# Methodology for Examining Differential Rates of Change for Longitudinal Data

Matthew Bryan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Patrick Heagerty, Chair

Elizabeth Brown

Elizabeth Sheppard

Program Authorized to Offer Degree:
School of Public Health - Biostatistics

University of Washington

**Abstract**

Methodology for Examining Differential Rates of Change for Longitudinal Data

Matthew Bryan

Chair of the Supervisory Committee:
Professor and Associate Chair Patrick Heagerty
Biostatistics

A common objective for longitudinal studies is to characterize differences in the rate of growth, or rate of change of an outcome across covariate-defined groups. The statistical challenges and potential extensions of models for comparing rates of change are intriguing with a broad scope for improving scientific research. We present research and newly proposed methodology on various scientific and statistical application of models for comparing rates of change in longitudinal outcomes across groups. We first discuss a generalized approach to modeling rates of change through direct structuring of differences in rates of change. The regression methodology offers a direct and parsimonious comparison of rates of change across groups and allows flexibility for structuring the underlying time trend of the outcome. By directly structuring rates of change relative to general time trend, power for detecting differences in the rate of change is improved compared to an equivalent linear models approach when the outcomes time trajectory is non-linear. Secondly, methodology for differentiating rates of change is extended to modeling multivariate longitudinal data. A common or global difference in the rate of change between groups is measured across all outcomes while separately structuring the time trend and mean level group differences for each outcome. When the true difference in the rate of change is similar for each outcomes, the global rate parameter method improves the ability to distinguish between groups compared to estimating separate rate effects for each outcome. Finally, the direct modeling of rates of change is made more robust to model misspecification by developing a semi-parametric

estimation approach. Non-parametric estimation of a smooth time trend function is incorporated with parametric estimation of differences in the rate of change. We describe methods for estimating the time trend non-parametrically based on penalized spline methodology. We illustrate the proposed methodology for longitudinal rates of change and its extensions using studies of growth in infant subjects. Models for comparing rates can also be applied to areas such as treatment trials and studies of environmental exposures. We conclude with a discussion of future areas of work and possible extensions for modeling rates of change using longitudinal data.

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

Table Number Page

# ACKNOWLEDGMENTS

The research for the following dissertation was partially supported by the NIH grants R01 HL072966 and ULI TR000423. The author wishes to thank the International Maternal Pediatric Adolescent AIDS Clinical Trials (IMPAACT) Group, grant UM1 AI068632, for providing access to the infant growth data from the HIVNET study, funded by National Institute of Allergy and Infectious Diseases of the NIH.

The author would also like to thank those who have provided him support during the process of writing this dissertation. Thank you to Patrick Heagerty for your advise and guidance for my dissertation and for my graduate studies in general. Thank you to Elizabeth Brown, Lianne Sheppard, and Sverre Vedal for serving as my committee members and offering helpful comments for improving my dissertation. To my parents, Jeff and Mary Kay, I offer thanks for everything you have provided me that has allowed me to get to where I am today. To my friends and family, I am grateful for all your support. Finally, I thank my wife, Michelle, for whom this dissertation could not be completed without. I am happy that we both completed this journey together.

# DEDICATION

To my beloved wife, Michelle, who is my constant inspiration for all that I do.

Chapter 1

# INTRODUCTION AND MOTIVATION

Longitudinal data is characterized by repeated measurements being taken over time on independent observations. By measuring observational units repeatedly, the information contributed by each unit increases relative to measuring each unit once which aids in distinguishing between independent units and, consequently, enhances differences between groups that categorize these units. In addition, repeated measurements make it possible for within-subject comparisons to be made in the sense that the trajectory of repeatedly measured variables can be mapped for each subject. Therefore, longitudinal data also provides information on the change over time of repeatedly measured variables.

Hence, when comparing differences between groups using longitudinal data, two basic and common scientific questions can be addressed. Do groups differ in the magnitude of an outcome, i.e. is one group bigger or smaller than the other(s)? Does the rate of change in the outcome over time differ between groups, i.e. does one group change or grow faster or slower than the other(s)? Methods for addressing the first question have been studied extensively, and standard methodology is well established. The second question has received considerably less attention, and a standard, general approach to the problem does not currently exist. Thus, we present methodology that offers a focused comparison of group differences in the rate of change for outcomes using longitudinal data. The proposed methodology provides a basis for discussing the advantages of direct consideration of differences in the rate of change across groups including promising extensions and adaptations for longitudinal rate models.

The work presented in this dissertation is organized into three chapters. Chapter 1 describes a general regression framework for modeling group differences in the rate of change for longitudinal data. In Chapter 2, we develop an extension to the regression methodology that can be used for multivariate longitudinal outcomes and testing for a global difference in the rate of change across groups. Finally, a semi-parametric adaptation for the proposed methodology which incorporates non-parametric estimation of a trend in time is presented in Chapter 3. Each chapter includes illustrations of the methodology using data from growth studies in infants since changes in growth is a natural way to interpret difference in the rate of change. In addition, we provide a final chapter to offer some concluding remarks and identify additional areas of work for modeling group differences in rate of change. As an overview of the topic under discussion, we summarize here the material of each chapter and the data applications used in this dissertation.

## *1.1 Overview of Chapter 2*

Methodology is presented for structured estimation of differences in the rate of change over time for an outcome across groups using a generalized framework. When developing a general model for comparing groups, it is common to begin by specifying structure for an underlying mean model. For example, methodology for a linear model specifies additive effects for mean level differences. The most common approach to modeling difference in the rate of change is by using a linear model and including appropriate interaction terms between covariate-defined groups and time. Thus, differences in rates of change are specified as mean level differences across groups and time. In contrast, the methodology introduced in this chapter begins with the direct structuring of the rate of change of an outcome with differences between groups being specified at the rate level. The direct structuring of rates of change facilitates construction of a model with a simple and parsimonious characterization for differences in the rate of change. In particular, comparisons of rates of change can be made while offering flexibility for specifying an underlying time trend for an outcome.

In the basic form of the method, differences in the rates of change across groups are structured as a proportional change in the rate between groups uniformly across time. We assume the underlying time trend structure is the same across groups and that only the magnitude of the rate of change across that time structure changes between groups. A full mean structure that corresponds to the specified rate structure is induced based on the rate assumption and includes a generalized adjustment for mean level differences across covariates. We outline both a mixed effect approach and an estimating equations approach for estimation of the mean structure. The model is generalized to allow time-varying covariates to be included in the model. The extension also provides a means for relaxation of the proportionality assumption. Diagnostic methods for model evaluation are discussed including a proposed score test for non-proportionality. The benefits of the longitudinal rate regression method are evaluated, illustrated, and discussed, particularly, in regard to the advantages of the method for outcomes with a non-linear trajectory over time.

## *1.2 Overview of Chapter 3*

One benefit of directly modeling of group differences in the rate of change of longitudinal data is its amenability to modeling group differences for multivariate longitudinal data. The goal of modeling group differences for multivariate longitudinal data is to borrow information across correlated outcome and across time in order to gain power to detect differences between the groups. The two common paradigms used in methods for study groups in a multivariate longitudinal setting are a latent variable approach and a global shared parameter approach. When the multivariate outcomes are viewed as describing some unmeasured or latent outcome, latent variable models can be used to directly model the association between the latent variable and covariates of interest with additional structure for relating the latent variable to the measured outcomes. Alternatively, when inference on the latent variable is not of interest or a latent variable does not exist for the outcomes, global shared parameter models specifies separate model structures for each outcome and each outcome model includes a parameter describing group differences that is assumed to be the same for each outcome in order to borrow strength across outcome. Of the two approaches, the global shared parameter scheme is the most straightforward extension of the methods discussed in Chapter 2. Therefore, we outline methods for examining a global difference in the rate of change for multivariate longitudinal data across covariate-defined groups.

The appeal of a global shared parameter model for multivariate longitudinal data is that different model structures can be specified for each outcome. However, the model structure for each outcome will contain a group parameter with a common interpretation for each outcome. In order to gain power to detect group differences, global shared parameter models assume that the effect size of the common group parameter is the same for each outcome and produces a common or global estimate of the group parameter for each outcome. For the proposed method for modeling rates of change, we describe a model where the difference in the rate of change is the same across outcome. Thus, the extended model will be most appropriate when it is reasonable to assume that the group effect on the rate of change is the same or similar across outcome. Separate structures are permitted for the time trend of an outcome and for mean level adjustment for covariates; a global estimate is produced only for rate level effects. We compare the power of the global estimation model to methods for estimating differences in the rate of change separately across outcomes.

### *1.3 Overview of Chapter 4*

When modeling rates of change for longitudinal data, the time structure of the outcome is an important consideration. Inference on the rate of change will be influenced by how the time trajectory is characterized. The methods outlined in Chapter 2 is appealing in regard to structuring changes in time since it allows a general function of time to be specified. The caveat of allowing a general time function to be specified is that the appropriateness of inference made on the rate of change may depend on a correct characterization of the time trend. When scientific understanding of an outcomes progression over time is limited or difficult to translate to a mathematical function of time, the likelihood of misspecifying the time trend for the outcome may be high. However, we may reduce the chances of a misspecified time trend by estimating a robust, smooth function of time that is generated based on the functional form of the data. We develop methodology for modeling group differences in the rate of change based on a robust time structure by estimating the trend in time non-parametrically. In addition to protecting against misspecification of the time trend, estimating a non-parametric function of time reduces burden on the user by no longer requiring a detailed time structure specification.

We adapt the methods from Chapter 2 to produce a semi-parametrically estimated model that estimates differences in the rates of change for a longitudinal outcome parametrically across groups based on a non-parametrically estimated trend in time. We propose using penalized splines for estimation of a non-parametric trend in time. Penalized splines is a reasonable means for non-parametric estimation in this instance since it is commonly used in existing semi-parametric methods. The time trend function is specified to be a linear combination of a large basis consisting of spline functions of time. The parameters of the linear combination are estimated by maximizing a likelihood function that is subject to a penalty. We describe the penalized likelihood estimation when using a smoothing penalty to control the roughness of the function. If *a priori* penalty parameter selection is not possible, the penalty parameter for the smoothing penalty can be selected by means of cross validation. The proposed methodology can similarly be implemented using alternative forms of penalization and procedures for selecting a penalty parameter.

## ***1.4 Summary of Data Examples***

### *1.4.1 Infant Growth Study*

The infant growth study is a secondary study from a clinical trial focusing on prevention of mother-to-child HIV transmission. Mothers were recruited during pregnancy and randomized to receive either zidovudine or nevirapine. The first infant born of the pregnancy was then followed and tested for HIV infection. Details and results for the primary aim of the clinical trial are presented in Jackson et al. [2003].

As a secondary aim, growth among the infants was measured longitudinally. A total of 622 infants were followed for 5 years from birth and measured as many as 16 times (7663 observations in total) for weight (Kg), crown-heel length (cm), and head circumference (cm). The goal of the study was to compare differences in growth across groups. Covariates used for comparison in this dissertation include sex (314 Female, 308 Male), treatment (306 zidovudine, 316 nevirapine), and whether HIV infection was detected (499 HIV negative, 123 HIV positive). Figure 1.1 provides Kaplan-Meier curves for time to HIV infection during the first 5 years after birth. The majority of HIV infections occurred pre- or peri-natally and therefore we treated HIV status as a baseline covariate in each illustration. The data from the infant growth study reflect the findings from Jackson et al. [2003] in that Nevirapine had a beneficial effect for the prevention of mother-to-child HIV transmission. There was little difference in the rate of HIV transmission between males and females.

Figure 1.1: Kaplan-Meier curves for HIV infection. Curves are separated into subgroups according to sex and treatment.

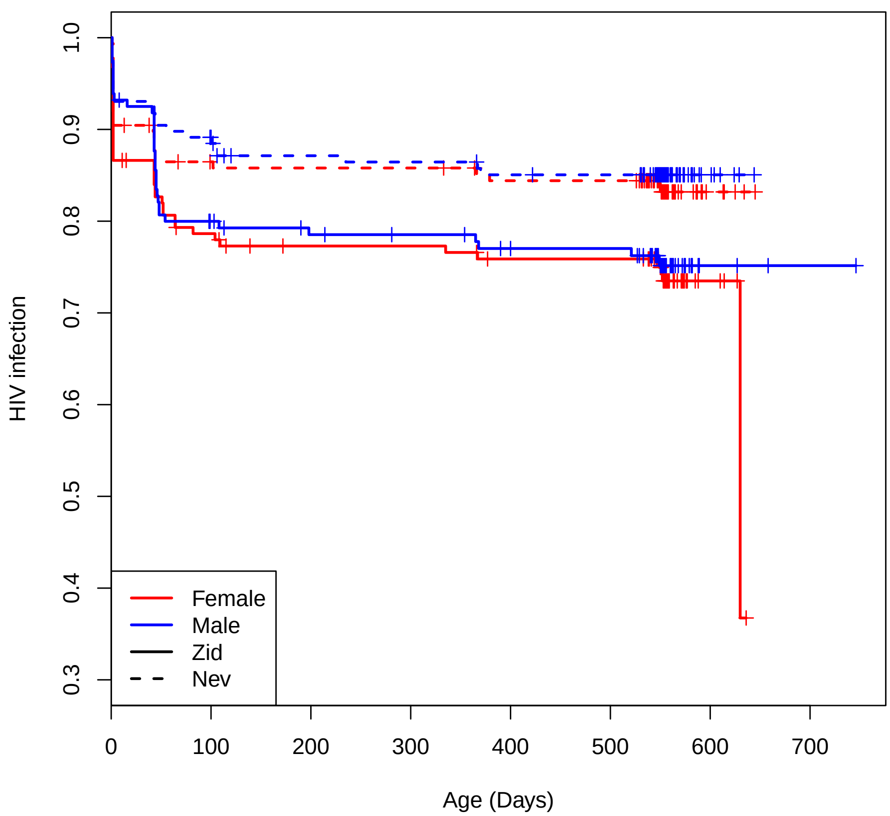


Figure 1.2: Scatter plots for growth outcomes over time. Unconstrained lowess smooth curves are included for subgroups defined by sex, treatment, and HIV infection status.

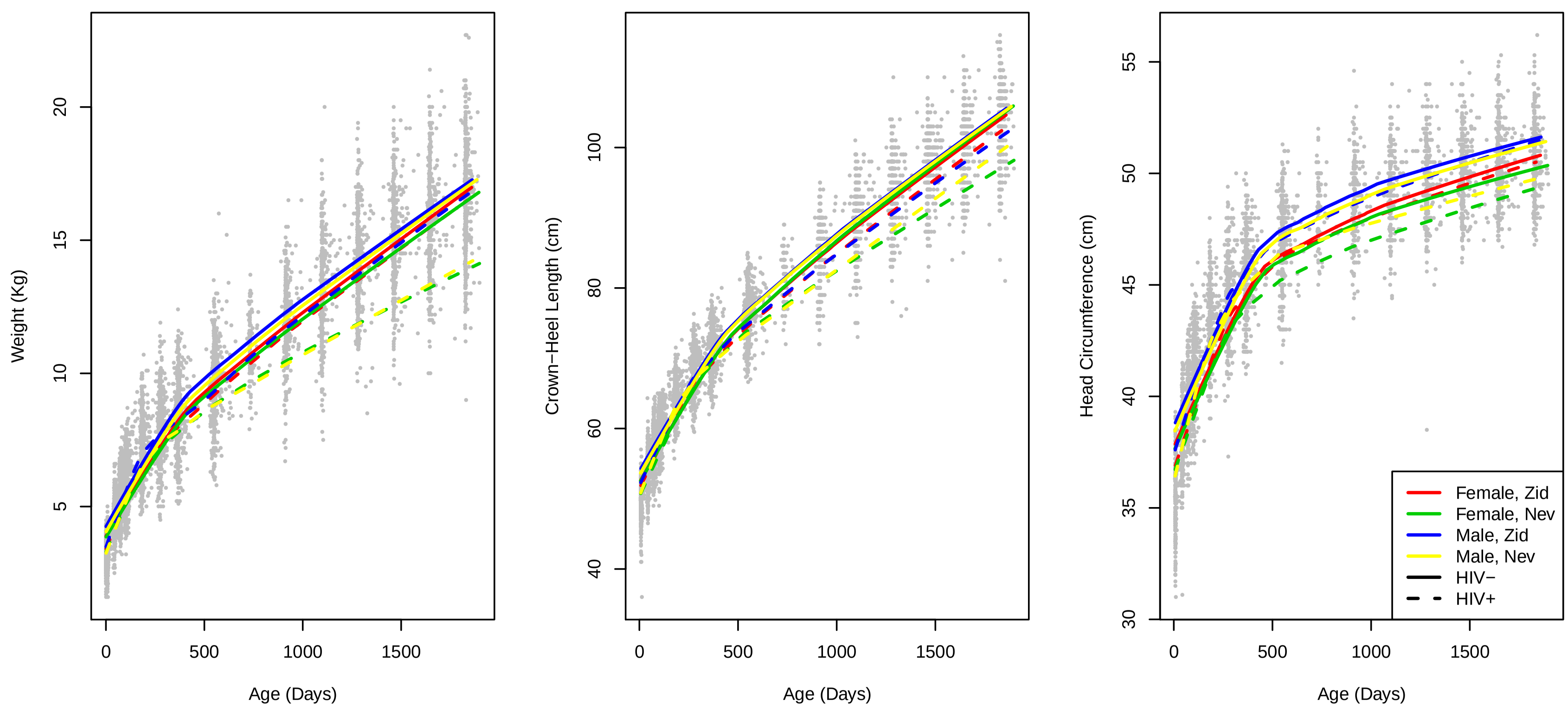


Table 1.1: Pairwise 5 year mean difference in growth outcomes among infants by Sex, treatment, and HIV infection status. Each observed growth outcome at 5 years was differed from the growth outcome observed at baseline for weight and at one week from basline for crown-heel and head circumference. Only infants who were observed at both time points were included in the analysis. Due to death and drop out, the mean differences were calculated based on 70.3%, 71.4%, and 71.3% of the sample who were observed for weight, crown-heel length, and head circumference respectively.

| | Sex | | Treatment | | HIV Status | |
|---|---|---|---|---|---|---|
| | Female | Male | Zidovudine | Nevirapine | HIV Negative | HIV Positive |
| Weight (Kg) | 13.30 | 13.57 | 13.52 | 13.36 | 13.56 | 12.37 |
| Crown-Heel Length (cm) | 53.63 | 53.36 | 53.34 | 53.64 | 53.90 | 50.26 |
| Head Circumference (cm) | 14.82 | 15.30 | 15.07 | 15.04 | 15.10 | 14.60 |

We now provide a descriptive comparison of the rate of change of the growth outcomes for infants across sex, treatment, and HIV infection status. Scatter plots of the three growth outcomes over time are presented in Figure 1.2. The most prominent differences in the rates based on the scatter plots is the decreased rate of growth in HIV positive infants compared to HIV negative infants. In Figure 1.2, the HIV positive lowess curve for a given subgroup represented by a dashed line grew slower than the corresponding lowess curve for HIV negative infants represented by a solid line. Differences in growth between the groups can be examined numerically by calculating mean differences between time points. In Table 1.1, we examine 5 year mean differences in the three outcomes and stratify by sex, treatment, and HIV infection status. In the sample, males experienced a larger mean difference in weight and head circumference over the five year period compared to females. Infants randomized to Nevirapine compared to those on Zidovudine had a larger mean difference for weight and head circumference. The growth trend for sex and treatment was reversed for crown-heel length with males and infants on nevirapine having a smaller mean difference in this outcome. The most striking result of Table 1.1 is the comparison of mean differences across HIV groups. For all three outcomes, infants who tested HIV positive experienced reduced growth on average compared to infants who tested negative. Growth was reduced in the first five years on average by approximately 1.20 Kg in weight, 3.64 cm in crown-heel length, and 0.51 cm in head circumference based on HIV infection status.

Some additional considerations include the fact that observations were missing for the growth variables: 20 for weight, 682 for crown-heel length, and 675 for head circumference. Majority of the missingness for crown-heel length and head circumference occurred because these measurements were not taken at baseline. For some infants, these outcomes were missing at all time points: this was the case for no infants for weight, 5 infants for crown-heel length, and 6 infants for head circumference. For the illustrations presented in this dissertation, analysis was carried out using the complete data without adjustment for missingness. Jackson et al. [2003] reported that some infants died during this study, but information on death was not included in the analyses for this dissertation. The presence of longitudinal truncation due to death potentially introduces bias in the analyses presented [Kurland and Heagerty, 2005].

### *1.4.2 Animal Nutrition Study*

In the animal nutrition study, 50 chicks were sampled and assigned to one of four protein diets. Diet 1 was assigned to 20 chicks, and 10 chicks were assigned to each of the other three diets. The chicks were measured for weight on the day of birth (Day 0) and every other day thereafter for the first 20 days as well as an additional measurement taken on day 21, i.e. 12 measurements for each chick. Five chicks in the sampled died during the study which potentially biases the analyses presented [Kurland and Heagerty, 2005]: 4 deaths in diet group 1, 1 death in diet group 4. Figure 1.3 displays a scatter plot of the observed weights over time. Lines are included for each diet group that connect the sample mean at each observed time point. Based on the plot, the rate of growth appeared to be largest for diet group 3 followed by diet groups 4, 2 and 1. There is some indication for changes in the rate of change for diet groups 3 and 4 with diet group 4 growing faster initially and diet group 3 growing faster later on. We examine group differences in growth numerically using 20 day mean differences from baseline presented in Table 1.2. The table results similarly

Figure 1.3: Scatter plot of chick weights during the first 21 days after birth. For each diet group, a line was drawn connecting the sample mean at each time point.

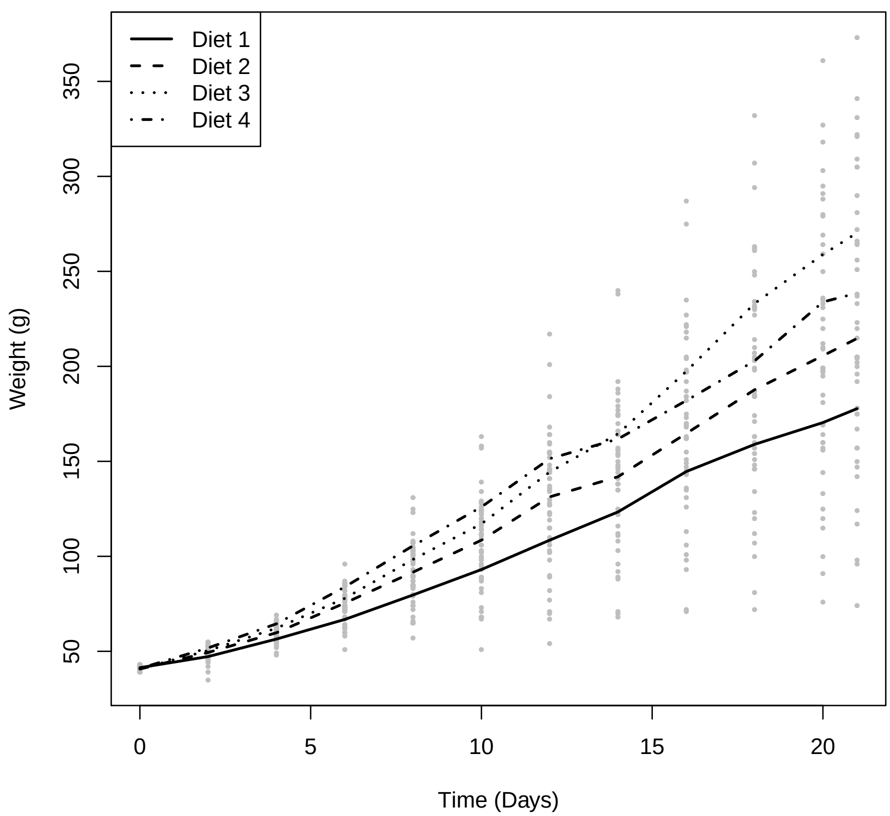


Table 1.2: Pairwise mean differences and standard deviations in weight among chicks for the first 20 days of life by diet groups. All chicks who were observed at baseline and day 20 were included in comparison (46 of 50 chicks).

| Diet | Mean Difference | Standard Deviation |
|---|---|---|
| 1 | 128.82 | 55.61 |
| 2 | 164.90 | 71.06 |
| 3 | 218.10 | 64.84 |
| 4 | 193.00 | 38.14 |

show the largest mean difference in weight over the first 20 days for diet group 3 followed by diet groups 4, 2, and 1.

## 1.5 Comments

The approach to modeling group differences in the rate of change for longitudinal data presented in this dissertation is motivated from a perspective of allowing inference for group differences that has a simple and meaningful scientific interpretation for differences in the rate of change. The straightforward interpretation is maintain while allowing added complexity to the model, particularly, with regard to the outcome's underlying time structure. Other approaches to modeling group differences in the rate of change can and should be considered. For example, the methods for comparing rates of change in this dissertation are very parametrically structured; therefore, an alternative of potential interest is a non-parametric approach that can compare rates of change with limited rate level assumptions. Ultimately, we hope that methodology presented in this dissertation will provide support for the utility of and the need for additional work in longitudinal methods that directly model rates of change.

Chapter 2

# REGRESSION MODELS FOR LONGITUDINAL RATES OF CHANGE

Comparing the rate of growth, or rate of change, across covariate-defined subgroups is a primary objective for many longitudinal studies. Few methods have directly addressed comparing differences in the rate of change for an outcome, particularly, when the longitudinal trajectory of the outcome is non-linear. In this chapter, we propose regression methodology for longitudinal rates of change that allows a direct, simple, and structured comparison of rates across subgroups even in the presence of a non-linear trend over time.

