbf{u}'_{1i}, \mathbf{u}'_{2i})'$ is Multivariate Normal with mean $\mathbf{0}$ and covariance matrix Σ_i in (4.6). We applied the Monte Carlo EM algorithm with the importance sampling method. A multivariate Student t distribution with 40 degrees of freedom was chosen as the instrumental distribution. We started with $L = 50$, and increased by $L = L + L/10$, until $L = 5000$.

An important issue in implementing the Monte Carlo EM algorithm is to assess the convergence of the algorithm. We used the criteria that when the relative change in the parameter values from successive iterations is small,

$$\max|\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}| < \delta,$$

where δ is predetermined constant. We set $\delta = 0.0001$. The method involves prespecified $\boldsymbol{\theta}^{(0)}$, and the resulting approximation is local in nature. Thus we iterate our procedure a few times by updating $\boldsymbol{\theta}^{(0)}$ to the current estimate of $\boldsymbol{\theta}$. We also evaluate the marginal likelihood at several parameter values.

The ML estimates of the regression parameters and the variance components are displayed in Table 4.3. The estimate of the time slope of the contin-

Parameter	Value	S.E.
β_{10}(Intercept)	95.9255	0.5846
β_{11}(time)	-8.6714	0.0924
β_{12}(Age)	-1.1467	0.0147
β_{13}(Gender)	14.3259	0.3856
β_{20}(Intercept)	-28.9379	2.2122
β_{21}(time)	3.8350	0.3345
β_{22}(Age)	0.2596	0.0425
β_{23}(Gender)	-4.3563	1.0924
σ_1^2	946.0577	153.5951
σ_2^2	98.0738	16.9252
σ	2.0720	0.5549
ρ	0.7030	0.0393
A	-0.3020	0.0107

Table 4.3: Estimates and standard errors from joint model for the transplant data

uous variable eGFR is -8.6714, which is statistically significant at the 0.001 level, and the negative sign indicates that eGFR decreases over time after the first transplant. In addition, the estimate of the time slope of the binary variable retransplant is 3.835, which is also statistically significant at the 0.001 level. The positive sign shows that the probability of getting a second transplant is increasing over time. The estimate of autocorrelation coefficient ρ in (4.6) is 0.703, which shows that there is a positive correlation in the latent processes U_{1_i} and U_{2_i}, $i = 1, \ldots, N$. The joint risk to a patient of a restransplant treatment and low eGFR reading can be estimated directly.

Amemiya (1985) [3] proposed a two-stage approach, which involves fitting two regression models. After modelling the first outcome as a function of covariates, one models the second outcome, including the predicted values from the first regression as covariates. The two-stage approach typically is used when interest focuses on modelling one of the outcomes, with appropriate

adjustment for the other. However, we are interested in jointly modelling both outcomes. Our approach allows for test on the covariate. For example, the covariate time is considered to affect both outcomes in the study.

4.6 Conclusion

In this chapter, we developed a flexible class of generalized linear latent variable models, random mean joint models, for bivariate mixed responses with clustered and longitudinal structures. We overcame the difficulty in joint modelling of mixed continuous and discrete responses by introducing two cross correlated latent processes. The Kronecker product is adopted in the formulation of the cross-covariance matrix, and the nice properties of the Kronecker product simplify the expression of the log-likelihood function of the joint latent variables, especially the inverse and determinant of the high dimensional covariance matrix. We also connected the joint modelling of two latent processes with autoregressive covariance structures to two time series satisfying a predictive equation. However, this simplicity of the cross-covariance matrix is sacrificed by the freedom of specifying different correlation coefficients of two latent processes. That is to say, random mean joint models require that two latent processes have the same correlation function.

For the statistical inference, we applied the MCEM algorithm to find the MLEs of regression coefficients and variance components of the random mean joint models, by treating the latent variables as missing data. The parameters that characterize the unobservable latent variables and coefficient A can be estimated directly. In the implementation of the MCEM algorithm, at each E-step, we approximated the Q function by using the importance sampling

approach and used the Laplace approximation of the posterior distribution of the latent variables given the responses to find the approximate the mean and covariance of the instrumental distribution.

We demonstrate the methodology with two simulations and a kidney study data set. The simulation results show that the estimates of parameters in the regression model and the variance components are unbiased. The Monte Carlo error is relatively smaller, when the cluster size is larger. The difference between the average of the standard error estimates and the standard deviations of the parameter estimates may reflect inadequacy of asymptotic standard error estimates for the cluster sizes used. For simulations with smaller cluster size, the result exhibits larger standard errors of the estimates and empirical standard errors, when compared to simulations with larger clustered size.

Chapter 5

Zero Modified Models

5.1 Introduction

In recent years there has been considerable interest in models for count data that allow for excess zeros. A large amount of attention has been paid to dealing with such data. Ridout et al. (1998) [78] review the literature and cite examples of applications from manufacturing, patent applications, road safety, biology, medical consultations, and the use of recreational facilities, and others.

The feature of count data with excess zeros is that they are often overdispersed relative to Poisson distribution. This overdispersion does not arise from heterogeneity, as is the case when the Poisson model is generalized to the negative binomial model. Instead, it arises from the large frequencies of zeros. In practice, the presence of overdispersion may come from one or both of these sources (Mullahy (1986) [69]). The variance-mean relationship must be correctly modelled.

Mullahy (1986) [69], Heilbron (1989), (1994) [46] [47], and Lambert (1992)

[55] pioneered regression models based on zero-inflated Poisson (ZIP) distribution. In a ZIP regression model, the count response variable is assumed to be distributed as a mixture of a Poisson(λ) distribution and a distribution with point mass of one at zero, with mixing probability p. Both p and λ are allowed to depend on covariates through the link functions. Hall (2000) [43] adapted Lambert's methodology to an upper bounded count situation, and introduced the zero-inflated binomial (ZIB) regression models.

In addition to cross-sectional data, zero inflation may also occur with repeated measures or longitudinal data. Many researchers have incorporated random effects into a wide variety of regression models to account for correlated responses and multiple sources of variance. In a mixture model context, Duijn and Bockenholt (1995) [23] presented a latent class Poisson model for analyzing overdispersed repeated count data. Zero-inflated regression models for continuous data with repeated measures have also been considered by Olsen and Schafer (2001) [72], Berk and Lachenbruch (2002) [5], Tooze et al. (2002) [96], and Yau et al. (2002) [103]. Hall (2000) incorporated random effects into the ZIP and ZIB models to accommodate the repeated measures, so the within-subject correlation and between-subject heterogeneity typical of repeated measures can be accommodated.

In many applications where a preponderance of zero counts is observed, it is important to assess whether the ZIP model assumption is indeed appropriate. In the literature, Broek (1995) [9] considered a score test for testing a standard Poisson regression model against a zero-inflated Poisson alternative under the framework of the zero-inflated Poisson regression model with a constant proportion of excess zeros. It was extended to the general situation where the zero probability is allowed to depend on covariates (see Jansakul and Hinde (2002)

[50]). Deng and Paul (2000) [21] developed the score tests for the generalized linear models against zero inflation. Ridout et al. (2001) [79] provide a score test for testing zero-inflated Poisson regression models against zero-inflated negative binomial alternatives. For correlated count data, Xiang et al. (2006) [101] develop score test for testing the zero-inflation Poisson mixed regression model.

In this chapter, we model the count data with excess zeros from a different point of view, and hence the statistical inferential method. The hurdle model (Mullahy (1986)), and the two-part model (Heilbron (1994)) for count data are special cases of our model, while ours can accommodate a broad class of distributions. For example, our model can handle a semicontinuous variable, which has a portion of responses equal to a single value (typically 0) and a continuous distribution among the remaining values (Feuerverger (1979) [27], Farewell (1986) [26], Meeker (1987) [66]).

We are seeking to investigate the modelling of count data with excess zeros through the $(a, b, 0)$ class and $(a, b, 1)$ class of distributions, the zero truncated distribution, and the zero modified distribution, which are commonly used in the actuarial and econometric literature. The zero-inflated distribution (Lambert (1992) [55], Hall (2000) [43]) is a reparametrization of the zero modified distribution. However, the statistical inference based on the two different formulations would be totally different as well as the interpretation of the parameters.

We start the discussion of modelling data from the truncated distribution. We verify that the zero truncated Poisson and zero truncated binomial distributions are members of the exponential family. For the statistical inference of the generalized linear models for the truncated distributions, the commonly

used algorithms Newton Raphson and Fisher Scoring still have the same form as the one of the generalized linear models, except for the mean and variance replaced by the truncated mean and the truncated variance. We then extend the modelling of the zero truncated data to data with excess zeros, by introducing an indicator variable to identify the responses as zero or nonzero. A regression model for the probability of nonzero response, and a conditional model for nonzero responses, are two model components which describe this type of data.

We also propose zero modified random effects models for clustered data with excess zeros. We discuss the maximum likelihood estimation of parameters in the zero modified random effects models by approximate Fisher Scoring algorithm based on the adaptive Gaussian quadrature approximations. Simulations for two regression models including random intercepts are conducted, and a real data example is used to illustrate the new method. We then extend random mean models introduced in Chapter 3 to data with excess zeros, and formulate the corresponding zero modified random mean models. A simulation is conducted to evaluate random mean model for the temporal count data with excess zeros.

5.2 Zero-inflated Poisson Regression

We shall consider data with excess zeros particularly in relation to the Poisson distribution, but the term may be used in conjunction with any discrete distribution to indicate that there are more zeros than would be expected on the basis of the non-zero counts. Of course it is also possible to have fewer zero counts than expected, but this is much less common in practice. Lambert

(1992) described zero-inflated Poisson distribution regression for count data with excess zeros. We present Lambert's model to elaborate the terminologies, which appear in this chapter.

The zero-inflated Poisson (ZIP) distribution is an extreme case of mixture distribution. A proportion p of data take value 0, and the remainder proportion $1-p$ follow Poisson distribution with parameter λ. The probability function of a zero-inflated Poisson random variable Y is

$$\Pr(Y=y) = \begin{cases} p + (1-p)e^{-\lambda}, & y = 0, \\ (1-p)e^{-\lambda}\lambda^y/y!\,, & y > 0. \end{cases} \tag{5.1}$$

It is possible for p in equation (5.1) to assume negative values, resulting in a zero-deflated distribution. Zero-deflated data seldom arise in practice, however, and we shall assume $0 \leq p < 1$. Zero-inflated forms of other count distributions, such as the negative binomial, can be defined similarly. Gupta et al. (1996) [42], for example, investigate the zero-inflated form of the generalized Poisson distribution.

For the zero-inflated Poisson distribution, the mean and the variance are

$$\begin{aligned} \mathrm{E}(Y) &= (1-p)\lambda = \mu, \\ \mathrm{Var}(Y) &= \mu + \left(\frac{p}{1-p}\right)\mu^2. \end{aligned}$$

Lambert (1992) considered models in which the Poisson and proportion parameters depend on the covariates

$$\begin{aligned} \log(\lambda_i) &= \mathbf{x}_i'\boldsymbol{\beta}, \\ \mathrm{logit}(p_i) &= \mathbf{g}_i'\boldsymbol{\gamma}, \end{aligned}$$

where $\mathbf{x}_i$ and $\mathbf{g}_i$ are vectors of covariates, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of parameters. The two sets of covariates may or may not coincide. When they do coincide, more parsimonious models may be developed by supposing that the two linear predictors are related in some way. Lambert proposed such a simple model, ZIP(τ) model, which has

$$\log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta},$$
$$\text{logit}(p_i) = \tau\mathbf{x}_i'\boldsymbol{\beta},$$

where τ is a scalar parameter. A great variety of alternative models can be generated by using different link functions for λ and/or p. Greene (1994) [39] gives details of analogous zero-inflated negative binomial regression models.

5.3 Modification and Truncation at Zero

In this section, the definitions of the $(a, b, 0)$ class, the $(a, b, 1)$ class, the zero truncated distribution, and the zero modified distribution from Klugman et al. (2009) [53] will be provided.

Definition: The distribution of a discrete random variable is said to be a member of the $(a, b, 0)$ class of distributions if its probability mass function p_k satisfies

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \ k = 1, 2, 3, \ldots,$$

for any constants a and b, where p_0 is a constant satisfying $0 < p_0 < 1$.

The binomial distribution, the Poisson distribution and the negative bino-

Distribution	a	b	p_0
Poisson	0	λ	$e^{-\lambda}$
Binomial	$-\frac{p}{1-p}$	$(m+1)\frac{p}{1-p}$	$(1-p)^m$
Negative Binomial	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$(1+\beta)^{-r}$

Table 5.1: Members of the $(a, b, 0)$ class

mial distribution belong to this class of distributions, with each distribution represented by a different sign of a. The parameters of these distributions are determined by both a and b. By substituting in the probability function for each of the Poisson, binomial, and negative binomial distributions on the left-hand side of the recursion, it can be seen that each of these three distributions satisfies the recursion.

At times, the distributions in the $(a, b, 0)$ class do not adequately describe the characteristics of data encountered in practice. This is because the tail of the negative binomial is not heavy enough or because the distributions in the $(a, b, 0)$ class cannot capture the shape of the data set in some other part of the distribution.

The problem of a poor fit at the left-hand end of the distribution, in particular, the probability at zero, is addressed through an adjustment of the probability at zero. It is easily handled for the Poisson, binomial, and negative binomial distributions.

Definition: Let p_k be the probability function of a discrete random variable. It is a member of the $(a, b, 1)$ class of distributions provided that there exists constants a and b such that

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \ k = 2, 3, 4, \ldots,$$

where p_1 is a constant satisfying $0 < p_1 < 1$.

The only difference between the $(a, b, 1)$ class and the $(a, b, 0)$ class is that the recursion starts at p_1 rather than p_0. The shape of the distribution function from $k = 1$ to $k = \infty$ is the same as the one in the $(a, b, 0)$ class, up to a scale constant. The effect of the scale constant is either to stretch or to contract the graph of the distribution function. It is flexible to set the scale constant, since the summation of $\sum_{k=1}^{\infty} p_k$ and p_0 equals to 1.

In the situation where $p_0 = 0$, we name such type of distribution as **zero truncated distribution**, noted as p_k^T. Specifically, there are zero truncated Poisson, zero truncated binomial, and zero truncated negative binomial distributions. In the other situation, where $p_0 > 0$, we name this type of distribution as **zero modified distribution**, noted as p_k^M. The zero modified distributions can be viewed as a mixture of an $(a, b, 0)$ distribution and a degenerate distribution with all the probability at zero. The zero truncated distributions can be considered as special case of the zero modified distributions, with $p_0 = 0$.

In general, the probability distribution function of a zero truncated distribution is

$$p_k^T = \frac{1}{1 - p_0} p_k, \ k = 1, 2, \ldots,$$

and the probability distribution function of a zero modified distribution is

$$p_k^M = (1 - p_0^M) p_k^T = \frac{1 - p_0^M}{1 - p_0} p_k, \ k = 1, 2, \ldots,$$

where p_0^M is an arbitrary number between 0 and 1, assigning to 0. The zero modified distribution is also the weighted average of a degenerate distribution

Distribution	Mean	Variance
Poisson	$\lambda/(1-e^{-\lambda})$	$\lambda[1-(\lambda+1)e^{-\lambda}]/(1-e^{-\lambda})^2$
Binomial	$\frac{mp}{1-(1-p)^m}$	$\frac{mp[(1-p)-(1-p+mp)(1-p)^m]}{[1-(1-p)^m]^2}$
Negative Binomial	$\frac{r\beta}{1-(1+\beta)^{-r}}$	$\frac{r\beta[(1+\beta)-(1+\beta+r\beta)(1+\beta)^{-r}]}{[1-(1+\beta)^{-r}]^2}$

Table 5.2: Mean and variance of zero truncated distributions

and the zero truncated member of the $(a, b, 0)$ class.

The mean and variance of zero truncated distributions are summarized in table 5.2. The mean and variance of the corresponding zero modified distribution are

$$\begin{aligned} \mathrm{E}(Y^M) &= (1-p_0^M)\mathrm{E}(Y^T), \\ \mathrm{Var}(Y^M) &= (1-p_0^M)\mathrm{Var}(Y^T) + p_0^M(1-p_0^M)\mathrm{E}(Y^T)^2. \end{aligned}$$

5.4 Truncated Distributions

In this section, we verify that zero truncated Poisson and zero truncated binomial distributions are members of the exponential family.

