A Comparison of Bayesian and Likelihood Joint Models for Longitudinal and Survival Data: An Application to Oral Cancer Study

Jintong Yan

April 3, 2023

Abstract

Oral cancer comprises 2%–4% of all cancer cases worldwide. To investigate the relationship between potential biomarkers and the development of oral malignant transformation, joint modeling of longitudinal and survival analysis is desirable. This report focuses on the application and comparison of joint models based on likelihood and Bayesian methods. A brief review of the models used for longitudinal and survival data analysis and joint models is provided at the beginning of the report. The application of joint models based on oral squamous cell carcinoma (OSCC) data is then discussed. In this dataset, oral lesion area is measured repeatedly for each subject, as well as the occurrence of malignant transformation. Model diagnosis is conducted using Q-Q plots and residual plots for joint models. For joint models based on Bayesian methods, sensitivity analyses, trace plots, and other common diagnostic diagrams are considered.

Keywords: joint models, longitudinal data, survival data.

Contents

List of Figures

List of Tables

1 Introduction

In biostatistics and medical research, it is very common to record observations of important variables repeatedly over time for each subject, along with dichotomous indicators to mark the time at which an event of interest occurs. The data measured repeatedly over time for each individual are called longitudinal data, and the times to an event of interest are called survival data.

1.1 Motivating Dataset

Oral cancer accounts for 2%–4% of all cancer cases worldwide. In some countries, the prevalence of oral cancer is higher, such as in Pakistan and India [1]. During 2022 in the United States, approximately 54,000 new cases of oral cavity or oropharyngeal cancer were diagnosed, while approximately 11,230 affected individuals died from these cancers [2]. Oral squamous cell carcinoma (OSCC) is the most common malignant epithelial neoplasm affecting the oral cavity, which is believed to arise through sequential stages of potentially malignant lesions (OPMLs), like hyperplasia, mild, moderate, severe dysplasia, and carcinoma in situ [3]. Currently, from a clinical point of view, there is no effective biomarker or diagnostic tool to guide triage or treatments. Hence, potential biomarkers such as size, appearance, and site are important to determine the cancer risk of OPMLs.

Among 41 patients with mild and moderate oral dysplasia lesions at the beginning of the study, 28 were non-progressors, and 13 progressed to severe dysplasia, carcinoma-in-situ, or squamous cell carcinoma until the end of the study. All of the variables were measured prior to the diagnosis result, including patient demographics, risk habit history, lesion clinical features, and DNA-ICM findings. Descriptive statistics of these variables are shown in Table 1.1.

Some research has identified that the larger the lesion area, the higher the risk of progression. The trajectories of oral lesion area for all individuals are shown in Figure 1.1. The trend of lesion area between non-progressors (individuals who did not experience cancer) and progressors (individuals who experienced cancer) is different, indicating that the trend of changes in oral lesion area could be a predictor for the occurrence of oral cancer.

This report mainly explores the factors that may correlate the occurrence of oral cancer based on the OSCC datasets.

1.2 Joint Modeling of Longitudinal and Survival Data

Common applications of joint models are longitudinal models with informative dropouts, survival models with measurement errors in time-dependent covariates, and longitudinal and survival processes influenced by common latent factors. Based on different cases and medical backgrounds, different association structures may be considered. Three popular association structures, the “shared parameters” association, the “current value”

Table 1.1: Variable descriptions for the dataset

Name	**Type**	**Description**
ID	Factor	Patient ID number, ranging from 1 to 41
Bxdiagnosis	Categorical	OPML Stages: H = hyperkeratosis; D1 = mild dysplasia; D2 = moderate dysplsia; D3 = severe dysplasia; SCC = squamous cell carcinoma; VC = verruous carcinoma. (Note that D3, SSC, and VC are viewed as progressors, while others are viewed as non-progressors.)
Months	Numerical	Measurement times (in month) from the beginning of the study
Age	Numerical	Age of the patients at the beginning of the study
Smoke	Factor	Smoking habits: 1 = habitual smoker; 0 = non-habitual smoker
Alcohol	Factor	Drinking habits: 1 = habitual drinker; 0 = non-habitual drinker
Obstime	Numerical	Measurement times (in year) from the beginning of the study
Lesion_site	Factor	Location of lesion inside oral cavity: 1= low-risk and 2 = high-risk
Lesion_area	Numerical	The area of lesion in the mouth(mm^2)
Aneuploid	Numerical	Number of aneuploid cells
Cycling %	Numerical	Percentage of cycling cells
Tetraploid %	Numerical	Percentage of tetraploid cells
Proliferation %	Numerical	Percentage of cancer cell proliferation
Type	Character	Diagnostic results coded as progressors and non-progressors
Progression	Factor	Diagnostic results: 1 = progressors and 2 = non-progressors

association, and the "current value plus slope" association, are commonly used. Furthermore, two popular inference methods can be used for estimating parameters: likelihood methods and Bayesian methods. For this specific dataset we analysed, with such a small sample size, it would be better to use Bayesian methods in order to incorporate prior information. Estimating parameters in joint models under these two methods is often challenging. This report introduces some computational methods corresponding to these two methods, such as Monte Carlo EM (MCEM) algorithm [4] for likelihood methods and Markov Chain Monte Carlo (MCMC) [5] for Bayesian methods.

In this report, we review how likelihood and Bayesian methods are used to estimate parameters. Among the five main joint models, two are based on likelihood methods with

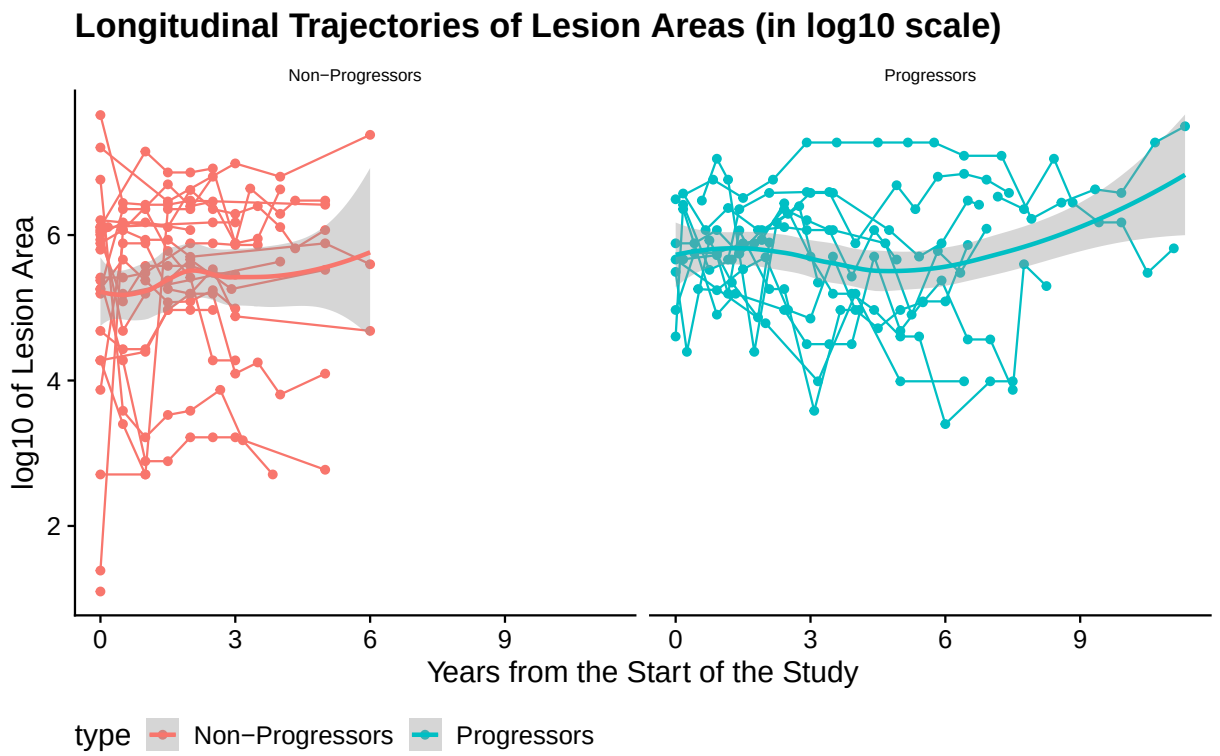

Figure 1.1: Longitudinal Trajectories of Lesion Areas (in log10 scale)

"current value" association and "current value plus slope" association. The other three are based on Bayesian methods with "shared parameters" association, "current value" association, and "current value plus slope" association. Software for joint modeling of longitudinal and survival data has been available for several years. We focus on the package *JM* for likelihood methods and *JMBayes* for Bayesian methods in R.

1.3 Outline

This report is organized as follows. In **Chapter 2**, we review common models and methods for longitudinal and survival analysis. In **Chapter 3**, we describe joint modelings of longitudinal and survival data, and different inference methods. In **Chapter 4**, we analyze the oral cancer data in detail. In this section, five joint models are built with two inference methods and three association structures. After that, the parameters based on likelihood and Bayesian methods are compared to detect the difference. **Chapter 5** concludes this report with some discussions.

2 Review of Models for Longitudinal Data and Survival Data

2.1 Models for Longitudinal Data

2.1.1 Introduction

Longitudinal studies are very common in various fields. In such studies, researchers collect a series of observations from each subject over a period of time. In clinical trials, researchers prefer to follow risk factors or health outcomes over time with continuous or repeated monitoring. For example: To study the incidence and risk factors of heterosexually transmitted HIV infection, SARACCO and the partners [6] followed a cohort of 343 seronegative women, stable, monogamous partners of infected men whose only risk of acquiring HIV was sexual exposure to the infected partner. In this study, seroconversion frequency, $CD4+$ cell number, p24 antigen, and other factors were measured many times on each woman.

In linear regression, all variables for all individuals are measured once. Hence, we can assume that they are independent without autocorrelation. But for longitudinal data, the repeated measurements of a variable on the same individual are likely to be correlated. Ignoring the correlation may lead to inefficient or biased results.

2.1.2 Linear Mixed-Effects Models

A linear mixed-effects(LME) model is an extension of the simple linear regression model. Specifically, a linear regression model works for independent data; a linear mixed-effects model, however, works for non-independent data. Consider a longitudinal data, let y_{ij} be the response variable for individual i at time t_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n_i$. The following shows a simple linear regression model for longitudinal data,

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + e_{ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, n_i,$$

where β_0 is the intercept, β_1 is the slope, and e_{ij} is a random error. But using a linear regression model to predict y_{ij} may cause bias as y_{ij} is a response variable measured multiple times on the same individual i. We need to incorporate the within-individual correlation. If we assume the intercept β_0 varies in different individuals, we may introduce a random effect b_{0i} for β_0. Then the intercept is individual-specific. A LME model can be written as

$$\begin{aligned} y_{ij} &= (\beta_0 + b_{0i}) + \beta_1 t_{ij} + e_{ij} \\ &= \beta_{0i} + \beta_1 t_{ij} + e_{ij} \\ &= (\beta_0 + \beta_1 t_{ij}) + b_{0i} + e_{ij} \\ &= \boldsymbol{T_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{e_i}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, n_i, \end{aligned}$$

where $\boldsymbol{T_i}$ and $\boldsymbol{Z_i}$ are design vectors usually contain covariates, and $\boldsymbol{e_i}$ is the random error matrix of the repeated measurements. In this formula,

$$\boldsymbol{T_i} = \begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix}, \quad \boldsymbol{Z_i} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \boldsymbol{e_i} = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{bmatrix}.$$

If we assume both the intercept β_0 and the slope β_1 varies for different individuals, we can introduce random effects b_{0i} and b_{1i} for β_0 and β_1, respectively.

$$\begin{aligned} y_{ij} &= (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + e_{ij} \\ &= \beta_{0i} + \beta_{1i}t_{ij} + e_{ij} \\ &= (\beta_0 + \beta_1 t_{ij}) + (b_{0i} + b_{1i}t_{ij}) + e_{ij} \\ &= \boldsymbol{T_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{e_i}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, n_i, \end{aligned}$$

where $\boldsymbol{T_i}$ and $\boldsymbol{Z_i}$ are design vectors usually containing covariates, and $\boldsymbol{e_i}$ is the random error matrix of the repeated measurements. In this formula,

$$\boldsymbol{T_i} = \begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix}, \quad \boldsymbol{Z_i} = \begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{bmatrix}, \quad \boldsymbol{e_i} = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{bmatrix}.$$

Generally, LME models can be expressed as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$ is the response variable with n_i measurements, $\boldsymbol{\beta} = (1, \beta_1, \ldots, \beta_p)^T$ is the fixed effects vector, $\boldsymbol{b}_i = (1, b_1, \ldots, b_q)^T$ is the random effects vector, $\boldsymbol{X_i}$ and $\boldsymbol{Z}_i$ are design vectors often containing covariates, and $\boldsymbol{e}_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})^T$ is the random error matrix for specific-individual measurements. And also

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & x_{11} & \ldots & x_{pn_i} \\ 1 & x_{12} & \ldots & x_{pn_i} \\ . & . & \ldots & . \\ 1 & x_{1n_i} & \ldots & x_{pn_i} \end{bmatrix}, \quad \boldsymbol{Z}_i = \begin{bmatrix} 1 & x_{11} & \ldots & x_{qn_i} \\ 1 & x_{12} & \ldots & x_{qn_i} \\ . & . & \ldots & . \\ 1 & x_{1n_i} & \ldots & x_{qn_i} \end{bmatrix},$$

where we assume that $\boldsymbol{b}_i \sim N(0, D)$, $\boldsymbol{e}_i \sim N(0, R_i)$, and $\boldsymbol{b}_i$ and $\boldsymbol{e}_i$ are independent. The covariance of the random effects D is a $(q+1) \times (q+1)$ matrix. The variance(the diagonal elements of D) of the $\boldsymbol{b}_i$ explains the variability of the longitudinal measurements between individuals that are unexplained by covariates. While this covariance matrix D is unstructured, it can sometimes be structured as a diagonal matrix. The covariance of random errors for within-individuals R_i is a $n_i \times n_i$ matrix. The variance (the diagonal elements of R_i) of $\boldsymbol{e}_i$ explains the variability of the repeated measurements within each individual. In practice, we often assume $R_i = \sigma^2 I$, which means that, given the random effects, the within-individual measurements are independent, and the variance (the diagonal elements of R_i) is constant.

Generally, there are four assumptions for a linear mixed-effects model [7]. The explanatory variables are related linearly to the response. The errors have constant variance. The errors are independent. And the random effects and the errors are Normally distributed.

Missing values in longitudinal data are common in practice. For instance, the repeated observations y_{i1}, y_{i2},..., y_{in_i} for different individuals can be measured at different time points. And the number of observations are often different among all the individuals. LME models allow longitudinal data with missing data in the responses, assuming missing data are missing at random.

Maximum likelihood methods(MLE) or restricted maximum likelihood methods are popular for estimating unknown parameters in LME models. Iterative algorithms, like EM algorithms or Newton-Raphson methods, are used to compute maximum likelihood estimates.

2.1.3 Other Types of Mixed-Effects Models

Nonlinear Mixed-Effects (NLME) Models

Nonlinear mixed-effects (NLME) models are extended from linear mixed-effects (LME) models. Compared with LME models, NLME models tend to be more complex to fit the observed data. They are capable of providing good predictions, however, for unobserved data. Generally, NLME models can be derived into two steps. In the first step, we model the mean and covariance structure of a given individual,

$$y_{ij} = g(t_{ij}, \boldsymbol{\beta_i}) + e_{ij}$$

where y_{ij} is the response variable of ith individual at time j, $g(\cdot)$ is a known nonlinear function, $\boldsymbol{\beta_i}$ is an individual-specific parameters, and e_{ij} is the random error within each individual.

In the second step, we model the between-individual variation through random effects, which means to specify the individual-specific parameters,

$$\boldsymbol{\beta_i} = h(x_i, \boldsymbol{\beta}, \boldsymbol{b_i}), \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i,$$

where $h(\cdot)$ is often a linear function, x_i is the covariate for each individual i, $\boldsymbol{\beta}$ is the the vector of fixed effects, $\boldsymbol{b_i}$ is the vector of random effects. The assumptions of NLME models are the same as LME models", $b_i \sim N(0, D)$, $e_i \sim N(0, R_i)$ and b_i and e_i are independent.

Statistical inference for NLME models is commonly based on likelihood methods. However, compared to LME models, marginal distributions and likelihood functions of the response variable in NLME models are not expressed analytically, which leads to computational challenges. Several methods have been introduced to overcome these challenges, including numerical or Monte Carlo integration methods, Expectation-Maximization (EM)

algorithms, and approximate methods.

Generalized Linear Mixed Models (GLMMs)

Both LME models and NLME models mentioned above are used when the response is continuous and approximately normally distributed. However, if the distribution of the response is from an exponential family, generalized linear mixed models (GLMMs) can be considered. The principle is similar to NLMEs, where random effects are introduced to account for the correlation among the repeated observations under each individual and the variation between individuals. Let $y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})$ correspond to the n_i repeated measurements for individual i. And y_{i1}, y_{i2}, $\ldots$, y_{in_i} are independent given random effects. GLMMs can be written as

$$g(\mu_{ij}) = \boldsymbol{x_{ij}}\boldsymbol{\beta} + \boldsymbol{z_{ij}^T}\boldsymbol{b_i},$$

where $\mu_{ij} = E(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b_i})$, $\boldsymbol{x_{ij}}$ and $\boldsymbol{z_{ij}}$ are design vectors, $\boldsymbol{\beta}$ is the vector contains fixed effects, and $\boldsymbol{b_i}$ contains random effects which we assume $\boldsymbol{b_i} \sim N(0, D)$.

Statistical inference for a GLMM also usually uses likelihood methods. However, similar to NLME models, inference for GLMMs can also be challenging due to random effects being nonlinear in the models. Numerical methods or Monte Carlo EM algorithms are often used for inference.

2.2 Models for Survival Data

2.2.1 Introduction

Survival analysis, also called time-to-event analysis, is a commonly used term for analysis of outcomes that are times to an event. Clinical fields prefer to use it to study the time to death, the onset of a disease, or the relapse of a condition, but the event can be anything. For example, the time until recidivism. In criminology, one main application of the survival model has been to analyze the time until recidivism. One example is to predict the time until recidivism for a sample of North Carolina prison releases [8]. The event is recidivism, and the time is the interval between the time of prison releases and the time of recidivism. In this example, the response variable of this sample is the time to an event.

In survival analysis, the response variable is the time until an event occurs. However, not all individuals experience the event before they are lost to follow-up or the study ends. In these cases, if individuals were to eventually experience the event, their survival times would be greater than the times recorded at the last observation. In other words, their true event time is unknown, but it is known to be greater than the recorded study time. These event times are referred to as censored. One goal of survival analysis is to properly handle censored data. One common type of censoring is **right-censoring**, which denotes that the event has not yet occurred as of time t, and may never occur.

Survival analysis methods discussed in this report assume that censoring is right-censored and non-informative, which means each subject has a censoring time that is statistically

independent of their failure time.

2.2.2 Survival and Hazard Functions

Survival function

In probability theory and statistics, we describe the distribution of any random variable X by its cumulative distribution function (CDF). The function given by $F(x) = P(X \leq x)$, which represents the probability that the random variable X takes on a value less than or equal to x.