## *2.1 Introduction*

With repeated measures data obtained in a longitudinal study the response profile for different groups of subjects can be compared over time. In particular, longitudinal designs allow the direct linking of changes in exposure, such as medical treatments or environmental factors, and the corresponding changes in health indicators, such as measures of disease progression, symptom burden, or disease-specific functional status. Powerful regression approaches have been developed [Laird and Ware, 1982, Liang and Zeger, 1986] that directly allow inference regarding the mean outcome, and by incorporating time as a key covariate, longitudinal models also permit inference regarding rates of change. However, the standard regression approach *directly* focuses on the mean response through the inclusion mean level differences associated with each covariate. By allowing inclusion of time as a covariate, these regression models can then *indirectly* permit structuring of the differences in the rate of change over time by considering appropriate interaction terms at the mean level. The primary focus of this chapter is to directly specify a regression relationship linking the longitudinal rate of change with covariates. We shift the applied focus from starting with an emphasis on the mean response to beginning with a direct focus on the rate of change and on how this may differ across covariate subgroups, and then we indirectly obtain an induced mean model that characterizes mean profiles. The primary advantage of our approach is the ability to allow general time functions, and in a parsimonious regression fashion, we can directly and simply structure comparisons in the rate of longitudinal change.

Standard approaches to estimating models for longitudinal data include the linear mixed effect (LME) model [Laird and Ware, 1982], that incorporates random effects to account for correlation among repeated measures, and the generalized estimating equation (GEE) model [Liang and Zeger, 1986], where correlation is directly modeled and semi-parametric estimation is adopted. When the time trend in the mean is assumed to be linear, both LME and GEE approaches can characterize differences in rates of change through the inclusion of a group-by-time interaction. The coefficient for the interaction term provides a direct, and simple, way to compare rates across key subgroups. Typically, when a non-linear time trend is necessary, the outcome will be regressed on multiple functions of time such as with inclusion of polynomial terms or a more general parametric spline basis. In the non-linear situation, the comparison of longitudinal rates of change across subgroups requires the inclusion of multiple interactions between covariates and each function of time. In addition, comparison of rates of change is determined by the derivative of the multiple functions of time and no longer directly characterized by key parameters such as the interaction term when time is modeled linearly.

A general non-linear model framework [Davidian and Giltinan, 1995] can be used for comparing rates across covariate-defined groups. Applications of non-linear longitudinal models have typically been focused in biologically motivated settings such as with pharmacokinetic models, or other compartment models. For example, Wu, Ding, and De Gruttola [1998] used hierarchical non-linear methods to characterize disease process dynamics based on longitudinal measures of HIV viral loads. In the model suggested by Wu et al. [1998], change in infectious T-cells over time for patients on HIV treatment is expressed in terms of the current number of infectious T-cells, and similarly for infectious and non-infectious virions. Note that, in this approach, the rate at time $t$ is linked directly to the magnitude of the state (outcome) at time $t$. Non-linear models are also used for estimation of mechanistic growth models [Huang, 2011] where the rate of change is again expressed in terms of the current state. Huang [2011] discuss mechanistic growth models for characterizing growth of bacteria in relation to temperature. Pharmacokinetics (PK) is another important area of application for non-linear models where the primary focus is on the estimation of absorption and elimination rates in early clinical testing. For example, one simple approach for estimating PK parameters is through a one-compartment non-linear model as discussed by Lindsey et al. [2000].

Another approach to modeling longitudinal processes is provided by accelerated time (AT) methods where covariates are assumed to impact the time scale and either accelerate or decelerate longitudinal progression [Brumback and Lindstrom, 2004, Gray and Brookmeyer, 1998, 2000]. Shape invariant models (SIM) are an example of AT methods and have been applied in the area of speech and hearing. A detailed example is discussed by Brumback and Lindstrom [2004] where the position history of the tongue is recorded for participants repeatedly uttering a phrase. Brumback and Lindstrom [2004] allow a general monotonic transformation of the time scale. In addition, Gray and Brookmeyer [1998, 2000] used AT models as a means to link multivariate longitudinal outcomes in treatment trials for Alzheimer's Disease (AD). Linking multivariate outcomes allows evaluation of whether a treatment accelerates or decelerates deterioration associated with underlying disease progression. An accelerated time structure is particularly useful when the ultimate magnitude of progression for an outcome is considered to be fixed, perhaps by floor or ceiling effects, and covariates are expected to potentially affect the speed at which progression occurs. However, in many other applications a common magnitude of ultimate disease progression is not anticipated and research generally focuses on differences in the absolute rate of change across subgroups.

The model introduced in this chapter provides a method for comparing covariate groups in terms of the rate of change for a longitudinal outcome. We allow for inference regarding rates of change while adopting a flexible specification for the reference shape or pattern over time. In Section 2.2, we outline the mean structure of our longitudinal rate regression (LRR) model, for comparing rates across groups, and detail methods for estimation using either likelihood-based or estimating equation-based approaches. Section 2.3 presents simulation results comparing the proposed method to a standard linear mixed model approach. Our method is illustrated in Section 2.4 using the infant growth study and animal nutrition study. Finally, in Section 2.5, discussion of the LRR method is provided.

### 2.2 Methods

We use $Y$, $t$, and $X$ generically to respectively denote a longitudinal outcome, time, and a covariate defining groups of interest. We use subscripts $i$ to specify an independent observation for $i = 1, \ldots, N$ and $j$ to denote the time point of an observation with $j = 1, \ldots, n_i$. For convenience of notation, we assume that $Y$ is continuous and that modeling of the expected value of the untransformed outcome is of primary interest. Specifically, we are interested in differences in the rate of change of the expected value of $Y$ across groups defined by $X$. Initially, we assume that the covariate $X$ is not time varying. We first express the expected value of $Y$ given $X$ and $t$ as follows

$$E[Y_{ij} \mid X_i = x, t_{ij} = t] = g(x) + \mu_x(t)$$

where $g(\cdot)$ is a function dependent only on $X$ and $\mu_x(\cdot)$ denotes some function of $t$ for a given value of $X$ with the constraint that $\mu_x(0) = 0$ for all values of $X$. In this framework, we refer to $g(\cdot)$ as the baseline function which defines the mean of $Y$ at a time origin, or time zero, for a given value of $X$. The function $\mu_x(\cdot)$ describes the change in the expected value of $Y$ from this baseline value for each value of $X$. In defining the expected value of $Y$ in this way, the rate of change of the expected value of $Y$ is equivalent to the rate of change of $\mu_x(t)$.

Using the above notation, the specific aim of interest can be restated as an interest in being able to describe the difference between $\frac{\partial}{\partial t}\mu_{x_1}(t)$ and $\frac{\partial}{\partial t}\mu_{x_2}(t)$ across time for all potential values of $x_1$ and $x_2$. In order to structure the difference in rates, we consider a simple common relationship across time that assumes rates are proportional. We will also introduce non-proportional models that include covariate-by-time interactions, but first we assume that for any $X = x$ there exists a parameter $\theta$ such that

$$\frac{\partial \mu_x(t)}{\partial t} = (1 + \theta x)\frac{\partial \mu_0(t)}{\partial t} \tag{2.1}$$

for all values of $x$ (and $t$) where $\mu_0(t)$ is the time function for a preselected reference group, defined by $X = 0$. In this model, the rate of change in the expected value of $Y$ for a group defined by $X$, relative to the rate of change in the reference group ($X = 0$), is given by $(1 + \theta X)$ for any time $t$. We will refer to this simple regression assumption as the Proportional Rate (PR) assumption. The parameter $\theta$ can be interpreted as a percent increase (decrease) in the rate of change of the mean when $X = (x + 1)$ relative to when $X = x$. For example, if the value of $\theta$ is 0.1, then the group defined by $X = 1$ is associated with a 10% increase in the rate of change of $E[Y \mid X]$ compared to the rate of change in the group defined by $X = 0$. Note that the reference group is used to anchor the mean (and rate of change) over time, and we refer to $\mu_0(t)$ as the reference time function.

Under the PR assumption, a full specification for the mean structure of $Y$ given $X$ and $t$ can be induced by integrating the rate model over the interval $[0, t]$. That is,

$$\begin{aligned} E[Y_{ij} \mid X_i = x, t_{ij} = t] &= \int_0^t \frac{\partial \mu_x(s)}{\partial s} ds = \int_0^t (1 + \theta x) \frac{\partial \mu_0(s)}{\partial s} ds \\ &= g(x) + (1 + \theta x)\mu_0(t) \ . \end{aligned}$$

By generating a mean structure using the PR assumption, we are assuming that time periods of faster (or slower) change for an outcome occur at the same time across groups, but that groups may differ in their relative degree of change within any specific time period.

More generally, we can consider an outcome $Y$ measured at time $t$, and a vector of covariates $\mathbf{X} = (X_1, \cdots, X_p)$ where each covariate impacts the rate of change in the expected value of $Y$. We similarly define the conditional expected value of $Y$ as $E[Y_{ij} \mid \mathbf{X}_i = \mathbf{x}, t_{ij} = t] = g(\mathbf{x}) + \mu_{\mathbf{x}}(t)$ for functions $g(\cdot)$ and $\mu_{\mathbf{x}}(\cdot)$, where $\mu_{\mathbf{x}}(0) = 0$ for all values of $\mathbf{x}$. The extended PR assumption based on the vector of parameters $\boldsymbol{\theta}$ and the resulting mean structure becomes

$$\frac{\partial \mu_{\mathbf{x}}(t)}{\partial t} = (1 + \boldsymbol{\theta}^T \mathbf{x}) \frac{\partial \mu_0(t)}{\partial t} \tag{2.2}$$

$$E[Y_{ij} \mid \mathbf{X}_i = \mathbf{x}, t_{ij} = t] = g(\mathbf{x}) + (1 + \boldsymbol{\theta}^T \mathbf{x})\mu_0(t) \ . \tag{2.3}$$

Throughout this chapter, we will assume that the reference time function, $\mu_0(\cdot)$, can be adequately expressed as a linear combination of a parametric basis of time such as a polynomial basis or a regression spline basis: $\mu_0(t_{ij}) = \boldsymbol{\beta}^T \mathbf{T}_{ij}$ where $\mathbf{T}_{ij}$ is a vector of functions evaluated at $t_{ij}$. We will also typically express the baseline function as a linear combination of covariates i.e. $g(\mathbf{X}_i) = \boldsymbol{\alpha}^T\mathbf{X}_i$, although more general regression structures are possible.

We consider methods for estimation of the model where all mean parameters are estimated jointly. In some settings, joint estimation of the mean parameters may not be optimal for inference on the rate of change such as when longitudinal data is too sparse to estimate an accurate underlying time trend [Bennet and Wakefield, 2001]. To aid in accurate estimation of a reference time function, the time trend coefficients are estimated using all the data. Therefore, the estimated reference time function for a model actually represents the time trend for the reference group based on an averaged time trend across all groups. An averaged reference trend approach is a common technique in regression models, most notably in Cox Proportional Hazards regression [Anderson and Gill, 1982]. Estimation based solely on the reference group may also be considered particularly when strong *a priori* knowledge exists about the time trend among the reference group.

### *2.2.1 Mixed Effects Model*

We first detail estimation for the LRR model using a natural mixed effects approach that characterizes individual variation in both the level at baseline and in the rate of change. The model using a mixed effects approach is specified as follows:

$$Y_{ij} = (g(\mathbf{X}_i) + b_{0i}) + (1 + \boldsymbol{\theta}^T\mathbf{X}_i + b_{1i})\mu_0(t) + \epsilon_{ij} \tag{2.4}$$

where $b_{0i} \sim N(0, \tau_0^2)$, $b_{1i} \sim N(0, \tau_1^2)$, $\text{Corr}(b_{0i}, b_{1i}) = \rho$, and $\epsilon_{ij} \sim N(0, \sigma^2)$. Thus, the mixed model specification allows for individual random intercepts and individual variation in rates of change since, for subject $i$,

$$\frac{\partial}{\partial t}E[Y_{ij} \mid \mathbf{X}_i = \mathbf{x}, t_{ij} = t, b_{1i}] = (1 + \boldsymbol{\theta}^T\mathbf{x} + b_{1i})\frac{\partial \mu_0(t)}{\partial t}.$$

Although including separate intercepts and slopes is advantageous for interpretation, the inclusion of the random effect for the rate of change creates score equations for the mean and variance components that are not orthogonal as in the standard linear mixed model. Details regarding calculation of the score equations and the hessian matrix for this likelihood function are provided in Appendix A to this dissertation. The non-orthogonality is a result of the induced interaction between the random slope and the reference time function and is primarily an issue in regard to implementation. Due to the induced dependency, maximizing the LRR model using mixed effects can be challenging and highly reliant on appropriate initial values.

### *2.2.2 Estimating Equations Approach*

Alternatively, the LRR model can be estimated using semi-parametric methods. First, since the natural mixed model specification includes random effects on the linear scale, an induced marginal mean retains the same parametric form as the random effects mean structure. Taking expectations of the outcome vector, $\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{in_i})$, over $\mathbf{b}_i$ and $\epsilon_{ij}$ for the conditional mean given in Equation (2.4) yields the marginal mean

$$E[\mathbf{Y}_i \mid \mathbf{X}_i = \mathbf{x}, \mathbf{t}_i = \mathbf{t}] = \mathbf{1}_{n_i} g(\mathbf{X}_i) + (1 + \boldsymbol{\theta}^T \mathbf{X}_i)\mu_0(\mathbf{t}_i).$$

where $\mu_0(\mathbf{t}_i) = \mathbf{T}_i \boldsymbol{\beta}$ is a vector of linear combinations of a parameteric basis for the time vector $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})$ and $g(\mathbf{X}_i) = \alpha^T \mathbf{X}_i$. Second, provided a working correlation or covariance model can be adopted, solutions to the following estimating equations are used to estimate all mean parameters

$$\mathbf{G}(\boldsymbol{\theta}, \mu_0(t), \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{i=1}^{N} \mathbf{D}_i^T \mathbf{W}_i^{-1} \{\mathbf{Y}_i - [\mathbf{1}_{n_i} g(\mathbf{X}_i) + (1 + \boldsymbol{\theta}^T \mathbf{X}_i)\mu_0(\mathbf{t}_i)]\} \tag{2.5}$$

where $\mathbf{D}_i$ is the vector of derivatives for the mean structure for parameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$; and $\mathbf{W}_i = \mathbf{W}_i(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ is the working covariance model that is possibly dependent on some additional parameters $\boldsymbol{\gamma}$. Finally, provided a consistent estimate, $\hat{\boldsymbol{\gamma}}$, for the working covariance parameters, results from Liang and Zeger [1986] show that the solutions for $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ to the estimating equation given in (2.5) converge in distribution to a normal distribution under mild regularity conditions (such as growth of information), parameter estimates are consistent, and the asymptotic variance is given by the standard sandwich formula:

$$\mathbf{V} = \left(\sum_{i=1}^{m} \mathbf{D}_i^T \mathbf{W}_i^{-1} \mathbf{D}_i\right)^{-1} \left[\sum_{i=1}^{m} \mathbf{D}_i^T \mathbf{W}_i^{-1} cov(\mathbf{Y}_i) \mathbf{W}_i^{-1} \mathbf{D}_i\right] \left(\sum_{i=1}^{m} \mathbf{D}_i^T \mathbf{W}_i^{-1} \mathbf{D}_i\right)^{-1}. \tag{2.6}$$

The necessary regularity conditions are outlined by Liang and Zeger [1986] and are explored in greater detail by Crowder [1986]. The variance given by Equation (2.6) is typically estimated by substituting the empirical estimate of $cov(\mathbf{Y}_i)$ into the equation .

The estimating equations approach is often attractive when the primary scientific focus is on the regression parameters since robustness to variance model and/or distributional assumptions is provided. In contrast, consistency of both point and variance estimates cannot be guaranteed in the LRR mixed effects modeling framework unless the model is correctly specified.

### *2.2.3 Time Varying Covariates*

For fixed covariates such as demographic characteristics or fixed treatment groups, the proposed LRR modeling framework can be useful for quantifying differences in rates of longitudinal change. However, there are many potential exposures of scientific interest that vary over time, and allowing incorporation of time-varying covariates into analysis is important. Although caution must be exercised with time-dependent covariate analysis (see Diggle et al. [2002], chapter 12 for overview), we outline methods for including time varying rate covariates in the LRR model. One important use of a time-dependent covariate is to allow a relaxation of the proportional rate assumption by including covariate-by-time interactions in the LRR model.

In order to characterize a LRR model with time-dependent covariates, we first consider a simple binary time varying covariate that represents an exogenous and discrete exposure that is delayed from baseline. An example exposure would be a treatment given in a controlled crossover trial. Let $X_i(t)$ denote the covariate status for individual $i$ who was exposed at specified time $t_1$. That is,

$$X_i(t) = \begin{cases} 0 & \text{for } t < t_1 \\ 1 & \text{for } t \geq t_1 \end{cases} .$$

Adopting the PR assumption of equation (2.1) with parameter $\theta$, the induced mean structure can be obtained by dividing the integration of the time function over the key time periods associated with changes in exposure. For simplicity, we focus on the case of no baseline mean differences in the outcome across groups (i.e. $g(x) = \alpha_0$ for all values of $x$). Extension to the case of baseline differences is straightforward. The mean function at times prior to $t_1$ will be identical to the reference mean structure up to a constant defined by the baseline function $g(x)$. For the outcome of individual $i$ observed at time $t_j \geq t_1$, the expected value of $Y_{ij}$ is calculated as

$$\begin{aligned} E[Y_{ij} \mid X_i(t) = x(t), t_{ij} = t] &= \int_0^t \left\{[1 + \theta x(s)] \frac{\partial \mu_0(s)}{\partial s}\right\} ds \\ &= \int_0^{t_1} \frac{\partial \mu_0(s)}{\partial s} ds + \int_{t_1}^t \left[(1 + \theta) \frac{\partial \mu_0(s)}{\partial s}\right] ds \\ &= \alpha_0 + \mu_0(t_1) - \mu_0(0) + (1 + \theta)\mu_0(t) \\ &\quad -(1 + \theta)\mu_0(t_1) \\ &= \alpha_0 + (1 + \theta)\mu_0(t) - \theta\mu_0(t_1). \end{aligned}$$

By induction, we can then extend the induced mean to the scenario involving a time varying covariate taking multiple values across multiple time points: $t_0 = 0, t_1, \cdots, t_p$. For an outcome measurement time $t$ contained in any time interval $t \in (t_k, t_{k+1})$, the mean becomes

$$\begin{aligned} E[Y_{ij} \mid X_i(t) = x(t), t_{ij} = t] &= \alpha_0 + [1 + \theta x(t)]\mu_0(t) - \theta x(t_k)\mu_0(t_k) \\ &\quad -\theta \sum_{l=1}^{k} x(t_l)[\mu_0(t_l) - \mu_0(t_{l-1})]. \end{aligned} \tag{2.7}$$

The third and fourth term in equation (2.7) ensures that the mean function remains continuous at the time points where the covariate is changing. For covariates that only change at discrete time points, the LRR model can easily incorporate such variables into the rate model. However, if covariates are given by a continuous process then additional computational burden is required to numerically derive the induced mean function. The integration across time would need to be done with consideration of the continuous process for the time-varying covariate. Finally, for some covariates that are only measured at select times for which an underlying continuous process represents their true time-varying state, covariate values can be considered to be measured with error in between measurements using the values of the nearest measurement. Additional work is needed to incorporate the resulting covariate measurement error associated with incomplete measurement (see Carroll et al. [2006], chapter 11).

### 2.2.4 Diagnostics

When using the LRR model, it is appropriate to assess the adequacy of the proportional rate assumption. Standard graphical evaluation of residuals, and formal testing approaches to model checking are both feasible. First, the PR assumption can be evaluated graphically by generating standard residual plots against time for each covariate group defined by a rate covariate $X$. Any underlying trend in the residuals would provide evidence against the validity of the PR assumption. Alternatively, residuals can be generated that are specifically designed to illustrate violations of the proportionality of rates. Given a LRR model with mean structure specified as in Equation (2.3), we can generate an observational-level rate residual for parameter $\theta_q$ for grouping variable $X_q$ based on the equation

$$r_{qij} = \frac{Y_{ij} - g(\mathbf{X}_i)}{\mu_0(t_{ij})} - 1 - \boldsymbol{\theta}^T_{(-q)}\mathbf{X}_{(-q)i} \tag{2.8}$$

where the subscript $(-q)$ denotes a vector with the $q$th entry removed. The rate residual given by Equation (2.8) has the property that observations in the reference group for the $q$th covariate, $X_q = 0$, will be centered around 0. In addition, observations in the group defined by $X_q = x$ will be centered around the value $\theta_q x$. Therefore, the PR assumption can be evaluated using the rate residuals by plotting the residuals against time and examining whether the clusters of rate residuals are parallel across groups defined by $X_q$. One drawback of the proposed rate residual is that they will be unstable for observations at or near the time origin, $t = 0$, since $\mu_0(0) = 0$. Therefore, observations at the origin must be discounted for this comparison, and observations near the origin must either be discounted or extremely down-weighted for their influence on diagnostic decisions.

Graphic displays of residuals can be highly useful for providing visual validation for modeling assumptions. However, diagnostic residual plots are only subjectively interpreted. Therefore, formal tests that consider focused departures from the proportional rate assumption provide an objective model evaluation tool. Paralleling methods developed for the Cox model [Grambsch and Therneau, 1994], we outline an approach for testing the adequacy of the PR assumption using a score test where the alternative is given by a linear change over time in the difference of rates of change across covariate groups. For this test, we specify an extension to the LRR method that includes a group-by-time interaction in the

PR assumption:

$$\frac{\partial \mu_x(t)}{\partial t} = (1 + \theta x + \psi t x)\frac{\partial \mu_0(t)}{\partial t}. \tag{2.9}$$

In addition to detecting linear violations, the proposed structured alternative is suitable for detecting monotonic deviations from the PR assumption. Tests for more complex violations may require development of additional diagnostic tests.

If the standard PR assumption given by Equation (2.1) is correct, then the interaction parameter, $\psi$, given in Equation (2.9) would be zero. Thus, we test the hypothesis $H_0$ : $\psi = 0$. The advantage of using a score test is that test statistic is calculated under the null permitting inference without the need for an extended (alternative model) fit. The score test only depends on the ability to compute the score equations and information for the extended model. Below, we derive the needed analytical components by integrating the modified rate assumption to obtain the induced mean model. Use of integration by parts yields the following general mean structure:

$$E[Y_{ij} \mid X_i = x, t_{ij} = t] = g(x) + (1 + \theta x + \psi t x)\mu_0(t) - \psi x A_\mu(t)$$

where $A_\mu(t) = \int_0^t \mu_0(s)ds$ i.e. $A_\mu(t)$ is the area under our reference time function over the interval $[0, t]$.

Here, we provide results necessary for calculating score equations for parameter $\psi$; all other score equations will be zero under the null hypothesis. Define

$$\mathbf{D}_{\psi i} = X_i d_i \boldsymbol{\mu}_i - X_i \mathbf{A}_{\boldsymbol{\mu} i}$$

where $d_i$ is a diagonal matrix whose diagonal entries are the time values associated with $\mathbf{Y}_i$, and $\boldsymbol{\mu}_i$ and $\mathbf{A}_{\mu i}$ denote the vectors whose values are the functions $\mu_0(t)$ and $A_\mu(t)$ respectively evaluated at each time value. The matrix $\mathbf{D}_{\psi i}$ here denotes the vector of derivatives of the mean function with respect to the parameter $\psi$ evaluated at each time point for individual $i$. If $\widehat{\theta}$, $\widehat{\beta}$, and $\widehat{\alpha}$ denote the estimates from the standard rate regression model or the null model, then we express the score equation for the parameter $\psi$, denoted by $U_\psi(\theta, \beta, \alpha, \psi)$, evaluated under the null as

$$U_\psi(\widehat{\theta}, \widehat{\beta}, \widehat{\alpha}, 0) = \sum_{i=1}^{m} \widehat{\mathbf{D}}_{\psi i}^T \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{R}}_i$$

where $\widehat{\mathbf{D}}_{\psi i}$, $\widehat{\mathbf{V}}_i$, and $\widehat{\mathbf{R}}_i$ are the estimated derivative vector, variance matrix, and residual vector for $\mathbf{Y}_i$ based on the estimates from the null model. The score test statistic can then be expressed as

$$S_\psi = U_\psi(\widehat{\theta}, \widehat{\beta}, \widehat{\alpha}, 0)^T I_{22\cdot 1}^{-1} U_\psi(\widehat{\theta}, \widehat{\beta}, \widehat{\alpha}, 0)$$

where $I_{22\cdot 1}$ is the information for $\psi$ under the null given by the formula $I_{22\cdot 1} = (I_{22} - I_{12}^T I_{11}^{-1} I_{12})$ based on the decomposition of the information matrix. The matrix $I_{11}$ can be estimated by the hessian matrix from the null LRR model. Estimating the cross information between the score for $\psi$ and the other mean parameters, given by $I_{12}$, and the marginal information of the score for $\psi$, given by $I_{22}$, requires taking the derivative of $U_\psi(\theta, \beta, \alpha, \psi)$ with respect to $\theta$, $\beta$, $\alpha$, and $\psi$.

For the single rate covariate test outlined above, the score statistic can be compared to a $\chi^2$ distribution with degree of freedom equal to the dimension of $X$. The test provides a means for testing the PR assumption based on a structured alternative. In this case, the alternative structure specifies linear changes in the rate of change due to group across times. Other structured alternatives could be considered.

### *2.3 Simulation Studies*

The LRR model using likelihood-based estimation was compared to a LME model approach in simulation studies. The LME model was chosen for comparison since it is commonly used, and can be adapted to compare rates of change by including appropriate covariate-by-time interactions. We focused evaluation on whether comparison of rates of change across groups using a direct and parsimonious LRR model provided more power than an LME approach which may require additional covariate-by-time interactions to characterize group differences. The two methods were compared where the reference trend over time was assumed to be a cubic polynomial function. A single, binary covariate was used for comparison of rates across group. The LRR and LME parameterizations can be expressed respectively as follows:

$$Y_{ij} = \alpha_0 + \alpha_1 X_i + b_{0i}^* + (1 + \theta X_i + b_{1i}^*)[\beta_1^* t_{ij} + \beta_2^* t_{ij}^2 + \beta_3^* t_{ij}^3] + \epsilon_{ij}^* \tag{2.10}$$

$$\begin{aligned} Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \beta_4 X_i + \beta_5 X_i t_{ij} + \beta_6 X_i t_{ij}^2 + \beta_7 X_i t_{ij}^3 \\ + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}. \end{aligned} \tag{2.11}$$

The goal of the simulation studies were to evaluate any potential gain in power to detect group differences in the rate of change of an outcome for the LRR model relative to the LME model. Gain in power was anticipated due to the difference in degrees of freedom of the test for these two approaches. For the LRR model in Equation (2.10), differences in rates of change were tested by the hypothesis $H_0 : \theta = 0$, a one degree of freedom test, while the LME specification in Equation (2.11) required a test of the hypothesis $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$, a three degree of freedom test.