5.4.1 Truncated Poisson Distribution

The probability mass function of zero truncated Poisson distribution is

$$f(y_i) = \frac{1}{1-e^{-\lambda_i}} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \ y_i = 1, 2, 3, \ldots,$$

for some $\lambda_i > 0$. Taking logs we find

$$\log\, f(y_i) = y_i \log(\lambda_i) - \lambda_i - \log(1-e^{-\lambda_i}) - \log(y_i!\,).$$

Looking at the coefficient of y_i, we see immediately that the canonical parameter is

$$\theta_i = \log(\lambda_i).$$

The second and third terms in the p.d.f are

$$b(\theta_i) = e^{\theta_i} + \log(1 - e^{-e^{\theta_i}}).$$

The last remaining term is a function of y_i only, so we identify

$$c(y_i, \phi) = -\log(y_i!\,),$$

and take $a_i(\phi) = \phi$ and $\phi = 1$. We have verified that zero truncated Poisson distribution belongs to the exponential family. Also, we can show the mean and variance are the first and the second derivatives of the cumulant function $b(\theta_i)$

$$\begin{aligned} \mathrm{E}(Y_i) &= b'(\theta_i) = \frac{e^{\theta_i}}{1 - e^{-e^{\theta_i}}} \\ \mathrm{Var}(Y_i) &= b''(\theta_i) = e^{\theta_i}[1 - (e^{\theta_i} + 1)e^{-e^{\theta_i}}]/(1 - e^{-e^{\theta_i}})^2. \end{aligned}$$

5.4.2 Truncated Binomial Distribution

The probability mass function of zero truncated binomial distribution is

$$f(y_i) = \frac{1}{1 - (1 - \pi_i)^{m_i}} \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}, \;\; y_i = 1, 2, \ldots, m_i,$$

for some $\pi_i > 0$. Taking logs we find that

$$\log\ f(y_i) = y_i \log \frac{\pi_i}{1-\pi_i} - \log[(1 + e^{\log \frac{\pi_i}{1-\pi_i})^{m_i}} - 1] + \log \binom{m_i}{y_i}.$$

Looking at the coefficient of y_i, we see immediately that the canonical parameter is

$$\theta_i = \log \frac{\pi_i}{1-\pi_i}.$$

The second term in the p.d.f is

$$b(\theta_i) = \log[(1 + e^{\theta_i})^{m_i} - 1].$$

The last remaining term is a function of y_i only, so we identify

$$c(y_i, \phi) = \log \binom{m_i}{y_i},$$

and take $a_i(\phi) = \phi$ and $\phi = 1$. We have verified that the truncated binomial distribution belongs to the exponential family. Also, we can show the mean and variance are the first and the second derivatives of the cumulant function $b(\theta_i)$

$$\begin{aligned}
\mathrm{E}(Y_i) &= b'(\theta_i) = \frac{m_i e^{\theta_i}(1 + e^{\theta_i})^{m_i - 1}}{(1 + e^{\theta_i})^{m_i} - 1} \\
\mathrm{Var}(Y_i) &= b''(\theta_i) = \frac{m_i e^{\theta_i}(1 + e^{\theta_i})^{2m_i - 2} - m_i e^{\theta_i}(1 + m_i e^{\theta_i})(1 + e^{\theta_i})^{m_i - 2}}{[(1 + e^{\theta_i})^{m_i} - 1]^2}.
\end{aligned}$$

5.5 Generalized Linear Models for Truncated Data

The Generalized Linear Models defined in Nelder and Wedderburn (1972) [71] are characterized by

- A dependent variable y whose distribution with parameter θ is one of the class in the exponential family.
- A set of independent variables $x_1, \ldots, x_p$ is related to the linear predictor through $\eta = \sum \beta_i x_i$.
- A linking function $\eta = g(\theta)$ connecting the parameter θ of the distribution of y with the η's of the linear model.

In this section, we mainly discuss the truncated Poisson and the truncated binomial distributions. However, the truncated distributions for other members in the exponential family can be formulated accordingly.

5.5.1 Truncated Poisson Regression Models

First, the likelihood function of responses from zero truncated Poisson distribution is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Pr(Y_i = y_i) = \prod_{i=1}^{n} \frac{1}{1 - e^{-\lambda_i}} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!},$$

with parameter λ_i of Poisson distribution related to various predictors through the log link function

$$\log(\lambda_i) = \mathbf{x}_i' \boldsymbol{\beta}.$$

Then the log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log(\lambda_i) - \lambda_i - \log(1 - e^{-\lambda_i})\} - \sum_{i=1}^{n} \log(y_i!).$$

To find the maximum likelihood estimate of $\boldsymbol{\beta}$, we first calculate the score function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 - e^{-e^{\mathbf{x}_i'\boldsymbol{\beta}}}} \right),$$

and then solve the equation $S(\boldsymbol{\beta}) = 0$ to get $\hat{\boldsymbol{\beta}}$. We may not find the closed form of the MLEs in most cases. Newton Raphson and Fisher Scoring algorithms are used to find the numeric solution. The observed Fisher information matrix is

$$I(\boldsymbol{\beta}) = \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i,$$

where W_i has the truncated variance as the diagonal element, $e^{\mathbf{x}_i'\boldsymbol{\beta}}[1 - (e^{\mathbf{x}_i'\boldsymbol{\beta}} + 1)e^{-e^{\mathbf{x}_i'\boldsymbol{\beta}}}]/(1 - e^{-e^{\mathbf{x}_i'\boldsymbol{\beta}}})^2$. The updating formula for the Newton Raphson and the Fisher-Scoring algorithm is given by

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + I(\boldsymbol{\beta}^{(r)})^{-1} S(\boldsymbol{\beta}^{(r)}).$$

5.5.2 Truncated Binomial Regression Models

First, the likelihood function of responses from zero truncated binomial distribution is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Pr(Y_i = y_i) = \prod_{i=1}^{n} \frac{1}{1-(1-\pi_i)^{m_i}} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i},$$

with parameter π_i of binomial distribution related to various predictors through the logit link function

$$\text{logit}(\pi_i) = \mathbf{x}_i' \boldsymbol{\beta}.$$

Then the log-likelihood function is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \text{log} \frac{\pi_i}{1-\pi_i} + m_i \text{log}(1-\pi_i) - \text{log}[1-(1-\pi_i)^{m_i}] - \sum_{i=1}^{n} \text{log} \binom{m_i}{y_i}.$$

To find the maximum likelihood estimate of $\boldsymbol{\beta}$, we first calculate the score function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \frac{m_i e^{\mathbf{x}_i' \boldsymbol{\beta}} (1+e^{\mathbf{x}_i' \boldsymbol{\beta}})^{m_i - 1}}{(1+e^{\mathbf{x}_i' \boldsymbol{\beta}})^{m_i} - 1} \right),$$

and then solve the equation $S(\boldsymbol{\beta}) = 0$ to get $\hat{\boldsymbol{\beta}}$. We may not find the closed form of the MLEs in most cases. Newton Raphson and Fisher Scoring algorithms are used to find the numeric solution. The observed Fisher information matrix is

$$I(\boldsymbol{\beta}) = \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i,$$

where W_i has the truncated variance as the diagonal element

$$\frac{m_i e^{\mathbf{x}_i'\boldsymbol{\beta}}(1+e^{\mathbf{x}_i'\boldsymbol{\beta}})^{2m_i-2} - m_i e^{\mathbf{x}_i'\boldsymbol{\beta}}(1+m_i e^{\mathbf{x}_i'\boldsymbol{\beta}})(1+e^{\mathbf{x}_i'\boldsymbol{\beta}})^{m_i-2}}{[(1+e^{\mathbf{x}_i'\boldsymbol{\beta}})^{m_i}-1]^2}.$$

The updating formula for the Newton Raphson and the Fisher-Scoring algorithm is given by

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + I(\boldsymbol{\beta}^{(r)})^{-1} S(\boldsymbol{\beta}^{(r)}).$$

5.5.3 Discussion

Zero-inflated forms of other count distributions, such as the negative binomial, can be defined similarly. The negative binomial distribution belongs to $(a, b, 0)$ class, however, is not a member of the exponential family. Only the negative binomial distribution with known stopping-time parameter belongs to the exponential family. Grogger and Carson (1991) [40] discuss the fitting of zero truncated negative binomial models. The Newton Raphson or Fisher Scoring algorithms for the maximum likelihood estimates of the negative binomial regression models can still be derived. Greene (1994) [39] gives details of zero modified negative binomial models. There is always a connection between the negative binomial regression and the Poisson regression. If we consider a random effects model where the multiplicative random effect is a gamma variable with unit mean, accordingly, the marginal distribution of the response is negative binomial (Greene (1994)). In this chapter, we will skip the negative binomial distribution.

5.6 Zero Modified Regression Models

For the modelling of data with excess zeros in cross-sectional study, we introduce an indicator variable to identify the responses as zero or nonzero. A regression model for the probability of zero response, and a conditional model for nonzero responses are components to describe this type of data.

5.6.1 Zero Modified Poisson Regression Model

The probability function of zero modified Poisson distribution is

$$f_i(y_i) = \begin{cases} \frac{1-p_i^M}{1-e^{-\lambda_i}} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, & y_i = 1, 2, 3, \ldots, \\ p_i^M, & y_i = 0. \end{cases} \tag{5.2}$$

When $p_i^M = 1$, we have the trivial case in which only zeros occur. When $p_i^M = 0$, it is simply zero truncated Poisson distribution. We model λ_i and p_i^M with log-linear and logistic regression models

$$\log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta}$$
$$\text{logit}(p_i^M) = \mathbf{g}_i'\boldsymbol{\gamma}, \ i = 1, \ldots, n,$$

where $\mathbf{x}_i$ and $\mathbf{g}_i$ are vectors of known covariate values associated with the response, Y_i.

Let $\boldsymbol{\psi} = (\boldsymbol{\gamma}', \boldsymbol{\beta}')'$ be the parameter vector. We introduce an indicator variable $I_i = 1$ if $Y_i = 0$, and $I_i = 0$ otherwise. The likelihood function of $\boldsymbol{\psi}$ is given by

$$L(\boldsymbol{\psi}; y) = \prod_{i=1}^{n} (p_i^M)^{I_i} (1 - p_i^M)^{1-I_i} \left[\frac{e^{-\lambda_i} \lambda_i^{y_i}}{(1 - e^{-\lambda_i})\ y_i!} \right]^{1-I_i}.$$

We break zero modified Poisson regression model into two separate ones: the logistic regression model for observations identified as either zero or nonzero; and zero truncated Poisson regression model for nonzero observations. The fitting of parameters of the latter was discussed in the previous section.

When the same covariates affect λ_i and p_i^M, it is useful to consider a model that involves the complementary-log-log link function for p_i^M

$$\log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta}$$
$$\log[-\log(1-p_i^M)] = \mathbf{g}_i'\boldsymbol{\gamma},\ i = 1, \ldots, n.$$

This model was first proposed by Mullahy (1986) [69]. The above model reduces to the Poisson regression model, when $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the same. A likelihood ratio test for testing Poisson regression model against zero modified Poisson regression model as the alternative can be derived for the complementary-log-log link function for the Bernoulli success probability of zeros. Twice the difference of log-likelihoods should be approximately chi-square. The degrees of freedom should be equal to the difference in the number of parameters in the two models.

5.6.2 Zero Modified Binomial Regression Model

The probability function of zero modified binomial distribution is

$$f_i(y_i) = \begin{cases} \frac{1-p_i^M}{1-(1-\pi_i)^{m_i}} \binom{m_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{m_i-y_i}, & y_i = 1, 2, \ldots, m_i, \\ p_i^M, & y_i = 0. \end{cases}$$

When $p_i^M = 1$, we have the trivial case in which only zeros occur. When

$p_i^M = 0$, it is simply zero truncated binomial distribution. We model λ_i and p_i^M with logistic regression models

$$\text{logit}(\pi_i) = \mathbf{x}_i'\boldsymbol{\beta}$$
$$\text{logit}(p_i^M) = \mathbf{g}_i'\boldsymbol{\gamma},$$

where $\mathbf{x}_i$ and $\mathbf{g}_i$ are vectors of known covariate values associated with the response, $Y_i, i = 1, \ldots, N$.

Let $\boldsymbol{\psi} = (\boldsymbol{\gamma}', \boldsymbol{\beta}')'$ be the parameter vector. We introduce an indicator variable $I_i = 1$ if $Y_i = 0$, and $I_i = 0$ otherwise. The likelihood function of $\boldsymbol{\psi}$ is given by

$$L(\boldsymbol{\psi}; y) = \prod_{i=1}^{n} (p_i^M)^{I_i}(1 - p_i^M)^{1-I_i}\left[\frac{1}{1-(1-\pi_i)^{m_i}}\binom{m_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{m_i-y_i}\right]^{1-I_i}.$$

We break zero modified binomial regression model into two separate ones: the logistic regression model for observations identified as either zero or nonzero; and zero truncated binomial regression model for the nonzero observations. The fitting of parameters in the latter was discussed in the previous section.

5.6.3 General Hypothesis Testing: Likelihood Ratio Test

The hypothesis testing problem is to compare two nested models: the model under H_0 (nested model) with p_1 parameters, and the model under H_A with p_2 parameters (full model). Suppose $\hat{\theta}_{H_0}$ is the MLE of θ under null hypothesis, and $\hat{\theta}_{H_A}$ is the MLE under alternative hypothesis. The likelihood ratio test is

defined as

$$\Lambda = 2\{l(\hat{\theta}_{H_A}) - l(\hat{\theta}_{H_0})\}.$$

Since the likelihood is always larger for the full model, we will have that $\Lambda \geq 0$. If the data come from the model under H_0, the two models almost give the same fitting to the data and therefore the statistic Λ should not take too large values. On the other hand, if the data come from the model under H_A, the full model provides much better fitting and therefore Λ should be quite large.

Theorem 5.1: *When the sample size is large, under usual regularity conditions*

$$\Lambda \sim \chi^2_\nu \text{ approximately,}$$

where $\nu = p_2 - p_1$ *degrees of freedom.*

5.6.4 Example

The biologists are interested in how many fish are caught by fishermen at a state park. Visitors are surveyed how long they stayed, how many people were in the group, whether there were children in the group and how many fish were caught. Some visitors do not fish, but there is no information on whether a person fished or not. Some visitors who did fish but did not catch any fish. So there are excess zeros in the data set, because some people did not fish.

We have data on 250 groups that have been to the park. Each group was questioned about how many fish they caught (count), how many children were in the group (child), how many people were in the group (persons), and

Model	Parameter	Estimate	Standard Error	P-value
Log-liner	intercept(β_0)	-0.8262	0.1723	<0.0001
	persons(β_1)	0.8348	0.0441	<0.0001
	child(β_2)	-1.1390	0.0929	<0.0001
	camper(β_3)	0.7336	0.0934	<0.0001
Logit	intercept(γ_0)	2.3087	0.4612	<0.0001
	persons(γ_1)	-1.1104	0.1911	<0.0001
	child(γ_2)	2.1380	0.3107	<0.0001
	camper(γ_3)	-1.0179	0.3246	0.0019

Table 5.3: Estimates of parameters in the zero modified Poisson regression model

whether or not they brought a camper to the park (camper).

Besides predicting the number of fish caught, biologists are interested in predicting the occurrence of excess zeros. Both bad luck fishing and no fishing during the visit are the possible sources. We include child, persons, and camper as covariates in the zero modified Poisson regression model. We model the Poisson parameter and Bernoulli probability through the log-linear and logistic regression models

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{persons} + \beta_2 \text{child} + \beta_3 \text{camper}$$
$$\text{logit}(p_i^M) = \gamma_0 + \gamma_1 \text{persons} + \gamma_2 \text{child} + \gamma_3 \text{camper}, \ \ i = 1, \ldots, 250.$$

The first block of Table 5.3 includes estimates and standard errors of coefficients in the log-linear model. The second block corresponds to the logit model predicting the zeros. All the predictors in both the log-linear and inflation portions models are statistically significant. The interpretation of the parameters in zero truncated Poisson regression model is different from that of the standard Poisson regression. The exponential of the coefficient is the relative change in Poisson parameter, for one-unit increase of the corresponding

Model	Parameter	Estimate	Standard Error	P-value
Log-liner	intercept(β_0)	-0.8262	0.1723	<0.0001
	persons(β_1)	0.8348	0.0441	<0.0001
	child(β_2)	-1.1390	0.0929	<0.0001
	camper(β_3)	0.7336	0.0934	<0.0001
Complementary-log-log	intercept(γ_0)	1.0106	0.2639	0.0002
	persons(γ_1)	-0.7034	0.1266	<0.0001
	child(γ_2)	1.3260	0.1791	<0.0001
	camper(γ_3)	-0.6183	0.1942	0.0016

Table 5.4: Estimates of parameters in the zero modified Poisson regression model with the complementary-log-log link function

predictor. Based on the logit model, we can predict the probability of zeros and non zeros. The mean of the modified distribution is the product of the predicted probability of non zeros and the mean of the truncated distribution.

The change in $\log(\lambda)$ for one-unit increase of child was -1.1390. Groups with campers (camper = 1) had $\log(\lambda)$ 0.7336 higher than groups without campers (camper = 0).

The log-likelihoods of the full model and the null model are -751.61823 and -1127.02294, respectively. The chi-squared value is 2(-751.61823 + 1127.0229) = 750.8094. Since we have six predictor variables in the full model, the degrees of freedom for the chi-squared test is 6. This yields a p-value <.0001. Thus, the overall model is statistically significant.

We also model the data through the log-linear and complementary-log-log link functions

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{persons} + \beta_2 \text{child} + \beta_3 \text{camper},$$

$$\log[-\log(1 - p_i^M)] = \gamma_0 + \gamma_1 \text{persons} + \gamma_2 \text{child} + \gamma_3 \text{camper}, \ i = 1, \ldots, 250,$$

where λ_i is the Poisson parameter.