If the variable is a random event time T, the CDF is $F(T) = P(T \leq t)$, which represents the probability of the event occurring prior to or at time t. However, in survival analysis, what we are interested in is the survival function $S(t)$, which is the probability that an individual has survived past t, instead of before time t. This turns out to be a cumulative probability after t, which can be written as

$$S(t) = 1 - F(t) = P(T > t).$$

Hazard function

In survival analysis, a risk of an event is the probability that an event may occur within a given time period, while a rate of an event is the risk per unit time. Hazard is a concept related to these two measures and specifically refers to the instantaneous risk of an event occurring at a specific time, t, among those who have not yet experienced the event. Mathematically, a hazard at time t, $h(t)$, can be defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}.$$

This function means given that a person has not yet experienced an event as of time t, the probability of the event occurring in a time period after t per unit time. With this function, we can explore the individual's hazard of experiencing the event, the change of hazard rates over time, and how the hazard rate differs between groups or depends on other risk factors.

A hazard function is related to the rate of change of a survival function S(t). When a survival function is flat, the hazard function is close to 0. This means that the risk of experiencing the event is very small, as the probability of survival is not changing much. Conversely, when a survival function is decreasing, the hazard function increases. This indicates that the risk of experiencing the event is larger as the probability of survival is dropping. Mathematically, the relationship can be written as

$$h(t) = \frac{dlogS(t)}{dt},$$

$$S(t) = exp[-\int_0^t h(s)ds].$$

2.2.3 Overview of Common Models for Survival Data

Four survival models, Kaplan-Meier methods, Exponential models, Weibull models, and Cox Proportional-Hazards models are commonly used. This section gives a brief review of the first three models, and more details about Cox Proportional-Hazards models are shown in the next section.

A Kaplan-Meier method is classified as a non-parametric method, which makes no assumptions about the distribution of a data, and it is used to estimate survival functions $S(t)$. It is intuitive and easy to interpret.

An Exponential model is a parametric model, which assumes the survival data follow an exponential distribution. An Exponential model assumes that the hazard rate is constant, which means the risk of an event occurring remains the same throughout the period of observation. However, the strong assumption that the hazard rate is constant at any given time may not be consistent to all data of interest.

Another parametric model is the Weibull model. Unlike Exponential models which assume that the hazard rate is constant, Weibull models assume the change in hazard rate is linear. The hazard rate can always increase, always decrease, or always stay the same. But it also depends on a strong assumption that the hazard rate changes linearly across time.

2.2.4 Cox Proportional Hazards Models

Cox regression refers to a semi-parametric method that was introduced by D. R. Cox in 1972 [9]. It is called semi-parametric because no assumption is made about the distribution of event times, but the hazard function depends on a set of parameters (regression coefficients) that explain the association between the hazard and a set of predictors. Cox models can be used in survival analysis to examine survival data with respect to multiple variables simultaneously, and can be written as

$$h(t) = h_0(t)e^{\beta_1 X_1 + \cdots + \beta_K X_K},$$

where $h_0(t)$ is the baseline hazard function which represents the hazard for individuals whose covariate values are all 0 or at their reference level, similar to the intercept in a linear regression model. There is no intercept in Cox regression model. Vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ is multiplied by predictors. e^{β}(coefficients) represents a hazard ratio(HR) comparing the hazard of experiencing the event at time t between individuals with $X = x + 1$ versus those with $X = x$ if X is a continuous predictor, or between individuals with a specific level of X and its reference level if X is a categorical predictor), holding all other variables fixed. For example, consider the hazards at $X_1 = x + 1$ and $X_1 = x$ with other predictors fixed:

$$h(t|X_1 = x_1 + 1, \ldots, X_K = x_K) = h_0(t)e^{\beta_1(x_1+1) + \cdots + \beta_K x_K},$$

$$h(t|X_1 = x_1, \ldots, X_K = x_K) = h_0(t)e^{\beta_1 x_1 + \cdots + \beta_K x_K}.$$

Taking the ratio of these two, everything cancels out except e^{β_1} which is the HR for X_1. Importantly, HR does not depend on time. The assumption of proportional hazards defines the hazard functions for any two individuals that have a constant proportion over time. For the explanation of the hazard ratio, we can learn from the meaning of the odds ratio in logistic regression. For example, $HR = 1.20$ implies that one group has 20% greater hazard than another.

3 Joint Modeling of Longitudinal and Survival Data

3.1 Introduction

Longitudinal and survival data are common in many studies. For example, researchers may be interested in the time it takes for patients to recover from an illness and collect their medical information, such as CD4 cell counts, multiple times over a period of time. In such cases, it is often important to model both longitudinal data and survival data simultaneously, such as survival models with error-prone time-dependent covariates. Inferences based on separate models may be biased and inefficient when these two types of data have a strong association.

Joint models can be used for simultaneous inference to produce valid and unbiased results. There has been extensive research in joint models in the past few decades (e.g., Faucett and Thomas [10], Wulfsohn and Tsiatis [11]). The survival model and the longitudinal model can be associated in different ways, such as "current value" association, "current value plus slope" association, and "shared parameters" association. Further details about these association structures will be explained in **Section 3.2.3**.

In this chapter, we provide an overview of basic ideas behind joint modeling for longitudinal and survival data. In **Section 3.2**, we illustrate statistical frameworks of joint models. Then, in **Section 3.3** and **Section 3.4**, we provide more details on likelihood methods and Bayesian methods, respectively, for estimating parameters in joint models.

3.2 Statistical Framework of Joint Models

In this section, we focus on survival models and consider longitudinal models as secondary. This situation often arises when time-dependent covariates in survival models have measurement errors or missing data. Therefore, secondary longitudinal models should be used to address these measurement errors and missing data.

For individual $i(i = 1, 2, \ldots, N)$, let s_i be the event time, and c_i the censoring time. Since we assume the censors are all randomly right-censored, the observed time t_i can be written as $t_i = min\{s_i, c_i\}$. The censoring indicator δ_i can be defined as $\delta_i = I(s_i \leq c_i)$ such that $\delta_i = 0$ when the survival time for individual i is right censored and $\delta_i = 1$ otherwise.

3.2.1 Survival and Longitudinal Submodels

A Cox PH model is used for analyzing survival data, which can be expressed as

$$h_i(t) = h_0(t)exp(\boldsymbol{\gamma}^T \boldsymbol{w_i}), \quad i = 1, 2, \ldots, n, \tag{2}$$

where $h_0(t)$ is the baseline hazard function at time t, $\boldsymbol{w_i}$ is a vector of covariates, possibly time-varying, and $\boldsymbol{\gamma}$ is the corresponding regression parameters.

In the aforementioned survival model, let y be a time-varying covariate measured with error at several time points. To address this measurement error, we model the y-data using a linear mixed-effects model. In the LME model, the response variable $y_i(t) = (y_{i1}(t), y_{i2}(t), \ldots, y_{in_i}(t))^T$ denotes the observed measurements for individual i at time $t(t = 1, 2, \ldots, T_i^*)$. The unobserved variable $\mu_i(t) = (\mu_{i1}(t), \mu_{i2}(t), \ldots, \mu_{in_i}(t))^T$ contains the repeated measurements for each individual i over time without measurement errors. We assume that the hazard $h(t)$ in the survival model depends on the true covariate $\mu(t)$ rather than the observed but mis-measured value $y(t)$, and the relationship between $y_i(t)$ and $\mu_i(t)$ can be written as

$$y_i(t) = \mu_i(t) + e_i, \qquad i = 1, 2, \ldots, n. \tag{3}$$

Linear mixed-effects models of the measurements $y_i(t)$ can be written as

$$y_i(t) = \mu_i(t) + e_i = X_i(t)\boldsymbol{\beta} + Z_i(t)\boldsymbol{b_i} + e_i, \quad i = 1, 2, \ldots, n, \tag{4}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects and $\boldsymbol{b_i}$ is the vector of random effects. The design vectors $X_i(t)$ and $Z_i(t)$ are known design matrices which often contain covariates.

3.2.2 Joint Models

Then, we focus on joint models to associate the linear mixed-effects model and the Cox model together, which can be written as

$$h_i(t) = h_0(t)exp(\boldsymbol{\gamma}^T\boldsymbol{w_i} + f\{\mu_i(t), \boldsymbol{b}_i, \alpha\}), \tag{5}$$

where $\boldsymbol{w_i}$ is a vector of covariates with corresponding regression coefficients $\boldsymbol{\gamma}$, $\boldsymbol{b}_i$ is the vector of random effects for individual i defined in Equation 5, and α is a parameter quantifies the association between features of the longitudinal data up to time t and the hazard of an event at time t. Various options for the $f(\cdot)$ function may lead to different forms of joint models. The following three association functions are frequently used in joint modeling.

The "Current Value" Association

The "current value" association suggests that the true value, $\mu_i(t)$, of the longitudinal data at a given time t, is a predictor of the hazard risk, $h_i(t)$, at that same time. A joint model under this structure is written as

$$h_i(t) = h_0(t)exp(\boldsymbol{\gamma}^T\boldsymbol{w_i} + \alpha_1\boldsymbol{\mu_i(t)}), \tag{6}$$

where α_1 measures the strength of the association between time-dependent longitudinal variables μ_i at time t and the hazard at the same time. For individual i at time t, if the current value $\mu_i(t)$ of the longitudinal part increases by one unit, the risk of this event occurring will increase by $exp(\alpha_1)$. It is important to note that α_1 does not change across all individuals.

The "Current Value Plus Slope" Association

The "current value plus slope" association extends the "current value" association by adding the slope(the rate of the change) of the measurements in longitudinal data at time t. This structure incorporates the increasing or decreasing changes of the longitudinal trajectories. In mathematical way, it can be written as

$$h_i(t) = h_0(t)exp(\boldsymbol{\gamma}^T \boldsymbol{w_i} + \alpha_1 \boldsymbol{\mu_i(t)} + \alpha_2 \boldsymbol{\mu_i'(t)}), \tag{7}$$

where α_1 expresses the same as in the "current value" association. The slope

$$\mu_i'(t) = \frac{d\mu_i(t)}{dt},$$

means the rate of change of the measurements at time t, and α_2 corresponds to the association between μ_i' at time t and the hazard at the same time. For individual i at time t with the same level of $\mu_i(t)$, the slope value $\mu_i'(t)$ increase one-unit lead to a $exp(\alpha_2)$-fold hazard ratio increase in the survival submodel.

The "Shared Parameters" Association

The "shared parameters" association also names the "shared random-effects" association. In this structure, we only use random effects from longitudinal submodel as linear predictors of the hazard function. This can be written as

$$h_i(t) = h_0(t)exp(\boldsymbol{\gamma}^T \boldsymbol{w_i} + \boldsymbol{\alpha}^T \boldsymbol{b}_i), \tag{8}$$

$$\boldsymbol{\alpha}^T \boldsymbol{b}_i = \alpha_1 b_{i1} + \alpha_2 b_{i2} + \cdots + \alpha_m b_{im},$$

where $\boldsymbol{b}_i = (b_{i1}, b_{i2}, \ldots, b_{im})^T$ is from the longitudinal submodel represents the random effects for individual i, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T$ expresses the association between the random effects from longitudinal trajectory and the survival process at the same time. The random effects from the longitudinal LME model are viewed as "covariates" in the survival model, since these random effects represent individual characteristics of the longitudinal process.

3.3 Statistical Inference

3.3.1 Likelihood Methods

Likelihood methods estimate all parameters in joint models of the longitudinal and survival data simultaneously. Let vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, D, \alpha)$ denote all the unknown parameters in the longitudinal and survival submodels. The likelihood for the joint model Equation 5 can be written as

$$
\begin{aligned}
\text{Ł}(\boldsymbol{\theta}) &= \prod_{i=1}^{N}\left[\int f(t_i,\delta_i|\boldsymbol{b}_i,\boldsymbol{\theta},\alpha)f(\boldsymbol{y}_i|\boldsymbol{b}_i,\boldsymbol{\beta},\boldsymbol{\sigma})f(\boldsymbol{b}_i|D)d\boldsymbol{b}_i\right] \\
&= \prod_{i=1}^{N}\int\Big\{\Big[h_0(t_i)\exp\{\boldsymbol{\gamma}^T\boldsymbol{w_i}+f\{\mu_i(t),\boldsymbol{b}_i,\alpha\}\}\Big]^{\delta_i} \\
&\qquad\qquad \times\exp\Big[-\int_0^{t_i}\Big\{h_0(u)\exp\{\boldsymbol{\gamma}^T\boldsymbol{w_i}+f\{\mu_i(t),\boldsymbol{b}_i,\alpha\}\Big\}du\Big] \\
&\qquad \times f(y_i|\boldsymbol{b}_i,\boldsymbol{\beta},\alpha)f(\boldsymbol{b}_i|D)\Big\}d\boldsymbol{b}_i.
\end{aligned}
\tag{9}
$$

As indicated in Equation 9, joint likelihoods often have complex forms due to the presence of unobserved random effects, censoring, and semi-parametric models. Monte Carlo EM (MCEM) algorithms can be used to do the estimation. It is a variant of EM algorithms, where the expectation in the E-step is calculated through Monte Carlo simulations.

EM Algorithm

An EM (expectation-maximization) algorithm is an approach for finding (local) maximum likelihood estimates of parameters in statistical models that contain latent variables, which are hidden or unobserved variables. The EM algorithm replaces such a single difficult optimization problem by a sequence of easy optimization steps: the E-step (expectation step) and the M-step (maximization step). These two steps alternate and iterate to finally find the local or global maximum likelihood. In the E-step, a function is created for the expectation of the log-likelihood evaluated using the current estimate for the parameters. In the M-step, parameters are computed to maximize the expected log-likelihood found in the E-step. These parameter estimates are then used to determine the distribution of the latent variables in the next E-step.

Specifically, let $\mathbf{Y}$ denote incomplete data (real data), $\mathbf{X}$ denote augmented data (data we constructed), $(\mathbf{Y},\mathbf{X})$ denote complete data, and $\boldsymbol{\theta}$ denote all the unknown parameters we want to estimate. The E-step contains several following specific steps.
The first is the incomplete data log-likelihood(real):

$$l(\boldsymbol{\theta};\mathbf{y}) = \log[f(\mathbf{y};\boldsymbol{\theta})].$$

The second is the complete data log-likelihood(theoretical):

$$l(\boldsymbol{\theta};\mathbf{y},\mathbf{x}) = \log[f(\mathbf{y},\mathbf{x};\boldsymbol{\theta})].$$

Then, the expected log-likelihood is:

$$\tilde{l}\left(\boldsymbol{\theta}\mid\mathbf{y};\boldsymbol{\theta}^{(k)}\right) = E_{\mathbf{X}|\mathbf{y};\boldsymbol{\theta}^{(k)}}\{l(\boldsymbol{\theta};\mathbf{y},\mathbf{X})\} = \int\cdots\int\log[f(\mathbf{y},\mathbf{x};\boldsymbol{\theta})]h\left(\mathbf{x}\mid\mathbf{y};\boldsymbol{\theta}^{(k)}\right)\mathbf{dx}.$$

Note: The expected log-likelihood is taken under the conditional distribution of $\mathbf{X}$ given

$\mathbf{Y}$ where

$$h(\mathbf{x} \mid \mathbf{y}; \boldsymbol{\theta}) = \frac{f(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})}{f(\mathbf{y}; \boldsymbol{\theta})}.$$

The unknown parameter $\boldsymbol{\theta}$ is set equal to $\boldsymbol{\theta}^{(k)}$(the current value of the parameter at step k):

$$h\left(\mathbf{x} \mid \mathbf{y}; \boldsymbol{\theta}^{(k)}\right) = \frac{f\left(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}^{(k)}\right)}{f\left(\mathbf{y}; \boldsymbol{\theta}^{(k)}\right)}.$$

In M-Step, we are interested in the $\boldsymbol{\theta}$ that maximizes the expected log-likelihood:

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\theta} \widetilde{l}\left(\theta \mid \mathbf{y}; \boldsymbol{\theta}^{(k)}\right).$$

MCEM Algorithm

For the joint likelihood

$$Ł(\boldsymbol{\theta}) = \prod_{i=1}^{N} \int \Big\{ f(t_i, \delta_i | \boldsymbol{b}_i, \boldsymbol{\theta}, \alpha) f(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\beta}, R_i) f(\boldsymbol{b}_i | D) \Big\} d\boldsymbol{b}_i,$$

the E-step computes the conditional expectation of the complete-data log-likelihood given the observed data and current parameter estimates. The complete-data can be defined as $\{(\boldsymbol{y_i}, \boldsymbol{w_i}, \delta_i, t_i, \boldsymbol{b_i}), i = 1, 2, \ldots, n\}$, and the complete-data log-likelihood can be written as

$$l_c^{(i)}(\boldsymbol{\theta}) = logf(t_i, \delta_i | \boldsymbol{b}_i, \boldsymbol{\theta}, \alpha) + logf(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\beta}, R_i) + logf(\boldsymbol{b}_i | D),$$

where $\boldsymbol{\theta}$ denote all the unknown parameters, $\boldsymbol{\beta}$ and $\boldsymbol{b_i}$ are the fixed effects and fixed effects, respectively, R_i represents the covariance matrix of measurement errors, and D is the covariance matrix of random effects $\boldsymbol{b_i}$. The tth EM iteration for individual i can be written as

$$Q_i(\boldsymbol{\theta} | \boldsymbol{\theta^{(t)}}) = \int \Big\{ logf(t_i, \delta_i | \boldsymbol{b}_i, \boldsymbol{\theta}, \alpha) + logf(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\beta}, R_i) + logf(\boldsymbol{b}_i | D) \Big\} f(\boldsymbol{b_i} | \boldsymbol{y_i}, \boldsymbol{w_i}, \boldsymbol{\theta^{(t)}}) d\boldsymbol{b}_i.$$

Computing the integral $Q_i(\boldsymbol{\theta} | \boldsymbol{\theta^{(t)}})$ can be challenging, which is where Monte Carlo methods come in. The integral can be approximated by simulating many samples of the unobservable random effects $\boldsymbol{b_i}$ from its conditional distribution given the observed data $f(\boldsymbol{b_i} | \boldsymbol{y_i}, \boldsymbol{w_i}, \boldsymbol{\theta^{(t)}})$. Then we can use $\boldsymbol{b_i}$ to approximate the integral with an empirical mean. Here, we generate samples $(\tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{(1)}, \tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{(2)}, \cdots, \tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{(m_t)})$ at tth EM iteration, and the $(t+1)$th iteration of the EM algorithm can then be written as follows:

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta^{(t)}}) &= \Sigma_{i=1}^{n} Q_i(\boldsymbol{\theta} | \boldsymbol{\theta^{(t)}}) \\ &\approx \Sigma_{i=1}^{n} \Big\{ \frac{1}{m_t} \Sigma_{j=1}^{m_t} \Big[logf(\boldsymbol{y_i} | \boldsymbol{w_i}, \boldsymbol{\beta}, R_i, \tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{(j)}) + logf(\tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{(j)} | D) + logf(t_i, \delta_i | \tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{(j)}, \boldsymbol{\theta}, \alpha) \Big] \Big\}, \end{aligned}$$

The M-step is to maximize $Q_i(\boldsymbol{\theta} | \boldsymbol{\theta^{(t)}})$, which is similar to the complete-data maximization. Iterating the E-step and M-step until convergence, we finally obtain the MLE (or a local

maximizer) $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

3.3.2 Bayesian Methods

The joint modeling of longitudinal and survival data often involves numerous parameters that can lead to poor estimation and identifiability issues. Bayesian methods are well-suited to handle these types of problems as they enable the incorporation of prior information about population parameters and can reduce identifiability issues. However, it is crucial to choose an appropriate prior distribution as it can significantly impact the final results. Furthermore, certain approaches should be used to address computational challenges.