Two simulations were conducted based on data generated under the LRR model and under the LME model using a mean specification that was compatible with both models. That is, using notation in Equations (2.10) and (2.11), $\alpha_0 = \beta_0$, $\beta_1^* = \beta_1$, $\theta\beta_1^* - \beta_1^* = \beta_5$, and so forth. First we simulated data under the LRR mixed model specification given in Equation (2.10). However, due to the non-linear LRR specified random effects structure, standard choices for linear mixed model random effects resulted in misspecification of the covariance structure for the LME model. Therefore, robust sandwich estimators were used for the LME model under this scenario to ensure the test was the correct size. Alternatively, one may specify a more compatible variance structure for the LME model by including four random effects for intercept, linear time, quadratic time and cubic time. We chose the sandwich variance approach since including random effects for intercept and linear time only is more common practice and the more involved random effects structure will be less likely to converge. Second, we simulated date under the LME model specified in Equation (2.11). Under this simulation, the random effects structure of the LRR model will be misspecified, but no adjustment for misspecification was made for this simulation since the size of the test for the LRR model was reasonably close to the nominal 5% level (see Table 2.1). All estimates for the LME model were calculated using the *lme* function in R. Simulations involved 1000 replicated datasets.

Table 2.1: Parameter settings and results for two simulations comparing the LRR model to the LME model. Simulation 1 generated data based on the LRR model with a random effect for the intercept and slope. The variance for LME model was adjusted using the sandwich estimator to ensure appropriate size in simulation 1. Simulation 2 generated data based on the LME model with a random effect for the intercept and linear time. For both simulations, a cubic polynomial equation was used to model time and bivariate covariate was used to estimate differences in the rate of change of the outcome. *Variance for the random effect for slope. **Variance for the random effect for linear time.

| | Simulation 1 | Simulation 2 |
|---|---|---|
| True Model Structure | LRR | LME |
| Sample Size | 100 | 100 |
| Time Structure ($\beta_1^*, \beta_2^*, \beta_3^*$) | 0.11, -0.0002, -0.00005 | 0.12, -0.002, -0.00002 |
| Intercept and Main Effect ($\alpha_0, \alpha_1$) | 4, -0.06 | 4, 0.02 |
| Difference in Rates ($\theta$) | -0.1 | -0.05 |
| Random Effects Covariance ($\tau_0^2$, $\tau_1^2$, $\rho$) | 0.005,0.05*,-.78 | 0.006,0.00001**,0 |
| Measurement Error Variance ($\sigma^2$) | 0.005 | 0.07 |
| Size of Test (LRR, LME) | .047, .051 | .044, .041 |
| Power of Test (LRR, LME) | .637, .455 | .941, .848 |
| Failure Rate (LRR, LME) | 0, 0 | .012,.008 |

Results from both simulations are presented in Table 2.1. For the first simulation scenario using the specified parameters and sample size given in Table 2.1, we find that LRR testing has 63.7% power using a focused one degree of freedom test while LME has only 45.5% power using a required three degree of freedom test. In the second simulation scenario when data was generated from the LME model, the one degree of freedom test from LRR model again showed higher power at 94.1% compared to 84.8% for the three degree of freedom test using the LME model. These simulation results demonstrate the potential advantages of a regression model that directly structures the longitudinal rate of change when this aspect is the primary target of inference.

Additional simulations (not shown) were conducted to verify that the model diagnostic test outlined in subsection 2.4 obtained proper size and had adequate power to detect violations of the proportional rate assumption. In simulations, the type I error for the score test was appropriate (observed as 4.2% for a nominal 5% level test based on 1000 replicates) and reliably detected linear violations of the PR assumption when the null model did not hold.

## *2.4 Application*

### *2.4.1 Study of Infant Growth*

For illustration using the infant growth example, we focused on weight as the outcome of interest to address whether the rate of change of weight among these infants differs across groups defined by sex, treatment, and HIV infection status. A LRR model for infant weight was constructed using a natural cubic spline basis for the reference time function with knots at 150 and 500 days. Both main effects and rate effects were included for sex, treatment and HIV infection status.

The results from the LRR model with two knots are presented in Table 2.2. The main effect estimates indicated large baseline differences in mean weight associated with sex and treatment. Males were estimated to weigh 0.26 Kg more than females on average at birth (95% CI = (0.17, 0.34)). Infants whose mothers were randomized to nevirapine treatment tended to be 0.12 Kg lighter than those infants on zidovudine (95% CI = (-0.21, -0.03)). There was little mean difference at baseline in weight among groups defined by HIV status. However, there was strong evidence of a decrease in the rate of change for weight among infants infected with HIV. Infants who were HIV positive were estimated to have a decreased rate of change in weight of 15% (95% CI = (-18%, -12%) compared to infants that were HIV negative. In other words, the LRR model estimates that, for a period of time where HIV negative infants would be expect to increase in weight by 1 Kg, the average weight of HIV positive infants is estimated to increase by 0.85 Kg. Therefore, the model provides evidence that HIV infection reduces growth among infants. Results from the primary trial for the infant growth study published in Jackson et al. [2003] showed that the rate of death was significantly higher among the HIV positive infants. Since death was not accounted

Table 2.2: Model results comparing rates of change for weight among infants of HIV infected mothers. Weight was regressed on sex, treatment, and HIV infection status in two LRR model. The model estimated main effects and rate effects for the three covariates using the standard LRR model. The reference time function for the model was estimated using a natural cubic spline basis with knots at 150 and 500 days.

| | Estimate | 95% CI |
|---|---|---|
| Main Effects | | |
| Intercept | 3.38 | (3.29, 3.46) |
| Sex (Male) | 0.26 | (0.17, 0.34) |
| Treatment (Nev) | -0.12 | (-0.21, -0.03) |
| HIV Status | 0.04 | (-0.07, 0.16) |
| Time Trend | | |
| Basis 1 | 6.41 | (6.24, 6.58) |
| Basis 2 | 15.23 | (14.87, 15.59) |
| Basis 3 | 11.41 | (11.14, 11.68) |
| Rate Effects | | |
| Sex (Male) | 0.02 | (-0.01, 0.04) |
| Treatment (Nev) | -0.01 | (-0.03, 0.02) |
| HIV Status | -0.15 | (-0.18, -0.12) |

for in the LRR model and weight is most likely negatively associated with death, the rate estimate for HIV infection status is likely conservative. However, the bias in the estimate as a result of infant death will be reduced as a result of the random effects structure used to capture individual variation. Since the random effects structure generates an individual time curve for each individual, each individual's time trajectory will influence the overall mean curve even at times when an individual was not observed due to death. There was little difference in the rate of change across groups defined by sex and treatment. The time trend among HIV positive and negative infants is illustrated in Figure 2.1.

To evaluate the goodness-of-fit of the LRR model, we produced residual plots for the

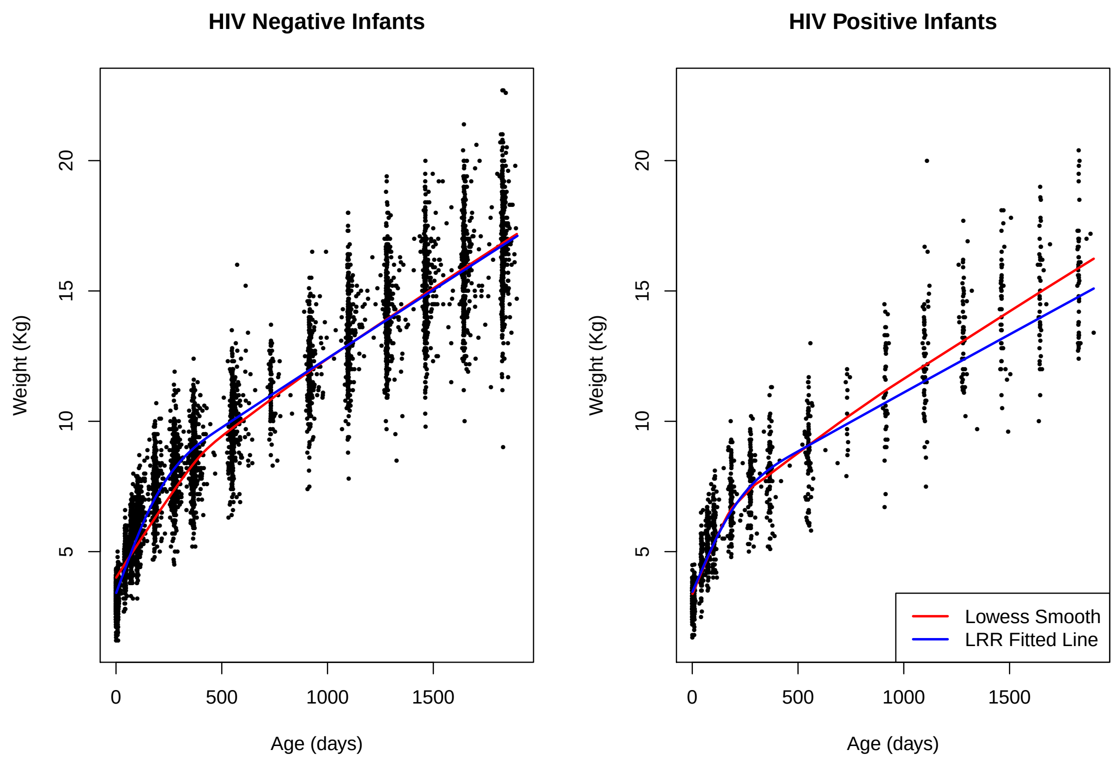



Figure 2.1: Changes in infant weight by HIV status. The red line represents a lowess smooth curve for the scatter plot. The blue line represents the fitted line for the LRR model presented in Table 2.2.

infant data based on the fitted model. Figure 2.2 shows the three sets of residual plots for the covariates sex, treatment, and HIV status. The residual plots show little evidence for a poor model fit except for a small lack of fit for the HIV positive group. We can also evaluate for violations of the PR assumption through plots of the rate residuals described by Equation (2.8). Rate residual plots for the rate parameter for HIV status are provided in Figure 2.3. The first plot illustrates the spread of the rate residuals with relatively large variation among points near the origin. The lowess smooth curves presented in each plot heavily down-weighted observations near day zero since these residuals are highly unstable. The trend among the residuals in the second plot suggests mild violations of the proportionality

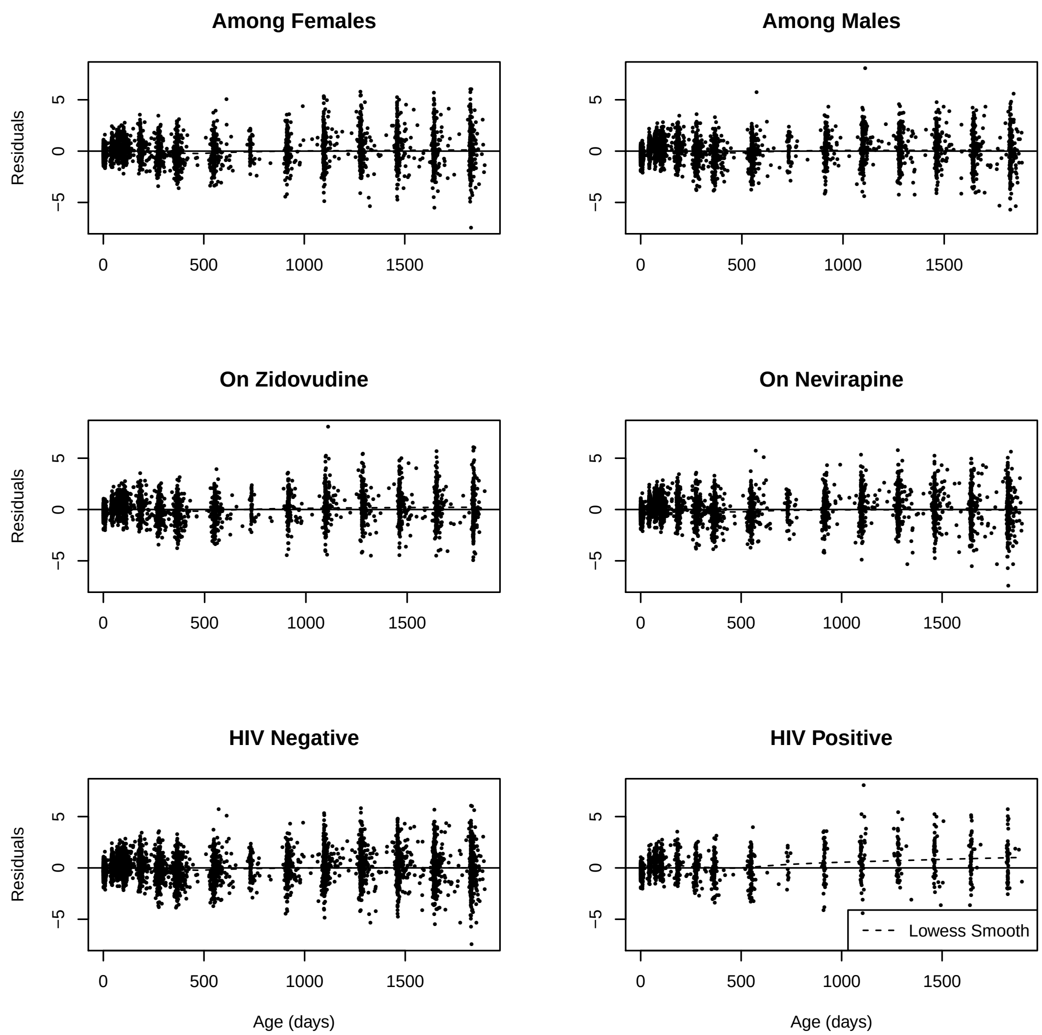



Figure 2.2: Residual plots for the model of weight among infants exposed to HIV infection across grouping variables based on the LRR results from Table 2.2. The dashed line in each plot represents a lowess smooth curve for the residuals.

assumption since the lowess lines taper toward each other at older age. The residual plot is more suggestive of misspecification of the reference time trend since both lowess smooth lines have highly non-linear shapes. The rate residual plots for sex and treatment (not presented) were less informative for evaluating the PR assumption since their rate effect sizes were relatively small. The score test for the PR assumption is highly significant for

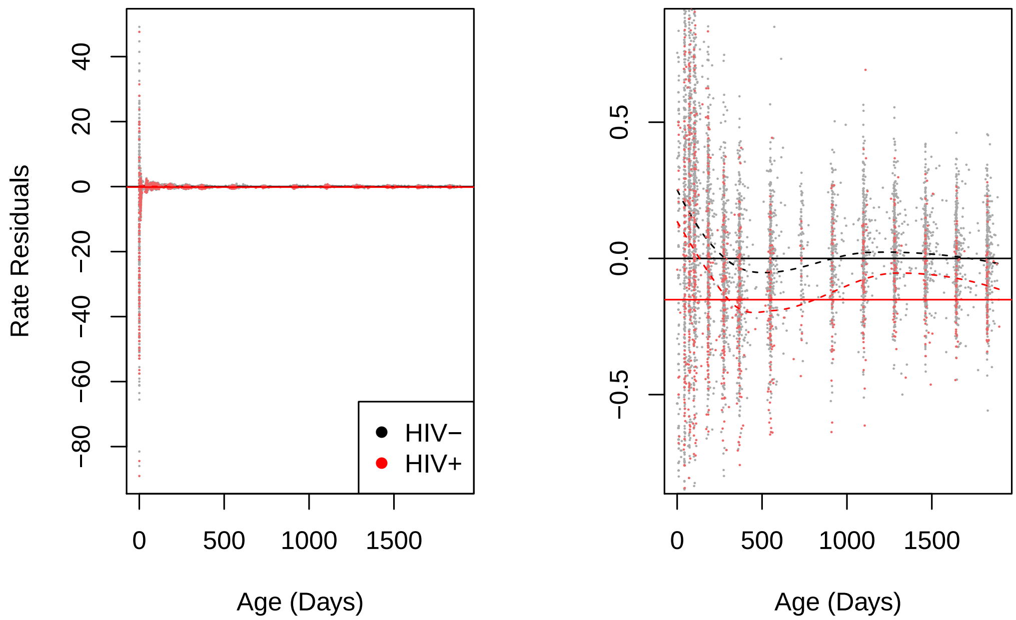



Figure 2.3: Rate residuals for the HIV infection status rate parameter associated with changes in weight from the LRR model presented in 2.2. Observations are colored based on HIV infection status. Solid horizontal lines are drawn at zero and at the rate effect for HIV status (-0.15). Lowess smooth lines were drawn for both HIV positive and negative subjects with points about the origin down-weighted. Two versions of the plot are displayed: one that contains the full range of rate residuals and one where the residual range was restricted to focus on residual values near zero.

each of the three grouping variables. Each test resulted in a p-value less than 0.001. The strong evidence of a deviation from the null is possibly a result of the large sample size coupled with modest departures and therefore may not be clinically relevant.

In response to Figure 2.3 which suggests additional time structure that has not been accounted for, a second LRR model was run that included an additional knot in the natural cubic spline at 1100 days. Results for the single rate three knot LRR model are presented in Table 2.3. The modified time trend altered main effect estimates for Sex and HIV infection status, though only mildly. The primary difference in the single rate three knot model

Table 2.3: Results from two LRR models for weight among infants exposed to HIV infection. The single rate model provides estimates for main effects and rate effects for sex, treatment, and HIV infection status. The split rate model included interactions in the rate effects between the three covariates and an indicator for observations measured after 500 days from birth. The rate effects for the split rate model can be interpreted as the difference in the rate of change associated with a given covariate prior to 500 days. The rate interaction effects estimate the change in the rate effect after 500 days. Both models specified a reference time structure as a natural cubic spline basis with knots at 150, 500, and 1100 days.

| | Single Rate Model | | Split Rate Model | |
|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI |
| Main Effects | | | | |
| Intercept | 3.26 | (3.17, 3.34) | 3.29 | (3.21, 3.38) |
| Sex (Male) | 0.25 | (0.16, 0.34) | 0.18 | (0.09, 0.28) |
| Treatment (Nev) | -0.12 | (-0.21, -0.03) | -0.11 | (-0.21, -0.02) |
| HIV Status | 0.05 | (-0.06, 0.17) | -0.02 | (-0.14, 0.10) |
| Time Trend | | | | |
| Basis 1 | 5.80 | (5.65, 5.96) | 5.71 | (5.54, 5.87) |
| Basis 2 | 9.14 | (8.92, 9.36) | 9.16 | (8.94, 9.39) |
| Basis 3 | 16.85 | (16.45, 17.24) | 16.79 | (16.38, 17.19) |
| Basis 4 | 10.51 | (10.26, 10.76) | 10.58 | (10.32, 10.84) |
| Rate Effects | | | | |
| Sex (Male) | 0.02 | (-0.01, 0.04) | 0.05 | (0.02, 0.08) |
| Treatment (Nev) | -0.01 | (-0.03, 0.02) | -0.01 | (-0.04, 0.02) |
| HIV Status | -0.15 | (-0.19, -0.12) | -0.12 | (-0.16, -0.08) |
| Rate Interaction | | | | |
| Sex (Male) | – | – | -0.06 | (-0.08, -0.03) |
| Treatment (Nev) | – | – | 0.01 | (-0.02, 0.03) |
| HIV Status | – | – | -0.08 | (-0.12, -0.04) |

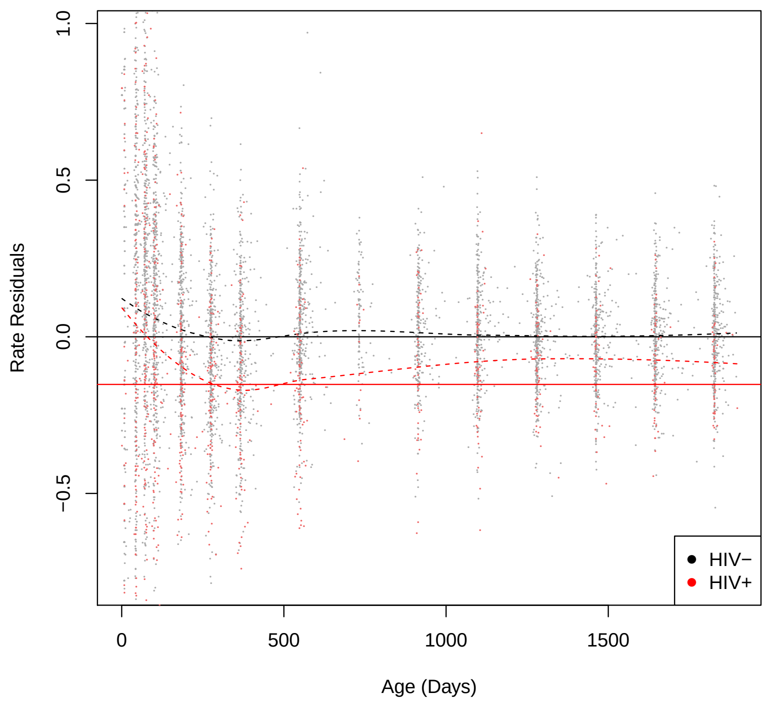



Figure 2.4: Plot of rate residuals for the effect of HIV infection status on the rate of change for weight based on the single rate LRR model presented in Table 2.3. Observations have been colored based on HIV infection status. Solid horizontal lines were drawn at 0 and -0.15. The lowess smooth curve provided for each HIV status group down-weighted observations near the baseline.

compared to the two knot model is the estimate of the intercept. The rate effects for the three knot model were nearly identical to the two knot model with only a small loss in precision for the rate effect associated with HIV infection status. The fit of the single rate three knot model was evaluated using diagnostic plots of the rate residuals for the three covariates. The rate residual plot for the rate effect for HIV infection status is depicted in Figure 2.4. The figure illustrates how the distinct time trend in the residuals for the reference group previously found in the two knot model has been captured by the three knot model. The trend in the residuals for the reference group coincides well with the zero line. The lowess smooth curve for the HIV positive infants shows a clear violation of the PR assumption since the line tapers to zero toward the end of the observed time period. Therefore, the rate of change for weight does not seem to be proportional across groups defined by HIV infection status.

We consider here an adjusted PR assumption by including select interactions with time in order to illustrate the LRR model under more relaxed assumptions. Specifically, we allow differing estimates for the rate parameters over selected time intervals. We allow the difference in the rate across groups to change before and after 500 days by including an interaction between each grouping variable and an indicator for whether an observation was observed after 500 days. Results from a LRR model with a split rate estimate before and after 500 days are presented in Table 2.3. We again use a natural cubic spline function to model time with knots at 150, 500, and 1100 days. For the split rate model, the rate effect estimates can be interpreted as the estimated difference in the rate of change associated with a given grouping variable in the first 500 days from birth. The rate interaction can be interpreted as the change in the rate effect for observations after day 500 compared to observations before day 500.

The results from the split rate model indicate a larger initial rate effect across sex groups in the first 500 days compared to the effect in the single rate model (see Table 2.3). The rate effect for sex is negated by the rate interaction after 500 days. The estimated difference in the rate of change based on treatment was small both before and after 500 days. Infants that were identified as HIV positive still had a large decrease in the rate of change before 500 days and an even larger deficit after 500 days. The alteration to the rate structure did have some impact on main effect estimates as well, particularly for treatment and HIV status. For the purposes of making inference on the rate of change, impact of rate assumptions on main effect estimates will typically not be of concern. However, some applications may wish to make simultaneous inference on rate level and mean level differences in which case the rate structure and also the reference time structure should be carefully consider for their impact on model estimates.

### *2.4.2 Study of Animal Nutrition*

In the chick weight study, we examined differences in rates of change of weight across the four dietary groups. Due to individual fluctuations in growth, the trend in weight was extremely heteroskedastic over time. Thus, weight was modeled on the logarithmic scale to reduce the non-constant variance which also accentuated the non-linear aspects of the outcome over time. The trend across time for this example was modeled using a natural cubic spline with two interior knots on days 7 and 14. Indicator variables were used for each dietary group with the fourth diet group used as the reference group. The indicator variables were included as main effects to allow for differences in log weight at birth and as covariates to allow differences in rates of change in log weight. To adjust for correlation within chicks, random effects were included for the intercept and the rate of change in line with the likelihood model specified in Equation (2.4).

Estimates for differences in rates of change between the four diet groups are provided in Table 2.4. Note that in this model, there is little mean differences between the four groups based on the main effect estimates and that the primary difference between the groups is in the rate of change over time. The LRR model showed a significant difference in the rate of change of log weight between the reference group and diet group 1 with chicks on diet 1 experiencing a 22% reduction (95% CI = (-37%, -7%)) in the rate of change in log weight in comparison to chicks on the reference diet. Diets 2 and 3 showed minimal evidence for a difference in rates of change compared to the reference diet with 7% reduction (95% CI = (-27%, 12%)) in the rate for diet 2 and a 9% increase (95% CI = (-12%, 30%) in the rate for diet 3. To illustrate the estimated trend in time between these four groups, the regression line fit by the rate regression line for untransformed weight is depicted in Figure 2.5 with lines connecting observed values for each chick in a given diet group.