Model	Parameter	Estimate	Standard Error	P-value
Log-liner	intercept(β_0)	-1.9818	0.1523	<0.0001
	persons(β_1)	1.0913	0.0393	<0.0001
	child(β_2)	-1.6900	0.0810	<0.0001
	camper(β_3)	0.9309	0.0891	<0.0001

Table 5.5: Estimates of parameters in the Poisson regression model

The first block of Table 5.4 includes estimates and standard errors of the log-linear model. The second block presents estimates and standard errors of the complementary-log-log model predicting the zeroes. All the predictors in both log-linear and inflation portions of the model are statistically significant. We can predict the probability of zeros and non zeros from the complementary-log-log model.

To assess whether the zero modified Poisson model assumption is indeed appropriate, we fit the Poisson regression model as the null model to the alternative zero modified Poisson regression model with the complementary-log-log link function for the probability of zeros. The -2log-likelihoods = 2(1674.14-1505.9) = 336.48. The degrees of freedom for the chi-squared test is 4. This yields a p-value <.0001. Thus, the overall model is statistically significant.

5.7 Zero Modified Random Effects Models

Starting from this section, we will discuss zero modified regression models for longitudinal and clustered data. We first formulate zero modified random effects models for repeated measures with clumping at zero.

5.7.1 Zero Modified Poisson Random Effects Models

Suppose the response $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_N')'$ contains data from N clusters, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, $i = 1, \ldots, N$. We assume that conditional on the random effect $\mathbf{b}_i$, the response Y_{ij} is from zero modified Poisson distribution with probability function

$$\Pr(Y_{ij} = k) = \begin{cases} \frac{1-p_{ij}^M}{1-e^{-\lambda_{ij}}} \frac{\lambda_{ij}^k e^{-\lambda_{ij}}}{k!}, & k = 1, 2, 3, \ldots \\ p_{ij}^M, & k = 0, \end{cases}$$

for $i = 1, \ldots, N,\ j = 1, \ldots, n_i$. We model λ_{ij} and p_{ij}^M through the log-linear and logistic regression models

$$\log(\lambda_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i$$
$$\text{logit}(p_{ij}^M) = \mathbf{g}_{ij}'\boldsymbol{\gamma},$$

where $\mathbf{x}_{ij}$, $\mathbf{z}_{ij}$ and $\mathbf{g}_{ij}$ are vectors of known covariate values associated with the response, Y_{ij}. We assume the random effect $\mathbf{b}_i$ is the random variable with a parametric density function $q_i(\mathbf{b}_i; \Sigma_i(\boldsymbol{\sigma}^2))$.

Let $\boldsymbol{\psi} = (\boldsymbol{\gamma}', \boldsymbol{\beta}', \boldsymbol{\sigma}^2)'$ be the parameter vector. We introduce an indicator variable to identify the response as zero or nonzero, $I_{ij} = 1$ if $Y_{ij} = 0$, and $I_{ij} = 0$ otherwise.

The likelihood function of $\boldsymbol{\psi}$ is given by

$$L(\boldsymbol{\psi}; \mathbf{y}) = \prod_{i=1}^{N} \int f(\mathbf{y}_i | \mathbf{b}_i; \boldsymbol{\gamma}, \boldsymbol{\beta})\ q_i(\mathbf{b}_i; \Sigma_i)\ \mathrm{d}\mathbf{b}_i,$$

where

$$f(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\gamma},\boldsymbol{\beta}) = \prod_{j=1}^{n_i} (p_{ij}^M)^{I_{ij}} (1-p_{ij}^M)^{1-I_{ij}} \left[\frac{e^{-\lambda_{ij}} \lambda_{ij}^{y_{ij}}}{(1-e^{-\lambda_{ij}})\ y_{ij}!} \right]^{1-I_{ij}}.$$

5.7.2 Zero Modified Binomial Random Effects Models

Suppose the response vector $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_N')'$ contains data from N clusters, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, $i = 1, \ldots, N$. We assume that, conditional on the random effect $\mathbf{b}_i$, the response Y_{ij} is from zero modified binomial distribution with probability function

$$\Pr(Y_{ij} = k) = \begin{cases} \frac{1-p_{ij}^M}{1-(1-\pi_{ij})^{m_{ij}}} \binom{m_{ij}}{y_{ij}} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{m_{ij}-y_{ij}}, & k = 1, 2, \ldots, m_{ij}, \\ p_{ij}^M, & k = 0, \end{cases}$$

for $i = 1, \ldots, N,\ j = 1, \ldots, n_i$. We model π_{ij} and p_{ij}^M through the logistic regression models

$$\begin{aligned} \text{logit}(\pi_{ij}) &= \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i \\ \text{logit}(p_{ij}^M) &= \mathbf{g}_{ij}'\boldsymbol{\gamma}, \end{aligned}$$

where $\mathbf{x}_{ij}$, $\mathbf{z}_{ij}$ and $\mathbf{g}_{ij}$ are vectors of known covariate values associated with the response, Y_{ij}. We assume the random effect $\mathbf{b}_i$ is the random variable with a parametric density function $q_i(\mathbf{b}_i; \Sigma_i(\boldsymbol{\sigma}^2))$.

Let $\boldsymbol{\psi} = (\boldsymbol{\gamma}', \boldsymbol{\beta}', \boldsymbol{\sigma}^2)'$ be the parameter vector. We introduce a variable $I_{ij} = 1$ if $Y_{ij} = 0$, and $I_{ij} = 0$ otherwise.

The likelihood function for $\boldsymbol{\psi}$ is given by

$$L(\boldsymbol{\psi};\mathbf{y}) = \prod_{i=1}^{N} \int f(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\gamma},\boldsymbol{\beta})\ q_i(\mathbf{b}_i;\Sigma_i)\ \mathrm{d}\mathbf{b}_i,$$

where

$$f(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\gamma},\boldsymbol{\beta}) = \prod_{j=1}^{n_i} (p_{ij}^M)^{I_{ij}} (1-p_{ij}^M)^{1-I_{ij}} \left[\frac{1}{1-(1-\pi_{ij})^{m_{ij}}} \binom{m_{ij}}{y_{ij}} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{m_{ij}-y_{ij}} \right]^{1-I_{ij}}.$$

5.8 Zero Modified Random Mean Models

The induced correlation structure from random effects may be unrealistic in some cases. When there is time series correlation in the data, the random mean model would be a better option, which takes the possible serial dependence within subject-specific measurements into account.

5.8.1 Zero Modified Poisson Random Mean Models

Suppose the response vector $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_N')'$ contains data from N clusters, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, $i = 1, \ldots, N$. We assume that, conditional on the latent variable, u_{ij}, the response Y_{ij} is from zero modified Poisson distribution

$$\Pr(Y_{ij} = k) = \begin{cases} \frac{1-p_{ij}^M}{1-e^{-\lambda_{ij}}} \frac{\lambda_{ij}^k e^{-\lambda_{ij}}}{k!}, & k = 1, 2, 3, \ldots \\ p_{ij}^M, & k = 0, \end{cases}$$

where λ_{ij} and p_{ij} are modelled through the log-linear and logistic regression models

$$\log(\lambda_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_{ij}$$
$$\text{logit}(p_{ij}) = \mathbf{g}'_{ij}\boldsymbol{\gamma},$$

where $\mathbf{x}_{ij}$ and $\mathbf{g}_{ij}$ are vectors of known covariate values associated with the response, Y_{ij}, for $i = 1, \ldots, N,\ j = 1, \ldots, n_i$. To complete the model specification, some covariance structure is imposed on the latent variable $\mathbf{u}_i$.

5.8.2 Zero Modified Binomial Random Mean Models

Suppose the response vector $\mathbf{Y} = (\mathbf{Y}'_1, \ldots, \mathbf{Y}'_N)'$ contains data from N clusters, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, $i = 1, \ldots, N$. We assume that, conditional on the latent variable, u_{ij}, The response Y_{ij} is from zero modified binomial distribution

$$\Pr(Y_{ij} = k) = \begin{cases} \frac{1-p_{ij}^M}{1-(1-\pi_{ij})^{m_{ij}}} \binom{m_{ij}}{y_{ij}} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{m_{ij}-y_{ij}}, & k = 1, 2, \ldots, m_{ij}, \\ p_{ij}^M, & k = 0, \end{cases}$$

where λ_{ij} and p_{ij} are modelled through the logistic regression models

$$\text{logit}(\lambda_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_{ij}$$
$$\text{logit}(p_{ij}) = \mathbf{g}'_{ij}\boldsymbol{\gamma},$$

where $\mathbf{x}_{ij}$ and $\mathbf{g}_{ij}$ are vectors of known covariate values associated with the response, Y_{ij}, for $i = 1, \ldots, N,\ j = 1, \ldots, n_i$. To complete the model specification, some covariance structure is imposed on the latent variable $\mathbf{u}_i$.

5.9 Adaptive Gaussian Quadrature Approximations

In this section, we describe the adaptive Gaussian Quadrature approximation to evaluate the integral of the log-likelihood in the zero modified random effects models. The log-likelihood of the marginal distribution function of the zero modified random effects models can be written as

$$l(\boldsymbol{\psi};\mathbf{y}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i} I_{ij}\log(p_{ij}^{M}) + (1-I_{ij})\log(1-p_{ij}^{M}) + \sum_{i=1}^{N} \log \int f^{T}(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\beta})\; q_i(\mathbf{b}_i;\Sigma_i(\boldsymbol{\sigma^2}))\; \mathrm{d}\mathbf{b}_i, \tag{5.3}$$

where $f^{T}(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\beta})$ is the conditional probability function of zero truncated distribution given the random effect $\mathbf{b}_i$. We are interested in the approximation of the integral in (5.3)

$$\sum_{i=1}^{N} \log \int f^{T}(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\beta})\; q_i(\mathbf{b}_i;\Sigma_i(\boldsymbol{\sigma^2}))\; \mathrm{d}\mathbf{b}_i. \tag{5.4}$$

Let $f^{T}(\mathbf{y}_i,\mathbf{b}_i)$ denote the joint distribution of the positive responses and the random effect. The simple approximate formulas for the mean and variance of $f^{T}(\mathbf{y}_i,\mathbf{b}_i)$ can be derived using Laplace approximation (de Bruijn (1981) [19]). Firstly, find the first and the second derivatives of the log-likelihood function of $f^{T}(\mathbf{y}_i,\mathbf{b}_i)$ with respect to $\mathbf{b}_i$

$$\frac{\partial \log f^{T}(\mathbf{y}_i,\mathbf{b}_i)}{\partial \mathbf{b}_i} = \sum_{j=1}^{n_i} \mathbf{z}_{ij}(y_{ij}-\mu_{ij}^{T}) - \Sigma_i^{-1}\mathbf{b}_i \tag{5.5}$$

$$\frac{\partial^2 \log f^{T}(\mathbf{y}_i,\mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^{T}} = -\mathbf{Z}_i'\mathbf{W}_i\mathbf{Z}_i - \Sigma_i^{-1}, \tag{5.6}$$

where μ_{ij}^T is the mean of zero truncated distribution, and the diagonal matrix $\mathbf{W}_i$ has zero truncated variance $\text{Var}(Y_{ij}^T)$ as the diagonal element.

It follows from (5.6) that $\frac{\partial^2 \log f^T(\mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}$ is negative-definite and, as a result, $\log f^T(\mathbf{y}_i, \mathbf{b}_i)$ is a strictly concave function of $\mathbf{b}_i$. Therefore, there is a unique point of maximum $\hat{\mathbf{b}}_i$ corresponding to $\frac{\partial \log f^T(\mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i} = 0$. By taking a second-order Taylor expansion of $\log f^T(\mathbf{y}_i, \mathbf{b}_i)$ around $\hat{\mathbf{b}}_i$, the integrand is approximately $\mathcal{N}(\hat{\mathbf{b}_i}, -\frac{\partial^2 \log f^T(\mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}^{-1}|_{\mathbf{b}_i = \hat{\mathbf{b}}_i})$ up to a normalizing constant.

The critical step for the success of importance sampling is the choice of an importance distribution that approximates the integrand. For the integral (5.4), the integrand $f^T(\mathbf{y}_i, \mathbf{b}_i)$ is approximated by $\text{N}(\hat{\mathbf{b}_i}, -\frac{\partial^2 \log f^T(\mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}^{-1}|_{\mathbf{b}_i = \hat{\mathbf{b}}_i})$ density, after accounting for some constant coefficient. This is the importance distribution used in the adaptive Gaussian quadrature rule, and the grid of abscissas is centered around the conditional modes $\hat{\mathbf{b}_i}$ and $\sqrt{2}(-\frac{\partial^2 \log f^T(\mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}|_{\mathbf{b}_i = \hat{\mathbf{b}}_i})^{-\frac{1}{2}}$ is used for scaling.

Define $\mathbf{u}_k = (u_{k_1}, \ldots, u_{k_{n_i}})'$, where u_{k_l} and ω_{k_l}, $k_l = 1, \ldots, N_{GQ}$, are the abscissas and the weight for the one-dimensional Gaussian quadrature rule based on the $\mathcal{N}(0,1)$ kernel. Centering and scaling the abscissas $\mathbf{u}_k$ according to

$$\tilde{\mathbf{b}}_{ik} = \hat{\mathbf{b}}_i + \sqrt{2}\left(-\frac{\partial^2 \log f^T(\mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}\bigg|_{\mathbf{b}_i = \hat{\mathbf{b}}_i}\right)^{-\frac{1}{2}} \mathbf{u}_k.$$

The adaptive Gaussian quadrature approximation of the integral (5.4) is

$$l_{AGQ}(\boldsymbol{\psi}; \mathbf{y}) = \sum_{i=1}^{N} \frac{q}{2}\log 2 - \log\left(-\frac{\partial^2 \log f^T(\mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T}\bigg|_{\mathbf{b}_i = \hat{\mathbf{b}}_i}\right)^{\frac{1}{2}} + \log \sum_{k}^{N_{GQ}} f^T(\mathbf{y}_i, \tilde{\mathbf{b}}_{ik}) W_k,$$

where $W_k = \exp(\|\mathbf{u}_k\|^2) \prod_{l=1}^{q} \omega_{k_l}$. The optimization of the likelihood function

can also be implemented through proc nlmixed in SAS.

5.9.1 Empirical Estimates

Given ML estimates of $\boldsymbol{\psi}$, the random effect $\mathbf{b}_i$ can be predicted as follow

$$\hat{\mathbf{b}}_{i_{EE}} = \mathrm{E}(\mathbf{b}_i|\mathbf{y}_i;\hat{\boldsymbol{\psi}}).$$

That is, the predicted random effect for the ith subject is simply "estimated" as the conditional mean of $\mathbf{b}_i$ given $\mathbf{y}_i$. It is the by-product from the adaptive Gaussian quadrature approximation. The empirical estimate of the random effect is simply $\hat{\mathbf{b}}_i$, the solution to

$$\frac{\partial \log f^T(\mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i} = 0,$$

as in (5.6).

The adaptive Gaussian Quadrature approximation to the integral of the log-likelihood function of the zero modified random mean models is similar to that of the random effects models. Hence, we omit the formulation of the details of its approximation.

5.10 Simulation Study

We conducted simulation studies to investigate performance of the zero modified random effects models and the zero modified random mean models.

5.10.1 Simulation Study I

In the first simulation, we consider the following zero modified Poisson regression model with random intercept

$$\Pr(Y_{ij}=k)=\begin{cases}\frac{1-p_{ij}^M}{1-e^{-\lambda_{ij}}}\frac{\lambda_{ij}^k e^{-\lambda_{ij}}}{k!}, & k=1,2,3,\ldots\\ p_{ij}^M, & k=0,\end{cases} \tag{5.7}$$

where λ_{ij} and p_{ij}^M are modelled with the log-linear and logistic regression models

$$\log(\lambda_{ij})=\mathbf{x}_{ij}'\boldsymbol{\beta}+b_i, \tag{5.8}$$

$$\text{logit}(p_{ij}^M)=\mathbf{g}_{ij}'\boldsymbol{\gamma},\ i=1,\ldots,N, j=1,\ldots,n_i. \tag{5.9}$$

Each of the 100 simulation data sets was comprised of $N=150$ clusters with size $n_i=5$. The observations in each cluster are conditionally independent given the random intercept. The data were simulated according to the above model with fixed effects of the log-linear model $(\beta_0,\beta_1)=(1,-2.2)$, the regression coefficients of the logit model $(\gamma_0,\gamma_1)=(1.1,-8.5)$, the variance of the random intercept $\sigma^2=1$, and the covariates $\mathbf{x}_{ij}=(1,j/15)'$ and $\mathbf{g}_{ij}=(1,j/15)'$.

In table 5.6, we summarized the parameter estimates and standard errors from the 100 simulations. The parameter estimate is the average of the 100 estimates, and the estimated SE is the average of the 100 estimated standard errors. The empirical SE is standard deviation of the 100 parameter estimates. We also calculated the actual coverage of the 95% confidence intervals are $95\%, 94\%, 90\%, 94\%$, and 94% for $\beta_0,\beta_1,\sigma^2,\gamma_0$ and γ_1, respectively.