General Concepts

The basic idea for Bayesian methods is to assume the prior distributions for the unknown parameters in the models; then make inference based on the posterior distributions of the parameters given the observed data. Let $\boldsymbol{y}$ be the data we observed with probability density function $f(\boldsymbol{y}|\boldsymbol{\theta})$, which is also called likelihood as $f(\boldsymbol{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$, where $\boldsymbol{\theta}$ denote the unknown parameters, which are random variables with probability density function(prior distribution) $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$, and $\boldsymbol{\theta}_0$ are the hyper-parameters which are frequently obtained from similar studies or expert options. Bayesian methods can be written as

$$\begin{aligned} f(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}), \end{aligned} \tag{10}$$

where $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the likeliood, $f(\boldsymbol{\theta})$ is the prior, and Bayesian methods of $\boldsymbol{\theta}$ is based on the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$. Based on Bayesian methods, the estimation of $\boldsymbol{\theta}$ is the posterior mean:

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\boldsymbol{y}) = \int \boldsymbol{\theta} f(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta},$$

and the accuracy of the estimation measured by the posterior variance:

$$Cov(\hat{\boldsymbol{\theta}}) = Cov(\boldsymbol{\theta}|\boldsymbol{y}) = \int (\boldsymbol{\theta} - E(\boldsymbol{\theta}|\boldsymbol{y}))(\boldsymbol{\theta} - E(\boldsymbol{\theta}|\boldsymbol{y}))^T f(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}.$$

Based on the mean and variance of the posterior distribution, credible intervals can be computed. Unlike frequentist confidence intervals which are measures of uncertainty around effect estimates, a Bayesian 95% credible interval means that there is a 95% probability that the true (unknown) estimate lies within the interval, given the evidence provided by the observed data. However, sometimes high dimensional parameters $\boldsymbol{\theta}$ lead to computational challenges. Markov chain Monte Carlo (MCMC) methods are popular to make such cases feasible by obtaining a sequence of observations from a probability distribution, especially when direct sampling is difficult.

Prior Distribution

The assumptions of the prior distributions of unknown parameters have a high influence on the final results. If we are able to gain some information about the study we are conducting from the past, such as previous experiments, or learn from some experienced experts, we may get reliable prior distributions. However, it is hard to decide which prior is preferred. In such cases, non-informative conjugate priors should be considered, meaning that the resulting posterior distribution belongs to the same family as the prior distribution. For instance, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes s in n independent experiments. The probability mass function can be written as

$$p(s) = \begin{pmatrix} n \\ s \end{pmatrix} q^s (1-q)^{n-s}.$$

Consider a common conjugate prior, a Beta distribution with parameters (α, β),

$$p(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{B(\alpha,\beta)},$$

where α and β are hyperparameters(parameters of the prior). If we consider the unknown probability of success q as random variable, and let $f = n - s$ as the number of failures, we have known that

$$P(s, f|q = x) = \begin{pmatrix} s+f \\ s \end{pmatrix} x^s (1-x)^f,$$

$$P(q = x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}.$$

The threshold of the probability of success q is $[0, 1]$. Then the posterior distribution of s should be

$$\begin{aligned} P(q = x|s, f) &= \frac{P(s, f|x)P(x)}{\int P(s, f|y)P(y)dy} \\ &= \frac{\begin{pmatrix} s+f \\ s \end{pmatrix} x^{s+\alpha-1}(1-x)^{f+\beta-1}/B(\alpha,\beta)}{\int_{y=0}^{1}(\begin{pmatrix} s+f \\ s \end{pmatrix} x^{s+\alpha-1}(1-x)^{f+\beta-1}/B(\alpha,\beta))dy} \\ &= \frac{x^{s+\alpha-1}(1-x)^{f+\beta-1}}{B(s+\alpha, f+\beta)}, \end{aligned}$$

which is another Beta distribution with parameters $(\alpha + s, \beta + f)$.

Markov Chain Monte Carlo(MCMC) methods

Posterior distributions in Bayesian methods are often highly complicated. A Markov chain

Monte Carlo (MCMC) method is an increasingly popular method for estimating posterior distributions in Bayesian methods. A particular strength of the MCMC method is that it can be used to draw samples from distributions even when all the information we have about the distribution is how to calculate the density for different samples [12].

MCMC methods consist of two parts, Monte Carlo methods and Markov chains. Monte Carlo methods are used to estimate the properties of a distribution by examining random samples from the distribution. For instance, the expected value of a normal distribution is of interest. Instead of calculating the mean directly from the distribution's equations, by a Monte Carlo approach, a large number of random samples from a normal distribution will be drawn, and the mean of the samples can be calculated as the expected value. Hence, Monte Carlo approaches lead to easier calculation, especially when the distribution's equations are unable to work out analytically. Markov chains describe sequences of random variables or vectors which have the Markov property: evolution of the Markov process in the future only depends on the present state and is independent of the past. In mathematical way, this can be written as

$$P(X_{n+1} = k | X_n = k_n, X_{n-1} = k_{n-1}, \ldots, X_1 = k_1) = P(X_{n+1} = k | X_n = k_n).$$

The basic idea is to construct a Markov chain that has a desired distribution as its stationary distribution. Stationary distribution has the property: $\pi = \pi P$, which means for a given transition probability matrix P, there is an initial distribution π such that the distribution of all the terms of the chain is equal to the initial distribution π. After a number of steps, we will end up creating a Markov Chain and the state of the chain is then used as a sample of the desired distribution. The larger the number of the steps, the higher the quality of the sample. Hence, in MCMC methods, the random samples are generated by a special sequential process, and each random sample is used as a premise to generate the next random sample [12].

Estimating posterior distributions, $f(\boldsymbol{\theta}|\boldsymbol{y})$, is usually difficult in Bayesian methods. In most cases, we can find the form of $f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$, but computing the marginalized probability (normalizing constant) $f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$ tends to be computationally expensive in lower dimensions and impossible in higher dimensions. MCMC methods can be used to estimate $f(\boldsymbol{\theta}|\boldsymbol{y})$ by drawing samples from it, while avoiding the need to estimate $f(\boldsymbol{y})$ simultaneously.

Model Selection

Various structures are used to link longitudinal and survival models. To select the most appropriate structure for a given dataset when using Bayesian methods, one popular statistical criterion is the deviance information criterion (DIC) [13].

DIC is used to compare the relative fit of a set of Bayesian hierarchical models. Similar to Akaike's information criterion (AIC), it combines a measure of goodness-of-fit and a measure of complexity, both based on the deviance. DIC is particularly useful in Bayesian

model selection problems where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation. However, DIC is only valid when the posterior distribution is approximately multivariate normal. Define the deviance of the unknown parameters θ of the model as

$$D(\theta) = -2log(p(y|\theta)) + C,$$

where y is the data we have, $p(y|\theta)$ is the likelihood function, and C is a constant. The constant can be ignored because it will cancel out when comparing different models. Two calculations are often used for the effective number of parameters of the model. The first can be written as

$$p_D = \overline{D(\theta)} - D(\bar{\theta}),$$

where $\bar{\theta}$ is the expectation of θ. The second can be written as

$$p_D = p_V = \frac{1}{2}\overline{\text{var}\,(D(\theta))}.$$

More effective parameters can make the model fit the data easier, and so the deviance needs to be penalized. The deviance information criterion is calculated as

$$DIC = p_D + \overline{D(\theta)},$$

or

$$DIC = D(\bar{\theta}) + 2p_D.$$

An Example Based on Bayesian Methods

Now we consider a joint model of a primary survival model and a secondary longitudinal model. For individual $i(i = 1, 2, \ldots, N)$, let s_i be the event time, and c_i be the censoring time. Thus, the observed time t_i can be written as $t_i = min\{s_i, c_i\}$. The censoring indicator δ_i can be defined as $\delta_i = I(s_i \leq c_i)$, in which $\delta_i = 0$ means the individual i is right censored and $\delta_i = 1$ otherwise. The time-dependent variable with measurement errors defined as $y_i(t) = (y_{i1}(t), y_{i2}(t), \ldots, y_{in_i}(t))^T$, and μ_i denotes the repeated unknown measurements for each individual i over time without measurement errors. A Cox proportional hazards model can be written as

$$h_i(t_i) = h_0(t_i)exp(\boldsymbol{\gamma}^T \boldsymbol{w_i}),$$

where $\boldsymbol{w_i}$ is a vector of baseline covariates without measurement errors, and $\boldsymbol{\gamma}$ is the regression parameters. A mixed-effects model of longitudinal data can be written as

$$y_i(t) = \mu_i(t) + e_i, \qquad i = 1, 2, \ldots, n,$$

$$\mu_i(t) = X_i(t)\boldsymbol{\beta} + Z_i(t)\boldsymbol{b_i} \quad i = 1, 2, \ldots, n,$$

where $b_i \sim N(0, D)$, $e_i \sim N(0, R_i)$, and b_i and e_i are independent. Joint models can be

shown as

$$h_i(t_i) = h_0(t_i)exp(\boldsymbol{\gamma}^T \boldsymbol{w_i} + f\{\mu_i(t), \alpha, \boldsymbol{b}_i\}),$$

where α denotes the association between the longitudinal data and the risk of the event. *JMBayes* package is a package used for Bayesian joint models. To ensure the flexibility of joint models, it uses B-spline approaches for the baseline hazard function $h_0(t_i)$ in the survival model. Mathematically, the logarithm of the baseline hazard function can be written as

$$logh_0(t) = \gamma_{h_0,0} + \Sigma_{q=1}^{Q} \gamma_{h_0,q} B_q(t, \boldsymbol{v}), \tag{11}$$

where $B_q(t, \boldsymbol{v})$ denotes the qth basis function of a B-spline with knots $\boldsymbol{v} = (v_1, \ldots, v_Q)$, and $\gamma_{h_0,0}$ denotes the spline coefficients. Hence, the larger the number of knots Q, the greater the flexibility in the function $logh_0(t)$. One thing that needs to be carefully considered is the risk of over-fitting, which can arise with an overly large number of knots.

The likelihood of the joint model can be written as

$$Ł(\theta) = \prod_{i=1}^{N} \Big[\int f(t_i, \delta_i | \mu_i, h_0, \boldsymbol{\gamma}, \alpha) f(y_i | \boldsymbol{b}_i, \boldsymbol{\beta}, R_i) f(\boldsymbol{b}_i | D) d\boldsymbol{b}_i \Big],$$

where

$$\int f(t_i, \delta_i | \mu_i, h_0, \boldsymbol{\gamma} = \Big[h_0(t_i) exp(\boldsymbol{\gamma}^T \boldsymbol{w_i} + f\{\mu_i(t), \alpha, \boldsymbol{b}_i\}) \Big]^{\delta_i} exp\Big[- \int_0^{t_i} h_0(x) exp(\boldsymbol{\gamma}^T \boldsymbol{w_i} + f\{\mu_i(t), \alpha, \boldsymbol{b}_i\}) dx \Big],$$

$$f(y_i | \boldsymbol{b}_i, \boldsymbol{\beta}, R_i) = f(y_i | \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-m_i/2} exp[-(y_i - \mu_i)^T (y_i - \mu_i)/2\sigma^2],$$
$$f(\boldsymbol{b}_i | D) = (2\pi |D|)^{-1/2} exp[-(\boldsymbol{b}_i^T D^{-1} \boldsymbol{b}_i)/2],$$

and

$$R_i = \sigma^2.$$

For Bayesian methods, we need to assume the prior distribution of the parameters. We take standard prior distributions for the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{b}_i, \boldsymbol{\beta}, \alpha)$. We assume

$$\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \Sigma_\gamma),$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \Sigma_\beta),$$
$$\alpha \sim N(\boldsymbol{0}, \Sigma_\alpha),$$

and

$$\gamma_{h_0} \sim N(\boldsymbol{0}, \Sigma_{\gamma_{h_0}}).$$

The mean of the normal distribution is centered at zero meaning that the prior assumes the explanatory variables are not associated with the response. Moreover, we assume an inverse Wishart prior for the covariance matrix D of the random effects $\boldsymbol{b}_i$. And for a joint model with a normally distributed longitudinal outcome, we take an inverse-Gamma prior for the variance of the error terms $\sigma^2 (\sigma^2 \sim IG(a, b))$, where $IG(a, b)$ is the inverse Gamma distribution with parameters a and b.

Once prior distributions of the parameters have been defined, posterior distributions of the parameters given the observed data can be expressed. To compute the intractable integral, various computational techniques such as Markov Chain Monte Carlo (MCMC) methods can be used.

3.3.3 Summary

In general, joint modeling for survival data and longitudinal data are usually based on likelihood methods and Bayesian methods. Likelihood methods have high correlation with Bayesian methods. If we use a non-informative (uniform distribution) prior to get maximum likelihood estimates in a Bayesian method, it is equivalent to a likelihood method. Mathematical speaking, for a prior distribution $f(\boldsymbol{\theta}) \propto 1$, the Bayesian method is

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})f(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y}),$$

which is comparable to the likelihood method. Due to the similar form, Bayesian methods and likelihood methods encounter similar computational challenges and use similar computational tools while facing the intractable high-dimensional data.

While comparing with likelihood methods, the major difference of Bayesian methods is that the model parameters are treated as random variables. Bayesian methods actually provide much more information than point estimators like MLE with considering prior and posterior distributions, while the extra information also leads to more complex calculation.

4 Joint Models for Oral Cancer Data Analysis

4.1 Background

Recall that oral cancer accounts for 2%–4% of all cancer cases worldwide. Among all of these oral cancers, oral squamous cell carcinoma(OSCC) is the most common malignant epithelial neoplasm affecting the oral cavity, which is believed to arise through sequential stages of potentially malignant lesions (OPMLs) i.e hyperplasia, mild, moderate, severe dysplasia and carcinoma in situ [3]. The presence and grade of dysplasia are considered the gold standard to assess the risk of malignant transformation; higher grades of dysplasia lead to higher risk. Overall, without treatment, most of the high-grade(severe dysplasia and carcinoma) dysplasia progress to cancer, while most of the low-grade (mild and moderate) dysplasia (LGDs) remain stable. Thus, treatment strategies comply with the gold standard: high-grade dysplasia requires treatment. However, it is challenging to predict LGDs risk since only $5-15\%$ undergo malignant transformation [14]. There is no identical and standard method to work on such cases. In British Columbia, LGDs are followed in specialty clinics, and the severe dysplasia, carcinoma-in situ and SCC(squamous cell carcinoma) are treated surgically. Oral LGDs may cause high or low cancer risk over time ranging widely from 7 to 177 months, which is hard to decide on an appropriate intervention. Non-invasive biomarkers are needed to triage LGDs according to their risk of malignant transformation.

Currently, from a clinical point of view, there is no effective biomarker or diagnostic tool to guide triage or treatments. Hence, clinical risk indicators like size, appearance, and site are important to determine the cancer risk of OPMLs. Studies have been carried on to find biomarkers which enable identifying lesions at a high-risk of malignant transformation in LGDs. DNA aneuploidy has been confirmed to be a marker of various malignancies including oral cancers.

Based on the background, an analysis based on OSCC datasets with patient demographic, clinical features and the corresponding diagnosis results during the study time is discussed in this section. A time varying variable is considered as the response of a longitudinal model and the occurrence of the event of interest is fitted by a survival model. Parameters in joint models are then estimated based on likelihood methods and Bayesian methods. Finally, we focus on the comparison of these two inference methods.

For the longitudinal data, we build mixed-effects models using the **lme(·)** function from the *nlme* package. For the survival data, we build Cox proportional hazards models using the **coxph(·)** function from the *survival* package. For joint modeling of longitudinal and survival data, we use the basic model fitting function called **jointModel(·)** in the *JM* package based on likelihood methods, and **jointModelBayes(·)** in the *JMbayes* package based on Bayesian methods.

4.2 Data Description

There are 408 observations for 41 patients in the dataset. Among these patients who had mild and moderate oral dysplasia lesions at the start of the study, 28 were non-progressors and 13 progressed to severe dysplasia, carcinoma-in-situ, or squamous cell carcinoma by the end of the study. The variables have been summarized in Table 1.1.

lesion_area is the only numeric variable that is measured multiple times on each individual at different time points. Therefore, a longitudinal model is necessary to fit the longitudinal trajectory of **lesion_area**. Patients with **study_id** *2094* and *2046* have no **lesion_area** data, so they should be excluded. Patients *1888* and *1913* should also be dropped as they have missing demographic data. Furthermore, observations taken after the diagnosis time should not be included when building joint models. Thus, measurements tested after the diagnosis time need to be dropped. After filtering, 287 observations for 37 patients aged between 40 and 88 years are available for future analysis. The longitudinal trajectory of **lesion_area** for patients with different diagnoses is shown separately in Figure 4.1. As can be seen from the trajectory plot, *Non-Progressors* who did not develop malignant transformation have relatively lower lesion area compared to the *Progressors* who did develop malignant transformation. The area of lesions in different patients varies over time, and even within the same patient, the lesion area can change over time. Therefore, a longitudinal model with fixed and random effects may be used to explain both between-individual differences and within-individual differences.

Based on the dataset, this section primarily focuses on investigating the factors that could potentially be associated with the incidence of oral cancer. These factors comprise time-dependent variables with measurement errors and time-independent variables. Therefore, we develop a longitudinal model for the time-dependent variable **lesion_area**, and a survival model to fit the time of the event of interest. To obtain a more precise fit, we use joint modeling for longitudinal and survival data based on likelihood methods and Bayesian methods.

4.3 Longitudinal and Survival Analysis

In this section, we aim to analyze the longitudinal data of patients' lesion area over time and the event time data of their oral cancer occurrence by using mixed-effect models and Cox PH models.

4.3.1 Longitudinal Analysis

In this section, we focus on a longitudinal analysis for variable **lesion_area**. First, we normalize the distribution of **lesion_area**, as shown in Figure 4.2.

It is well known that repeated measurements of **lesion_area** for the same individual are correlated. Therefore, a mixed-effects model can be used to analyze the longitudinal data. This model accounts for between-person variability by using invariant fixed effects and

Figure 4.1: Longitudinal Trajectories of Lesion Areas (in log10 scale)

within-person variability by using specific random effects. As shown in Figure 4.1, there are individual differences in the rates of change of lesion areas over time. To account for this, we included the variable **obstime** as a random effect in our model. Moreover, previous research has suggested a high prevalence of oral mucosal lesions and diseases in the geriatric population, with the risk increasing with age [15]. Hence, we also considered age as a possible factor that may be correlated with the rate of change in lesion area over time.