To evaluate the goodness-of-fit of the LRR model, we produced residual plots of the data for the model. Figure 2.6 shows the residual plots of the data across time within each of the four diet groups. The plots suggest a possible issue with the PR assumption of the model. This is particularly the case for the reference diet which has a distinct non-linear trend in the residuals with minimal variation. Thus, the PR assumption may not be appropriate for this

Table 2.4: Model results comparing rates of change for log weight among chicks on different protein diets. Weight was regressed on the four dietary groups in two LRR model. The single rate model estimated main effects and rate effects for diet using the standard LRR model. The split rate model included interactions in the rate effects between the diet and an indicator for observations measured after day 11. The rate effects for the split rate model can be interpreted as the difference in the rate of change due to diet prior to day 11. The rate interaction effects estimate the change in the rate effect after day 11. Both models accounted for time using a natural cubic spline basis with knots at days 7 and 14.

| | Single Rate Model | | Split Rate Model | |
|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI |
| Main Effects | | | | |
| Intercept | 3.76 | (3.71, 3.81) | 3.74 | (3.69, 3.80) |
| Diet 1 | -0.06 | (-0.12, 0.01) | -0.04 | (-0.11, 0.03) |
| Diet 2 | -0.05 | (-0.12, 0.03) | -0.05 | (-0.12, 0.03) |
| Diet 3 | -0.08 | (-0.16, -0.01) | -0.04 | (-0.12, 0.04) |
| Time Trend | | | | |
| Basis 1 | 1.18 | (1.01, 1.35) | 1.19 | (1.02, 1.36) |
| Basis 2 | 2.24 | (1.93, 2.56) | 2.28 | (1.96, 2.61) |
| Basis 3 | 1.29 | (1.11, 1.47) | 1.27 | (1.10, 1.45) |
| Rate Effects | | | | |
| Diet 1 | -0.22 | (-0.37, -0.07) | -0.24 | (-0.39, -0.09) |
| Diet 2 | -0.07 | (-0.27, 0.12) | -0.08 | (-0.27, 0.12) |
| Diet 3 | 0.09 | (-0.12, 0.30) | 0.00 | (-0.21, 0.20) |
| Rate Interaction | | | | |
| Diet 1 | – | – | 0.06 | (-0.04, 0.16) |
| Diet 2 | – | – | 0.01 | (-0.12, 0.13) |
| Diet 3 | – | – | 0.23 | (0.09, 0.37) |

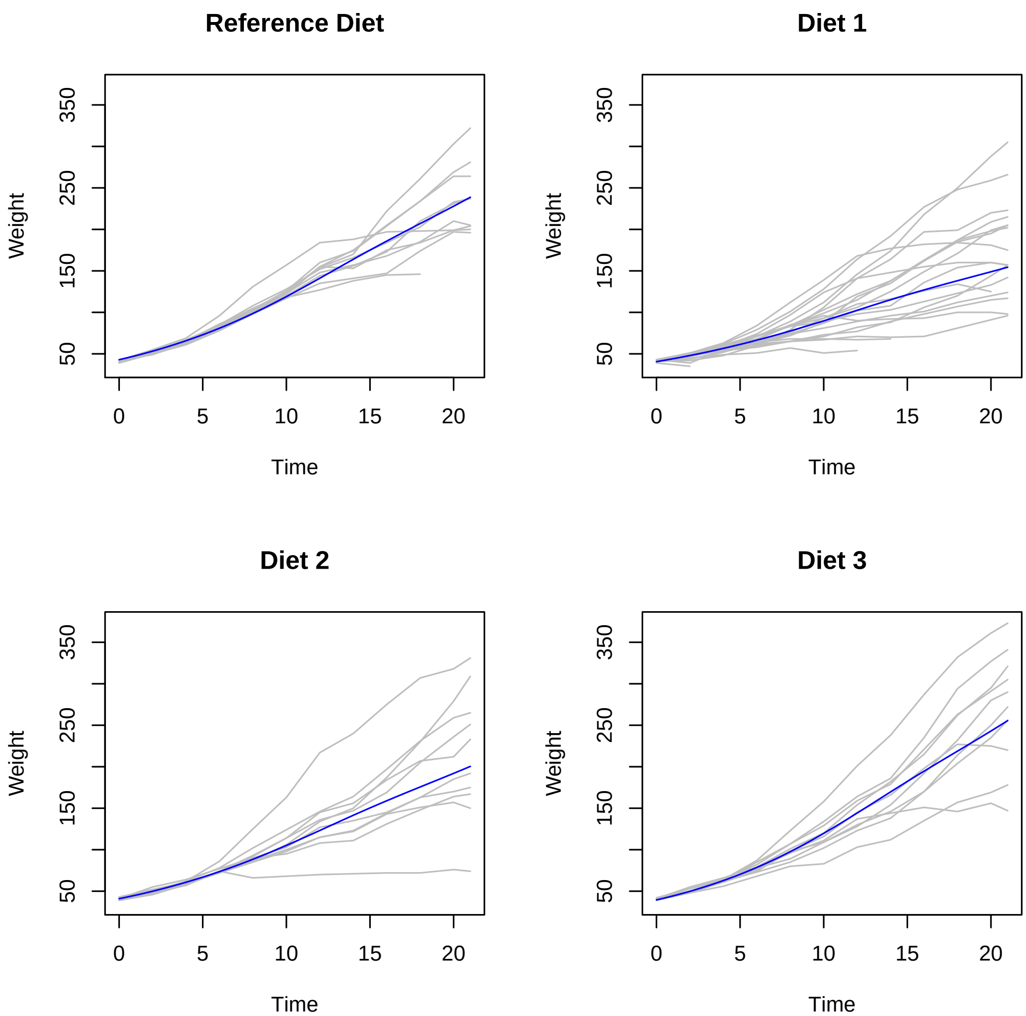



Figure 2.5: Spaghetti plots of chick weight over time separated by diet groups. The blue line depicts the fitted line for each diet group from the single rate LRR model for log weight presented in Table 2.4.

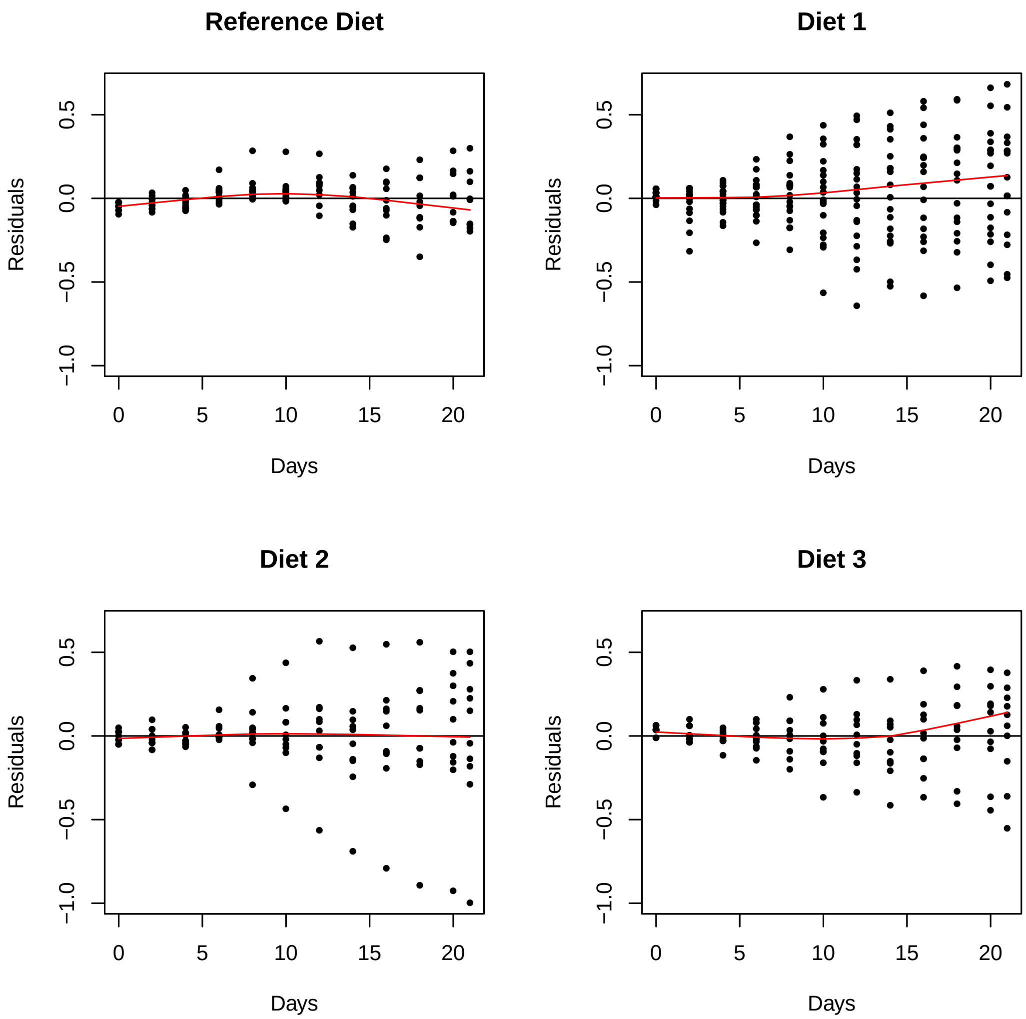



Figure 2.6: Residual plots for the single rate LRR model for weight among chicks split by diet groups. A red line was drawn for the lowess smooth curve for the residual points.

dataset. Violation of the PR assumption is supported by the score test for a linear change in the difference in the rate across groups. The score statistic had a value of 15.50 which when compared to a chi-squared distribution with three degrees of freedom corresponded to a p-value of 0.0014.

In reaction to evidence against the PR assumption, we relax the PR assumption by allowing a variable rate difference over time. We constructed a new LRR model that included an interaction in the rate effects between the dietary indicator variables and an indicator for observations measured after day 11. The indicator at day 11 was selected since it represented the approximate extrema of the u-shaped trend in the residuals depicted in Figure 2.6. The rate coefficient estimated for each diet group can be interpreted as the difference in the rate of change compared to the reference diet group in the first 11 days of the study. The coefficient for the interaction term for each diet can be interpreted as the difference in the difference in rates after day 11 relative to before day 11. Estimates for the split rate model have been provided in Table 2.4. Comparing these results to the single rate model, it can be seen that the main effect differences across groups and the reference time trend are roughly the same in the two models. Also, the difference in the rate of change for diet groups 1 and 2 relative to the reference group is roughly the same before and after day 11. However, the rate estimates for diet group 3 are very different in the split rate model with no estimated difference in the rate of change for diet group 3 compared to the reference diet prior to day 11, and after day 11, the difference in the rate between these two groups was estimated to increase by 22%. Thus, there is little difference in the rate of change between the reference diet and diet 3 early, but a large difference in the rate of change later with those on diet 3 growing faster.

## *2.5 Discussion*

Modeling longitudinal rates of change is important in many biomedical settings and, in particular, for pediatric applications where growth is characterized. To our knowledge, the proposed LRR method is the first attempt to focus regression directly on estimating differences in rates of change relative to a generally specified time trend. The defining characteristic of longitudinal data is that outcomes are measured over time making the study of change a natural use for such data.

We have discussed existing methods that can be used for estimating differences in rates. Linear models [Laird and Ware, 1982, Liang and Zeger, 1986] suffer from complex interpretation and potential losses in power when an outcome's trend over time is non-linear. Methods developed for HIV viral load counts [Wu et al., 1998], mechanistic growth models [Huang, 2011], and pharmacokinetics [Lindsey et al., 2000] are tailored to settings where rates are linked to underlying states. However, non-linear methods [Davidian and Giltinan, 1995] is a general framework that may be useful for modeling rates, and our method can be considered as a focused special case. Accelerated time models [Brumback and Lindstrom, 2004, Gray and Brookmeyer, 1998, 2000] are another novel non-linear approach to modeling rates of change through transformations of the time scale. Applications where the magnitude of an outcome is constant but progression of the outcome can be considered to accelerate and decelerate in relation to covariate are suitable for an AT approach. In contrast, our LRR model directly links covariates to the magnitude of an outcomes progression.

The LRR method allows for a general specification of a reference trend over time. Multiple covariates can be examined in a single comprehensive model with corresponding rate parameters that are simple to interpret. On the other hand, the standard linear mixed model approach will be suitable when comparing rates using linear trends in time. The mean structures for the LME and LRR models are equivalent in this setting. One limitation of the LRR method under a non-linear trend in time is the amount of data needed for estimation. Estimating differences in rates for outcomes with non-linear trends over time requires enough time points and density of data for the non-linear time trend to be adequately estimated. Thus, in settings where data is sparse or the non-linear trend is approximately linear, an appropriate linear model approach will likely perform better than the LRR method. However, the dependence on dense data exists primarily for the estimation of the reference time function since our proposed estimating procedures estimate this function jointly with the differences in the rate of change. Since the primary focus of the LRR model is on comparisons of rates of change, other estimation approaches could be considered that are less reliant on dense data across all groups. In some applications, information exterior to the group data may exist that can inform the estimation of the underlying time trend. For example, in a treatment trial, previously collected longitudinal data may exist on untreated subjects that could be used to estimate a reference time function. The estimated time trend could then be used for comparisons of rates of change between treatment and control groups in which case the rate comparison will be less dependent on having dense data for the two groups.

We applied the LRR method to examples of growth research in human subjects as studies of adolescent and juvenile development seemed to be a particularly natural use for the model. However, there are numerous other areas for which this method may be useful including treatment trials where LRR could be utilized to examine any outcome whose rate is impacted by treatment. Modeling environmental risk factors could also consider utilizing the LRR method for exposures that have an acute effect on outcomes.

Longitudinal rate regression may also be useful for applications with multivariate longitudinal outcomes. One goal with multivariate longitudinal data is to borrow information across related outcomes and across time in order to gain power to detect group differences. Structuring both the mean and the variance structure present interesting challenges in this area of methods research. One approach for specifying the mean structure is to construct a global test for group differences for a multivariate outcome [Gray and Brookmeyer, 1998, 2000, Travison and Brookmeyer, 2007]. The accelerated time model proposed by Gray and Brookmeyer [1998, 2000] was suggested under this premise where the acceleration parameter was estimated uniformly across correlated outcomes. The LRR method can be extended similarly to estimate a single rate parameter for a multivariate outcome measured over time. Estimating a common difference in the rate of change for multivariate outcomes is a convenient way to link outcomes rate level differences are scale free with regard to each outcome. We explore multivariate longitudinal applications of the LRR model in greater detail in Chapter 3.

The estimation of the trend over time will typically be viewed as secondary to the estimation of differences in rates in the LRR model. Therefore, it will be advantageous to pursue nonparametric methods for estimating a trend over time so as to make the model more robust to misspecification of the time trend when reliable data is available. A parametric estimation of rate coefficients combined with the nonparametric time trend estimation is a concept similar to the semiparametric Cox proportional hazard model [Anderson and Gill, 1982]. Several approaches to nonparametric approaches exist for adapting to the LRR method. One such approach is using penalized splines [Eilers and Marx, 1996] which has been previously adopted for semi-parametric methods [Zhang, Lin, and Sowers, 2000]. Methods for incorporating penalized splines into estimation of the LRR model are described in Chapter 4.

# Chapter 3

# GLOBALLY SHARED PARAMETERS FOR DIFFERENCES IN RATES OF CHANGE FOR MULTIVARIATE LONGITUDINAL DATA

We expand the topic of discussion to consider methods for multivariate longitudinal data. The exploration of methods for modeling rates of change leads naturally into this area. Since the rate of change as it has been defined in this dissertation is a characteristic that is scale free, linking group differences in the rate of change across related outcomes, that potentially differ in mean level scales, is straightforward. Therefore, we extend the methods of the previous chapter as a basis for discussion of modeling rates of change for multivariate longitudinal data.

## *3.1 Introduction*

One goal of examining group differences across multivariate longitudinal outcomes is to borrow information across time and across outcomes. Specifically, when multivariate longitudinal methods are used appropriately, power can be gained to detect differences between groups at the mean level or in the rate of change for related measures.

Two common approaches for modeling multivariate longitudinal data exist: linking outcomes through a longitudinal, latent variable; and estimating a common group parameter across separately specified outcome models. Methods that use latent variables to model multivariate longitudinal data have been proposed by Roy and Lin [2000] and Proust-Lima, Letenneur, and Jacqmin-Gadda [2007]. Under the latent variable approach, each outcome characterizes the behavior of an unmeasured latent variable. For example, in autism research, several behavioral outcome measures exist that characterize the severity of repetitive behavior in children with autism. Roy and Lin [2000] propose use of a hierarchical model for making inference on the latent variable that is assumed to drive the measured outcomes. A two part model is constructed where a longitudinal model structure is assumed for the latent variable representing the underlying “state”, and a second, measurement model structure is assumed for each outcome linked to the latent variable. Roy and Lin [2000] outline the use of mixed effect models for both the longitudinal, latent variable structure and the multivariate outcome measurement structure when all outcomes are continuous. Let $Y_{ijk}$ denote the observed measurement of the $k$th outcome for individual $i$ at time $t_j$, and let $U_{ij}$ denote the latent variable for the related outcomes. Given a fixed effect design matrix $\mathbf{X}_{ij}$ and random effect design matrix $\mathbf{Z}_{ij}$, Roy and Lin [2000] express the latent variable model and the measured outcomes model as follows:

$$\begin{aligned} U_{ij} &= \mathbf{X}_{ij}^T\boldsymbol{\alpha} + \mathbf{Z}_{ij}^T\mathbf{a}_i + \epsilon_{ij} \\ Y_{ijk} &= \beta_{0k} + U_{ij}\beta_{1k} + b_{ik} + e_{ijk}. \end{aligned}$$

By combining the two mean structures together, the association between the covariates of interest and the latent variable can be estimated. Proust-Lima et al. [2007] also propose methodology for a longitudinal latent variable model where the multivariate outcome consists of a mixture of continuous and categorical outcomes. When direct inference on the underlying latent variable is of interest, the latent variable approach will characterize each covariates effect. However, in many applications with multivariate longitudinal data, interest is in inference on the measured outcome even when an underlying latent variable may describe longitudinal and multivariate dependence. For example, clinical studies of patients recovering from back pain are focused on direct improvements in pain and function although both outcomes are dimensions of a patient's quality of life. When inference on each outcome is of primary interest, it is often beneficial to allow for separate structures for each outcome since a uniform longitudinal and multivariate structure may be difficult to justify.

A second approach to analysis of multivariate longitudinal data is to allow each outcome to be modeled separately, but to link each outcome model by a common parameter that describes shared differences between groups defined by a covariate of interest [Gray and Brookmeyer, 1998, 2000, Jia and Weiss, 2009, Travison and Brookmeyer, 2007]. By generating a common, or global, estimate for this parameter across each outcome, a global test for group differences across the multivariate outcome can be constructed. We refer to such models as global shared parameter models which can generally be expressed as

$$E(Y_{ijk} \mid X_i = x, t_{ij} = t) = f_k[\lambda(x), \boldsymbol{\beta}_k(t)] \tag{3.1}$$

where the global parameter is denoted by $\lambda$ for a grouping variable $X_i$ and is fixed across all outcomes, $k = 1, \cdots, K$. A global test for multivariate longitudinal data based on a global shared parameter model was proposed by Gray and Brookmeyer [1998, 2000] using accelerated time (AT) models. In AT models, the time scale of a longitudinal outcome is assumed to differ across groups. Coinciding with the notation in Equation (3.1), the mean structure for the AT model under a third degree polynomial time structure is given by the equation

$$E(Y_{ijk} \mid X_i = x, t_{ij} = t) = \beta_{0k} + \beta_{1k}(\lambda^x t) + \beta_{2k}(\lambda^x t)^2 + \beta_{3k}(\lambda^x t)^3.$$

Thus, longitudinal progression of an outcome is accelerated or decelerated across groups of interest. The multivariate AT method introduced by Gray and Brookmeyer [1998, 2000] allows separate specification for the time structure of each outcome, but estimates a global group parameter that alters the time scale uniformly across outcomes. Travison and Brookmeyer [2007] proposed a global test for a treatment across multivariate outcomes by linking the marginal distributions of treatment groups by a shared treatment parameter across outcomes. The treatment parameter is used to relate the marginal distribution for observations on treatment to the marginal distribution of control observations in a manner that allows the use of estimation from survival analysis methodology. Jia and Weiss [2009] constructed a multivariate longitudinal model with common additive effects for covariates of interest. The linear combination of covariates for the additive model is multiplied by a unique parameter for each outcome that accounts for the scale of the outcome.

Existing global shared parameter approaches focus on different types of association between a multivariate outcome and covariate-defined groups: group differences in the time scale [Gray and Brookmeyer, 1998, 2000], group differences in the quantiles of the marginal distribution [Travison and Brookmeyer, 2007], and scaled mean level group differences [Jia and Weiss, 2009]. In this chapter, we propose a new global shared parameter model that estimates group differences in the rate of change of a multivariate longitudinal outcome. The proposed methodology is an extension of the longitudinal rate regression (LRR) model developed in Chapter 2. The LRR method characterizes longitudinal change across groups through direct structuring of longitudinal rates of change. The LRR method links rates of change across groups by assuming a Proportional Rate (PR) assumption where rates differ across groups by a constant proportion over time. A mean model is then induced based on the rate assumptions coupled with a flexible specification for a reference time trend. In Chapter 2, we demonstrated potential advantages of a direct approach to modeling rates of change as compared to linear mixed effects model approach when the underlying time trend is non-linear. The LRR method is similar to the AT model developed by Gray and Brookmeyer [1998, 2000] in that each approach estimate difference in rates of change across covariate-defined groups. However, the AT model assumes a fixed range or magnitude for the outcome, and focuses on potential accelerations and decelerations in the progression of an outcome along a common trajectory. In contrast, the LRR model directly links covariates to the magnitude of longitudinal change and does not assume a common, bounded trajectory.

Our proposed multivariate longitudinal data extension to the LRR model permits estimation of a global shared parameter representing the difference in the rate of change associated with covariate-defined groups. We maintain a separate specification for the reference time trend and adjustment for baseline covariates for each outcome. In Section 3.2, we detail methodology for the Multivariate Longitudinal Rate Regression (MLRR) model and discuss options for specifying a multivariate longitudinal covariance structure. In Section 3.3, the asymptotic power of the MLRR model using a global shared parameter is compared to the power for testing for differences in the rate of change for each outcome separately. We illustrate the multivariate rate regression method in Section 3.4 on a study of growth among infants exposed to HIV. Finally, in Section 3.5, we offer discussion and concluding remarks for studying differences in rates of change for multivariate longitudinal data.

## 3.2 Methods

We use the notation $Y_{ijk}$ to denote the $k$th outcome of individual $i$ observed at time $t_{ijk}$ for $i = 1, \dots, N$, $j = 1, \dots, n_{ik}$, and $k = 1, \dots, K$. We use $\mathbf{X}_i = (X_{i1}, \dots, X_{iQ})$ to denote a vector of covariates. We are interested in detecting differences in the rate of change of $Y_{ijk}$

across groups defined by $\mathbf{X}_i$. Let

$$E[Y_{ijk} \mid \mathbf{X}_i = \mathbf{x}, t_{ij} = t] = g_k(\mathbf{x}) + \mu_{\mathbf{x}k}(t)$$

where $g_k(\cdot)$ is a function dependent only on $\mathbf{X}$ and $\mu_{\mathbf{x}k}(\cdot)$ denotes some function of $t$ for a given value of $\mathbf{X}$ with the constraint that $\mu_x(0) = 0$ for all values of $\mathbf{X}$. The function $g_k(\cdot)$ describes all mean level differences in $Y_{ijk}$ at a time origin, or time zero. The function $\mu_{\mathbf{x}k}(t)$ specifies the change in the expected value of the $k$th outcome over time from baseline for a given $\mathbf{X}_i = \mathbf{x}$. We generalize the PR assumption from Chapter 2 to a global multivariate proportional rate (GMPR) assumption which can be expressed as

$$\frac{\partial \mu_{\mathbf{x}k}(t)}{\partial t} = (1 + \boldsymbol{\theta}^T \mathbf{x}) \frac{\partial \mu_{0k}(t)}{\partial t}$$

where $\mu_{0k}(\cdot)$ is the time function for the $k$th outcome for some preselected reference group, defined by $\mathbf{X}_i = \mathbf{0}$. Thus, the MLRR model assumes that the rate of change in the expected value of the multivariate outcome, $\mathbf{Y}_{ij}$, for given values of $\mathbf{X}_i = \mathbf{x}$, relative to the reference group ($\mathbf{X}_i = \mathbf{0}$), is given by $(1 + \boldsymbol{\theta}^T \mathbf{x})$ for all time $t_{ij}$. Note that $\boldsymbol{\theta}$, the vector of rate parameters for $\mathbf{X}_i$, does not depend on the specific outcome since we assume a global rate effect across all outcomes $k = 1, \ldots, K$. We can interpret the rate parameters, $\boldsymbol{\theta}$, as the global difference in the rate of change in the mean of $\mathbf{Y}_{ij}$ associated with a unit difference in the covariates, $\mathbf{X}_i$. That is, the parameter $\theta_q$ is interpreted as the percent increase (decrease) in the rate of change in the mean of $\mathbf{Y}_{ij}$ when $X_{iq} = (x + 1)$ relative to when $X_{iq} = x$. The MLRR model may also be constructed using a non-proportional rate assumption by including time varying covariates. Chapter 2 discusses methods for non-proportional rate models for the univariate situation, and extending these methods to multivariate outcomes is relatively straightforward.

A full mean structure is induced by the GMPR assumption through integrating over a time given interval, $[0, t]$. For outcome $Y_{ijk}$, the full mean structure is specified as

$$E(Y_{ijk} \mid \mathbf{X}_i = \mathbf{x}, t_{ij} = t) = g_k(\mathbf{x}) + (1 + \boldsymbol{\theta}^T \mathbf{x}) \mu_{0k}(t). \tag{3.2}$$

For the model defined by Equation (3.2), we refer to the functions $g_k(\cdot)$ and $\mu_{0k}(\cdot)$ respectively as the baseline function and the reference time function for the $k$th outcome. We emphasize that both the baseline function $g_k(\cdot)$ and the reference time function $\mu_{0k}(\cdot)$ are allowed to differ across outcomes; only the rate parameters are assumed to be common across outcomes. Similarly to the Chapter 2, the reference time structure for each outcome will be modeled as a linear combination of a flexible parametric basis for time: $\mu_{0k}(t_{ij}) = \boldsymbol{\beta}_k^T \mathbf{T}_{ijk}$ where $\mathbf{T}_{ijk}$ is a vector of functions evaluated at $t_{ij}$. The baseline function for each outcome will be specified as a linear combination of the covariates: $g_k(\mathbf{X}_i) = \boldsymbol{\alpha}_k^T \mathbf{X}_i$. Model estimation in this chapter is done based on a joint estimation of all mean parameters. As discussed in Chapter 2, alternative estimation approaches may be more beneficial in some scenarios.