Model	Parameter(True)	Average	Estimated SE	Empirical SE
Log-linear	$\beta_0(1)$	1.00989	0.12298	0.11917
	β_1(-2.2)	−2.21878	0.34748	0.34600
Logit	$\gamma_0(1.1)$	1.08914	0.18494	0.18074
	γ_1(-8.5)	−8.48799	0.91613	0.91946
Variance				
Component	σ^2 (1)	0.98072	0.16473	0.18325

Table 5.6: Simulation results for 100 data sets from zero modified Poisson regression model with random intercept

Mean	Std Dev	Min	Max
0.09012	0.81877	−1.17984	2.25768

Table 5.7: Empirical estimates of the random intercept from a simulated data set

We calculate the empirical estimates of the random intercept in a simulated data set. Table 5.7 summarizes the mean, the standard deviation, the minimum and the maximum of the predicted random intercepts from the 150 clusters. We locate the minimum and the maximum empirical estimates occurred at cluster 65 and cluster 122. We observe that cluster 122 has observations 18, 22, 0, 13, 16, which has much higher level of response than others; and cluster 65 has observations 1, 1, 1, 0, 1.

5.10.2 Simulation Study II

In the second simulation, we consider the following the zero modified binomial regression model with random intercept

$$\Pr(Y_{ij}=k)=\begin{cases}\frac{1-p_{ij}^M}{1-(1-\pi_{ij})^{m_{ij}}}\binom{m_{ij}}{y_{ij}}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{m_{ij}-y_{ij}}, & k=1,2,\ldots,m_{ij},\\ p_{ij}^M, & k=0,\end{cases} \tag{5.10}$$

Model	Parameter(True)	Average	Estimated SE	Empirical SE
Logit(π)	β_0(1)	1.00122	0.13118	0.12224
	β_1(-2.2)	-2.21911	0.41088	0.42202
Logit(p^M)	γ_0(1.1)	1.11836	0.18537	0.18391
	γ_1(-8.5)	-8.62215	0.91898	0.91983
Variance				
Component	σ^2 (1)	0.98694	0.14937	0.15103

Table 5.8: Simulation results for 100 data sets from zero modified binomial regression model with random intercept

where λ_{ij} and p_{ij}^M are modelled with logistic regression models

$$\text{logit}(\pi_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta} + b_i,$$
$$\text{logit}(p_{ij}^M) = \mathbf{g}_{ij}'\boldsymbol{\gamma},\ i = 1, \ldots, N.$$

Each of the 100 simulation data sets was comprised of $N = 150$ clusters with size $n_i = 5$. The observations in each cluster are conditionally independent given the random intercept. The data were simulated according to the above model with the fixed effects of the logit model $(\beta_0, \beta_1) = (1, -2.2)$, the regression coefficients in the logit model $(\gamma_0, \gamma_1) = (1.1, -8.5)$, the variance of the random intercept $\sigma^2 = 1$, and the covariates $\mathbf{x}_{ij} = (1, j/15)'$ and $\mathbf{g}_{ij} = (1, j/15)'$.

In table 5.8, we summarized the parameter estimates and standard errors from the 100 simulations. The parameter estimate is the average of the 100 estimates, and the estimated SE is the average of the 100 estimated standard errors. The empirical SE is standard deviation of the 100 parameter estimates. We also calculated the actual coverage of the 95% confidence intervals are $94\%, 95\%, 93\%, 95\%$ and 96%, for $\beta_0, \beta_1, \sigma^2, \gamma_0$, and γ_1 separately.

We calculate the empirical estimates of the random intercept in a simu-

Mean	Std Dev	Min	Max
−0.00894	0.83084	−2.16234	1.71253

Table 5.9: Empirical estimates of the random intercept of the simulated data set

lated data set. Table 5.9 summarizes the mean, the standard deviation, the minimum and the maximum of the predicted random intercepts from the 150 clusters. We locate the minimum and the maximum empirical estimates occurred at cluster 42 and cluster 36. We observe that cluster 36 has observations 0, 0, 15, 20, 0 with corresponding m equal to 2, 11, 17, 20, 14, which has higher level of response than others; and cluster 42 has observations 0, 0, 1, 0, 1 with corresponding m equal to 13, 4, 12, 15, 14.

5.10.3 Simulation Study III

In the third simulation, we intend to compare the zero modified random effects model and the zero modified random mean model, especially when there is serial pattern in the clustered data. We consider the following zero modified Poisson random mean model

$$\Pr(Y_{ij}=k)=\begin{cases}\frac{1-p_{ij}^{M}}{1-e^{-\lambda_{ij}}}\frac{\lambda_{ij}^{k}e^{-\lambda_{ij}}}{k!}, & k=1,2,3,\ldots\\ p_{ij}^{M}, & k=0,\end{cases} \tag{5.11}$$

where λ_{ij} and p_{ij}^{M} are modelled with the log-linear and logistic regression models

$$\log(\pi_{ij})=\mathbf{x}_{ij}'\boldsymbol{\beta}+u_{ij},$$
$$\operatorname{logit}(p_{ij}^{M})=\mathbf{g}_{ij}'\boldsymbol{\gamma},\ i=1,\ldots,N.$$

Model	Parameter(True)	Average	Estimated SE	Empirical SE
Log-linear	$\beta_0(1)$	1.02756	0.17745	0.16971
	β_1(-2.2)	-2.31844	0.82266	0.83880
Logit	$\gamma_0(1.1)$	1.11744	0.21131	0.19658
	γ_1(-8.5)	-8.65723	1.20624	1.17635
Variance				
Component	σ^2 (1)	0.99166	0.15716	0.15745
	ρ (0.8)	0.79380	0.05883	0.06058

Table 5.10: Simulation results for 100 data sets from zero modified Poisson random mean model

Mean	Std Dev	Min	Max
0.10841	0.66448	-1.01132	3.17191

Table 5.11: Empirical estimates of the latent variables of a simulated data set

Each of the 100 simulation data sets was comprised of $N = 150$ clusters with size $n_i = 4$. The observations in each cluster are conditionally independent given the latent variable. The data were simulated according to the above model with the fixed effects of the log-linear model $(\beta_0, \beta_1) = (1, -2.2)$, the regression parameters in the logit model $(\gamma_0, \gamma_1) = (1.1, -8.5)$, the variance of the latent variable $\sigma^2 = 1$, the correlation coefficient $\rho = 0.8$, and the covariates $\mathbf{x}_{ij} = (1, j/15)'$ and $\mathbf{g}_{ij} = (1, j/15)'$.

In table 5.10, we summarized the parameter estimates and standard errors from the 100 times simulations. The parameter estimate is the average of the 100 estimates, and the estimated SE is the average of the 100 estimated standard errors. The empirical SE is standard deviation of the 100 parameter estimates. We also calculated the actual coverage of the 95% confidence intervals are $95\%, 90\%, 95\%, 91\%, 94\%$, and 94% for $\beta_0, \beta_1, \gamma_0, \gamma_1, \sigma^2$, and ρ separately.

We calculate the empirical estimates of the latent variable in a simulated data set. Table 5.11 summarizes the mean, the standard deviation, the mini-

Model	Parameter(True)	Average	Estimated SE	Empirical SE
Log-linear	$\beta_0(1)$	1.10740	0.13137	0.20221
	β_1(-2.2)	−2.27877	0.50381	1.01597
Logit	$\gamma_0(1.1)$	1.11744	0.21131	0.19658
	γ_1(-8.5)	−8.65727	1.20624	1.17635
Variance Component	σ^2	0.83711	0.14914	0.15138

Table 5.12: Simulation results for 100 data sets fitted by the zero modified Poisson random intercept model

mum and the maximum of the predicted latent variables of the 150 clusters. We locate the minimum and the maximum empirical estimates occurr at cluster 54 and cluster 105. We observe that cluster 54 has observations 4, 0, 30, 38, which has much higher level of responses than others; and cluster 105 has observations 1, 1, 1, 0.

For each of the simulated data sets, we also fit the zero modified Poisson random intercept model. In Table 5.12, the estimates of parameters in the log-linear model are unbiased. However, the estimated standard errors of estimates are much smaller than the corresponding empirical standard errors. It indicates that the zero modified random intercept model may underestimate the standard errors of the estimates in the log-linear model, and the zero modified random intercept model may be insufficient to model the data with serial correlation.

5.11 Measles Data

We revisit the measles data studied by Sherman and le Cessie (1997) [87]. The annual measles data were collected for each of 15 counties in the United States between 1985 and 1991. For each county, the annual number of preschoolers

cases	0	1	2	3-5	6-10	11-20	21-30	31-100	>100
freq	31	19	5	9	8	7	6	6	14

Table 5.13: Histogram of cases

with measles was recorded as well as factors possibly affecting measles incidence: immunization rate and density of preschoolers per county. Table 5.13 are preliminary summary of the annual number of preschoolers with measles.

We study the relationship between the annual number of preschoolers with measles and the immunization rate in two year old children. The immunization rate is assumed to be constant during the period 1985-1991. There seems a strong negative relationship between the immunization rate and the incidence of measles, but there is also wide variability in measles incidence from year to year. For example, in county 6, with the lowest immunization rate, there are four years with a large number of preschool measles, but three years with a negligible number of cases.

Let Y_{ij} be the number of cases in county i and in year j, $i = 1, \ldots, 15, j = 1, \ldots, 7$. Let n_{ij}, be the total number of preschool children in county i and year j. We model the measles data through zero modified Poisson regression model with random county effects b_i

$$\Pr(Y_{ij} = k) = \begin{cases} \frac{1-p_{ij}^M}{1-e^{-\lambda_{ij}}} \frac{\lambda_{ij}^k e^{-\lambda_{ij}}}{k!}, & k = 1, 2, 3, \ldots \\ p_{ij}^M, & k = 0, \end{cases} \tag{5.12}$$

where λ_{ij} and p_{ij}^M are modelled with the log-linear and logistic regression mod-

Model	Parameter	Estimate	Std Err	P-value
Log-linear	β_0	1.1008	1.7180	0.5320
	β_1	−0.1331	0.0248	<.0001
Logit	γ_0	−2.6010	1.7412	0.1574
	γ_1	0.0249	0.0247	0.3309
Variance component	σ^2	0.6362	0.2507	0.0237

Table 5.14: Parameter estimates from the zero modified Poisson random intercept model for Measles data

Mean	Std Dev	Min	Max
0.0107	0.7764	−1.4326	1.0682

Table 5.15: Summaries of empirical estimates of the random intercept of Measles data

els

$$\log(\lambda_{ij}) = \beta_0 + \beta_1 \text{rate}_{ij} + b_i + \log(n_{ij}), \tag{5.13}$$

$$\text{logit}(p_{ij}^M) = \gamma_0 + \gamma_1 \text{rate}_{ij},\ i = 1, \ldots, 15, j = 1, \ldots, 7. \tag{5.14}$$

Table 5.14 summarizes the parameter estimates and standard errors by fitting the zero modified Poisson random intercept model. Table 5.15 summarizes the mean, the standard deviation, the minimum and the maximum of the predicted random intercepts of the 15 counties. We locate the minimum and the maximum empirical estimates occurred at county 12 and county 11. The county 11 has observations 6, 43, 1, 0, 0, 136, 0; while county 12 has observations 1, 6, 0, 1, 7, 58, 1.

5.12 Conclusion

In this chapter, we propose regression models for count data with excess zeros in cross-sectional and longitudinal studies. Aided with zero truncated distri-

bution, and zero modified distribution, we establish a class of distributions to model data with excess zeros.

Firstly, we formulate the generalized linear models for truncated data, and mainly discuss zero truncated Poisson and zero truncated binomial regression models. We then extend the zero truncated regression models to data with excess zeros, by introducing an indicator variable to identify the responses as zero or nonzero. It includes a regression model for the probability of zero response, and a conditional model for nonzero data.

Secondly, we propose zero modified random effects models for clustered data with excess zeros. In contrast to the fitting of ZIP and ZIB mixed effects models through the use of EM algorithm as in Hall (2000), the fitting of the zero modified random effects models is simpler, in that we need only the numerical approximation method to evaluate the integral in the marginal likelihood function. In particular, the adaptive Gaussian quadrature method is applied in this chapter. The optimization of likelihood function can also be implemented through proc nlmixed in SAS. Simulations for two zero modified random intercept models are conducted, and Measles data is used to illustrate the new method.

Lastly, we apply the random mean models introduced in Chapter 3 to clustered data with excess zeros, and formulate the corresponding zero modified random mean models. A simulation is conducted to evaluate the zero modified random mean model for the temporal count data. For each of the simulated data sets, we also fit the zero modified Poisson random intercept model. We find the zero modified random intercept model may underestimate the standard errors of the estimates in the log-linear model, and may be insufficient to model the data with serial correlation.

The limit of the implementation of zero modified random mean models is the computational issue. To fit zero modified random effects models and zero modified random mean models, we use adaptive Gaussian quadrature method, which works well when the dimension of the integration is small, say 5. But for higher dimension of the integration, we need to seek other approximation methods.

Chapter 6

Log-Gamma Linear Mixed Models for Multiple Characteristics

6.1 Introduction

The multiple characteristics model is used when responses on two or more characteristics are observed over time for each individual in longitudinal studies. Such data are commonly collected in the health sciences and epidemiological studies. The methodology described in this chapter is motivated by a glaucoma study. The purpose of this study was to investigate the longitudinal structure and function association in glaucoma and the evolution of this association over time, including an assessment of rates of change. Glaucoma is an optic neuropathy characterized by progressive neuroretinal rim thinning, excavation, and loss of the retinal nerve fiber layer. These structural changes are usually accompanied by functional losses. Although there is an unques-

tionable relationship between structural and functional damage in glaucoma, the precise association and the evolution of this association over time are still unclear. The elucidation of the longitudinal relationship between structural and functional tests and their rates of change over time is essential in order to enhance our understanding of the glaucomatous process and to determine the relative utility of these tests in monitoring different stages of the disease. The evaluation of rates of visual field change during follow-up was performed using the visual field index (VFI). The VFI represents the percentage of normal age-corrected visual function, and it is intended for use in calculating the rates of progression and the staging of glaucomatous functional damage. The VFI can range from 100% (normal visual field) to 0% (perimetrically blind field). Retinal nerve fiber layer (RNFL) retardation measurements were obtained on a 3.2-mm-diameter calculation circle around the optic nerve head. The global average RNFL thickness (calculated as the average of the RNFL measurements obtained on the 360° around the optic nerve) was used in this study.

There has been a great deal of literature dealing with multivariate longitudinal data. Reinsel (1984) [77] considered the random effects model for multiple characteristics with a complete and balanced design. Shah, Laird, and Schoenfeld (1997) extended linear mixed models to allow for multiple longitudinal outcomes in the case where the number and timing of observations may differ from individual to individual. Roy and Lin (2000) [82] extended latent variable models to multivariate longitudinal data. There are two sources of within-subject correlation that should be reflected in the multiple-characteristics model: i) among different characteristics; ii) among repeated measures of the same outcome over time. Although one can perform separate

analyses of the two outcomes, this does not address the main question of interest: how overall treatment practices have changed over time. Moreover, the analysis will have increased power if information from all of the outcomes is used (Pocock, Geller, and Tsiatis (1987) [75]).

The normality assumption for random effects in the linear mixed model may be unrealistic, raising concerns about the validity of inferences on fixed effects and random effects if it is violated. For single-characteristic longitudinal data, it has been shown (Verbeke and Lesaffre (1996), (1997) [97] [98]) that deviations from the normality assumption have little impact on the estimation of the fixed effects and variance components, and much more on the empirical Bayes estimates for random effects in linear mixed models. Allowing the random effects distribution to have more complex features than a symmetric normal density may provide insights into the underlying heterogeneity. Verbeke and Lesaffre (1996) and Magder and Zeger (1996) [62] have extended the linear mixed model with a mixture of normals as the random effects distribution. Pinheiro et al. (2001) [74] considered the multivariate t distribution for the modelling of both the random effects and the within-subject errors, and they demonstrated its robustness against outliers through examples and simulations. Zhang et al. (2008) [111] considered the log-gamma distribution for the modelling of the skewed random effects.

From the preliminary study of the glaucoma data set, we found the distribution of responses from one characteristic is skewed. We propose to extend the mixed effects model to accommodate skewed responses and associations among multiple characteristics by the use of the log-gamma distributed random variable. For the multiple characteristics model, there are only a few results related to the consequences of misspecifying the random-effects dis-

tribution. Ghosh et al. (2007) [35] developed a Bayesian approach to the bivariate mixed effects model through the use of a multivariate skew-normal distribution. However, there is no closed form of the marginal distribution of the responses when the random effects are assumed to be nonnormal. In this chapter, we present a non-Gaussian linear mixed effects model for multiple outcomes, which is feasible mathematically and has less computational complexity.

We express the reordered random effects through the product of the linear transformation matrix with unknown entries and the random vector with independent components, and we specify that one component in the random vector follows skewed distribution with the others coming from the multivariate normal distribution. The introduction of the linear transformation matrix not only accounts for correlated random effects, but also ensures that the dimension of the numerical integration of the marginal likelihood is one, since the convolution of two normal distributions is still normal. Various techniques can be chosen to compute the integral approximation of the marginal log-likelihood: quadrature, Laplace, Monte Carlo, and Markov chain Monte Carlo methods. We consider density functions in the family of the log-gamma distributions with mean zero for modelling of skewed random effects. We allow the number and time of repeated measures to differ for different characteristics and units. We propose a lack-of-fit test for comparing the log-gamma model and the Gaussian model, based on the profile likelihood function of the shape parameter.

