Methods for Longitudinal Data Measured at Distinct Time Points

by

Xiaoqin Xiong

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2010

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

For longitudinal data where the response and time-dependent predictors within each individual are measured at distinct time points, traditional longitudinal models such as generalized linear mixed effects models or marginal models cannot be directly applied. Instead, some preprocessing such as smoothing is required to temporally align the response and predictors.

In Chapter 2, we propose a binning method, which results in equally spaced bins of time for both the response and predictor(s). Hence, after incorporating binning, traditional models can be applied. The proposed binning approach was applied on a longitudinal hemodialysis study to look for possible contemporaneous and lagged effects between occurrences of a health event (i.e., infection) and levels of a protein marker of inflammation (i.e., C-reactive protein). Both Poisson mixed effects models and zero-inflated Poisson (ZIP) mixed effects models were applied to the subsequent binned data, and some important biological findings about contemporaneous and lagged associations were uncovered. In addition, a simulation study was conducted to investigate various properties of the binning approach.

In Chapter 3, asymptotic properties have been derived for the fixed effects association parameter estimates following binning, under different data scenarios. In addition, we propose some leave-one-subject-out cross-validation algorithms for bin size selection.

In Chapter 4, in order to identify levels of a predictor that might be indicative of recently occurred event(s), we propose a generalized mixed effects regression tree (GMRTree) based method which estimates the tree by standard tree method such as CART and estimates the random effects by a generalized linear mixed effects model. One of the main steps in this method was to use a linearization technique to change the longitudinal count response into a continuous surrogate response. Simulations have shown that the GMRTree method can effectively detect the underlying tree structure in an applicable longitudinal dataset, and

has better predictive performance than either a standard tree approach without random effects or a generalized linear mixed effects model, assuming the underlying model indeed has a tree structure. We have also applied this method to two longitudinal datasets, one from the aforementioned hemodialysis study and the other from an epilepsy study.

Acknowledgements

I would like to express my deep and sincere gratitude to my supervisor, Dr. Joel A. Dubin for his insight, wise direction, constant encouragement and for his support in many aspects. I am not only respectful for his wide knowledge and logical way of thinking toward research, but also very appreciated for his generosity to me. His understanding, encouraging and personal guidance have provided a good basis for the present thesis.

My special thanks go to Dr. Grace Y. Yi, Dr. Richard J. Cook, Dr. Phil Bigelow and Dr. Gary Sneddon for their valuable advice and for serving as thesis committee members.

I also would like to thank both Dr. George A. Kaysen and Dr. Patrick S. Romano at the University of California, Davis for providing the the hemodialysis application data.

I am so indebted to my parents Mingshan Xiong and Duolian Fang for their support and encouragement throughout my whole life. Last, but not least, I am deeply grateful for my husband, Jun Liu, and my children, Andrew Liu and Terry Liu. Without their understanding, love and patience, I would not have been able to finish this thesis.

Dedication

For my husband Jun Liu, my children Andrew, Terry, and the rest of my family for their support and love in my research and beyond.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to longitudinal studies

Longitudinal studies are increasingly common in many areas of research including medicine, public health, and the social sciences. The defining feature of longitudinal studies is that individuals are repeatedly measured over time. This is in contrast to cross-sectional studies, where a single outcome is measured for each individual. The primary goal of a longitudinal study is often to characterize the change in responses over time as well as factors that influence this change.

To analyze longitudinal data, mainly three types of models are used: mixed effects (conditional) models, marginal models and transition models.

For mixed effects models, Harville (1977) introduced a general class of two-stage linear mixed effects models and variance component estimation procedures via maximum likelihood and restricted maximum likelihood. Laird and Ware (1982), and Lindstrom and Bates (1988) proposed iterative estimation procedures for linear mixed effects models. Vonesh and Carter (1987) proposed a non-iterative procedure for estimation of model parameters.

Cnaan et al. (1997) provided a detailed review of linear mixed effects models with an application to a schizophrenia clinical trial. Non-linear mixed effects models have been proposed by many authors for different applications; for example, Lindstrom and Bates (1990) proposed an iterative EM-type algorithm for parameter estimation for a two-stage non-linear mixed effects model, which can be applied to non-linear longitudinal data such as in pharmacodynamics and pharmacokinetics.

Generalized linear mixed models (GLMMs) are widely used in longitudinal studies where the responses are non-normal, such as binary and count responses. When estimating parameters for GLMMs, the most common strategy is to obtain the marginal likelihood by integrating the random effects out from the joint likelihood of the observable responses and random effects. However, it is very computationally intensive due to evaluation of high-dimensional integrals, and often cannot be solved in closed form. To handle these concerns, Stiratelli et al. (1984) discussed an EM algorithm approach for binary response data with Gaussian random effects. Longford (1993) discussed an approach based on direct maximization of the likelihood. Anderson and Aitkin (1985) proposed to use adaptive Gaussian quadratures for the evaluation of integrals over the random effects. See also Crouch and Spiegelman (1990), Press et al. (1992), McCulloch(1994), and Liu and Pierce (1994) for more information on quadrature method. But the quadrature methods can have trouble with many random effects and/or with too many nested levels of random effects. As an alternative, Tierney and Kadane (1986), Solomon and Cox (1992), Liu and Pierce (1993) proposed Laplace approximations of the integrated likelihood. Breslow and Clayton (1993) proposed the penalized quasi-likelihood (PQL) approach as an attempt to extend quasi-likelihood to GLMMs. See also Schall (1991), Wolfinger and O'Connell (1993), McGilchrist (1994) for similar approaches as PQL. Breslow and Lin (1995), Lin and Breslow (1996) have shown that PQL estimators can be biased and inconsistent for highly non-normal (e.g., binary) responses. Goldstein (1991) proposed marginal quasi-likelihood (MQL) approach for estimating the parameters for GLMMs. Lee and Nelder

(1996, 2001) proposed a hierarchical likelihood model that avoids numerical evaluation of the integral in the likelihood function. Sutradhar et al. (2008), Jowaheer et al. (2009) and Sutradhar (2010) proposed a generalized quasi-likelihood (GQL) approach for familial-longitudinal data such as binary and count data. GQL method will not only account for familial random effects, but also allows for specification of the correlation structure of the longitudinal observations. See also Sutradhar (2003) for GQL method. There are also other approaches to estimate the parameters of GLMMs, such as Monte Carlo ML methods (McCulloch, 1997; Booth and Hobert, 1999), Gibbs sampling Bayesian method (Zeger and Karim, 1991) and stochastic approximation method (Gu and Kong, 1998).

For marginal modeling, Liang and Zeger (1986) and Zeger and Liang (1986) introduced generalized estimating equations (GEE). The pioneering paper of Godambe (1960) elucidated the optimality theory for estimating functions which laid a theoretical foundation for GEE. Prentice (1988) and Zhao and Prentice (1990) proposed extensions of GEE methodology to incorporate assumptions and estimation of higher-order moments. The resulting methods have been called GEE2.

Transition models examine the effect of past response and covariates on the transition patterns across responses (continuous or discrete) over time. Various autoregressive-type regression models exist for modeling transitional patterns in binary, categorical and ordinal longitudinal data. See Cox (1970), Muenz and Rubinstein (1985), Zeger and Qaqish (1988) on autoregressive models for binary data; Diggle et al. (2002), Fahrmeir and Kaufmann (1987) on ordinal and repeated categorical data respectively. Cook et al. (2004) proposed a continuous Markov model that incorporated multiplicative random effects to reflect the clustering in processes within subjects.

1.2 Mixed effects models

Mixed-effects models have been developed in a variety of disciplines, with varying names and terminologies: random-effects models (statistics, econometrics), variance and covariance-component models (statistics), hierarchical linear models (education), multi-level models (sociology), contextual-effects models (sociology), random-coefficient models (econometrics), repeated-measures models (statistics, psychology).

Mixed models extend classical linear/generalized linear models by including random or subject-specific effects in the structure for the mean. The random effects not only determine (at least partially) the structure of correlation between observations on the same subject, they also take account of heterogeneity among subjects, due to unobserved characteristics.

1.2.1 Linear mixed effects models (LMEs)

Since the appearance of Laird and Ware (1982), linear mixed models have been widely used for analysis of clustered data, including longitudinal data. A linear mixed effects model assumes that each subject's measurements follow a linear regression model with random effects, such as a random intercept and/or slope. A linear mixed effects model takes the form

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i,\ i = 1, \ldots, N\ . \tag{1.1}$$

- $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ is a $n_i \times 1$ vector of repeated measures for subject i.
- $\mathbf{X}_i = (\mathbf{X}'_{i1}, \ldots, \mathbf{X}'_{in_i})'$ is a $n_i \times p$ matrix of covariates associated with fixed effects $\boldsymbol{\beta}$, where $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})', j = 1, 2, \ldots, n_i$.
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown population level fixed effects.
- $\mathbf{Z}_i = (\mathbf{Z}'_{i1}, \ldots, \mathbf{Z}'_{in_i})'$ is a $n_i \times q$ matrix of covariates associated with random effects $\mathbf{u}_i$ where $\mathbf{Z}_{ij} = (Z_{ij1}, Z_{ij2}, \ldots, Z_{ijq})', j = 1, 2, \ldots, n_i$.

- $\mathbf{u}_i$ is a $q \times 1$ vector of unknown subject-specific random effects. In principle, any multivariate distribution for $\mathbf{u}_i$ could be assumed; however, in practice, it is typical to assume that $\mathbf{u}_i \sim N_q(0, \mathbf{D})$, where $\mathbf{D}$ is an unknown $q \times q$ positive-definite covariance matrix.

- $\mathbf{e}_i$ is a $n_i \times 1$ vector of within-subject error terms for subject i. In practice, we typically assume $\mathbf{e}_i \sim N_{n_i}(0, \mathbf{\Sigma}_i)$, where $\mathbf{\Sigma}_i$ is an unknown $n_i \times n_i$ positive-definite covariance matrix.

- It is generally assumed $\mathbf{u}_1, \ldots, \mathbf{u}_N$ are independent of each other, and $\mathbf{e}_1, \ldots, \mathbf{e}_N$ are independent of each other. We also assume $\mathbf{u}_i$ are independent of $\mathbf{e}_i$.

In (1.1), $\mathbf{Z}_i$ is a known design matrix linking the vector of random effect $\mathbf{u}_i$ to $\mathbf{Y}_i$. $\mathbf{Z}_i$ is most often a subset of columns of $\mathbf{X}_i$. The random effect $\mathbf{u}_i$ induces a correlation structure for repeated measurements within a subject.

If $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$, then conditional on the random effects $\mathbf{u}_i$, the Y_{ij} are independent for subject i, i.e., the Y_{ij} are conditionally independent given $\mathbf{u}_i$ (conditional independence). However, marginally, the Y_{ij} are not independent (for a given subject i), with the correlation among the Y_{ij} for subject i being induced by the $\mathbf{u}_i$. Also, $var(\mathbf{Y}_i)$ accounts for both within-subject (or intrasubject) variability via $\mathbf{\Sigma}_i$, and between-subject (or intersubject) variability via $\mathbf{Z}_i \mathbf{D} \mathbf{Z}_i^{'}$, suggesting additional dependence among the $\mathbf{Y}_i$ can be introduced through $\mathbf{\Sigma}_i$ such as autocorrelated structure.

Usually the parameters $(\boldsymbol{\beta}, \mathbf{\Sigma}, \mathbf{D})$ of LMEs are estimated by maximum likelihood or restricted maximum likelihood (REML). The prediction of random effects $\mathbf{u}_i$ are given by the "empirical best linear unbiased predictor" (empirical BLUP) or the "empirical Bayes" (EB) estimates. Refer to Verbeke and Molenberghs (2000), Fitzmaurice et al. (2004), for example, for more details on estimation and inference for LMEs.

1.2.2 Generalized linear mixed effects models (GLMMs)

For non-normal longitudinal data, especially for binary and count longitudinal data, generalized linear mixed effects models are developed which extend the generalized linear model by the inclusion of random effects into the model (see, for example, Breslow and Clayton (1993)). Conditional on the random effects, it is assumed that responses from a subject are independent observations from a distribution belonging to an exponential family. We express the GLMMs as follows:

$$\begin{aligned} Y_{ij}|\mathbf{u}_i &\sim f_{Y_{ij}|\mathbf{u}_i}(y_{ij}|\mathbf{u}_i) \\ f_{Y_{ij}|\mathbf{u}_i}(y_{ij}|\mathbf{u}_i) &= exp\{\frac{y_{ij}\gamma_{ij} - d(\gamma_{ij})}{\phi} - c(y_{ij},\phi)\} \\ E[Y_{ij}|\mathbf{u}_i] &= \mu_{ij} = d^{'}(\gamma_{ij}) \\ g(\mu_{ij}) &= \mathbf{X}_{ij}^{'}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{'}\mathbf{u}_i \end{aligned} \tag{1.2}$$

where $i = 1, \ldots, N; j = 1, \ldots, n_i$. $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ are the covariate vectors for fixed effects $\boldsymbol{\beta}$ and random effects $\mathbf{u}_i$ respectively, which are similarly defined as in the linear mixed effects model.

The assumptions for GLMMs are

- $g(\cdot)$ is some known link function, linking the conditional mean of Y_{ij} and the linear form of the predictors (and random effects).
- Random effects $\mathbf{u}_i$ is a $q \times 1$ vector. The conditional distribution of Y_{ij}, given $\mathbf{u}_i$, belongs to the exponential family. It is also assumed that $Y_{ij}|\mathbf{u}_i$ are independent of one another.
- $var(Y_{ij}|\mathbf{u}_i) = \phi v(\mu_{ij})$ where $v(\cdot)$ is a known variance function.
- The random effects are assumed to have some probability distribution. In practice, it is typical to assume $\mathbf{u}_i \sim N_{q\times q}(0, \mathbf{D})$.

Let $f(\mathbf{u}_i|\mathbf{D})$ be the density function of the multivariate normally distributed random effects with mean 0 and covariance matrix $\mathbf{D}$. The joint probability for $\mathbf{Y}_i$ and $\mathbf{u}_i$ can be expressed as:

$$f(\mathbf{Y}_i, \mathbf{u}_i) = f(\mathbf{Y}_i|\mathbf{u}_i)f(\mathbf{u}_i) \text{ where } f(\mathbf{Y}_i|\mathbf{u}_i) = f(Y_{i1}|\mathbf{u}_i)f(Y_{i2}|\mathbf{u}_i)\dots f(Y_{in_i}|\mathbf{u}_i) \quad .$$

Then marginal probability for subject i is

$$f_i(\mathbf{Y}_i|\boldsymbol{\beta}, \mathbf{D}, \phi) = \int \prod_{j=1}^{n_i} f(Y_{ij}|\mathbf{u}_i, \boldsymbol{\beta}, \phi) f(\mathbf{u}_i|\mathbf{D}) d\mathbf{u}_i \quad .$$

So the likelihood function for $\boldsymbol{\beta}, \mathbf{D}, \phi$ is

$$L(\boldsymbol{\beta}, \mathbf{D}, \phi) = \prod_{i=1}^{N} f_i(\mathbf{Y}_i|\boldsymbol{\beta}, \mathbf{D}, \phi) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(Y_{ij}|\mathbf{u}_i, \boldsymbol{\beta}, \phi) f(\mathbf{u}_i|\mathbf{D}) d\mathbf{u}_i \tag{1.3}$$

In order to solve the likelihood, integration over the random-effects distribution must be performed. As a result, estimation is much more complicated than the case of linear mixed effect models where the solution can be expressed in a closed form. Various approximations for evaluating the integral over the random-effects distribution have been proposed in the literature such as Gauss-Hermit quadrature method, Laplace method and penalized-quasi likelihood (PQL). These approaches are summarized elsewhere including in Molenberghs and Verbeke (2005) and McCulloch et al. (2008).

In this thesis, the approximation approach we emphasize is adaptive Gauss-Hermit quadrature (AGQ), since this approach gives more accurate estimates than the Laplace and PQL approaches, and it calculates the true likelihood rather than the pseudo-likelihood from PQL. AGQ is more computationally intensive than PQL and Laplace. In fact, Laplace is simply a version of Gaussian quadrature with a single quadrature point.

1.3 Introduction to tree-based methods

The purpose of the analyses via tree-based algorithm is to determine a set of *if-then* logical (split) conditions that permit accurate prediction or classification of cases. A classification tree is used when we wish to predict or explain a categorical response variable, while a regression tree is for a quantitative response variable.

The origins of classification trees date back to Morgon and Sonquist (1963) who developed AID (Automatic Interaction Detection). In the early 1970s, Morgan and Messenger (1973) developed a sequential analysis program THAID to deal with classification problems. Breiman et al. (1984) introduced the famous monograph "Classification and regression trees" (CART). For CART, the regression tree is a piecewise constant estimate of a regression function, constructed by recursively partitioning the data and sample space. Thus the estimated regression function is unsmoothed. Chaudhuri et al. (1994) developed smooth and unsmooth piecewise polynomial regression trees (SUPPORT) for continuous data. However, the Chaudhuri et al. (1994) approach did not support categorical predictors. Loh (2002) proposed the generalized unbiased interaction detection and estimation (GUIDE), which supported categorical predictors and had the ability to detect pairwise interactions and curvatures. CART performs splitting in a greedy fashion. It searches all over the potential splitting points and splitting variables, then finds the best one according to some splitting criterion. Both SUPPORT and GUIDE use signs of the residuals to separate the observations into two classes and then uses two-sample t-tests or chi-square tests for variable selection. The rationale behind GUIDE or SUPPORT is that if a fitted model is unsatisfactory, the lack of fit would be reflected in the distributional pattern of the residuals. GUIDE and SUPPORT do model fitting just once at each node and hence allow for fast tree construction. Chaudhuri et al. (1995) extended Chaudhuri et al. (1994) to support more generalized data types such as binary data, count data and survival data. Bayesian approaches have also been developed for fitting tree structured

models, e.g. Chipman et al. (1998) and Denison et al. (1998).

There are several tree-based methods for multiple responses, such as Segal (1992) for longitudinal continuous response and Zhang (1998) for multiple binary responses. De'Ath (2002) proposed multivariate regression trees for longitudinal environment data, which is available as the R package *mvpart* (De'Ath 2006). Larsen and Speckman (2004) is a variation of Segal (1992). Lee (2005, 2006) and Lee et al. (2005) proposed generalized multivariate decision trees, which used generalized estimating equations and supported general types of response variables such as binary and count. Hsiao and Shih (2007) extended the GUIDE (Loh, 2002) approach to multivariate normal responses. However, most of the above tree-based methods assume the covariates are time-invariant. Galimberti and Montanari (2002) proposed tree-based models for longitudinal continuous data with time-varying covariates. Mixed effects model have also been used to build trees for longitudinal data. Abdolell (2002) used tree models to find clusters based on a single predictor and a longitudinal continuous outcome variable. The splitting criterion was based on a likelihood ratio statistic (deviance) from the linear mixed effects model. This method has been implemented in R package *longRPart*. Sela and Simonoff (2009) proposed a RE-EM tree algorithm for continuous longitudinal data, which combined the flexibility of tree-based estimation method CART with the structure of linear mixed effects models for longitudinal data. Sela and Simonoff (2009) has been implemented in the R package *REEMtree*.

In order to build a tree, we need to answer three basic questions:

1. Which criterion should we use to split a parent node into its two children nodes? (splitting criterion)

2. How do we decide when a node become a terminal node? (stopping criterion)?

3. How to decide the right-size of a tree? (pruning)

1.3.1 CART: classification and regression tree

Breiman et al. (1984) introduced the "Classification and regression trees" (CART). Suppose $\mathbf{Y}$ is a response variable, and $\mathbf{X}$ is a set of predictors where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)$. The general structure of CART model can be described by the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. This model has two main components: a tree T with V terminal nodes, and a parameter $\Theta = (\theta_1, \theta_2, \ldots, \theta_V) \in R^V$ which associates the parameter values θ_v with the v^{th} terminal node. Thus a tree model is fully specified by the pair (T, Θ). If $\mathbf{X}$ lies in the region corresponding to the v^{th} terminal node then $\mathbf{Y}|\mathbf{X}$ has the distribution $f(y|\theta_v)$, where we use $f(\cdot)$ to represent a conditional distribution indexed by θ_v. The model is called a *regression tree* or a *classification tree* according to whether the response $\mathbf{Y}$ is quantitative or qualitative, respectively.

Splitting a tree

Non-terminal node is split into two children nodes according to some rules s. If the observations meet the rules, then they are assigned to the left child node, otherwise to the right child node. For quantitative predictors, the splitting rules are $\{X_i <= s\}$ or $\{X_i > s\}$. For qualitative predictors, the splitting rule is based on a category subset C, and assign observations for which $\{X_i \in C\}$ or $\{X_i \notin C\}$ to the left or right child node, respectively.

Assume the response Y is a factor taking outcomes $1, 2, \ldots, K$. For node τ, we define the *node impurity function* $i(\tau)$ (Izenman, 2008) as

$$i(\tau) = \phi(p(1|\tau), \ldots, p(K|\tau)) \quad .$$

where $p(k|\tau)$ is an estimate of $p(Y_i = k|\tau)$, the conditional probability of an observation Y_i in class k in terminal node τ. Usually $p(k|\tau)$ is estimated by

$$\hat{p}(k|\tau) = \frac{1}{N_\tau} \sum_{x_i \in R_\tau} I(y_i = k) \quad ,$$

where N_τ is the number of observations in node τ, R_τ is the region (covariate space) represented for node τ. For classification tree, two commonly used functions for ϕ are

$$Gini\ \ index\ \ :\ \ i(\tau) = -\sum_{k=1}^{K} p(k|\tau)\log[p(k|\tau)]\ \ .$$

$$Entropy\ \ function\ \ :\ \ i(\tau) = \sum_{k\neq k'} p(k|\tau)(1-p(k'|\tau)) = 1-\sum_{k=1}^{K} p(k|\tau)\ \ .$$

For binary outcomes, if $p = p(2|\tau)$, the proportion of the second class, then the two measures are $2p(1-p)$, and $-p\log(p)-(1-p)\log(1-p)$, respectively.

Now if τ is the terminal node, we will classify the observations in terminal node τ to class $k(\tau) = \text{argmax}_k p(k|\tau)$, the majority class in node τ. Then, an estimate of misclassification rate in node τ is defined as

$$r(\tau) = 1 - \text{argmax}_k p(k|\tau) = 1 - p(k(\tau)|\tau)\ \ .$$

For a tree T with terminal nodes $\tilde{T} = \{\tau_1, \tau_2, \ldots, \tau_V\}$, the estimated *misclassification rate* is

$$R(T) = \sum_{v=1}^{V} r(\tau_v)p(\tau_v) = \sum_{v=1}^{V} R(\tau_v)\ \ , \tag{1.4}$$

where $p(\tau_v)$ is the proportional of observations that fall into τ_v, $R(\tau_v) = r(\tau_v)p(\tau_v)$. $R(T)$ will be used for subsequently pruning.

For a regression tree, the conventional algorithm models the response in each region R_v as a constant c_v. Thus, the overall tree model can be expressed similarly as Hastie, Tibshirani and Friedman (2001)

$$f(\mathbf{X}) = \sum_{v=1}^{V} c_v I(\mathbf{X} \in R_v)\ \ ,$$

where $R_v, v = 1, 2, \ldots, V$, consists of a partition of the predictors space, and therefore represents the space of V terminal nodes. If we adopt the method of minimizing the

estimate of prediction error

$$R(T) = \frac{1}{N}\sum(Y_i - f(\mathbf{X}_i))^2 \quad ,$$

as our criterion to characterize the best split, it is easy to see that the best $\hat{c}_v$ is just the average of Y_i in region R_v:

$$\hat{c}_v = ave(Y_i|X_i \in R_v) = \frac{1}{N_v}\sum_{X_i \in R_v} Y_i \quad ,$$

where N_v is the number of observations fall into terminal node v. Thus

$$R(T) = \frac{1}{N}\sum_{v=1}^{V}\sum_{X_i \in R_v}(Y_i - \hat{c}_v)^2 = \sum_{v=1}^{V} R(\tau_v) \quad , \tag{1.5}$$

where

$$R(\tau_v) = \frac{1}{N}\sum_{X_i \in R_v}(Y_i - \hat{c}_v)^2 = p(\tau_v)s^2(\tau_v) \quad .$$

$p(\tau_v) = N_v/N$ is the proportion of observations in node v, $s^2(\tau_v) = \sum_{X_i \in R_v}(Y_i - \hat{c}_v)^2/N_v$ is the sample variance of node v. We can treat $s^2(\tau_v)$ as the *impurity function* at node τ_v.

To split a node τ into left node τ_L and right node τ_R, *goodness of split measure* for a regression tree is

$$\Delta(s, \tau) = R(\tau) - R(\tau_L) - R(\tau_R) \quad , \tag{1.6}$$

and it is the following for a classification tree

$$\Delta(s, \tau) = i(\tau) - p_L(\tau_L)i(\tau_L) - p_R(\tau_R)i(\tau_R) \ . \tag{1.7}$$

where p_L is the proportion of observations on the left child node, and p_R is the proportion of observations on the right child node. The best splitting point is the one that maximizes $\Delta(s, \tau)$:

$$\Delta(s^*, \tau) = \max_{s \in S}\Delta(s, \tau) \quad .$$

Pruning a Tree

To declare a node is a terminal node, one commonly used stopping rule is that the number of observations in a node must be less than a certain predefined number. Another rule is that if the largest goodness of split value in a node is smaller than a certain threshold. However, these two rules do not turn out to be the best ideas in practice. A better approach in Breiman et al. (1984) is to let the tree grow to overly large and then prune it back by some pruning rules which are introduced in the following.

A tree too large might overfit the data, while a small tree may not be able to capture the important structure. For CART, the tree is pruned by some *cost-complexity* criterion.

We define a subtree $T \subset T_0$ to be any tree that can be obtained by pruning T_0. Define $\tilde{T}$ to be the set of terminal nodes of T. As before, we index the terminal nodes by v, with node v representing region R_v. Let $|\tilde{T}|$ be the number of terminal nodes in T. The cost complexity criterion is defined as (Breiman et al., 1984)

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \quad . \tag{1.8}$$

where $\alpha(> 0)$ is the complexity parameter. $R(T)$ defined in (1.4) is for classification tree, and $R(T)$ defined in (1.5) is for regression tree.

For each α, find the subtree $T_\alpha \subset T_0$ to minimize $R_\alpha(T)$. α is a tuning parameter. Larger values of α result in smaller tree T_α. $\alpha = 0$ suggests the full tree T_0, and $\alpha = \infty$ for no splits at all. Cross validation (such as 5-fold or 10-fold) is typically used to find a best value of α.

Splitting criterion for Poisson data

For discrete response such as Poisson data, the data model is $\lambda = f(\mathbf{X})$ where λ is an event rate and $\mathbf{X}$ is some set of predictors. Let c_i be the observed event count for observation i,

t_i be the observation time, and $X_{ij}, j = 1, \ldots, p$ be the predictors. To split a node (parent) into two child nodes (left and right), the splitting criterion is based on deviance D and defined as

$$\Delta I = D_{\text{parent}} - (D_{\text{left child}} + D_{\text{right child}}) \quad , \tag{1.9}$$

where

$$\hat{\lambda} = \frac{\#\text{ events}}{\text{total time}} = \frac{\sum c_i}{\sum t_i} \quad , \text{ and } D = \sum c_i \log(\frac{c_i}{\hat{\lambda} t_i}) - (c_i - \hat{\lambda} t_i) \quad .$$

For each node to be split, we will find the split point and splitting variable which maximizes ΔI, similarly as continuous or binary response. More details can be found on Therneau, Atkinson and Foundation (2007) regarding the R package *rpart*.

Advantages

The advantages of tree-based methods include

- Simplicity of results: in most cases, the interpretation of results summarized in a tree is very simple. A new observation can be rapidly classified by using just a few if-then conditions. Also it often yields a much simpler "model" for explaining why observations are classified or predicted in a particular manner.

- Tree methods are nonparametric or nonlinear. Therefore there is no implicit assumptions that the underlying relationship between the response and predictors are linear, or follow some non-linear link functions, or even monotonic in nature. And in a tree, the same variables can be split multiple times which indicates a non-monotonic relationship between the response and this predictor.

- Invariant to monotone transformations: changing one or several splitting variables to its logarithm or square root, for example, will not change the structure of the tree; only the splitting values (but not variables) will be different.

- Have the ability to handle more complex interactions between predictors than a standard linear model can handle.

1.4 Binning and smoothing

Binning has been used in kernel smoothing such as density estimation by histogram (Wand and Jones, 1995), where the real line is divided into equally spaced intervals (or bins). A smaller binwidth leads to a relatively jagged histogram while a larger binwidth results in a smoother looking histogram; however, too large a binwidth will result in lost underlying structure. Histograms are also subject to the bias-variance tradeoff phenomenon as other kernel density estimation or nonparametric kernel regression methods, such that bias increases and variance decreases as binwidth increases, while the opposite occurs when the binwidth decreases.

1.5 Motivation

This thesis research is originated from a hemodialysis study where the response (health events, e.g. infection) and covariates (protein measurements, e.g., C-reactive protein) are not measured on the same time points. Details of this data are described in Chapter 2. The first goal is to find out the temporal association between health events and proteins markers, such as contemporaneous association or lagged association. In order to do this, we first need to align the response and covariates on the same time points. We have proposed *binning* method prior to implementing longitudinal modeling. GLMMs and mixed zero-inflated models have been applied on the binned data.

An additional goal is to find out whether different levels of C-reactive protein are indicative of the number of occurrences of infection events. Thus we have proposed generalized

mixed effects regression tree method which uses CART and GLMMs.

1.6 Outline of thesis

The organization of the thesis is as follows. In Chapter 2, the binning method are introduced for longitudinal data where the response and time-dependent predictor(s) are not measured on the same time points. Then the proposed binning approach are applied on a longitudinal hemodialysis study to look for possible contemporaneous and lagged effects between occurrences of a health event (i.e., infection) and levels of a protein marker of inflammation (i.e., C-reactive protein or crp). Both Poisson mixed effects models and zero-inflated Poisson (ZIP) mixed effects models are applied to the subsequent data. In addition, extensive simulations are conducted to investigate various properties of the binning approach including within-subject variability, mismatch rate, correlation structure, edge effects, and etc.

In Chapter 3, asymptotic properties of the binning estimates are derived. Leave-one-subject-out cross-validation algorithms are proposed for bin size selection.

In Chapter 4, a generalized mixed effects regression tree based method (GMRTree) is proposed. This proposed GMRTree approach is applied to a binned hemodialysis dataset and an epilepsy seizure dataset. Simulations are done to evaluate the predictive performance of GMRTree.

Chapter 5 provides a discussion of approaches proposed in the thesis, and future work is enumerated.

Chapter 2

A binning method for longitudinal data measured at distinct time points

[Part of the work in this chapter has been published in Xiong and Dubin (2010)]

2.1 Introduction

For a longitudinal study, individuals are repeatedly measured over time. Typically, one or more predictors, some possibly time-varying, and a longitudinal response of interest are recorded across the same set of time points. To analyze a longitudinal dataset, traditionally either mixed effects models (Laird and Ware, 1982; Breslow and Clayton, 1993) or marginal models using generalized estimating equations (Liang and Zeger, 1986) are most often implemented to help determine the association between the predictors, possibly including time, and the longitudinal response. For certain longitudinal problems, transition models may also be appropriate (Ware et al., 1988; Zeger and Qaqish, 1988; Cook et al., 2004). However, in the scenario that the response and longitudinal predictors within each individual are not measured at the same time points, we cannot directly apply traditional

longitudinal modeling. Instead, some form of preprocessing such as smoothing of the data is required in order to align the response and predictors to the same time points.

This work is motivated by two issues. The main motivation is based on the interests of nephrologists who are curious about the timing of certain events of interest (such as an infection) and serum protein markers of inflammation (such as C-reactive protein) for patients on hemodialysis. Do such events and rises in these proteins occur simultaneously or is there any evidence suggesting one process precedes the other?