### *3.2.1 Estimation*

A necessary secondary aspect to multivariate longitudinal modeling is the structuring of the covariance to account for correlation across time and across outcomes. Here, we outline select alternatives for specifying the covariance structure for multivariate longitudinal data. When using a latent variable modeling structure, Roy and Lin [2000] and Proust-Lima et al. [2007] discuss an approach to modeling the covariance structure using hierarchical linear mixed effects. Part of the hierarchical covariance structure of the latent variable models includes random effects specified at the latent variable level. Since we focus on global shared parameter methods which do not incorporate latent variables, a hierarchical linear mixed effects model will not be directly applicable. Previous work [Dubin and Muller, 2005] has structured a dynamic correlation for functional longitudinal data with multivariate outcomes. We focus our discussion on approaches to covariance specification when the outcomes are measured discretely in time as is often done in designed clinical studies.

Galecki [1994] proposed a multivariate longitudinal covariance structure that separately specifies the correlation due to time and correlation due to outcome, and combines them to form a covariance matrix. Two matrices are combined using a Kronecker product where one matrix captures correlation across time and the other captures correlation across outcome [Galecki, 1994, Genton, 2007]. Let $\text{Cov}(\mathbf{Y}_i) = \Sigma \otimes \Phi$ where $\Sigma$ specifies the correlation across time and $\Phi$ specifies the correlation across outcome in that $\text{Cov}(\mathbf{Y}_{i \cdot k}) = \phi_{kk}\Sigma$ and

$\text{Cov}(\mathbf{Y}_{ij\cdot}) = \sigma_{jj}\Phi$. The Kronecker product structure proposed by Galecki [1994] is equivalent to a conditional independence correlation structure where we assume that, for $j \neq j'$ and $k \neq k'$, $(Y_{ijk} \perp\!\!\!\perp Y_{ij'k'}|Y_{ij'k})$ and $(Y_{ijk} \perp\!\!\!\perp Y_{ij'k'}|Y_{ijk'})$. The implications of structuring the covariance in this way is that the correlation over time is the same across outcomes and that correlation across outcomes is the same over time. Only the total variation is allowed to vary across outcome and across time in the Kronecker product structure. In practice, these assumptions may be difficult to justify, in particularly, assuming that the correlation over time is the same for each outcome is restrictive.

Another approach to covariance modeling is to incorporate longitudinal random effects that are correlated across outcomes. The use of a mixed effects structure to specify a covariance structure is a common approach for modeling multivariate longitudinal data [Fieuws, Verbeke, and Molenbergh, 2007]. For a mixed effects structure, we specify the longitudinal structure of outcome $k$ with individual intercepts and rates of change as

$$Y_{ijk} = [g_k(\mathbf{X}_i) + b_{0ik}] + (1 + \boldsymbol{\theta}\mathbf{X}_i + b_{1ik})\mu_{0k}(t_{ij}) + \epsilon_{ijk} \tag{3.3}$$

where, for outcome $k$, the vector of random effects for individual $i$, $\mathbf{b}_{ik} = (b_{0ik}, b_{1ik})^T$, is normally distributed with mean $\mathbf{0}$ and covariance $\mathbf{D}_{kk}$ and the vector of random errors for individual $i$, $\boldsymbol{\epsilon}_{ik} = (\epsilon_{i1k}, \cdots, \epsilon_{iTk})$, is normal distributed with mean $\mathbf{0}$ and covariance $\sigma_k^2 I_T$. The covariance is then connected across outcomes by assuming that $Cov(\mathbf{b}_{ik}, \mathbf{b}_{ik'}) = D_{kk'}$. Therefore, the mixed effects model for the MLRR method characterizes individual variation both in the level at baseline and in the rate of change, and this variation is correlated across outcomes. The multivariate longitudinal mixed effects modeling approach is analogous to the LRR mixed effects model proposed in Subsection 2.2.1.

Using random effects to structure the multivariate longitudinal covariance structure provides an intuitive interpretation of the model at the individual level with each subject having their own baseline value and rate of change. The disadvantage of a mixed effects approach is that four additional variance parameters are introduced by the across outcome covariance matrix for each pair of outcomes modeled. More explicitly, given $K$ outcomes, the mixed effects multivariate structure requires the estimation of $3K + 4\frac{K(K-1)}{2}$ variance parameters for the random effects. If we instead model the outcomes separately in independent models using random effects, we would only estimate $3K$ parameters for the random effects. If only one global rate parameter is being estimated by the multivariate model, then the number of rate parameters being estimated is reduced by $K-1$. However, in total, the multivariate mixed effects model approach would estimate $2K^2 - K + 1$ more parameters than when using separate models for each outcome. In addition, the difference between joint and separate models in the number of parameters estimated is reduced when multiple covariates are included with a global rate parameter. The increase in model parameters is a common issue with modeling multivariate longitudinal outcomes, but can also be advantageous for maintaining flexibility for modeling the covariance structure. By specifying an unstructured cross outcome random effect covariance matrix, the mixed effects model can robustly model a correlation structure that differs across outcome.

A third approach is to construct a covariance matrix by considering a transition model concept where an observation at a given time point is assumed to be a linear combination of previous observations plus random error. First, for outcome $k$, let the observation for the first time point be given by $Y_{1k} = \beta_{0k}^{(1)} + \epsilon_{1k}$ where $\epsilon_{1k} \sim N(0, \tau_{1k}^2)$. Then, at the second time point, let $Y_{2k} = \beta_{0k}^{(2)} + \beta_{1k}^{(2)} Y_{1k} + \epsilon_{2k}$ where $\epsilon_{2k} \sim N(0, \tau_{2k}^2)$ and $\epsilon_{2k}$ is independent of $\epsilon_{1k}$. Generally, for the $j$th time point, we would specify $Y_{jk} = \beta_{0k}^{(j)} + \beta_{1k}^{(j)} Y_{1k} + \cdots + \beta_{(j-1)k}^{(j)} Y_{(j-1)k} + \epsilon_{jk}$ where $\epsilon_{jk} \sim N(0, \tau_{jk}^2)$ and is independent of the previous error terms for the $k$th outcome. The recursive procedure induces a covariance structure across time. To incorporate correlation across outcomes, the error terms are assumed to be correlated across outcomes e.g. $\mathrm{Corr}(\epsilon_{jk}, \epsilon_{jk'}) = \rho_j$. The recursive covariance structure is illustrated in Figure 3.1 for two outcomes observed over three time points. Constructing a covariance structure based on the outlined telescoping procedure is convenient for providing an intuitive interpretation of the covariance structure as illustrated in Figure 3.1. The structure implies that variation in an outcome at a given time point can partially be explained by the history of the outcome at previous time points. An outcome is then assumed to be correlated with related outcomes only through the remaining measurement error once the longitudinal history has been taken into account. However, the recursive structure can be challenging from an implementation stand point since the covariance matrix will become increasingly complex with each additional observed time point. The covariance structure is also restrictive in

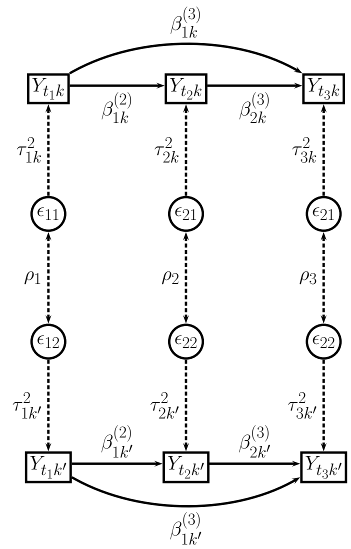



Figure 3.1: The recursive multivariate longitudinal covariance structure for two outcomes and three time points. Random variables are framed by boxes. Random errors are framed by circles. Solid arrows are used to denote fixed effects. Dashed arrows indicate sources of variation with two way arrows to indicate correlation.

that each outcome must be measured at the same time point in order for the cross-outcome correlation structure to be appropriate.

As is common in statistical methodology, selection of a covariance structure for a multivariate longitudinal model is an exercise in balance between restrictive modeling assumptions and flexible though parametrically exhaustive structure. The kronecker product covariance structure proposed by Galecki [1994] is an example of a model that relies heavily on assumptions about the multivariate longitudinal correlation structure. By assuming a constant correlation structure across outcomes and across time, the number of degrees of freedom used to estimate a covariance structure is relatively small though this assumption may be difficult to justify for many applications. In contrast, the proposed mixed effects modeling approach introduces a large number of parameters into the variance model. The modeling assumptions of mixed effects models are more commonly accepted in the scientific literature, and the approach offers a convenient interpretation at the individual level. Another flexible approach to modeling multivariate longitudinal data is through the use of estimating equations. Methodology for modeling multivariate longitudinal data using estimating equations was not provided here since methods for estimating equation methods presented in Subsection 2.2.2 can be extended in a straightforward manner. The recursive approach to multivariate longitudinal covariance structure is generated based on intuitive modeling assumptions and can flexibly characterize variation. In addition, estimation of the recursive covariance structure will often be complex and may necessitate additional modeling assumptions. For the remainder of this chapter, we use the proposed mixed effects model structure to evaluate and illustrate the MLRR model. The mixed effects was deemed most suitable since it is a common modeling approach and can be implemented in a straightforward manner. Using a mixed effects approach also simplifies comparisons with results presented in Chapter 2 where a univariate mixed effects approach was implemented.

### *3.3 Power*

A primary goal of the global shared parameter model is to potentially gain power to detect group differences by borrowing information across outcomes. For the MLRR model in particular, the global shared parameter of interest is the group difference in the rate of change. Specifically, we hope to gain power to detect group differences in the rate of change for a multivariate outcome compared to examining group differences for each outcome separately. To evaluate this potential gain in power, we compare the global shared parameter MLRR model to two alternatives: a univariate LRR model for one pre-selected outcome; and a MLRR model with separate rate parameters for each outcome. The univariate LRR model approach represents an investigator selecting a single outcome for testing e.g. the outcome of greatest scientific interest or the outcome most anticipated to show a difference across groups. Alternatively, an investigator may decide to test all outcomes, but not assume the difference in the rate of change is the same for each outcome. In this case, the MLRR model with separate rate parameters for each outcome would be appropriate although power to detect group differences may be sacrificed since each rate parameter test results in additional degrees of freedom. For the remainder of this section, we refer to the three competing models as the univariate model, the joint multivariate model, and the global multivariate model for simplicity.

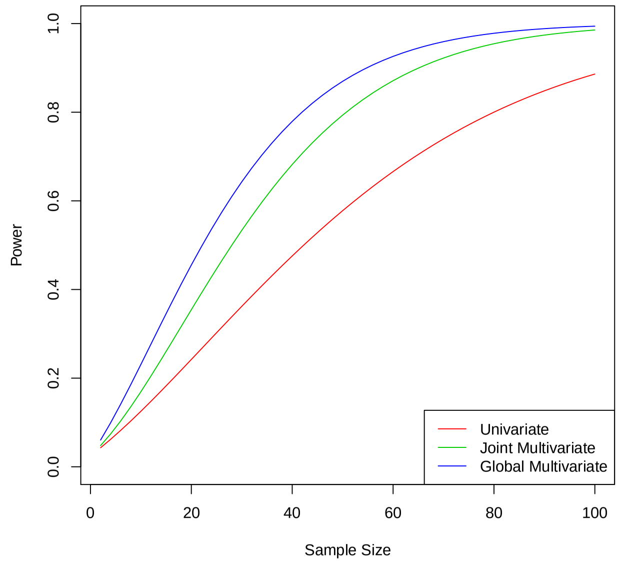
We compare the power of the three modeling approaches using a bivariate continuous outcome and a single binary covariate for comparing rates of change. The reference time trend is modeled using a cubic polynomial equation, and the covariance matrix was structured using the mixed effects model approach outlined in subsection 3.2.1. The three models for each comparison can generally be expressed as

| | |
|---|---|
| Univariate: | $E(Y_{ij1} \mid X_i = x, t_{ij} = t) = \alpha_{01} + \alpha_{11}x + (1 + \theta_1 x)(\beta_{11}t + \beta_{21}t^2 + \beta_{31}t^3)$ |
| Joint Multivariate: | $E(Y_{ijk} \mid X_i = x, t_{ij} = t) = \alpha_{0k} + \alpha_{1k}x + (1 + \theta_k x)(\beta_{1k}t + \beta_{2k}t^2 + \beta_{3k}t^3)$ |
| Global Multivariate: | $E(Y_{ijk} \mid X_i = x, t_{ij} = t) = \alpha_{0k} + \alpha_{1k}x + (1 + \theta x)(\beta_{1k}t + \beta_{2k}t^2 + \beta_{3k}t^3)$ |

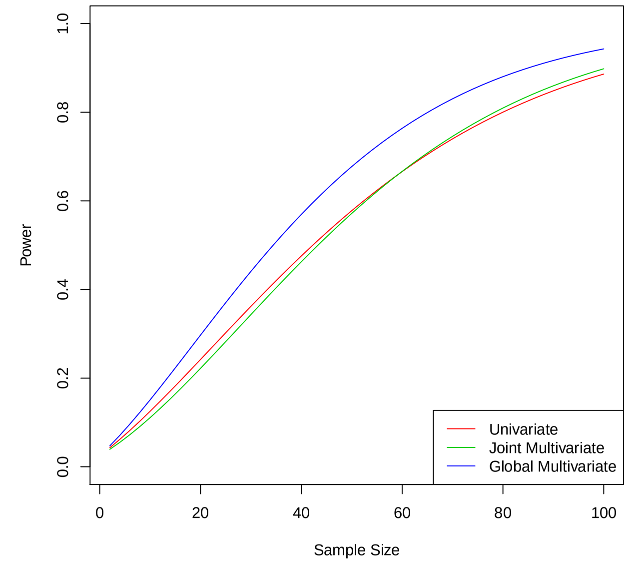
where $k = 1, 2$. We consider scenarios both where the GMPR assumption is correct and where the assumption is incorrect. When the GMPR assumption is correct, the rate parameters for each outcome is the same, $\theta_k = \theta$ for $k = 1, 2$, and all three models are correctly specified. Hence, asymptotic power for the three models is based directly on the hessian of the log-likelihood. When the GMPR assumption is incorrect, the rate parameters differ between the outcomes, $\theta_1 \neq \theta_2$, and the global multivariate model is misspecified. The univariate model and the joint multivariate model are still correctly specified since we still assume that the rate of change for each outcome is proportional across groups. Power for the univariate model and the joint multivariate model are again generated using the hessian. To calculate the power for the global multivariate model, results for the asymptotic behavior of a longitudinal estimate under model misspecification are used based on White [1982] and Heagerty and Kurland [2001]. Asymptotic estimates for the misspecified model are generated using Monte Carlo methods consisting of $10^4$ individuals measured 12 times. Based on a sample of 10 replicates, we estimate that the Monte Carlo procedure produced a power curve that has maximum point-wise standard deviation of approximate 3%.

Figure 3.2 displays three plots of power curves comparing the three modeling approaches. For all three plots, the value of the rate parameter for the outcome tested using the univariate model is the same, and greater than or equal to the rate effect for the outcome that is not tested by the univariate model. Therefore, the curve for the univariate model is identical in each plot. In Figure 3.2(a), the curves were calculated where the true rate parameter was the same for both outcomes, $\theta_1 = \theta_2 = \theta$. When the GMPR assumption is true, the two multivariate models show a clear advantage over the univariate model for detecting group difference. Estimating a global rate parameter is advantageous compared to estimating separate rate parameters which was expected since the global rate model is correct in this scenario. Figures 3.2(b) and 3.2(c) depict scenarios where the GMPR assumption is incorrect. The rate parameter for the second outcome was reduced by 50% relative to the first outcome (e.g. $\theta_2 = 0.5\theta_1$) for the power curves depicted in Figure 3.2(b). The multivariate models show a substantial reduction in power in Figure 3.2(b) relative to Figure 3.2(a). The joint multivariate model had roughly the same power to detect group differences as the univariate model. The global multivariate model still showed gains in power for detecting group differences relative to the competing models. In Figure 3.2(c), the rate parameter for the second outcome was zero, $\theta_2 = 0$, or equivalently, was reduced by 100% relative to the rate parameter for the first outcome, $\theta_1$. The power to detect group differences suffers for both multivariate models relative to the correctly chosen univariate model under this scenario. The global multivariate model performs the worst in this instance.

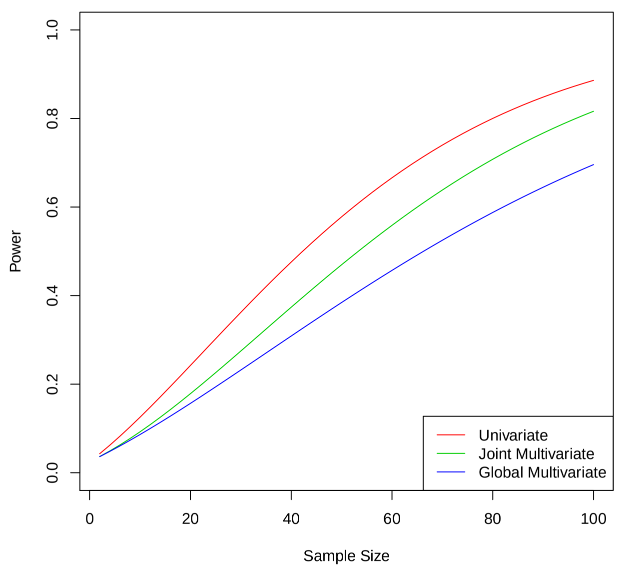
The power curves results illustrate the gains in power of the global shared parameter MLRR model when the rate parameter is the same across outcomes and how these gains are negated as the difference between the rate parameters for each outcome is increased. The results are encouraging for showing that gains in efficiency can be obtained even when the rate parameters are not the same, but similar. Defining what similar means for a given application will be dependent on rate parameter effect size and the values of the variance components although this characteristic is not illustrated here. The illustrated advantage of the global shared parameter MLRR model over the univariate LRR model is likely conservative since in each scenario the univariate analysis always used the outcome with the largest effect size. We could instead introduce a probability distribution on which outcome was selected for the univariate model where there was a non-zero probability of selecting the outcome with the small effect size. If this were the case, the power curve for the univariate model would be reduced in the second and third scenarios. Therefore, the global rate parameter MLRR model is also advantageous over a univariate approach since




(a)




(b)




(c)

Figure 3.2: Power curves for testing group differences in the rate of change for two outcomes between two groups for the univariate LRR model (red), the MLRR model with separate rate parameters for each outcome (green), and the MLRR model with a global rate parameter (blue). **(a)** The true data was generated from a model where the rate parameter for each outcome was the same and the GMPR assumption was correct. The global rate effect size is 25%. **(b)** The true data was generated from a model where the rate parameter for the second outcome was half the size of the rate parameter for the first outcome. The rate effect sizes for the two outcomes are 25% and 12.5%. **(c)** The true rate parameter is 25% for the first outcome and is 0% for the second outcome. The power curve for the univariate LRR model was generated from testing the first outcome whose rate effect was 25% in each scenario.

it avoids the need for outcome selection.

### *3.4 Application*

We illustrate the global shared parameter MLRR method on the infant growth study. For simplicity, we focus on two outcomes: weight and crown-heel length. A global shared parameter MLRR model was constructed to examine global differences in rates of change of growth outcomes across groups defined by sex, treatment, and HIV infection status. Table 3.1 presents coefficient estimates and 95% confidence intervals for the MLRR model for weight and crown-heel length. The three covariates were used to estimate both mean level and rate level differences. In concordance with the more appropriate time trend structure described in Subsection 2.4.1, the reference time function was modeled using a natural cubic spline with knots at 150, 500, and 1100 days. A mixed effects model was used to structure the covariance matrix. Separate estimates were produced for main effects and time trend coefficients associated with each outcome, and global estimates were produced for rate effects of the MLRR model.

In Table 3.1, the main effect estimates for the global shared parameter MLRR model showed large mean differences at birth due to sex and treatment. Males were estimated to be 0.30 Kg (95% CI = (0.24, 0.37)) heavier and 1.15 cm (95% CI = (0.95, 1.35)) taller at birth on average compared to females. Infants whose mothers were assigned to nevirapine were estimated to have lower weight (Mean Diff. = -0.12 Kg, 95% CI = (-0.19, -0.05)) and crown-heel length (Mean Diff. = -0.32 cm, 95% CI = (-0.53, -0.12)) at birth on average compared to those on zidovudine. There was no evidence that infants who tested HIV positive differed in weight at birth from infants who tested negative (Mean Diff. = 0.07 cm, 95% CI = (-0.02, 0.16)), but there was evidence for a difference in crown-heel length at birth (Mean Diff. = 0.57 cm, 95% CI = (0.31, 0.83)) with HIV positive infants estimated to be taller. The time trend for these two outcomes is illustrated in Figure 3.3 which presents scatter plots of both outcomes among HIV negative and HIV positive infants with the corresponding fitted line from the global rate parameter MLRR model. The global rate effects estimated a small difference in the rate of change in the two growth outcomes due to sex and treatment with males growing at 1% slower rate and essentially no difference in the

Table 3.1: Group differences in the rate of change for growth outcomes among infants exposed to HIV infection based on a global shared parameter MLRR model. Covariates were included for mean level and rate level differences. A natural cubic spline with knots at 150, 500, and 1100 days was used as a reference time trend for both outcomes. The covariance structure was specified using a mixed effects model. Main effect estimates and time trend coefficient estimates are provided for each outcome. A global rate effect was estimated for each covariate.

| | Weight (Kg) | | Crown-Heel Length (cm) | |
|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI |
| Main Effects | | | | |
| Intercept | 3.36 | (3.29, 3.43) | 49.32 | (49.12, 49.53) |
| Sex (Male) | 0.30 | (0.24, 0.37) | 1.15 | (0.95, 1.35) |
| Treatment (Nev) | -0.12 | (-0.19, -0.05) | -0.32 | (-0.53, -0.12) |
| HIV Status | 0.07 | (-0.02, 0.16) | 0.57 | (0.31, 0.83) |
| Time Trend | | | | |
| Basis 1 | 5.64 | (5.52, 5.77) | 25.62 | (25.24, 26.00) |
| Basis 2 | 8.97 | (8.81, 9.13) | 37.14 | (36.66, 37.61) |
| Basis 3 | 16.34 | (16.06, 16.61) | 66.05 | (65.24, 66.87) |
| Basis 4 | 10.46 | (10.28, 10.63) | 44.09 | (43.55, 44.63) |
| Rate Effects | Estimate | | 95% CI | |
| Sex (Male) | -0.01 | | (-0.03, -0.00) | |
| Treatment (Nev) | 0.00 | | (-0.01, 0.02) | |
| HIV Status | -0.11 | | (-0.12, -0.09) | |

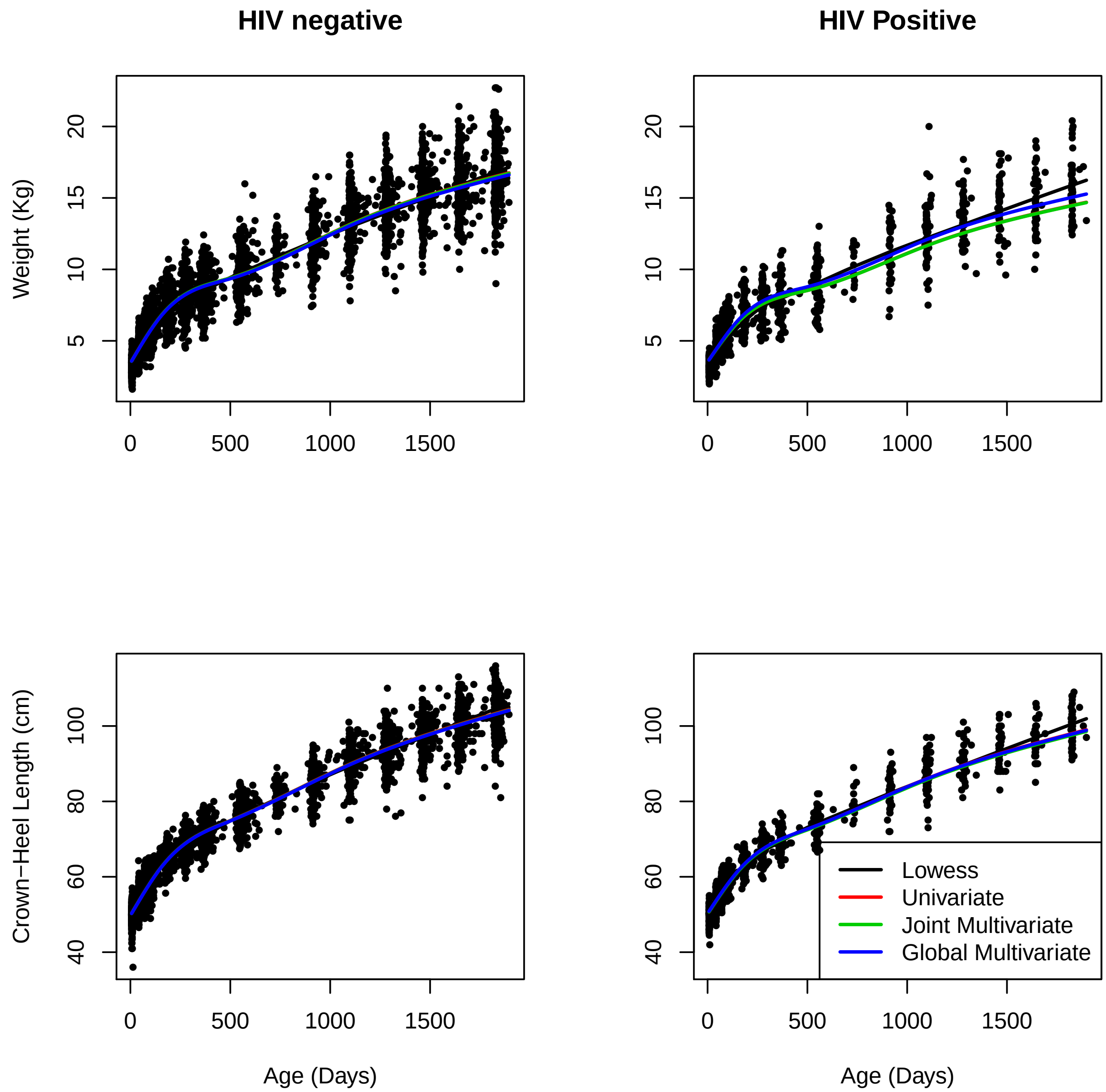
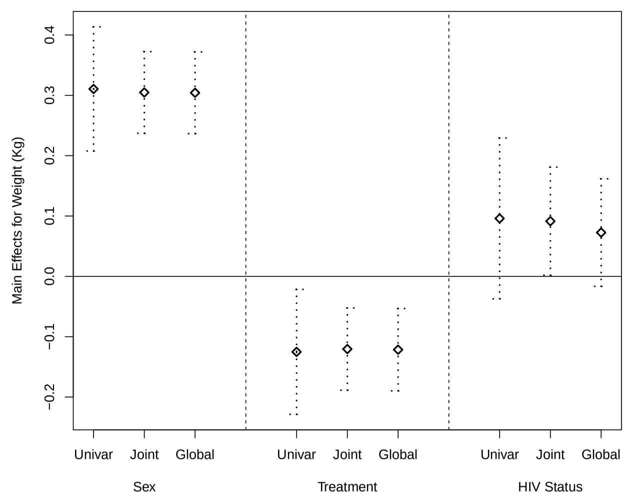


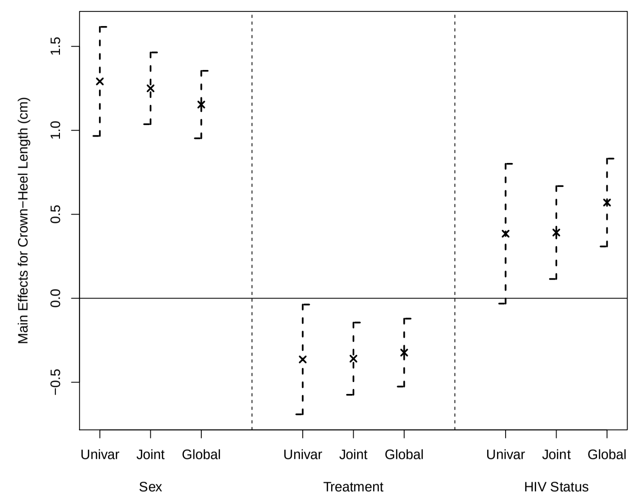

Figure 3.3: Change in growth outcomes over time by HIV status. A lowess smooth line for each scatter plot is included in red. The fitted line for the univariate models is depicted in red, the joint MLRR model in green, and the global MLRR model in blue.

rate of change for infants on nevirapine compared to zidovudine (see Table 3.1). The rate effect for sex was borderline significant with a 95% confidence interval of (-2.7%, -0.2%). There was strong evidence for a difference in the rate of change in growth outcomes due to HIV infection status with infants infected with HIV estimated to have 11% decrease in the rate of change (95% CI = (-12%, -9%)) compared to non-infected infants. Therefore, the model shows evidence that weight and crown-heel length differ across groups defined by sex and treatment at the mean level primarily where as HIV infection status primarily impacted the rate of change for these outcomes.