6.2 Model and Assumptions

Let $\mathbf{Y}_i = (\mathbf{Y}'_{i1}, \ldots, \mathbf{Y}'_{ip})'$ be the vector of stacked responses for the i^{th} subject with p characteristics, where $\mathbf{Y}_{ik} = (Y_{i1k}, \ldots, Y_{in_{i_k}k})'$ is the collection of n_{i_k} observations for the k^{th} characteristic. We assume a linear mixed effects model to be of the form

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, N, \tag{6.1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_p)'$ is a vector of fixed effects, $\mathbf{b}_i = (\mathbf{b}'_{i1}, \ldots, \mathbf{b}'_{ip})'$ is a vector of random effects, $\mathbf{X}_i = \text{diag}(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})$ is the design matrix of fixed effects, $\mathbf{Z}_i = \text{diag}(\mathbf{Z}_{i1}, \ldots, \mathbf{Z}_{ip})$ is the design matrix of random effects, and $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}'_{i1}, \ldots, \boldsymbol{\varepsilon}'_{ip})'$ is the vector of error terms. The vectors of responses, $\mathbf{Y}'_i s$, for the N subjects are assumed to be independent of one another. In addition, we assume that the $\mathbf{Y}_i$ are conditionally independent, i.e., given the random effects $\mathbf{b}_i$, the components in $\mathbf{Y}_i$ are independent of one another. The usual distributional assumption for the random effects is multivariate normal, but the maximum likelihood would work for other distributions as well.

We model some components of $\mathbf{b}_i$ through the log-gamma distribution to account for the skewed random effects. Suppose $\mathbf{b}_i$ is a $q \times 1$ vector with components $b_{i1}, b_{i2}, \ldots, b_{iq}$. To develop the model, we rearrange the components in $\mathbf{b}_i$ and divide them into two parts. Denote the reordered vector $\mathbf{b}^*_i = (b^*_{i1}, \ldots, b^*_{i,q_0} \mid b^*_{i,q_0+1}, \ldots, b^*_{i,q})'$. The distributions of the first q_0 components are skewed, and those of the remaining $q - q_0$ components are symmetric. The reordering also results in the interchange of the columns in the design matrix $\mathbf{Z}_i$ according to the subscripts of $\mathbf{b}^*_i$. To allow correlated random intercepts

and random slopes, we define $\mathbf{b}_i^* = \mathbf{R}\mathbf{s}_i^*$, where the linear transformation matrix $\mathbf{R}$ is an upper triangular matrix with unknown entries r_{ij}, $1 \leq i < j \leq q$, and $1's$ on the diagonal, and the components of the vector $\mathbf{s}_i^*$ are independent of one another. Expressing the transformation in matrix notation gives

$$\begin{pmatrix} b_{i1}^* \\ b_{i2}^* \\ \vdots \\ b_{iq}^* \end{pmatrix} = \begin{pmatrix} 1 & r_{12} \cdots r_{1q} \\ 0 & 1 \cdots r_{2q} \\ \cdots\cdots\cdots\cdots \\ 0 & 0 \cdots 1 \end{pmatrix} \begin{pmatrix} s_{i1}^* \\ s_{i2}^* \\ \vdots \\ s_{iq}^* \end{pmatrix}. \tag{6.2}$$

For ease of computation, we specify that just one component in $\mathbf{s}_i^*$, say s_{i,q_0}^*, is from the log-gamma distribution. The remaining components of $\mathbf{s}_i^*$ are assumed to be multivariate normally distributed with mean $\mathbf{0}$ and a diagonal covariance matrix $\mathbf{G}$. If the distribution of the random effect b_{i,q_0}^* is positively skewed, then a minus log-gamma distribution should be selected for s_{i,q_0}^* because the log-gamma distribution is negatively skewed. For other skewed components b_{ij}^*, $1 \leq j < q_0$, the sign of the corresponding coefficient r_{j,q_0} in front of s_{i,q_0}^*, determines the direction of skewness. The advantage of introducing the linear transformation matrix $\mathbf{R}$ rather than specifying correlated random effects directly is that it produces a comparatively simple model and also makes it easy to implement the log-gamma mixed effects model.

The log-gamma distribution represents a flexible class of symmetric, negatively skewed, positively skewed, and very skewed distributions. The probability density function of the standard log-gamma random variable W is

$$f_W(w) = \frac{1}{\Gamma(\kappa)} e^{-e^w + \kappa w}, \; w \in \mathbb{R}, \tag{6.3}$$

where $\kappa > 0$ is the shape parameter. For the location parameter $\mu \in \mathbb{R}$ and the scale parameter $\eta > 0$, the log-gamma location-scale family of probability density functions $(1/\eta) f_W((w-\mu)/\eta)$ has mean $\mu + \eta\psi(\kappa)$ and variance $\eta^2\psi'(\kappa)$, where ψ and ψ' are the digamma and trigamma functions. Due to the identifiability consideration, the mean of the log-gamma random variable s^*_{i,q_0} is set to zero by choosing the location parameter as $\mu = -\eta\psi(\kappa)$. The density functions with mean zero in the location-scale family $(1/\eta) f_W((w-\mu)/\eta)$ with standard pdf $f_W(w)$ in (6.3) are chosen to model s^*_{i,q_0}.

6.3 Inference

Since the joint distributions of both the responses and the random effects are fully specified, we base the estimation and inference on the likelihood function. We first discuss the maximum likelihood estimation of the fixed effects $\boldsymbol{\beta}$, the covariance components in $\mathbf{G} = \mathrm{Cov}(s^*_{i,1}, \ldots, s^*_{i,q_0-1}, s^*_{i,q_0+1}, \ldots, s^*_{i,q})$, the elements in the transformation matrix $\mathbf{R}$, the log-gamma shape parameter κ, the log-gamma scale parameter η, and the conditional covariance parameters $\boldsymbol{\Sigma}_i = \mathrm{Cov}(\boldsymbol{\varepsilon}_i) = \mathrm{diag}(\sigma_1^2 I_{n_{i_1}}, \sigma_2^2 I_{n_{i_2}}, \ldots, \sigma_p^2 I_{n_{i_p}})$. We denote all the above parameters $\boldsymbol{\theta}$. The inferences for the fixed effects, the components in the covariance matrix, and the parameters in the log-gamma distribution are based on the marginal likelihood function, which is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int \Bigg\{ \int \cdots \int \phi(\mathbf{y}_i | \mathbf{s}_i^*) \cdot \phi(s^*_{i,1}, \ldots, s^*_{i,q_0-1}, s^*_{i,q_0+1}, \ldots, s^*_{i,q}) \, ds^*_{i,1} \cdots ds^*_{i,q_0-1} ds^*_{i,q_0+1} \cdots ds^*_{i,q} \Bigg\} \cdot f(s^*_{i,q_0}) ds^*_{i,q_0}, \tag{6.4}$$

where $f(s^*_{i,q_0})$ represents the density function from the location-scale family of the log-gamma distributions with standard pdf in (6.3). However, there are no simple, closed-form solution to the integral. Instead, numerical integration techniques are required to maximize the likelihood function. Under suitable conditions (discussed later), the MLEs of $\hat{\boldsymbol{\theta}}$ are consistent and asymptotically normal with the asymptotic covariance matrix equal to the inverse of the Fisher information matrix.

Newton Raphson algorithm is often the choice for finding the MLE. In many cases, deriving the Hessian matrix or the Fisher information matrix is analytically intricate, and therefore alternative numerical strategies are desirable. We invoke a Gauss-Newton algorithm (Ruppert (2005) [83]), in which the Fisher information matrix is approximated by $\mathbf{B}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \cdot l_i(\mathbf{y}_i;\boldsymbol{\theta}) \cdot l_i(\mathbf{y}_i;\boldsymbol{\theta})'$. Thus, only the first-order derivatives of the log likelihood are involved. The key step of this algorithm is to halve the value of δ, which guarantees a steady increase in the likelihood from the previous iteration. To be more precise, the $(k+1)$th iteration should proceed as

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \delta\{\mathbf{B}(\boldsymbol{\theta}^k)\}^{-1} \cdot l(\boldsymbol{\theta}^k), \tag{6.5}$$

where δ is the step-halving term. The step-halving starts with $\delta = 1$ and halves δ until $l(\boldsymbol{\theta}^{k+1}) > l(\boldsymbol{\theta}^k)$. The algorithm stops when an increase in the likelihood is no longer possible, or the difference between two consecutive updates is smaller than a prespecified precision level. The starting values can be found either using an inefficient, but easily calculated, estimator, or by maximizing the likelihood on some grid. Thiébaut et al. [93] (2002) discussed how to fit the bivariate linear mixed models using SAS proc MIXED. The results from

the SAS output can be used to set starting values for our new algorithm, but there are many other choices to consider.

The estimates of random effects reflect between-subject variability, which makes them helpful for detecting outlying individuals or individuals evolving differently in time. The posterior distribution of $\mathbf{b}_i^*$ is given by

$$f(\mathbf{b}_i^*|\mathbf{y}_i) = \frac{\phi(\mathbf{y}_i|\mathbf{b}_i^*)f(\mathbf{b}_i^*)}{\int \phi(\mathbf{y}_i|\mathbf{b}_i^*)f(\mathbf{b}_i^*)d\mathbf{b}_i^*}. \tag{6.6}$$

The posterior mean of (6.6) is used to predict the random effects $\mathbf{b}_i^*$, with the unknown parameters replaced with their maximum likelihood estimates. We use the Markov chain Monte Carlo to simulate direct draws from the posterior distribution in (6.6) to obtain the posterior means.

6.4 Lack-of-fit Test

For the lack-of-fit test, we use the property that the limiting distribution of the log-gamma distribution is standard normal (Bartlett and Kendall (1946) [4]; Prentice (1974) [76]). The mean and variance of the standard log-gamma distribution W in (6.3) are respectively $E(W)=\psi(\kappa)$ and Var(W)=$\psi'(\kappa)$, where ψ and ψ' are the digamma and trigamma functions, and they behave like logκ and κ^{-1} for large κ. The transformed log-gamma variate $Z = \kappa^{1/2}(W - \log\kappa)$ has density function

$$f_0(z) = \frac{\kappa^{\kappa-1/2}}{\Gamma(\kappa)} e^{-\kappa e^{\kappa^{-1/2}z} + \kappa^{1/2}z},\ z \in \mathbb{R}. \tag{6.7}$$

The generalized log-gamma model is then the three-parameter family of distributions for which $Z = (S - u)/b$ has pdf (6.7). It can be shown that as

$\kappa \to \infty$, Z converges to the standard normal distribution.

Consider the generalized log-gamma version of the marginal likelihood function for the scaled response $\mathbf{y}^*_{\kappa_i} = \kappa^{1/2}\mathbf{y}_i$,

$$f(\mathbf{y}^*_{\kappa_i}, \boldsymbol{\theta}) = \int \Big\{ \int \cdots \int \phi(\mathbf{y}^*_{\kappa_i} | \mathbf{s}^*_i) \cdot \phi(\kappa^{1/2} s^*_{i,1}, \ldots, \kappa^{1/2} s^*_{i,q_0-1}, \kappa^{1/2} s^*_{i,q_0+1}, \ldots, \kappa^{1/2} s^*_{i,q}) \\ ds^*_{i,1} \cdots ds^*_{i,q_0-1} ds^*_{i,q_0+1} \cdots ds^*_{i,q} \Big\} \cdot \frac{1}{b} f_0\left(\frac{s^*_{i,q_0} - u}{b}\right) ds^*_{i,q_0}, \tag{6.8}$$

where $u = -b\kappa^{1/2}(\psi(\kappa) - \log\kappa)$. Then, the generalized log-gamma version of the marginal likelihood function of the response $\mathbf{y} = \frac{1}{\kappa^{1/2}} \mathbf{y}^*_\kappa$ is

$$L_0(\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\kappa^{n_i/2}} f\left(\frac{1}{\kappa^{1/2}} \mathbf{y}^*_{\kappa_i}, \boldsymbol{\theta}\right). \tag{6.9}$$

Now, the parameter of interest in $\boldsymbol{\theta}$ is κ, and the remaining parameters in $\boldsymbol{\theta}$ are treated as nuisance parameters and we denote them $\boldsymbol{\theta}_1$. It is straightforward to maximize (6.9) with fixed κ to obtain $\tilde{\boldsymbol{\theta}}_1(\kappa)$. The profile log-likelihood function is $l_p(\kappa) = \log L_0(\tilde{\boldsymbol{\theta}}_1(\kappa), \kappa)$ for κ. Tests of the hypothesis $H_0 : \kappa = \kappa_0$ vs. $H_a : \kappa \neq \kappa_0$ can be based on the likelihood ratio statistic $\Lambda(\kappa_0)$

$$\Lambda(\kappa_0) = 2l_p(\hat{\kappa}) - 2l_p(\kappa_0), \tag{6.10}$$

where $\hat{\kappa}$ is the MLE that maximizes $l_p(\kappa)$. For finite κ_0 the distribution of $\Lambda(\kappa_0)$ is asymptotically χ^2_1 under H_0. A slight technical difficulty arises in testing the normal model, since $\kappa_0 = \infty$ is on the boundary of the parameter space. However, the nuisance parameters $\boldsymbol{\theta}_1$ with true values are not on the boundary, if the variance components in $\boldsymbol{\theta}_1$ are transformed onto the log scale. When $\kappa_0 = \infty$, the asymptotic distribution of $\Lambda(\kappa_0)$ is a $50 : 50$ mixture of χ^2_0

and χ^2_1; then for $a \geq 0$, $Pr(\Lambda(\infty) \leq a) = 0.5 + 0.5Pr(\chi^2_1 \leq a)$ (Self and Liang 1987).

6.5 Asymptotic Properties

The asymptotic properties of the maximum likelihood estimates in the linear mixed model were discussed in Weiss (1973) and Jiang (2001). In the Appendix, we state the conditions for the asymptotic consistency and normality of the MLEs of the marginal likelihood function of the mixed effects model, and then show that the marginal likelihood function derived from the Log-Gamma mixed effects model satisfies those conditions.

Let s denote the dimension of $\boldsymbol{\theta}$. The full log-likelihood function is given by

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{N} l_i(\boldsymbol{\theta}, \mathbf{y}_i). \tag{6.11}$$

Theorem 6.1: *Let the observations be* $\mathbf{y}_1, \ldots, \mathbf{y}_N$*, each with the marginal distribution function* $f_i(\mathbf{y}_i|\boldsymbol{\theta})$ *in (6.4). Under the conditions* $(A0) - (A6)$*, with probability tending to* 1 *as* $N \to \infty$*, there exist solutions* $\hat{\boldsymbol{\theta}}_N = (\hat{\theta}_{1N}, \ldots, \hat{\theta}_{sN})$ *of the likelihood equations*

$$\frac{\partial}{\partial \theta_j} l(\boldsymbol{\theta}) = 0, \quad j = 1, \ldots, s, \tag{6.12}$$

such that

(a) $\hat{\boldsymbol{\theta}}_N$ *is a consistent estimator for* $\boldsymbol{\theta}_0$*;*

(b) $\mathbf{B}_N^{1/2}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0) \to MVN(\mathbf{0}, \mathbf{I})$*, where* $\boldsymbol{\theta}_0$ *is the true parameter point, and*

$$\mathbf{B}_N = \sum_{i=1}^{N} E_{\boldsymbol{\theta}_0}\left\{ \left(\frac{\partial}{\partial \boldsymbol{\theta}} l_i\right) \left(\frac{\partial}{\partial \boldsymbol{\theta}} l_i\right)^T \right\}. \tag{6.13}$$

Remark: The marginal distributions of the $\mathbf{Y}_i's$ are not identical. The standard conditions for the asymptotic properties of the maximum likelihood estimation cannot apply directly in the mixed models. (A6) is the condition of the central limit theorem for the sum of independent but not identically distributed random variables. The marginal distribution derived from the new class of mixed models has nice properties and satisfies all the conditions required for asymptotic consistency and normality. It is then convenient to conduct hypothesis tests or to find confidence intervals for the parameters in the model.