We finally tested seven linear mixed-effects submodels: the first based on fixed effects for intercept and linear slope and random effects for only the intercept, the second adding to the first a random effect for linear slope (to describe linear change over time), the third adding to the first a new predictor **Age**, the fourth adding to the third a random effect for linear slope (to describe linear change over time), the fifth adding to the fourth a quadratic slope (to describe quadratic change over time), and the sixth and the seventh with spline effects of time and with different numbers of nodes. Let $y_i(t) = (y_{i1}(t), y_{i1}(t), \ldots, y_{in_i}(t))$ be the continuous response of lesion area on a $log10$ scale for individual $i(i = 1, \ldots, 37)$ at time t. $obstime_i(t)$ is the measurement times (in year) from the beginning of the study.

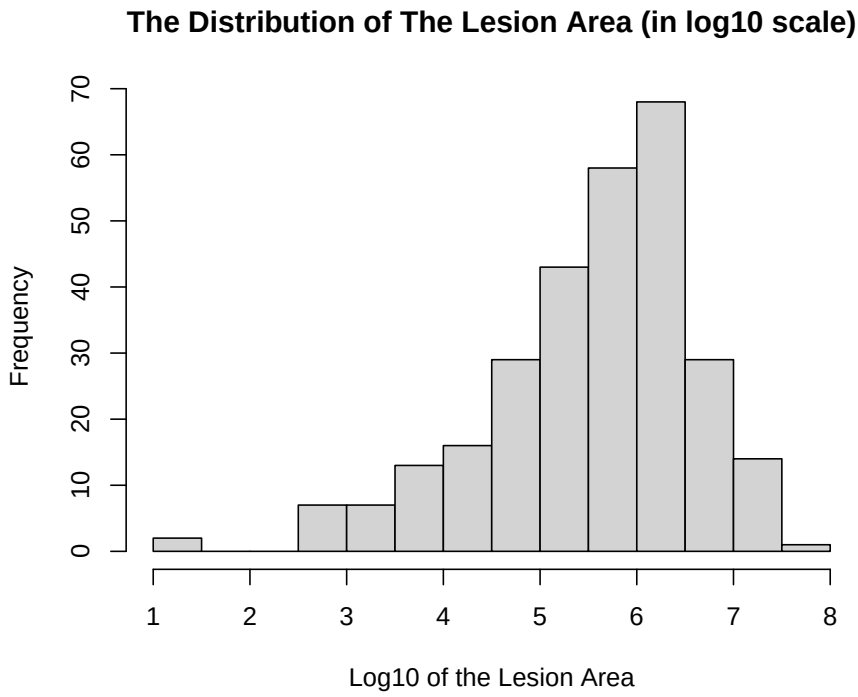

Figure 4.2: The Distribution of The Lesion Area (in log10 scale)

Age_i is the age of the patients. Mathematically, these seven submodels can be expressed as

$$
\begin{aligned}
&lmeFit1 : y_i(t) = (\beta_0 + b_{0i}) + \beta_1 obstime_i(t) + \epsilon_i(t),\\
&lmeFit2 : y_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) obstime_i(t) + \epsilon_i(t),\\
&lmeFit3 : y_i(t) = (\beta_0 + b_{0i}) + \beta_1 obstime_i(t) + \beta_3 Age_i + \epsilon_i(t),\\
&lmeFit4 : y_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) obstime_i(t) + \beta_3 Age_i + \epsilon_i(t),\\
&lmeFit5 : y_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) obstime_i(t) + \beta_2 obstime_i(t)^2 + \beta_3 Age_i + \epsilon_i(t),\\
&lmeFit6 : y_i(t) = f(obstime_i(t), 3) + b_{0i} + b_{1i} obstime_i(t) + \beta_3 Age_i + \epsilon_i(t),\\
&lmeFit7 : y_i(t) = f(obstime_i(t), 4) + b_{0i} + b_{1i} obstime_i(t) + \beta_3 Age_i + \epsilon_i(t).
\end{aligned}
$$

$$log(lesion\_area_i)(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) time_i(t) + \beta_2 time_i^2(t) + \epsilon_i(t)$$

Among the seven submodels, we select the one that provides the most satisfying fitting of the lesion area data based on the Bayesian Information Criterion(BIC) [16]. BIC is a criterion for model selection, and the models with lower BIC are generally preferred. When fitting the data, adding parameters tend to increase the likelihood, but uncontrolled adding might cause overfitting. This problem can be resolved by using a penalty term for the number of parameters in the model, like BIC and Akaike information criterion (AIC) [17].

The BIC is formally defined as

$$\text{BIC} = k\ln(n) - 2\ln(\widehat{L}),$$

where $\hat{L}$ is the maximized value of the likelihood function of the model M, n is the sample size, k is the number of parameters estimated by the model. For example, in a multiple linear regression $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$, the estimated parameters are the intercept(β_0), the p slope parameters($\beta_1, \beta_2, \ldots, \beta_p$), and the constant variance of the errors ϵ_i. Thus, the number of parameters $k = q + 2$. Another reason for using BIC to do model selection is that using likelihood ratio tests to compare models with different fixed effects and random effects should be performed using maximum likelihood estimation (ML). However, we build the longitudinal submodel with the *lme(·)* function from the **nlme** package [18], in which the default estimation method is Restricted Maximum Likelihood(REML). Unlike ML, REML produces unbiased estimates of variance which may lead to an incomparable likelihood. Hence, it is more reasonable to compare longitudinal submodels by comparing BIC which are computed by ML.

The BIC for all the models are displayed in Table 4.1. After comparing them, we choose

Table 4.1: A Comparison of the BIC for Mixed-Effects Models

Model	**df**	**BIC**
m1	4	711.1971
m2	6	703.8555
m3	5	716.6006
m4	7	709.0695
m5	8	712.3846
m6	9	715.6708
m7	10	721.5631

the second model as the best fitting model, which is expressed as

$$y_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})obstime_i(t) + \epsilon_i(t). \tag{12}$$

Then using the function *lme(·)* from **nlme** package to fit the data based on Equation 12. The results and the diagnosis are shown in the following.

Model Diagnosis

There are several general assumptions for a linear mixed-effects model [7]. To assess whether the explanatory variables are linearly related to the response, a plot of the residuals against the explanatory variable can be used to determine if the fitted model is adequate, or if a higher-order term is needed to explain the relationship. To assess whether the errors have constant variance, a plot of the residuals in order can reveal any seasonal patterns or autocorrelation that may exist. To evaluate whether the errors are

independent, a plot of the residuals against the fitted values can be used to check for non-constant error variance. For example, this plot can show if the variance increases with the mean and if the residuals become more spread out as the fitted value increases. A lag plot with no discernible pattern is strong evidence of error independence. To assess whether random effects and errors are normally distributed, a normal probability plot (also known as a quantile-quantile plot), a histogram of the residuals, or a Wilk-Shapiro test can be used to check for normality in the random effects and errors. A normal probability plot is a graphical technique used to evaluate whether a dataset is approximately normally distributed. The points should form a roughly straight line if the data is from a normal distribution.

Hence, several assumptions should be diagnosed after the longitudinal analysis. First, we check the heteroscedasticity of the residuals. As mentioned earlier, inspecting diagnostic plots of the residuals is a crucial tool, and plotting the residuals against the fitted values detects heteroscedasticity. If the model fits the data well, the points should be randomly scattered with no specific pattern. According to Figure 4.3, patients with **ID** equal to **36**, **37**, and **32** have two outliers each. Another assumption of linear mixed-effects models is that there is no heteroscedasticity among different levels of the random effects. We can verify this assumption by plotting the residuals by splitting all observations into different individuals. According to Figure 4.4, nearly all the data for each patient are around the middle line, except for patient *3011* ($ID = \mathbf{30}$), who has one outlier data point. We dropped this point in the following analysis. Normal distributions of random effects and errors are also basic assumptions for a linear mixed-effects model. We can check this assumption using Q-Q plots. Ideally, the points (errors or random effects) from a normal distribution on the Q-Q plot will lie alongside (or close to) the straight line. According to the Q-Q plot of the errors shown in Figure 4.5, most of the points fall along a line in the middle of the graph but curve off in the extremities. Such behavior means that there are more extreme values of the errors than would be expected if they truly came from a normal distribution. Hence, compared to a theoretical normal distribution, the distribution of the errors has "heavy tails". Even though there are some outliers in the data frame, no significant violations of the assumptions are detected.

Results

The parameter estimates are summarized in Table 4.2. In the table, $\boldsymbol{\beta} = (\beta_0, \beta_1)$ represents the fixed effects corresponding to the global intercept and the coefficients of **obstime**, respectively. The global intercept is the mean of all the log of lesion areas when the numerical covariates are zero, which equals 5.6215. The p-value of this estimate is 0.0000. If we define the significance level as 0.05, then the p-value of the global intercept is lower than what we desired, which means we should reject the null hypothesis that the global intercept is zero. The coefficient β_1 is small, which suggests that the linear association (slope) between **lesion_area** and **obstime** is not very strong. For instance, the fixed effect of **obstime** equals -0.0465, which means that a one-unit change of **obstime** will cause a 0.9546 change of **lesion_area**. The p-value of this estimate is larger than 0.05. However, since we only have 286 observations for a small sample size, which is not large enough to

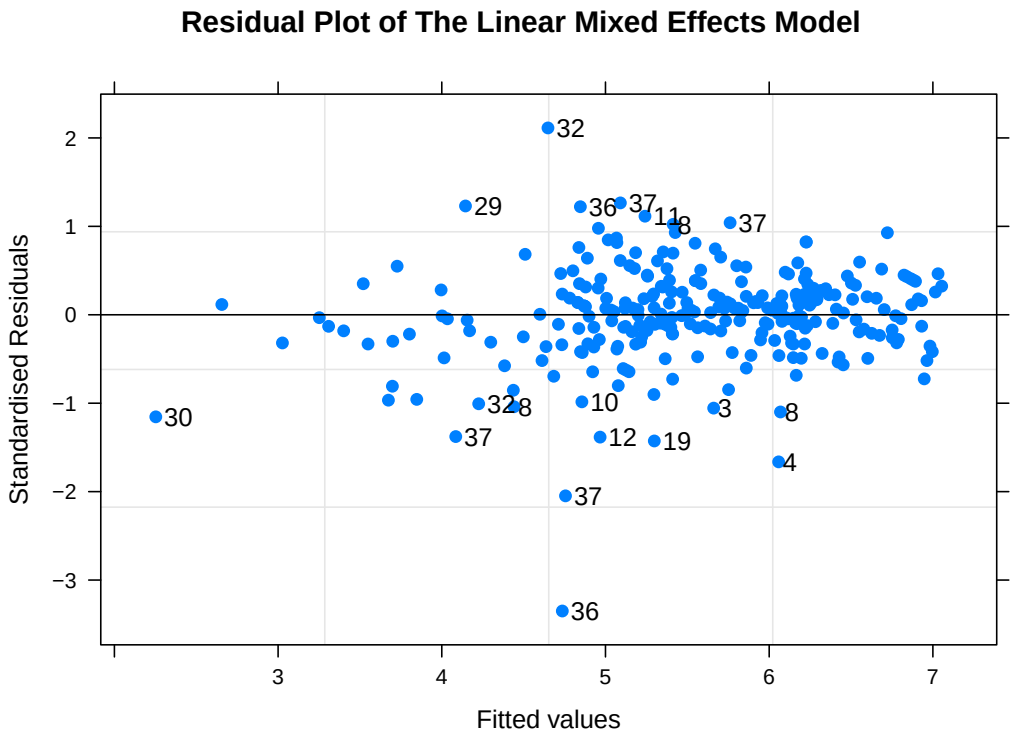

Figure 4.3: Residuals Plot of The Linear Mixed-Effects Model

conclude that the slopes are different from 0. The standard error of the estimate indicates the average deviation of the observed values from the regression line, reflecting the average accuracy of the model fit. Smaller values indicate a closer fit. In this case, the standard errors of the fixed effects are small, indicating a good fit of the model.

Table 4.2: Summary Table of The Linear Mixed-Effects Model of OSCC Data

Parameter	**Estimation**	**Standard Error**	**P-value**
β_0	5.6215	0.1451	0.0000
β_1	-0.0465	0.0447	0.2991

To assess the goodness of the model fits, nine patients were randomly selected to visualize the model-fitting. In Figure 4.6, the lines represent the fitted values based on the model, and the observed measurements are shown as dots. Based on the figure, the model appears to fit well for each of the randomly selected patients.

4.3.2 Survival Analysis

In this section, we shift our focus to a survival submodel. We perform a Cox PH model to model the hazard of developing oral squamous cell carcinoma. According to previous

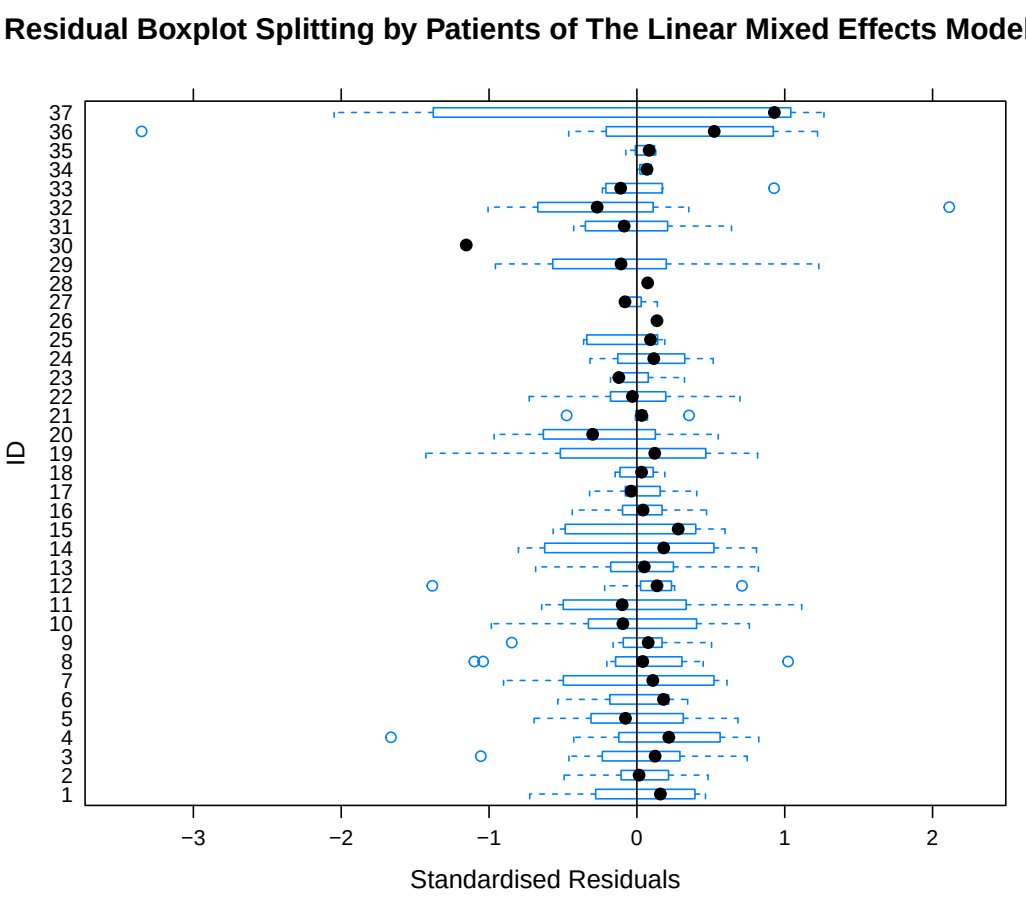

Figure 4.4: Residual Boxplot Splitting by Patients of The Linear Mixed-Effects Model

research, heavy alcohol consumption, especially when combined with cigarette smoking, increases the risk of oral squamous cell carcinoma [19]. Hence, in the COX PH model, we consider **cycling_mean**, **Smoke** and **Alcohol** as explanatory variables and the time to diagnosis is response.

Let $h_i(t)$ be the hazard of the event for individual i at time point t, and $h_0(t)$ is the baseline hazard at time t. The Full Model can be written as

$$h_i(t) = h_0(t)exp(\gamma_1 w_1 + \gamma_2 w_2 + \gamma_3 w_3 + \gamma_4 w_4), \tag{13}$$

where w_1 denotes the variable **cycling_mean**, w_2 the variable **Smoke**, w_3 the variable **Alcohol**, and w_4 the interaction of **Alcohol** and **Smoke**. The boxplot of the log of **cycling_mean** shows in Figure 4.7. For the x-axis, 1 denotes the **Progressor** who developed oral squamous cell carcinoma, and 0 denotes the **Non-Progressor** who did not develop oral squamous cell carcinoma. According to Figure 4.7, some differences exist between these two groups. The mean of the number of cycling cells in these two groups seems to be different, and the range of these two groups are as well different. From the

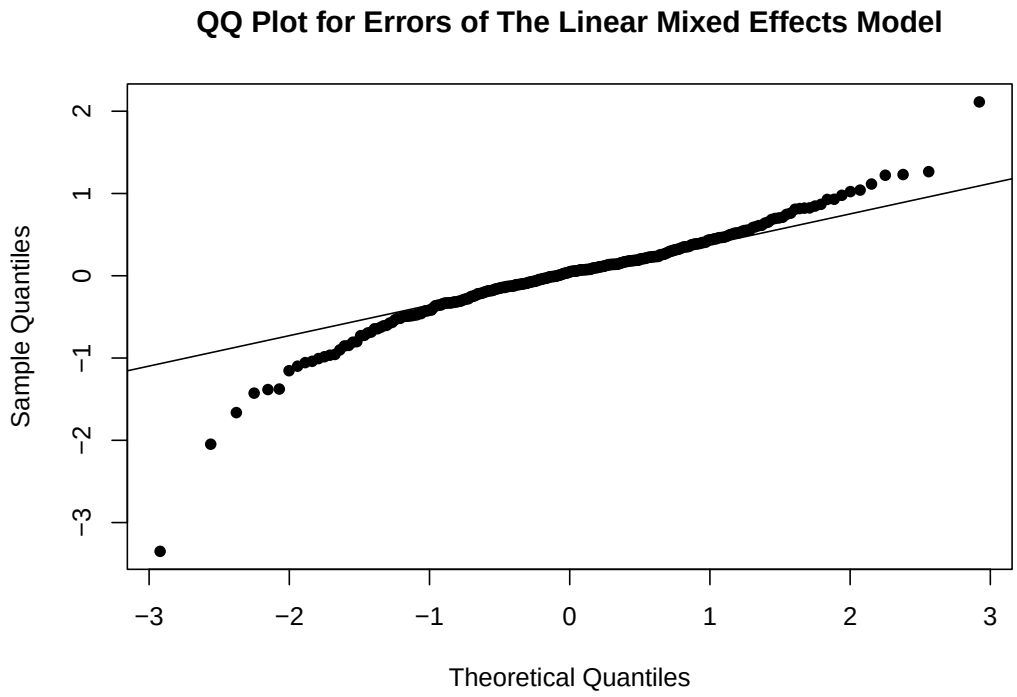

Figure 4.5: QQ Plot for Errors of The Linear Mixed-Effects Model

Kaplan-Meier estimates in Figure 4.8 and Figure 4.9, it seems that the non-alcohol group has slightly higher survival than the alcohol group, and non-smoke group has slightly higher survival than the smoke group.