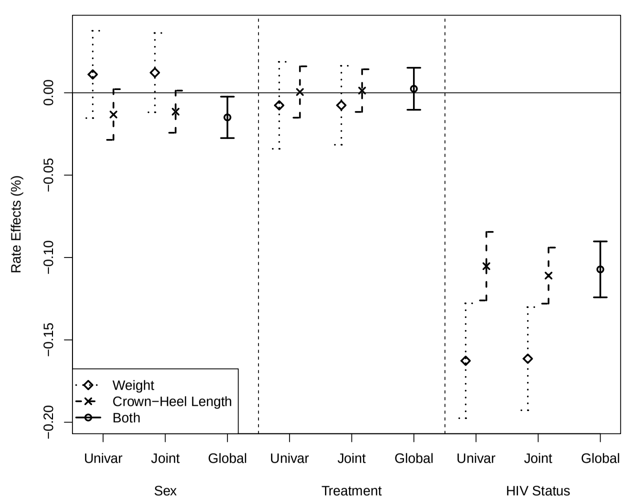
We illustrate the impact global shared parameter model has on model estimates in Figure 3.4 by comparing the confidence intervals for the above model to the confidence intervals from the equivalent univariate LRR models and MLRR models with separate rate parameter estimates for each outcome. Confidence intervals for the main effect estimates are provided for each outcome in Figure 3.4(a) and (b). The main effect plots show the gain in precision for the main effect estimates obtained by both multivariate models compared to the univariate model. There were also some mild to moderate differences in main effect estimates for global multivariate model compared to the univariate estimates. Conclusions for significant main effect differences from zero changed only for HIV status in relation to crown-heel length and also for weight in the case of the joint multivariate model. There were also some mild differences between estimates and confidence intervals for the intercept (not shown). The rate effect estimates also showed gains in precision for the multivariate models although there were no apparent additional gains in precision for the global MLRR model (see Figure 3.4(c)). The gain in precision was particularly evident for HIV status where the confidence bounds were reduced by a full percentage point. Differences in rate effect size between the two outcomes may partially explain why no additional precision was obtained by the global MLRR model. The rate estimates for the global MLRR model were most similar to the estimates for crown-heel length provided by the univariate LRR and joint MLRR models. The global estimate may have been dominated by crown-heel length outcome since the standard error for the rate estimates were smaller for crown-heel length than they were for weight. The global MLRR model altered the scientific conclusion for differences in the rate of change for sex where a borderline significant global effect was




(a)




(b)




(c)

Figure 3.4: Parameter estimates and 95% confidence intervals are plotted for several models for estimating group differences in growth outcomes for infants exposed to HIV infection. Each model regressed the growth outcome(s) against sex, treatment, and HIV infection status. Results are provided for univariate LRR models for weight and crown-heel length, a MLRR model with separate rate estimates for each outcome, and a global shared parameter MLRR model. **(a)** Main effect estimates and confidence intervals for group differences in weight (Kg). **(b)** Main effect estimates and confidence intervals for group differences in crown-heel length (cm). **(c)** Rate effect estimates and confidence intervals for the two growth outcomes. Each estimate and interval is indicated for whether they correspond to weight, crown-heel length, or both.

estimated. In addition to the time trend for the global MLRR model, the time trend for the univariate and joint multivariate models are plotted in Figure 3.3. The scatter plots indicate very little difference in the estimated time trend for these models.

In Chapter 2, diagnostic plots and testing procedures for validating the PR assumption for the LRR model were discussed. These procedures may also be used to evaluate the proportionality of rates for the MLRR model. Additionally, we can also evaluate the global rate parameter assumption for the MLRR model using diagnostic tests. To test the global parameter assumption for the infant growth data, the estimates from the joint MLRR model were used to test for a difference between rate effects for each outcome. The test for differences suggested a significant difference in the rate effects across growth outcomes for sex (P-value $= 0.01$) and HIV (P-value $< 0.01$). There was no evidence for a difference in the rate effects associated with treatment (P-value $= 0.33$). These test support violations of the global shared parameter assumption and suggest that the MLRR model with separate rate effect estimates may be more appropriate for this application.

## 3.5 Discussion

Methodology for multivariate longitudinal data has received limited attention in the statistical literature. Since the routine collection of detailed longitudinal data is becoming more common in the scientific community, more applications where multivariate longitudinal methods are potentially advantageous will likely arise. The work discussed in this chapter has focused on utilizing multivariate longitudinal data to model group differences. Comparing longitudinal changes across groups is a common scientific aim, and established methodology has separately shown the utility of longitudinal data and multivariate data for making inference. Using both data types presents challenges, but also provides gains in power and precision that counter the added model complexity.

When modeling multivariate longitudinal data, two paradigms of multivariate data have been used. In the latent variable approach proposed by Roy and Lin [2000] and Proust-Lima et al. [2007], the multivariate outcomes are considered to describe an unmeasured variable related to each of the measured outcomes. The latent variable methodology is less appropriate for inference on the measured outcomes since a covariates effect can only be interpretted indirectly in terms of the association with the latent variable. Since we were primarily interested in direct inference on measured outcomes, we chose not to focus on latent variable methodology for the work proposed in this chapter. In order to make appropriate inference directly on correlated outcomes, separate modeling structure for each outcome will often need to be considered to account for the unique characteristics of each outcome. In particular, outcome specific time structures will be important. Methods for allowing separate modeling of multivariate outcomes have been proposed by Gray and Brookmeyer [1998, 2000], Travison and Brookmeyer [2007], Jia and Weiss [2009]. These methods exploit the multivariate data by generating an estimate for a global shared parameter associated with covariate-defined groups. The methodology developed in this chapter was formulated under a similar direct global shared parameter paradigm.

We proposed an extension to the LRR model that was introduced in Chapter 2. The LRR method directly and parsimoniously models differences in rates of change for longitudinal data. The model allows for a flexible time structure and a simple interpretation for the difference in the rate of change which makes the LRR method advantageous when the underlying time trend is non-linear. The extended MLRR model allows separate specifications for baseline covariate adjustment and reference time trend for each outcome. The mean of each outcome is linked by a global parameter for the difference in the rate of change for the multivariate outcome. Group differences can then be interpreted as global or overall difference in the rate of change of the multivariate outcome associated with the group.

Comparison of the power to detect group differences for the global shared parameter MLRR approach illustrated gains in power over testing a single outcome or testing multiple outcomes with separate rate parameters for each outcome. Gains in power were even possible when the true difference in the rate of change across groups is not the same for each outcome. However, a multivariate comparison of rates of change lost power relative to a univariate comparison when major differences existed between the rate effects for each outcome. Power for the univariate test for group differences was reliant on always selecting the outcome with the largest rate effect. The gains in power for the univariate model would be less when there is a non-zero probability of testing outcomes with a smaller rate effect. Therefore, for applications where differences in the rate of change between groups is expected to be reasonably similar across outcomes, a global shared parameter MLRR model will be beneficial for increasing the power and precision of the group comparison.

The application of the MLRR method to the growth study for infants exposed to HIV also illustrates the potential power gains of a multivariate approach. The width of the confidence interval for the globally estimated rate effect was noticeably smaller than the confidence intervals based on univariate estimates. Examining differences in infant growth was a natural application for this method since there are multiple measures that can be considered to quantify growth. In addition, the rate of change is a useful scale for comparison since they are invariant to scale. The non-linear nature of infant growth data make the flexible time structure of the LRR method appealing. There are other application where the multivariate rate regression approach would be useful for the comparison of groups such as longitudinal studies of aging and treatment trials with multiple end points of interest. A global shared parameter MLRR approach is also potentially applicable in areas where effect sizes are commonly small such as in studies of environmental exposures.

The utility of the MLRR method is dependent on the non-linearity of the outcomes over time. If all outcomes change linearly or approximately linearly over time, then a linear mixed model approach will be more suitable for comparing groups. Whether the MLRR approach is advantageous when some outcomes behave linearly and others are non-linear is worthy of further research. The GMPR assumption is also a limitation of the MLRR method when the difference in the rate of change is not proportional over time or is not the same across outcome. Diagnostic approaches for assessing the proportional rate assumption for a univariate LRR model are presented in Chapter 2 and can be similarly applied for model checking at the multivariate level. Relaxing the proportional rate assumption in regard to time is discussed further in Chapter 2. The assumption of a global rate parameter can also be tested based on multivariate models with separate rate parameter estimates. When the global rate assumption is violated, both the univariate LRR and the MLRR approaches with separate rate parameters for each outcome could be considered.

# Chapter 4

# A SEMI-PARAMETRIC METHOD FOR MODELING DIFFERENTIAL RATES OF CHANGE USING LONGITUDINAL DATA

We now shift focus back to the univariate setting for modeling rates of change using longitudinal data. The goal of this chapter is to expand the utility of methodology proposed in Chapter 2. We do this by incorporating non-parametric estimation for the reference time function into the parametric estimation of differences in rates of change. Using non-parametric estimation for a trend in time can potentially address issues of model misspecification and makes the model more user-friendly since the resulting semi-parametric model would not require the user to specify the form of the reference time function.

## *4.1 Introduction*

A common concern in longitudinal data analysis is estimating population or individual-specific averages over time when the trend in time is non-linear. When the trend of an outcome over time is considered to be smooth and non-linear, two general approaches are used. A parametric approach such as a linear model on a polynomial basis of time offers the advantage of relatively straightforward estimation and inference. However, misspecification of the time structure can be a common issue since the progress of an outcome over time may not conform to simple parametric basis. Therefore, non-parametric estimation for longitudinal data is a useful alternative for offering flexibility in estimating time structure. The non-parametric approach will provide protection against model misspecification provided that the data are reliable for reflecting the underlying trend in time. Flexibility is particularly advantageous for longitudinal methods where inference on the time structure is not of direct interest. For instance, the longitudinal rate regression (LRR) method proposed in Chapter 2 estimates a non-linear, reference time structure in order examine differences in the rate of change for an outcome across groups. In this chapter, we explore the use of non-parametric methods for time structure estimation in the LRR method through the use of penalized splines.

Methodology for penalized splines is an extension of smoothing spline where piecewise polynomial functions are connected across a collection of knots defined by a covariate [Fitzmaurice et al., 2009]. In the case of longitudinal data, the covariate will typically be a measure of time. For a smoothing spline function, knots are placed at all observed values of the covariate and placed under smoothing constraints. Typically, these constraints include continuity and differentiability conditions. The major disadvantage of smoothing splines is the high dimensionality of the method. As a result, the basic smoothing spline approach will typically have high degrees of freedom and be computationally intensive. One means for compensating for this high dimensionality is incorporating a pre-specified penalty parameter and function into the score equations to restrict the estimation of the penalized spline function. It is common practice to select the penalty function such that the penalty parameter controls the roughness of the smooth spline where the smoothing spline will become more linear as the penalty parameter increases.

Using a penalized likelihood approach for smoothing splines presents the issue of deciding how to specify the penalty parameter. One approach is to use cross-validation (CV) to minimize the mean squared error for select subsamples of the data relative to the value of the penalty parameter. Jacqmin-Gadda et al. [2002] used a leave-out-one-subject CV approach for selecting a penalty parameter for correlated longitudinal data. Specifying the penalty parameter using CV will some times be influenced by leverage points. Thus, when analyzing independent data, Craven and Wahba [1979] proposed generalized cross-validation (GCV) in order to downweight the influence of these points. An alternative to CV and GCV approaches is to use general maximum likelihood (GML) [Wahba, 1985] which takes advantage of the equivalence between smoothing splines estimates and the best linear unbiased prediction (BLUP) from a linear mixed-effects (LME) model. Under this approach, the BLUP estimates are generated using the restricted maximum likelihood (REML) algorithm. The use of a LME model to specify the smoothing parameter in a non-parametric model is counter intuitive since the non-parametric methods are generally intended for applications where a parametric approach is not justifiable. However, Wahba [1985] showed that estimation of the penalty parameter using the GML approach performed similarly to a GCV approach.

The introduction of the penalty term into the estimation of a smoothing spline function serves the purpose of reducing the effective degrees of freedom in a high-dimensional space. Despite this reduction, estimation of a smoothing spline function will remain computationally intensive when a large number of knots is used for a given dataset i.e. when the number of distinct covariate values is large. To reduce the computation burden in these situations, it is common to estimate smoothing splines using a reduced number of knots [Fitzmaurice et al., 2009]. Such an approach is commonly referred to as penalized splines. Eilers and Marx [1996] first suggested the use of penalized splines. The proposed P-spline method [Eilers and Marx, 1996] uses a B-spline basis with a penalty function to control high-order difference of the coefficients of adjacent B-splines. Other specifications have also been used, the most common approach being a B-spline basis with a cubic smoothing spline penalty [Fitzmaurice et al., 2009]. The advantage of this specification is that the resulting spline estimate is equivalent to the unique natural cubic spline fit to the data when all unique covariate values are used as knots.

The appeal of penalized splines is in offering a flexible specification to the functional relationship between an outcome and a covariate similar to that of smoothing splines with the ability to control the effective degrees of freedom and the computational burden of the resulting estimates. These characteristics make penalized splines ideal for use in longitudinal data in instances where the time trend of an outcomes is non-linear and not easily specified. Furthermore, penalized splines is useful in semi-parametric longitudinal applications where a longitudinal trend is estimated, but is not the primary focus of the methodology. Semi-parametric models for longitudinal data using penalized splines have been proposed previously by Zhang et al. [2000] and Ruppert et al. [2003].

In the LRR method proposed in Chapter 2, a reference trend over time is estimated so that the rate of change of the mean of an outcome can be connected across groups such that the difference in the rate of change is proportional across groups relative to the estimated trend in time. The use of penalized splines to estimate a reference trend for the LRR model will be advantageous for protecting against model misspecification when the underlying time trend structure is not well known. In addition, a more data-driven non-parametric approach will be more susceptible to data intricacies, and therefore, should be used primarily for highly dense data where outlying observations will have less influence. We introduce methodology here for extending the LRR method to a semi-parametric model using penalized splines. In Section 4.2, the methodology for semi-parametric estimation of the LRR model with reference time functions specified as a penalized spline is described. We also outline a means of penalizing the estimation and selecting an appropriate penalty parameter value. The semi-parametric LRR method is illustrated in Section 4.3 using the infant growth study. Finally, we offer discussion of the semi-parametric approach to estimating differences in the rate of change in Section 4.4.

## *4.2 Methods*

The LRR model was constructed by restricting the rate of change for a longitudinal outcome to be proportional across covariate-defined groups. Let $Y$, $t$, and $\mathbf{X}$ respectively denote a longitudinal outcome, the time of measurement for the outcome, and a vector of covariates defining groups. We use subscript $i$ to denote an individual for $i = 1, \ldots, N$, and subscript $j$ to denote a time point of measurement with $j = 1, \ldots, n_i$. Suppose that the aim of scientific interest is to compare how the rate of change of $Y$ differs across groups defined by $\mathbf{X}$. The LRR method proposes to model this difference by imposing the Proportional Rate (PR) assumption which can be expressed as

$$\frac{\partial \mu_{\mathbf{x}}(t)}{\partial t} = (1 + \boldsymbol{\theta}^T \mathbf{x}) \frac{\partial \mu_0(t)}{\partial t}$$

for all values of $\mathbf{X} = \mathbf{x}$, where $\frac{\partial \mu_{\mathbf{x}}(t)}{\partial t}$ is a function of $t$ that describes the rate of change of the expected value of $Y$ for a given value of $\mathbf{X} = \mathbf{x}$ over time and $\frac{\partial \mu_0(t)}{\partial t}$ describes the rate of change of the expected value of $Y$ for a preselected group defined by $\mathbf{X} = \mathbf{0}$. The percent increase (decrease) in the rate of change of the expected value of $Y$ that is attributable group differences in $\mathbf{X}$ is quantified by the parameter vector $\boldsymbol{\theta}$. Given the PR assumption, the full mean structure of the LRR model can be specified as

$$E[Y_{ij} \mid \mathbf{X}_i = \mathbf{x}, t_{ij} = t] = g(\mathbf{x}) + (1 + \boldsymbol{\theta}^T \mathbf{x}) \mu_0(t)$$

where $g(\cdot)$ is a function that specifies the expected value of $Y$ across groups defined by $\mathbf{X}$ at a time origin, or time zero, and we specify $\mu_0(0) = 0$. We here on refer to $g(\cdot)$ as the baseline function and $\mu_0(\cdot)$ as the reference time function. Typically, we express the baseline function as a linear combination of $\mathbf{X}$ i.e. $g(\mathbf{X}) = \boldsymbol{\alpha}^T\mathbf{X}$.

See Chapter 2 for additional detail into the LRR method. In the previous chapters, we estimated the reference time function using a linear combination of a parameteric basis i.e. $\mu_0(t) = \boldsymbol{\beta}^T\mathbf{T}$ where $\mathbf{T}$ is a vector of functions evaluated at time $t$. We now outline a modified approach for estimating the LRR model where the reference time function is estimated non-parametrically using penalized splines.

### *4.2.1 Penalized Spline Estimation*

Given longitudinal data observed over the time interval $[a, b]$, we estimate a reference time function for the LRR model using a penalized spline function. First, we select knots $\nu_1, \cdots, \nu_q$ such that $a < \nu_1 < \cdots < \nu_q < b$. Typically, we specify knots at all uniquely observed time points in the data provided this is computationally feasible. Using these knots, we may construct a spline basis which we denote $B_1(t), \cdots, B_p(t)$ where $p$ depends on $q$ and the type of spline basis chosen e.g. a cubic B-spline basis. Then, we can express the reference time function as a linear combination of this spline basis:

$$\mu_0(t) = \sum_{k=1}^{p} \beta_k B_k(t). \tag{4.1}$$

When the number of spline functions, $B_k(t)$ and consequently the number of coefficients, $\beta_k$, is large relative to the size of the data, we introduce a penalty function, $P(\cdot)$, and penalty parameter, $\lambda$, into the estimation procedure to restrict the estimation of these parameters. The penalty function and parameter are incorporated into the estimation by maximizing a penalized likelihood equation, $l_P(\cdot)$ that is equal to the standard likelihood equation $l(\cdot)$ plus the penalty term. That is,

$$l_P(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = l(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \lambda P(\boldsymbol{\beta}). \tag{4.2}$$

The penalized estimation process limits the effective degrees of freedom [Hastie and Tibshirani, 1990] spent on estimating these parameters. The manner by which the estimation of coefficients for $\mu_0(t)$ is restricted depends on the penalty function selected. The penalty parameter controls the degree to which the estimation is restricted.

In this chapter, we use a cubic smoothing spline penalty function to outline estimation of the LRR model with a penalized spline reference time function. The cubic smoothing spline penalty was selected since it is commonly used in practice. Modifying the estimation procedure for other penalty functions is straightforward. The cubic smoothing spline penalty is defined by the expression

$$P(\boldsymbol{\beta}) = \int_a^b \left[ \sum_{k=1}^{p} \beta_k B_k^{(2)}(t) \right]^2 dt$$

where $B_k^{(2)}(t)$ denotes the second derivative of the $k$th B-spline function evaluated at time $t$ [Fitzmaurice et al., 2009]. In other words, the cubic smoothing spline penalty penalizes the second derivative or the curvature of the spline function. Therefore, the degree of curvature estimated for the reference time function will vary relative to the penalty parameter, $\lambda$. When $\lambda = 0$, the curvature of the reference time function is unrestrained. As $\lambda \to \infty$, curvature is reduced, and the estimated reference time function becomes more linear.

An appealing feature of the cubic smoothing spline penalty function is that it may be re-expressed into a quadratic form that is convenient for incorporating the penalty term into score and hessian equations used for estimation. The equivalent quadratic form is written as

$$P(\boldsymbol{\beta}) = \boldsymbol{\beta}^T D \boldsymbol{\beta}$$

where $D$ is a matrix with $(k, l)$th entry

$$D_{kl} = \int_a^b B_k^{(2)}(t) B_l^{(2)}(t) dt.$$

The entries of $D$ can be calculated by hand or through the use of established software such as the *getbasispenalty* function in the *fda* package in R. For a given value of $\lambda$, the penalty matrix can be incorporated into the penalized likelihood, and parameter estimates can be obtained by maximizing the penalized likelihood using standard maximization procedures.

### *4.2.2 Selecting the Penalty Parameter*

Prespecifying an appropriate value for the penalty parameter, $\lambda$, can often be problematic since it is difficult to know *a priori* how smooth an estimated function should be and what value of $\lambda$ will correspond to a specific degree of smoothness. Thus, it is common to use data based methods to select an appropriate value for the penalty parameter. We caution that data based methods for parameter selection are generally to be avoided when one is interested in inference on model parameters, and so, choosing a value for $\lambda$ before consulting the data should be done when possible. However, since the LRR model is focused on inference on the rate of change, using methods such as CV to generate an appropriate underlying time trend is a reasonable approach when little is known about the behavior of the outcome. For the LRR model, we propose a penalty parameter selection procedure that is adapted from an approach for smooth curves for longitudinal data suggested by Jacqmin-Gadda et al. [2002]. The procedure uses cross-validation to estimate the predicted mean squared error for an array of models with varying values of $\lambda$.

Let $\boldsymbol{\eta}$ denote the vector of parameters in the mean function of the LRR model. We use $\widehat{\boldsymbol{\eta}}_{-i}(\lambda)$ and $\widehat{V}_{-i}(\lambda)$ to respectively denote the penalized maximum likelihood estimates (PMLEs) for the mean parameters and covariance matrix obtained from the sample with subject $i$ omitted and with a fixed penalty, $\lambda$. Define the function $f_i(\cdot)$ as the fitted mean vector for $\mathbf{Y}_i$ given parameter values $\boldsymbol{\eta}$. For a given value of $\lambda$, calculate the approximate cross-validation (aCV) criterion that is expressed as

$$\text{aCV}(\lambda) = \sum_{i=1}^{N} \left( \{\mathbf{Y}_i - f_i[\widehat{\boldsymbol{\eta}}_{-i}(\lambda)]\}^T [\widehat{V}_{-i}(\lambda)]^{-1} \{\mathbf{Y}_i - f_i[\widehat{\boldsymbol{\eta}}_{-i}(\lambda)]\} \right). \tag{4.3}$$

For many applications precise estimation of the PMLEs for the mean and variance parameters for each subsample and each value of $\lambda$ will be computationally intensive. Thus, Jacqmin-Gadda et al. [2002] offer two suggestions for reducing computational burden. The first simplification is to assume that the variance is constant across all subsamples and values of $\lambda$. Secondly, Jacqmin-Gadda et al. [2002] propose a one-step estimation procedure for generating approximate PMLEs for the mean parameters for a given subsample and value of $\lambda$ based on the true PMLEs for the mean parameters based on the full sample and a preselected value of $\lambda$. For the proposed approximation procedure, a fixed penalty parameter value, $\lambda_0$, is selected and the PMLEs, $\widehat{\boldsymbol{\eta}}(\lambda_0) = \widehat{\boldsymbol{\eta}}_0$ and $\widehat{\mathbf{V}}(\lambda_0) = \widehat{\mathbf{V}}_0$, are calculated based on the full sample. Jacqmin-Gadda et al. [2002] recommend selecting a value between 0 and 100 for $\lambda_0$ in the approximation procedure. Then, for the subsample that removes individual $i$ and a general penalty parameter value, $\lambda$, the approximation assumes that $\widehat{V}_{-i}(\lambda) = \widehat{\mathbf{V}}_0$ and $\widehat{\boldsymbol{\eta}}_{-i}(\lambda) = h(\widehat{\boldsymbol{\eta}}_0)$ for a one-step formula defined by the function $h(\cdot)$. A one-step estimator for a linear model was presented by Jacqmin-Gadda et al. [2002], and we extend this one-step estimator for non-linear models in Appendix B. In summary of the one-step estimator, the PMLEs for the mean parameters based on a sample that omits individual $i$ and a given value of $\lambda$ can be calculated using the formula

$$\widehat{\boldsymbol{\eta}}_{-i}(\lambda) = \left(\sum_{j \neq i} \widehat{f}_j^{(1)T} \widehat{\mathbf{V}}_{0j}^{-1} \widehat{f}_j^{(1)} + \lambda \Omega\right)^{-1} \left[\sum_{j \neq i} \widehat{f}_j^{(1)} \widehat{\mathbf{V}}_{0j}^{-1} (\mathbf{Y}_j + \widehat{f}_j^{(1)} \widehat{\boldsymbol{\eta}}_0 - \widehat{f}_j)\right] \tag{4.4}$$

where $f_j^{(1)}$ denotes the vector of first derivatives of the mean structure with respect to the mean parameters for individual $j$, and $\Omega$ is a matrix with entries corresponding to the values of the penalty matrix $D$, described previously, for the coefficients of the reference time function and zeros everywhere else. The resulting PMLEs can be used in Equation (4.3) to approximate the aCV criterion for each value of lambda. By searching for a value of $\lambda$ that minimizes the approximated aCV criterion, a suitable value for the penalty parameter can be selected.