6.6 Data Analysis

A total of 1939 repeated measures for VFI and RNFL from 203 patients were used for the analysis. The measures of the two characteristics were collected from each of the $N = 203$ patients at different times. The VFI scores are proportional data, so that a logit transformation is applied in the analysis. Measures of RNFL are regular continuous values. It is believed that the disease progression is a linear function of time, and the slope depends on the individual patient. For the i^{th} patient, the baseline measures of the k^{th} $(k = 1, 2)$ characteristic recorded can be expressed as $\beta_{k0} + b_{ik0}$, where β_{k0} is an unknown parameter and b_{ik0} is the random intercept. Several measures are collected at time t_{ijk}, for $j = 1, \ldots, n_{i_k}$. The coefficient β_{k1} is an unknown parameter and

b_{ik1} is the random slope for the ith patient, and the subject-specific progression rate can be expressed as $\beta_{k1} + b_{ik1}t_{ijk}$. Furthermore, a random error is associated with each time measurement. Using the vector and matrix notation, the linear mixed effects model can be written as

$$\begin{pmatrix} \mathbf{Y}_{i1} \\ \mathbf{Y}_{i2} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{i2} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_{i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{i2} \end{pmatrix} \begin{pmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_{i1} \\ \boldsymbol{\varepsilon}_{i2} \end{pmatrix} \tag{6.14}$$

where

$$\mathbf{X}_{ik} = \mathbf{Z}_{ik} = \begin{pmatrix} 1 & t_{i1k} \\ 1 & t_{i2k} \\ \vdots & \vdots \\ 1 & t_{in_{i_k}k} \end{pmatrix}, k = 1, 2. \tag{6.15}$$

Based on findings from the preliminary analysis, the intercepts of VFI scores, the first characteristic, show a negatively skewed pattern when only the linear mixed model for the $\mathbf{Y}_{i1}'s$ is fitted. Histograms of regression coefficients for the within-subject regressions of two characteristics on time are shown in Figure 6.1. The scatterplots of intercepts and slopes obtained from the single subject models between the two characteristics (VFI on the x-axis and RNFL on the y-axis) are displayed in Figure 6.2. They not only verify that the distribution of subject-specific intercepts for VFI is negatively skewed but also clearly indicate the relationship between the two random intercepts and that of the two slopes.

The random intercept of the first characteristic b_{i10} in $\mathbf{b}_i$ is the only skewed element. It is in the first position, so there is no need to reorder $\mathbf{b}_i$ to obtain

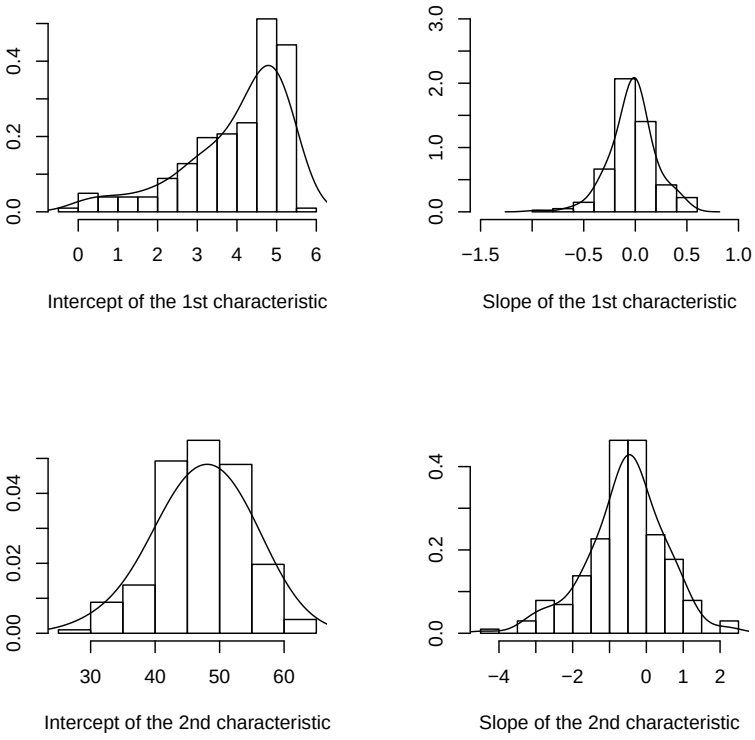

Figure 6.1: Coefficients for the within-subject regressions of two characteristics on time

Parameter	Value	Std Err	P-value
β_{10}	3.9194 (3.9671)	0.0623 (0.0919)	<0.0001
β_{11}	−0.0265 (−0.0290)	0.0128 (0.0148)	0.0378
β_{20}	47.5025 (47.4261)	0.2167 (0.4630)	<0.0001
β_{21}	−0.5645 (−0.5746)	0.0681 (0.0713)	<0.0001
$\sigma^2_{b_{i10}}$	0.9117 (1.5998)	0.0609 (0.1705)	<0.0001
$\sigma^2_{b_{i11}}$	0.0026 (0.0104)	0.0006 (0.0040)	<0.0001
$\sigma^2_{b_{i20}}$	35.4466 (41.6805)	0.2269 (4.3238)	<0.0001
$\sigma^2_{b_{i21}}$	0.3243 (0.3948)	0.0807 (0.1034)	<0.0001
σ^2_1	0.2201 (0.2396)	0.0083 (0.0119)	<0.0001
σ^2_2	2.1456 (2.4707)	0.1229 (0.1778)	<0.0001

Table 6.1: Estimates of the fixed effects and the variance components of the log-gamma and normal random effects models (The numbers in parentheses are estimates from the normal model)

the desired form. Then, we define the linear transformation $\mathbf{b}_i = \mathbf{R}\mathbf{s}_i$ with

$$\mathbf{R} = \begin{pmatrix} 1 & r_1 & r_2 & r_3 \\ 0 & 1 & r_4 & r_5 \\ 0 & 0 & 1 & r_6 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{6.16}$$

where $\mathbf{s}_i = (s_{i10}, s_{i11}, s_{i20}, s_{i21})'$, with independent components. We model s_{i10} through the log-gamma distribution, and the remaining components in $\mathbf{s}_i$ are multivariate normal. Since the distribution of the log-gamma is negatively skewed, it results in b_{i10} being negatively skewed also. Conditioning on the random effects, the responses for a particular individual are assumed to be independent and conditionally normal:

$$\mathbf{Y}_i|\mathbf{s}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{R}\mathbf{s}_i, \boldsymbol{\Sigma}_i),\ i = 1, \ldots, N. \tag{6.17}$$

We use MATLAB for the computation. Table 6.1 provides the estimates

of the parameters in the log-gamma and normal random effects models. The estimates of the parameters in the log-gamma model are close to those in the normal model. In the single-characteristic case, deviations from the normality assumption have little impact on the estimation of the fixed effects and the variance components. Based on the parameter estimates in Table 6.1, the same observation is suggested for the multiple case. The multivariate delta method and the asymptotic properties of the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ are applied to calculate the standard errors of the variance components of $\mathbf{b}_i$. The introduction of the linear transformation matrix $\mathbf{R}$ does not lead to any difficulties in the computation of the standard errors of the variance component estimates. The computational complexity is similar to that for the setting of the random effects in the normal case; they are usually specified with a nondiagonal covariance matrix. The estimated average regression coefficients for VFI and RNFL are -0.0265 and -0.5645, respectively. For both characteristics, larger values of the regression coefficients of time indicate slower deterioration. Therefore, negative slopes indicate disease progression over time.

The profile likelihood values $l_p(\kappa)$ are based on maximizing

$$l_p(\kappa) = \log L_0(\tilde{\boldsymbol{\theta}}_1(\kappa), \kappa),$$

with κ fixed. The profile log-likelihood is maximized at $\hat{\kappa} = 0.41$. The likelihood-ratio statistic $\Lambda(\infty)$ is 2402.4. The p-value is extremely small: less than 0.0001. There is strong evidence to reject the null hypothesis of the normal random effects model, suggesting that the log-gamma model is a better fit to the data.

Table 6.2 lists the means and standard errors of the $N = 203$ predicted

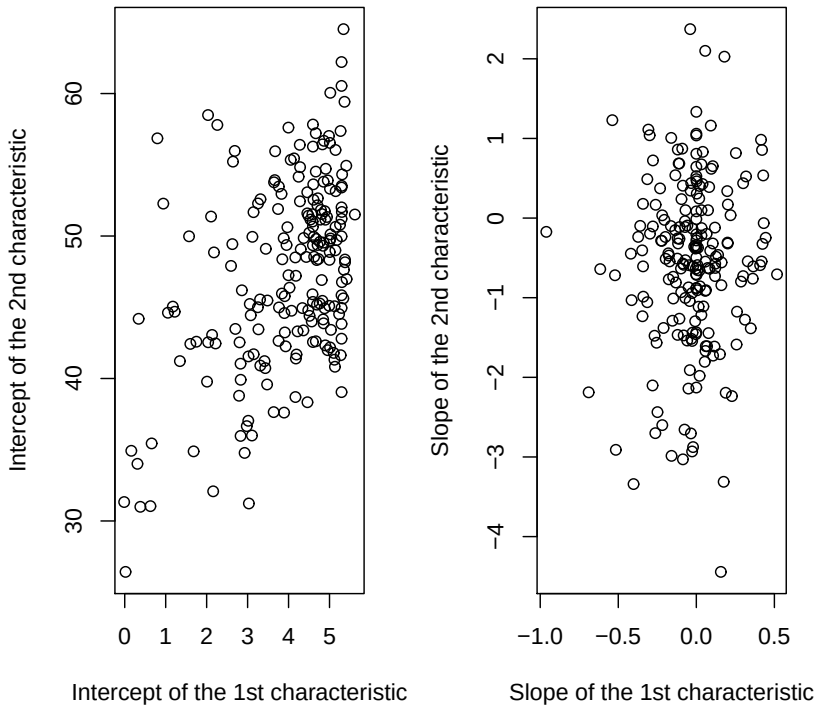

Figure 6.2: Scatterplots of intercepts and slopes for the within-subject regressions of two characteristics on time

	Log-gamma		Normal	
Random effect	Mean	SE	Mean	SE
b_{i10}	0.0313	1.2022	-0.0000	1.2444
b_{i11}	0.0039	0.0434	0.0000	0.0538
b_{i20}	-0.0951	6.3831	-0.0002	6.3864
b_{i21}	-0.0062	0.3875	0.0001	0.3978

Table 6.2: Results for the posterior distribution of random effects under the log-gamma and normal models

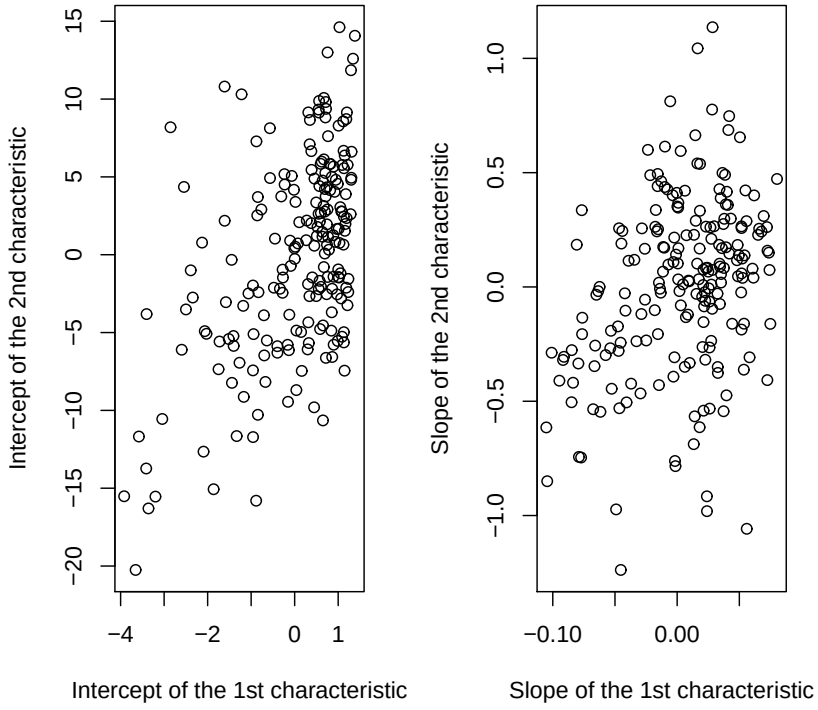

Figure 6.3: Scatterplots of empirical estimates of the intercepts and slopes from the posterior distributions of the random effects of the log-gamma model

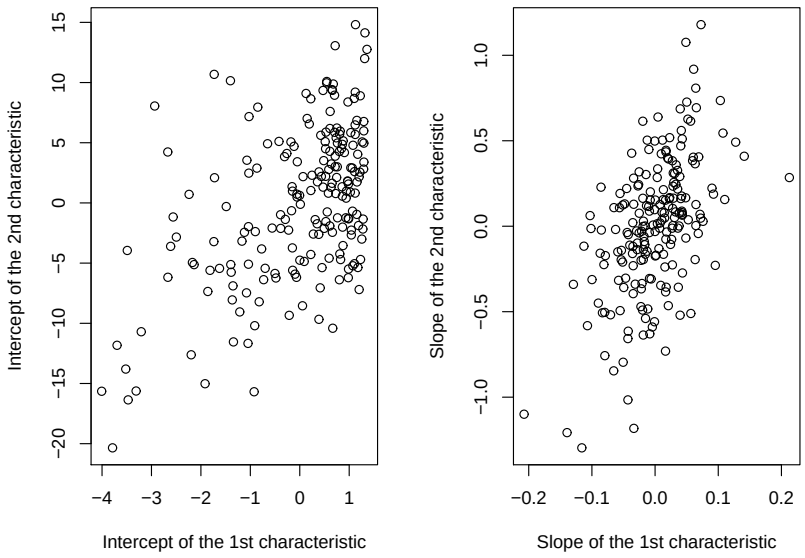

Figure 6.4: Scatterplots of empirical estimates of the intercepts and slopes from the posterior distributions of the random effects of the normal model

random effects under the log-gamma and normal models, the means of the posterior distributions of $\mathbf{b}_i$ given $\mathbf{Y}_i$. In the log-gamma model, given the ML estimates of fixed effects, variance components, and $\mathbf{R}$, the random effects $\mathbf{b}_i's$ were predicted from 18,000 sample points of 3 chains generated from the posterior distribution $f(\mathbf{b}_i|\mathbf{y}_i)$, with the first 4000 sample points discarded for burn-in. In the normal model, the random effects $\mathbf{b}_i's$ were predicted from Empirical Bayes estimates. Figure 6.3 displays distribution of the estimate of the random effect $\mathbf{b}_i$ from the $N = 203$ posterior means under the log-gamma model. We found that the distribution of the predicted random effects $b_{i10}'s$ is negatively skewed, which confirms the observed skewness from the preliminary study. The estimate of the skewness parameter $\kappa = 3.4212$ is moderately high. Figure 6.4 shows the scatterplot of the empirical estimates of the intercepts and slopes from the normal model.

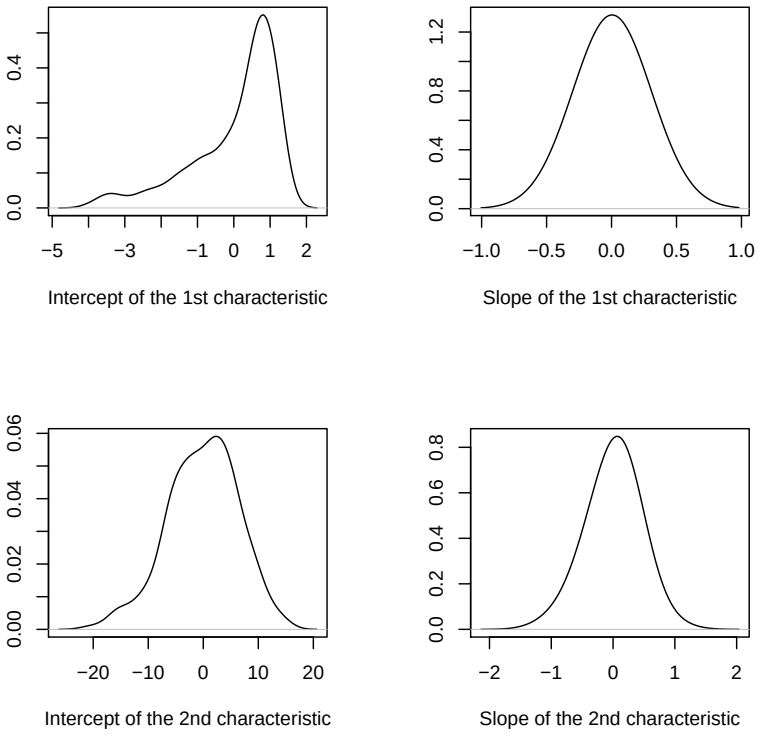

Figure 6.5: Distributions of the empirical estimates of the random effects under the log-gamma model

	Log-gamma		Normal	
corr	Estimate	SE	Estimate	SE
(b_{i10}, b_{i20})	0.4421	0.0210	0.5044	0.0568
(b_{i11}, b_{i21})	0.1896	0.4728	0.3280	0.2210

Table 6.3: Correlations of random effects under the log-gamma and normal models

Correlations between two random intercepts, ρ_0, and between two random slopes, ρ_1, of the two characteristics are of most interest. Correlations of pairs of random effects and their standard errors are obtained from the log-gamma and normal mixed effects models in Table 6.3. Under the log-gamma model, for the null hypothesis $H_0 : \rho_0 = 0$, a Wald test gives $z = 21.0803$ with p-value 0, which shows that there is significant positive relationship between the two random intercepts of the regressions of the two tests results on time. The positive relationship between the intercepts of the two characteristics is also indicated in the scatterplot of the intercepts from the single-patient regression analysis in Figure 6.2. The Wald test for hypothesis regarding the correlation of slopes $H_0 : \rho_1 = 0$ reports $z = 0.4009$ with p-value 0.3442. The intercepts represent the level of disease at the start of the study. The positive correlation coefficient of the two random intercepts indicates that the structural and functional characteristics are consistent with each other. The patient has a high level baseline in one test will be associated with a high level in the other. However, the rates of progressive RNFL loss are not significantly associated with the rates of functional change in glaucoma.