The second motivation is based on the characteristics of a specific dataset provided to attempt to answer the above questions. That is, the events and serum proteins in a particular hemodialysis study (Kaysen et al., 2000 for discussion of the hemodialysis study and longitudinal analysis) of serum proteins occurred at different measurement frequencies, and, as a result, the health events and proteins were measured at different time points in general. We describe a binning method in Section 2.3 to allow us to answer questions of interest of the nephrologists in spite of the non-standard form of the original data. More importantly, this method could be more generally applied to problems where a longitudinal response and longitudinal predictors are mis-aligned in time and questions regarding temporal association arise.

Event history analysis such as Cook and Lawless (2007), Aalen et al. (2008) could also be applied to this type of data. However, for the problem at hand, scientifically we are not as interested in the association of certain predictors and the (instantaneous) probability of an event occurring in a window of time. Instead, we have interest in temporal ordering of association including evidence that a longitudinal predictor precedes or follows (or occurs contemporaneously) with an event of interest. For this reason, joint longitudinal and survival modeling (e.g., Wulfsohn and Tsiatis, 1997) is not of interest for the particular application under study, as we would like to consider lags of equidistant time span between the event process and longitudinal process. Furthermore, the hemodialysis data has an

event process that contains recurrent events for many subjects, further complicating any consideration of joint modeling, even if equidistant time span between lags was not of importance. Instead, we will use traditional longitudinal modeling to attempt to achieve our modeling (and scientific) goals following a temporal alignment of the response and any longitudinal predictors via binning.

The paper will be organized as follows. In Section 2.2, we will fully describe the hemodialysis study that motivated the proposed methodology. In Section 2.3, we describe the proposed binning methods to align the data, and the modeling steps based on the resulting binned data. Section 2.4 presents the results of an analysis incorporating binning from the hemodialysis study. In Section 2.5, we present a simulation study, which looks at the effect of various conditions on the success of the binning approach. Finally, we offer some concluding remarks in Section 2.6.

2.2 Description of a hemodialysis dataset

During hemodialysis, acute phase serum proteins can be measured longitudinally. These include C-reactive protein (crp), albumin (alb), transferrin (trf), α-1 acid glycoprotein (aag), and ceruloplasmin (cer). Both alb and trf belong to the class of negative acute-phase proteins, whereas crp, aag, and cer belong to the class of positive acute phase proteins. Proteins in the same class are positively correlated, while proteins in different classes are negatively correlated, a relationship established over time (Dubin and Müller, 2005).

In the hemodialysis study focused upon in this paper, 53 subjects were longitudinally measured for a period of up to 2 years. Information for these individuals included measurement time of proteins, values of the proteins, and other patient data such as gender, ethnicity, and body mass index (bmi) at study entry. The proteins were measured weekly for each patient for the first seven weeks under observation, then monthly thereafter.

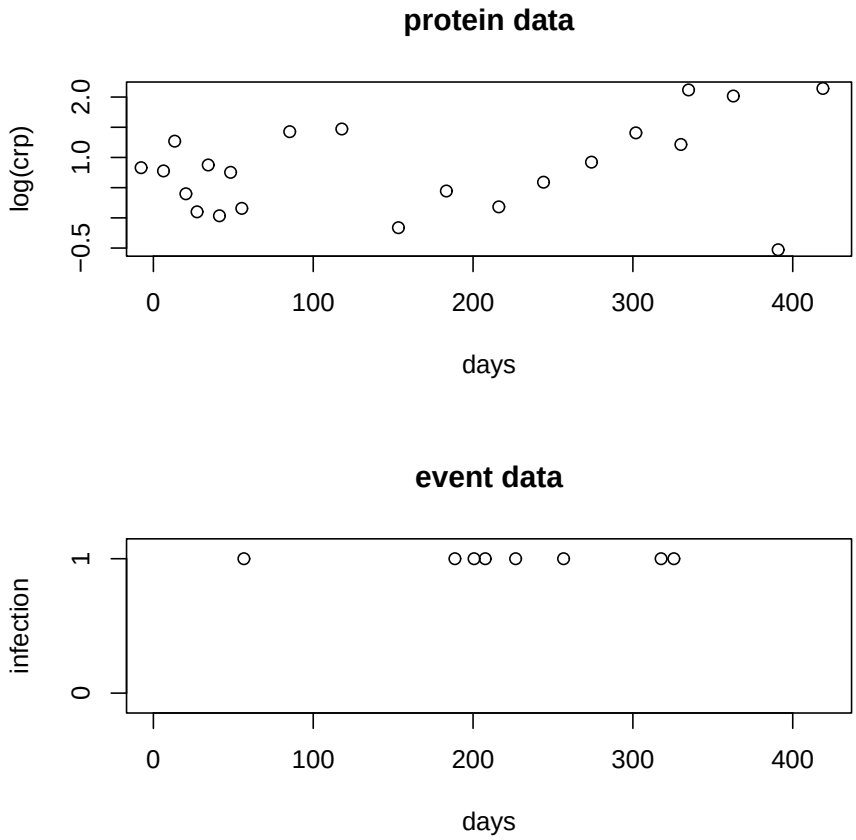

Figure 2.1: log(crp) and infection event over time for one subject

In addition to this protein data, event data, including information on cardiovascular events, access events, and infections, were also available through patient chart records measured three times a week. Good information was available on infection events (116 in total), and, hence, infection events will be the focus in the analysis in this thesis. The event information was not part of the original protein dataset and was not, in general, measured at the same time, or frequency, as the proteins.

Intermittent missing data, on either the proteins or on chart record event occurrence, was not an issue from this study sample. There were a few subjects who died during follow-

up, so not all intended data, by design, was collected. There was no loss to follow-up for reasons other than death.

To provide an illustration of the type of data we will be analyzing, we show a plot of the main event of interest, infection, and the main protein of interest, crp, for one of the patients in Figure 2.1. As the distribution of crp is skewed, the log transformation of crp is taken to make the distribution closer to symmetric. Specifically, Figure 2.1 shows log(crp) and infection events, respectively, for this patient followed for approximately 420 days. In general, the proteins and events were measured at different time points.

There is another key feature of the infection event data that may not be immediately apparent in Figure 2.1. Taking, for example, day 0 to day 50, there are no events displayed on the plot in the lower panel, which simply represents no infection events occurring during this period. Note that this is how the data were recorded, i.e., only events (and the days on which they occurred) were entered in the original event dataset, whereas days where no events were observed were never entered in the dataset, even though event data were obtained three times per week for each subject. Also, it seems that for at least this patient, after an infection event occurred, there was generally some increase in log(crp) levels, which indicated that there could be a positive lagged association between log(crp) and infection events. We have proposed a method in the next section to allow us to see if indeed this result holds on average, across all patients.

2.3 Methodology

2.3.1 Binning method

For each subject, we implement a smoothing step by defining equally spaced bins of time for the longitudinal response and predictor. Within each bin, we average the covariate and

this will be used as new predictor for the remainder of the analysis. In addition, we sum up the events within each bin, and this will be used as the new response. The notation is defined as follows:

- Let Y_{ij} be the number of health events (e.g., infection) observed for subject i at time j, where $j = 1, 2, \ldots, n_i^{(\mathbf{Y})}$. In continuous time, we assume Y_{ij} is a binary variable, but in discrete time (e.g., on a given day), Y_{ij} may be a count variable.

- Let X_{ik} be a continuous time-varying predictor (e.g., log(crp)) for subject i observed at time k for $k = 1, 2, \ldots, n_i^{(\mathbf{X})}$, where, in general, the times represented by k do not equal the times represented by j and $n_i^{(\mathbf{X})} \neq n_i^{(\mathbf{Y})}$.

- We then bin $(\mathbf{X}, \mathbf{Y})$ in equidistant units of time.

- Next we take the unweighted or weighted average (or sum) of variables within each bin, for each individual.

- The resulting data will be $(X_{i,m}, Y_{i,m})$, where $m = 1, 2, , \ldots, n_i^{(\mathbf{X},\mathbf{Y})}$. There are a total of $n_i^{(\mathbf{X},\mathbf{Y})}$ bins for individual i.

In the above, we could bin multiple (say p) longitudinal predictors of interest, creating $\mathbf{X}_{i,m} = (X_{1,i,m}, X_{2,i,m}, ..., X_{p,i,m})$. In this paper, for the data analysis in Section 2.4, we will use the unweighted bin average of one protein, focusing on crp, and the unweighted bin sum of one event, focusing on infection events. In this new dataset after binning, each subject has only one entry for each bin, i.e., $(X_{i,m}, Y_{i,m})$, and each subject has several repeated measurements (i.e., bins observed over time). This bears modest relation to histogram smoothing, since the bins are not moving along with time once the bin size is fixed. Each original data point is used only a single time, i.e., in the defined bin in which it resides.

Instead of equi-distant bins, variable-length binning could also be considered, though this would not easily allow for consideration of a lagged association between binned response

and predictor. We could also consider a weighted average within each bin, where the weights could be based on, say, the number of original observations within a bin or based on a measure of variability, and etc. In our analysis and simulation in this Chapter, we will focus on equal bin width and unweighted average within bins, but these considerations could be an area of future research.

Through this binning method, the longitudinal predictor and response become aligned in time, and, hence, standard longitudinal modeling can be applied. Also, we can potentially predict the expected total number of events occurring in a period of time using the average predictor level during this time. For a lagged model, which will be further discussed in Section 2.3.4, we could use the average current predictor levels to predict the total number of events in a period of time in the future or provide guidance on event occurrences in the past, assuming a lagged association can even be identified.

Consideration of different bin sizes will be application and study specific. We will discuss our method of bin size selection for the hemodialysis dataset in Section 2.4. Automatic bin size selection choices could be used, e.g., using cross-validation, or pre-defined bin sizes may be appropriate, assuming these are sensible for the data application and bounded below in size by the data collection procedure. We mention one possible bin choice selection approach in the Discussion. It will also be important to consider sensitivity of bin size choice on any resulting model fit.

2.3.2 Generalized linear mixed effect model

For non-normal longitudinal data, generalized linear mixed effects models (Breslow and Clayton, 1993) are often applied which extend the generalized linear model by inclusion of random effects into the model. Marginal models using generalized estimating equation (Liang and Zeger, 1986) are also a possibility. But we will here emphasize a mixed effects modeling approach, which will allow for subject-specific interpretation of association

parameter estimates.

Recall that, after binning, the response is the sum of the events within each bin, which is now count data. For a given bin size, conditional on random effects, we assume for each subject, the event counts for each bin is independent and distributed as Poisson. The reason the Poisson distribution is used is that, theoretically, we want to allow for the fact that greater than one event can occur on any day, so we cannot impose a cap on the number of events for a given bin size; if such a cap were known, we could use a Binomial distribution assumption instead.

Given the above Poisson assumption, a generalized linear mixed effects Poisson regression model could be initially used to model this data with a standard log link function and a single normally distributed random effect, i.e., a random intercept. To be more specific, we consider the following GLMM

$$\begin{aligned} Y_{i,m}|u_i &\sim Poisson(\lambda_{i,m}) \\ \log(\lambda_{i,m}) &= \beta_0 + \beta_1 * X_{i,m} + u_i \\ u_i &\sim N(0, \sigma_u^2) \quad \text{where i } = 1, \ldots, n; \quad m = 1, \ldots, n_i^{(\mathbf{X},\mathbf{Y})} \quad . \end{aligned}$$

Here, response $Y_{i,m}$ is the sum of events, and $X_{i,m}$ is the averaged continuous predictor, respectively, within the m^{th} bin for individual i. The random u_i terms are assumed independent between subjects. Note that we are using the time notation m, as described in Section 2.3.1, to emphasize the change from original time scale to the binned time scale, and each subject i then has a total of $n_i^{(\mathbf{X},\mathbf{Y})}$ (binned) repeated measurements. $n_i^{(\mathbf{X},\mathbf{Y})}$ is treated as fixed variable here.

2.3.3 Mixed zero-inflated Poisson model with normal random effects

Count data with excess zeros relative to a Poisson distribution are commonly encountered in biomedical and public health applications (and appears to be a feature of our example hemodialysis event data, with at least 85% of bins with 0 counts with the bins sizes considered in our analysis in Section 2.4). One popular model to account for excess zeros is the zero-inflated Poisson (ZIP) model (Mullahy, 1986; Lambert, 1992). Applications for the ZIP model can be found in Böhning (1998), Ridout et al. (1998), Böhning et al. (1999), Dietz and Böhning (2000), and Lee et al. (2001). Test for zero-inflation or overdispersion of ZIP models can be found in Van den Broek (1995), Deng and Paul (2000), Jansakul and Hinde (2002), Lee et al. (2004), and Ugarte et al. (2004). Hall (2000) has extended the ZIP model to incorporate random effects. See also Yau and Lee (2001), Wang et al. (2002), Xiang et al. (2006) and Xiang et al. (2007) for mixed ZIP models.

The standard ZIP model is a mixture of a Poisson component and a degenerate component at point mass 0. Using general notation, let Y be a count response variable, then

$$Y \sim \begin{cases} 0 & \text{with probability } p \\ \text{Poisson}(\lambda) & \text{with probability } (1-p) \ . \end{cases}$$

That is,

$$P(Y=y) = \begin{cases} p + (1-p)e^{-\lambda} & , \quad y = 0 \\ (1-p)\frac{e^{-\lambda}\lambda^y}{y!} & , \quad y > 0 \ , \end{cases}$$

meaning zeros are not only coming from the Poisson distribution.

In the longitudinal settings with covariates and random effects, usually we assume p follows a logistic model and λ follows a log-linear model. With random effects $\mathbf{v}_i$ and $\mathbf{u}_i$,

the mixed ZIP model for longitudinal data is as follows:

$$\begin{aligned} \text{logit}(p_{ij}) &= \mathbf{W}_{ij}^{'}\boldsymbol{\alpha} + \mathbf{Z}_{1ij}^{'}\mathbf{v}_i \quad (2.1) \\ \log(\lambda_{ij}) &= \mathbf{X}_{ij}^{'}\boldsymbol{\beta} + \mathbf{Z}_{2ij}^{'}\mathbf{u}_i \ , \quad (2.2) \end{aligned}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$,

$\mathbf{W}_{ij}$ and $\mathbf{X}_{ij}$ are the covariate vectors for the fixed effects $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ respectively, where some covariates may overlap between $\mathbf{W_{ij}}$ and $\mathbf{X}_{ij}$. $\mathbf{Z}_{1ij}$ and $\mathbf{Z}_{2ij}$ are the covariate vectors for random effects $\mathbf{v}_i$ and $\mathbf{u}_i$ respectively. The random effects $\mathbf{u}_i$ and $\mathbf{v}_i$ are assumed jointly normal and possibly correlated,

$$\begin{pmatrix} \mathbf{v}_i \\ \mathbf{u}_i \end{pmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^{'} & \Sigma_{22} \end{bmatrix} \right) \ ,$$

where Σ_{11}, Σ_{12} and Σ_{22} are unknown positive-definite matrices. In practice, the simple random intercept form of models is often adequate, in which $\mathbf{v}_i = v_i$ and $\mathbf{u}_i = u_i$ are univariate and $\mathbf{Z}_{1ij} = \mathbf{Z}_{2ij} = 1$ (see Min and Agresti, 2005).

2.3.4 Contemporaneous and lagged models

A longitudinal response may change simultaneously with the longitudinal predictor of interest. In this case, we use a standard contemporaneous approach where $Y_{i,m}$ is fitted against $\mathbf{X}_{i,m}$.

It is also possible that the longitudinal predictor changes prior to an individual experiencing the longitudinal response. Of course, such a lagged relationship may occur in the opposite direction, whereby a response occurs prior to a change in the predictor. In model notation, we will be fitting $Y_{i,m}$ versus $\mathbf{X}_{i,m-q}$ as well as $Y_{i,m-q}$ versus $\mathbf{X}_{i,m}$, the former suggesting the predictor leads the response and the latter suggesting the response leads

the predictor. Here q is the lag order. Consideration of lags using the binned methodology from Section 2.3.1 is quite straightforward, as the bins are of equal size across the observational time period of each subject.

We discuss these models further, with regard to the hemodialysis example, in Section 2.4.

2.4 Analysis of the hemodialysis data

2.4.1 Analysis setup

For the hemodialysis dataset, we will be fitting both the generalized linear mixed effects models and mixed ZIP models presented in Sections 2.3.2 and 2.3.3. For bin size selection, we used bin sizes that made sense given the data collection. Specifically, bin sizes included 30 days, and 45 days, noting that after the first seven weeks of follow up, protein (crp) measurements were obtained every 30 days. However, there is a chance that an infection and crp level may be associated within a period of time closer than 30 days, especially if we want to consider a lagged association. So a 7-day bin size is also used, but only on the data from the first seven weeks of the study, as such a small bin size would not be possible following the first seven weeks of follow-up, unless we further interpolated the data, say using a curve-based approach. Here, we will consider two datasets for analysis:

1. the full dataset including all longitudinal observations, with a minimum bin size of 30 days, and

2. a subset of the full dataset, where only the first seven weeks of follow-up for each patient is considered, with a minimum bin size of 7 days.

The lagged investigation is of key interest to the principle investigator of the nephrology study. Specifically, lagged models are used to check whether the infection events appear to occur before or after a rise in the crp levels, as this could help answer an important bidirectional question about which process possibly precedes the other, if any. Due to the large bin sizes in the full dataset, it is not biologically meaningful to consider lags (i.e., lag orders q) greater than 1 (i.e., 30 days). In addition, for the subset dataset with just seven weeks of follow-up, we do not want to throw away too much data. So again, we will maintain a max lag consideration of one bin (i.e., 7 days). Lags of size greater than one could very well be considered for other datasets, as appropriate.

For each of the two datasets (full and 7-week data) discussed just above, three Poisson GLMM and three mixed ZIP models are fit for binned data at a given bin size and direction. In each case (i.e., either Poisson GLMM or mixed ZIP), one of these three models is for contemporaneous direction (i.e., 0 lag), the other two are for lagged directions. To accommodate these three models in the mixed ZIP modeling, we adapt the logistic setting of (2.1) and log-linear setting of (2.2), as necessary:

$$\text{Contemporaneous model:} \quad \begin{cases} \text{logit}(p_{i,m}) &= \alpha_0 + v_i \\ \log(\lambda_{i,m}) &= \beta_0 + \beta_1 * X_{i,m} + u_i \end{cases}$$

$$\text{Lag -1 model:} \quad \begin{cases} \text{logit}(p_{i,m}) &= \alpha_0 + v_i \\ \log(\lambda_{i,m}) &= \beta_0 + \beta_1 * X_{i,m-1} + u_i \end{cases}$$

$$\text{Lag +1 model:} \quad \begin{cases} \text{logit}(p_{i,m-1}) &= \alpha_0 + v_i \\ \log(\lambda_{i,m-1}) &= \beta_0 + \beta_1 * X_{i,m} + u_i \end{cases}$$

The response $Y_{i,m}$ is the number of infection events within each bin. The predictor $X_{i,m}$ is only log(crp). We are not considering other predictors in this particular analysis, including in the logistic part of the model, as our main interest is in investigating an association between infection event and crp as well as the direction of such association. Further work could be considered to produce a fuller explanatory model for the longitudinal infections.

Normal random effects are included in both the logit model and log linear model.

We assume the covariance matrix between u_i and v_i is

$$\begin{bmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix}$$

Our modeling has shown that ρ is either not statistically different from 0 or including a ρ different than 0 does not improve the AIC for a given lag and bin size. So we will treat ρ as 0 in the following analysis.

In our analyses, we implemented the binned data longitudinal modeling using PROC NLMIXED in SAS (Version 9.1.3). Also, to account for zero-inflation, the correlated longitudinal data, and possible further overdispersion, we have also fit mixed zero-inflated negative binomial (ZINB) models (Yau et al., 2003; Xiang et al. 2007) to the longitudinal hemodialysis infection data. However, for almost all such model fits, the overdispersion α parameter of the negative binomial approach was either not significant or the AIC for a fixed bin size (when compared to the less complex mixed ZIP model) was not smaller. In addition, we found computational difficulties with a few mixed ZINB fits, in that the NLMIXED procedure did not always converge for the mixed ZINB models, whereas this problem did not arise for comparable mixed ZIP modeling for this data. Hence, for these issues, we will not further discuss the mixed ZINB model, though it may be useful for other data of this type.

2.4.2 Analysis results

Table 2.1 shows the analysis results of the Poisson GLMMs. The analysis results for mixed ZIP models are shown on Table 2.2. A single random intercept is used for the Poisson GLMMs, whereas, in the mixed ZIP models discussed in Section 2.3.3, we used a random intercept in both the logistic model of p and the log-linear model of λ. We have tried more

complicated models including adding additional random effect terms, but these did not improve upon the models fit with the random intercepts only.

In both tables, model "0" stands for contemporaneous model, model "-1" for "lag -1" model, and model "+1" for "lag +1" model. Bin sizes 30 and 45 days are applied to the full dataset, while bin sizes 7 is applied to the 7-week data, where each dataset was described in Section 2.3.1. We also had run a bin size 10-day model with the 7-week dataset, but the results of this model within the mixed ZIP modeling framework were not terribly stable for the logistic part of the model. Also, the log-linear results for these 10-day models were consistent with that of the 7-day bin size models, both in Table 1 (mixed Poisson) and Table 2 (mixed ZIP), so we did not include these results in the displayed output.

2.4.3 Contemporaneous analysis focus

Based on the contemporaneous fit (model "0") of the Poisson GLMMs in Table 2.1, and looking at the estimated β_1 values, we can see that log(crp) is significant for all bin sizes (30 and 45 day bins for the full dataset, and 7-day bin for the 7-week data). This suggests that the occurrence of infection is simultaneously associated with log(crp) levels. The positive sign of $\hat{\beta}_1$ means that the estimated association between infection and log(crp) is positive. Larger log(crp) values in a bin are associated with the occurrence of infection, and smaller log(crp) values are associated with lack of events. Again, this contemporaneous result is consistent, as it is seen across various bin sizes, in both sets of analyzed data.

The mixed ZIP model contemporaneous analysis results are shown in Table 2.2 model "0". By looking at the β_1 estimates, we can see the significance of log(crp) in the log-linear model for bin sizes 30 and 45, which suggests that there is strong positive contemporaneous association between infection and log(crp) for the full data. The significance of log(crp) is not quite as consistent for the smaller bin sizes, however, in the contemporaneous mixed ZIP models. The p-value for log(crp) for the 7-day binned data is .1282, not statistically

Table 2.1: Generalized linear mixed effects model (Poisson GLMM) for infection versus log(crp)

bin size	model	estimate $\hat{\beta}_1$(S.E.)	p-value	Random effects $\hat{\sigma}_u$(S.E)	AIC
30	0	0.5885(0.08819)	<0.0001	1.1328(0.1955)	630.3
	-1	0.06858(0.1072)	0.5251	1.0656(0.1905)	629.8
	+1	0.07141(0.1116)	0.5250	1.0133(0.2149)	569.6
45	0	0.5469(0.09149)	<0.001	1.1147(0.1948)	579.8
	-1	0.1347 (0.1099)	0.2257	1.0666(0.1908)	566.8
	+1	0.08745(0.1195)	0.467	1.0756(0.2233)	518.5
7	0	0.5413(0.2493)	0.0345	1.0594(0.5273)	106.4
	-1	0.3680(0.2989)	0.2240	1.2116(0.563)	91.4
	+1	1.2430(0.3556)	0.0010	1.3314(0.6957)	80.4

significant at the traditional .05 level. Still, the association estimate is .4899 which is consistent with its corresponding Poisson GLMM analysis (where $\hat{\beta}_1$=.5413, with a p-value of .0345). But the standard error for log(crp) is much higher in the mixed ZIP model.

For the contemporaneous analyses, the AIC from any mixed ZIP model is always smaller than that of its comparable Poisson GLMM (i.e., same "0" model, dataset, and bin size), regardless of bin size 30, 45 and 7. The lower AIC of comparable mixed ZIP models means we should focus on the mixed ZIP model results for the contemporaneous analysis, though the positive association between log(crp) and event occurrence within the same bin is seen across all models investigated (except, as mentioned above, there is lack of traditional statistical significance from the 7-day binned data mixed ZIP model association).

We also investigated if there was an interaction between time (via bin number) and log(crp), but this interaction was never significant, in any of the Poisson GLMM or mixed ZIP models we had fit (for the contemporaneous models, or the lagged models fit in Section 2.4.4).

Table 2.2: Mixed ZIP model for infection versus log(crp); note "-" for $\hat{\sigma}_v$ entries means its associated random effect did not make the final model.

bin	model	logistic			log-linear			AIC
		$\hat{\alpha}_0$(S.E.)	p-value	$\hat{\sigma}_v$(S.E.)	$\hat{\beta}_1$(S.E.)	p-value	$\hat{\sigma}_u$(S.E.)	
30	0	0.8555(0.4217)	0.0477	1.071(0.4656)	0.5333(0.1018)	<0.0001	0.7144(0.2310)	602.5
	-1	1.2048(0.4468)	0.0095	0.7741(0.415)	0.2120(0.1590)	0.1882	0.7171(0.283)	588.2
	+1	1.6405(0.4401)	0.0001	1.1882(0.3356)	0.1647(0.1082)	0.1342	0.039(3.1453)	538.2
45	0	1.1312(0.3393)	0.0016	1.0256(0.3508)	0.4413(0.0936)	<0.0001	0.4959(0.1967)	541.1
	-1	1.1018(0.3483)	0.0026	0.8133(0.3756)	0.3971(0.1639)	0.0196	0.7177(0.2368)	513.7
	+1	1.6136(0.3326)	<0.0001	1.1192(0.3353)	0.1988(0.1133)	0.0853	0.1736(0.5000)	473.6
7	0	1.6916(0.9343)	0.0761	-	0.4899(0.3168)	0.1282	0.9723(0.7394)	101.9
	-1	1.6253(1.2188)	0.1884	-	0.3500(0.3438)	0.3136	0.932(0.8173)	89.7
	+1	0.0947(1.4565)	0.9484	-	1.1148(0.3641)	0.0035	1.0025(0.7251)	81.3

2.4.4 Lagged analysis focus

Based on the lagged models' fits (model "+1" and model "-1") in Tables 2.1 and 2.2, log(crp) is significant for bin size 7, but only for lagged "+1" model (in both the Poisson GLMM and mixed ZIP models). This is an important biological finding, i.e., that is the occurrence of infection events appears to occur prior to an increase in log(crp). In addition, this result is not seen in the full dataset with the larger bin sizes, suggesting the need to focus on smaller bin sizes to capture the dynamics of the ordering of the association between infection event and log(crp) levels, at least with the hemodialysis patients in the application study.

In an outlying result, there is also significance of log(crp) for bin size 45 "-1" mixed ZIP model. Given its unusual result compared to all other "-1" model results, we suspect this is an aberrant finding that we cannot trust as meaningful. To further emphasize this point, we also ran a 60-day binned mixed ZIP model for lagged log(crp), and the association between lagged log(crp) and infection events was not significant (p-value of .3836). It is not biologically plausible that there would be a positive association between lagged log(crp) and infection events with 45-day bins but not at either 30-day or 60-day bins.

For the full data, for bin size 30 or 45, lagged or contemporaneous analysis, the AIC of mixed ZIP model is always much smaller than that of the Poisson GLMM model. However, the results for the 7-week binned data are not so clear in terms of using AIC criterion to choose between the Poisson GLMM and mixed ZIP model for comparable models (i.e., same bin size and lag order). For 7-day "+1" lagged models, which show the important association between infection events and log(crp) levels, the Poisson GLMM is better (lower AIC) for the 7-day. The opposite is true for the 7-day "-1" lagged models and contemporaneous cases, however, non-significance is detected for log(crp). This set of results suggests the possible need for a more automated approach of selecting models, something we briefly cover in the Discussion.

2.4.5 Analysis summary

In summary, for the full data, there is a strong positive contemporaneous association between infection events and log(crp) levels. For the 7-week subset of data, the existence of the contemporaneous association appears to exist as well, though there is one exception on the 7-day binned data mixed ZIP model. However, for the subset data, there is an important lagged "+1" association across all models, which conveys a biologically meaningful relationship between infection event and crp levels, i.e., infection events appear to occur prior to increases in crp in this dataset. The general non-significant results for the "-1" model, with the exception of one of the 45-day binned models, suggests it is unlikely the opposite temporal ordering exists, i.e., that rises in crp occur prior to infection events.

2.5 Simulation

We present a simulation study to evaluate various properties of binning. The goal of this study is to demonstrate longitudinal data scenarios under which binning should be successful and other scenarios in which binning might not work well. We will focus on a single longitudinal count response and a single time-varying continuous predictor. There are several considerations for this simulation study:

- within-subject variability of the response and time-varying predictor, as well as possible serial correlation structure
- mismatch structure between the days that the response and time-varying predictor are observed
- bin size
- number of observations within a fixed bin size.

2.5.1 Simulation setup

We initially assume that the count response and the continuous predictor are always observed on the same days, with the predictor X_{ij} following a linear mixed effects model, and the response Y_{ij} following a conditional Poisson GLMM distribution. The setup is:

$$\begin{aligned} X_{ij} &= \alpha_0 + v_i + e_{ij} \\ Y_{ij}|u_i &\sim \text{Poisson}(\lambda_{ij}) \\ \log(\lambda_{ij}) &= \beta_0 + \beta_1 * X_{ij} + u_i \ , \end{aligned}$$

where $\alpha_0 = -0.45$, $v_i \sim N(0, \sigma_v^2)$, $\sigma_v = 0.8$, $e_{ij} \sim N(0, \sigma_e^2)$, $\sigma_e = 0.6$, $\beta_0 = -2.0$, $\beta_1 = 0.5$, $\sigma_u = 0.8$, $u_i \sim N(0, \sigma_u^2)$. These parameter choices are based on the hemodialysis application from Section 2.2. In particular, α_0, σ_v and σ_e are chosen to be close to the estimates when we performed a linear mixed effects model on log(crp). β_0, β_1 and σ_u are chosen to be close to what we found with the 30-day infection data. Note that we are initially trying to keep the conditional Poisson rate fairly low, so that the generated data is consistent with the hemodialysis application. We will investigate a few deviations from this above scenario.

We assume N=50 subjects and each subject's follow-up time is 40 weeks. When we impose serial correlation, we provide the following assumption, simply an AR(1) structure, through the specification of the predictor X_{ij}:

$$e_{ij} = \rho e_{ij-1} + w_{ij} \ ,$$

where $|\rho| < 1$. If we set the initial error $e_{i0} \sim N(0, \sigma_e^2)$, and $w_{ij} \sim N(0, \sigma_e^2(1-\rho^2))$, then it can be shown (Fitzmaurice et al., 2004) that for such a process

$$\text{Var}(e_{ij}) = \sigma_e^2, \quad \text{Cov}(e_{ij}, e_{ik}) = \sigma_e^2 \rho^{|j-k|} \ .$$

2.5.2 Additional simulated data details

Again, to be close to consistent with the application, we assume that for the first 8 weeks, we see 4 days of observed data each 7-day week. These are days 1, 2, 3, and 6. After the first 8 weeks, days 1, 3 and 6 are still observed weekly, but day 2 is observed only every four weeks. If we treat day 2 as the day when crp is measured, this means crp is measured weekly for the first 8 weeks, then once every four weeks thereafter. Using the same notion, days 1, 3 and 6 are the days for health events which are measured three times per week throughout the follow-up of each subject. There will be variants of this general data structure, as discussed just below when we talk about various forms of matching (and mismatching) of the days when the response and predictor are observed.

In our modeling of the simulated data, note that we use the first 8-week data for analysis as well as the full 40-week data, similar to the hemodialysis data application (when we had (i) a 7-week analysis and (ii) a second analysis that included the entire follow-up period of each subject).