Model Diagnosis

An important assumption of Cox PH models is that the baseline hazard functions for model predictors are proportional [20]. Hence, after fitting the data by a Cox PH model, we need to check the PH assumption for each covariate. This can be done based on the Schoenfeld Residuals Test (SRT) [21], which evaluates the independence between model residuals and the time variable **obstime**. The summary of the tests are shown in Table 4.3. The test result indicates that the PH assumption holds for all of the predictors (with $p - values > 0.05$) and for the whole survival submodel (with a $p - value > 0.05$).

Results

After fitting the survival submodel Equation 13, the parameter estimates are summarized in Table 4.4. It should be noted that the p-value for all three overall tests (likelihood, Wald, and score) are not significant.

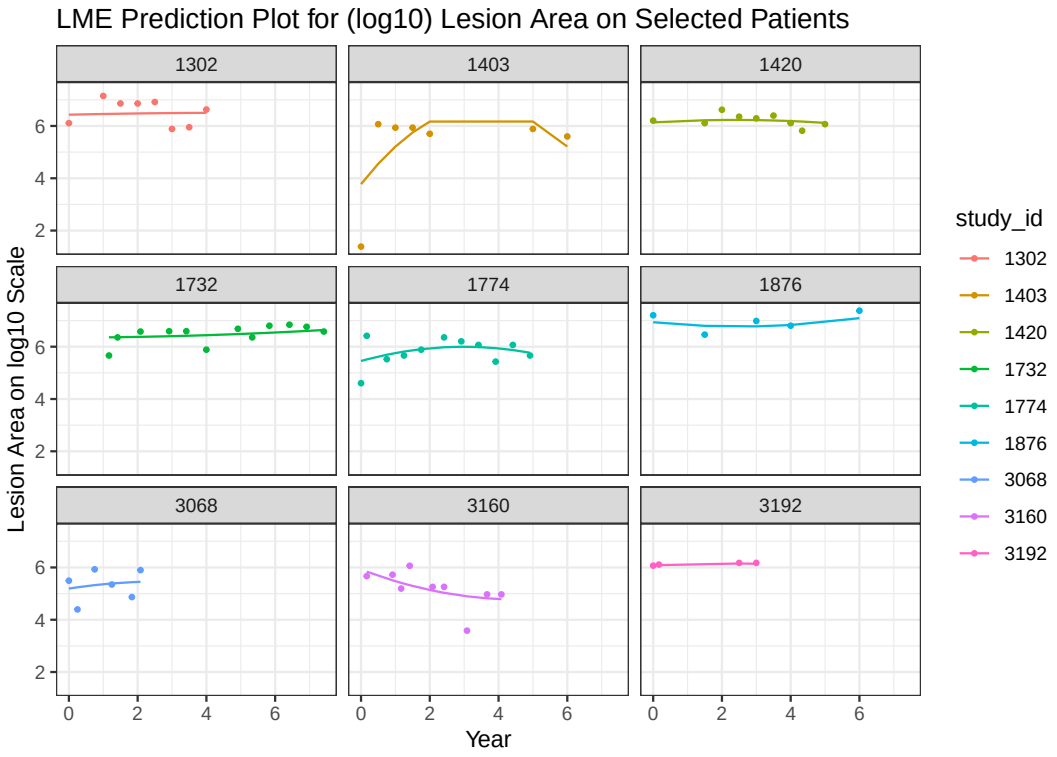

Figure 4.6: LME Prediction Plot for (log10) Lesion Area on Selected Patients

Table 4.3: Schoenfeld Residuals Test for the Survival Model of OSCC Data

Parameter	**Chisq**	**P-value**
Alcohol	0.235	0.63
Smoke	0.844	0.36
cycling_mean	0.720	0.40
Alcohol:Smoke	1.067	0.30
GLOBAL	6.119	0.19

Table 4.4: Parameter Estimates Based on the Survival Model of OSCC Data

Parameter	**Coefficient**	**Standard Error**	**P-value**
Alcohol1	1.2132	1.0867	0.2642
Smoke1	2.6852	1.4310	0.0606
cycling_mean	-0.0054	0.0040	0.1776
Alcohol1:Smoke1	-3.4026	1.8223	0.0619

The p-value of **Alcohol** is 0.26, with a hazard ratio $HR = \exp(1.2132) = 3.36$, and a 95% confidence interval of 0.3998 to 28.306. The p-values of **Smoke** and the interaction between **Smoke** and **Alcohol** are around 0.06, which is relatively low. The hazard ratio of

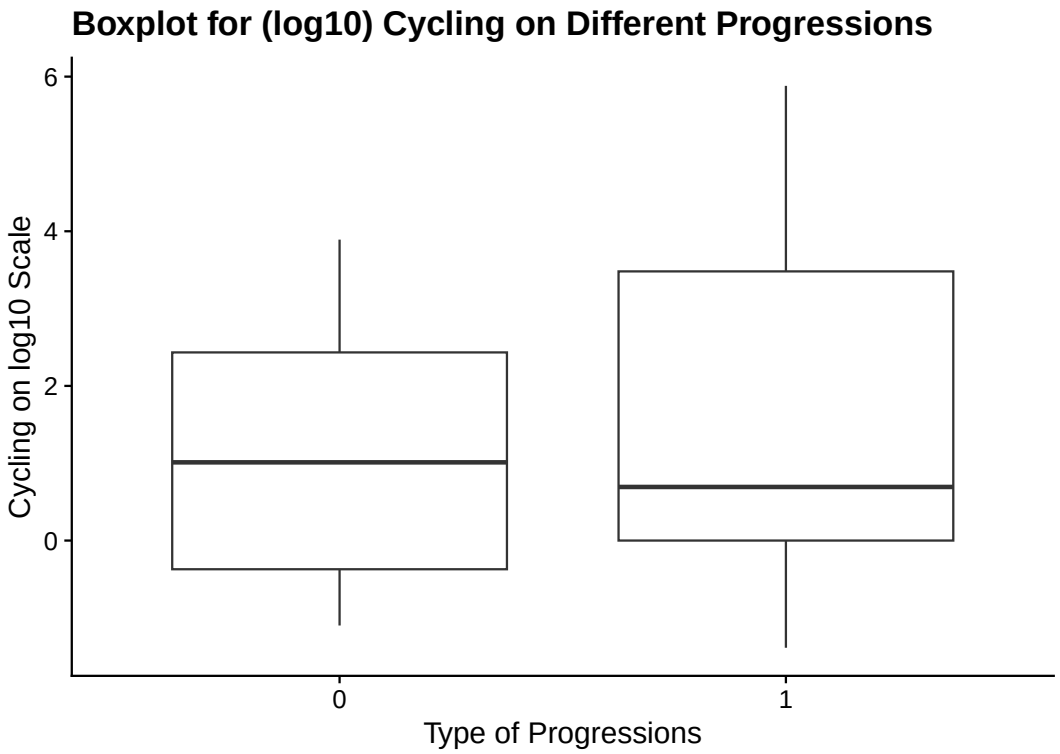

Figure 4.7: The Boxplot for (log10) Cycling on Different Progressions

Smoke is $HR = \exp(2.6852) = 14.66$, hence smoking has a strong influence on developing progression. However, the hazard ratio of the interaction between **Smoke** and **Alcohol** is $HR = \exp(-3.4026) = 0.033$, which implies smoking and drinking may lead to lower risk of experiencing progression. The p-value of **cycling_mean** is 0.18, with a hazard ratio $HR = \exp(-0.0054) = 0.99$, indicating a relationship between the number of cycling cells and decreased risk of malignant transformation. Holding other predictors constant, a higher value of **cycling_mean** is associated with a lower risk of malignant transformation. All p-values of these parameters are larger than 0.05, which means all of these estimates are not statistically significant. A Null Cox PH Model with no explanatory variables is then built to compare nested survival models. The Null COX PH Model can be written as

$$h_i(t) = h_0(t). \tag{14}$$

We use the *anova(·)* function to perform an F-test. The F-test is a statistical test that is commonly used to compare different statistical models fitted to observed data sets. It is used to identify the model that best fits the population from which the data were sampled [22]. The F-test result is summairzed in Table 4.5. The p-value is approximately 0.22, which is larger than 0.05, indicating that there is no significant difference between the Null model and the Full model. Hence, the Null Cox PH model should be selected to do the future analysis.

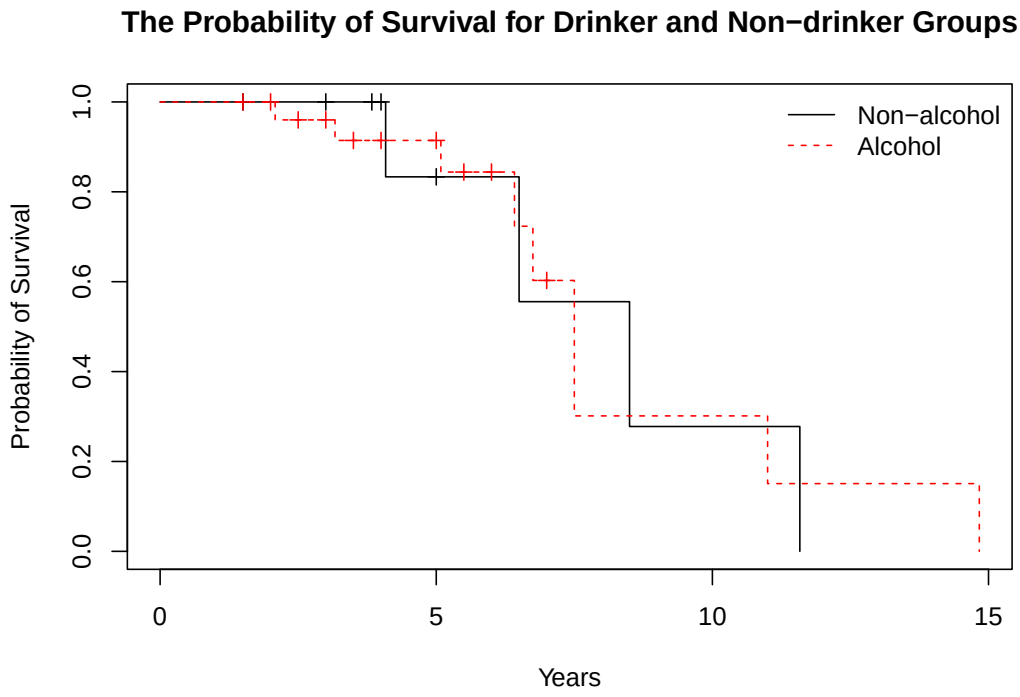

Figure 4.8: Kaplan-Meier Estimates of the Probability of Survival for Drinker and Non-drinker Groups

Table 4.5: A Comparison of the Full and Null COX PH Models of OSCC Data

Model	**loglik**	**Chisq**	**Pr**$(>\|Chi\|)$
Null Model	-25.938	-	-
Full Model	-23.071	5.734	0.2199

4.4 Likelihood Joint Models

A Cox PH model is used to fit the time it takes for a patient to develop malignant transformation. However, some research suggests that a higher **lesion_area**, especially greater than $200mm^2$, is associated with a higher risk of progression. In the following two sections, we proceed by specifying and fitting joint models that explicitly include the linear mixed-effects model Equation 12 for the **lesion_area** and a null Cox PH model Equation 14 for the survival data.

In this section, we present an analysis of joint models based on likelihood methods which is implemented by the **jointModel(**$\cdot$**)** function in R package *JM*. As we elaborated before,

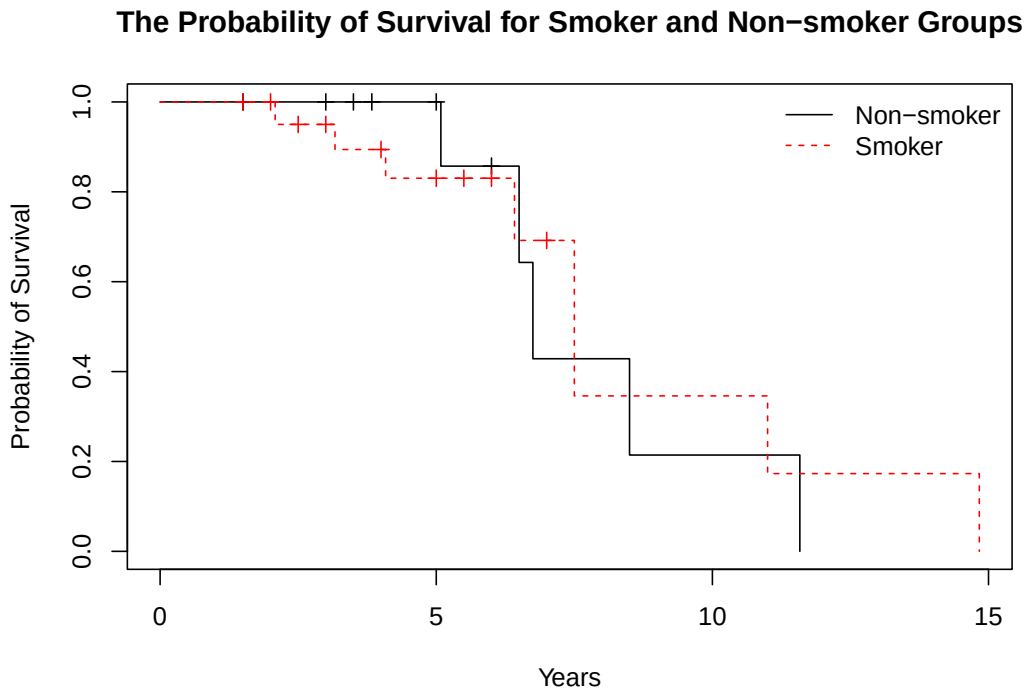

Figure 4.9: Kaplan-Meier Estimates of the Probability of Survival for Smoker and Non-smoker Groups

the joint model can be shown as

$$h_i(t) = h_0(t)exp(f\{\mu_i(t), \boldsymbol{b_i}, \boldsymbol{\alpha}\}), \tag{15}$$

where $h_i(t)$ denotes the hazard of the event for individual i at time point t, $h_0(t)$ is the baseline hazard at time t, $\mu_i(t)$ is the true and unobserved value of the longitudinal outcome **lesion_area** at time t, $\boldsymbol{b}_i$ is the fixed effects, and $\boldsymbol{\alpha}$ measures the association strength of the **lesion_area** to the risk for developing malignant transformation. Note that $\mu_i(t)$ is different from $y_i(t)$, with the latter being contaminated with measurement errors at time t. The density function of joint models is of the following form:

$$p(T_i, \delta_i, y_i) = \int p(y_i|\boldsymbol{b}_i)\{h(T_i|\boldsymbol{b}_i)^{\delta_i} S(T_i|\boldsymbol{b}_i)\}p(\boldsymbol{b}_i)d\boldsymbol{b}_i, \tag{16}$$

where $\boldsymbol{b}_i$ is the vector of fixed effects, $p(\cdot)$ is the density function, and $S(\cdot)$ is the survival function which can be expressed as

$$S(t|\boldsymbol{b}_i) = exp\Big\{-\int_0^t h_0(u)exp(f\{\mu_i(t), \boldsymbol{b_i}, \boldsymbol{\alpha}\})du\Big\}.$$

4.4.1 Likelihood Joint Models Based on Varying Associations

Maximum likelihood estimation for Equation 15 is based on the maximization of the log-likelihood corresponding to the joint distribution of the survival and longitudinal outcomes T_i, δ_i, y_i, where T_i is the observed failure time for the subject $i (i = 1, 2, \ldots, 36)$, δ_i is the indicator of censoring, and y_i is the observed longitudinal outcome. Under the following two assumptions, the longitudinal outcome is independent of the survival outcome; the repeated measurements in the longitudinal outcome are independent of each other. We have

$$p(T_i, \delta_i, y_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = p(T_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{\theta})p(y_i|\boldsymbol{b}_i; \boldsymbol{\theta}), \tag{17}$$

$$p(y_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = \prod_j p\{y_i(t_{ij})|\boldsymbol{b}_i; \boldsymbol{\theta}\}, \tag{18}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_t, \boldsymbol{\theta}_y, \boldsymbol{\theta}_b)$ denote all of the parameters, $\boldsymbol{\theta}_t$ contain the parameters in the event time model, $\boldsymbol{\theta}_y$ contain the parameters in longitudinal model, $\boldsymbol{\theta}_b$ denote the unique parameters of the random-effects covariance matrix, and $p(\cdot)$ denotes an appropriate probability density function of the linear mixed-effects model. The joint log-likelihood function for the i-th subject can be expressed as

$$log\{p(T_i, \delta_i, y_i; \boldsymbol{\theta})\} = log \int p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) \Big[\prod_j^{n_i} p\{y_i(t_{ij})|\boldsymbol{b}_i; \boldsymbol{\theta}_y\}\Big] p(\boldsymbol{b}_i; \boldsymbol{\theta}_b) d\boldsymbol{b}_i, \tag{19}$$

where $p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta})$ is the likelihood of the survival model, which can be formulated as

$$p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) = \Big[h_0(t_i) \exp(f\{\mu_i(t), \boldsymbol{b_i}, \boldsymbol{\alpha}\})\Big]^{\delta_i} \times \exp\Big[-\int_0^{t_i} h_0(u) \exp(f\{\mu_i(t), \boldsymbol{b_i}, \boldsymbol{\alpha}\}) du\Big]. \tag{20}$$

There are three common structures for modeling the association between the longitudinal submodel and the risk of the event of interest: the "current value" association, the "current value plus slope" association, and the "shared parameters" association. However, the **jointModel(·)** function in the R package *JM* can only handle the first two structures. On the other hand, the **jointModelBayes(·)** function in the R package *JMbayes* is able to handle all three structures. If we assume that the longitudinal submodel and survival submodel associate with current value, it can be written as

$$f\{\mu_i(t), \boldsymbol{b_i}, \alpha\} = \alpha\mu_i(t). \tag{21}$$

If we assume that the longitudinal submodel and survival submodel associate with current value and the slope, it can be written as

$$f\{\mu_i(t), \boldsymbol{b_i}, \boldsymbol{\alpha}\} = \alpha_1\mu_i(t) + \alpha_2\mu_i'(t). \tag{22}$$

The method argument of **jointModel(·)** function specifies several types of the baseline hazard function $h_0(t)$ of the survival submodel. One common method is ***method* = "*piecewise* − *PH* − *GH*"**, which means the joint model Equation 15 has a

piecewise-constant baseline risk function. In particular, the function can be written as

$$h_0(t) = \Sigma_{q=1}^{Q} \xi_q I(v_{q-1} < t \leq v_q), \tag{23}$$

where ξ_q denotes the value of the hazard in the interval $(v_{q-1}, v_q]$, $0 = v_0 < v_1 < \cdots < v_Q$ denotes a split of the time scale, and v_Q is larger than the last observed time. GH means the Gauss-Hermite integration rule of approximating the integral. We set $\boldsymbol{method = \text{“}piecewise - PH - GH\text{”}}$ in this analysis.

Maximization of the log-likelihood function Equation 19 involves intractable integrals that may arise due to high dimensions. One effective method is to use Monte Carlo EM algorithms to tackle such a challenging computation. One important advantage of such algorithms is that they only simulate random effects in the first iteration; thus, the computational burden can be reduced, and the likelihood that the increasing property of the original EM algorithm can be preserved [23].