Under the normal model, the test for $H_0 : \rho_0 = 0$ gives $z = 8.8771$ with p-value 0, and the test for slopes $H_0 : \rho_1 = 0$ has $z = 1.4845$ with p-value 0.1377. Apparently the model with misspecified distributions of random effects reports less reliable conclusion on the correlations of random effects. The normal model overestimates the correlation of random slopes of the two characteristics, and underestimates its standard error, which results in a large value in the test statistic and a small p-value (0.067 for a one-sided alternative hypothesis). The performance of the log-gamma model and normal model in terms of estimating the correlations of random effects can also be illustrated by the comparison

of scatterplots of subject-specific effects in Figure 6.2 to 6.4. The analysis from the single patient regressions shows that the two random intercepts are positively correlated, but the two random slopes may be uncorrelated. The log-gamma model draws the same conclusion as the single patient regression analysis. The normal model gives a similar scatterplot for the two random intercepts, but indicates a positive correlation of the two random slopes, which is against the preliminary analysis.

6.7 Discussion

We propose a class of log-gamma linear mixed models for modelling multiple characteristics longitudinal data, applying it to a glaucoma study. To address the main scientific questions of whether the structural and functional measurements are associated and how the progressions of the two characteristics related, hypotheses are tested based on the fitting of the proposed model. We conclude that the progressions of the structural and functional characteristics are positively correlated, but the rates of changes in the two are uncorrelated. Further studies, either methodological or new investigations with more data, are warranted.

Besides that fact that they reflect the covariance structure of the multivariate responses, the main advantage of the class of log-gamma mixed effects models lies in their simplicity in accounting for the skewness of the random effects, which results in more efficient estimation of the parameters. The expression of the reordered random effects by the product of the linear transformation matrix and the random vector of independent components avoids high-dimensional integration in the marginal log likelihood function and makes

the implementation feasible. This new class provides a generalized method for estimating the correlation between two or more slopes. In the case of a single response, the model reduces to that in Zhang et al. (2008). The family of log-gamma distributions possesses the nice property that the limiting distribution is normal and a lack-of-fit test on the adequacy of the log-gamma distributional assumption of the random effects can be derived. It would be of interest to extend the model to the case of mixed multivariate responses in which the continuous characteristic response shows a skewed pattern.

Chapter 7

Future Work

7.1 Joint Models for Multivariate Mixed Outcomes of Repeated Measurements.

Observations of multiple outcomes across space and over time occur often in environmental and ecological studies. Compared to the number of spatial models for a single outcome variable in the exponential family of distributions, fewer statistical tools are available for multiple outcome variables that are not necessarily Gaussian. Wang and Wall (2003) [99] considered the exponential family of distributions for multiple-response variables, but only a univariate spatial process for a single latent variable. Zhu et al. (2005) [112] extended the model in Wang and Wall (2003) and developed a flexible class of generalized linear latent variable models for multivariate spatial-temporal data.

We have discussed the joint modelling of bivariate mixed outcomes of repeated measurements. The random mean joint models proposed for the joint analysis of two longitudinal sequences are less applicable to more sequences.

Even if a plausible joint model can be formulated in terms of the random mean joint models, the fitting of these high-dimensional models can be very cumbersome. Working on extending the random mean joint models to allow for multivariate outcomes, containing more than one continuous response and more than one discrete response, is a possible direction for future work.

One possible solution is to extend the random mean joint models for bivariate outcomes to multiple outcomes by using the composite likelihood methods. The composite likelihood is an inference function derived by multiplying a collection of component likelihood functions. The application areas include geostatistics, spatial extremes, and clustered and longitudinal data. We can consider the modelling of each pair of different types of outcomes by the random mean joint models. Assuming that, conditional on the latent variables, the responses are independent. A multivariate distributional assumption is imposed on the latent variables to account for the association between different outcomes.

For the model fitting, we can consider pairwise fitting. It is a pairwise pseudo-likelihood approach. Molenberghs and Verbeke (2005) [32] proposed a method of pairwise fitting, a composite marginal likelihood constructed from all pairs of outcomes. Instead of maximizing the log-likelihood function of the full joint model, all pairwise bivariate models are fitted separately in the first step. In the second step, the parameters obtained from the pairwise models are combined to obtain single estimate for each parameter in the full joint model. The pseudo-likelihood methods are often less efficient that the full maximum likelihood function. However, the simulation results of Fieuws and Verbeke (2006) [28] suggest that the loss of efficiency is negligible, if there is any.

7.2 Goodness of Fit for the Zero Modified Regression Model with Random Effects

Zero-inflated Poisson mixed regression models are popular approaches to analyzing clustered count data with excess zeros. Prior to application of these models, it is essential to examine the necessity of the adjustment for zero outcomes. The existing literature, however, has focused only on score tests for testing the suitability of zero-inflated models for correlated count data or other alternative approaches to the test, such as, the Wald and likelihood ratio tests for zero-inflation in correlated count data Xiang and Teo (2011) [102]. Under the zero modified regression model setting, it is essential to check whether or not the data are indeed zero-inflated relative to what are predicted by the standard models.

Investigating the goodness of fit under the zero modified regression model with random effects is of interest. For the independent count data, when the same covariates affect the probability of zeros and the Poisson parameter, it is useful to consider the model that involves the complementary-log-log link function for the probability of zeros and the log link for the Poisson parameter. This model can reduce to the generalized linear Poisson regression model when the coefficients of the two models are equal. In the zero modified regression models with random effects, there is no such result.

7.3 Zero Modified Regression Models for Semicontinuous Distributions

A semicontinuous variable, which has a portion of responses equal to 0, and a continuous distribution among the remaining values, has been described by the lognormal and gamma distributions for the continuous component. The zero modified regression models can easily be extended to semicontinuous distributions. For example, we can define the zero modified regression models with the normal distribution, and the discrete distribution with all mass at 0. By assigning an arbitrary value of the probability of zeros, we can scale the normal distribution according to the adjusted value.

A Poisson-gamma model has been introduced to account for between-subjects heterogeneity and within-subjects serial correlation occurring in longitudinal count data in Henderson and Shimakura (2003) [48]. The model extends the usual time-constant shared frailty approach to allow time-varying serially correlated gamma frailty whilst retaining standard marginal assumptions. They illustrate pairwise likelihood inference for the composite likelihood model in Lindsay (1988) [58] with the analysis of a clinical trial on the number of analgesic doses taken by hospital patients in successive time intervals following abdominal surgery. We can investigate the Poisson-correlated gamma-frailty zero modified model for data with excess zeros.

Appendix A

Derivatives of Log-likelihood Function

Suppressing the dependence on i, the first and second derivatives of (4.9), the log-likelihood function of the latent variable $\mathbf{u}$, with respect to $\boldsymbol{\theta}$ in R, and ρ in T are

$$\frac{\partial l}{\partial \theta_k} = -\frac{n}{2}\frac{1}{|R|}\frac{\partial |R|}{\partial \theta_k} - \frac{1}{2}\mathbf{u}'\left(\frac{\partial R^{-1}}{\partial \theta_k} \otimes T^{-1}\right)\mathbf{u},$$

$$\frac{\partial l}{\partial \rho} = -\frac{1}{|T|}\frac{\partial |T|}{\partial \rho} - \frac{1}{2}\mathbf{u}'\left(R^{-1} \otimes \frac{\partial T^{-1}}{\partial \rho}\right)\mathbf{u},$$

and

$$\frac{\partial^2 l}{\partial \theta_k^2} = \frac{n}{2}\frac{1}{|R|^2}\left(\frac{\partial |R|}{\partial \theta_k}\right)^2 - \frac{n}{2}\frac{1}{|R|}\frac{\partial^2 |R|}{\partial \theta_k^2} - \frac{1}{2}\mathbf{u}'\left(\frac{\partial^2 R^{-1}}{\partial \theta_k^2} \otimes T^{-1}\right)\mathbf{u},$$

$$\frac{\partial^2 l}{\partial \theta_k \partial \theta_j} = \frac{n}{2}\frac{1}{|R|^2}\frac{\partial |R|}{\partial \theta_j}\frac{\partial |R|}{\partial \theta_k} - \frac{n}{2}\frac{1}{|R|}\frac{\partial^2 |R|}{\partial \theta_k \partial \theta_j} - \frac{1}{2}\mathbf{u}'\left(\frac{\partial^2 R^{-1}}{\partial \theta_k \partial \theta_j} \otimes T^{-1}\right)\mathbf{u},$$

$$\frac{\partial^2 l}{\partial \theta_k \partial \rho} = -\frac{1}{2}\mathbf{u}'\left(\frac{\partial R^{-1}}{\partial \theta_k} \otimes \frac{\partial T^{-1}}{\partial \rho}\right)\mathbf{u},$$

$$\frac{\partial^2 l}{\partial \rho^2} = \frac{1}{|T|^2}\left(\frac{\partial |T|}{\partial \rho}\right)^2 - \frac{1}{|T|}\frac{\partial^2 |T|}{\partial \rho^2} - \frac{1}{2}\mathbf{u}'\left(R^{-1} \otimes \frac{\partial^2 T^{-1}}{\partial \rho^2}\right)\mathbf{u},$$

where $\frac{\partial |R|}{\partial \theta_k}$, $\frac{\partial |T|}{\partial \rho}$, $\frac{\partial R^{-1}}{\partial \theta_k}$, $\frac{\partial T^{-1}}{\partial \rho}$, $\frac{\partial^2 |R|}{\partial \theta_k^2}$, $\frac{\partial^2 |R|}{\partial \theta_k \partial \theta_j}$, $\frac{\partial^2 R^{-1}}{\partial \theta_k^2}$, $\frac{\partial^2 R^{-1}}{\partial \theta_k \partial \theta_j}$, $\frac{\partial^2 T^{-1}}{\partial \rho^2}$, and $\frac{\partial^2 |T|}{\partial \rho^2}$ are easy to be calculated.

Appendix B

Asymptotic Properties of the Log-Gamma Mixed Effects Models

We first state the conditions:

(A0) The distributions of the observations $\mathbf{Y}_i's$ have common support, so that the elements of the set $A_i = \{\mathbf{y}_i : f_i(\mathbf{y}_i|\boldsymbol{\theta}) > 0\}$ are independent of $\boldsymbol{\theta}$ for all i.

(A1) The distributions $f_i(\mathbf{y}_i|\boldsymbol{\theta})$ are identifiable.

(A2) The observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are independent.

(A3) There exists an open subset ω of Ω containing the true parameter point $\boldsymbol{\theta}_0$ such that for almost all $\mathbf{y}_i's$, the density $f_i(\mathbf{y}_i|\boldsymbol{\theta})$ has third derivatives with respect to $\boldsymbol{\theta} \in \omega$.

(A4) The first and second derivatives of l_i satisfy the equations

$$E_{\boldsymbol{\theta}_0}\left(\frac{\partial}{\partial\theta_j}l_i\right)=0 \quad \text{for } 1\le i\le N, 1\le j\le s \tag{2.1}$$

and

$$E_{\boldsymbol{\theta}_0}\left(\frac{\partial}{\partial\theta_j}l_i\cdot\frac{\partial}{\partial\theta_k}l_i\right)=E_{\boldsymbol{\theta}_0}\left(-\frac{\partial^2}{\partial\theta_j\partial\theta_k}l_i\right). \tag{2.2}$$

(A5) There exists a function $M(\mathbf{y}_i)$ such that

$$\left|\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}l_i(\boldsymbol{\theta},\mathbf{y}_i)\right|\le M(\mathbf{y}_i) \quad \text{for all}\,\boldsymbol{\theta}\in\omega, \tag{2.3}$$

where $E_{\boldsymbol{\theta}_0}\{M(\mathbf{Y}_i)\}<\infty$.

(A6) $E_{\boldsymbol{\theta}_0}(|\frac{\partial}{\partial\theta_j}l_i|^d)$, $1\le i\le N, 1\le j\le s$, are bounded for some $d>2$.

Proof. (a) Existence and consistency follow from the fact that the marginal probability (6.4) clearly satisfies the conditions (A0)-(A5).

(b) Asymptotic Normality. This part of the proof is basically to verify condition $(A6)$ for the marginal probability densities. Consider the integrand in the braces of (6.4):

$$\int\cdots\int \phi(\mathbf{y}_i|\mathbf{s}_i^*)\cdot\phi(s_{i,1}^*,\ldots,s_{i,q_0-1}^*,s_{i,q_0+1}^*,\ldots,s_{i,q}^*)\,ds_{i,1}^*\cdots ds_{i,q_0-1}^*ds_{i,q_0+1}^*\cdots ds_{i,q}^*.$$

It is still normal, since the convolution of normal distributions is normal. Let $\mathbf{V}_i$ denote the covariance matrix of the above multivariate normal distribution. Expressing the log likelihood function in a general way, we have

$$l_i(\boldsymbol{\theta},\mathbf{y}_i)=\ln\int\frac{1}{(2\pi)^{\frac{n_i}{2}}\mid\mathbf{V}_i\mid^{\frac{1}{2}}\Gamma(\kappa)}\cdot\exp\{F_i(s)\}\,ds, \tag{2.4}$$

where

$$F_i(s) = -\frac{1}{2}\mathbf{M}_i'\mathbf{V}_i^{-1}\mathbf{M}_i - e^{\frac{s+\eta\psi(\kappa)}{\eta}} + \frac{\kappa}{\eta}\{s + \eta\psi(\kappa)\} \tag{2.5}$$

and

$$\mathbf{M}_i = \mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{G}_0(s), \tag{2.6}$$

the components in $\mathbf{G}_0(s)$ are polynomials in s of at most degree one, and n_i is the number of observations of the p characteristics for the ith individual.

Differentiating the log likelihood with respect to the parameters, we obtain the following equations:

$$\frac{\partial l_i(\boldsymbol{\theta}, \mathbf{y}_i)}{\partial \boldsymbol{\beta}} = \frac{\int \exp\{F_i(s)\} \cdot \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{M}_i ds}{\int \exp\{F_i(s)\}\, ds} \tag{2.7}$$

$$\frac{\partial l_i(\boldsymbol{\theta}, \mathbf{y}_i)}{\partial \theta_j} = \frac{\int \exp\{F_i(s)\} \cdot \{-\frac{1}{2}\mathbf{M}_i'\frac{\partial \mathbf{V}_i^{-1}}{\partial \theta_j}\mathbf{M}_i - \frac{1}{2}|\mathbf{V}_i|^{-1}\frac{\partial |\mathbf{V}_i|}{\partial \theta_j}\}\, ds}{\int \exp\{F_i(s)\}\, ds} \tag{2.8}$$

$$\frac{\partial l_i(\boldsymbol{\theta}, \mathbf{y}_i)}{\partial \kappa} = \frac{\int \exp\{F_i(s)\} \cdot \{\frac{s}{\eta} - \psi'(\kappa)e^{\frac{s+\eta\psi(\kappa)}{\eta}} + \kappa\psi'(\kappa)\}\, ds}{\int \exp\{F_i(s)\}\, ds} \tag{2.9}$$

$$\frac{\partial l_i(\boldsymbol{\theta}, \mathbf{y}_i)}{\partial \eta} = \frac{\int \exp\{F_i(s)\} \cdot \{\kappa s - se^{\frac{s+\eta\psi(\kappa)}{\eta}} + \eta\}\, ds}{\eta^2 \int \exp\{F_i(s)\}\, ds} \tag{2.10}$$

where θ_j is the variance component in $\mathbf{G}, \mathbf{R}$, and $\boldsymbol{\Sigma}_i$.

We first show the steps for proving that the first derivatives of the log likelihood function with respect to the fixed effects $\boldsymbol{\beta}$ satisfy condition $(A6)$. The proofs for the variance components and the log-gamma parameters are

similar.

$$
\begin{aligned}
E_{\boldsymbol{\theta}_0}(|\frac{\partial}{\partial \boldsymbol{\beta}} l_i|^d) &= \int \left| \frac{\partial l_i(\boldsymbol{\theta}, \mathbf{y}_i)}{\partial \boldsymbol{\beta}} \right|^d \cdot \left[\int \frac{1}{(2\pi)^{\frac{n_i}{2}} \mid \mathbf{V}_i \mid^{\frac{1}{2}} \Gamma(\kappa)} \cdot \exp\{F_i(s)\}\, ds \right] d\mathbf{y}_i \\
&= \int \left[\frac{\left| \int \exp\{F_i(s)\} \cdot \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{M}_i ds \right|^d}{(2\pi)^{\frac{n_i}{2}} \mid \mathbf{V}_i \mid^{\frac{1}{2}} \Gamma(\kappa) \left| \int \exp\{F_i(s)\}\, ds \right|^d} \cdot \int \exp\{F_i(s)\}\, ds \right] d\mathbf{y}_i \\
&= \int \frac{\left| \int \exp\{F_i(s)\} \cdot \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{M}_i ds \right|^d}{(2\pi)^{\frac{n_i}{2}} \mid \mathbf{V}_i \mid^{\frac{1}{2}} \Gamma(\kappa) \left| \int \exp\{F_i(s)\}\, ds \right|^{d-1}} d\mathbf{y}_i \\
&= \int \frac{\left| \int \exp\{F_i(s)\} \cdot \mathbf{G}_1(s, \mathbf{y}_i)\, ds \right|^d}{(2\pi)^{\frac{n_i}{2}} \mid \mathbf{V}_i \mid^{\frac{1}{2}} \Gamma(\kappa) \left| \int \exp\{F_i(s)\}\, ds \right|^{d-1}} d\mathbf{y}_i,
\end{aligned}
$$

where the vector $\mathbf{G}_1(s, \mathbf{y}_i) = \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{M}_i$ has as components polynomials in s or elements in $\mathbf{y}_i$ of at most degree one.