In order to evaluate the effect of binning, we simulate 5 different types of datasets. The first one is complete matched data and the other four are some form of mismatched data. For complete matched, both predictor X_{ij} and response Y_{ij} are available on each day that is observed. This is an ideal scenario within this simulation setup, used as a gold-standard, and not one that we observed with the application data. Specifically, in each week of the first eight weeks, there are four days that both X_{ij} and Y_{ij} are observed. After the first 8 weeks, either three or four days are observed where both X_{ij} and Y_{ij} are obtained each week; the four-day observation weeks occur only once every four weeks. For mismatched data, some of the X_{ij} or Y_{ij} are not available. Specifically, the 5 different simulated datasets fall under the 5 following scenarios:

1. complete matched data: both X_{ij} and Y_{ij} are available for each day that either one

is observed.

2. data I: the response initially matched with one of the X_{ij} in a given week is not available, such as day 2.

3. data II: the predictor initially matched with one of the Y_{ij} in a given week is not available, such as day 1.

4. data III: the predictor initially matched with two of the Y_{ij} in a given week is not available, such as both days 1 and 6.

5. data IV: the X_{ij} and Y_{ij} are totally mismatched, such as day 2 having no response, and days 1, 3 and 6 having no predictors. In this scenario, the predictor and response are never observed on the same days.

The number of mismatches is 0 in the complete data case and increases as data goes from I to IV, with data IV being the extreme case where there is complete mismatching. We generated 500 datasets under each scenario.

2.5.3 Simulated results

Results of the simulation study described above can be seen in Tables 2.3 through 2.5, which reflect an analysis on the 8-week data only (in Table 2.3), and the entire 40-week follow-up period (in Tables 2.4 and 2.5), respectively. In each of these three tables, it can be seen that for complete matched data, the estimated β_1 is close to 0.5 (i.e., the original β_1 before binning). Not surprisingly, this binning method works well when the data is all matched.

For mismatched data I, II, III, the estimated β_1 may be close to those of matched data, especially with low within-subject variability and/or high serial correlations (e.g., $\rho = 0.8$

Table 2.3: Estimates for matched data, and data I, II, III and IV for the first 8 weeks, with bin sizes 7, 14 and 28 days. Before binning, $\beta_0 = -2, \beta_1 = 0.5$, and $\sigma_u = 0.8$. After binning, the target β_1 value remains at 0.5 and σ_u remains at 0.8, whereas the target β_0 value will be a function of bin size based on an offset, increasing as bin size increases.

Data	binsize	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$	$\hat{\sigma}_u(S.E.)$
complete matched	7	-0.577(0.151)	0.501(0.140)	0.771(0.134)
	14	0.119(0.152)	0.495(0.158)	0.769(0.134)
	28	0.814(0.154)	0.494(0.171)	0.767(0.134)
I	7	-0.869(0.158)	0.499(0.150)	0.767(0.142)
	14	-0.172(0.160)	0.494(0.171)	0.765(0.142)
	28	0.525(0.161)	0.496(0.183)	0.764(0.142)
II	7	-0.602(0.152)	0.443(0.133)	0.772(0.135)
	14	0.103(0.152)	0.459(0.152)	0.770(0.134)
	28	0.806(0.154)	0.476(0.167)	0.768(0.135)
III	7	-0.639(0.153)	0.356(0.117)	0.779(0.136)
	14	0.078(0.153)	0.401(0.142)	0.773(0.135)
	28	0.794(0.154)	0.447(0.161)	0.770(0.133)
IV	7	-1.023(0.162)	0.156(0.104)	0.815(0.151)
	14	-0.282(0.161)	0.250(0.134)	0.793(0.147)
	28	0.464(0.161)	0.360(0.158)	0.777(0.143)

Table 2.4: Estimates for matched data, and data I, II, III and IV for 40 weeks with different bin sizes and different within-subject variabilities. Before binning, β_0 = -2 and β_1 = 0.5. After binning, the target β_1 value remains at 0.5, whereas the target β_0 value will be a function of bin size based on an offset, increasing as bin size increases.

		$\sigma_e = 0.6$		$\sigma_e = 1.5$	
Data	binsize	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$
complete matched	28	0.654(0.132)	0.498(0.126)	0.870(0.125)	0.497(0.065)
	40	1.012(0.132)	0.499(0.133)	1.233(0.125)	0.499(0.075)
	56	1.347(0.134)	0.495(0.143)	1.572(0.129)	0.494(0.093)
I	28	0.528(0.133)	0.497(0.129)	0.744(0.127)	0.497(0.070)
	40	0.887(0.133)	0.498(0.135)	1.108(0.128)	0.499(0.079)
	56	1.224(0.135)	0.495(0.144)	1.447(0.131)	0.494(0.097)
II	28	0.622(0.130)	0.430(0.117)	0.813(0.124)	0.369(0.060)
	40	0.987(0.129)	0.446(0.124)	1.179(0.124)	0.380(0.066)
	56	1.328(0.132)	0.454(0.135)	1.521(0.126)	0.385(0.082)
III	28	0.569(0.127)	0.320(0.103)	0.745(0.125)	0.223(0.049)
	40	0.938(0.127)	0.344(0.112)	1.108(0.125)	0.231(0.057)
	56	1.291(0.129)	0.376(0.127)	1.454(0.125)	0.247(0.072)
IV	28	0.315(0.134)	0.062(0.062)	0.518(0.135)	0.008(0.026)
	40	0.682(0.134)	0.081(0.072)	0.877(0.136)	0.013(0.030)
	56	1.038(0.132)	0.120(0.088)	1.216(0.134)	0.017(0.038)

chosen in one of the simulation scenarios in Table 2.5). The totally mismatched case IV has the largest bias for β_1. Only in special cases such as a high autocorrelation and/or very low levels of within-subject variability, will we see possibly acceptable levels of association bias toward the null.

In Table 2.3, for bin size 28, and under data scenario IV, the estimated β_1 is 0.360 with standard error 0.158. Although this estimate is biased downward from 0.5, it is still statistically significant showing that the response and predictors are associated. Though bias may be the result under high mismatching, it will not necessarily remove a detected signal. It may just provide a (possibly very) conservative estimate of association.

As bin size increases, the estimated β_0 increases since the number of observations within a bin increases. More importantly, the estimated β_1 increases toward its target value, i.e., bias decreases, as bin sizes increase for mismatched data. However, there is a bias-variance tradeoff working here, noting the standard deviation of β_1 increases with increasing bin size, which is likely due to the number of bins decreasing, i.e., the resulting number of repeated measurements is decreasing.

From Table 2.3, we can see that the standard error of $\hat{\beta}_1$ for data I is larger than that of complete matched data. This may due to that there are fewer responses within each bin for data I than that of complete matched data, but they have the same number of covariates values within each bin. As data goes from data II to data IV, there are fewer number of covariates values, but the response values are the same as those of the completed matched data. As data goes I to IV, the bias of $\hat{\beta}_1$ increases but the standard error of $\hat{\beta}_1$ decreases. The bias effect of $\hat{\beta}_1$ is much more dramatic than the standard error. The mean square error (MSE) increases and the coefficient of variation increases, each as mismatching increases.

For complete matched data and data I to IV, the days where the misaligned cases occurred are fairly fixed and the same for all the subjects. To make the misaligned cases

Table 2.5: Estimates for matched data, and data I, II, III and IV for 40-week data with different positive serial correlation values in X_{ij}. Before binning, β_0 = -2 and β_1 = 0.5. After binning, the target β_1 value remains at 0.5, whereas the target β_0 value will be a function of bin size based on an offset, increasing as bin size increases.

		$\rho = 0.4$		$\rho = 0.8$	
Data	binsize	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$
complete matched	28	0.648(0.125)	0.5006(0.095)	0.634(0.122)	0.499(0.070)
	40	1.005(0.127)	0.4964(0.107)	0.994(0.122)	0.496(0.079)
	56	1.346(0.126)	0.5024(0.112)	1.336(0.122)	0.500(0.086)
I	28	0.525(0.127)	0.5014(0.099)	0.510(0.122)	0.501(0.072)
	40	0.882(0.128)	0.4982(0.110)	0.871(0.123)	0.498(0.081)
	56	1.223(0.128)	0.5045(0.114)	1.213(0.123)	0.503(0.088)
II	28	0.633(0.124)	0.4639(0.090)	0.631(0.121)	0.491(0.069)
	40	0.994(0.125)	0.4690(0.103)	0.992(0.122)	0.491(0.078)
	56	1.335(0.125)	0.4770(0.109)	1.334(0.121)	0.495(0.086)
III	28	0.604(0.123)	0.3940(0.084)	0.625(0.121)	0.476(0.069)
	40	0.965(0.124)	0.3999(0.098)	0.984(0.121)	0.473(0.077)
	56	1.314(0.125)	0.4254(0.106)	1.330(0.122)	0.484(0.086)
IV	28	0.359(0.130)	0.1133(0.061)	0.415(0.125)	0.259(0.060)
	40	0.721(0.128)	0.1242(0.070)	0.783(0.124)	0.282(0.067)
	56	1.077(0.128)	0.1658(0.083)	1.139(0.124)	0.322(0.078)

Table 2.6: Estimates for different bin sizes and various mismatch rates, 8 weeks data. Before binning, β_0 = -2 and β_1 = 0.5. After binning, the target β_1 value remains at 0.5, whereas the target β_0 value will be a function of bin size based on an offset, increasing as bin size increases.

Mismatch	bin size = 7		bin size = 14		bin size = 28	
(%)	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$
0	-0.577(0.151)	0.501(0.140)	0.119(0.152)	0.495(0.158)	0.814(0.154)	0.494(0.170)
10	-0.610(0.152)	0.483(0.137)	0.090(0.154)	0.487(0.158)	0.787(0.154)	0.490(0.170)
20	-0.648(0.154)	0.455(0.134)	0.058(0.154)	0.473(0.155)	0.759(0.155)	0.484(0.171)
40	-0.731(0.154)	0.390(0.130)	-0.013(0.154)	0.439(0.154)	0.697(0.155)	0.472(0.173)
60	-0.819(0.158)	0.321(0.121)	-0.087(0.159)	0.398(0.146)	0.627(0.160)	0.447(0.164)
80	-0.916(0.158)	0.243(0.113)	-0.177(0.157)	0.335(0.142)	0.554(0.158)	0.414(0.163)
100	-1.023(0.162)	0.156(0.104)	-0.282(0.161)	0.250(0.134)	0.464(0.161)	0.360(0.158)

more flexible, we also simulated the data by misalignment rate. For example, for 10% mismatched rate, we assume that 10% of those days where the Y's are observed do not have corresponding predictors observed, and 10% of those days where X's are observed do not have corresponding responses observed.

The results can be seen in Tables 2.6 and 2.7, where it shows that as mismatch rate increases, the estimated β_1 decreases, for a fixed bin size. This is to be expected as the estimates do not reflect the original data as well when we have less matched information. Specifically, the result is that the estimates for β_1 tend more toward the null as we have less and less information. Also, a larger bin size results in an estimated β_1 closer to its target, which likely reflects more information (more observations) within a bin when bin size is larger, though with an increased variance tradeoff, due to fewer corresponding repeated measures, as seen earlier.

Table 2.7: Estimates for different bin sizes and various mismatch rates, for 40 weeks data. Before binning, β_0 = -2 and β_1 = 0.5. After binning, the target β_1 value remains at 0.5, whereas the target β_0 value will be a function of bin size based on an offset, increasing as bin size increases.

Mismatch	bin size = 28		bin size = 40		bin size = 56	
(%)	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$	$\hat{\beta}_0(S.E.)$	$\hat{\beta}_1(S.E.)$
0	0.654(0.132)	0.498(0.126)	1.012(0.132)	0.499(0.133)	1.347(0.134)	0.495(0.143)
10	0.632(0.130)	0.477(0.123)	0.993(0.131)	0.484(0.131)	1.330(0.132)	0.484(0.139)
20	0.609(0.131)	0.455(0.121)	0.973(0.131)	0.466(0.129)	1.311(0.133)	0.470(0.140)
40	0.558(0.129)	0.397(0.116)	0.926(0.130)	0.419(0.122)	1.271(0.132)	0.435(0.135)
60	0.495(0.130)	0.319(0.107)	0.869(0.131)	0.353(0.117)	1.222(0.133)	0.385(0.131)
80	0.414(0.127)	0.203(0.092)	0.794(0.128)	0.248(0.101)	1.154(0.129)	0.297(0.119)
100	0.315(0.134)	0.062(0.062)	0.682(0.134)	0.081(0.072)	1.038(0.132)	0.120(0.088)

2.5.4 Simulation summary

1. If the mismatch rate is fairly small, then we would suggest using a smaller bin, since the estimated β_1 would likely not be very different from its true value, but the standard deviation of the estimate would be smaller. We state this with more confidence when within-subject variability is small or when positive serial correlation is present. The apparent reason for smaller bin sizes resulting in smaller SE's for β_1 is that the smaller the bin size for a given follow-up period, the greater number of effective repeated measures, i.e., bins.

2. However, if there is large mismatch rate such as case IV (heading toward 100% mismatch), then a larger bin size would be suggested, noting larger bin sizes in this scenario result in less bias, though still with greater variability (due to a lower number of effective repeated measures per individual as stated in point #1 above).

3. When we have a greater number of predictor observations within a fixed bin size, the

estimates tend closer to the true β_1, as compared to the case for a smaller number of predictor observations within a fixed bin. We were able to identify this unsurprising result through simulation study, by looking at one dataset that had four days, including day 2, that were observed every week across the entire span of follow-up, whereas another dataset was similar but day 2 was only observed for the first 8 weeks, then once every 4 weeks thereafter. We also looked at more extreme versions of this scenario (not shown here), and the results were always improved for more predictor observations within a fixed bin size. This said, sufficient signal can still be captured with a small number of predictor observations within a bin, at least under various scenarios investigated in our simulations. This sufficient signal case for a low number of observations within a bin was also found in our modeling of the hemodialysis data presented in Section 2.4.

4. To evaluate bin edge effects, we used a smaller bin size for the first bin and last bin, compared to the rest of the data. For example, if there were 40 days of follow-up, and originally there were four bins of 10 days each (i.e., ([0,10], (10,20],(20,30],(30,40])), we would instead create five bins: [-5,5], (5,15],(15,25],(25,35],(35,45]. These newly created bins effectively had half-size lower and upper bins due to no data collected below 0 or above 40 days. In our simulation of 40 weeks (280 days) follow-up, for bin size 28 days, we let the first bin be [-21,7] (similarly, [-33,7] for bin size 40, and [-49, 7] for bin size 56). In this case, there is only a slight change on the estimated β_1. For example, for complete matched data, bin size 28, the estimated β_1 changed from 0.498 to 0.467, while it changed from 0.062 to 0.057 under data IV.

5. We should note that the hemodialysis data analyzed in Section 2.4 most closely resembles mismatched data type III. Hence, based on the simulation results, this would suggest that our estimates of β_1 from the binned modeling of hemodialysis data are biased toward 0, but that for the contemporaneous association for 30-day

bins in the larger dataset, and for the lagged association for 7-day bins in the smaller dataset, we still were able to capture sufficient signal of the association between longitudinal CRP and infection events via the proposed binning approach.

2.6 Discussion

The response and predictors of a longitudinal dataset, including the motivating example longitudinal hemodialysis data, may be measured at different time points. We have proposed a binning method to preprocess this data before doing any longitudinal modeling. Binning as a statistical technique is not a new method, and has been used in the density estimation literature for histograms and for faster processing of nonparametric curve estimators (see, for example, Fan and Marron, 1994). However, to our knowledge, binning has not been used for aligning within-individual longitudinal measures recorded at different time points as introduced here, and in particular, for investigating ordered associations between a longitudinal measured event and a time-dependent covariate, each recorded at different time points within each individual.

This binning approach has allowed us to answer important scientific questions on the hemodialysis dataset and could be applied to other longitudinal datasets with similar non-standard data structures. After binning the application dataset, we have applied the generalized linear mixed effects model and an extension, i.e., the zero-inflated Poisson model with random effects, which accounted for excess zeros in the binned response. We considered analysis of two versions of the dataset, each of which allowed us to target different biomedical questions. The larger dataset allowed us to consider the complete data which showed a contemporaneous association between infection events and crp levels, whereas the smaller dataset allowed detection of a lagged effect.

The latter was especially an important biological result, as it appears that infection

events occur prior to elevation of crp levels, with the smaller bin sizes (7-day) needed to detect the dynamics of this lagged association.

Though, through the binning method, we were able to uncover some meaningful findings with the hemodialysis study data, as well as evaluate properties of the approach through simulation, it is important to consider asymptotic properties of the resulting regression estimators following binning. This is a current area of our research, with some developments presented in Chapter 3. We are also developing a more formal model selection approach that takes into account bin size, lag order, model type, and resulting model fit. Regarding model type, we might consider models discussed here, such as the Poisson GLMM or mixed ZIP model, as well as models allowing for a more flexible association between the response and time-dependent predictor, such as generalized additive mixed effect models. We currently use AIC for relative model fit for comparing models (e.g., Poisson GLMM vs. mixed ZIP) for a fixed bin size, but this needs to be more fully investigated as a tool for model selection in the binning context in future work, and additional work is also required to consider tools for absolute model fit. We do not have a specific recommendation regarding how to optimally select bin sizes, an area of future research. However, initial work suggests that a variation of cross-validation may be an approach worth consideration, in particular, a variation of leave-one-subject-out cross-validation that is often used in estimating a mean curve across a sample of subjects (e.g.,). Define the leave-one-subject-out predicted residual sum squares PRSS= $\sum_{i=1}^{n} \sum_{m=1}^{n_i^{(X,Y)}} (Y_{i,m} - \hat{E}_{(-i)}(Y_{i,m}|\mathbf{u}_i))^2$, where $Y_{i,m}$ is the binned response in m^{th} bin for i^{th} subject, and $\hat{E}_{(-i)}(Y_{i,m}|\mathbf{u}_i)$ is the predicted conditional mean response in the m^{th} bin based on model estimates from the dataset without subject i. The best bin size could be the one that minimizes PRSS. One issue with the above approach is that it ignores the direct impact of the original data points Y_{ij}, only considering the individual points indirectly following binning, i.e., focusing on the $Y_{i,m}$. Implementing the above approach on the hemodialysis data suggests the smallest bin size considered

(i.e., 30-day bins in the full dataset, and 7-day bins in the subset 7-week dataset) is best using this PRSS criterion, though only by a small margin. Much more investigation of this approach is necessary, including the possible consideration of the original data points in the evaluation of bin size selection. We add further comments on bin size selection in Chapter 3.

Our simulation studies show that when the data is completely matched, the estimated association parameter β_1 will be close to the true value. Several factors will affect this estimate when the data is mismatched, such as proportion of mismatches, bin size, within-subject variability, and serial correlation with the predictors. For mismatched data, the association parameter is biased toward the null. However, with low within-subject variability and high serial correlation, the bias may be small. Although the bias is larger for completely mismatched data, there may still be sufficient signal in the data such that a statistically significant association can be detected, even if the association parameter is clearly biased downward. So, the goals of the modeling effort will need to be kept in mind when there exist high mismatch rates of the days when the predictor(s) and response are observed. If establishing an accurate association estimate is the goal, then binning when mismatch rates are high may not be recommended. However, if establishing a significant association between a predictor and response is the goal, including for lags of interest, as was the case for the hemodialysis data application, then binning can be recommended, even when mismatching is sizable.

Binning is a relatively straightforward approach that has proved useful in other areas of statistics, and has allowed us to answer some important biomedical questions for the longitudinal hemodialysis study discussed here. After binning, one is able to use more traditional longitudinal models and standard statistical packages. The approach, as evidenced in the simulation study, can have very useful properties, even under some less than ideal data settings.

Oversmoothing is rarely a good idea, for example, in histogram smoothing (Wand, 1997), where binning is relevant. A very large bin width results a histogram with a single block. For our binning method, we would also not want to choose bin sizes too large in the longitudinal modeling setting that exists in this thesis. One consequence, surely, is that any lagged associations will be lost, and possibly even a contemporaneous association. The large bins in the example (30-day and larger) did not let us obtain a lagged association, for example. But the smaller bins, i.e., 7-days, when allowed (in the subset dataset) resulted in a lagged association that was lost in the larger bin sizes.

Binning is not limited to a single predictor. For multiple predictors, there is actually a greater opportunity to use binning, as there could be even greater misalignment than with just one predictor. However, the bias and efficiency of the estimated parameters need to be further investigated for binning method for multiple predictors.

For longitudinal data measured at distinct time points, binning has been shown to be useful to find the temporal association between predictors and response. Apart from binning method, curve-based methods can also serve as potential approaches to model the relationship between the predictors and response. Also, methods in missing data framework such as EM algorithm may be used to find the association between response and predictors under the misalignment setting described in this chapter.

Chapter 3

Inference for binning estimates and bin size selection

3.1 Estimation and inference for longitudinal data

Let Y_{ij} be the response variable for subject i at time point j, $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})'$ be the $p \times 1$ covariate vector associated with Y_{ij}, where $i = 1, \ldots, n$, $j = 1, \ldots, n_i$. Denote $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ as the vector of all responses for subject i, and $\mathbf{X}_i = (\mathbf{X}'_{i1}, \mathbf{X}'_{i2}, \ldots, \mathbf{X}'_{in_i})'$ as the corresponded $n_i \times p$ covariate matrix. Assume $\mathbf{Y}_i$ is multivariate normally distributed with mean $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix $var(\mathbf{Y}_i) = \mathbf{V}_i$, and assume $\mathbf{Y}_i$ are independent of each other for different i. We specify the marginal regression model as

$$\mathbf{Y_i} = \mathbf{X_i}\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \quad , \tag{3.1}$$

where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i1}, \ldots, \epsilon_{in_i})$, and $\boldsymbol{\epsilon}_i \sim N(0, \mathbf{V}_i)$. $\boldsymbol{\epsilon}_i$ in (3.1) accounts for all the correlations between repeated measurements for subject i. If $\mathbf{V}_i = \sigma^2 I_{n_i}$ where I_{n_i} is the identity matrix of dimension n_i, then the repeated observations for subject i are assumed independent. In this case model (3.1) becomes a univariate linear regression model.

Usually $\boldsymbol{\beta}$ and $\mathbf{V}_i$ are unknown and $\mathbf{V}_i$ is assumed to be positive-definite. Suppose $\mathbf{V}_i = \mathbf{V}_i(\boldsymbol{\alpha})$, and $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\alpha}')$. Then the marginal likelihood function takes the form

$$L(\boldsymbol{\theta}) = \Pi_{i=1}^{n} \left\{ (2\pi)^{-n_i/2} |\mathbf{V}_i(\boldsymbol{\alpha})|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta})' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) \right] \right\} \quad .$$

Estimation of $\boldsymbol{\theta}$ requires joint maximization of the above likelihood with respect to all elements in $\boldsymbol{\theta}$.

For a given $\boldsymbol{\alpha}$, the MLE of $\boldsymbol{\beta}$ is given by a generalized least squares (GLS) estimator

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{W}_i \mathbf{Y}_i \quad ,$$

where $\mathbf{W}_i = \mathbf{V}_i^{-1}$. However, $\boldsymbol{\alpha}$ is usually unknown and can be replaced by its MLE $\hat{\boldsymbol{\alpha}}$. Most often, $\boldsymbol{\alpha}$ is estimated by the restricted maximum likelihood (REML) estimator, which allows estimation of $\boldsymbol{\alpha}$ without having to estimate $\boldsymbol{\beta}$ first.

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})) &= \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{W}_i E(\mathbf{Y}_i) = \boldsymbol{\beta} \quad . \\
\mathrm{Var}(\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha})) &= \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{W}_i \mathrm{Var}(\mathbf{Y}_i) \mathbf{W}_i \mathbf{X}_i \right) \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i \right)^{-1} \\
&= \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{W}_i \mathbf{X}_i \right)^{-1} \quad .
\end{aligned}$$

provided the mean and variance are correctly specified for the model, i.e. $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ and $\mathrm{Var}(\mathbf{Y}_i) = \mathbf{V}_i$.

Still define $\mathbf{Y}_i$ as a response vector for subject i, and $\mathbf{X}_i$ as its covariate matrix for fixed effects $\boldsymbol{\beta}$. For a linear mixed effects model defined in Section 1.2,

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

where $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iq})'$ is $q \times 1$ vector of random effects for subject i; $\mathbf{Z}_i = (\mathbf{Z}_{i1}', \mathbf{Z}_{i2}', \ldots, \mathbf{Z}_{in_i}')'$ is the $n_i \times q$ design matrix for subject specific random effects $\mathbf{u}_i$ and $\mathbf{Z}_{ij} = (Z_{ij1}, Z_{ij2}, \ldots,$

$Z_{ijq})'$. $\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})'$ is the random error term. Typically we assume $\mathbf{u}_i$ are distributed as $N(0, \mathbf{D})$ and independent of each other. $\mathbf{e}_i$ are distributed as $N(0, \Sigma_i)$ and independent of each other. $\mathbf{u}_i$ are assumed to be independent of $\mathbf{e}_i$. $\mathbf{Z}_i$ is most often a subset of $\mathbf{X}_i$.

Conditional on random effects $\mathbf{u}_i$, $\mathbf{Y}_i$ are distributed as $N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$. Marginally, $\mathbf{Y}_i$ is distributed as $N(\mathbf{X}\boldsymbol{\beta}, \mathbf{ZDZ}' + \boldsymbol{\Sigma}_i)$. Thus a linear mixed effects model implies a marginal model as (3.1) but with a special parametric form of covariance matrix $\mathbf{V}_i(\boldsymbol{\alpha}) = \mathbf{ZDZ}' + \boldsymbol{\Sigma}_i$. Vector $\boldsymbol{\alpha}$ includes the unique parameters in $\mathbf{D}$ and $\boldsymbol{\Sigma}_\mathbf{i}$. Fitting of a linear mixed effects model is usually based on the marginal model and its marginal likelihood as above. $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are still estimated by ML or REML estimator as marginal models. Random effects $\mathbf{u}_i$ usually are predicted by empirical Bayes estimator (Laird and Ware, 1982).

3.2 Binning method for longitudinal continuous data

For traditional longitudinal data, a covariate X_{ij} and response Y_{ij} are measured at the same time points $j, j = 1, 2, \ldots, n_i$. However, in the case that the response and covariates are not measured on the same time points such as the hemodialysis data discussed in Chapter 2, we use the binning method which was introduced in Chapter 2 to align them to common time points. In hemodialysis data, the response Y_{ij} is binary. Here we will treat response Y_{ij} as continuous. Most notation and steps of the binning method are the same for both binary and continuous response, except step 4 below. For continuous data, the new binned response in step 4 is the average of the original responses. However, for binary data, it is the sum of the original responses in a bin.

1. Let Y_{ij} be the continuous response observed for subject i at time j, where $j = 1, 2, \ldots, n_i^{(\mathbf{Y})}$.

2. Let X_{ik} be a covariate for subject i observed at time k for $k = 1, 2, \ldots, n_i^{(\mathbf{X})}$, where, in general, the times represented by k do not equal the times represented by j and $n_i^{(\mathbf{X})} \neq n_i^{(\mathbf{Y})}$.

3. We then bin $(\mathbf{X}, \mathbf{Y})$ in equidistant units of time.

4. Next we take the unweighted or weighted average of response or covariate within each bin, for each individual.

5. The resulting data will be $(X_{i,m}, Y_{i,m})$, where $m = 1, 2, , \ldots, n_i^{(\mathbf{X},\mathbf{Y})}$. That is, there are a total of $n_i^{(\mathbf{X},\mathbf{Y})}$ bins for individual i.

In the above, we could bin multiple (say p) longitudinal covariates of interest, creating $\mathbf{X}_{i,m} = (X_{1,i,m}, X_{2,i,m}, ..., X_{p,i,m})$, the average then for each covariate separately within each bin.

To investigate the properties of binning estimates, we use the notation of *gold standard data* (GSD). For this GSD, it includes all the distinct time points for both response and covariates, and at each time point, both the response and covariate values are available. This is a setup where binning would be unnecessary, but allows for a comparison when some of the observed data (either the response or covariate or both) are not all actually observed at the same time points. Assume the GSD Y_{ij} and $\mathbf{X}_{ij}$, $i = 1, 2, \ldots, n$, $j = 1, 2, 3, \ldots, n_i$, are modeled as in (3.1); we then have

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \text{ where } \boldsymbol{\epsilon}_i \sim N(0, \mathbf{V}_i) \quad .$$

However, in general, the covariates and response are not measured on the same time points; a setting for which the binning process is applied. Binning can be considered as a type of transformation of GSD. At each time point, we apply a weight on the GSD. If the observed response or covariates are not available, then we apply a zero weight to the corresponded response or covariates in the GSD. These weights will be captured in weight

matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ which will be defined in the following sections. We use $\mathbf{X}_i$ and $\mathbf{Y}_i$ to denote the GSD, and use $\mathbf{X}_i^\star$ and $\mathbf{Y}_i^\star$ as the resulting data after binning.