The advantage of the approach described by Jacqmin-Gadda et al. [2002] is that it is computationally simple since the longitudinal model only needs to be maximized once for a single value of $\lambda$. The reduced computation burden is achieved by assuming a constant variance across subsamples and values of $\lambda$ and by using a one-step estimator to approximate the PMLEs for the mean parameters. We provide evaluation of the impact of these assumptions on the penalty parameter selection procedure in the following application. Though the constant variance assumption may impact the estimation of the aCV criterion, one may argue that the assumption is more reasonable for selecting a penalty parameter than if the variance were allowed to vary across values of $\lambda$. In selecting a penalty parameter, we are primarily interested on the penalty parameters impact on the mean parameters, but the covariance matrix will also change with values of $\lambda$. Thus, calculating the aCV criterion based on varying mean and variance parameters will evaluate the impact of $\lambda$ on both the mean and variance parameters. If we instead fix the variance across values of $\lambda$, we more directly evaluate the impact of $\lambda$ on the mean parameters. The aCV criterion in this case could be considered to be a CV selection procedure based on a fixed Mahalanobis distance. Based on this consideration, the impact of the one-step estimator on the penalty parameter procedure is considered of greater interest to the validity of the approach. Further work is needed to separately evaluate the influence of these assumptions. The CV procedure above was outlined based on a leave-out-one-subject CV procedure. We implement this procedure based on leaving-out-several-subjects procedure to reduce computational burden due to the large sample size of our illustration.

## *4.3 Application*

The univariate analysis of the infant growth study using the LRR method that was presented in Chapter 2 is re-evaluated here using the semi-parametric approach to the LRR model. Weight is regressed as the outcome with main effects and rate effects for sex, treatment, and HIV infection status. Penalized spline estimation using a B-spline basis was used to generate a reference time function. Interior knots were included at 1, 6, 10, and 15 weeks from birth; 6 months from birth, 1 year from birth, and every half year there after for a total of 17 knots including the boundary knots. Linear time was also included in the linear combination of the B-spline basis. A cubic smoothing spline penalty was used, and the penalty parameter was selected using the cross-validation procedure described in subsection 4.2.2 and sub-sampling by removing a tenth of the dataset at a time. Results from the semi-parametric LRR model were compared to results from the parametric LRR model (previously described in Subsection 2.4.1) in Table 4.1.

The results of the CV selection procedure for the infant growth study is depicted in Figure 4.1. An initial penalty parameter value of 1 was used to generate the fixed variance estimate and the initial values of the mean parameters for the one-step estimator. The aCV was estimated to be minimized at a penalty parameter value of 811 based on the approximation procedure. Select points were also included for estimating the true aCV criterion value based on a procedure where the actual PMLEs for the mean and variance parameters for each subsample and value of $\lambda$ were calculated. The results of the true aCV criterion indicate a minimizing value of $\lambda$ of approximately 1400. Therefore, the assumptions of the simplified CV selection procedure did impact the selection of $\lambda$. Though the results

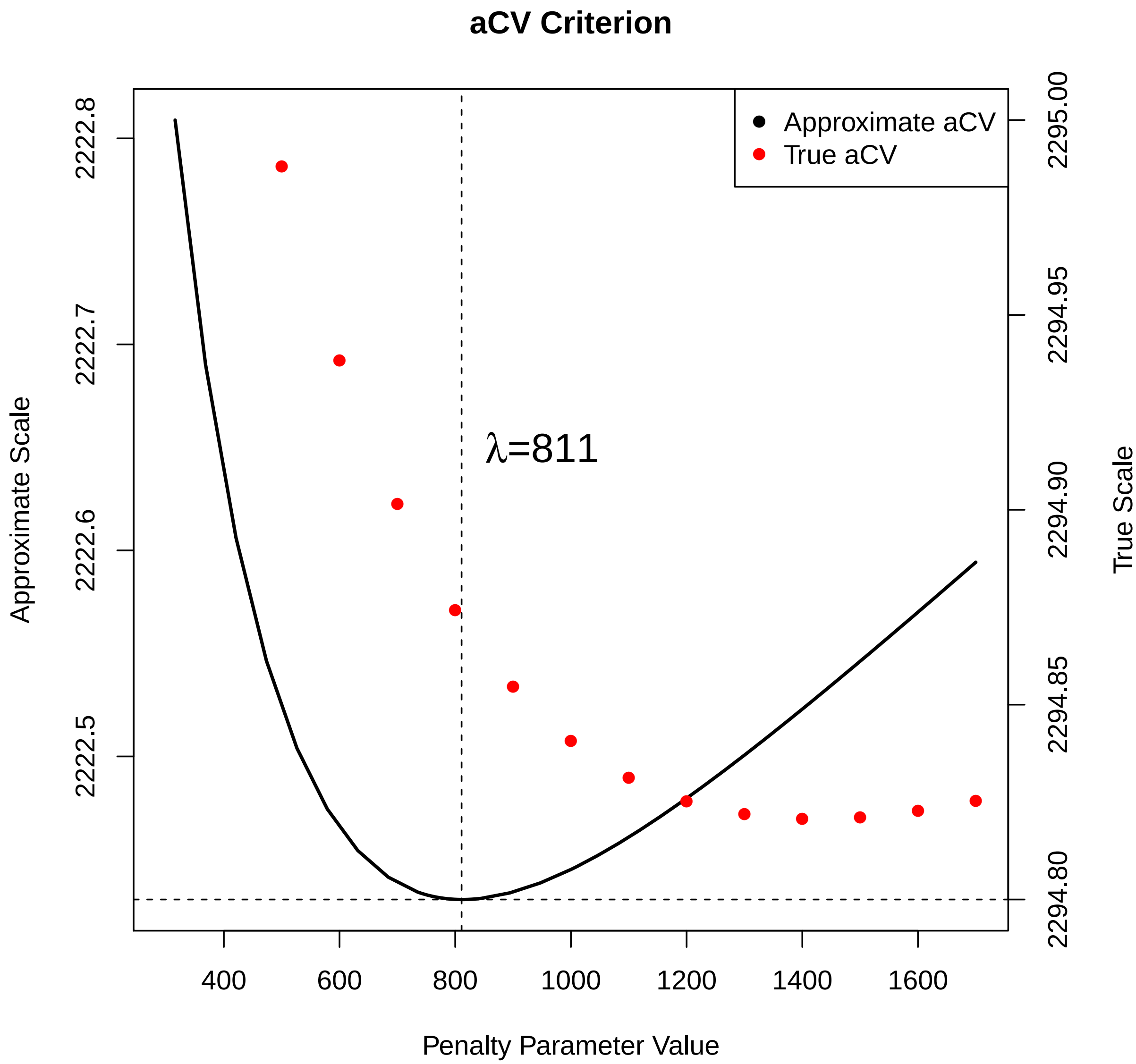



Figure 4.1: Graph of approximate values of aCV for models with a given penalty parameter value for the infant growth dataset. The approximate estimates were calculated based on the computationally simplified CV procedure where a tenth of the dataset was removed at a time and a reference model that had a penalty parameter of 1 was used. The dashed lines highlight the minimum aCV value and the penalty parameter where the minimum value was obtained based on the approximated curve. The aCV criterion was minimized by a penalty parameter value of 811. Select points were also included where the true aCV value was calculated based on a CV procedure that calculated fully iterated mean and variance estimates for each subsample and each penalty parameter value. The scale of the true aCV values is displayed to the right of the plot.

of these two procedures differ, the preferred penalty parameter value can be debated as discussed in Subsection 4.2.2. For this example, the semi-parametric LRR model was run using the results of the approximated aCV, and so, a penalty parameter value of 811 was used for the model results. The two computational approaches for the aCV differed greatly in computational speed. The approximate aCV approach generated results for 20 different penalty parameter values in approximately 5 or 6 hours where as the true aCV approach took between 25 to 30 hours of processing time to generate the aCV for a single value of $\lambda$.

We first provide an interpretation of the results from the semi-parametric LRR model presented in Table 4.1. Mean weight for the infants at birth was estimated to differ by 0.24 Kg (95% CI=(0.15, 0.33)) between Males and Females. Infants on nevirapine weighed 0.13 Kg less (95% CI=(-0.21, -0.04)) on average than infants on zidovudine. There was a small difference in weight at birth between children infected with HIV versus those that are HIV negative with HIV positive infants weighing 0.09 Kg more (95% CI=(-0.03, 0.20)) than HIV negative infants. There was little difference in the rate of change for weight due to sex and treatment. Males were estimated to grow at a 2% faster rate (95% CI=(-0%, 5%)) than females. Infants exposed to nevirapine were estimated to have a 1% reduction in the rate of change (95% CI=(-3%, 2%)) compared to infants exposed to zidovudine. A large difference in the rate of change was estimated for HIV status with HIV positive infants growing at a 16% reduced rate (95% CI=(-19%, -13%)) relative to HIV negative infants. The estimated penalized spline function for the reference time function is illustrated in Figure 4.2 along with 99% confidence interval for the reference curve. We discuss details for generating an asymptotically valid confidence interval for the reference function when estimated via a penalized spline equation in Appendix C to this dissertation.

In comparing the results of the semi-parametric LRR model to the parametric model results, the semi-parametric estimation approach had minimal impact on model conclusions (see Table 4.1). Mild differences in main effect estimates were presents with the largest difference for the coefficient for HIV status (0.04 versus 0.09) though there was also a large difference in the intercept value. There was also little evidence for a difference in precision for these estimates. In general, differences in the main effects are not of great concern since the focus of the LRR model will typically be on the rate effect estimates. The rate

Table 4.1: Estimates and 95% confidence intervals for the parametric LRR model and the semi-parametric LRR model for weight among infants exposed to HIV infection. The parametric model used parametric cubic spline bases with two knots to estimate the reference time function. The semi-parametric model estimated a reference time function using a penalized spline equation with 17 knots. A cubic smoothing spline penalty function with a penalty parameter value of 811 was used to restrict estimation for the semi-parametric model. Coefficient estimates are provided for main effects and rate effects for the covariates sex, treatment, and HIV infection status.

| | Parametric Model | | Semi-Parametric Model | |
|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI |
| Main Effects | | | | |
| Intercept | 3.38 | (3.29, 3.46) | 3.03 | (2.95, 3.12) |
| Sex (Male) | 0.26 | (0.17, 0.34) | 0.24 | (0.15, 0.33) |
| Treatment (Nev) | -0.12 | (-0.21, -0.03) | -0.13 | (-0.21, -0.04) |
| HIV Status | 0.04 | (-0.07, 0.16) | 0.09 | (-0.03, 0.20) |
| Rate Effects | | | | |
| Sex (Male) | 0.02 | (-0.01, 0.04) | 0.02 | (-0.00, 0.05) |
| Treatment (Nev) | -0.01 | (-0.03, 0.02) | -0.01 | (-0.03, 0.02) |
| HIV Status | -0.15 | (-0.18, -0.12) | -0.16 | (-0.19, -0.13) |

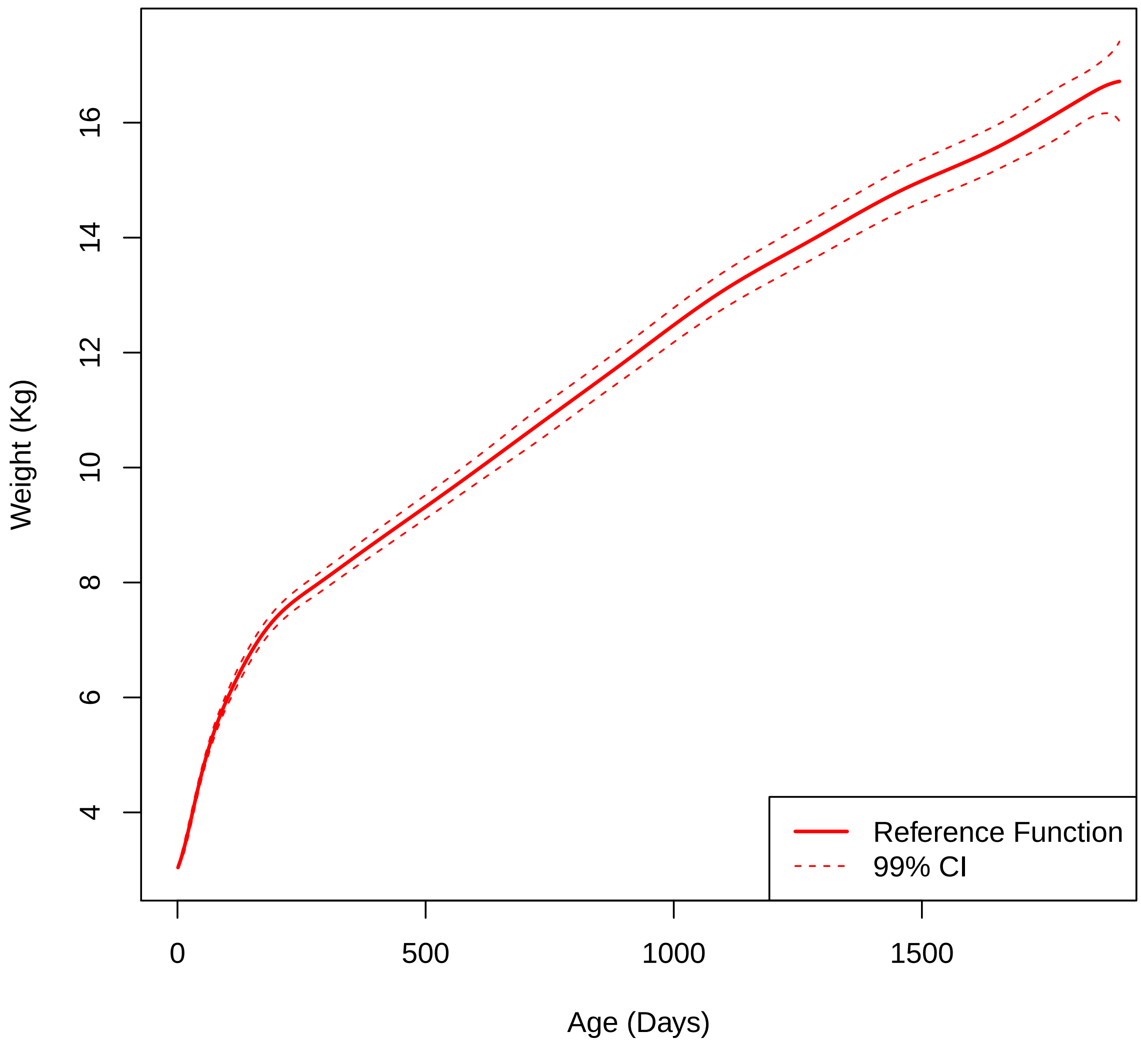



Figure 4.2: Plot of the estimated reference time function in red for the semi-parametric LRR model. The red dashed lines indicate the 99% Confidence interval for the curve.

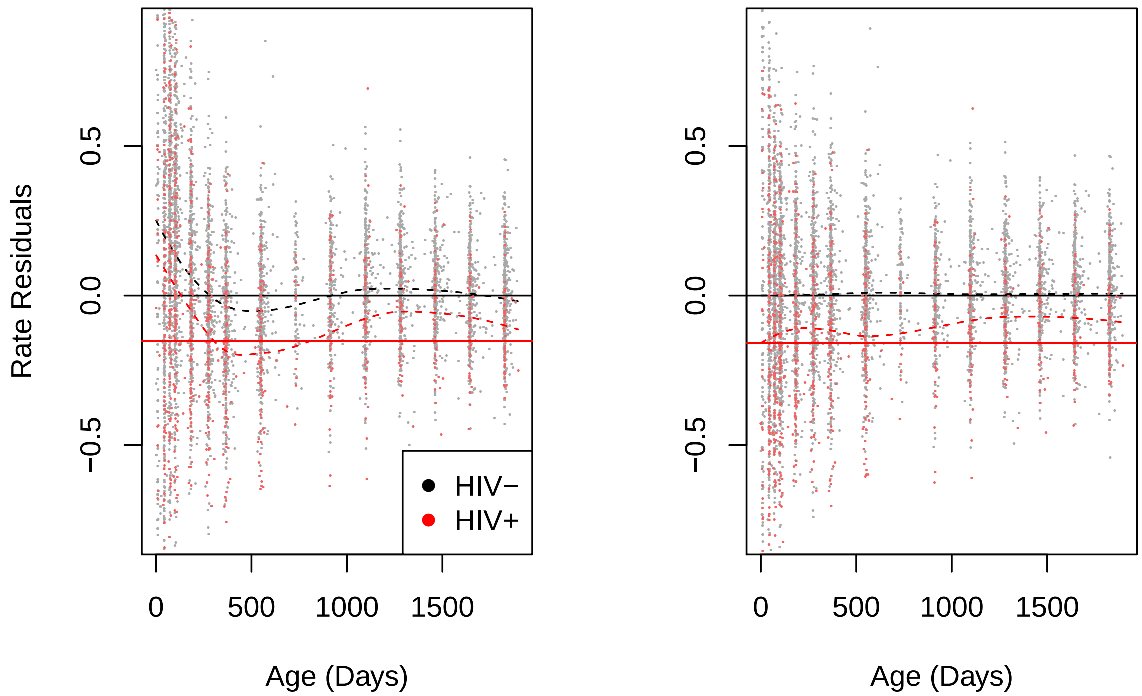



Figure 4.3: Rate residual plots for the rate effects associated with HIV status from the parametric model (left) and the semi-parametric model (right). Observations are marked by their HIV status. Horizantal lines are drawn at zero and at the rate effect size for each model. Lowess smooth lines were drawn for both HIV positive and negative infants.

effects for the two models were nearly identical apart from a one percentage point increase in the effect for HIV status and a small in the rate effect for sex conveyed by a shift in the confidence interval. Therefore, the semi-parametric model appeared to be equally effective for inference on the rate of change for this example.

As in Chapter 2, the modeling assumptions of the LRR method using a semi-parametric estimation can be evaluated using graphical and hypothesis testing diagnostic methods. We compare the validity of model assumptions for the semi-parametric model to validity in the parametric model by comparing plots of the rate residuals of each plots. The rate residuals were calculated based on the methods outlined in Subsection 2.2.4. A plot of the rate residuals for the semi-parametric model was plotted and compared to the rate residual plot for the parametric model in Figure 4.3. Comparison of the two plots show how the semi-parametric model dramatically reduces the model misspecification of the time trend. The rate residuals for the semi-parametric model also indicate some violations of the PR assumption since the lowess smooth curve shows that the residuals for HIV positive infants tend towards zero at older age.

## *4.4 Discussion*

The focused comparison of rates of change across groups is dependent on the underlying time trajectory of the outcome over which the rates of change are being compared. In order for inferences on differences in the rate of change to be effective, it is beneficial to ensure that the specification for the time trend for the outcome is correct. Justifying the correctness of an estimated time structure can depend either on an accurate and detailed function specification that is ideally supported by strong *a priori* knowledge of the outcome or on a robust, data-driven approach. The methodology described in Chapter 2 generated a model that compared rates based on a structured parametric time function that was specified by the user. Alternatively, estimating the time trend non-parametrically allows one to generate a generic, smooth curve for comparing rates of change without requiring additional information from the user.

A convenient approach to non-parametric estimation of a longitudinal curve is by using a penalized spline function since the non-parametric method has previously been adopted for longitudinal, semi-parametrically estimated mean structures [Zhang et al., 2000, Ruppert et al., 2003]. Several approaches can be taken for constructing a penalized spline function; some of the most common of these approaches are described in Fitzmaurice et al. [2009]. One primary consideration for using penalized splines is the penalization of the estimation of the function. For majority of applications, smoothing spline estimation must be penalized in order to reduce computational intensity and minimize the effective degrees of freedom used by the function [Hastie and Tibshirani, 1990]. The functional characteristic that is most commonly penalized is the roughness of the function. How strictly the roughness of a function is penalized will be dependent on the specified penalty parameter. The value of the penalty parameter can be preselected or selected based on the data using approaches such as CV, GCV, and GML [Jacqmin-Gadda et al., 2002, Craven and Wahba, 1979, Wahba, 1985]. The computational burden of estimation of a smooth spline function can also be moderated by the number of knots selected for the spline basis which has been used to distinguish penalized splines from a standard smoothing spline estimation approach [Fitzmaurice et al., 2009].

Penalized spline estimation of a smooth reference time function was incorporated with parametric estimation of differences in rates in order to model group differences in the rate of change of a longitudinal outcome. The proposed semi-parametric model was adapted from the LRR method proposed in Chapter 2. Methodology was described for a model that used a b-spline basis for the penalized spline function and a cubic smoothing spline penalty function to restrict estimation. Both of these modeling decisions are commonly used in practice [Fitzmaurice et al., 2009]. Other modeling choices can be implemented using similar methodology. We also offered a CV approach for penalty parameter selection that is an extension of the selection procedure described by Jacqmin-Gadda et al. [2002]. By assuming a constant variance and using a one-step formula to estimate mean parameters for varying subsamples and varying values of $\lambda$, the CV procedure can be implemented with limited computational burden. Results for the semi-parametric LRR model for the infant growth study illustrated how suitable inference for differences in the rate of change can be obtained based on a non-parametrically estimated time trend.

Modeling the reference time function of the LRR method using a penalized spline helps improve the robustness of the model, but it also increases the methods dependency on having rich data. When the number of independent observations is low or measurement time points are sparse, precise estimation of an appropriate reference curve will be difficult. In addition, an estimated penalized spline curve will be more susceptible to influential points and data intricacies in this scenario compared to a more structured parametric estimation. Thus, one can argue that in some scenarios the semi-parametric approach will be more susceptible to model misspecification when the data is unreliable for estimation of a trend in time. Therefore, to an even greater extent than that which was discussed in Section

2.5, inferences on differences in rates of change across groups using a semi-parametric LRR model will suffer in comparison to a linear models approach when longitudinal data is sparse. Additional work is needed to determine the amount of data that is needed for a semi-parametric LRR approach to be an effective means of comparing rates of change.

Chapter 5

# CONCLUSIONS AND FUTURE WORK

The work presented in this dissertation has explored the ability of longitudinal data to evaluate differences in the rate of change of an outcome across groups. The introduced methodology deviates from the tradition focus on mean-level differences for longitudinal data to a focus on rate-level differences. The ultimate goal of the work is to convey the potential of methodology for making direct comparisons of rates of change in longitudinal data. Numerous areas of scientific research can benefit from methods for comparing rates of change. Studies of human growth, as has been illustrated, is a natural application for rate comparisons particularly among infants, adolescents, and aging populations where longitudinal trajectories are likely to be non-linear. Research on treatment trials and environment risk factors are among the other possible fields where rate comparisons may be useful.

Methodology was presented in Chapter 2 that directly structured differences in the rate of change relative to a general longitudinal trajectory. For the basic Longitudinal Rate Regression (LRR) model, a Proportional Rate (PR) assumption is used where rates differ between groups by a fixed proportion across time. Estimates for differences in the rate of change based on the specified structure had a simple and scientifically meaningful intrepretation as a percent difference in the rate of change. The rate structure was used to induce a full mean model specification that can be estimated using both mixed effect and estimating equation methodology. The validity of the modeling assumptions of the LRR model can be verified using diagnostic methods and can be relaxed by incorporating time-varying covariates in the rate structure. The LRR model demonstrated improvements for inference on differences in rates of change between groups compared to a linear model approach. The advantage of using the LRR method to compare rates primarily exists for applications to longitudinally dense data that is non-linear across time.

Since rate level differences described by the LRR model are scale free, the developed methodology for examining rates can be extended in a straightforward manner to multivariate longitudinal outcomes. In Chapter 3, we presented a multivariate longitudinal extension of the LRR model that estimated a global difference in the rate of change across groups. The Multivariate Longitudinal Rate Regression (MLRR) method incorporated a global rate effect with reference time functions and mean-level covariate adjustments that were uniquely specified for each outcome. Various variance structures were outlined for forming a full likelihood structure for estimating the MLRR method, and a mixed effect variance structure with random effects that were correlated across outcomes was implemented. The global rate effect estimate from the MLRR model had increased power and precision for detecting group differences in the rate of change compared to estimating separate rate effects for each outcome in both simulated and real applications. When major differences existed between the true rate effect associated with each outcome, the power for detecting group differences using a global rate effect was reduced and a separate rate estimate approach was more appropriate. Therefore, provided the rate effects are reasonably similar for each outcome, a global rate test for group differences in multivariate longitudinal data will be advantageous.