By Hölder's inequality, for $d > 1$, we have

$$
\begin{aligned}
&\left| \int \exp\{F_i(s)\} \cdot \mathbf{G}_1(j)\, ds \right| \\
&\leq \left(\int \left[(\exp\{F_i(s)\})^{\frac{d-1}{d}} \right]^{\frac{d}{d-1}} ds \right)^{\frac{d-1}{d}} \cdot \left(\int \left| (\exp\{F_i(s)\})^{\frac{1}{d}} \cdot \mathbf{G}_1(j) \right|^d ds \right)^{\frac{1}{d}} \\
&= \left[\int \exp\{F_i(s)\}\, ds \right]^{\frac{d-1}{d}} \cdot \left[\int \left| \exp\{F_i(s)\} \cdot \mathbf{G}_1^d(j) \right| ds \right]^{\frac{1}{d}},
\end{aligned}
$$

where $\mathbf{G}(j)$ is the jth component of $\mathbf{G}_1(s, \mathbf{y}_i)$. Thus,

$$
\begin{aligned}
&\int \frac{\left|\int \exp\{F_i(s)\} \cdot \mathbf{G}_1(j)\, ds\right|^d}{(2\pi)^{\frac{n_i}{2}} \mid \mathbf{V}_i \mid^{\frac{1}{2}} \Gamma(\kappa) \left|\int \exp\{F_i(s)\}\, ds\right|^{d-1}}\, d\mathbf{y}_i \\
&\leq \frac{1}{(2\pi)^{\frac{n_i}{2}} \mid \mathbf{V}_i \mid^{\frac{1}{2}} \Gamma(\kappa)} \int \left[\int \left|\exp\{F_i(s)\} \cdot \mathbf{G}_1^d(j)\right|\, ds\right] d\mathbf{y}_i \\
&= \frac{1}{(2\pi)^{\frac{n_i}{2}} \mid \mathbf{V}_i \mid^{\frac{1}{2}} \Gamma(\kappa)} \int \left[\int \exp(-\frac{1}{2}\mathbf{M}_i' \mathbf{V}_i^{-1} \mathbf{M}_i) \cdot |\mathbf{G}_1^d(j)|\, d\mathbf{y}_i\right] \\
&\quad \cdot \exp\left\{-e^{\frac{s+\eta\psi(\kappa)}{\eta}} + \frac{\kappa}{\eta}(s + \eta\psi(\kappa))\right\} ds \\
&< \infty.
\end{aligned}
$$

The verifications of condition (A6) for the remaining parameters in $\boldsymbol{\theta}$ are similar to those of $\boldsymbol{\beta}$. □

Bibliography

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. United States Department of Commerce, 1964.

[2] Paul S. Albert and Dean A. Follmann. Modeling repeated count data subject to informative dropout. *Biometrics*, 56(3):pp. 667–677, 2000.

[3] Takeshi Amemiya. *Advanced Econometrics*. Oxford, U.K.: Basil Blackwell, 1 edition, 1985.

[4] M. S. Bartlett and D. G. Kendall. The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society*, 8(1):pp. 128–138, 1946.

[5] K. N. Berk and P. A. Lachenbruch. Repeated measures with zeros. *Statistical Methods in Medical Research*, 11:pp. 303–316, 2002.

[6] James G. Booth and James P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(1):pp. 265–285, 1999.

[7] Kurt Brännäs and Per Johansson. Time series count data regression. *Communications in Statistics - Theory and Methods*, 23(10):2907–2925, 1994.

[8] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):pp. 9–25, 1993.

[9] Jan van den Broek. A score test for zero inflation in a poisson distribution. *Biometrics*, 51(2):pp. 738–743, 1995.

[10] M. J. Campbell. Time series regression for counts: An investigation into the relationship between sudden infant death syndrome and environmental temperature. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 157(2):pp. 191–208, 1994.

[11] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, 2 edition, 2001.

[12] Paul J. Catalano and Louise M. Ryan. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87(419):pp. 651–658, 1992.

[13] K. S. Chan and Johannes Ledolter. Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):pp. 242–252, 1995.

[14] Kung Sik Chan and Charles J. Geyer. Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):pp. 1747–1758, 1994.

[15] D. R. Cox, Gudmundur Gudmundsson, Georg Lindgren, Lennart Bondesson, Erik Harsaae, Petter Laake, Katarina Juselius, and Steffen L. Lauritzen. Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 8(2):pp. 93–115, 1981.

[16] D. R. Cox and Nanny Wermuth. Response models for mixed binary and quantitative variables. *Biometrika*, 79(3):pp. 441–461, 1992.

[17] Michael J. Daniels and Sharon-lise T. Normand. Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics*, 7(1):1–15, 2006.

[18] Richard A. Davis, William T. M. Dunsmuir, and Ying Wang. On autocorrelation in a poisson regression model. *Biometrika*, 87(3):pp. 491–505, 2000.

[19] N. G. de Bruijn. *Asymptotic Methods in Analysis*. Courier Dover Publications, 1981.

[20] A. R. De Leon and B. Wu. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, 30(2):175–185, 2011.

[21] Dianliang Deng and Sudhir R. Paul. Score tests for zero inflation in generalized linear models. *The Canadian Journal of Statistics*, 28(3):pp. 563–570, 2000.

[22] Peter J. Diggle, Kung-Yee Liang, and Zeger Scott L. *Analysis of Longitudinal Data*. Oxford Science Publications, 1995.

[23] Marijtje A. J. van Duijn and Ulf Böckenholt. Mixture models for the analysis of repeated count data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):pp. 473–485, 1995.

[24] David B. Dunson. Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(2):pp. 355–366, 2000.

[25] Christel Faes, Marc Aerts, Geert Molenberghs, Helena Geys, Greet Teuns, and Luc Bijnens. A high-dimensional joint model for longitudinal outcomes of different nature. *Statistics in Medicine*, 27(22):4408–4427, 2008.

[26] Vernon T. Farewell. Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262, 1986.

[27] A. Feuerverger. On some methods of analysis for weather experiments. *Biometrika*, 66(3):655–658, 1979.

[28] Steffen Fieuws and Geert Verbeke. Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431, 2006.

[29] Garrett M. Fitzmaurice and Nan M. Laird. Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, 90(431):pp. 845–852, 1995.

[30] Garrett M. Fitzmaurice and Nan M. Laird. Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, 53(1):pp. 110–122, 1997.

[31] Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied Longitudinal Analysis*. John Wiley & Sons, 1 edition, 2004.

[32] Molenberghs Geert and Verbeke Geert. *Models for Discrete Longitudinal Data*. Springer, 1 edition, 2005.

[33] John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica*, 57(6):pp. 1317–1339, 1989.

[34] Charles J. Geyer. Practical markov chain monte carlo. *Statistical Science*, 7(4):pp. 473–483, 1992.

[35] Pulak Ghosh, Marcia D. Branco, and Hrishikesh Chakraborty. Bivariate random effect model using skew-normal distribution with application to hiv-rna. *Statistics in Medicine*, 26(6):1255–1267, 2007.

[36] Harvey Goldstein. Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, 78(1):pp. 45–51, 1991.

[37] Gene H. Golub. Some modified matrix eigenvalue problems. *SIAM Review*, 15(2):pp. 318–334, 1973.

[38] Gene H. Golub and John H. Welsch. Calculation of gauss quadrature rules. *Mathematics of Computation*, 23(106):pp. 221–230+s1–s10, 1969.

[39] William H. Greene. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. *Working Paper EC-94-10, Department of Economics, New York University*, 1994.

[40] J. T. Grogger and Richard T. Carson. Models for truncated counts. *Journal of Applied Econometrics*, 6(3):225–38, July-Sept 1991.

[41] Ralitza V. Gueorguieva and Alan Agresti. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association*, 96(455):pp. 1102–1112, 2001.

[42] Pushpa L. Gupta, Ramesh C. Gupta, and Ram C. Tripathi. Analysis of zero-adjusted count data. *Computational Statistics and Data Analysis*, 23(2):207 – 218, 1996.

[43] Daniel B. Hall. Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*, 56(4):pp. 1030–1039, 2000.

[44] A. C. Harvey and C. Fernandes. Time series models for count or qualitative observations. *Journal of Business & Economic Statistics*, 7(4):pp. 407–417, 1989.

[45] J. L. Hay and A. N. Pettitt. Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. *Biostatistics*, 2(4):433–444, 2001.

[46] D. Heilbron. Generalized linear models for altered zero probabilities and overdispersion in count data. Technical report, Department of Epidemiology and Biostatistics, University of California, San Francisco, 1989.

[47] David C. Heilbron. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36(5):531–547, 1994.

[48] Robin Henderson and Silvia Shimakura. A serially correlated gamma frailty model for longitudinal count data. *Biometrika*, 90(2):pp. 355–366, 2003.

[49] James P. Hobert and Charles J. Geyer. Geometric ergodicity of gibbs and block gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis*, 67(2):414–430, November 1998.

[50] N. Jansakul and J. P. Hinde. Score tests for zero-inflated poisson models. *Computational Statistics and Data Analysis*, 40(1):75 – 96, 2002.

[51] H. Kahn and A. W. Marshall. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):pp. 263–278, 1953.

[52] Heinz Kaufmann. Regression models for nonstationary categorical time series: Asymptotic estimation theory. *The Annals of Statistics*, 15(1):pp. 79–98, 1987.

[53] S. A. Klugman, H. H. Panjer, and G. E. Willmot. *Loss Models: From Data to Decisions*. Wiley, 3 edition, 2009.

[54] Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):pp. 963–974, 1982.

[55] Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):pp. 1–14, 1992.

[56] Emmanuel Lesaffre and Bart Spiessens. On the effect of the number of quadrature points in a logistic random-effects model: An example. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50(3):pp. 325–335, 2001.

[57] Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):pp. 13–22, 1986.

[58] Bruce G. Lindsay. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988.

[59] Mary J. Lindstrom and Douglas M. Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):pp. 673–687, 1990.

[60] Qing Liu and Donald A. Pierce. A note on gauss-hermite quadrature. *Biometrika*, 81(3):pp. 624–629, 1994.

[61] Thomas A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):pp. 226–233, 1982.

[62] Laurence S. Magder and Scott L. Zeger. A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association*, 91(435):pp. 1141–1151, 1996.

[63] P. McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall, 2 edition, 1989.

[64] Charles E. McCulloch. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437):pp. 162–170, 1997.

[65] C. A. McGilchrist. Estimation in generalized mixed models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):pp. 61–69, 1994.

[66] Jr. Meeker, William Q. Limited failure population life tests: Application to integrated circuit reliability. *Technometrics*, 29(1):pp. 51–65, 1987.

[67] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the hastings and metropolis algorithms. *The Annals of Statistics*, 24(1):pp. 101–121, 1996.

[68] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag, London, 1 edition, 1993.

[69] J. Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):pp. 341–365, 1986.

[70] Per Mykland, Luke Tierney, and Bin Yu. Regeneration in markov chain samplers. *Journal of the American Statistical Association*, 90(429):pp. 233–241, 1995.

[71] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384, 1972.

[72] Maren K. Olsen and Joseph L. Schafer. A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):pp. 730–745, 2001.

[73] Jos C. Pinheiro and Douglas M. Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):pp. 12–35, 1995.

[74] Jos C. Pinheiro, Chuanhai Liu, and Ying Nian Wu. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics*, 10(2):pp. 249–276, 2001.

[75] Stuart J. Pocock, Nancy L. Geller, and Anastasios A. Tsiatis. The analysis of multiple endpoints in clinical trials. *Biometrics*, 43(3):pp. 487–498, 1987.

[76] R. L. Prentice. A log gamma model and its maximum likelihood estimation. *Biometrika*, 61(3):pp. 539–544, 1974.

[77] Gregory Reinsel. Estimation and prediction in a multivariate random effects generalized linear model. *Journal of the American Statistical Association*, 79(386):pp. 406–414, 1984.

[78] M. Ridout, C. G. B. Demétrio, and J. Hinde. Models for count data with many zeros. *Proceedings of the XIXth International Biometric Conference, Cape Town*, pages 179–192, 1998.

[79] Martin Ridout, John Hinde, and Clarice G. B. Demétrio. A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57(1):pp. 219–223, 2001.

[80] Gareth O. Roberts and Jeffrey S. Rosenthal. Convergence of slice sampler markov chains. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):pp. 643–660, 1999.

[81] G. Rodríguez. Lecture notes on generalized linear models, 2007.

[82] Jason Roy and Xihong Lin. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, 56(4):pp. 1047–1054, 2000.

[83] David Ruppert. Discussion of "maximization by parts in likelihood inference," by song, peter x.-k. and fan, yanqin and kalbfleisch. *Journal of the American Statistical Association*, 100(472):pp. 1145–1167, 2005.

[84] Mary Dupuis Sammel, Louise M. Ryan, and Julie M. Legler. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3):pp. 667–678, 1997.

[85] Robert Schall. Estimation in generalized linear models with random effects. *Biometrika*, 78(4):pp. 719–727, 1991.

[86] Amrik Shah, Nan Laird, and David Schoenfeld. A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, 92(438):pp. 775–779, 1997.

[87] Michael Sherman and Saskia le Cessie. A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics - Simulation and Computation*, 26(3):901–925, 1997.

[88] Peter X.-K. Song, Mingyao Li, and Ying Yuan. Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, 65(1):60–68, 2009.

[89] Robert Stiratelli, Nan Laird, and James H. Ware. Random-effects models for serial observations with binary response. *Biometrics*, 40(4):pp. 961–971, 1984.

[90] Martin A. Tanner. *Tools for Statistical Inference*. Springer, 3 edition, 1996.

[91] Peter F. Thall. Mixed poisson likelihood regression models for longitudinal interval count data. *Biometrics*, 44(1):pp. 197–209, 1988.

[92] Peter F. Thall and Stephen C. Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46(3):pp. 657–671, 1990.

[93] R. Thiébaut, H. Jacqmin-Gadda, G. Chêne, C. Leport, and D. Commenges. Bivariate linear mixed models using sas proc mixed. *Computer Methods and Programs in Biomedicine*, 69(3):249 – 256, 2002.

[94] Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):pp. 1701–1728, 1994.

[95] Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):pp. 82–86, 1986.

[96] Janet A. Tooze, Gary K. Grunwald, and Richard H. Jones. Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research*, 11(4):341–355, 2002.

[97] Geert Verbeke and Emmanuel Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):pp. 217–221, 1996.

[98] Geert Verbeke and Emmanuel Lesaffre. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis*, 23(4):541–556, 1997.

[99] Fujun Wang and Melanie M. Wall. Generalized common spatial factor model. *Biostatistics*, 4(4):569–582, 2003.

[100] Greg C. G. Wei and Martin A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):pp. 699–704, 1990.

[101] Liming Xiang, Andy H. Lee, Kelvin K. W. Yau, and Geoffrey J. McLachlan. A score test for zero-inflation in correlated count data. *Statistics in Medicine*, 25(10):1660–1671, 2006.

[102] Liming Xiang and Guo Shou Teo. A note on tests for zero-inflation in correlated count data. *Communications in Statistics: Simulation and Computation*, 40(7):992–1005, 2011.

[103] Kelvin K. W. Yau, Andy H. Lee, and Angus S. K. Ng. A zero-augmented gamma mixed model for longitudinal data with many zeros. *Australian and New Zealand Journal of Statistics*, 44(2):177–183, 2002.

[104] Scott L. Zeger. A regression model for time series of counts. *Biometrika*, 75(4):pp. 621–629, 1988.

[105] Scott L. Zeger and M. Rezaul Karim. Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):pp. 79–86, 1991.

[106] Scott L. Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):pp. 121–130, 1986.

[107] Scott L. Zeger, Kung-Yee Liang, and Paul S. Albert. Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4):pp. 1049–1060, 1988.

[108] Scott L. Zeger, Kung-Yee Liang, and Steven G. Self. The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, 72(1):pp. 31–38, 1985.

[109] Scott L. Zeger and Bahjat Qaqish. Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, 44(4):pp. 1019–1031, 1988.

[110] Peng Zhang. *Contributions to Mixed Effects Models for Longitudinal Data*. PhD thesis, University of Waterloo, May 2006.

[111] Peng Zhang, Peter X.-K. Song, Annie Qu, and Tom Greene. Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models. *Biometrics*, 64(1):pp. 29–38, 2008.

[112] J. Zhu, J. C. Eickhoff, and P. Yan. Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. *Biometrics*, 61(3):pp. 674–683, 2005.