After performing binning, for i^{th} subject and m^{th} bin, assume the number of time points for observed X is $c_{im}^{(\mathbf{X})}$, the number of observed Y is $c_{im}^{(\mathbf{Y})}$, and the number of elements in the union of distinct time points for both response and covariates are $c_{im}^{(\mathbf{X},\mathbf{Y})}$ where $i = 1, 2, \ldots, n$; $m = 1, 2, \ldots, n_i^{(\mathbf{X},\mathbf{Y})}$. So, for example, say for subject i and bin m, covariate X is observed at time point (1, 4, 5) and response is observed at time point (1, 4, and 6). Then $c_{im}^{(\mathbf{X})} = 3$, $c_{im}^{(\mathbf{Y})} = 3$, and $c_{im}^{(\mathbf{X},\mathbf{Y})} = 4$. In general,

$$\sum_{m=1}^{n_i^{(\mathbf{X},\mathbf{Y})}} c_{im}^{(\mathbf{X})} = n_i^{(\mathbf{X})}, \quad \sum_{m=1}^{n_i^{(\mathbf{X},\mathbf{Y})}} c_{im}^{(\mathbf{Y})} = n_i^{(\mathbf{Y})}, \quad \sum_{i=1}^{n} \sum_{m=1}^{n_i^{(\mathbf{X},\mathbf{Y})}} c_{im}^{(\mathbf{X},\mathbf{Y})} = \sum_{i=1}^{n} n_i \quad .$$

3.2.1 Binning estimates for complete matched data

First, we assume the observed data before binning is completely matched, i.e. at each observed time point, both response and covariates values are available. In this case, the observed data is the same as GSD. For the complete matched case, after performing binning, then $c_{im}^{(\mathbf{X})} = c_{im}^{(\mathbf{Y})} = c_{im}$ and $n_i^{(\mathbf{X})} = n_i^{(\mathbf{Y})} = n_i$. However for the more general mismatched case in the next section, usually $c_{im}^{(\mathbf{X})} \neq c_{im}^{(\mathbf{Y})}$ and $n_i^{(\mathbf{X})} \neq n_i^{(\mathbf{Y})}$. The vector of binned response for the i^{th} subject is denoted as $\mathbf{Y}_i^\star = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n_i^{(X,Y)}})'$, the binned covariate design matrix is $\mathbf{X}_i^\star = (\mathbf{X}_{i,1}, \mathbf{X}_{i,2}, \ldots, \mathbf{X}_{i,n_i^{(\mathbf{X},\mathbf{Y})}})'$. We can think of the binning process as a transformation of the original data

$$\mathbf{Y}_i^\star = \mathbf{A}_i \mathbf{Y}_i \quad ,$$

$$\mathbf{X}_i^\star = \mathbf{B}_i \mathbf{X}_i \quad .$$

However, for complete matched data, $\mathbf{A}_i = \mathbf{B}_i$, where

$$\mathbf{A}_i = \mathbf{B}_i = \begin{bmatrix} \overbrace{\frac{1}{c_{i1}} \cdots \frac{1}{c_{i1}}}^{c_{i1} \text{ elements}} & & & \\ & \overbrace{\frac{1}{c_{i2}} \cdots \frac{1}{c_{i2}}}^{c_{i2} \text{ elements}} & & \\ & & \cdots & \\ & & & \overbrace{\frac{1}{c_{in_i^{(\mathbf{X},\mathbf{Y})}}} \cdots \frac{1}{c_{in_i^{(\mathbf{X},\mathbf{Y})}}}}^{c_{in_i^{(\mathbf{X},\mathbf{Y})}} \text{ elements}} \end{bmatrix}_{n_i^{(\mathbf{X},\mathbf{Y})} \times n_i} . \tag{3.2}$$

After binning, we will fit the model

$$\mathbf{Y}_i^\star = \mathbf{X}_i^\star \boldsymbol{\beta}^\star + \boldsymbol{\epsilon}_i^\star \quad ,$$

where $\boldsymbol{\epsilon}_i^\star \sim F(\cdot)$, $E(\boldsymbol{\epsilon}_i^\star) = 0$ and $var(\boldsymbol{\epsilon}_i^\star) = \mathbf{V}_i^\star$ with $\mathbf{V}_i^\star$ being a positive definite covariance matrix. Note $var(\boldsymbol{\epsilon}_i^\star) = \mathbf{V}_i^\star = var(\mathbf{Y}_i^\star) = var(\mathbf{A}_i\mathbf{Y}_i) = \mathbf{A}_i var(\mathbf{Y}_i)\mathbf{A}_i^{'} = \mathbf{A}_i\mathbf{V}_i\mathbf{A}_i^{'}$. So

$$\mathbf{V}_i^\star = \mathbf{A}_i\mathbf{V}_i\mathbf{A}_i^{'} = \mathbf{B}_i\mathbf{V}_i\mathbf{B}_i^{'}, \text{ as } \mathbf{A}_i = \mathbf{B}_i \text{ in the complete matched case.}$$

Let $\mathbf{W}_i^\star = (\mathbf{V}_i^\star)^{-1}$. Then

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^\star &= \left(\sum_{i=1}^n \mathbf{X}_i^{\star'}\mathbf{W}_i^\star\mathbf{X}_i^\star\right)^{-1} \sum_{i=1}^n \mathbf{X}_i^{\star'}\mathbf{W}_i^\star\mathbf{Y}_i^\star \\
&= \left(\sum_{i=1}^n (\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^\star(\mathbf{B}_i\mathbf{X}_i)\right)^{-1} \sum_{i=1}^n (\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^\star\mathbf{B}_i\mathbf{Y}_i \quad . \\
E(\hat{\boldsymbol{\beta}}^\star) &= \left(\sum_{i=1}^n (\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^\star(\mathbf{B}_i\mathbf{X}_i)\right)^{-1} \sum_{i=1}^n (\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^\star\mathbf{B}_i E(\mathbf{Y}_i) \\
&= \left(\sum_{i=1}^n (\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^\star(\mathbf{B}_i\mathbf{X}_i)\right)^{-1} \sum_{i=1}^n (\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^\star\mathbf{B}_i\mathbf{X}_i\boldsymbol{\beta} = \boldsymbol{\beta} \quad . \\
var(\hat{\boldsymbol{\beta}}^\star) &= \left(\sum_{i=1}^n \mathbf{X}_i^{\star'}\mathbf{W}_i^\star\mathbf{X}_i^\star\right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i^{\star'}\mathbf{W}_i^\star \mathrm{Var}(\mathbf{Y}_i^\star)\mathbf{W}_i^\star\mathbf{X}_i^\star\right) \left(\sum_{i=1}^n \mathbf{X}_i^{\star'}\mathbf{W}_i^\star\mathbf{X}_i^\star\right)^{-1} \\
&= \left(\sum_{i=1}^n \mathbf{X}_i^{\star'}\mathbf{W}_i^\star\mathbf{X}_i^\star\right)^{-1} = \left(\sum_{i=1}^n (\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^\star(\mathbf{B}_i\mathbf{X}_i)\right)^{-1} \quad .
\end{aligned}$$

So $\hat{\boldsymbol{\beta}}^{\star}$ is an unbiased estimate of $\boldsymbol{\beta}$ for complete matched data.

We can compare $var(\hat{\boldsymbol{\beta}})$ with $var(\hat{\boldsymbol{\beta}}^{\star})$.

$$var(\hat{\boldsymbol{\beta}}) = (\sum_{i=1}^{n} \mathbf{X}_i^{'} \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} = (\sum_{i=1}^{n} \mathbf{X}_i^{'} \mathbf{W}_i \mathbf{X}_i)^{-1}, \text{ where } \mathbf{W}_i = \mathbf{V}_i^{-1}, \text{ and}$$

$$var(\hat{\boldsymbol{\beta}}^{\star}) = \left(\sum_{i=1}^{n} (\mathbf{X}_i^{\star'} \mathbf{W}_i^{\star} \mathbf{X}_i^{\star}) \right)^{-1} = \left(\sum_{i=1}^{n} \mathbf{X}_i^{'} \mathbf{B}_i^{'} (\mathbf{B}_i \mathbf{V}_i \mathbf{B}_i^{'})^{-1} \mathbf{B}_i \mathbf{X}_i \right)^{-1} .$$

The difference between the $var(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{\beta}}^{\star})$ is mainly due to the difference between$\mathbf{V}_i^{-1}$ and $\mathbf{B}_i^{'}(\mathbf{B}_i \mathbf{V}_i \mathbf{B}_i^{'})^{-1} \mathbf{B}_i$. In order to see the difference between $var(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{\beta}}^{\star})$ more closely, we look into special cases as follows. Assume $c_{i1} = c_{i2} = \ldots = c_{in_i^{(\mathbf{X},\mathbf{Y})}} = c$, i.e., the number of original observations in each bin are the same. And assume there are a total of M bins for each subject which means $n_i^{(\mathbf{X},\mathbf{Y})} = M$ for all i. So, total number of observations for each subject is cM. Let $J = cM$. The assumptions of equal numbers of observations within each bin and equal number of bins for each subject are just for demonstration only. They are not required by the binning method itself.

Case I: assume $\mathbf{V}_i = \sigma^2 I_{n_i}$. Then

$$var(\hat{\boldsymbol{\beta}}^{\star}) = \frac{\sigma^2}{c} \left(\sum_{i=1}^{n} \mathbf{X}_i^{'} \mathbf{B}_i^{'} \mathbf{B}_i \mathbf{X}_i \right)^{-1} .$$

$$var(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\sum_{i=1}^{n} \mathbf{X}_i^{'} \mathbf{X}_i \right)^{-1} .$$

The difference between $var(\hat{\boldsymbol{\beta}}^{\star})$ and $var(\hat{\boldsymbol{\beta}})$ will be reflected by matrix $\mathbf{B}_i^{'} \mathbf{B}_i$. And given the above assumptions,

$$\mathbf{B}_i^{'} \mathbf{B}_i = diag(\frac{1}{c} \mathbf{1}_{c \times c}, \ldots, \frac{1}{c} \mathbf{1}_{c \times c})_{J \times J} .$$

a block diagonal matrix with dimension $J \times J$. $\mathbf{1}$ is a matrix with all elements of 1.

Assume we have one single covariate X_{ij}, then $\boldsymbol{\beta} = (\beta_0, \beta_1)$,

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}) &= \sigma^2 (\sum_{i=1}^{n} \mathbf{X_i}' \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \\ &= \sigma^2 \left[\sum_{i=1}^{n} \begin{pmatrix} 1 & X_{i1} \\ \cdots & \cdots \\ 1 & X_{iJ} \end{pmatrix}' \begin{pmatrix} 1 & X_{i1} \\ \cdots & \cdots \\ 1 & X_{iJ} \end{pmatrix}^{-1} \right] \\ &= \frac{\sigma^2}{\sum_{i=1}^{n} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{\cdot\cdot})^2} \begin{pmatrix} \frac{1}{nJ} \sum_{i=1}^{n} \sum_{j=1}^{J} X_{ij}^2 & -\bar{X}_{\cdot\cdot} \\ -\bar{X}_{\cdot\cdot} & 1 \end{pmatrix}, \end{aligned}$$

where $\bar{X}_{\cdot\cdot} = \frac{1}{nJ} \sum_{i=1}^{n} \sum_{j=1}^{J} X_{ij}$. And

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}^{\star}) &= \frac{\sigma^2}{c} \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{B}_i' \mathbf{B}_i \mathbf{X}_i \right)^{-1} = \frac{\sigma^2}{c} \left(\sum_{i=1}^{n} (\mathbf{X}_i^{\star})' \mathbf{X}_i^{\star} \right)^{-1} \\ &= \frac{\sigma^2/c}{\sum_{i=1}^{n} \sum_{m=1}^{M} (X_{i,m} - \bar{X}_{\cdot\cdot}^{\star})^2} \begin{pmatrix} \frac{1}{nM} \sum_{i=1}^{n} \sum_{m=1}^{M} X_{i,m}^2 & -\bar{X}_{\cdot\cdot}^{\star} \\ -\bar{X}_{\cdot\cdot}^{\star} & 1 \end{pmatrix}, \end{aligned}$$

where $X_{i,m}$ is the covariate in the m^{th} bin for the i^{th} subject which has been defined in Section 3.2. $\bar{X}_{\cdot\cdot}^{\star} = \frac{1}{nM} \sum_{i=1}^{n} \sum_{m=1}^{M} X_{i,m}$, the overall mean of the binned covariates. We can see that $var(\hat{\boldsymbol{\beta}})$ and $var(\hat{\boldsymbol{\beta}}^{\star})$ depend on the sum squares of X_{ij} and $X_{i,m}$, respectively. And $var(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^{n} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{\cdot\cdot})^2 = \sigma^2/((nJ-1)S_X^2)$, where $S_X^2 = \sum_{i=1}^{n} \sum_{j=1}^{J} (X_{ij} - \bar{X}_{\cdot\cdot})^2/(nJ-1)$, the sample variance of X_{ij}, and $var(\hat{\beta}_1^{\star}) = \sigma^2/(c \sum_{i=1}^{n} \sum_{m=1}^{M} (X_{i,m} - \bar{X}_{\cdot\cdot}^{\star})^2) = \sigma^2/(c(nM-1)S_{X^{\star}}^2)$ where $S_{X^{\star}}^2 = \sum_{i=1}^{n} \sum_{m=1}^{M} (X_{i,m} - \bar{X}_{\cdot\cdot}^{\star})^2/(nM-1)$, the sample variance of $X_{i,m}$. Because we obtain the average in each bin, the sample variance S_X^2 would usually be greater than $S_{X^{\star}}^2$. Also there will be fewer resulting repeated measurements following binning, as compared to the original set of observations. So $var(\hat{\beta}_1)$ would usually be smaller than $var(\hat{\beta}_1^{\star})$. This is consistent with the simulations in Chapter 2. As bin size increases, $var(\hat{\beta}_1{}^{\star})$ will generally increase too.

Case II: assume $\mathbf{V}_i$ is a compound symmetric matrix and has the form as

$$\mathbf{V}_i = \begin{bmatrix} \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \dots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \dots & \sigma_b^2 \\ \dots & \dots & & \dots \\ \sigma_b^2 & \dots & \sigma_b^2 & \sigma_b^2+\sigma_e^2 \end{bmatrix}_{J\times J} = (\sigma_b^2+\sigma_e^2) \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & & \dots \\ \rho & \dots & \rho & 1 \end{bmatrix}_{J\times J} .$$

where $\rho = \frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2}$. Then

$$\mathbf{V}_i^{-1} = a_0 \begin{bmatrix} 1 & \rho' & \dots & \rho' \\ \rho' & 1 & \dots & \rho' \\ \dots & \dots & & \dots \\ \rho' & \dots & \rho' & 1 \end{bmatrix}_{J\times J} .$$

where $a_0 = \frac{1+(J-2)\rho}{(1+(J-2)\rho-(J-1)\rho^2)(\sigma_b^2+\sigma_e^2)}$, and $\rho' = \frac{-\rho}{1+(J-2)\rho}$.

$$\sum_{i=1}^{n} \mathbf{X_i}'\mathbf{V}_i^{-1}\mathbf{X}_i = \sum_{i=1}^{n} \begin{pmatrix} 1 & X_{i1} \\ \dots & \dots \\ 1 & X_{iJ} \end{pmatrix}' a_0 \begin{pmatrix} 1 & \rho' & \dots & \rho' \\ \rho' & 1 & \dots & \rho' \\ \dots & \dots & & \dots \\ \rho' & \dots & \rho' & 1 \end{pmatrix} \begin{pmatrix} 1 & X_{i1} \\ \dots & \dots \\ 1 & X_{iJ} \end{pmatrix}$$

$$= a_0 \begin{pmatrix} nJ(1-\rho'+J\rho') & (1-\rho'+J\rho')\sum_{i=1}^{n}\sum_{j=1}^{J} X_{ij} \\ (1-\rho'+J\rho')\sum_{i=1}^{n}\sum_{j=1}^{J} X_{ij} & (1-\rho')\sum_{i=1}^{n}\sum_{j=1}^{J} X_{ij}^2 + \rho'\sum_{i=1}^{n}(\sum_{j=1}^{J} X_{ij})^2 \end{pmatrix} .$$

Then

$$var(\hat{\boldsymbol{\beta}}) = (\sum_{i=1}^{n} \mathbf{X_i}'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1} =$$

$$\frac{1}{\Delta} \begin{pmatrix} \frac{1}{nJ}\left((1-\rho')\sum_{i=1}^{n}\sum_{j=1}^{J} X_{ij}^2 + J^2\rho'\sum_{i=1}^{n}\bar{X}_{i\cdot}^2\right) & -(1-\rho'+J\rho')\bar{X}_{\cdot\cdot} \\ -(1-\rho'+J\rho')\bar{X}_{\cdot\cdot} & 1-\rho'+J\rho' \end{pmatrix} ,$$

where $\Delta = (1-\rho'+J\rho')a_0\left\{(1-\rho')\sum_{i=1}^{n}\sum_{j=1}^{J}(X_{ij}-\bar{X}_{\cdot\cdot})^2 + \rho'\sum_{i=1}^{n}\sum_{j=1}^{J}(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^2\right\}$.
And $\bar{X}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} X_{ij}$, $\bar{X}_{\cdot\cdot} = \frac{1}{nJ}\sum_{i=1}^{n}\sum_{j=1}^{J} X_{ij}$.

We know that $var(\hat{\boldsymbol{\beta}}^{\star}) = (\sum_{i=1}^{n} \mathbf{X}_i^{\star'} \mathbf{W}_i^{\star} \mathbf{X}_i^{\star})^{-1} = (\sum_{i=1}^{n} \mathbf{X}_i^{\star'} (\mathbf{B}_i \mathbf{V}_i \mathbf{B}_i)^{-1} \mathbf{X}_i^{\star})^{-1}$. The calculation of $var(\hat{\boldsymbol{\beta}}^{\star})$ will be very similar to the calculation of $var(\hat{\boldsymbol{\beta}})$. We just need to change the $\mathbf{X}_i$ in $var(\hat{\boldsymbol{\beta}})$ to $\mathbf{X}_i^{\star}$, and change the $\mathbf{V}_i$ to $\mathbf{B}_i \mathbf{V}_i \mathbf{B}_i$.

$$\mathbf{B}_i \mathbf{V}_i \mathbf{B}_i^{'} = (\sigma_b^2 + \sigma_e^2/c) \begin{bmatrix} 1 & \rho_1 & \dots & \rho_1 \\ \rho_1 & 1 & \dots & \rho_1 \\ \dots & \dots & & \dots \\ \rho_1 & \dots & \rho_1 & 1 \end{bmatrix}_{M \times M} ,$$

and here the correlation $\rho_1 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2/c}$ for matrix $\mathbf{B}_i \mathbf{V}_i \mathbf{B}_i^{'}$. Note here, the covariance matrix dimension is $M \times M$, since the effective repeated measurements are the binned responses. Similarly, we can get the variance of $\hat{\boldsymbol{\beta}}^{\star}$ as follows

$$var(\hat{\boldsymbol{\beta}}^{\star}) = \\ \frac{1}{\Delta_1} \begin{pmatrix} \frac{1}{nM} \left((1 - \rho_1^{'}) \sum_{i=1}^{n} \sum_{m=1}^{M} X_{i,m}^2 + M^2 \rho_1^{'} \sum_{i=1}^{n} \bar{X}_{i\cdot}^{\star 2} \right) & -(1 - \rho_1^{'} + M\rho_1^{'}) \bar{X}_{\cdot\cdot}^{\star} \\ -(1 - \rho_1^{'} + M\rho_1^{'}) \bar{X}_{\cdot\cdot}^{\star} & 1 - \rho_1^{'} + M\rho_1^{'} \end{pmatrix} ,$$

where $\Delta_1 = (1 - \rho_1^{'} + M\rho_1^{'}) a_1 \{ (1 - \rho_1^{'}) (\sum_{i=1}^{n} \sum_{m=1}^{M} (X_{i,m} - \bar{X}_{\cdot\cdot}^{\star})^2$ $+ \rho_1^{'} \sum_{i=1}^{n} \sum_{m=1}^{M} (\bar{X}_{i\cdot}^{\star} - \bar{X}_{\cdot\cdot}^{\star})^2 \}$ and $\bar{X}_{i\cdot}^{\star} = \frac{1}{M} \sum_{m=1}^{M} X_{i,m}$, $\bar{X}_{\cdot\cdot}^{\star} = \frac{1}{nM} \sum_{i=1}^{n} \sum_{m=1}^{M} X_{i,m}$, $\rho_1 = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2/c}$, $\rho_1^{'} = \frac{-\rho_1}{1 + (M-2)\rho_1}$, $a_1 = \frac{1 + (M-2)\rho_1}{(1 + (M-2)\rho_1 - (M-1)\rho_1^2)(\sigma_b^2 + \sigma_e^2/c)}$. Here $X_{i,m}$ is the covariate value after binning. Here $var(\hat{\beta}_1^{\star}) = 1/(a_1 \{ (1 - \rho_1^{'}) \sum_{i=1}^{n} \sum_{m=1}^{M} (X_{i,m} - \bar{X}_{\cdot\cdot}^{\star})^2$ $+ \rho_1^{'} \sum_{i=1}^{n} \sum_{m=1}^{M} (\bar{X}_{i\cdot}^{\star} - \bar{X}_{\cdot\cdot}^{\star})^2 \})$. If the within-subject variability of X_{ij} is increased, then $\sum_{i=1}^{n} \sum_{m=1}^{M} (X_{i,m} - \bar{X}_{\cdot\cdot}^{\star})^2$ and $\sum_{i=1}^{n} \sum_{m=1}^{M} (\bar{X}_{i\cdot}^{\star} - \bar{X}_{\cdot\cdot}^{\star})^2$ will be increased, thus $var(\hat{\beta}_1^{\star})$ will be decreased.

3.2.2 Binning estimates for mismatched data

Assume now the more general case where the response or covariates are mismatched, where at some observed time points, either only response or covariate is observed, but not for

both. We now write

$$\mathbf{Y}_i^\star = \mathbf{A}_i \mathbf{Y}_i,\ \mathbf{X}_i^\star = \mathbf{B}_i \mathbf{X}_i \quad ,$$

where $\mathbf{Y}_i$ and $\mathbf{X}_i$ denote the GSD for subject i, as defined in the previous section.

$\mathbf{Y}_i^\star$ requires its own weight matrix $\mathbf{A}_i$ whereas $\mathbf{X}_i^\star$ requires a distinct matrix $\mathbf{B}_i$. In general, for the mismatched case, $\mathbf{A}_i \neq \mathbf{B}_i$. As denoted earlier in Section 3.2, $c_{im}^{(\mathbf{X})}$ is the number of observed time points for covariates in m^{th} bin, $c_{im}^{(\mathbf{Y})}$ is the number of observed time points for response in m^{th} bin, and $c_{im}^{(\mathbf{X},\mathbf{Y})}$ is the number of elements in the union of distinct time points for both response and covariates in m^{th} bin. For example, for m^{th} bin, assume the covariates are observed at time points 0, 3, 6, and the response is observed at time points 0, 2, 4, 6. Then the union of time points are (0, 2, 3, 4, 6). Thus $c_{im}^{(\mathbf{X})} = 3$, $c_{im}^{(\mathbf{Y})} = 4$ and $c_{im}^{(\mathbf{X},\mathbf{Y})} = 5$. Let $c_i = \sum_{m=1}^{n_i^{(\mathbf{X},\mathbf{Y})}} c_{im}^{(\mathbf{X},\mathbf{Y})}$.

$$\mathbf{A}_i = \left[\overbrace{\frac{1}{c_{i1}^{(\mathbf{Y})}}, \frac{1}{c_{i1}^{(\mathbf{Y})}}, 0, \ldots, \frac{1}{c_{i1}^{(\mathbf{Y})}}}^{c_{i1}^{(\mathbf{X},\mathbf{Y})} \text{ elements}} \quad \overbrace{\frac{1}{c_{i2}^{(\mathbf{Y})}}, \frac{1}{c_{i2}^{(\mathbf{Y})}}, 0, \ldots, \frac{1}{c_{i2}^{(\mathbf{Y})}}}^{c_{i2}^{(\mathbf{X},\mathbf{Y})} \text{ elements}} \quad \cdots \quad \overbrace{\frac{1}{c_{in_i^{(\mathbf{X},\mathbf{Y})}}^{(\mathbf{Y})}}, \frac{1}{c_{in_i^{(\mathbf{X},\mathbf{Y})}}^{(\mathbf{Y})}}, 0, \ldots, \frac{1}{c_{in_i^{(X,Y)}}^{(Y)}}}^{c_{in_i^{(\mathbf{X},\mathbf{Y})}} \text{ elements}} \right]_{n_i^{(\mathbf{X},\mathbf{Y})} \times c_i} . \tag{3.3}$$

$$\mathbf{B}_i = \begin{bmatrix} \overbrace{\frac{1}{c_{i1}^{(\mathbf{X})}}, 0, \frac{1}{c_{i1}^{(\mathbf{X})}}, \ldots, \frac{1}{c_{i1}^{\mathbf{X}}}}^{c_{i1}^{(\mathbf{X},\mathbf{Y})} \text{ elements}} & & & \\ & \overbrace{\frac{1}{c_{i2}^{(\mathbf{X})}}, 0, \frac{1}{c_{i2}^{(\mathbf{X})}} \ldots, \frac{1}{c_{i2}^{(\mathbf{X})}}}^{c_{i2}^{(\mathbf{X},\mathbf{Y})} \text{ elements}} & & \\ & & \ldots & \\ & & & \overbrace{\frac{1}{c_{in_i^{(\mathbf{X},\mathbf{Y})}}^{(\mathbf{X})}}, 0, \frac{1}{c_{in_i^{(\mathbf{X},\mathbf{Y})}}^{(\mathbf{X})}}, \ldots, \frac{1}{c_{in_i^{(\mathbf{X},\mathbf{Y})}}^{(\mathbf{X})}}}^{c_{in_i^{(\mathbf{X},\mathbf{Y})}}^{(\mathbf{X},\mathbf{Y})} \text{ elements}} \end{bmatrix}_{n_i^{(\mathbf{X},\mathbf{Y})} \times c_i} . \tag{3.4}$$

We can think of matrices (3.3) and (3.4) as weight matrices which connect the data before and after binning via GSD. GSD contains all the distinct time points for both the original observed response and covariates before binning. In the m^{th} bin, for those time points available in the GSD but where only the observed covariate values are available and the responses are not available, the weights of those time points on $\mathbf{B}_i$ are set to 0, otherwise set to $1/c_{im}^{(\mathbf{Y})}$; on the other hand, for those time points available in the GSD but where only the observed responses are available and covariates are not available, then the weights of those time points on $\mathbf{A}_i$ are set to 0, otherwise set to $1/c_{im}^{(\mathbf{X})}$. For the example above, the weight vector for covariate in the m^{th} bin is $(1/3, 0, 1/3, 0, 1/3)$; the weight vector for the response in the m^{th} bin is $(1/4, 1/4, 0, 1/4, 1/4)$.

Then the estimated $\hat{\beta}^{\star}$ is

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\star} &= (\sum_{i=1}^{n} \mathbf{X}_i^{\star\prime} \mathbf{W}_i^{\star} \mathbf{X}_i^{\star})^{-1} \sum_{i=1}^{n} (\mathbf{X}_i^{\star\prime} \mathbf{W}_i^{\star} \mathbf{Y}_i^{\star}) \\ &= (\sum_{i=1}^{n} (\mathbf{B}_i \mathbf{X}_i)' \mathbf{W}_i^{\star} (\mathbf{B}_i \mathbf{X}_i))^{-1} \sum_{i=1}^{n} (\mathbf{B}_i \mathbf{X}_i)' \mathbf{W}_i^{\star} \mathbf{A}_i \mathbf{Y}_i, \text{ where} \\ E(\hat{\boldsymbol{\beta}}^{\star}) &= (\sum_{i=1}^{n} (\mathbf{B}_i \mathbf{X}_i)' \mathbf{W}_i^{\star} (\mathbf{B}_i \mathbf{X}_i))^{-1} \sum_{i=1}^{n} (\mathbf{B}_i \mathbf{X}_i)' \mathbf{W}_i^{\star} \mathbf{A}_i \mathbf{X_i} \boldsymbol{\beta} \\ &= (\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{B}_i' \mathbf{W}_i^{\star} \mathbf{B}_i \mathbf{X}_i)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{B}_i' \mathbf{W}_i^{\star} \mathbf{A}_i \mathbf{X_i} \boldsymbol{\beta} \quad . \end{aligned}$$

$$
\begin{aligned}
var(\hat{\boldsymbol{\beta}}^{\star}) &= \left(\sum_{i=1}^{n} \mathbf{X}_i^{\star'}\mathbf{W}_i^{\star}\mathbf{X}_i^{\star}\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{X}_i^{\star'}\mathbf{W}_i^{\star}\mathrm{Var}(\mathbf{Y}_i^{\star})\mathbf{W}_i^{\star}\mathbf{X}_i^{\star}\right)\left(\sum_{i=1}^{n} \mathbf{X}_i^{\star'}\mathbf{W}_i^{\star}\mathbf{X}_i^{\star}\right)^{-1} \\
&= \left(\sum_{i=1}^{n} \mathbf{X}_i^{\star'}\mathbf{W}_i^{\star}\mathbf{X}_i^{\star}\right)^{-1} = \left(\sum_{i=1}^{n} (\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^{\star}(\mathbf{B}_i\mathbf{X}_i)\right)^{-1} \\
&= \left(\sum_{i=1}^{n} (\mathbf{B}_i\mathbf{X}_i)'(\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i')^{-1}(\mathbf{B}_i\mathbf{X}_i)\right)^{-1} \quad ,
\end{aligned}
$$

under the assumption that $var(\mathbf{Y}_i^{\star}) = (\mathbf{W}_i^{\star})^{-1} = var(\mathbf{A}_i\mathbf{Y}_i) = \mathbf{A}_i\mathbf{V}_i\mathbf{A}_i'$. Since $\mathbf{A}_i \neq \mathbf{B}_i$, thus most likely $E(\hat{\boldsymbol{\beta}}^{\star}) \neq \boldsymbol{\beta}$, meaning the binning method produces a biased estimate for $\boldsymbol{\beta}$ in general. This is consistent with the simulation findings from Chapter 2. And in Chapter 2, simulations have shown that the $\hat{\boldsymbol{\beta}}^{\star}$ is biased toward 0, which indicates that $\hat{\boldsymbol{\beta}}^{\star}$ is not a consistent estimator of $\boldsymbol{\beta}$.

3.2.3 Examples

To better understand the mismatched case, we use a simple example as follows. In this example, we assume bin size is b and all subjects have equal number of bins, i.e., $n_i^{(X,Y)} = M$ for all i. In each bin, assume the data structures are the same for each subject, such that there are L matched pairs, and J response-only data points, and K predictor-only data points. Assume there is one single covariate X. Table (3.1) shows a clearer picture of the observed data for the m^{th} bin. We denote the GSD data before binning as $X_{i,mk}$ and $Y_{i,mk}$ which means the k^{th} observations for X and Y in the m^{th} bin for subject i, and $i = 1, \ldots, n$. We also arrange the original observed data so that in each bin, the data is shown as L time points of matched pairs $(X_{i,mk}, Y_{i,mk})$, J time points of Y-only data $(\cdot, Y_{i,mk})$ and K time points of X-only data $(X_{i,mk}, \cdot)$. The $(\cdot)$ signifies there is no observed corresponding predictor or response. There are a total of (L+J+K) original days on which we have observed data in each bin. The special cases of equal number of observations within each bin and equal number of observations for each subject are just for illustration

only. Binning method itself allows for variable number of observations within each bin and for each subject.