When uncertainty in the underlying time trend of an outcome exists, modeling differences in the rate of change relative to a time structure that is robust and flexible will be beneficial. Thus, we discussed a semi-parametric approach to estimating the LRR model in Chapter 4 that incorporated a non-parametric reference time function. Time was modeled as a penalized spline over which parametric differences in the rate of change could be estimated. Methodology for implementing the semi-parametric model was described using a b-spline basis with a cubic smoothing spline penalty. Both modeling choices are common in the research literature, and modifying the model for other common approaches is straightforward. Also, we described a computationally simple cross-validation procedure for penalty parameter selection. The semi-parametric LRR method preformed similarly to the basic parametric approach and is a useful approach to modeling differences in rates of change for applications with rich longitudinal data. In addition to increasing the robustness of inference to model misspecification given reliable data, the semi-parametric LRR method benefits the user by requiring fewer *a priori* decisions for the modeling structure.

Further research is needed on the limitations of the LRR method and its extensions to better understand where direct rate modeling approach is useful. Assessing the richness of data need for the LRR method to be advantageous over a linear models approach is a major concern. Primarily, extensive longitudinal data for describing a non-linear trend over time is needed for the LRR model. However, particularly for the semi-parametric modeling approach, the number of independent observations required for precise estimation is also of concern. Exploration would be useful for applications where other sources of information can be used model a reference time structure such as when exterior dense longitudinal data exists on the reference group for a dataset. Under such settings, the LRR model will be less limited by the need for dense longitudinal data for group comparisons. In the multivariate setting, the implications of linking related outcomes in varying scenarios can be explored. For instance, linking outcomes with a mix of linear and non-linear trajectories or linking bivariate outcomes. Also, additional detail would be useful for the similarities in rate effect sizes needed for a global rate effect estimate to be beneficial and how power gains relate to the variance structure especially in terms of measurement error imbalance across outcomes. Evaluation of the PR assumption made by the basic LRR method is another area of consideration. The ability to detect violations in the PR assumption may vary based on modeling assumptions. Specifically, the semi-parametric LRR method could mask deviations from proportionality more than a parametric approach as a result of the more data driven specification. Modifications to the PR assumption through time varying indicators alters model results in regard to rate effects, mean effects, and the estimated reference time function. Better understanding of the impact relaxation of the PR assumption has on model estimates is needed.

Other potential statistical application exists for modeling rates of change for longitudinal data. For example, models for rate of change can be viewed as a combination of estimation and prediction. When generating models that rely on estimation and prediction, the quality of inference on estimates of interest can depend on the approach to combining estimation and prediction [Bennet and Wakefield, 2001, Wakefield and Shaddick, 2006, Sheppard et al., 2010]. Thus, models for rates of change can be used to evaluate inference under competing approaches. For example, by necessity, a model for comparing rates of change in longitudinal data must predict an underlying time trajectory in addition to estimating rate differences. The methodology presented in this dissertation conducted prediction of the time trend and estimation of differences in rates jointly. Alternatively, the LRR model could first predict a reference time trend, and then, estimate differences in the rate of change conditional on the predicted time function. Comparison of these two approaches in regard to making inference on the rate of change would improve understanding of the impact of strategies for incorporate estimation and prediction. The modeling of rates of change can also be useful in developing novel study design for follow-up studies. Biased sampling schemes can be used to improve detection of group differences by identifying subjects with extreme or unusual behavior for a previously measured outcome. By sampling these subjects for additional follow-up, power for detecting group differences will be gained for newly measured covariates. For longitudinal data, outcome dependent sampling schemes have been developed based on the extreme intercept and slope values from a fitted line for each individual [Schildcrout, Garbett, and Heagerty, 2012]. The proposed methodology for modeling rates of change could also be used for a biased sampling scheme for longitudinal data where individuals with extreme rates of change are sampled for follow-up. Sampling subjects based on their individual rate would likely improve power to detect differences in the rate of change for additionally measured covariates.

In conclusion, methodology for comparing rates of change across groups for longitudinal data has produced intriguing developments in statistical research. The presented work represents the first attempt at developing a generalized approach to modeling differences in the rate of change. The direct modeling of rates of change showed improve detection of group differences for non-linear outcomes, can be used to borrow information across multivariate outcomes, and can examine group differences even when the underlying longitudinal trajectory is not well understood. Exploration a regression based approach for structuring rates of change lead to numerous areas that are available for additional methods work. Therefore, continued work on longitudinal methods for modeling rates of change is promising for providing significant contributions to the scientific community.

## BIBLIOGRAPHY

P. Anderson and R. Gill. Cox's regression model for counting processes, a large sample study. *Annals of Statistics*, 10:1100–1120, 1982.

J. Bennet and J. Wakefield. Errors-in-variables in joint population pharmacokinetic/pharmacodynamic modeling. *Biometrics*, 57:803–812, 2001.

L.C. Brumback and M.J. Lindstrom. Self modeling with flexible, random time transformations. *Biometrics*, 60:461–470, 2004.

R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC Press, 2006.

P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross validation. *Numerical Mathematics*, 31:377–403, 1979.

M. Crowder. On consistency and inconsistency of estimating equations. *Econometric Theory*, 2:303–330, 1986.

M. Davidian and D.M. Giltinan. *Nonlinear models for repeated measurement data*. Chapman and Hall, 1995.

P. Diggle, P. Heagerty, K.Y. Liang, and S.L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

J.A. Dubin and H.G. Muller. Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, 100:872–881, 2005.

P.H.C. Eilers and B.D. Marx. Flexible smoothing with b-splines and penalties. *Statistics in Science*, 11:89–121, 1996.

S. Fieuws, G. Verbeke, and G. Molenbergh. Random-effects models for multivariate repeated measures. *Statistical Methods in Medical Research*, 16:387–397, 2007.

G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal Data Analysis*. Chapman and Hall/CRC, 2009.

A.T. Galecki. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods*, 23:3105–3119, 1994.

M.G. Genton. Separable approximations of space-time covariance matrices. *Environmetrics*, 18:681–695, 2007.

P.M. Grambsch and T.M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526, 1994.

S.M. Gray and R. Brookmeyer. Estimating a treatment effect from multidimensional longitudinal data. *Biometrics*, 54:976–988, 1998.

S.M. Gray and R. Brookmeyer. Multidimensional longitudinal data: Estimating a treatment effect from continuous, discrete, or time-to-event response variables. *Journal of the American Statistical Association*, 95:396–406, 2000.

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

P.J. Heagerty and B.F. Kurland. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88:973–985, 2001.

L. Huang. A new mechanistic growth model for simultaneous determination of lag phase duration and exponential growth rate and a new belehdradek-type model for evaluating the effect of temperature on growth rate. *Food Microbiology*, 28:770–776, 2011.

B.J. Jackson, P. Musoke, T. Fleming, L.A. Guay, D. Bagenda, M. Allen, C. Nakabiito, J. Sherman, P. Bakaki, M. Owor, C. Ducar, M. Deseyve, A. Mwatha, L. Emel, C. Duefield, M. Mirochnick, M.G. Fowler, L. Mofenson, P. Miotti, M. Gigliotti, D. Bray, and F. Mmiro.

Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of hiv-1 in kampala, uganda: 18-month follow-up of the hivnet 012 randomised trial. *Lancet*, 362:859–868, 2003.

H. Jacqmin-Gadda, P. Joly, D. Commenges, C. Binquet, and C. Genevieve. Penalized likelihood approach to estimate a smooth mean curve on longitudinal data. *Statistics in Medicine*, 21:2391–2402, 2002.

J. Jia and R.E. Weiss. Common predictor effects for multivariate longitudinal data. *Statistics in Medicine*, 28:1793–1804, 2009.

B.F. Kurland and P.J. Heagerty. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 6:241–258, 2005.

N.M. Laird and J.H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38: 963–974, 1982.

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

J.K. Lindsey, W.D. Byrom, J. Wang, P. Jarvis, and B. Jones. Generalized nonlinear models for pharmacokinetic data. *Biometrics*, 56:81–88, 2000.

C. Proust-Lima, L. Letenneur, and H. Jacqmin-Gadda. A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome. *Statistics in Medicine*, 26:2229–2245, 2007.

J. Roy and X. Lin. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*, 56:1047–1054, 2000.

D. Ruppert, M.P. Wand, and R.J. Carroll. *Semi-parametric Regression*. Cambridge: Cambridge University Press, 2003.

J.S. Schildcrout, S.P. Garbett, and P.J. Heagerty. Outcome vector dependent sampling with longitudinal continuous response data: stratied sampling based on summary statistics. *Biometrics*, 2012. To Appear.

L. Sheppard, R.T. Burnett, A.A. Szpiro, S.Y. Kim, M. Jerrett, C.A. Pope, III, and B. Brunekreef. Confounding and exposure measurement error in air pollution epidemiology. *Air Quality, Atmosphere and Health*, 5:203–216, 2010.

T.G. Travison and R. Brookmeyer. Global effects estimation for multidimensional outcomes. *Statistics in Medicine*, 26:4845–4859, 2007.

G. Wahba. A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline problem. *Annals of Statistics*, 13:1378–1402, 1985.

J. Wakefield and G. Shaddick. Health-exposure modeling and the ecological fallacy. *Biostatistics*, 7:438–455, 2006.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25, 1982.

H.L. Wu, A.A. Ding, and V. De Gruttola. Estimation of hiv dynamic parameters. *Statistics in Medicine*, 17:2463–2485, 1998.

D. Zhang, X. Lin, and M. Sowers. Semiparametric regression for periodic longitudinal hormone data from multiple menstrual cycles. *Biometrics*, 56:31–39, 2000.

# Appendix A

# SCORE AND HESSIAN EQUATIONS

Consider a normal distribution likelihood function for a longitudinal outcome $\mathbf{Y} = (Y_1, \cdots, Y_n)$ measured on an individual observed at times $t_{1i}, \cdots, t_{ni}$. Let $\boldsymbol{\mu}$ denote the mean structure and $\mathbf{V}$ denote the variance structure. Denote the parameters of the model generally as $\eta$, $\nu$ and $\phi$ where $\eta$ represent parameters in the mean structure, $\nu$ represent parameters in the variance structure, and $\phi$ represent parameters in both the mean and variance structure. The likelihood function can then be specified as follows:

$$l(\eta, \nu, \phi) = \text{constant} - \frac{1}{2}\text{log}|\mathbf{V}| - \frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}).$$

Taking the first derivative of the above equation with respect to each parameter provides the following score equations:

$$\begin{aligned}
\dot{l}_\eta &= \left(\frac{\partial \boldsymbol{\mu}}{\partial \eta}\right)^T \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \\
\dot{l}_\nu &= -\frac{1}{2}\text{trace}\left[\mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \nu}\right)\right] + \frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \nu}\right)\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \\
\dot{l}_\phi &= -\frac{1}{2}\text{trace}\left[\mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \phi}\right)\right] + \frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \phi}\right)\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) + \left(\frac{\partial \boldsymbol{\mu}}{\partial \phi}\right)^T \mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}).
\end{aligned}$$

To calculate the hessian matrix, take the negative expected value of the derivatives of the score equations with respect to each parameter. These equations can be simplified by utilizing the properties that $E[\mathbf{Y} - \boldsymbol{\mu}] = 0$ and $E[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T] = V$. Several other properties of traces and expected values are also important in these calculations. Simplifying these equations provides the following hessian equations:

$$
\begin{aligned}
H_{\eta,\eta'} &= \left(\frac{\partial \boldsymbol{\mu}}{\partial \eta}\right)^T \mathbf{V}^{-1} \left(\frac{\partial \boldsymbol{\mu}}{\partial \eta'}\right) \\
H_{\eta,\nu} &= 0 \\
H_{\eta,\phi} &= \left(\frac{\partial \boldsymbol{\mu}}{\partial \eta}\right)^T \mathbf{V}^{-1} \left(\frac{\partial \boldsymbol{\mu}}{\partial \phi}\right) \\
H_{\nu,\nu'} &= \frac{1}{2}\text{trace}\left[\mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \nu}\right)\mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \nu'}\right)\right] \\
H_{\nu,\phi} &= \frac{1}{2}\text{trace}\left[\mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \nu}\right)\mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \phi}\right)\right] \\
H_{\phi,\phi'} &= \frac{1}{2}\text{trace}\left[\mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \phi}\right)\mathbf{V}^{-1}\left(\frac{\partial \mathbf{V}}{\partial \phi'}\right)\right] - \left(\frac{\partial \boldsymbol{\mu}}{\partial \phi}\right)^T \mathbf{V}^{-1} \left(\frac{\partial \boldsymbol{\mu}}{\partial \phi'}\right).
\end{aligned}
$$

To apply these general results to the longitudinal rate regression model proposed in Subsection 2.2.1, we only need to calculate the partial derivatives of the mean structure and variance structure with respect to each parameter specified in the model. We carry out these calculation using the following notation: let $\mathbf{M}$ denote the vector of length $n$ consisting of values for $\mu_0(t)$ evaluated at each time point $t$, $\mathbf{X}$ be a 1 by $q+1$ matrix whose first entry is 1 and the remaining $q$ entries are variable values for a set of variables of interest for their effect on the rate of change of $\mathbf{Y}$, and $\mathbf{Z}$ by a 1 by $r$ matrix of variable values for variables of interest for main effect adjustment possibly overlapping with variables in $\mathbf{X}$. The parameters of the model consist of the vector $\boldsymbol{\beta}$ for parameters of $\mu_{\mathbf{x}}(t)$; the vector $\boldsymbol{\theta}$ of length $q+1$ whose first entry is 1 and remaining entries are coefficients for the rate of change associated with variables in $\mathbf{X}$; the vector $\boldsymbol{\alpha}$ of length $r$ consisting of coefficients for variables in $\mathbf{Z}$; and variance components $\tau_0^2$, $\tau_1^2$, $\rho$, and $\sigma^2$ for the variation of the random intercept, the variation in the random slope, the correlation between the random effects, and the variation in the random error as specified in Equation (2.4). The mean and variance structure can then be expressed as follows:

$$
\begin{aligned}
\boldsymbol{\mu}(\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\beta}) &= \mathbf{1}_n\mathbf{Z}\boldsymbol{\alpha} + \mathbf{M}\mathbf{X}\boldsymbol{\theta} \\
\mathbf{V}(\tau_0^2,\tau_1^2,\rho,\sigma^2,\boldsymbol{\beta}) &= \tau_0^2\mathbf{1}_n\mathbf{1}_n^T + \tau_1^2\mathbf{M}\mathbf{M}^T + \rho\tau_0\tau_1(\mathbf{1}_n\mathbf{M}^T + \mathbf{M}\mathbf{1}_n^T) + \sigma^2\mathbf{I}_n.
\end{aligned}
$$

The partial derivatives of these equations with respect to each parameter can then be expressed as follows:

$$\left(\frac{\partial \boldsymbol{\mu}}{\partial \alpha_k}\right) = \mathbf{1}_n X_k \quad \left(\frac{\partial \boldsymbol{\mu}}{\partial \theta_k}\right) = \mathbf{M} X_k \quad \left(\frac{\partial \boldsymbol{\mu}}{\partial \beta_k}\right) = \left(\frac{\partial \mathbf{M}}{\partial \beta_k}\right) \mathbf{X}^T \boldsymbol{\theta}$$

$$\begin{aligned}
\left(\frac{\partial \mathbf{V}}{\partial \tau_0^2}\right) &= \tau_0^2 \mathbf{1}_n \mathbf{1}_n^T + \frac{\rho\tau_1}{2\tau_0}(\mathbf{1}_n \mathbf{M}^T + \mathbf{M}\mathbf{1}_n^T) \\
\left(\frac{\partial \mathbf{V}}{\partial \tau_1^2}\right) &= \tau_1^2 \mathbf{M}\mathbf{M}^T + \frac{\rho\tau_0}{2\tau_1}(\mathbf{1}_n \mathbf{M}^T + \mathbf{M}\mathbf{1}_n^T) \\
\left(\frac{\partial \mathbf{V}}{\partial \rho}\right) &= \tau_0\tau_1(\mathbf{1}_n \mathbf{M}^T + \mathbf{M}^T\mathbf{1}_n^T) \\
\left(\frac{\partial \mathbf{V}}{\partial \sigma}\right) &= \mathbf{I}_n \\
\left(\frac{\partial \mathbf{V}}{\partial \beta_k}\right) &= \tau_1^2 \left[\left(\frac{\partial \mathbf{M}}{\partial \beta_k}\right)\mathbf{M}^T + \mathbf{M}\left(\frac{\partial \mathbf{M}}{\partial \beta_k}\right)^T\right] + \rho\tau_0\tau_1 \left[\mathbf{1}_n \left(\frac{\partial \mathbf{M}}{\partial \beta_k}\right)^T + \left(\frac{\partial \mathbf{M}}{\partial \beta_k}\right)\mathbf{1}_n^T\right].
\end{aligned}$$

By substituting these partial derivatives into the score and hessian equations, the score vector and hessian matrix for an individual can be calculated. By setting initial parameter values and averaging the score vector and hessian matrix across individuals, the Newton-Raphson iterations can be applied to obtain Maximum Likelihood Estimates for all parameters.

Appendix B

# CROSS-VALIDATION ONE-STEP ESTIMATION

The one-step estimate for the mean parameters for a given subsample and under a given penalty parameter presented in Equation (4.4) was constructed by extending the estimate discussed by Jacqmin-Gadda et al. [2002] for a linear mean model. Given a sample of size $N$, let $f_i(\cdot)$ denote the mean function for an outcome vector $\mathbf{Y}$ evaluated for individual $i$, with $i = 1, \cdots, N$, that is dependent on mean parameters $\boldsymbol{\eta}$. The matrix $V$ denotes the covariance of $\mathbf{Y}$ which we assume to be fixed across all subsamples and for all values of the penalty parameter, $\lambda$, for the purposes of this cross-validation procedure. We use the functions $U_p(\boldsymbol{\eta})$ and $I_p(\boldsymbol{\eta})$ to denote the score equations and information equations for the penalized likelihood for the mean parameters $\boldsymbol{\eta}$ with a fixed covariance matrix. Let $\widehat{\boldsymbol{\eta}}_{-i}(\lambda)$ denote the penalized maximum likelihood estimates (PMLEs) for the subsample with the $i$th individual removed and a given penalty parameter value, $\lambda$. We can express the PMLEs as follows:

$$
\begin{aligned}
\widehat{\boldsymbol{\eta}}_{-i}(\lambda) &= \widehat{\boldsymbol{\eta}}_{-i}(\lambda) + I_p[\widehat{\boldsymbol{\eta}}_{-i}(\lambda)]^{-1} U_p[\widehat{\boldsymbol{\eta}}_{-i}(\lambda)] \\
&= \widehat{\boldsymbol{\eta}}_{-i}(\lambda) + \left[\sum_{j \neq i} \widehat{f}_j^{(1)T} V^{-1} \widehat{f}_j^{(1)} + \lambda\Omega\right]^{-1} \left[\sum_{j \neq i} \widehat{f}_j^{(1)T} V^{-1}(\mathbf{Y}_j - \widehat{f}_j) - \lambda\Omega\widehat{\boldsymbol{\eta}}_{-i}(\lambda)\right] \\
&= \left[\sum_{j \neq i} \widehat{f}_j^{(1)T} V^{-1} \widehat{f}_j^{(1)} + \lambda\Omega\right]^{-1} \left[\sum_{j \neq i} (\widehat{f}_j^{(1)T} V^{-1} \widehat{f}_j^{(1)} + \lambda\Omega)\widehat{\boldsymbol{\eta}}_{-i}(\lambda) + \right. \\
&\qquad \left. \widehat{f}_j^{(1)T} V^{-1}(\mathbf{Y}_j - \widehat{f}_j) - \lambda\Omega\widehat{\boldsymbol{\eta}}_{-i}(\lambda)\right] \\
&= \left[\sum_{j \neq i} \widehat{f}_j^{(1)T} V^{-1} \widehat{f}_j^{(1)} + \lambda\Omega\right]^{-1} \left[\sum_{j \neq i} \widehat{f}_j^{(1)T} V^{-1} \mathbf{Y}_j + \widehat{f}_j^{(1)T} V^{-1}(\widehat{f}_j^{(1)} \widehat{\boldsymbol{\eta}}_{-i}(\lambda) - \widehat{f}_j)\right]
\end{aligned}
$$

where $f_j^{(1)}$ is the matrix of derivatives of $f_j(\cdot)$ with respect to the vector $\boldsymbol{\eta}$. The first line of the above expression utilizes the fact that the penalized score equation will be zero when evaluated at the PMLEs. The remainder of the calculation is a result of algebraic manipulation. When the mean structure is linear in $\boldsymbol{\eta}$, then $\widehat{f}_j^{(1)}\widehat{\boldsymbol{\eta}}_{-i}(\lambda) = \widehat{f}_j$ and the second term in the last line is eliminated which results in the expression presented in Jacqmin-Gadda et al. [2002]. However, the second term will generally not be eliminated for a non-linear mean structure as is the case for the semi-parametric LRR model. By substituting a reference estimate into the last line, the above result can be used as a one-step estimate for approximating the PMLEs for each subsample and $\lambda$ value.

Appendix C

## DISTRIBUTIONAL PROPERTIES OF THE REFERENCE TIME FUNCTION WHEN USING PENALIZED SPLINE ESTIMATION

We can re-express the form of the reference time function given in Equation (4.1) as a product of matrices. Let $T_B$ denote the vector with entry values corresponding to the spline basis $B_1(t), \cdots, B_p(t)$ evaluated at a given time and $\boldsymbol{\beta}$ be the vector of length $p$ consisting of coefficients for each spline function. Then,

$$\mu_0(t) = \boldsymbol{\beta}^T T_B.$$

Given a penalized maximum likelihood estimate (PMLE) for the coefficients denoted by $\widehat{\boldsymbol{\beta}}$, a $(1-\alpha) \times 100\%$ confidence interval for $\mu_0(\cdot)$ across all values of $t$ can be constructed based on the expression

$$\widehat{\mu}_0(t) \pm z_{1-\alpha/2}\sqrt{T_B^T \text{var}(\widehat{\boldsymbol{\beta}}) T_B} \tag{C.1}$$

where $z_a$ denotes the quantile corresponding to probability $a$ from either a normal distribution or a student-t distribution depending on sample characteristics. In order to ensure that Expression (C.1) provides a valid confidence interval, the distributional properties of the PMLE need to be considered.

Let $\boldsymbol{\eta}$ be the vector of parameters for semi-parametric LRR model, and let $\widehat{\boldsymbol{\eta}}$ be the PMLE for these parameters. To describe the distributional behavior of $\widehat{\boldsymbol{\eta}}$ as an estimate of $\boldsymbol{\eta}$, consider that PMLE is defined by the solution to the penalized likelihood score equation i.e.

$$U(\widehat{\boldsymbol{\eta}}) - \lambda \Omega \widehat{\boldsymbol{\eta}} = 0 \tag{C.2}$$

where $U(\cdot)$ is the score equation for the standard, unpenalized likelihood equation. We can relate the PMLE to the true parameter vector by considering the first-order Taylor Series approximation of the score equation based on the definition in Equation (C.2). That is, it follows from Equation (C.2) that

$$U(\boldsymbol{\eta}) + U^{(1)}(\boldsymbol{\eta})(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}) - \lambda\Omega\widehat{\boldsymbol{\eta}} = 0. \tag{C.3}$$

If the hessian matrix, $H$, for the standard likelihood equation is substituted for the derivative of the score equation where $H(\boldsymbol{\eta}) = -U^{(1)}(\boldsymbol{\eta})$, we can derive from Equation (C.3) the result

$$[H(\boldsymbol{\eta}) + \lambda\Omega]\{\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta} + [H(\boldsymbol{\eta}) + \lambda\Omega]^{-1}\lambda\Omega\boldsymbol{\eta}\} = U(\boldsymbol{\eta}). \tag{C.4}$$

Standard distribution theory for score equations tells us that $U(\boldsymbol{\eta}) \sim N[\mathbf{0}, H(\boldsymbol{\eta})]$. Thus, based on Equation (C.4), the PMLE will be normally distributed with mean and variance given by

$$\begin{aligned} E(\widehat{\boldsymbol{\eta}}) &= \boldsymbol{\eta} - [H(\boldsymbol{\eta}) + \lambda\Omega]^{-1}\lambda\Omega\boldsymbol{\eta}, \text{ and} \\ \text{var}(\widehat{\boldsymbol{\eta}}) &= [H(\boldsymbol{\eta}) + \lambda\Omega]^{-1}H(\boldsymbol{\eta})[H(\boldsymbol{\eta}) + \lambda\Omega]^{-1}. \end{aligned}$$

The distributional results for the PMLE show that, for a fixed sample size, $\widehat{\boldsymbol{\eta}}$ will be biased estimate for $\boldsymbol{\eta}$ with bias given by $\text{Bias}(\widehat{\boldsymbol{\eta}}) = -[H(\boldsymbol{\eta}) + \lambda\Omega]^{-1}\lambda\Omega\boldsymbol{\eta}$. A biased estimate was anticipated for the PMLE since the penalization of the parameters forces deviations from the unbiased maximum likelihood estimate. Since $\Omega$ has zero entries values for all entries corresponding to parameters not in the penalized spline equation, this bias is non-zero only for the estimated reference time function. However, the PMLE is a consistent estimate since $\lambda$ and $\Omega$ are fixed and $H(\boldsymbol{\eta})$ increases relative to increased sample size. Nonetheless, for a fixed sample size, the bias in estimating the reference time function must be corrected for when constructing a confidence interval. Therefore, we update Expression (C.1) to construct an asymptotically valid confidence interval based on the expression

$$(\widehat{\boldsymbol{\beta}} + \text{Bias}(\widehat{\boldsymbol{\beta}}))^T T_B \pm z_{1-\alpha/2}\sqrt{T_B^T \text{var}(\widehat{\boldsymbol{\beta}}) T_B}.$$

The bias and the variance of the PMLE estimates can be approximated by substitute the PMLE estimates for the true parameter values in the respective bias and variance equations given above.