4.4.2 Results Based on Likelihood Joint Models

The summary of fitting the two likelihood joint models based on different association structures are shown in Table 4.6. One is to fit Equation 15 based on “current value” association, which formally can be expressed as

$$h_i(t) = h_0(t) exp(\alpha_1 \mu_i(t)). \tag{24}$$

The other is to fit Equation 15 based on current and slope value association, which formally can be expressed as

$$h_i(t) = h_0(t) exp(\alpha_1 \mu_i(t) + \alpha_2 \mu_i'(t)), \tag{25}$$

where

$$\mu_i'(t) = \frac{\partial \mu_i(t)}{\partial obstime} = \beta_1 + b_{1i}.$$

In Table 4.6, coefficient α_1 represents the strength of the correlation between the current true value of oral lesion area and the risk of developing the disease, while α_2 represents the strength of the correlation between the true rate of change of oral lesion area and the risk of experiencing the disease. For the “current value” association, $\alpha_1 = -0.3943$ can be interpreted as a one unit increase in the current true value of oral lesion area causing a 67.42% decrease in the risk of progression, calculated as $(1 - exp(-0.3943)) \times 100$. For the “current value plus slope” association, $\alpha_2 = 3.9933$ can be interpreted as a one unit increase in the rate of change of the true value of $\mu_i(t)$ causing a 5325.35% increase in the risk of progression, given the constant current value. But all of these estimates are not statistically significant for all p-values of these estimates are not less than 0.05.

We conducted a comparison between a standard time-dependent Cox proportional hazards (PH) model with the logarithm of ’lesion_area’ and a joint model that incorporates the ’current value’ association. The results showed that the joint model has a larger standard

Table 4.6: Likelihood Joint Models Based on Different Associations

Association		α_1	α_2	σ
Current Value	Estimate	-0.3943	-	0.5957
	S.E.	0.2747	-	-
	P-value	0.1512	-	-
Current Value + Slope	Estimate	-0.7657	3.9933	0.5970
	S.E.	0.4638	4.1546	-
	P-value	0.0988	0.3365	-

error, indicating a stronger bias in the estimation of the effect of lesion area. However, the estimates of these two models are similar. The estimated regression coefficient for lesion area in the time-dependent Cox model is -0.3427, while the estimated coefficient for lesion area in the joint model is -0.3943.

AIC and BIC are good for comparing the performance and accuracy of these two likelihood joint models based on different associations. To be specific,

$$AIC = 2k - 2\ln(\hat{L}),$$

and

$$BIC = k\ln(n) - 2\ln(\widehat{L}),$$

where $\hat{L}$ is the maximized value of the likelihood function of the model we built under observed values, n is the number of data points in the observed data(sample size), and k is the number of parameters estimated by the model. Typically, the penalty term for the number of parameters in the model is larger in BIC than in AIC. The results are summarized in Table 4.7. The table shows that the "current value" association joint model has a better fit, as indicated by its lower AIC and BIC.

Table 4.7: AIC and BIC of the Likelihood Joint Models

Association	**AIC**	**BIC**
Current Value	743.36	765.53
Current Value + Slope	744.87	768.62

4.4.3 A Diagnosis of Likelihood Joint Models

After comparing the joint models Equation 24 and Equation 25 using the *anova(·)* function, no significant difference was found. Therefore, we prefer the simpler model, the likelihood joint models based on the "current value" association. Figure 4.10 shows the diagnostic plots for the fitted likelihood joint models based on the "current value" association. The top two panels show the diagnostics of the longitudinal analysis. Specifically, the top-left panel describes the within-subject residuals for the longitudinal

analysis versus their corresponding fitted values. The top-right panel shows the Q-Q plot of the standardized subject-specific residuals for the longitudinal analysis. The bottom-left panel displays estimates of the marginal survival function for the survival analysis. The bottom-right panel describes an estimate of the marginal cumulative risk function for the survival analysis. The plots above indicate that the model is a good fit.

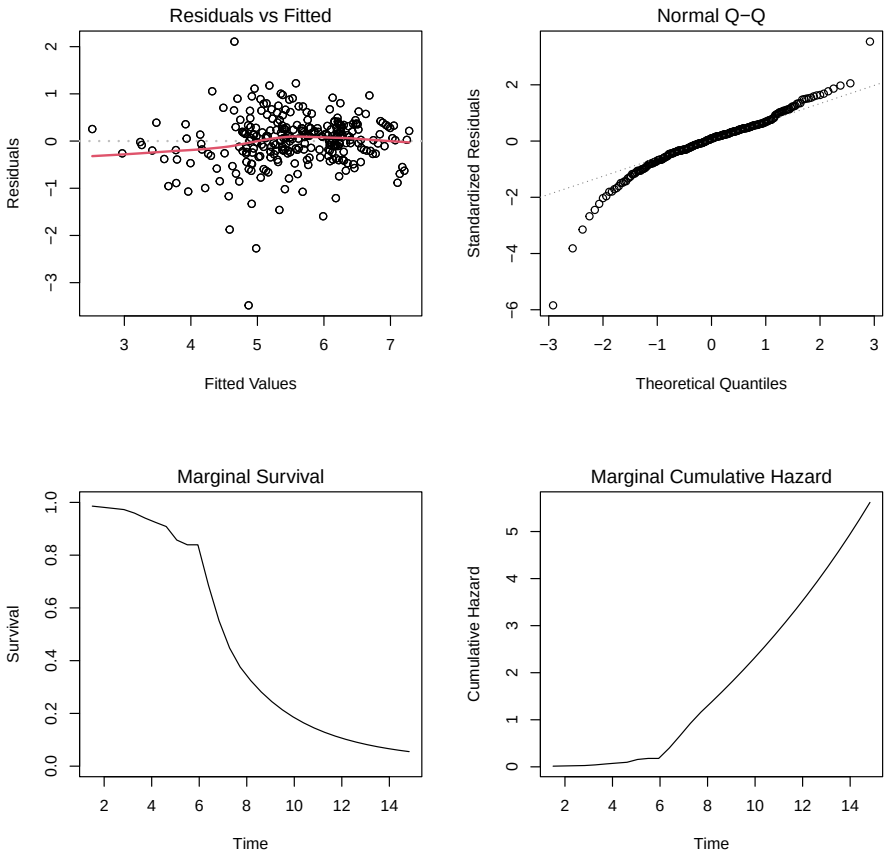

Figure 4.10: Diagnostic Plots for the Fitted Likelihood Joint Model

4.5 Bayesian Joint Models

In this section, we discuss an analysis of Bayesian joint models which is implemented by the **jointModelBayes(·)** function in R package *JMbayes*.

4.5.1 Joint Models Based on "Shared Parameters" Association

There are three structures of association available for joint modeling based on Bayesian methods: the "current value" association, the "current value plus slope" association, and the "shared parameters" association. In **Section 4.4**, we focus on the first two structures

based on likelihood methods. Now, we will switch gears to the "shared parameters" association, also known as the "shared random effects parameters" association. As the name suggests, the vector of time-independent random effects $\boldsymbol{b}_i$ exists in both the longitudinal and survival analysis. The joint model based on the "shared parameters" association can be written as

$$h_i(t) = h_0(t)exp(\boldsymbol{\alpha}^T\boldsymbol{b}_i) = h_0(t)exp(\alpha_0 b_{0i} + \alpha_1 b_{1i}), \tag{26}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^T$ denote the strength of association between the random effects from longitudinal analysis and the survival analysis, and $\boldsymbol{b}_i = (b_{0i}, b_{1i})^T$ denote the random effects. For each specific subject i, the random effect $\boldsymbol{b}_i$ is unique. In this study, we assume the random effects $\boldsymbol{b}_i \sim N(0, D)$, the error in longitudinal analysis $e_i \sim N(0, R_i)$, and b_i and e_i are independent.

For the joint model based on the "shared parameters" association, the posterior distribution of the unknown parameters is derived under the assumption that given the random effects, both the longitudinal and survival processes are independent, and the longitudinal trajectories for different subjects are also independent.

The expression of the likelihood of Equation 26 can be written as

$$p(T_i, \delta_i, y_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = p(T_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{\theta})p(y_i|\boldsymbol{b}_i; \boldsymbol{\theta}), \tag{27}$$

$$p(y_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = \prod_j p\{y_i(t_{ij})|\boldsymbol{b}_i; \boldsymbol{\theta}\}, \tag{28}$$

where T_i is the observed event time for the subject $i(i = 1, 2, \ldots, n)$, δ_i is the indicator of censoring, and y_i is the observed longitudinal outcome. Vector $\boldsymbol{\theta} = (\theta_t, \boldsymbol{\theta}_y, \boldsymbol{\theta}_b)$ contains all of the unknown parameters. Specifically, $\boldsymbol{\theta}_t = \boldsymbol{\alpha}$ denotes the parameters in the event time model, $\boldsymbol{\theta}_y = (\boldsymbol{b}_i, \boldsymbol{\beta}, D, R_i)$ denote the parameters in the longitudinal model, and $\boldsymbol{\theta}_b = (\boldsymbol{b}_i, D)$ denote the unique parameters of the random-effects covariance matrix. After assuming the prior distribution $p(\boldsymbol{\theta})$, the posterior distribution can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{b}) \propto \prod_i^n \prod_j^{n_i} p(y_i(t_{ij})|\boldsymbol{b}_i; \boldsymbol{\theta})p(T_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{\theta})p(\boldsymbol{b}_i|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{29}$$

where $y_{ij} = y_i(t_{ij})$ represents the longitudinal observed value for individual $i(i = 1, 2, \ldots, n)$ at time $j = (1, 2, \ldots, n_i)$. Recall that we assume the Cox PH model and the LME model share the same random effects $\boldsymbol{b}_i$, hence, the likelihood of the survival part can be formulated as

$$p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_t) = \Big[h_0(t_i)\exp\{\boldsymbol{\alpha}^T\boldsymbol{b}_i\}\Big]^{\delta_i} \times \exp\Big[-\int_0^{t_i} h_0(u)\exp\{\boldsymbol{\alpha}^T\boldsymbol{b}_i\}du\Big]. \tag{30}$$

4.5.2 Results Based on the Bayesian Joint Model

After fitting the model Equation 26 using the **jointModelBayes(·)** function in the *JMbayes* package, the summary are presented in Table 4.8. The coefficient α_0 represents the correlation between the random intercept and the survival analysis, while α_1 represents the correlation between the random effect of **obstime** and the survival analysis. The standard deviation of the residuals of the joint model is represented by σ. According to the parameters of $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$, patients with lower baseline levels of the longitudinal outcome (intercept) and steeper decreases in their longitudinal trajectories (slope) are more likely to experience the event. However, the results suggest that both the baseline levels of the underlying lesion area and the longitudinal evolution are not strongly related to the hazard of progression, as the p-values are larger than 0.05 and thus not significant.

Table 4.8: The Joint Model with Bayesian Methods on "Shared Parameters" Associations

Parameter	**Coefficient**	**Standard Error**	**2.5%**	**97.5%**	**P-value**
α_0	-0.6765	0.0448	-1.6303	0.2678	0.168
α_1	-0.7421	0.0365	-3.1076	1.2063	0.493
σ	0.5737	-	-	-	-

In addition, several model selection criteria have been obtained, including the log pseudo marginal likelihood value (LPML) of -427.06, the deviance information criterion (DIC) of 824.99, and the effective number of parameters component of DIC (pD) of 74.50.

4.5.3 Joint Models Based on Other Associations

To compare the parameter estimates based on different inference methods under different associations, two more models were built based on Bayesian Methods, which is also completed by the function **jointModelBayes(·)** in *JMbayes* package in R. In Table 4.9, α_1 quantifies the association between the current features of the oral lesion area at time t and the hazard of progression at the same time point, while α_2 describes the strength of the correlation between the slope of the oral lesion area at time t and the hazard of progression. The value of $\alpha_1 = -0.47$ in the "current value" association indicates that for patients with the same underlying level of the lesion area at time t, if the lesion area increases by 50%, the corresponding hazard ratio of developing progression is 0.83 (95% CI: [0.68, 1.01]). This is because a difference of 0.41 ($log(1.5)$) in the log-scale for lesion area corresponds to a ratio of 1.5, and thus $exp(0.41 \times \alpha_1)$ corresponds to the hazard ratio for a 50% increase in lesion area. The values of α_1 in the "current value" and "current value plus slope" associations are similar. The value of α_2 in the model with the "current value plus slope" association is relatively large, indicating a strong correlation between the rate of change of the lesion area and survival. This means that individuals with a higher rate of change have a 32% greater risk of developing progression than those with lower rates of change. The residual standard deviation (σ) is used to describe the difference in standard deviations between observed and predicted values in a regression model. The values of σ in

these two models are similar.

Table 4.9: A Comparison of Bayesian Joint Models Based on Varying Associations

Association		α_1	α_2	σ
Current Value	Estimate	-0.4710	-	0.5744
	S.E.	0.0309	-	-
	2.5%	-0.9612	-	-
	97.5%	0.0157	-	-
	P-value	0.473	-	-
Current Value + Slope	Estimate	-0.5348	0.2751	0.5748
	S.E.	0.0857	0.1869	-
	2.5%	-1.0484	-2.4652	-
	97.5%	0.0319	2.6918	-
	P-value	0.068	0.807	-

The following comparison of the three Bayesian joint models based on different associations relies on the DIC, where smaller values indicate better model fitting to the data. The results are summarized in Table 4.10. According to the DIC, these three models show very similar performance, although the joint model with the "current value" association has the best performance.

Table 4.10: DIC of Bayesian Joint Models Based on Different Associations

Association	**DIC**
Shared Parameters	824.99
Current Value	823.51
Current Value + Slope	824.61

4.5.4 A Diagnosis of Bayesian Joint Models

Different diagnostic approaches are conducted to assess the performance of the three Bayesian joint models. A number of parameters play a role in this section: β_0 represents the fixed effects for the intercept, while β_1 represents the fixed effects for the variable **obstime**. In joint models based on "shared parameters" association, α_0 denotes the correlation between the random effects of the intercept and the survival analysis, and α_1 denotes the correlation between the random effects of the **obstime** and the survival analysis. In joint models based on "current value" association and "current value plus slope" association, α_1 represents the correlation between the current longitudinal measurement and the survival analysis. Lastly, in joint models based on "current value plus slope" association, α_2 denotes the correlation between the current rate of change in

the longitudinal measurement and the survival analysis.

Sensitivity Analysis

The *JMBayes* package uses B-spline approaches to model the baseline hazard function $h_0(t_i)$ in survival models, and the number of knots can affect the model's flexibility. To examine the impact of the number of knots on the final results, we built 12 models, with each association building 4 models with 11 knots, 13 knots, 15 knots, and 17 knots. Table 4.11 summarizes the posterior means for the fixed effects of the longitudinal submodel for these 12 models. The values of β_0 are consistent across all the models within each association and across different associations. The β_1 values in the "shared parameters" and "current value plus slope" joint models are stable, but they show erratic behavior in the "current value" joint models, suggesting that they may not have converged well.

Table 4.11: Posterior Means for β Based on Different Baseline Hazard Function

Association	**Par.**	**11 knots**	**13 knots**	**15 knots**	**17 knots**
Shared Parameters	β_0	5.6268	5.6197	5.6323	5.6313
	β_1	-0.0571	-0.0674	-0.0607	-0.0618
Current Value	β_0	5.6293	5.6266	5.6301	5.6338
	β_1	-0.0602	-0.0459	-0.0619	-0.0473
Current Value + Slope	β_0	5.6314	5.6289	5.6235	5.6329
	β_1	-0.0535	-0.0582	-0.0551	-0.0549

Table 4.12 presents the posterior means for the strength of association between the longitudinal and survival submodels for the 12 models. The parameters in the joint models based on "shared parameters" and "current value" associations are consistent. However, in the joint models based on the "current value plus slope" association, an increase in the number of knots corresponds to a larger value of α_1 and α_2, particularly for α_2. This suggests that α_1 and α_2 are sensitive to the number of knots in the "current value plus slope" association joint models.

Table 4.12: Posterior Means for α Based on Different Baseline Hazard Function

Association	**Par.**	**11 knots**	**13 knots**	**15 knots**	**17 knots**
Shared Parameters	α_0	-0.7127	-0.6175	-0.6765	-0.7026
	α_1	-0.8585	-0.8350	-0.7421	-0.7955
Current Value	α_1	-0.5157	-0.4508	-0.4710	-0.5056
Current Value + Slope	α_1	-0.5030	-0.5130	-0.5348	-0.5729
	α_2	0.1361	0.1786	0.2751	-0.3962

The joint models based on "shared parameters" and "current value" associations are not sensitive to changes in the number of knots, while the joint model with the "current value

plus slope" association has sensitive parameters.

In the following, we analyze the sensitivity of the parameter estimates based on different prior distribution assumptions. The default value for the prior variance of all unknown parameters in the **jointModelBayes(·)** function is set to 100. As previously stated, if a non-informative prior is used, meaning that all unknown parameters follow uniform distributions, Bayesian methods will be comparable to likelihood methods. In other words, the larger the variance of the prior, the closer the results from these two methods will be. To investigate the impact of prior assumptions, 12 models are built, with each association building 4 models with prior variances equal to 10, 100, 1000, and 10000. Table 4.13 shows the posterior means for the fixed effects of the longitudinal submodel for these 12 models. The values of β_0 are consistent across all the models within each association and across different associations. The β_1 values in the "current value" joint models are stable, but show a little erratic behavior in the "shared parameters" and "current value plus slope" joint models. Table 4.14 represents the posterior means for the strength of association between the longitudinal and survival submodel for these 12 models. According to Table 4.14, the choice of prior variance does not significantly affect the inferences based on the "shared parameters" and "current value" association joint models. However, the parameter α_2 in the "current value plus slope" association joint models shows a noticeable difference with different prior variances, suggesting that the "current value plus slope" association joint model is sensitive to the choice of prior and the estimate of α_2 lacks credibility.

Table 4.13: Posterior Means for β Based on Different Prior variance

Association	**Par.**	**Var.= 10**	**Var.= 100**	**Var.= 1000**	**Var.= 10000**
Shared Parameters	β_0	5.6137	5.6323	5.6257	5.6235
	β_1	-0.0717	-0.0607	-0.0493	-0.0592
Current Value	β_0	5.6126	5.6301	5.6352	5.6343
	β_1	-0.0518	-0.0619	-0.0549	-0.0589
Current Value + Slope	β_0	5.6186	5.6235	5.6299	5.6339
	β_1	-0.04596	-0.0551	-0.0626	-0.0618

Table 4.14: Posterior Means for α Based on Different Prior variance

Association	**Par.**	**Var.= 10**	**Var.= 100**	**Var.= 1000**	**Var.= 10000**
Shared Parameters	α_0	-0.5872	-0.6765	-0.6837	-0.6162
	α_1	-0.3467	-0.7421	-1.1267	-1.3112
Current Value	α_1	-0.4529	-0.4710	-0.5216	-0.5199
Current Value + Slope	α_1	-0.4785	-0.5348	-0.4132	-0.6510
	α_2	0.1117	0.2751	-0.1484	0.5001

Trace Plots

Trace plots show the history of a parameter value across the iteration of the chain. It can be used to make sure that the prior distribution is well calibrated, which is indicated by the parameters having sufficient state changes as a MCMC algorithm runs. In each trace plot for one model parameter, the x-axis represents the iterations and the y-axis represents the value of the parameter we are interested in. Figure 6.1 to Figure 6.11 are trace plots for the three different joint models based on "shared parameters", "current value", and "current value plus slope" associations based on Bayesian methods, respectively. All the trace plots can be found in the Appendix.