Let GSD data $\mathbf{Y}_i = (Y_{i,m1}, \ldots, Y_{i,m,L+J+K})^{'}$, and $\mathbf{X}_i = (\mathbf{1}^{'}_{L+J+K}, (X_{i,m1}, \ldots, X_{i,m,L+J+K})^{'})$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{'}$, $\boldsymbol{\epsilon}_i = (\epsilon_{i,m1}, \ldots, \epsilon_{i,m,L+J+K})^{'}$. $\boldsymbol{\epsilon}_i \sim N(0, \mathbf{V}_i)$. $\mathbf{W}_i = \mathbf{V}_i^{-1}$. And the model for GSD data by matrix form is

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \quad .$$

And we have the expectation

$$E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta} \quad .$$

After binning, the binned response and predictors for the m^{th} bin are $Y_{i,m}$ and $X_{i,m}$:

$$\begin{aligned} Y_{i,m} &= \frac{Y_{i,m1} + Y_{i,m2} + \ldots + Y_{i,mL} + Y_{i,m,L+1} + \ldots + Y_{i,m,L+J}}{L+J} \quad . \\ X_{i,m} &= \frac{X_{i,m1} + X_{i,m2} + \ldots + X_{i,mL} + X_{i,m,L+J+1} + \ldots + X_{i,m,L+J+L}}{L+K} \quad . \end{aligned}$$

Now, after binning, we need to fit the binned data by model

$$\mathbf{Y}_i^{\star} = \mathbf{X}_i^{\star}\boldsymbol{\beta}^{\star} + \boldsymbol{\epsilon}_i^{\star}$$

where $\boldsymbol{\epsilon}_i^{\star} \sim N(0, \mathbf{V}_i^{\star})$, $\mathbf{W}_i^{\star} = (\mathbf{V}_i^{\star})^{-1}$. Now we let

$$\begin{aligned} Z1_{i,m} &= \frac{X_{i,m1} + X_{i,m2} + \ldots + X_{i,mL}}{L} \\ Z2_{i,m} &= \frac{X_{i,m,L+1} + X_{i,m,L+2} + \ldots + X_{i,m,L+J}}{J} \\ Z3_{i,m} &= \frac{X_{i,m,L+J+1} + X_{i,m,L+J+2} + \ldots + X_{i,m,L+J+K}}{K} \quad . \end{aligned}$$

$Z1_{i,m}$ is the averaged covariate value for matched pairs in the m^{th} bin. $Z2_{i,m}$ is the averaged covariate values for $Y - only$ data in the m^{th} bin. $Z3_{i,m}$ is the averaged covariate values for $X - only$ data in the m^{th} bin. Then we will have

$$X_{i,m} = \frac{L}{L+K}\mathbf{Z1}_{i,m} + \frac{K}{L+K}\mathbf{Z3}_{i,m} \quad .$$

	type	observed data	
m^{th}'s bin $(X_{i,m}, Y_{i,m})$	matched pairs	$(X_{i,m1}, Y_{i,m1})$ $(X_{i,m2}, Y_{i,m2})$ $(X_{i,mk}, Y_{i,mk})$ $\ldots$ $(X_{i,mL}, Y_{mL})$ $k = 1, \ldots, L$	$Z1_{i,m}$
	Y-only	$(\cdot, Y_{i,m,L+1})$ $(\cdot, Y_{i,m,L+2})$ $(\cdot, Y_{i,mk})$ $\ldots$ $(\cdot, Y_{i,m,L+J})$ $k = L+1, \ldots, L+J$	$Z2_{i,m}$
	X-only	$(X_{i,m,L+J+1}, \cdot)$ $(X_{i,m,L+J+2}, \cdot)$ $(X_{i,mk}, \cdot)$ $\ldots$ $(X_{i,m,L+J+K}, \cdot)$ $k = L+J+1, \ldots, L+J+K$	$Z3_{i,m}$

Table 3.1: The data in the m^{th} bin

And let

$$\begin{aligned}
\mathbf{X}_i^\star &= (\mathbf{1}_M^{'}, (X_{i,1}, X_{i,2}, \ldots, X_{i,M})^{'}) \ , \\
\mathbf{Y}_i^\star &= (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,M})^{'} \ , \\
\mathbf{Z1}_i &= (\mathbf{1}_M^{'}, (Z1_{i,1}, Z1_{i,2}, \ldots, Z1_{i,M})^{'}) \ , \\
\mathbf{Z2}_i &= (\mathbf{1}_M^{'}, (Z2_{i,1}, Z2_{i,2}, \ldots, Z2_{i,M})^{'}) \ , \\
\mathbf{Z3}_i &= (\mathbf{1}_M^{'}, (Z3_{i,1}, Z2_{i,2}, \ldots, Z3_{i,M})^{'}) \ .
\end{aligned}$$

where $\mathbf{1}_M$ denotes a column vector with all elements of 1, and the length of the column vector is M. Then

$$\mathbf{X}_i^\star = \frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i \ , \text{ and}$$

$$\begin{aligned}
E(Y_{i,m}) &= E(\frac{Y_{i,m1} + Y_{i,m2} + \ldots + Y_{i,mL} + Y_{i,m,L+1} + \ldots + Y_{i,m,L+J}}{L+J}) \\
&= \frac{E(Y_{i,m1}) + E(Y_{i,m2}) + \ldots + E(Y_{i,mL}) + E(Y_{i,m,L+1}) + \ldots + E(Y_{i,m,L+J})}{L+J} \\
&= \frac{X_{i,m1} + X_{i,m2} + \ldots + X_{i,mL} + X_{i,m,L+1} + \ldots + X_{i,m,L+J}}{L+J}\beta_1 + \beta_0 \\
&= \frac{LZ1_{i,m} + JZ2_{i,m}}{L+J}\beta_1 + \beta_0 \ .
\end{aligned}$$

If we write in matrix form, then

$$E(\mathbf{Y}_i^\star) = (\frac{L}{L+J}\mathbf{Z1}_i + \frac{J}{L+J}\mathbf{Z2}_i)\boldsymbol{\beta} \ .$$

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}^{\star} &= \left(\sum_{i=1}^{n} \mathbf{X}_i^{\star'} \mathbf{W}_i^{\star} \mathbf{X}_i^{\star}\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^{\star'} \mathbf{W}_i^{\star} \mathbf{Y}_i^{\star} \quad . \\
E(\hat{\boldsymbol{\beta}}^{\star}) &= \left(\sum_{i=1}^{n} \mathbf{X}_i^{\star'} \mathbf{W}_i^{\star} \mathbf{X}_i^{\star}\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^{\star'} \mathbf{W}_i^{\star} E(\mathbf{Y}_i^{\star}) \\
&= \left(\sum_{i=1}^{n} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)' \mathbf{W}_i^{\star} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)\right)^{-1} \\
&\quad \sum_{i=1}^{n} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)' \mathbf{W}_i^{\star} (\frac{L}{L+J}\mathbf{Z1}_i + \frac{J}{L+J}\mathbf{Z2}_i)\boldsymbol{\beta} \\
&= \frac{L+K}{L+J}\boldsymbol{\beta} + \left(\sum_{i=1}^{n} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)' \mathbf{W}_i^{\star} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)\right)^{-1} \\
&\quad \sum_{i=1}^{n} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)' \mathbf{W}_i^{\star} (\frac{J}{L+J}\mathbf{Z2}_i - \frac{K}{L+J}\mathbf{Z3}_i)\boldsymbol{\beta} \quad .
\end{aligned}
$$

$$
\begin{aligned}
var(\hat{\boldsymbol{\beta}}^{\star}) &= \left(\sum_{i=1}^{n} \mathbf{X}_i^{\star'} \mathbf{W}_i^{\star} \mathbf{X}_i^{\star}\right)^{-1} \\
&= \left(\sum_{i=1}^{n} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)' \mathbf{W}_i^{\star} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)\right)^{-1} \quad .
\end{aligned}
$$

1) If $K = 0$, which means in each bin, there are some days that covariates are not measured, but responses are all measured. Then

$$
E(\hat{\boldsymbol{\beta}}^{\star}) = \frac{L}{L+J}\boldsymbol{\beta} + \left(\sum_{i=1}^{n} \mathbf{Z1}_i' \mathbf{W}_i^{\star} \mathbf{Z1}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{Z1}_i' \mathbf{W}_i^{\star} \mathbf{Z2}_i \frac{J}{L+J}\boldsymbol{\beta} \quad .
$$

2) If $J = 0$, which means in each bin, there are some days that responses are not measured, but covariates are all measured, then

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}^{\star}) &= \left(\sum_{i=1}^{n} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)' \mathbf{W}_i^{\star} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)\right)^{-1} \\
&\quad \sum_{i=1}^{n} (\frac{L}{L+K}\mathbf{Z1}_i + \frac{K}{L+K}\mathbf{Z3}_i)' \mathbf{W}_i^{\star} \mathbf{Z1}_i \boldsymbol{\beta} \quad .
\end{aligned}
$$

3) If $L = 0$, which means in each bin, response and covariates are totally mismatched. There are no matched pairs in each bin at all. Then

$$E(\hat{\boldsymbol{\beta}}^{\star}) = \left(\sum_{i=1}^{n} \mathbf{Z3}_i^{'} \mathbf{W}_i^{\star} \mathbf{Z3}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{Z3}_i^{'} \mathbf{W}_i^{\star} \mathbf{Z2}_i \boldsymbol{\beta} \quad .$$

However, because $\mathbf{Z1}_i$, $\mathbf{Z3}_i$, and $\mathbf{Z3}_i$ are in matrix form, it is not easy to see explicitly the form of the expectation of $\hat{\boldsymbol{\beta}}^{\star}$. We can say the expectation may be affected by

- matched rate $\frac{L}{L+J}$ for response and $\frac{L}{L+K}$ for covariate.
- correlation between $Z1_{i,m}$, $Z2_{i,m}$ and $Z3_{i,m}$. This means the correlation structure of each covariate is important.

We use a special case $\mathbf{W}_i^{\star} = \frac{1}{\sigma^2} I_{n_i}$ as an example to explore the expectation of $\hat{\boldsymbol{\beta}}^{\star}$. This form of W_i means there is no dependence between repeated measurements for each subject, which is not generally realistic for longitudinal data but is used here for illustrative purposes. This also corresponds to the independent covariance structure for the response.

1. Assume $\mathbf{W}_i^{\star} = \frac{1}{\sigma^2} I_{n_i}$

   If $\mathbf{W}_i^{\star} = \frac{1}{\sigma^2} I_{n_i}$, then observations within each subject are independent, then GLS becomes ordinary least, and the longitudinal modeling becomes simple linear regression. Then, we can write estimated $\boldsymbol{\beta}^{\star}$ as $\hat{\boldsymbol{\beta}}^{\star}$

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\star} &= \left(\sum_{i=1}^{n} \mathbf{X}_i^{\star'} \mathbf{X}_i^{\star}\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^{\star'} \mathbf{Y}_i^{\star} \\ cov(\hat{\boldsymbol{\beta}}^{\star}) &= \left(\sum_{i=1}^{n} \mathbf{X}_i^{\star'} \mathbf{X}_i^{\star}\right)^{-1} \sigma^2 \end{aligned}$$

Since we assume there is only one single covariate X_{ij}, and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ thus we can write

$$\begin{aligned}
Y_{i,m} &= \beta_0^\star + \beta_1^\star X_{i,m} + \epsilon_{i,m}^\star \\
\hat{\beta}_0^\star &= \bar{Y}_{\cdot\cdot}^\star - \hat{\beta}_1^\star \bar{X}_{\cdot\cdot}^\star \\
\hat{\beta}_1^\star &= \frac{\sum_{i=1}^n \sum_{m=1}^M (X_{i,m} - \bar{X}_{\cdot\cdot}^\star) Y_{i,m}}{\sum_{i=1}^n \sum_{m=1}^M (X_{i,m} - \bar{X}_{\cdot\cdot}^\star)^2} \\
var(\hat{\beta}_1^\star) &= \frac{\sigma^2}{\sum_{i=1}^n \sum_{m=1}^M (X_{i,m} - \bar{X}_{\cdot\cdot}^\star)^2} \quad , \\
\text{where} \qquad & \bar{X}_{\cdot\cdot}^\star = \frac{\sum_{i=1}^n \sum_{m=1}^M X_{i,m}}{nM}, \bar{Y}_{\cdot\cdot}^\star = \frac{\sum_{i=1}^n \sum_{m=1}^M Y_{i,m}}{nM} \quad .
\end{aligned}$$

Since

$$E(\hat{\beta}_1^\star) = \frac{\sum_{i=1}^n \sum_{m=1}^M (\frac{L}{L+K} Z1_{i,m} + \frac{K}{L+K} Z3_{i,m} - \bar{X}_{\cdot\cdot}^\star)(\frac{L}{L+J} Z1_{i,m} + \frac{J}{L+J} Z2_{i,m})}{\sum_{i=1}^n \sum_{m=1}^M (\frac{L}{L+K} Z1_{i,m} + \frac{K}{L+K} Z3_{i,m} - \bar{X}_{\cdot\cdot}^\star)^2} \beta_1 \quad .$$

Let

$$\bar{Z1}_{\cdot\cdot}^\star = \frac{\sum_{i=1}^n \sum_{m=1}^M Z1_{i,m}}{nM}, \ \bar{Z3}_{\cdot\cdot}^\star = \frac{\sum_{i=1}^n \sum_{m=1}^M Z3_{i,m}}{nM}$$

. From the setup, we can derive that

$$\frac{L}{L+K} \bar{Z1}_{\cdot\cdot} + \frac{K}{L+K} \bar{Z3}_{\cdot\cdot} = \bar{X}_{\cdot\cdot}^\star \quad .$$

Let $E(\hat{\beta}_1^\star) = \frac{T1}{T2} \beta_1$ where

$$\begin{aligned}
T1 &= \sum_{i=1}^n \sum_{m=1}^M (\frac{L}{L+K} Z1_{i,m} + \frac{K}{L+K} Z3_{i,m} - \bar{X}_{\cdot\cdot}^\star)(\frac{L}{L+J} Z1_{i,m} + \frac{J}{L+J} Z2_{i,m}) \\
&\quad \sum_{i=1}^n \sum_{m=1}^M ((Z1_{i,m} - \bar{Z1}_{\cdot\cdot}) \frac{L}{L+K} + (Z3_{i,m} - \bar{Z3}_{\cdot\cdot}) \frac{K}{L+K})(\frac{L}{L+J} Z1_{i,m} + \frac{J}{L+J} Z2_{i,m}) \\
&= var(Z1) \frac{L^2}{(L+K)(L+J)} + cov(Z1, Z3) \frac{KL}{(L+K)(L+J)} \\
&+ cov(Z1, Z2) \frac{LJ}{(L+K)(L+J)} + cov(Z2, Z3) \frac{KJ}{(L+K)(L+J)} \quad .
\end{aligned}$$

$$
\begin{aligned}
T2 &= \sum_{i=1}^{n}\sum_{m=1}^{M}(\frac{L}{L+K}Z1_{i,m}+\frac{K}{L+K}Z3_{i,m}-\bar{X}^{\star}_{..})^2 \\
&= \sum_{i=1}^{n}\sum_{m=1}^{M}(\frac{L}{L+K}Z1_{i,m}+\frac{K}{L+K}Z3_{i,m}-\frac{L}{L+K}\bar{Z}1_{..}-\frac{K}{L+K}\bar{Z}3_{..})^2 \\
&= \sum_{i=1}^{n}\sum_{m=1}^{M}((Z1_{i,m}-\bar{Z}1_{..})\frac{L}{L+K}+(Z3_{i,m}-\bar{Z}3_{..})\frac{K}{L+K})^2 \\
&= \frac{L^2(nM-1)}{(L+K)^2}var(Z1)+\frac{K^2(nM-1)}{(L+K)^2}var(Z3)+\frac{2KL(nM-1)}{(L+K)^2}cov(Z1,Z3) \quad .
\end{aligned}
$$

Thus

$$
\begin{aligned}
E(\hat{\beta}^{\star}_1) &= \frac{T1}{T2}\beta_1=\beta_1\frac{var(Z1)\frac{L^2}{(L+K)(L+J)}+cov(Z1,Z3)\frac{KL}{(L+K)(L+J)}}{\frac{L^2}{(L+K)^2}var(Z1)+\frac{K^2}{(L+K)^2}var(Z3)+\frac{KL}{(L+K)^2}cov(Z1,Z3)} \\
&+ \frac{cov(Z1,Z2)\frac{LJ}{(L+K)(L+J)}+cov(Z2,Z3)\frac{KJ}{(L+K)(L+J)}}{\frac{L^2}{(L+K)^2}var(Z1)+\frac{K^2}{(L+K)^2}var(Z3)+\frac{KL}{(L+K)^2}cov(Z1,Z3)} \quad . \qquad (3.5) \\
var(\hat{\beta}^{\star}_1) &= \frac{\sigma^2}{T2}=\frac{\sigma^2/(nM-1)}{\frac{L^2}{(L+K)^2}var(Z1)+\frac{K^2}{(L+K)^2}var(Z3)+\frac{KL}{(L+K)^2}cov(Z1,Z3)} \quad . (3.6)
\end{aligned}
$$

- if $K=0$, then

$$
\begin{aligned}
E(\hat{\beta}^{\star}_1) &= \beta_1\frac{var(Z1)\frac{L}{L+J}+cov(Z1,Z2)\frac{J}{L+J}}{var(Z1)}=\frac{L}{L+J}\beta_1+\frac{cov(Z1,Z2)}{var(Z1)}\frac{J}{L+J}\beta_1 \\
var(\hat{\beta}^{\star}_1) &= \frac{\sigma^2}{(nM-1)var(Z1)} \quad ;
\end{aligned}
$$

  positive correlation between $Z1$, $Z2$ would increase the expectation of $\hat{\beta}^{\star}_1$.

- if $J=0$, then

$$
E(\hat{\beta}^{\star}_1)=\frac{var(Z1)\frac{L}{L+K}+cov(Z1,Z3)\frac{K}{L+K}}{\frac{L^2}{(L+K)^2}var(Z1)+\frac{K^2}{(L+K)^2}var(Z3)+\frac{KL}{(L+K)^2}cov(Z1,Z3)} \quad .
$$

$$
var(\hat{\beta}^{\star}_1)=\frac{\sigma^2/(nM-1)}{\frac{L^2}{(L+K)^2}var(Z1)+\frac{K^2}{(L+K)^2}var(Z3)+\frac{KL}{(L+K)^2}cov(Z1,Z3)} \quad .
$$

Summary: from equation (3.5) and (3.6), we can see that the $E(\hat{\beta}^{\star}_1)$, $var(\hat{\beta}^{\star}_1)$ are affected by

- The matching rate $\frac{L}{L+K}$ for X and for $\frac{L}{L+J}$ for Y.
- The sample covariance among $Z1$, $Z2$, $Z3$, which are the binned covariates values for matched data, Y-only data and $X-only$ data.
- The sample variance of $Z1$, $Z3$.
- Fixing all other conditions, if there is positive autocorrelation among the predictor values $X_{i,m}$, then $cov(Z1, Z2) > 0$, $cov(Z1, Z3) > 0$, and $cov(Z2, Z3) > 0$. Thus the $E(\hat{\beta}^{\star})$ would be increased compared to the case with no autocorrelation among the predictors.
- σ^2, the variance of the random error term.

2. Assume $\mathbf{W}_i^{\star}$ is a compound symmetry matrix

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}^{\star}) &= (\sum_{i=1}^{n}(\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^{\star}(\mathbf{B}_i\mathbf{X}_i))^{-1}\sum_{i=1}^{n}(\mathbf{B}_i\mathbf{X}_i)'\mathbf{W}_i^{\star}\mathbf{A}_i\mathbf{X_i}\boldsymbol{\beta} \\ var(\hat{\boldsymbol{\beta}}^{\star}) &= \left(\sum_{i=1}^{n}\mathbf{X}_i^{\star\prime}\mathbf{W}_i^{\star}\mathbf{X}_i^{\star}\right)^{-1} = \left(\sum_{i=1}^{n}(\mathbf{B}_i\mathbf{X}_i)'(\mathbf{A}_i\mathbf{V}_i\mathbf{A}_i')^{-1}(\mathbf{B}_i\mathbf{X}_i)\right)^{-1} . \end{aligned}$$

$\mathbf{W}_i^{\star}$ is a compound symmetry matrix with correlation matrix ρ and variance $\sigma_b^2+\sigma_e^2$

$$var(\hat{\boldsymbol{\beta}}^{\star}) = \frac{1}{\Delta_1}\begin{pmatrix} \frac{1}{nM}\left((1-\rho_1')\sum_{i=1}^{n}\sum_{m=1}^{M}X_{i,m}^2 + M^2\rho_1'\sum_{i=1}^{n}\bar{X}_{i\cdot}^{\star}\right) & -(1-\rho_1'+M\rho_1')\bar{X}_{\cdot\cdot}^{\star} \\ -(1-\rho_1'+M\rho_1')\bar{X}_{\cdot\cdot}^{\star} & 1-\rho_1'+M\rho_1' \end{pmatrix},$$

where $\Delta_1 = (1-\rho_1'+M\rho_1')a_1\left\{(1-\rho_1')\sum_{i=1}^{n}\sum_{m=1}^{M}(X_{i,m}-\bar{X}_{\cdot\cdot}^{\star})^2 + \rho_1'M^2\sum_{i=1}^{n}(\bar{X}_{i\cdot}^{\star}-\bar{X}_{\cdot\cdot}^{\star})^2\right\}$.

And $\bar{X}_{i\cdot}^{\star} = \frac{1}{M}\sum_{m=1}^{M}X_{i,m}$, $\bar{X}_{\cdot\cdot}^{\star} = \frac{1}{nM}\sum_{i=1}^{n}\sum_{m=1}^{M}X_{i,m}$, $\rho_1 = \frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2}$, $\rho_1' = \frac{-\rho_1}{1+(M-2)\rho_1}$, $a_1 = \frac{1+(M-2)\rho_1}{(1+(M-2)\rho_1-(M-1)\rho_1^2)(\sigma_b^2+\sigma_e^2)}$. Here $X_{i,m}$ is the covariate value after binning.

3.3 Conditional Poisson distribution

From the previous section, if the original observed data is completely matched (same as the gold standard data), then the expectation of estimated $\hat{\boldsymbol{\beta}}^{\star}$ for the binned data, i.e., $E(\hat{\boldsymbol{\beta}}^{\star})$, is the same as $\boldsymbol{\beta}$. This is true for multivariate normal responses. However, for non-normal data, such as conditional Poisson data, we would like to know the relationship between the estimated $\boldsymbol{\beta}^{\star}$ for the binned data and the $\boldsymbol{\beta}$ of the original gold standard data. Here we mainly investigate a simple random intercept model because of its simplicity and wide usage.

Assume Y_{ij} is the response for subject i and time j. For simplicity, suppose we have a random intercept conditional Poisson model for Y_{ij}, i.e., conditional on the random intercept u_i, Y_{ij} is distributed as Poisson(μ_{ij}). Assume there are k such original observations in the m^{th} bin. After binning, then the binned response $Y_{i,m}$ and covariate $X_{i,m}$ are

$$\begin{aligned} Y_{i,m} &= Y_{i1} + Y_{i2} + \ldots + Y_{ik} \ , \\ X_{i,m} &= \frac{X_{i1} + X_{i2} + \ldots + X_{ik}}{k} \ . \end{aligned}$$

By conditional Poisson GLMMs, we have

$$\begin{cases} \log(\mu_{i1}) &= \beta_0 + X_{i1}\beta_1 + u_i \ , \\ \log(\mu_{i2}) &= \beta_0 + X_{i2}\beta_1 + u_i \ , \\ & \vdots \ , \\ \log(\mu_{ik}) &= \beta_0 + X_{ik}\beta_1 + u_i \ . \end{cases}$$

Summing up the left hand-side and right hand-side of the above equations, and dividing each side by k, we have

$$1/k \ \log(\mu_{i1}\mu_{i2}\ldots\mu_{ik}) = \beta_0 + (X_{i1} + X_{i2} + \ldots + X_{ik})\beta_1/k + u_i = \beta_0 + X_{i,m}\beta_1 + u_i \ .$$

On the other hand, conditional on random effects u_i, Y_{i1}, Y_{i2}, ..., Y_{ik} are independent,

thus $Y_{i1} + Y_{i2} + \ldots + Y_{ik}|u_i$ are Poisson$(\mu_{i1} + \mu_{i2} + \ldots + \mu_{ik})$. That is

$$\begin{aligned} \log(\mu_{i1} + \ldots + \mu_{ik}) &= \log(e^{\beta_0 + X_{i1}\beta_1 + u_i} + e^{\beta_0 + X_{i2}\beta_1 + u_i} + \ldots + e^{\beta_0 + X_{ik}\beta_1 + u_i}) \\ &= \log(e^{X_{i1}\beta_1} + e^{X_{i2}\beta_1} + \ldots + e^{X_{ik}\beta_1}) + \beta_0 + u_i \quad . \end{aligned}$$

If $X_{i1} = X_{i2} = \ldots = X_{ik} = c$, i.e., for subject i, X_{ij} is time-invariant within each bin but can be the same or different between bins, then $\log((\mu_{i1} + \ldots + \mu_{ik})/k) = 1/k \log(\mu_{i1}\mu_{i2} \ldots \mu_{ik})$, thus $\log(\mu_{i1} + \ldots + \mu_{ik}) = \beta_0 + \beta_1 X_{i,m} + u_i$. So

$$Y_{i,m}|u_i \sim \text{Poisson}(\beta_0 + X_{i,m}\beta_1 + log(k) + u_i) \quad .$$

Under these conditions, then after binning, β_1 remains the same but there is an offset $log(k)$ in the intercept.

However, if X_{ij} is time-varying, but the within-subject variability of X_{ij} is sufficiently small, then within each bin, the X_{ij} would be very close to a constant. Then after binning, the conditional distribution of the binned response is still approximately Poisson.

We would like to establish a more general proof as future work, i.e., to establish the conditional distribution of $Y_{i,m}$ when X_{ij} is fully time-varying and to establish precisely what we mean by the within-subject variability of X_{ij} being sufficiently small. We have established in numerical studies that X_{ij} can be time-varying with a constant mean and the conditional distribution of $Y_{i,m}$ follows.

3.4 Bin size selection

In Chapter 2, we have discussed the potential important of bin size selection as well as having investigated the role of bin size influence on parameter estimation via simulation. We discuss this topic a bit further in this section.

3.4.1 Cross validation using binned data

Cross-validation can be used for bin size selection. Similar ideas have been used in non-parametric smoothing to choose smoothing parameters. For longitudinal data, we consider leave-one-subject-out cross validation instead of leaving-one-observation-out, the former also a common choice in bandwidth selection in nonparametric modeling of longitudinal data (e.g. Rice and Silverman, 1991).

In this approach, we use the binned data for cross validation. The loss function is defined as predicted residual sum of squares PRSS $= \sum_{i=1}^{n} \sum_{m=1}^{n_i^{(\mathbf{X},\mathbf{Y})}} (y_{i,m} - \hat{E}_{(-i)}(Y_{i,m}|\mathbf{u}_i))^2$. $Y_{i,m}$ is the binned response in the m^{th} bin for the i^{th} subject. $\hat{E}_{(-i)}(Y_{i,m}|\mathbf{u}_i)$ is the predicted conditional mean for the i^{th} subject in the m^{th} bin, based on estimates from the dataset without subject i. The best bin size is the one with minimal PRSS. To calculate $\hat{E}_{(-i)}(Y_{i,m}|\mathbf{u}_i)$, we need to predict the random effects of subject i based on the data without subject i. We can use the linearization method to predict the random effects.

For a generalized linear mixed effects model, conditional on the random effects, assume the mean of Y_{ij} is μ_{ij}. And with a known link function $g(\cdot)$, $g(\mu_{ij}) = g[E(Y_{ij}|\mathbf{u}_i)] = \mathbf{X}_{ij}'\boldsymbol{\beta} + \mathbf{Z}_{ij}'\mathbf{u}_i$. Here $\mathbf{X}_{ij}$ is the p-dimensional vector and $\mathbf{Z}_{ij}$ is the q-dimensional vector for fixed effects $\boldsymbol{\beta}$ and random effects $\mathbf{u}_i$ respectively where $\mathbf{u}_i \sim N(0, \mathbf{D})$

Let $v_{ij} = \mathrm{Var}(Y_{ij}|\mathbf{u}_i)$, Q_i be a diagonal matrix where $Q_i = \mathrm{diag}(v_{ij}g'(\mu_{ij})^2)$, and let Y_{ij}^* be a surrogate response whose elements are

$$Y_{ij}^* = g(\mu_{ij}) + (Y_{ij} - \mu_{ij})g'(\mu_{ij}) \quad .$$

$\mathbf{V}_i$ is a $n_i \times n_i$ matrix and defined as $\mathbf{V}_i = \mathbf{Q}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'$. The prediction of random effects is (Diggle et al., 2002, p.174)

$$\hat{\mathbf{u}}_i = \hat{\mathbf{D}}\mathbf{Z}_i'\hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i^* - \mathbf{X}_i\hat{\boldsymbol{\beta}}) \tag{3.7}$$

For the hemodialysis data, the predicted PRSS is 213.9695 for bin size 30, 218.7389 for bin size 45 and 218.9826 for bin size 60, which shows 30 day bin size would be selected.

We had run 200 simulations for complete matched data using bin size 28, 40 and 56. By the PRSS criteria, 59.5% of the time, bin size 28 is selected, while 33% for bin size 40 and 7.5% for bin size 56. Similarly, for complete mismatched data by bin size 28, 40 and 56, 52% of the time bin size 28 is selected, and 32% for bin size 40 and 16% for bin size 56.

3.4.2 Cross validation using original data before binning

In the above section, we have used binned data for bin size selection. However, binned data would be different once the bin size is changed. On the other hand, original observed data before binning is always the same no matter what bin size we use to perform binning. It is also desirable to use original data to conduct bin size selection instead of using binned data only. In this section, we will consider bin size selection by incorporating original data.

First approach

Let Y_{ij} denotes the observed response for subject i and time j where $j = 1, 2, \ldots, n_i^{(Y)}$. The PRSS is defined as $\sum_{i=1}^{n} \sum_{j=1}^{n_i^{(\mathbf{Y})}} (Y_{ij} - \hat{E}_{(-i)}(Y_{ij}|\mathbf{u}_i))^2$. We need to predict the conditional mean response $\hat{E}_{(-i)}(Y_{ij}|\mathbf{u}_i)$ for subject i by the data which exclude subject i. However, we fit models only on the binned data because the misalignment of the original response and covariates on time. Here we suggest calculating $\hat{E}_{(-i)}(Y_{ij}|\mathbf{u}_i)$ by the binned data. One possible way is to use the average predicted conditional mean response in the bin where $Y_{i,j}$ is located. Suppose original response Y_{ij} is located at the m^{th} bin, then we use $\frac{\hat{E}_{(-i)}(Y_{i,m}|\mathbf{u}_i)}{c_{im}^{(\mathbf{Y})}}$ for calculating $\hat{E}_{(-i)}(Y_{ij}|\mathbf{u}_i)$. Here $c_{im}^{(\mathbf{Y})}$ is the number of original responses in the m^{th} bin for subject i. For the hemodialysis data, the PRSS is 92.5235, 90.9055 and 92.3674 for bin size 30, 45 and 60 respectively which indicates bin size 45 will be selected.

Second approach

In the first approach, we try to calculate the $\hat{E}_{(-i)}(Y_{ij}|\mathbf{u}_i)$ by $\hat{E}_{(-i)}(Y_{i,m}|\mathbf{u}_i)/c_{im}^{(\mathbf{Y})}$. In this second approach, we propose to use marginal mean $\hat{E}_{(-i)}(Y_{ij})$ instead of the conditional mean $\hat{E}_{(-i)}(Y_{ij}|\mathbf{u}_i)$ to calculate the PRSS, which is defined as $\sum_{i=1}^{n}\sum_{j=1}^{n_i^{(\mathbf{Y})}}(Y_{ij}-\hat{E}_{(-i)}(Y_{ij}))^2$.

If the random effects structure of the conditional Poisson GLMMs contains only a single normal random intercept, then the marginal mean of $Y_{i,m}$ can be calculated explicitly. Assume u_i is distributed as $N(0,\sigma_u^2)$. By GLMM setting, we have

$$E(Y_{i,m}|u_i)=\mu_{i,m},\quad \log(\mu_{i,m})=\mathbf{X}_{i,m}^{'}\boldsymbol{\beta}+u_i \quad .$$

The marginal mean of $Y_{i,m}$ is calculated as

$$\gamma_{i,m}=E(Y_{i,m})=E(E(Y_{i,m}|u_i))=E(e^{\mathbf{X}_{i,m}^{'}\boldsymbol{\beta}+u_i})=e^{\mathbf{X}_{i,m}^{'}\boldsymbol{\beta}}E(e^{u_i})=e^{\mathbf{X}_{i,m}^{'}\boldsymbol{\beta}+\sigma_u^2/2} \quad .$$

However, this second approach is limited to the conditional intercept model. Assume Y_{ij} is located in the m^{th} bin, $c_{im}^{(\mathbf{Y})}$ is the number of original responses in the m^{th} bin for subject i. We suggest using $\frac{\gamma_{i,m}}{c_{im}^{(\mathbf{Y})}}$ for calculating $\hat{E}_{(-i)}(Y_{ij})$. We choose the bin size which minimizes the $PRSS$. For the hemodialysis data, the $PRSS$ values are 116.1346, 120.8958, 130.9015 for bin size 30, 45 and 60 days respectively, which indicates bin size 30 will be selected.

These two approaches using original data have good potential because we compare the $PRSS$ criteria based on the same original response. However, one limitation is that the response and covariates for original data are misaligned on time. Thus we cannot model the original data directly but have to do some pre-processing such as the proposed binning method. Thus for calculating the $PRSS$ criteria for the original data, we still have to borrow the binned data.

For hemodialysis among bin size 30, 45 and 60 days, approach 1 gives 45 days as the best bin sizes, while approach 2 suggests bin size 30 days as the best. So some future work is still required to choose between the two proposals using the original data.

3.4.3 Summary of bin size selection

In this section, we have proposed two methods for bin size selection, using either binned data or original data. The selection criteria is PRSS, the predicted residual sum squares. For bin size selection using original data, we also proposed two different approaches to construct the PRSS criteria. In total, we have three different PRSS criteria. When using original data for bin size selection, in order to build the PRSS, we still have to rely on the marginal/conditional mean of binned data and number of original responses within each bin (i.e., $c_{im}^{(\mathbf{Y})}$). We applied these three PRSS criteria into hemodialysis data. Two criteria selected bin size 30 and the other one selected bin size 45.

The reason to use original response for bin size selection is that as bin size changes, the binned data also change. Thus comparison of PRSS criteria for different bin sizes are based on different binned responses. However, when choosing smoothing parameter in nonparametric smoothing, comparison using cross-validation criteria are always based on the same responses. So it is worth considering original data in bin size selection, since our original responses will never change no matter which bin size we have applied.

It looks like bin size selection using binned data compared to using original data also has its advantages, since we have a model for the binned data (we fit the GLMM model on the binned data), but not for the original data. And the $PRSS$ criteria of original data still depends on the binned data, which uses either $\frac{\hat{E}_{(-i)}(Y_{i,m}|\mathbf{u}_i)}{c_{im}^{(\mathbf{Y})}}$ for $\hat{E}_{(-i)}(Y_{ij}|\mathbf{u}_i)$ or $\frac{\gamma_{i,m}}{c_{im}^{(\mathbf{Y})}}$ for $\hat{E}_{(-i)}(Y_{ij})$. However, we still need to do more investigation on choosing one approach versus the other.

In summary, further investigation is necessary to evaluate the $PRSS$ criteria for selecting bin size under the binned data modeling approach.

Chapter 4

A mixed effects regression tree-based method for longitudinal count data

4.1 Motivation

For the hemodialysis study described in Chapter 2, we established via the binning procedure that there was a contemporaneous relationship between log(crp) and infection for 30-day binned data. And there was also a lagged relationship between infection and log(crp) for the 7-week binned data; more specifically, the analysis indicated that infection occurred ahead of log(crp) change in the latter setting. This is an important biological finding. Now, from a biomedical standpoint, we are also interested in whether the rise in crp was obviously preceded by an infection event, and if such event could be well determined to have occurred in the recent past based on current readings of crp, a serum physiological measurement that is easy to collect. In fact, there may be noticeable cutoff levels for crp that are indicative of recent occurrences of infection event. For example, cut-off levels of crp can be used to predict potential cardiovascular disease. According to the guidelines of the American Heart Association (AHA) and the Center for Disease Control (CDC), risk

for cardiovascular disease is low if crp is 1mg/L or less; risk for cardiovascular disease is moderate if crp is between 1mg/L and 3mg/L; it is of high risk if crp $>$ 3mg/L. In this research, we are interested in finding out the cut-off levels of log(crp). It is also of interest to detect whether other covariates such as age, gender, ethnicity, BMI and etc. should be included in the model and what these covariate levels should be for best predicting number of event occurrences.

Optimally, we would like to know risk factors of a pending occurrence. It is easy to collect serum marker data, levels of which may result from a recent past event. In the hemodialysis study, chart records were available for each patient. However, in general, such chart records may be difficult to obtain, meaning only the marker of inflammation (here, crp) can be used (among the consideration of both a measure of "infection" and obtaining the crp measure itself) to help see if an infection may have recently occurred. If so, it is possible to monitor the patients for future events (e.g., additional infections, or possibly worse outcomes). Here there is a justification in identifying crp levels that may be a reflection of event occurrences in the recent past.

In order to answer this medical question, a tree-based method is proposed for longitudinal count data. To get the cut-off levels of log(crp), it is the same as finding the splitting point of log(crp) in the regression tree.

Breiman et al. (1984) introduced "Classification and regression trees" (CART) which is for univariate continuous response. Tree-based methods for longitudinal/clustering responses have also been proposed, such as Segal (1992). Segal (1992) suggested a tree that can analyze continuous longitudinal response using Mahalanobis distance for within-node homogeneity measures. Larsen and Speckman (2004) suggested a multivariate regression tree that is a variation of the approach of Segal (1992) for longitudinal continuous data. Zhang (1998) suggested a tree that can analyze multiple binary responses using generalized entropy criterion. Lee (2005) proposed generalized multivariate decision to support various

type of data such as count and binary using generalized estimating equations. Abdolell (2002), Sela and Simonoff (2009) are using random effects models to build tree. More specifically, Abdolell(2002) used linear mixed effects model for building splitting criterion. Sela and Simonoff (2009) proposed a random effects EM tree (REEMtree) method for longitudinal continuous data.

In summary, most of the existing classification and regression tree based methods are either dealing with uncorrelated continuous or discrete responses, or correlated binary data or count data but requiring the covariate value to be time-invariant. In this paper, we have to make the tree-based method work for time-varying covariate, and supporting longitudinal count data. Also we are more interested in subject-specific effects, so this tree-based model should also account for individual subject effects.

The organization of this Chapter is as follows. In Section 4.2, we will introduce the proposed generalized mixed regression tree-based method for longitudinal count data with time-varying covariates. Section 4.3 will present the results of an analysis on hemodialysis data and epilepsy seizure data. Some simulations are conducted in Section 4.4 to evaluate the proposed method. Some discussion and future work are provided on Section 4.5.

In order to answer the scientific question related with the cut-off levels for $\log(crp)$, a tree-based method is a reasonable choice. For the proposed GMRTree method in the following section, we will adopt the linearization idea of penalized-quasi likelihood (PQL) approximation in GLMMs and change the longitudinal count response into a surrogate continuous response. As the response is count, following binning, the bias of PQL (Breslow and Lin, 1995) should not be troubling as this affects mostly binary responses. Then, we estimate the tree by CART, and random effects by a generalized linear mixed effects model.

4.2 Model for generalized mixed effects regression tree (GMRTree)

Before we introduce the proposed method, we review some notations of recursive partition methodology for a univariate continuous response $\mathbf{Y}$ and a set of covariates $(X_1, X_2, \ldots, X_p)$. Suppose that $Y_i = f(X_{1,i}, X_{2,i}, \ldots, X_{p,i}) + e_i$. For recursive partition, the aim is to approximate f by a step function defined as

$$\hat{f}(X_i) = \sum_{v=1}^{V} a_v b_v(X_i) \quad .$$

The functions $b_v(X)$ takes the form

$$b_v(X) = I(X \in R_v) \quad ,$$

where $I(\cdot)$ is an indicator function, and $\{R_v\}_{v=1}^{V}$ represents a partition of the covariate space which include the variables to be split and the split point. After fitting with the regression tree, observations within the same region, i.e. partition R_v will have the same mean a_v.

Suppose Y_{ij} is the longitudinal response for subject i at time j, $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})^{'}$ is the corresponded covariate vector, and the q-dimensional random effects vector $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iq})^{'}$ where $i = 1, 2, \ldots, N; j = 1, 2, \ldots, n_i$. Conditional on the random effects $\mathbf{u}_i$, assume

$$\mathrm{g}(\mu_{ij}) = f(X_{ij1}, X_{ij2}, \ldots, X_{ijp}) + \mathbf{Z}_{ij}^{'}\mathbf{u}_i \quad . \tag{4.1}$$

where $E(Y_{ij}|\mathbf{u}_i) = \mu_{ij}$, $var(Y_{ij}|\mathbf{u}_i) = v(\mu_{ij})$. $\mathbf{u}_i$ is assumed to be distributed as $N_q(0, \mathbf{D})$. $\mathbf{Z}_{ij}$ is the covariate vector for the random effects. $g(\cdot)$ is a link function. For count data, most often $g(\cdot)$ is the *log* canonical function. If $f(\cdot)$ is a linear function of parameters such that $f(X_{ij1}, X_{ij2}, \ldots, X_{ijp}) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}$, then (4.1) is the form of a generalized linear mixed effects model. However, if different subsets of data have different

structures, then parametric form for $f(\cdot)$ might be too restrictive. Also, p may be large, so that including all of the predictors directly in parametric model may lead to overfitting and therefore poor predictions. In this case, we propose a nonparametric regression tree-based method estimate $f(\cdot)$. We use a piecewise constant model to fit $f(\cdot)$, such as by CART, and the model (4.1) is defined as

$$\log(\mu_{ij}) = \mathbf{Z}_{ij}^{'}\mathbf{u}_i + \sum_{v=1}^{V} I(\mathbf{X}_{ij} \in R_v)\beta_v \quad , \tag{4.2}$$

where R_v is the partition of the covariate space in the v^{th} terminal node. Here R_v and β_v are unknown. And the covariance matrix of $\mathbf{u}_i$ is also unknown.

For longitudinal continuous response, Sela and Simonoff (2009) have proposed REEMtree. For the tree-based model with random effects

$$Y_{ij} = f(X_{ij1}, X_{ij2}, \ldots, X_{ijp}) + \mathbf{Z}_{ij}^{'}\mathbf{u}_i + e_{ij} \quad ,$$

where $e_{i1}, e_{i2}, , \ldots, e_{in_i}$ are assumed to normally distributed with covariance matrix $\mathbf{R}_i$. And e_{ij} are independent across subjects. The algorithms to estimate $f(\cdot)$ and predict the random effects are as follows.

1. Initialize the predicted random effects $\hat{\mathbf{u}}_i$ to zero.

2. Iterate through the following steps until the predicted random effects, $\hat{\mathbf{u}}_i$ converges:

   (a) Estimate a regression tree approximating f, based on the target variable, $Y_{ij} - \mathbf{Z}_{ij}^{'}\hat{\mathbf{u}}_i$, and predictors, $X_{ij.} = (X_{ij1}, \ldots, X_{ijp})$, for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. Use this regression tree to create a set of indicator variables, $I(X_{ij.} \in g_v)$, where g_v ranges over all of the terminal nodes in the tree.

   (b) Fit the linear random effects model, $Y_{ij} = \mathbf{Z}_{ij}^{'}\mathbf{u}_i + I(X_{ij.} \in g_v)\mu_v + e_{ij}$. Extract $\hat{\mathbf{u}}_i$ from the estimated model.

For longitudinal non-normal data, a similar approach as REEMtree will be proposed to get $\hat{f}$ and $\hat{\mathbf{u}}_i$. However, here, we will use linearization to change the non-normal response into an approximated continuous normal response.

4.2.1 Estimation procedure

Assume $f(\cdot)$ is a linear function of parameters β_v, then (4.1) becomes the form of GLMMs

$$g(\mu_{ij}) = \mathbf{X}_{ij}^{'}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{'}\mathbf{u}_i \quad . \tag{4.3}$$

By linearization technique in penalized quasi-likelihood (Schall 1991; Breslow and Clayton, 1993), given the current estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}_i$, the surrogate response Y_{ij}^*, i.e.,

$$Y_{ij}^* \equiv g(\hat{\mu}_{ij}) + g^{'}(\hat{\mu}_{ij})(Y_{ij} - \hat{\mu}_{ij}) \approx \mathbf{X}_{ij}^{'}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{'}\mathbf{u}_i + e_{ij}^* \quad ,$$

approximately follows a linear mixed effects model with fixed effects $\boldsymbol{\beta}$ and random effects $\mathbf{u}_i$, with error term e_{ij}^*. e_{ij}^* is assumed to be normally distributed with mean 0 and diagonal covariance matrix $var(\mathbf{e}_i^*) = \text{diag}(g^{'}(\hat{\mu}_{ij})^2 var(Y_{ij}))$. Here $\mathbf{e}_i^* = (e_{i1}^*, \ldots, e_{in_i}^*)$.

Using similar idea, we derive the distribution of surrogate response Y_{ij}^* for GMRTree model in (4.1). Suppose $h(\cdot)$ is the inverse of link function $g(\cdot)$. Let $\eta_{ij} = g(\mu_{ij}) = f(X_{ij1}, X_{ij2}, \ldots, X_{ijp}) + \mathbf{Z}_{ij}^{'}\mathbf{u}_i$. Then we have $\mu_{ij} = h(\eta_{ij})$. Given current estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}_i$, Y_{ij} can be written as

$$Y_{ij} = \mu_{ij} + e_{ij} = h(\eta_{ij}) + e_{ij} \approx h(\hat{\eta}_{ij}) + h^{'}(\hat{\eta}_{ij})(\eta_{ij} - \hat{\eta}_{ij}) + e_{ij} \quad ,$$

by first order Taylor expansion, Y_{ij}. Here $var(e_{ij}) = v(\mu_{ij})$. Then reordering the above expression yields

$$Y_{ij}^* \quad \equiv \quad \hat{\eta}_{ij} + (Y_{ij} - h(\hat{\eta}_{ij}))/h^{'}(\hat{\eta}_{ij}) \approx \eta_{ij} + e_{ij}/h^{'}(\hat{\eta}_{ij}) \quad .$$

That is

$$Y_{ij}^* \equiv g(\hat{\mu}_{ij}) + g^{'}(\hat{\mu}_{ij})(Y_{ij} - \hat{\mu}_{ij}) \approx f(X_{ij1}, X_{ij2}, \ldots, X_{ijp}) + \mathbf{Z}_{ij}^{'}\mathbf{u}_i + e_{ij}^* \quad . \tag{4.4}$$

Note $1/h^{'}(\hat{\eta}_{ij}) = g^{'}(h(\hat{\eta}_{ij})) = g^{'}(\hat{\mu}_{ij})$ because $h(\cdot)$ is the inverse function $g(\cdot)$. Thus Y^*_{ij} approximately follows GMRTree model with tree function $f(X_{ij1}, X_{ij2}, \ldots, X_{ijp})$, random effects $\mathbf{u}_i$ and error term e^*_{ij}, where $var(\mathbf{e}^*_i) = \text{diag}(var(Y_{ij})g^{'}(\hat{\mu}_{ij})^2)$. The right side of equation (4.4) is the general form of the REEMTree model proposed by Sela and Simonoff (2009). Thus we can adopt the estimation procedures in REEMTree to our GMRTree.

In our proposed estimation method for GMRTree model for longitudinal count data, we will use the linearization technique in (4.4) to change the longitudinal response Y_{ij} into a surrogate response Y^*_{ij}. Since Y^*_{ij} is treated as continuous data, the tree-based methods for continuous data such as CART can be applied. The detailed algorithm is as follows:

1. Initial: Estimate a standard regression tree f using CART on count response Y_{ij} and predictors, $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^{'}$, for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$ by assuming $\hat{\mathbf{u}}_i = 0$.

2. Iterate through the following steps until the predicted random effects $\hat{\mathbf{u}}_i$ converges:

   (a) Use this estimated regression tree $\hat{f}$ to create a set of indicator variables, $I(\mathbf{X}_{ij} \in R_v)$, where R_v is the partition of the covariate space for each terminal node, $v = 1, \ldots, V$.

   (b) Fit the following conditional Poisson GLMM model and get the $\hat{\mathbf{u}}_i$

   $$\log(\mu_{ij}) = \mathbf{Z}^{'}_{ij}\mathbf{u}_i + \sum_{v=1}^{V} I(\mathbf{X}_{ij} \in R_v)\beta_v \quad .$$

   (c) Calculate the fitted $\hat{\mu}_{ij}$ from (b). Then calculate the surrogate response Y^*_{ij} by

   $$Y^*_{ij} = g(\hat{\mu}_{ij}) + (Y_{ij} - \hat{\mu}_{ij})g^{'}(\hat{\mu}_{ij}) \quad .$$

   (d) Estimate f by a weighted regression tree, based on the response $Y^*_{ij} - \mathbf{Z}^{'}_{ij}\hat{\mathbf{u}}_i$, and predictors $\mathbf{X}_{ij}$. The weight is $w_{ij} = (\hat{\mu}_{ij}g^{'}(\hat{\mu}_{ij})^2)^{-1} = \hat{\mu}_{ij}$ since $var(Y_{ij}) = \mu_{ij}$.

Here, a weighted regression tree is fitted in step 2(d), because after linearization of the GLMMs by (4.4), the covariance matrix of random error term is $\text{diag}(\hat{\mu}_{ij}g^{'}(\hat{\mu}_{ij})^2)$ instead of diagonal matrix $\sigma^2 I_{n_i}$. Normally in CART, we assume the error term has constant variance σ^2. For non-constant variance, we need to put a weight to fit the tree.

We call the above proposed approach Generalized Mixed Effects Regression Tree (or GMRTree). In this approach, we iteratively estimate the tree $f(\cdot)$ by CART (Classification and Regression Tree) and predict the random effects $\mathbf{u}_i$ by GLMM model until convergence. In the initial step, by assuming random effects $\mathbf{u}_i$ is 0, standard regression tree for Poisson data is estimated. The splitting criterion is based on maximizing the deviance between a parent node and two children nodes (see Chapter 1, Section 1.3.1, "Splitting criterion for Poisson data"). Once $f(\cdot)$ is estimated, then we will know R_v, the partition of the covariates space. We use this partition R_v information to fit a GLMM model and then predict the random effects $\mathbf{u}_i$ (step 2(b)). In step 2(c), a linearized response Y_{ij}^* is calculated. After step 2(c), a count response Y_{ij} is successfully changed into a surrogate continuous response Y_{ij}^*. Thus we can adapt tree-based method for continuous response to estimate the $f(\cdot)$. In the last step 2(d), $Y_{ij}^* - \mathbf{Z}_{ij}^{'}\hat{\mathbf{u}}_i$ is used to estimate the $f(\cdot)$ like REEMTree. The splitting criterion for step 2(d) is the goodness of split measurement in (1.6), which is based on maximizing the prediction errors between the parent node and two children nodes (See Chapter 1, Section 1.3.1, "Splitting a tree"). The pruning for Poisson data and continuous data are based on cost complexity criterion which is defined in (1.8). The cost complexity parameter is usually given by cross validation (See Chapter 1, Section 1.3.1, "Pruning a tree").

Within each iteration, from step 2(a) to 2(c), the partition R_v is fixed. R_v will be updated in step 2(d) since a new estimated $f(\cdot)$ is created. Before convergence, the partition of R_v will be updated after each iteration.

For the convergence, we first check whether the likelihood of step (b) converges. Most

of the time, the log-likelihood will converge in a few steps, especially if we fit the data with the correct model as shown in our simulation. However, there are also times that log-likelihood will not converge; but, the difference of log-likelihood between two consecutive iterations might be stable, in this case, we also claim it has converged.

CART theory is pretty mature and has been implemented in R package *rpart*. CART provides us splitting criteria and pruning algorithms, concepts briefly presented in Chapter 1. Parameter estimation of GLMMs has been implemented in R package *lme*4. Computation is not burdensome by using existing package *rpart* and *lme*4, and in fact there are useful purposes for utilizing tested software/algorithms.

4.3 Data analysis

4.3.1 Analysis of hemodialysis data

In Chapter 2, we introduced the binning method for aligning response and time-varying covariates that were originally collected at distinct time points within an individual. We utilize the same approach here. Specifically, we have applied the proposed GMRTree method to both the 30-day binned data and 7-day binned data (i.e., only for the first 7 week data), also described in Chapter 2. For the binned data, the response is the sum of infection events in each bin. Covariates are either logcrp only or logcrp and additional covariates such as age, gender, bmi and ethnicity. $log(crp)$ is the only time-varying covariate and its average within each bin becomes the binning method covariate.

Figure 4.1 is the regression tree for 30-day binned data by CART with one single covariate *logcrp*, while Figure 4.2 is the corresponded GMRTree. For tree built by CART (Figure 4.1), the value under each terminal nodes is the averaged response value within that terminal node. For example, the value 0.1073 for the node $logcrp < 0.01732$ is the

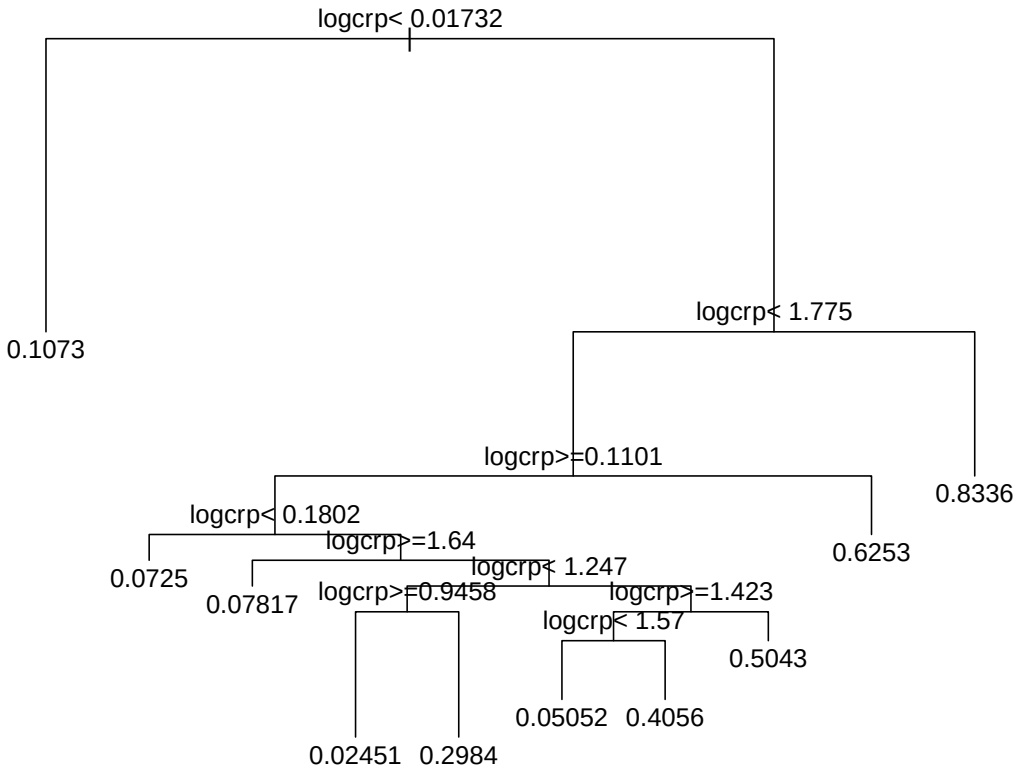

Figure 4.1: CART for infection for bin size 30; logcrp is the only covariate

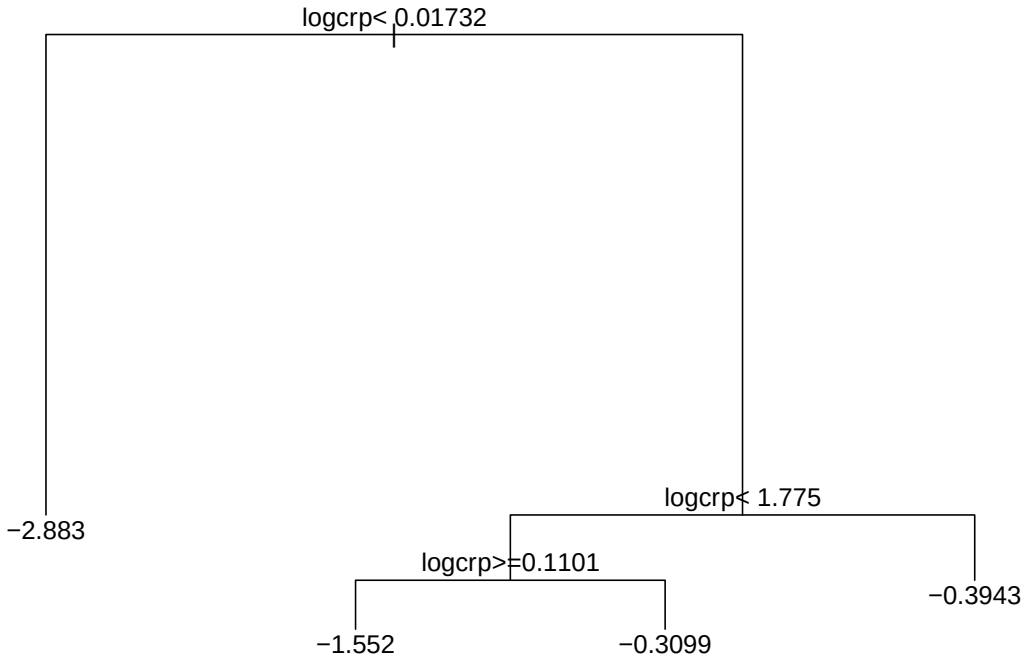

Figure 4.2: GMRTree for infection for bin size 30; logcrp is the only covariate

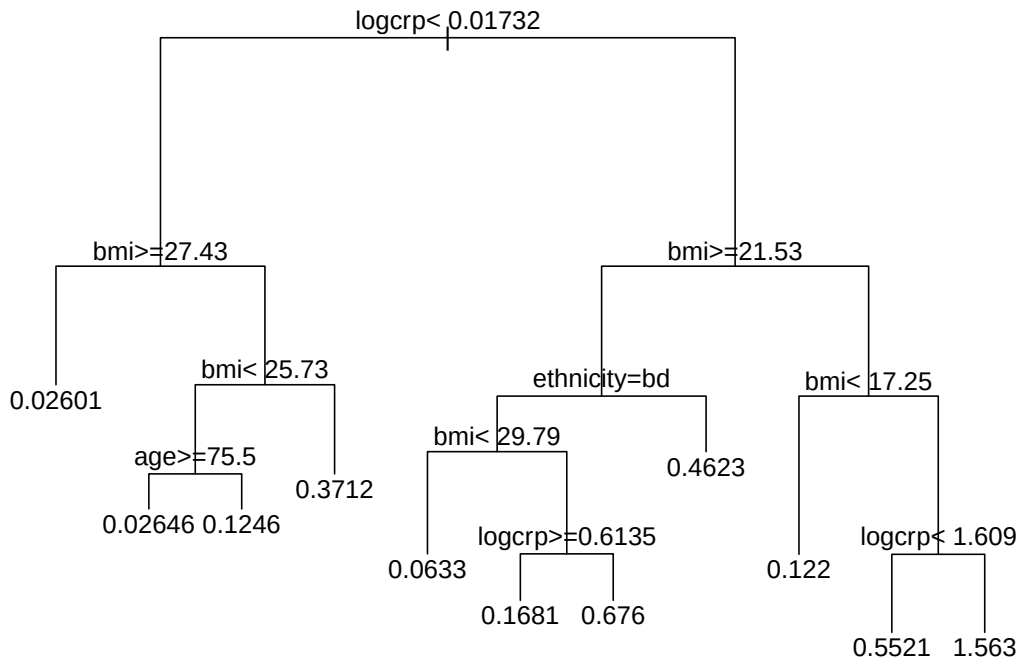

Figure 4.3: CART for infection for bin size 30; logcrp, bmi, age, gender, ethnicity are the covariates

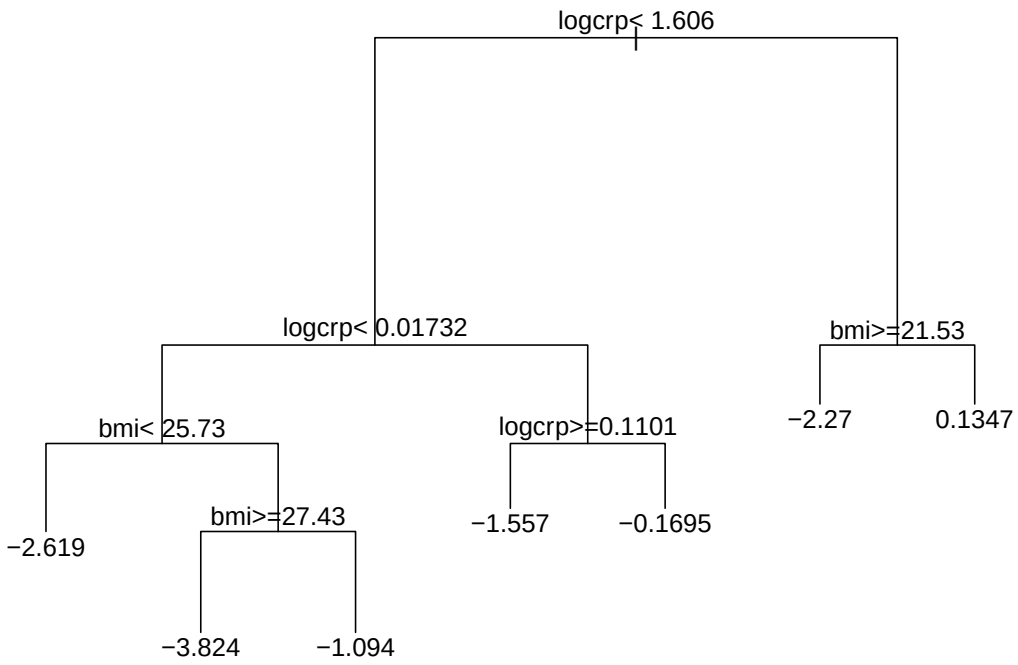

Figure 4.4: GMRTree for infection for bin size 30; logcrp, age, gender, bmi, ethnicity are the covariates

average of the responses whose logcrp meet the condition $logcrp < 0.01732$. For GMRTree, the value under each terminal node is the estimated fixed effects $\hat{\beta}_v$, assuming the node is the v^{th} terminal node with covariate partition R_v. And β_v is defined in

$$\log(\mu_{ij}) = \mathbf{Z}_{ij}^{'}\mathbf{u}_i + \sum_{v=1}^{V} I(\mathbf{X}_{ij} \in R_v)\beta_v \quad .$$

If we want to compare the terminal node values between CART and GMRTree, it is better to use the log transformation of value of CART for comparison.

Comparing Figure 4.1 with Figure 4.2, the size of GMRTree is smaller than that of CART. However, the first few splitting conditions of Figure 4.1 are the same as those of Figure 4.2. This may just be a coincidence.

If we include bmi, age, gender, ethnicity and logcrp as the covariates, the regression tree by CART for 30-day binned data is shown on Figure 4.3, while the regression tree by proposed GMRTree method is in Figure 4.4. In both of these two trees, $log(crp)$ is the leading splitting variable. However, variable ethnicity and age are significant splitting variables in CART (Figure 4.3), but not in GMRTree (Figure 4.4).

From GMRTree in Figure 4.4, we can see that both logcrp and bmi are selected as splitting variables. The first splitting point is $\{logcrp < 1.609\}$. The terminal node with condition $\{logcrp < 0.01732 \ \& \ bmi > 27.43\}$ has the smallest $\hat{\beta}_v$ -3.824, while the terminal node with condition $\{logcrp > 1.606\}$ & $\{bmi < 21.53\}$ has the largest $\hat{\beta}_v$ is 0.1347. This indicates that the conditional rate of the infection occurrences depends on logcrp level and bmi. logcrp level greater than 1.609 and and bmi less than 21.53 would have higher conditional expected rate of infection occurrences.

Figure 4.5 and Figure 4.7 are the regression trees for lagged infection by CART for 7-day binned data, while Figure 4.7 includes covariates other than logcrp. Figure 4.6 and Figure 4.8 are the corresponded GMRTrees.