Figure 6.1 to Figure 6.4 are trace plots for the "shared parameters" association joint model showing that β_0, β_1, and α_1 do not show any long-term trends and have roughly flat averages of the chains, appearing to follow a normal distribution. However, the trace plot of α_0 in Figure 6.3 shows a long-term trend, with different results obtained for the mean depending on the number of iterations used for estimation.

The trace plots for the "current value" association joint model in Figure 6.5 to Figure 6.7 show flat averages of the chains for the fixed parameters in the longitudinal submodel, following a normal distribution. However, the trace plot of α_1 in Figure 6.7 shows a long-term trend, indicating the need for more iterations.

The trace plots for the "current value plus slope" association model in Figure 6.8 to Figure 6.11 show flat averages of the chains for the fixed parameters in the longitudinal submodel, following a normal distribution. However, both α_1 and α_2 in the survival submodel show long-term trends, indicating the need for more iterations.

One or more unstable parameters are present in all three models based on Bayesian methods. Therefore, future analysis should consider using more iterations.

Autocorrelation Plots

Autocorrelation plots [24] are commonly used for checking randomness in a data set. An autocorrelation expressed as a number between -1 and 1, which measures the linear dependence of the current value in the chain to past values or lags. If the data is random, the autocorrelation should be near zero for all time-lag separations. This is important for Bayesian methods as it shows how much information is available from the Markov chain. Sampling 1000 iterations from a highly correlated Markov chain yields less information than we would obtain from 1000 independent samples drawn from the stationary distribution. The x-axis in an autocorrelation plot represents the past values or lags, and the y-axis represents the autocorrelation between the current value and the lags. The autocorrelation plots for the Bayesian joint models based on different associations are presented in the Appendix.

The autocorrelation plots for the parameters in the "shared parameters" joint model are

shown in Figure 6.12. The upper left panel shows the autocorrelation of β_0 in the longitudinal model. The upper right panel shows the autocorrelation of β_1 in the longitudinal model. The lower left panel shows the autocorrelation of α_0 in the survival model, and the lower right panel shows the autocorrelation of α_1 in the survival model. All the plots express randomness of the parameters except the lower left one, which shows that the value becomes less correlated as we go further along the chain. However, the correlation persists for many lags, even for the past 30 lags, meaning that there is less information available from the Markov chain to estimate the parameter α_0.

The autocorrelation plots for the parameters in the "current value" association joint model are shown in Figure 6.13. The upper left and right panels show the autocorrelation of β_0 and β_1 in the longitudinal model, respectively. The lower left panel shows the autocorrelation of α_1, which measures the strength of correlation between the longitudinal and survival models. The top two plots express the parameters are random, while the lower one shows a decrease in correlation as we go further along the chain. However, the correlation persists for many lags, even for the past 30 lags, meaning that there is less information available from the Markov chain to estimate the parameter α_1.

The autocorrelation plots for the parameters in the "current value plus slope" association joint model are shown in Figure 6.14. The upper left and right panels and the lower left panel depict the same as in the "current value" association joint model. The difference is the lower right panel, which shows the autocorrelation of α_2. The top two plots express the parameters are random, while the lower left one shows a decrease in correlation as we go further along the chain. However, the correlation persists at a high level for many lags, even for the past 30 lags, meaning that there is less information available from the Markov chain to estimate the parameter α_1. The lower right panel has a relatively low correlation value for many lags, meaning that there is also limited information available from the Markov chain to estimate the parameter α_2.

Kernel Density Estimation Plots

Kernel Density Estimation (KDE) solves a crucial problem in data smoothing, as it uses a finite data sample to make inferences about a population. The resulting plots depict the Probability Density Function (PDF) of continuous or non-parametric data variables. The KDE plots for Bayesian joint models based on different association are shown in the Appendix.

Figure 6.15 displays the KDE plots for the parameters in the "shared parameters" association joint model. The upper left panel shows the PDF of β_0 in the longitudinal model, which follows a normal distribution. Similarly, the upper right panel displays the PDF of β_1 in the longitudinal model, also following a normal distribution. The lower left panel shows the PDF of α_0 in the survival model, which also follows a normal distribution. The lower right panel displays the PDF of α_1 in the survival model, following a normal distribution. However, the density of α_0 has a relatively smooth peak.

Figure 6.16 presents the KDE plots for the parameters in the "current value" association joint model. The upper two panels are the same as in the "shared parameters" association joint model. The lower left panel displays the PDF of α_1, which follows a perfect normal distribution.

Figure 6.17 shows the KDE plots for the parameters in the "current value plus slope" association joint model. The upper two and lower left panels are the same as in the "current value" association joint model. The lower right panel displays α_2. The lower left panel, which shows the PDF of α_1, does not follow a normal distribution. As seen in the trace plot and autocorrelation plot, α_1 in the "current value plus slope" association joint model does not converge and is not estimated well.

4.6 A Comparison of Bayesian and Likelihood Joint Models

The preceding sections have explained the optimal associations based on Bayesian and likelihood methods. As a reminder, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are employed to compare likelihood joint models, while the Deviance Information Criterion (DIC) is used for comparing Bayesian joint models. Both methods lead to the conclusion that the "current value" association joint model is the most suitable model for this oral cancer data. In this section, we compare likelihood and Bayesian joint models based on the "current value" and "current value plus slope" association. The results are summarized in Table 4.15.

The parameters estimates are similar, except α_2 in the "current value plus slope" association. As discussed in **Section 4.5.4**, α_2 in Bayesian joint model based on the "current value plus slope" association is sensitive to the choice of prior and the number of knots, and its computation fails to converge well. The standard errors of the estimates based on Bayesian joint models and likelihood joint models differ significantly due to the differing definitions of standard errors in frequentist and Bayesian statistics. The value of

residual standard errors are very similar in these four models, while it is relatively small in the Bayesian joint model based on the "current value" association.

Table 4.15: Likelihood Joint Models and Bayesian Joint Models Based on Different Associations

		The Likelihood Method			**The Bayesian Method**		
Association	**Par.**	**Est.**	**S.E.**	**P-value**	**Est.**	**S.E.**	**C.I.**
Current Value	α_1	-0.3943	0.2747	0.1512	-0.4710	0.0309	$(-0.9612, 0.0157)$
	σ	0.5957	-	-	0.5744	-	-
Current Value + Slope	α_1	-0.7657	0.4638	0.0988	-0.5348	0.0857	$(-1.048, 0.0319)$
	α_2	3.9933	4.1546	0.3365	0.2751	0.1869	$(-2.4652, 2.6918)$
	σ	0.5970	-	-	0.5748	-	-

Par.: parameter, Est.: estimation, S.E.: standard error, C.I.: Credible Interval

5 Conclusion

This report is based on an oral cancer dataset. Our goal is to find the factors that may be correlated with the time it takes to develop oral cancer. To effectively analyze the longitudinal and survival data, we consider joint modeling for longitudinal and survival data. We assume that the longitudinal submodel provides certain characteristics that are used in combination with the survival submodel to form various joint models. Two inference methods are popular for joint modeling of longitudinal and survival data: likelihood methods and Bayesian methods. Compared with likelihood methods, Bayesian methods combine prior information with the observed data to guide statistical inference. Bayesian methods treat the model parameters as random variables with known prior distributions. It is worth noting that Bayes factors are related to likelihood ratios and can be considered equivalent to likelihood ratios under certain conditions, such as when the prior distribution is uniform. However, Bayesian methods provide more information than point estimators produced by maximum likelihood estimation (MLE) and can lead to more complex calculations. MLE produces a single fixed value, while Bayesian methods return a probability density (or mass) function. In practice, these two methods are implemented using different packages and programs, making it challenging to compare the performance and accuracy of models estimated using likelihood and Bayesian methods.

In this report, all the joint models are based on Bayesian and likelihood methods, using different association structures. Likelihood joint models are built using the **jointModel(·)** function, based on "current value" and "current value plus slope" associations, while Bayesian joint models are built using the **jointModelBayes(·)** function, based on "current value," "current value plus slope," and "shared parameters" associations. Model comparison is performed using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for the likelihood joint models, and the Deviance Information Criterion (DIC) for the Bayesian joint models. The results indicate that the average oral lesion area starts at around 5.6mm and does not change significantly over time. However, there is a relatively large individual difference, which requires a mixed-effects model to explore individual-specific effects. All of these joint models show a negative association between the longitudinal process and the event of interest, although the result is not statistically significant. These aforementioned results can help physicians personalize treatment plans for patients. Generally, the joint models based on the "current value" association outperform the others for both the likelihood and Bayesian methods. However, as mentioned earlier, comparing the performance and accuracy of likelihood and Bayesian joint models is challenging. Therefore, we focus on comparing the parameter estimates, diagnosing the models, and evaluating the residual standard errors. The results reveal that based on the "current value" association, these two methods provide similar parameter estimates, while based on the "current value plus slope" association, the estimates of α_2 differ due to convergence issues. Nevertheless, these two methods show very similar residual standard errors.

References

[1] H. Williams, "Molecular pathogenesis of oral squamous carcinoma," *Molecular Pathology*, vol. 53, no. 4, p. 165, 2000.

[2] A. C. Society, "Key statistics for oral cavity and oropharyngeal cancers," 2019.

[3] A. N. Vu and C. S. Farah, "Efficacy of narrow band imaging for detection and surveillance of potentially malignant and malignant lesions in the oral cavity and oropharynx: a systematic review," *Oral Oncology*, vol. 50, no. 5, pp. 413–420, 2014.

[4] R. A. Levine and G. Casella, "Implementations of the monte carlo em algorithm," *Journal of Computational and Graphical Statistics*, vol. 10, no. 3, pp. 422–439, 2001.

[5] D. Ruppert and D. S. Matteson, "Bayesian data analysis and mcmc," in *Statistics and Data Analysis for Financial Engineering*, pp. 581–644, Springer, 2015.

[6] A. Saracco, M. Musicco, A. Nicolosi, G. Angarano, C. Arici, G. Gavazzeni, P. Costigliola, S. Gafa, C. Gervasoni, R. Luzzati, *et al.*, "Man-to-woman sexual transmission of hiv: longitudinal study of 343 steady partners of infected men," *JAIDS Journal of Acquired Immune Deficiency Syndromes*, vol. 6, no. 5, pp. 497–502, 1993.

[7] A. L. Oberg and D. W. Mahoney, "Linear mixed effects models," *Topics in biostatistics*, pp. 213–234, 2007.

[8] C.-F. Chung, P. Schmidt, and A. D. Witte, "Survival analysis: A survey," *Journal of Quantitative Criminology*, vol. 7, no. 1, pp. 59–98, 1991.

[9] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[10] C. L. Faucett and D. C. Thomas, "Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach," *Statistics in medicine*, vol. 15, no. 15, pp. 1663–1685, 1996.

[11] M. S. Wulfsohn and A. A. Tsiatis, "A joint model for survival and longitudinal data measured with error," *Biometrics*, pp. 330–339, 1997.

[12] D. Van Ravenzwaaij, P. Cassey, and S. D. Brown, "A simple introduction to markov chain monte–carlo sampling," *Psychonomic bulletin & review*, vol. 25, no. 1, pp. 143–154, 2018.

[13] J. Miles, "Wiley statsref: Statistics reference online," *R squared, adjusted R squared, edited by N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and JL Teugels. Chichester, UK: Wiley*, 2014.

[14] M. Datta, D. M. Laronde, M. P. Rosin, L. Zhang, B. Chan, and M. Guillaud, "Predicting progression of low-grade oral dysplasia using brushing-based dna ploidy and chromatin organization analysispredicting progression in oral dysplasia using dna ploidy," *Cancer Prevention Research*, vol. 14, no. 12, pp. 1111–1118, 2021.

[15] M. Radwan-Oczko, K. Bandosz, Z. Rojek, and J. E. Owczarek-Drabińska, "Clinical study of oral mucosal lesions in the elderly—prevalence and distribution," *International Journal of Environmental Research and Public Health*, vol. 19, no. 5, p. 2853, 2022.

[16] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[17] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected papers of hirotugu akaike*, pp. 199–213, Springer, 1998.

[18] C. Proust-Lima, V. Philipps, A. Diakite, and B. Liquet, "lcmm: Extended mixed models using latent classes and latent processes," *R package version*, vol. 1, no. 7, 2017.

[19] S. M. Schwartz, D. R. Doody, E. D. Fitzgibbons, S. Ricks, P. L. Porter, and C. Chen, "Oral squamous cell cancer risk in relation to alcohol consumption and alcohol dehydrogenase-3 genotypes," *Cancer Epidemiology Biomarkers & Prevention*, vol. 10, no. 11, pp. 1137–1144, 2001.

[20] F. E. Harrell, "Cox proportional hazards regression model," in *Regression modeling strategies*, pp. 475–519, Springer, 2015.

[21] R. A. Huddart, E. Hall, S. A. Hussain, P. Jenkins, C. Rawlings, J. Tremlett, M. Crundwell, F. A. Adab, D. Sheehan, I. Syndikus, *et al.*, "Randomized noninferiority trial of reduced high-dose volume versus standard volume radiation therapy for muscle-invasive bladder cancer: results of the bc2001 trial (cruk/01/004)," *International Journal of Radiation Oncology* Biology* Physics*, vol. 87, no. 2, pp. 261–269, 2013.

[22] R. G. Lomax and D. L. Hahs-Vaughn, *Statistical Concepts-A Second Course*. Routledge, 2013.

[23] Y.-H. Chen, "Computationally efficient monte carlo em algorithms for generalized linear mixed models," *Journal of Statistical Computation and Simulation*, vol. 76, no. 9, pp. 817–828, 2006.

[24] G. E. Box and G. M. Jenkins, "Time series analysis: Forecasting and control san francisco," *Calif: Holden-Day*, 1976.

6 Appendix

6.1 R code

R code for data processing, visualization, and data analysis on the Oral Cancer Dataset can be found on my GitHub repository.

6.2 Plots

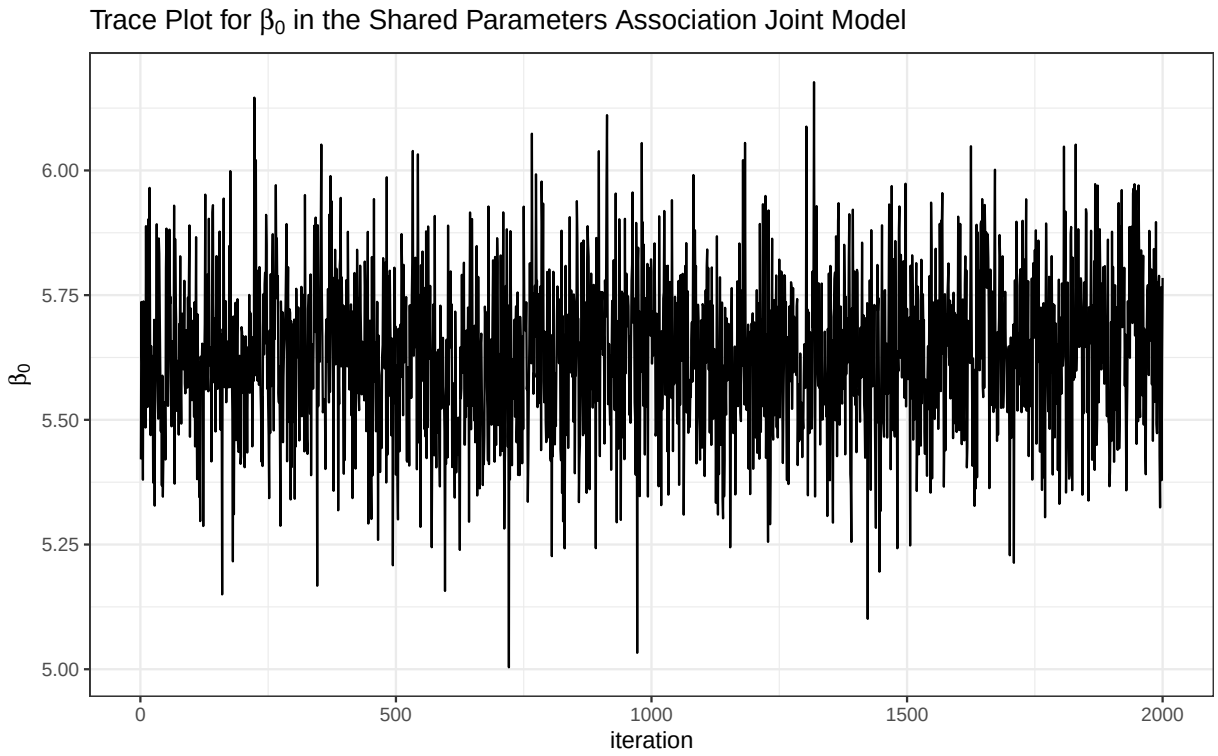

Figure 6.1: Trace Plot for β_0 of the Longitudinal Model in "Shared Parameters" Association Joint Model

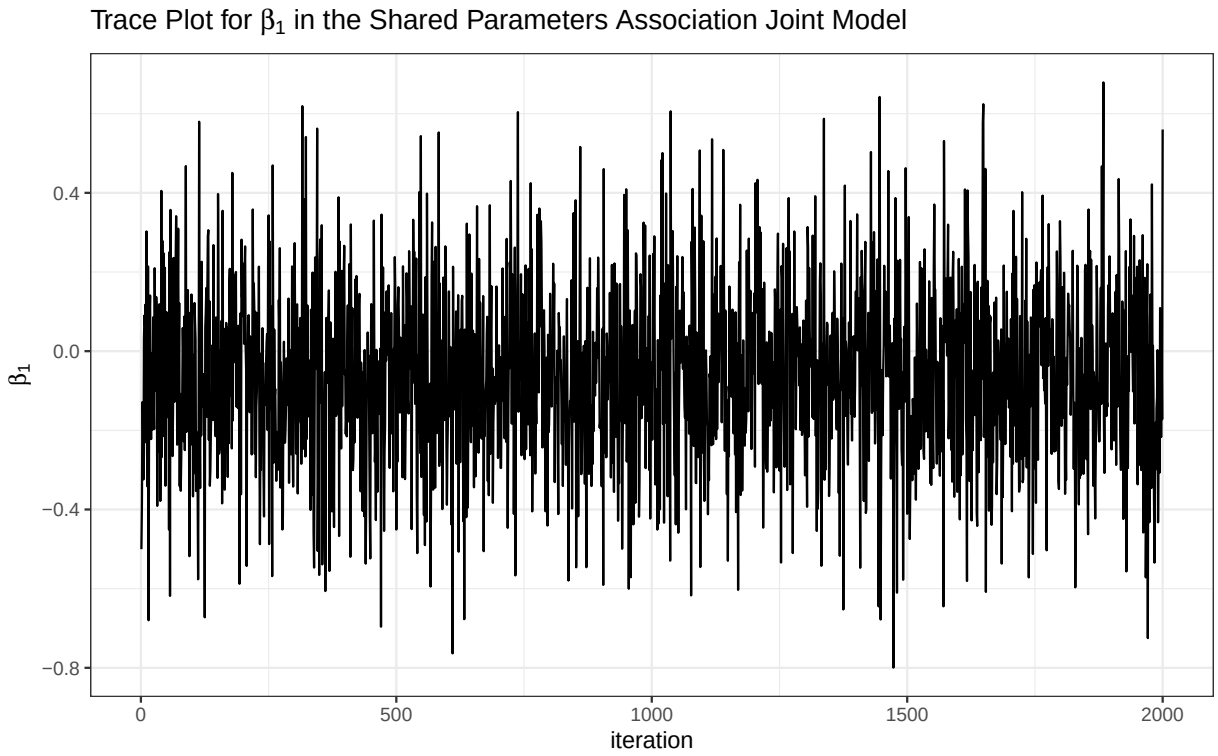

Figure 6.2: Trace Plot for β_1 of the Longitudinal Model in "Shared Parameters" Association Joint Model