If we just use logcrp as the only covariates such as Figure 4.6, the first three cut-off

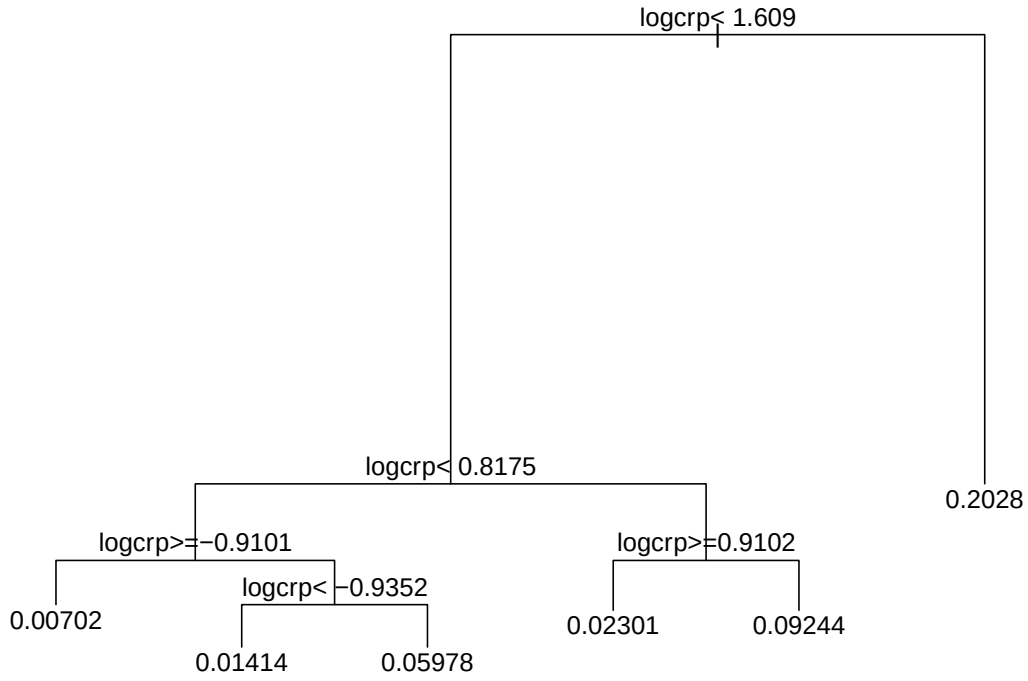

Figure 4.5: CART for lagged infection for bin size 7; logcrp is the only covariate

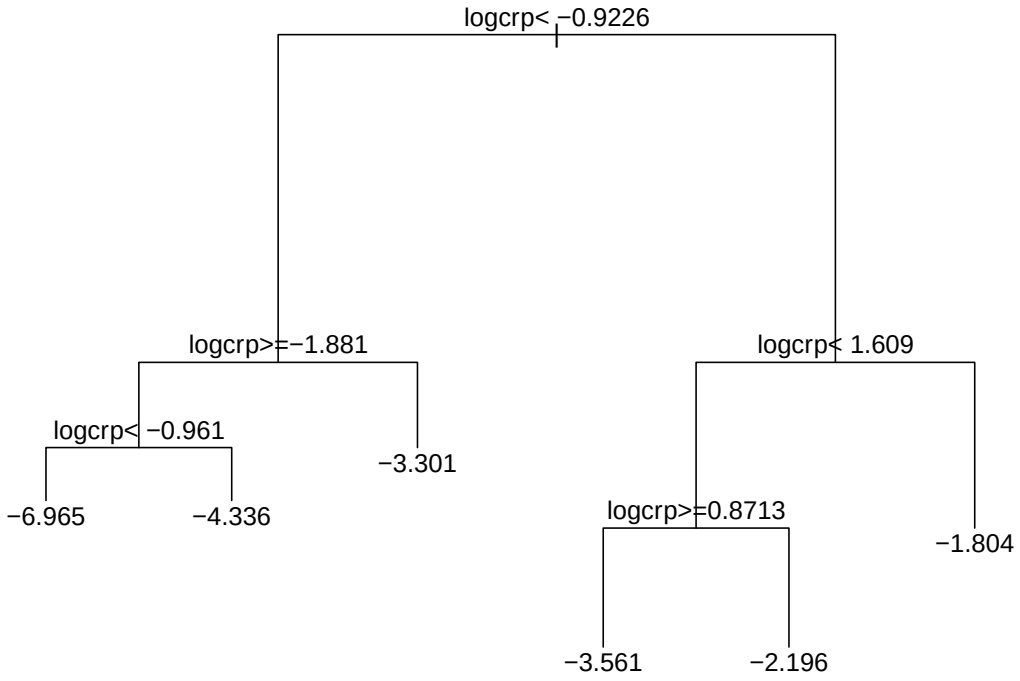

Figure 4.6: GMRTree for lagged infection for bin size 7; logcrp is the only covariate

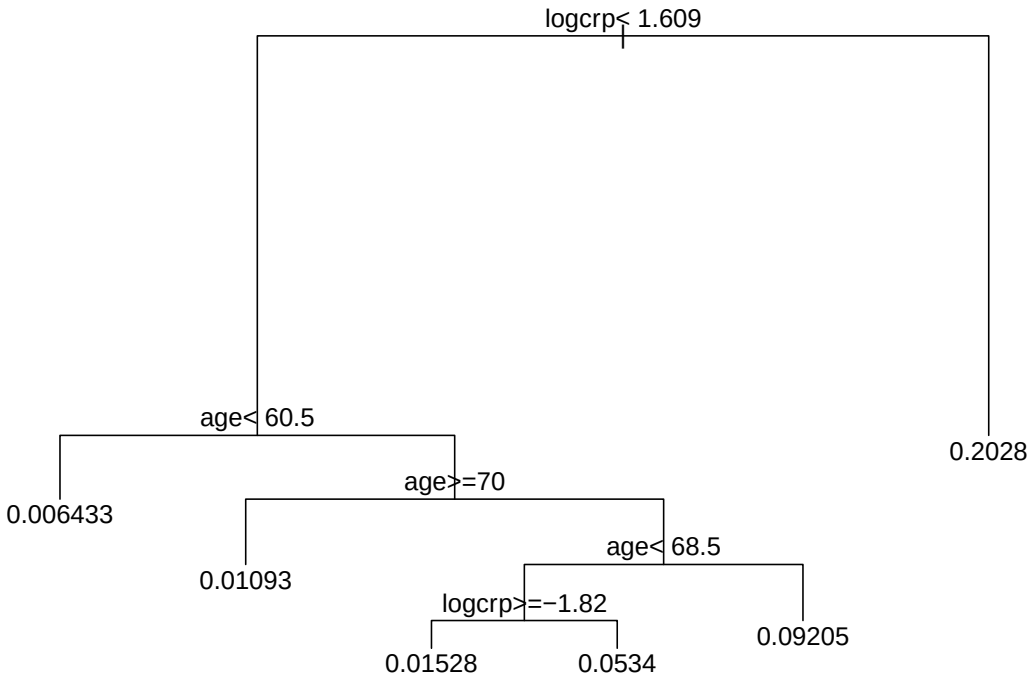

Figure 4.7: CART for lagged infection for bin size 7; logcrp, age, gender, bmi and ethnicity are the covariates

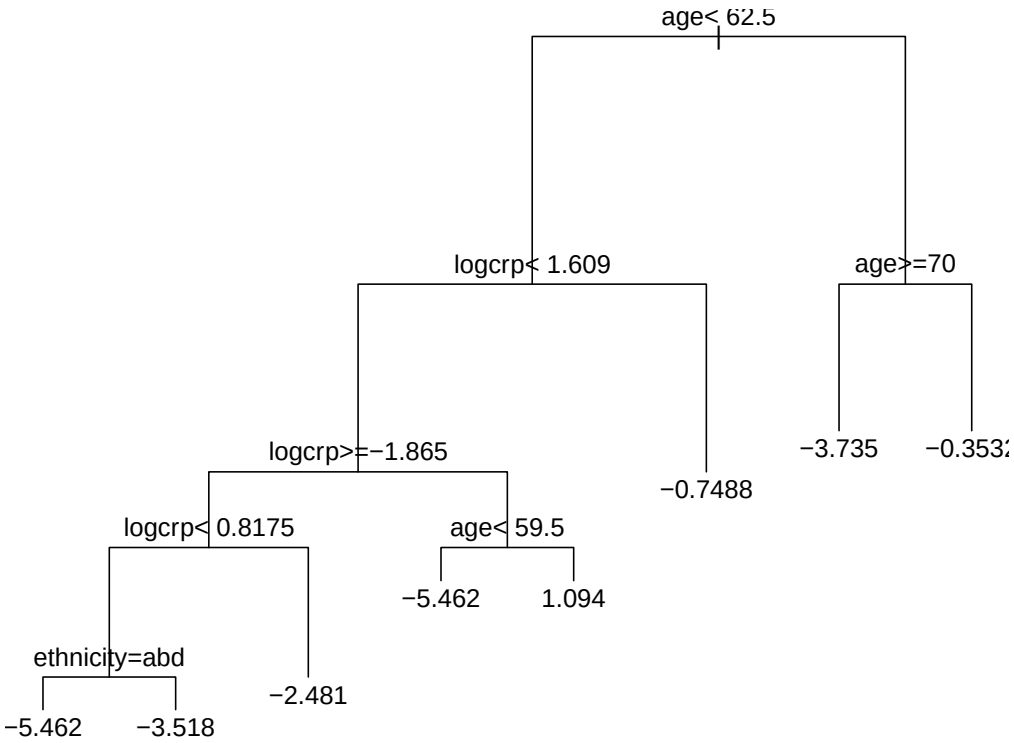

Figure 4.8: GMRTree for lagged infection for bin size 7; logcrp, age, gender, bmi, ethnicity are the covariates

levels for logcrp are -0.9226, -1.881, and 1.609. The conditions formed by these three cutoff levels are $\{logcrp > -1.881 \ \& \ logcrp < -0.9226\}$, $\{logcrp < -1.881\}$, and $\{logcrp > -0.9226 \ \& \ logcrp < 1.609\}$, $\{logcrp > 1.609\}$. However, there are still some cut points after these three levels. Branch $\{logcrp > 1.609\}$ has the highest estimated β_v among all the terminal nodes. This indicates when $\{logcrp > 1.609\}$, the patients would have recently experienced the highest conditional mean rate of infection occurrences. However, if we include more covariates to build the tree, then age, ethnicity and logcrp are all selected as splitting variables as shown on Figure 4.8.

4.3.2 Analysis of epilepsy seizure data

We next analyze a well-known count dataset consisting of epileptic seizure counts as presented originally by Thall and Vail (1990) and also analyzed by Diggle et al. (2002). Thall and Vail (1990) presented longitudinal data from a clinical trial of 59 epileptics who were randomized to a new anti-epileptic drug progabide with 31 under treatment (trt=1) and 28 under placebo (trt=0). The number of seizures was recorded for a baseline period of 8 weeks. Counts were then recorded for four successive two-week periods. Age was also recorded at the start of the trial. For this tree-based analysis, the covariates are logarithm of counts for the baseline period (lbase) centered to have zero mean, logarithm of age (lage) centered to have mean 0, Treatment (trt), an indicator for the fourth period (V4), i.e., if a given observation is in the fourth period after the baseline period. The response variable is the seizure count in each two-week period after baseline. We fit a GMRTree model to this data. The estimated tree is shown on Figure 4.10. From this figure, we can see that only *lbase* and *lage* have been selected to split the tree, while $V4$ and *trt* are not selected. There are a total of 7 terminal nodes from left to right, which are denoted as node number 1, 2, 3, 4, 5, 6 and 7. The value shown below each terminal node is the estimated β_v for that terminal node where β_v is defined in (4.2). The $\hat{\beta}_v$ for these 7 terminal nodes are

1.035, 0.6172, 1.696, 2.074, 2.073, 2.719 and 3.244 respectively. Node 2 has the smallest $\hat{\beta}_v$ 0.6172, while node 7 has the largest $\hat{\beta}_v$ 2.719.

Figure 4.11 and figure 4.12 shows the average seizure counts for all individuals within root node and each terminal node. This shows that terminal nodes 1, 2 and have noticeable lower average seizure counts than those of terminal node 4, 5, 6 and 7, while individuals in node 2 have the lowest seizure counts and node 7 has much higher ones. Individuals within each terminal node have different patterns of trajectories, which supports the use of GMRTree.

Figure 4.13 is the average seizure count overall subjects for each terminal node which shows that terminal node 6 has the smallest average seizure count, while terminal node 13 has the largest one. This result agrees with the estimated β_v in each terminal node ($\hat{\beta}_v$ for node 2 is 0.6172 and is the smallest among the seven $\hat{\beta}_v$, while node 7 has the largest $\hat{\beta}_v$ 2.719).

4.4 Simulation

We conduct some simulations to evaluate the performance of the proposed GMRTree. We would like to know whether GMRTree can effectively detect the underlying tree structure, and how it performs comparing with CART and GLMM on their own.

4.4.1 Setup: data generation

In this simulation, we consider two data generation scenarios: generate data from a GMRTree model with a random intercept or from a regression tree CART (no random effects).

Assume there are 4 covariates X_1, X_2, X_3 and X_4 which are independent and identically distributed as $Unif(0, 2)$. The split point for each variable is at value 1. The first

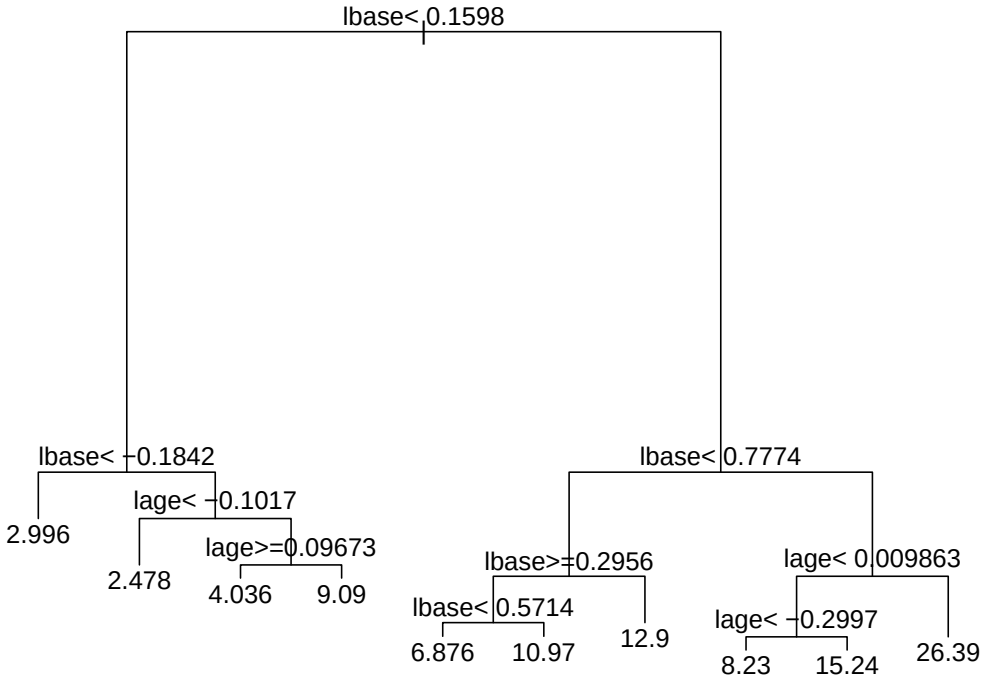

Figure 4.9: CART for seizure data

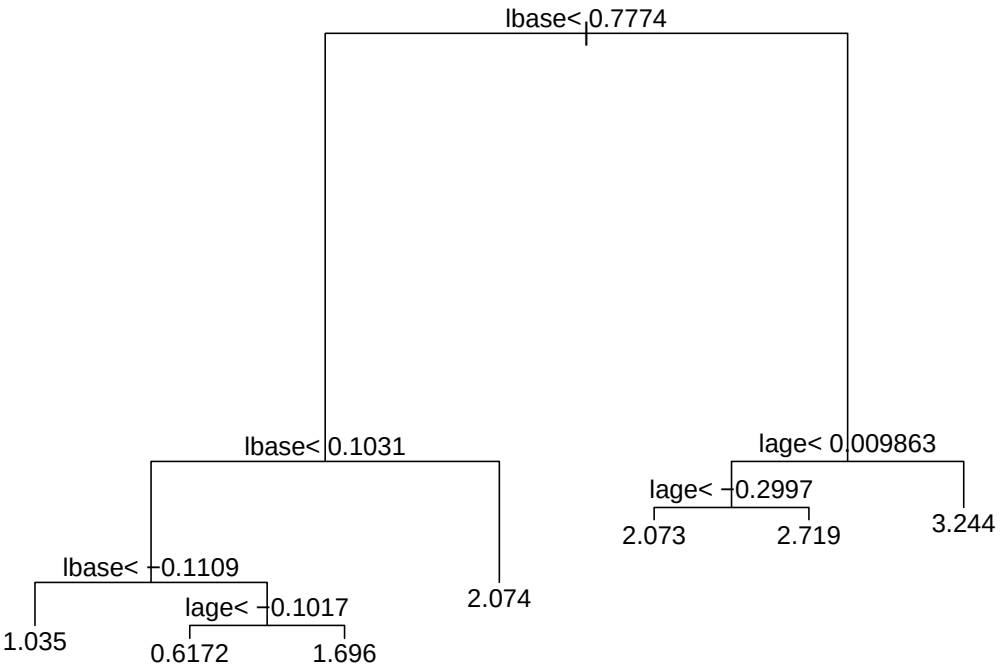

Figure 4.10: GMRTree for seizure data

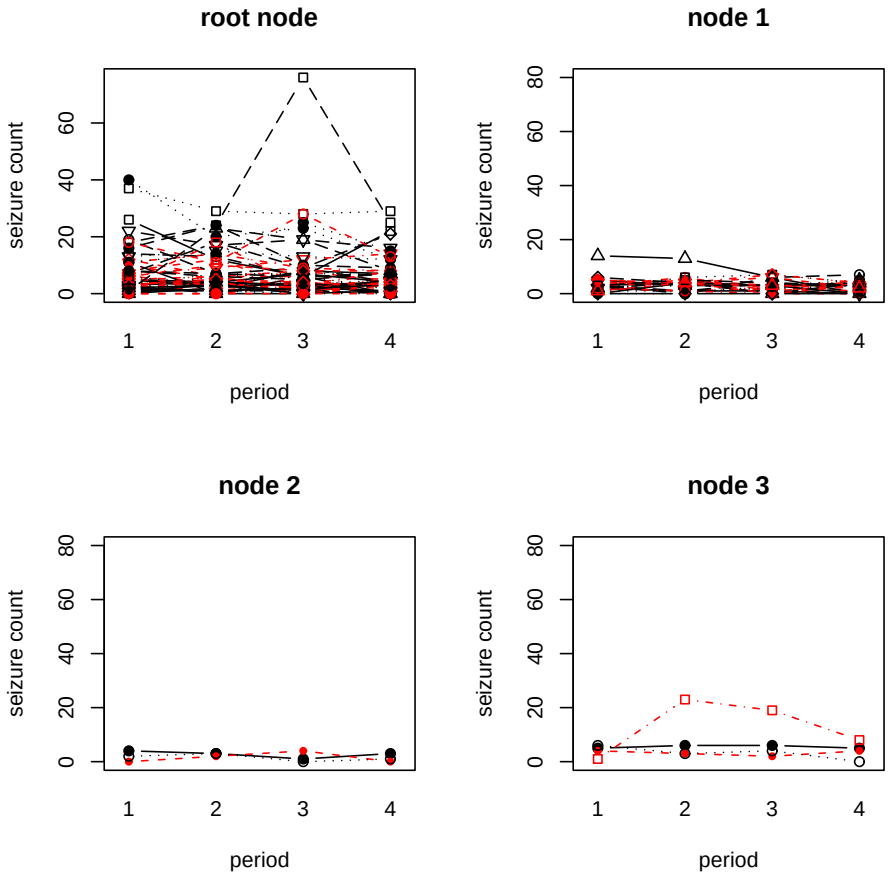

Figure 4.11: Patterns of number of epileptic seizures for individuals in GMRTree root node and nodes 1, 2 and 3

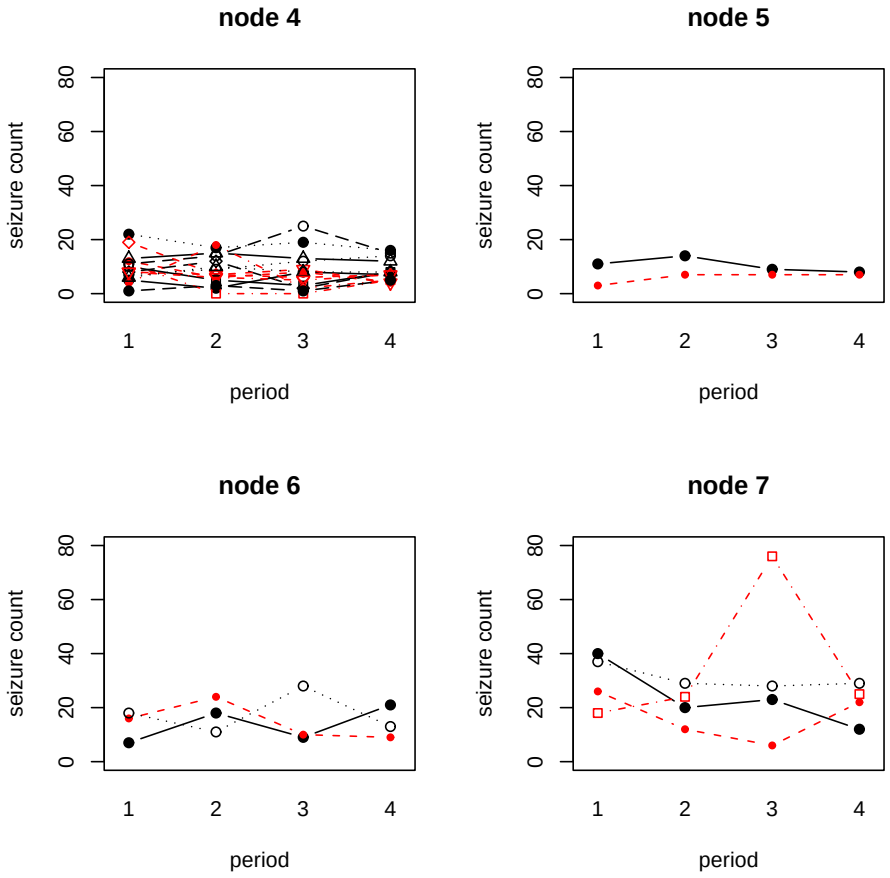

Figure 4.12: Patterns of number of epileptic seizures for individuals in GMRTree nodes 4, 5, 6, and 7

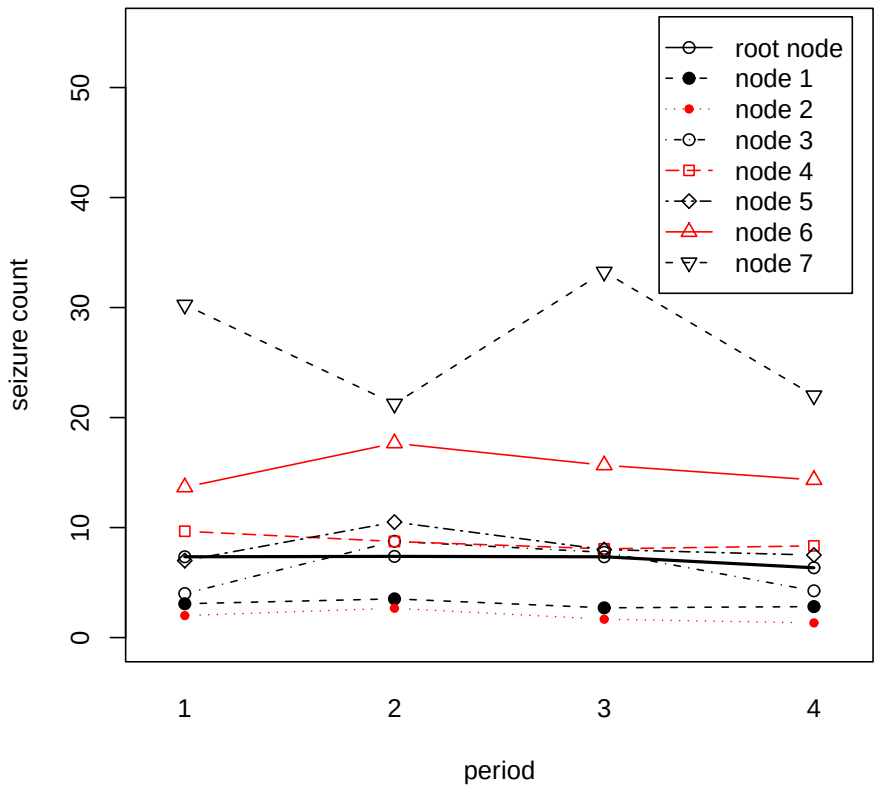

Figure 4.13: Average profile of number of epileptic seizure within each terminal node in the GMRTree

splitting variable is X_1. For each generate dataset, there are 100 subjects and 60 repeated observations for each subject. For each subject, we use the first 10 observations of the generated data as "training data". The next 50 observations of each subject are used as "testing data" which will be used for prediction.

1. If the data is generated by GMRTree model which is a regression tree plus a random intercept model, then the model for generating the simulated Y_{ij} is

$$\begin{aligned}\log(\mu_{ij}) &= \beta_1 I_{\{(X_{ij1}<=1)\cap(X_{ij2}<=1)\}} + \beta_2 I_{\{(X_{ij1}<=1)\cap(X_{ij2}>1)\}} \\ &+ \beta_3 I_{\{(X_{ij1}>1)\cap(X_{ij3}<=1)\}} + \beta_4 I_{\{(X_{ij1}>1)\cap(X_{ij3}>1)\cap(X_{ij4}<=1)\}} \\ &+ \beta_5 I_{\{(X_{ij1}>1)\cap(X_{ij3}>1)\cap(X_{ij4}>1)\}} + u_i \quad , \end{aligned} \tag{4.5}$$

   where $u_i \sim N(0, \sigma_u^2)$.

2. If the data is generated from CART, then the model is (4.5) without the random effects u_i. That is

$$\begin{aligned}\log(\mu_{ij}) &= \beta_1 I_{\{(X_{ij1}<=1)\cap(X_{ij2}<=1)\}} + \beta_2 I_{\{(X_{ij1}<=1)\cap(X_{ij2}>1)\}} \\ &+ \beta_3 I_{\{(X_{ij1}>1)\cap(X_{ij3}<=1)\}} + \beta_4 I_{\{(X_{ij1}>1)\cap(X_{ij3}>1)\cap(X_{ij4}<=1)\}} \\ &+ \beta_5 I_{\{(X_{ij1}>1)\cap(X_{ij3}>1)\cap(X_{ij4}>1)\}} \quad , \end{aligned} \tag{4.6}$$

Table 4.1 shows the parameter values of $\beta_k, k = 1, \ldots, 5$ and σ_u^2. There are four sets of parameters in this simulation study which are denoted as 1, 2, 3 and 4 in the first column of table 4.1. Two sets are for CART model and two for GMRTree model. For CART model, $\boldsymbol{\beta}$ values in the set 2 are higher than that of set 1. For GMRTree model, σ_u is set 4 is higher than that of set 3. So in total, there are 4 dataset is generated. However, for each generated data, the first 10 observations of each subject consist of the training data. The remainder of the 50 observations of each subject consists of the testing data.

Table 4.1: Parameters and models for data generation

Data Generation	model	Tree					Random Effects
		β_1	β_2	β_3	β_4	β_5	σ_u^2
1	CART	-2	-0.5	0.5	1	1.5	0
2	CART	0.5	1	2	3	4	0
3	GMRTree	-1	0.5	1	1.5	2	0.36
4	GMRTree	-1	0.5	1	1.5	2	2.25

4.4.2 Predictive performance

To evaluate the performance of the GMRTree, first four models are fitted to the training data, which are

- GMRTree model: a GMRTree model with random intercept. GMRTree model is defined in (4.1). Here we only use a random intercept for $\mathbf{u}_i$. The split points will be given by the fitted tree, which are most likely not the same as $X_{ijk} = 1$ where $k = 1, \ldots, 4$ but will be close.

- True model: true model is the model used by the data generation, which is either (4.5) or (4.6). In the true model, the split point is always at $X_{ijk} = 1$.

- GLMM: a generalized linear mixed effects model which contains only main effects and random intercept. No splitting point is considered in this case. If the data is generated by CART, then it is a GLM model.

- CART model: a standard Poisson regression tree model by *rpart*. If we fit CART model to longitudinal data, we will treat the data as uncorrelated, this ignore the correlation among the responses. No random effects are considered in CART model.

Then we make prediction for the four fitted models by using testing data. And the corresponded prediction errors are calculated for each fitted model.

There is no specific guidelines on how to compare the prediction performance of different non-nested fitted models for non-normal count data. However, for continuous data, Sela and Simonoff (2009) used RMSE (root mean square error) to evaluate the prediction performance. Here we propose three criteria, prediction of mean square error (PMSE), prediction of mean absolute deviation (PMAD) and prediction of mean absolute bias (PMAB), which are defined as follows

$$\begin{aligned} PMSE &= \frac{1}{N*n}\sum_{i=1}^{N}\sum_{j=1}^{n}(Y_{ij}-\hat{\mu}_{ij})^2 \quad , \\ PMAD &= \frac{1}{N*n}\sum_{i=1}^{N}\sum_{j=1}^{n}|Y_{ij}-\hat{\mu}_{ij}| \quad , \\ PMAB &= \frac{1}{N*n}\sum_{i=1}^{N}\sum_{j=1}^{n}|\hat{\mu}_{ij}-\mu_{ij}| \quad , \end{aligned}$$

where $N = 100, n = 50$. Y_{ij} is the response of the testing data. And μ_{ij} is the estimated conditional mean of the testing data by using the estimated parameters from the training data. Details of calculation of μ_{ij} are as follows: get the estimates $\hat{f}$ and predicted random effects $\hat{u}_i$ for each subject from the fitted model of the training data. If the fitted models are CART models, then $\hat{u}_i$ is 0. Then predict the $\hat{\mu}_{ij}$ for the testing data by

$$\hat{\mu}_{ij} = exp(\hat{f}(X_{ij1}\ldots,X_{ij4}) + \hat{u}_i)$$

Note that $X_{ijk}, k = 1,\ldots,4$ are the covariates in the testing data, not the ones generated from the original training data.

On the other hand, for the testing data, we know the true μ_{ij}. So we should be able to calculate the bias for the μ_{ij}, thus be able to calculate $PMAB$.

Table 4.2: Prediction error for different fitted models with different data generation models. The parameters for data generation are given by Table 4.1 and the number of subjects is 100. For the training data, the number of repeated measures per subject is 10. For the testing data, the number of repeated observations per subject is 50.

data	fitted		PMSE			PMAD			PMAB		
generation	model	$\|T\|$	min	mean	max	min	mean	max	min	mean	max
CART	GMRTree	5.09	1.408	1.555	1.768	0.838	0.876	0.916	0.037	0.094	0.241
1	True	5	1.400	1.510	1.664	0.831	0.865	0.910	0.015	0.065	0.140
	CART	5.01	1.402	1.548	1.748	0.839	0.873	0.916	0.028	0.085	0.238
	GLMM	5	2.116	2.309	2.488	1.051	1.094	1.143	0.619	0.647	0.683
CART	GMRTree	5.09	12.030	14.810	21.970	2.287	2.408	2.572	0.291	0.487	0.782
2	True	5	11.080	12.280	13.320	2.243	2.322	2.424	0.052	0.180	0.450
	CART	5.01	12.030	14.910	21.980	2.287	2.409	2.600	0.294	0.476	0.777
	GLMM	5	129.000	140.900	157.400	6.910	7.286	7.776	6.453	6.830	7.349
GMRTree	GMRTree	5.02	3.163	3.932	6.088	1.218	1.348	1.524	0.360	0.465	0.613
3	True	5	3.068	3.740	4.684	1.201	1.328	1.473	0.346	0.434	0.545
	CART	5.31	5.748	11.110	26.630	1.645	2.053	2.833	1.087	1.562	2.427
	GLMM	5	5.839	8.330	13.120	1.655	1.902	2.202	1.164	1.397	1.672
GMRTree	GMRTree	5	4.223	7.945	24.170	1.273	1.566	2.157	0.389	0.571	1.004
4	True	5	4.040	6.644	10.750	1.262	1.533	1.901	0.382	0.526	0.804
	CART	9.19	23.870	182.600	1401.000	2.951	5.621	11.430	2.624	5.355	11.200
	GLMM	5	11.140	47.110	273.500	1.947	2.894	4.476	1.493	2.405	3.993

4.4.3 Simulation results

Simulation results are shown in Table 4.2. The first column "data generation" denotes the model used to generate the training and testing data. The second column "fitted model" denotes the fitted model to the training data. And this fitted model is applied to the testing data to do prediction. The third column "$|T|$" is the average size of the tree if the fitted model is a tree model (either GMRTree or CART model). The "$|T|$" is 5 for GLMM, which is defined as the number of estimated fixed effects in the model. For the true model, "$|T|$" is 5 since there are always 5 terminal nodes in the tree, no matter data is generated by (4.6) or (4.6). Thus the tree size is 5.

We can see that when the data is generated by CART model, the fitted GMRTree model with random intercept still provides close PMSE/PMAD/PMAB results to those of the fitted CART model, and also close results to the fitted true model using the tree splitting points and splitting variables. The tree size of fitted GMRTree model is close to 5. The main reasons is that in this case, the predicted random intercept of the fitted GMRTree model will be very small and close to zero, thus has little noticeable effect on the final estimated conditional mean $\hat{\mu}_{ij}$. However, if the data is generated by a CART model, but we fit a GLMM model to this data, then the prediction errors for GLMM model would be the largest compared with those of the fitted GMRTree or CART model; this pattern is more obvious if the $\boldsymbol{\beta}$ increases. So if the data are generated by a model which has tree structure such as CART or GMRTree, GMRTree can efficiently detect the tree structure, even though the model for the generated data may be a CART model.

If the data is generated by GMRTree with random intercept, then the fitted GMRTree model would provide the smallest PMSE/PMAD/PMAB than those of fitted CART or fitted GLMM model. And the PMSE/PMAD/PMAB of fitted GMRTree model is close to those of the fitted true model, which indicates that the estimated parameters of GMRTree model would be close to its true values of the parameters. And the partitions of the

Table 4.3: For the generated data, $\boldsymbol{\beta} = c(-1, 0.5, 1, 1.5, 2)$, variance of random effects $\sigma_u^2 = 0.36$ and the number of subjects is 200. For the training data, the number of repeated observations per subject is 20. For the testing data, the number of repeated observations per subject is 50.

data	fitted		PMSE			PMAD			PMAB		
generation	model	$\|T\|$	min	mean	max	min	mean	max	min	mean	max
GMRTree	GMRTree	5	3.133	3.482	4.018	1.220	1.289	1.368	0.267	0.312	0.366
	True	5	3.070	3.444	4.013	1.212	1.284	1.367	0.262	0.305	0.361
	CART	5	6.852	10.270	19.100	1.753	1.989	2.360	1.199	1.492	1.885
	GLMM	5	6.180	7.620	10.480	1.688	1.839	2.000	1.171	1.316	1.489

GMRTree covariates space would be close to those of the true model.

In summary, GRMTree model can efficiently detect the tree structure in the data.

More simulations are done to explore the GMRTree performance by changing the number of subjects, parameter values of $\boldsymbol{\beta}$, and σ_u^2 value in the generated data. Also, instead of using four uniform covariates, we consider mixed types of covariates such as including both continuous and categorical covariates. Here we assume the both the testing and training data are generated from a GMRTree with random intercept. The prediction errors of the testing data are calculated by PMSE/PMAD/PMAB, the same way as the ones calculated in Table 4.2.

The numbers of subject in the training data in Table 4.3 is increased from 100 to 200, comparing with Table 4.2, and repeated measurements from each subject increased from 10 to 20. The tree parameter $\boldsymbol{\beta}$ in Table 4.4 is increased from (-1, 0.5, 1, 1.5, 2) to (1, 2, 3, 3.5, 4), comparing with Table 4.3. In Table 4.5, binomial and discrete uniform covariates are included.

The pattern of the simulation results in Table 4.3, Table 4.4 and Table 4.5 are similar to the results in Table 4.2 when the data generation model is GMRTree. That is, when the data is generated by a GMRTree model with a random intercept, if we fit a GMRTree with

Table 4.4: For the generated training and testing data, $\boldsymbol{\beta} = c(1, 2, 3, 3.5, 4), \sigma_u^2 = 0.36$, and the number of subjects is 200. For the training data, the number of repeated observations per subject is 20. For the testing data, the number of repeated observations per subject is 50.

data	fitted		PMSE			PMAD			PMAB		
generation	model	$\|T\|$	min	mean	max	min	mean	max	min	mean	max
GMRTree	GMRTree	5	22.070	25.000	29.840	3.154	3.350	3.534	0.706	0.843	0.993
	True	5	21.050	24.240	27.240	3.136	3.336	3.523	0.699	0.825	0.959
	CART	5	237.400	396.800	901.900	9.049	10.940	13.660	8.326	10.340	13.160
	GLMM	5	204.700	270.000	455.200	8.871	9.919	11.130	8.210	9.325	10.560

Table 4.5: Covariate $X_1 \sim Unif(0, 2)$, $X_2 \sim binom(1, 0.5)$, $X_3 \sim Unif(0, 1)$, $X_4 \sim$ discrete $Unif(1, 2, 3, 4)$. For the generated training and testing data, $\boldsymbol{\beta} = c(1, 2, 3, 3.5, 4)$, $\sigma_u^2 = 0.36$ and the number of subjects is 200. For the training data, the number of repeated observations per subject is 20. For the testing data, the number of repeated observations per subject is 50.

data	fitted		PMSE			PMAD			PMAB		
generation	model	$\|T\|$	min	mean	max	min	mean	max	min	mean	max
GMRTree	GMRTree	5	3.126	3.531	4.103	1.246	1.323	1.405	0.262	0.312	0.364
	True	5	3.102	3.507	4.100	1.240	1.321	1.404	0.261	0.308	0.364
	CART	5	6.929	10.340	19.250	1.799	2.032	2.409	1.226	1.522	1.928
	GLMM	5	5.931	7.316	9.811	1.675	1.806	1.989	1.104	1.237	1.413

random intercept to this generated data, then the PMSE/PMAD/PMAB values are very close to those of the fitted true model, and much smaller than those of a fitted GLMM model or a fitted CART model. This pattern is more obvious when we increase the tree parameter $\boldsymbol{\beta}$. When we increase the $\boldsymbol{\beta}$, it is more obvious that GMRTree outperforms CART or GLMM model. Based on the PMSE/PMAD/PMAB, the prediction error statistics for fitted GMRTree in Table 4.5 are close to those of the fitted true model, which indicates that GMRTree method also supports different type of covariates.

4.4.4 Predictive performance for new subjects

For all the above simulations, the subjects in the testing data are the same as the subjects in the training data. In this case, we use the $\hat{f}$ and random effects $\hat{u}_i$ from the training data to predict $\hat{\mu}_{ij}$ for the testing data. The predicted $\hat{\mu}_{ij}$ is calculated by

$$\hat{\mu}_{ij} = exp(\hat{f}(X_{ij1} \ldots, X_{ij4}) + \hat{u}_i) \quad ,$$

However, if the subjects in the testing data are different from the ones in the training data, the predicted random effects $\hat{u}_i$ will be unknown for the testing data. Thus we need to predict the random effects u_i for the subjects in the testing data, using equation (3.7) in Chapter 3. In this simulation, a training data is generated which has 100 subjects and 10 repeated observations for each subject. Another testing data is generated with 100 new subjects and 50 repeated observations per subject.

Simulation results are shown in Table 4.6. Even though the subjects are new for the testing data, the prediction errors (PMSE/PMAD/PMAB) of the proposed GMRTree are still very close to those of the true model, and much smaller than the predictor error for CART and GLMM. These results agree with the results in Table 4.2 where the data generation is GMRTree (data 3 & 4) and fitted model is GMRTree.

Table 4.6: Prediction error for new subjects. For the generated training data, there are 100 subjects with 10 repeated observations per subject. Fixed effects $\beta = c(-1, 0.5, 1, 1.5, 2)$, $\sigma_u^2 = 0.36$. Covariate X_1 to X_4 are all Unif(0, 2). For the generated testing data, number of subjects is 100, and the number of repeated observations per subject is 50.

data	fitted		PMSE			PMAD			PMAB		
generation	model	$\|T\|$	min	mean	max	min	mean	max	min	mean	max
GMRTree	GMRTree	5.026	2.574	3.289	4.489	1.135	1.252	1.412	0.178	0.245	0.404
	True	5	2.558	3.124	3.918	1.119	1.234	1.365	0.162	0.214	0.297
	CART	5.304	6.290	11.120	28.030	1.703	2.051	2.629	1.143	1.560	2.229
	GLMM	5	4.921	7.142	11.010	1.554	1.783	2.130	1.027	1.263	1.594

4.4.5 Summary

In summary, GMRTree method can efficiently detect the tree structure in the data. Simulation results have shown that if the generated data is from a tree model (either GMRTree or CART model), the prediction errors (i.e., PMSE/PMAD/PMAB) of the fitted GMRTree model are very close to the prediction errors of the true model which uses the true split point and true splitting variables, and the prediction errors of the fitted GMRTree model outperform those of fitted GLMM and CART model. Even when the data is generated from a CART model, fitted GMRTree still provides close result to the fit of the true CART model which uses the true splits.

4.5 Discussion

The GMRTree method utilizes the existing theories of CART and GLMM model, and uses existing R functions *lme*4 and *rpart*. There is no restricted requirements such that subjects have to have an equal number of observations and have measurements at a well-defined consistent schedule in Segal(1992). Time-dependent covariate can also be included

in GMRTree model, so that different observations of the same subjects can be located in different terminal nodes.

In GMRTree, one main step is to change the count response into a continuous response by linearization. Then we use the continuous response to estimate the tree and use GLMM model to predict the random effects. The performance of GMRTree also largely depends on the linearization. For GLMM model, only when the conditional distribution approximates normal, then the linearization works well. This will work better for count data than for binary data, for example.

Simulations have shown that GMRTree method can effectively detect the tree structure located in the data. If the data really has a tree structure, then GRMTree method works better than GLMM model. The estimated PMSE, PMAD and PMAB for fitted GMRTree model are very close to the ones for the true model, which indicates the fitted GMRTree model is close to the true model. Even though the underlying model is CART and we fit a GMRTree model, the fitted model would still be close to the true model, since at this time, the predicted random effects would be very close to zero.

However, GMRTree also has its shortcomings. When the data is generated from a GLMM model with main effects and we fit a GMRTree tree model to this data, sometimes the estimation algorithm of GMRTree does not converge. There are several reasons for this. First when we use a tree to approximate a true linear model, we already have some information lost. Second, we use this approximated tree to get the random effects, thus the predicted random effects may incur some bias. Third, linearization itself may also introduce some bias. So if we fit a GMRTree to a GLMM, and if we still use the convergence of log-likelihood, or convergence of fixed effects and random effects as convergence criteria, we may not be able get proper convergence at all. So, in future work, we will consider alternative convergence criteria, and see how GMRTree performs under more general model settings.

CART uses exhaustive search algorithm to find the best split point and best split variable by searching over all the split variables and all the split points. However this exhaustive search approach has selection bias toward variables which provide more split points (Doyle 1973). GUIDE (Loh, 2002) almost provides an unbiased variable selection and also detects curvature. Hence, as part of future work, we would like to investigate if we can use GUIDE to estimate $f(\cdot)$ instead of CART. Other tree-based algorithms for continuous data may also be considered to estimate $f(\cdot)$.

Chapter 5

Summary and future work

In this Chapter, we provide a summary of current contributions of this thesis and also a listing of some future work items.

5.1 Summary

The research topic from this thesis originated from a real hemodialysis study where the longitudinal response (such as infection event) and time-varying covariates (such as C-reactive protein (crp)) were not measured on the same time points. We were interested in finding out the temporal association between the response and covariates, such as the ordering of infection event occurrence and changes in *crp* levels. We were also interested in finding out the cut-off levels of *crp* which might be indicative of event occurrences.

In Chapter 2, a binning method was introduced for longitudinal data where the response and time-dependent predictors were not measured on the same time points. Then the proposed binning approach was applied on a longitudinal hemodialysis study to look for possible contemporaneous and lagged effects between occurrences of a health event (i.e.,

infection) and levels of a protein marker of inflammation (i.e., crp). Both Poisson mixed effects models and zero-inflated Poisson (ZIP) mixed effects models were applied to the full dataset and first 7-week subset data. We have found there were contemporaneous effects for mixed ZIP model between infection and log(crp) for the full data by using bin size 30 days and 45 days. Another important biological finding was that there was lagged association between infection and log(crp) when looking into the first 7-week data, when we could consider smaller bin sizes, such as bin-length of 7 days. More specifically, infection event occurred prior to a rise in log(crp). In addition, extensive simulations were conducted to investigate various properties of the binning approach including within-subject variability, mismatch rate, correlation structure, edge effects, etc. We also provided an initial approach on bin size selection. This work has been published recently (Xiong and Dubin, 2010).

In Chapter 3, we have derived asymptotic properties of the binning estimates for completed matched data and partially mismatched data, especially on the fixed effects. We also further investigated bin size selection criterions for the binning approach by using data both before binning and after binning.

In Chapter 4, we had interest in identifying levels of *crp* that might be indicative of recent infection occurrences in hemodialysis patient. Hence, in order to find out the potential cut-off levels of crp, we proposed a generalized mixed effects regression tree (GMRTree) method for longitudinal count data. The GMRTree model included a tree structure and also allowed for inclusion of random effects. We used weighted regression tree to estimate the tree, and used generalized linear mixed effects model to estimate the random effects. One of the main steps here was to change the longitudinal count response into a surrogate continuous response by linearization. We evaluated the prediction performance of the proposed GMRTree method by comparing the predicted of mean square error (PMSE), prediction of absolute mean deviation (PMAD) and prediction of absolute mean bias (PMAB) with those of CART and GLMM models. The proposed GMRTree

approach has been applied to binned hemodialysis data and epilepsy seizure data.

In the following section, we discuss some future work items which will extend what has been done in this thesis.

5.2 Future work

5.2.1 Binning method and bin size selection

Binning is a relatively straightforward approach that has proved useful in other areas of statistics, and has allowed us to answer some important biomedical questions for the longitudinal hemodialysis study discussed in Chapter 1. After binning, we are able to use traditional longitudinal models and standard statistical packages to do analysis. For longitudinal data measured at distinct time points, binning method has been shown to be useful to find the temporal association between coviariates and response. Apart from binning method, curve-based methods are also potential approaches to model the association between the covariates and response when they are not aligned on time. Also methods in missing data framework such as EM algorithm may be another choice. Measurement error models can also be a third choice, here we model the covariates by a measurement error model. Then we find out the temporal association between response and true covariates, by using the observed covariate values which have measurement errors.

So, in the future, more efforts are needed to investigate the potential of curve-based methods, EM algorithms and measurement error models for longitudinal data where the response and covariates are not measured on the same time points.

For bin size selection, we have proposed leave-one subject out cross-validation to select the best bin size by minimizing the PRSS value. The PRSS value is calculated by either using original data before binning or the binned data. For original data, PRSS can also be

calculated either using the marginal mean or conditional mean. For the binned hemodialysis data, the PRSS values are not very sensitive to different bin sizes since the PRSS values are close for different bin sizes. In the future, we would like to develop more sensible and more robust criterions for selecting bin sizes.

The approach of using binned data to calculate the PRSS ignores the direct impact of the original data points Y_{ij}, only considering the individual points indirectly following binning, i.e, focusing on the Y_{im}. Thus, we have proposed two approaches of bin size selection by using original data. However, for the PRSS using original data, we still need to calculate $\hat{E}_{-i}(Y_{ij}|\mathbf{u}_i)$ or $\hat{E}_{-i}(Y_{ij})$. However, we do not have model information for original data, thus we cannot calculate these expectations directly. The only information is the binned data. We proposed to predict these two expectations by $\hat{E}_{-i}(Y_{i,m}|\mathbf{u}_i)/c_{im}^{(\mathbf{Y})}$ or $\hat{E}_{-i}(Y_{i,m})/c_{im}^{(\mathbf{Y})}$ respectively, where m is the bin number where Y_{ij} is located, and $Y_{i,m}$ is the binned response, $c_{im}^{(\mathbf{Y})}$ is the total number of original responses in m^{th} bin for subject i. However, it is not known what is the implication of choosing one over the other between the two proposed approaches by using the original data. In the future, more investigation needs to be done to choose between these two approaches. We still need to evaluate the performance of using $\frac{\hat{E}_{(-i)}(Y_{i,m}|\mathbf{u}_i)}{c_{im}^{(\mathbf{Y})}}$ to calculate $\hat{E}_{(-i)}(Y_{ij}|\mathbf{u}_i)$, and $\frac{\gamma_{i,m}}{c_{im}^{(\mathbf{Y})}}$ to calculate $\hat{E}_{(-i)}(Y_{ij})$ respectively.

In summary, in the future, more efforts need to be done to build more sensible and robust bin size selection criteria by using both binned response $Y_{i,m}$ and original individual response Y_{ij}. More investigation needs to be done to choose among bin size selection approaches using binned data and original data.

5.2.2 Correlation structure in models of binned data

We would like to more specifically investigate the unconditional correlation structure of repeated measures following binning, including what effect the conditional correlation struc-

ture of the repeated measures may provide, as well as the correlation structure coming from the time-varying covariates. We only indirectly investigated, via simulation, the effect of the time-varying covariate structure, using only AR(1), on the resulting unconditional correlation structure of the repeated measures following binning.

Sutradhar (2003) proposed to use generalized quasi-likelihood (GQL) based estimating equations to estimate the regression parameters for longitudinal data and use methods of moments to estimate the longitudinal correlations which follow a general correlation structure. Sutradhar et. al (2008) and Sutradhar (2010) have extended Sutradhar (2003) into longitudinal familial study where single or multiple random effects are considered. This GQL approach for familial study can not only handle general correlation structure of the responses including AR(1), but also variance components of the random effects.

Much more work is necessary to get a handle on the resulting unconditional covariance (correlation) structure of the repeated measures following binning under a variety of realistic scenarios. GQL could be one of the approaches to be used.

5.2.3 Asymptotic properties of binning estimates

In this thesis, we here mainly looked into the asymptotic properties of the binning estimates $\hat{\boldsymbol{\beta}}^{\star}$ (fixed effects) by marginal models, based on continuous normal response (linear model) with only independent and compound symmetry covariance structures. For complete matched data, we can prove that the estimated $\hat{\boldsymbol{\beta}}^{\star}$ is an unbiased estimate of $\boldsymbol{\beta}$. And the variance of $\hat{\boldsymbol{\beta}}^{\star}$ can be derived explicitly. However, it is much more complicated for mismatched data where the estimated $\hat{\boldsymbol{\beta}}^{\star}$ is biased, and the variance of $\hat{\boldsymbol{\beta}}^{\star}$ cannot be displayed explicitly except using weight matrices $\mathbf{A}_i$ and $\mathbf{B}_i$. The number of matches and mismatches between the response and covariates affect $var(\hat{\boldsymbol{\beta}}^{\star})$. In the future, we would like to look more into the asymptotic properties of the binning estimates for mismatched cases.

For conditional Poisson model, we have shown that under certain conditions, after binning, the binned data still follows a conditional Poisson model. In the future, we will look more into the asymptotic properties for binning estimates for conditional Poisson model.

5.2.4 Tree estimation

On occasions, GMRTree algorithm may not always converge when the underlying true model for the data is a GLMM model with main effects, and we fit a GMRTree model. A few reasons including linearization step and convergence criteria that may cause this divergence. We plan to investigate these issues further in the future. We will develop flexible and reasonable convergence criteria. We also might do some adjustments on the linearization step to get better approximation if possible.

In the future, we hope to explore the extension of existing consistency results from regression trees and GLMMs to GMRTree, checking whether parameters for the tree and variance components of random effects are estimated consistently.

CART uses exhaustive search algorithm to find the best split point and best split variable by searching over all the split variables and all the split points. However, this exhaustive search approach has selection bias toward variables which provide more split points (Doyle 1973). GUIDE almost provides an unbiased variable selection and also detects curvature. So in the future, we will explore to use GUIDE to estimate $f(\cdot)$ instead of CART. Or any other tree-based algorithms for continuous data can also be considered to estimate $f(\cdot)$. Methods such as bagging and boosting build on a tree structure as a way to improve predictive performance (see for example, Hastie et al., 2001), noting these tend to perform better than just fitting a single tree.

Another possible approach is to develop our own splitting criteria and grow the tree, instead of using linearized surrogate response and CART to fit the tree. We could build

our splitting criteria by using our marginal likelihood of GLMMs by integrating out the random effects. This approach is computationally intensive because of the integration. So one possibility is that instead of using the first overlygrown tree and then pruning it back, i.e., the approach of CART, we could use conservative penalty to grow the tree and also avoid pruning. This is an area of future research.

Bibliography

[1] Aalen, O.O., Borgan, Ø. and Gjessing, H.K. (2008). *Survival and Event History Analysis*. Springer: New York.

[2] Abdolell, M., LeBlanc, M., Stephens, D., and Harrison, R.V. (2002). Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. *Statistics in Medicine*, **21**, 3395-3409.

[3] Anderson, D.A. and Aitkin, M. (1985). Variance components models with binary response: interviewer variability. *Journal of the Royal Statistical Society, Series B*, **47**, 203-210.

[4] Böhning, D. (1998). Zero-inflated Poisson models and C.A.MAN: a tutorial collection of evidence. *Biometrical Journal*, **40**, 833-843.

[5] Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L. and Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society, Series A*, **162**, 195-209.

[6] Booth, J.G. and Hobert, J.P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B*, **26**, 211-252.

[7] Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and regression trees. Wadsworth, Belmont.

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

[9] Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *B*iometrika, **82**, 81C91.

[10] Chaudhuri, P., Huang, M.-C., Loh, W.-Y., and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, **4**, 143-167.

[11] Chaudhuri, P., Lo, W.-D., Loh, W.-Y. and Yang, C.-C. (1995). Generalized regression trees. *Statistica Sinica*, **5**, 641-666.

[12] Chipman, H., George, E.I., and McCulloch, R.E. (1998). Bayesian CART model search (with discussion). *Journal of the American Statistical Association*, **93**, 935-960.

[13] Cnaan, A., Laird, N.M. and Slasor, P. (1997). Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, **16**, 2349-2380.

[14] Cook, R.J., Lawless J.F. (2007). *The statistical analysis of recurrent events*. Springer: New York.

[15] Cook, R.J., Yi, G.Y, Lee, K.-A., and Gladman, D.D. (2004). A conditional Markov model for clustered progressive multistate processes under incomplete observation. *Biometrics*, **60**, 436-443.

[16] Cox, D.R. (1970). The Analysis of Binary Data. London: Chapman & Hall.

[17] Crouch, A. and Spiegelman, E. (1990). The evaluations of integrals of the form $\int_{-\infty}^{\infty} f(t)exp(-t^2)dt$: application to logistic-normal models. *Journal of the American Statistical Association*, **85**, 464-469.

[18] De'Ath, G. (2002). Multivariate regression trees: a new technique for modeling species-environment relationships. *Ecology*, **83**, 1105-1117.

[19] De'Ath, G. (2006). mvpart: Multivariate partitioning. R package version 1.2-4.

[20] Denison, D.G., Mallick, B.K., and Smith, A.F.M. (1998). A Bayesian CART algorithm. *Biometrika*, **85**, 363-377.

[21] Deng, D., and Paul, S.R. (2000). Score tests for zero inflation in generalized linear models. *Canadian Journal of Statistics*, **28**, 563-570.

[22] Dietz, E. and Böhning, D. (2000). On estimation of the Poisson parameter in zero-inflated modified Poisson models. *Computational Statistics and Data Analysis*, **34**, 441-459.

[23] Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data (Second Ed.)*. New York: Oxford University Press.

[24] Doyle, P. (1973). The use of Automatic Interaction Detector and similar search procedures. *Operational Research Quarterly*, **24**, 465-467.

[25] Dubin, J.A. and Müller, H-G. (2005). Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, **100**, 872-881.

[26] Fahrmeir, L. and Kaufmann, H. (1987). Regression models for nonstationary categorical time series. *Journal of time Series Analysis*, **8**, 147-160.

[27] Fan, J. and Marron, J.S. (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, **3**, 35-56.

[28] Fitzmaurice, G.M., Laird, N.M., Ware, J.H. (2004). *Applied Longitudinal Analysis*. Wiley: Hoboken.

[29] Galimberti, G. and Montanari, A. (2002). Regression trees for longitudinal data with time-dependent covariates. In K. Jajuga, A. Sokolowski, and H.-H. Bock, editors, *Classification, Clustering and Data Analysis*, pp. 391-398. Springer

[30] Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, **31**, 1208-1212.

[31] Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika*, **78**, 45-51

[32] Gu, M.G. and Kong, F.H. (1998) A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences*, **95**, 7270-7274.

[33] Hall, D.B. (2000). Zero-inflated Poisson and Binomial regression with random effects: a case study. *Biometrics*, **56**, 1030-1039.

[34] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320-338.

[35] Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York.

[36] Hsiao, W.-C. and Shih, Y.-S. (2007). Splitting variable selection for multivariate regression trees. *Statistics and Probability Letters*, **77**, 265-271.

[37] Izenman, A.J. (2008). Modern Multivariate Statistical Techniques. Springer; 1 edition.

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

[39] Jowaheer, V., Sutradhar, B.C. and Sneddon, G. (2009). On familial Poisson mixed models with multi-dimensional random effects. *Journal of Statistical Computation and Simulation*, **79**, 1043-1062.

[40] Kaysen, G.A., Dubin, J.A., Müller, H.G., Rosales, L.M., Levin, N.W., and The HEMO Study Group. (2000). The acute-phase response varies with time and predicts serum albumin levels in hemodialysis patients. *Kidney International*, **58**, 346-352.

[41] Laird, N.M. and Ware, J.H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, **38**, 963-974.

[42] Lambert, D. (1992). Zero inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

[43] Larsen, D.R. and Speckman, P.L. (2004). Multivariate regression trees for analysis of abundance data. *Biometrics*, **60**, 543-549.

[44] Lee, A.H., Wang, K. and Yau, K.K.W. (2001). Analysis of zero-inflated Poisson data incorporating extent of exposure. *Biometrical Journal*, **43**, 963-975.

[45] Lee, A.H., Xiang, L. and Fung, W.K. (2004). Sensitivity of score tests for zero-inflation in count data. *Statistics in Medicine*, **23**, 2757-2769.

[46] Lee, S.K. (2005). On generalized multivariate decision tree by using gee. *Computational Statistics and Data Analysis*, **49**, 1105-1119.

[47] Lee, S.K., Kang, H.-C., Han, S.-T., and Kim, K.-H. (2005). Using generalized estimating equations to learn decision trees with multivariate responses. *Data Mining and Knowledge Discovery*, **11**, 273:293.

[48] Lee, S.K. (2006). On classification and regression trees for multiple responses and its application. *Journal of Classification*, **23**, 123-141.

[49] Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with Discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619-678.

[50] Lee, Y. and Nelder, J.A. (2001). Modeling and analysing correlated non-normal data. *Statistical modelling*, **1**, 3-16.

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

[52] Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Assoication*, **91**, 1007-1016.

[53] Lindstrom, M.J. and Bates, D.M. (1988). Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association*, **83**, 1014-1022.

[54] Lindstrom, M.J. and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures. *Biometrics*, **46**, 673-687.

[55] Liu, Q.,and Pierce, D.A. (1993). Heterogeneity in Mantel-Haenszel-type Models. *Biometrika*, **80**, 543-556.

[56] Liu, Q. and Piece, A.D. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, **81**, 624-629.

[57] Loh, W.Y. (2002). Regression trees With Unbiased Variable Selection and Interaction Detection. *Statistica Sinica*, **12**, 361-386.

[58] Longford, N.T. (1993). Random Coefficient Models. Oxford, UK: Oxford University Press.

[59] McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**, 330-335.

[60] McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162-170.

[61] McCulloch, C.E., Searle, S.R. and Neuhaus, J.W. (2008). Generalized, linear, and mixed Models (2nd ed.). New York: Wiley.

[62] McGilchrist, C.A. (1994). Estimation in generalized linear mixed models. *Journal of the Royal Statistical Society, Series B*, **56**, 61-69.

[63] Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1-19.

[64] Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York.

[65] Monahan, J.F. and Stefanski L.A. (1992). Normal Scale Mixture Approximations to $F^*(z)$ and Computation of the Logistic-Normal Integral. In *Handbook of the Logistic Distribution*(ed. N. Balakrishnan), p529-540. Marcel Dekker, New York.

[66] Morgon, J.N. and Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, **58**, 415-434.

[67] Morgan, J.N. and Messenger, R. (1973). THAID: a sequential analysis program for the analysis of nominal scale dependent variables. The University of Michigan Survey Research Center, Ann Arbor.

[68] Muenz, L.R. and Rubinstein, L.V. (1985). Markov models for covariate dependence of binary sequence. *Biometrics*, **43**, 863-871.

[69] Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, **34**, 341-365.

[70] Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033-1048.

[71] Press, W.H., Flannery, B.P., Teukolsky, S. A. and Vetterling W. T. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd ed.* Cambridge, England: Cambridge University Press.

[72] Rice, J.A and Silverman, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* , **53**, 233C243

[73] Ridout, M.S., Demétrio, C.G.B., and Hinde, J.P. (1998). Models for count data with many zeros. *Proceedings of the XIXth International Biometrics Conference*, Cape Town, Invited Papers, 179-192.

[74] Schall, R. (1991). Estimation in Generalized Linear Models With Random Effects. *Biometrika*, **40**, 917-927.

[75] Segal, M.R. (1992). Tree-structured models for longitudinal data. *Journal of the American Statistical Association*, **87**, 407-418.

[76] Sela, R.J., and Simonoff, J.S. (2009) RE-EM Trees: A New Data Mining Approach for Longitudinal Data. NYU Working Paper No. SOR-2009-03.

[77] Solomon, P.J., and Cox, D.R. (1992). Nonlinear Components of Variance Models. *Biometrika*, **79**, 1-11.

[78] Stiratelli, R., Laird, N.M. and Ware J.H. (1984). Random effects models for serial observations with binary response. *Biometrics*, **40**, 961-971.

[79] Sutradhar, B.C. (2003). An overview on regression models for discrete longitudinal responses. *Statistical Science*, **18**, 377C393.

[80] Sutradhar, B. C., Jowaheer, V. and Sneddon, G. (2008). On a unified generalized quasi-likelihood approach for familial-longitudinal non-stationary count data. *Scandinavian Journal of Statistics*, **35**, 597C612.

[81] Sutradhar, B.C. (2010). Inferences in generalized linear longitudinal mixed models. *Canadian Journal of Statistics*, **38**, 174C196.

[82] Thall, P.F., Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, **46**, 657-671.

[83] Therneau, T.M. and Atkinson, M. J., Foundation M. An Introduction to Recursive Partitioning Using the RPART Routines. Sept. 3, 1997

[84] Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Assoication*, **81**,

[85] Ugarte, M.D., Ibáñez, B. and Militino A.F. (2004). Testing for Poisson zero inflation in disease mapping. *Biometrical Journal*, **46**, 526-539.

[86] Van den Broek, J. (1995). A score test for Zero-inflation in a Poisson distribution. *Biometrics*, **51**, 738-743.

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

[88] Vonesh, E.F. and Carter, R.L. (1987). Efficient inference for random-coefficient growth curve models with unbalanced data. *Biometrics*, **43**, 617-628.

[89] Wand, M.P. and Jones, M.C. (1995). Kernel Smoothing. Chapman & Hall: London.

[90] Wand, M.P. (1997). Data-based choice of histogram bin width. *The American Statistician*, **51**, 59-64.

[91] Wang, K., Yau, K.K.W and Lee, A.H (2002). A zero-inflated Poisson mixed model to analyze diagnosis related groups with majority of same-day hospital stays. *Computer Methods and Program iin Biomedicine*, **68**, 195-203.

[92] Ware J.H., Lipsitz S. and Speizer F.E. (1988). Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine*, **7**, 95-107.

[93] Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach, *Journal of Statistical Computation and Simulation*, **48**, 233-243.

[94] Wulfsohn, M.S. and Tsiatis, A.A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* 1997, **53**, 330-339.

[95] Xiang, L., Lee, A.H., Yau, K.K.W and McLachlan, G.J. (2006). A score test for Zero-inflation in correlated count data. *Statistics in Medicine*, **25**, 1660-1671.

[96] Xiang, L., Lee, A.H., Yau K.K.W. and McLachlan, G.J. (2007). A score test for overdispersion in zero-inflated poisson mixed regression model. *Statistics in Medicine*, **26**, 1608-1622.

[97] Xiong, X., Dubin, J.A. (2010) A binning method for analyzing mixed longitudinal data measured at distinct time points. *Statistics in Medicine*, **29**, 1919-1931.

[98] Yau, K.K.W and Lee, A.H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, **20**, 2907-2920.

[99] Yau, K.K.W., Wang, K. and Lee A.H. (2003). Zero inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrial Journal*, **45**, 437-452.

[100] Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121-130.

[101] Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects. *Journal of the American Statistical Association*, **91**, 61-76.

[102] Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasilikelihood approach. *Biometrics*, **44**, 1019-1031.

[103] Zhang H. (1998) Classification trees for multiple binary responses. *Journal of the American Statistical Association*, **93**, 180-193.

[104] Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a generalized quadratic model. *Biometrika*, **77**, 642-648.