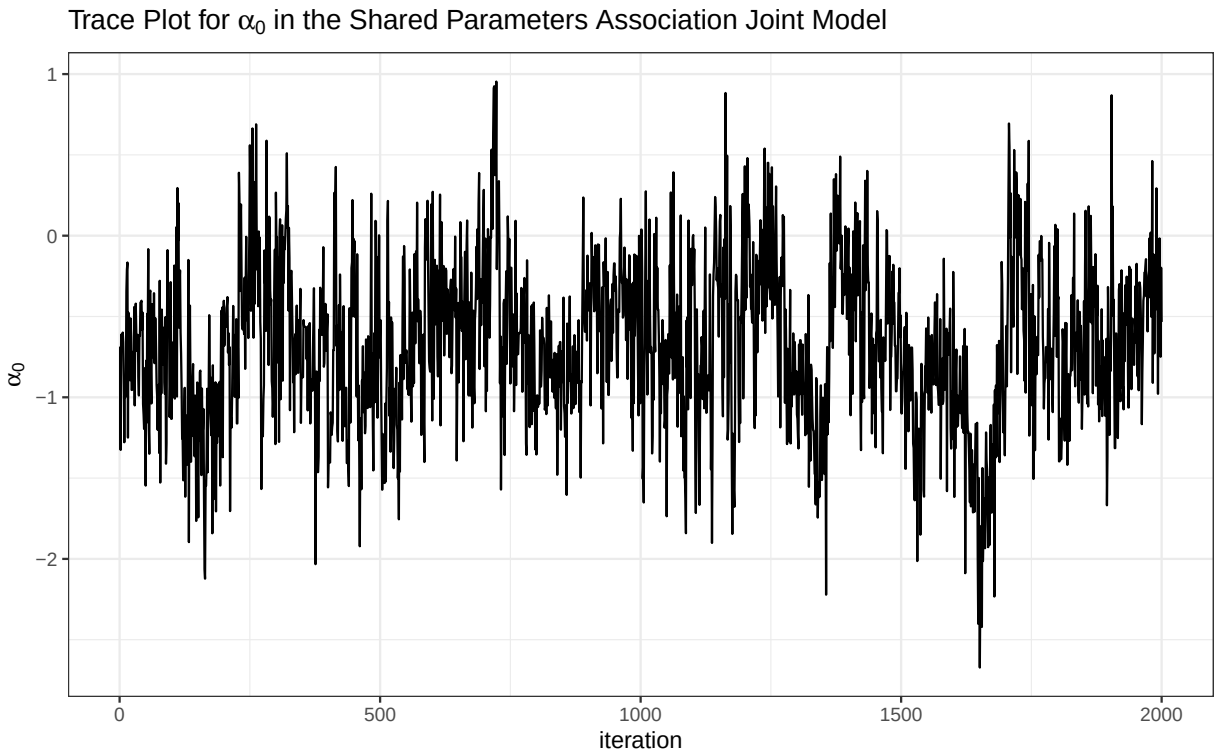

Figure 6.3: Trace Plot for α_0 of the Survival Model in "Shared Parameters" Association Joint Model

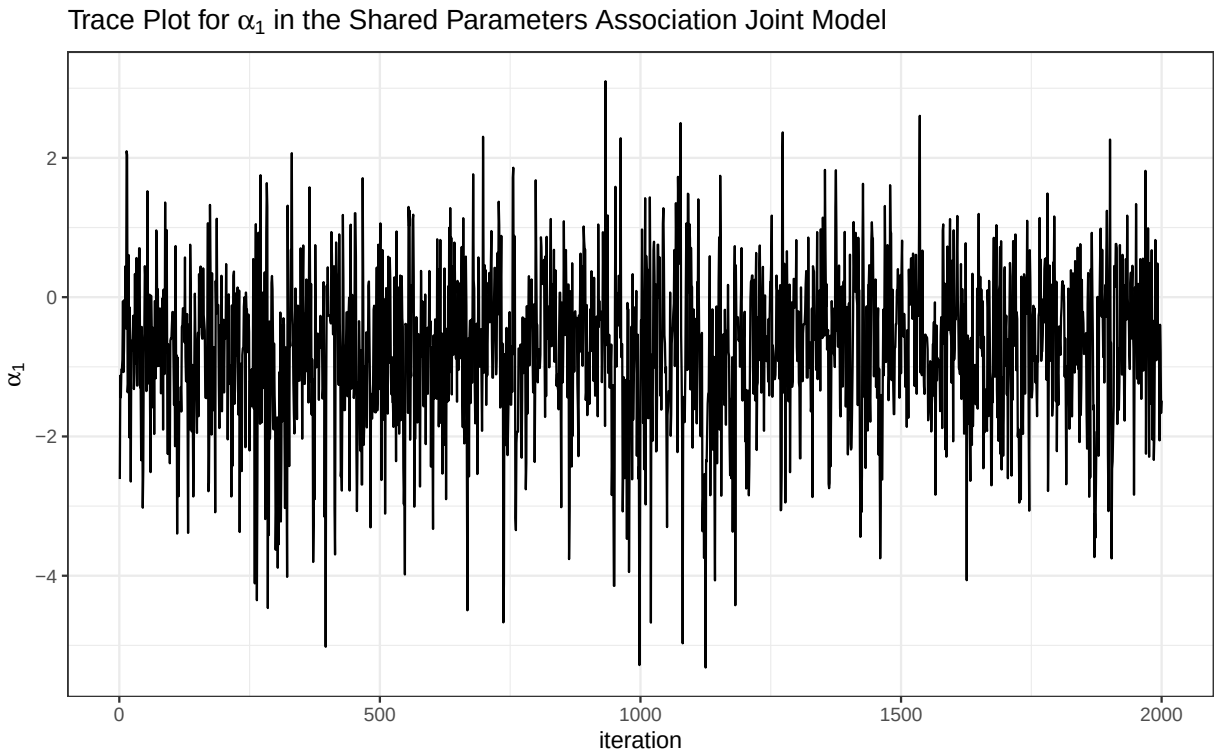

Figure 6.4: Trace Plot for α_1 of the Survival Model in "Shared Parameters" Association Joint Model

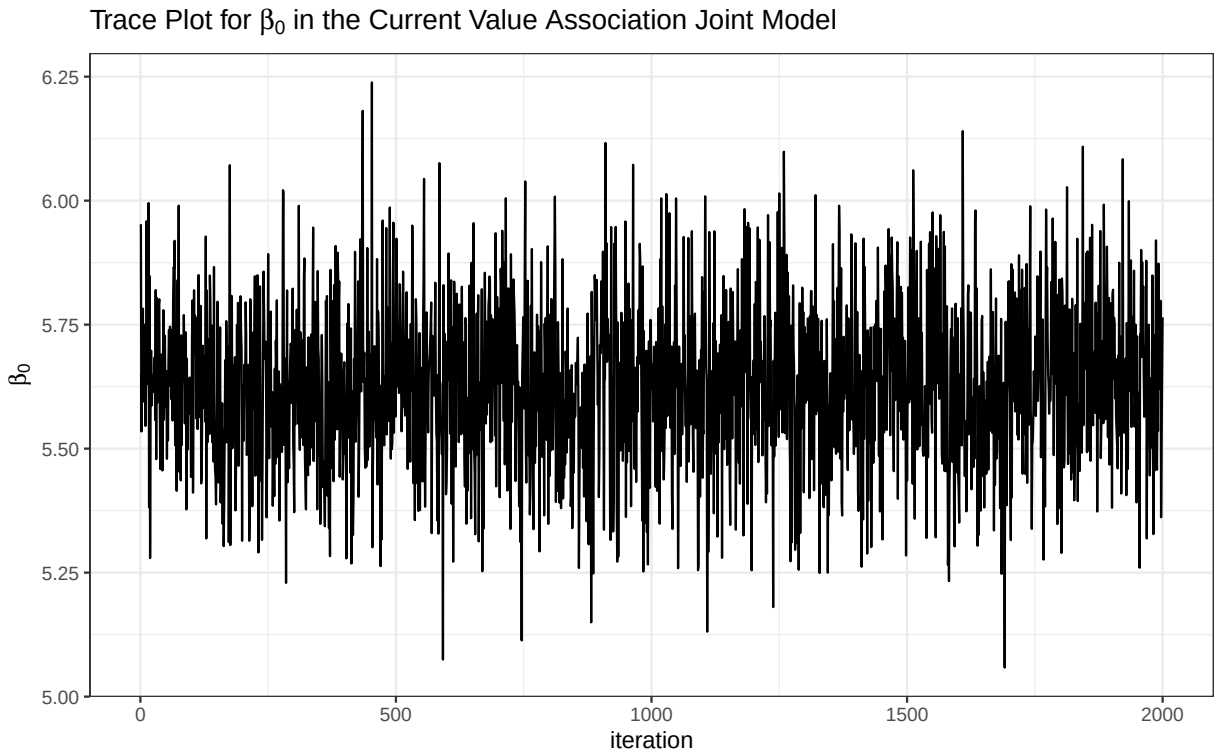

Figure 6.5: Trace Plot for β_0 of the Longitudinal Model in "Current Value" Association Joint Model

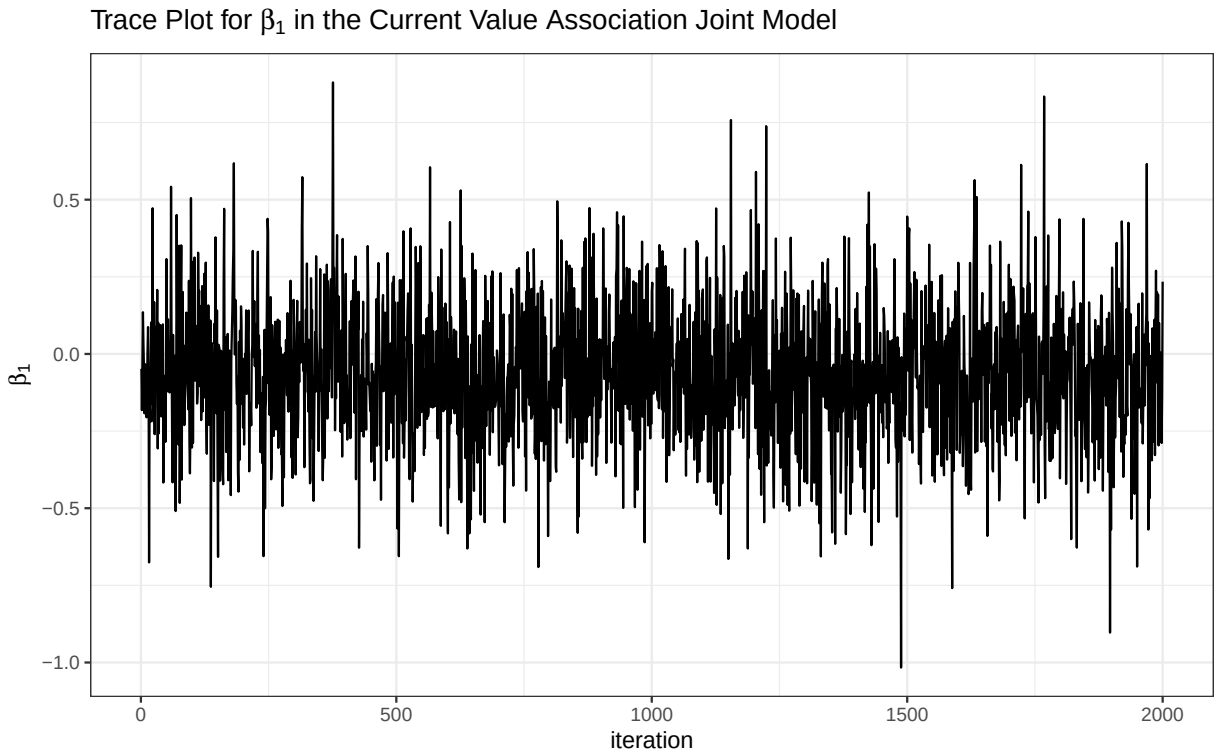

Figure 6.6: Trace Plot for β_1 of the Longitudinal Model in "Current Value" Association Joint Model

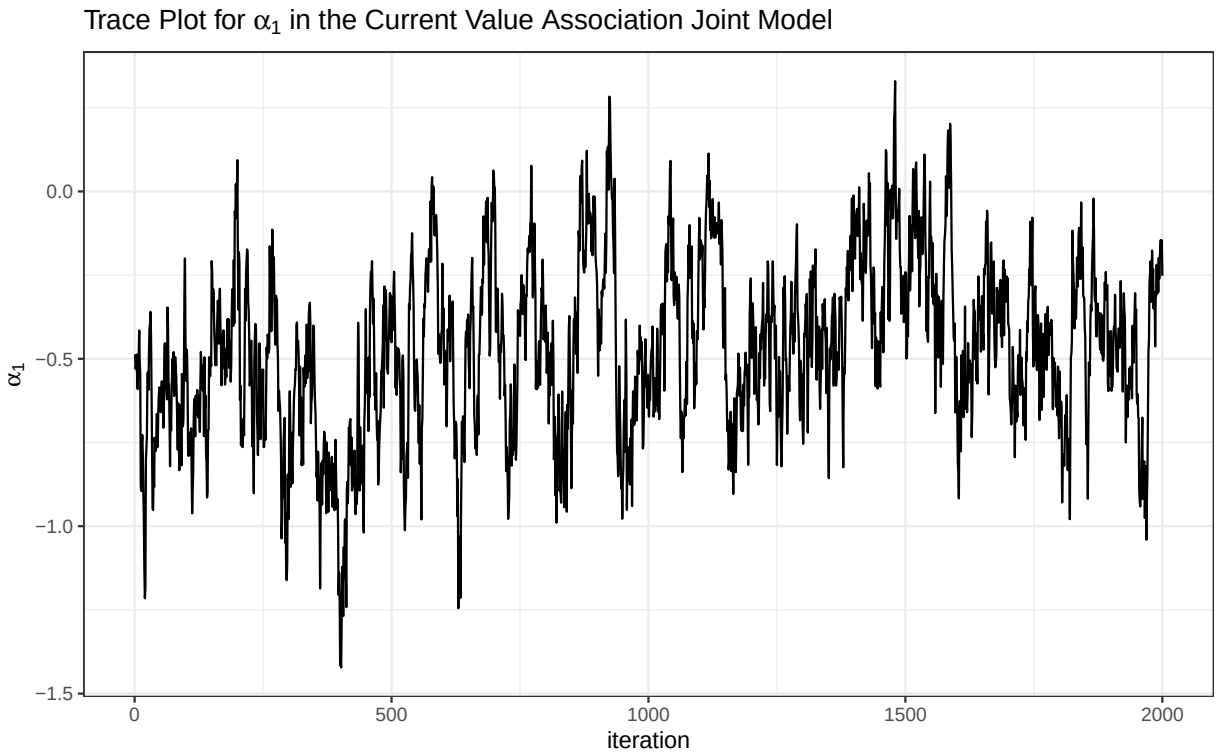

Figure 6.7: Trace Plot for α_1 of the Survival Model in "Current Value" Association Joint Model

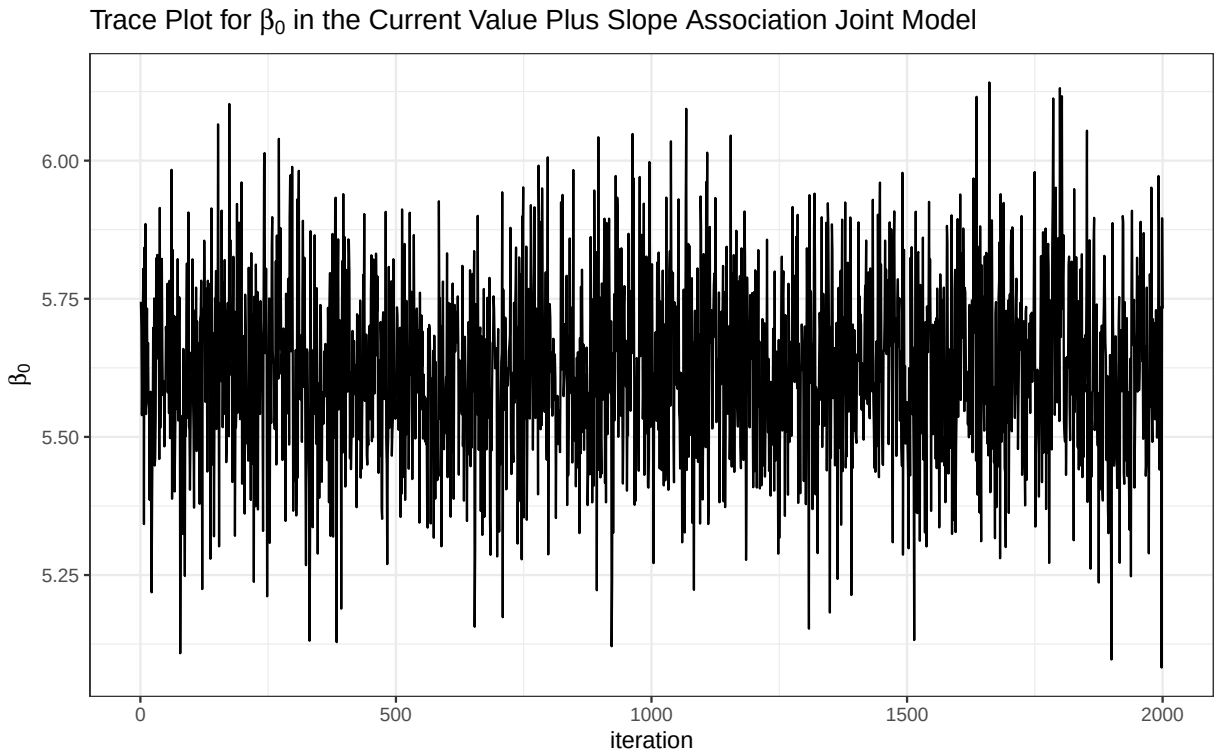

Figure 6.8: Trace Plot for β_0 of the Longitudinal Model in "Current Value Plus Slope" Association Joint Model

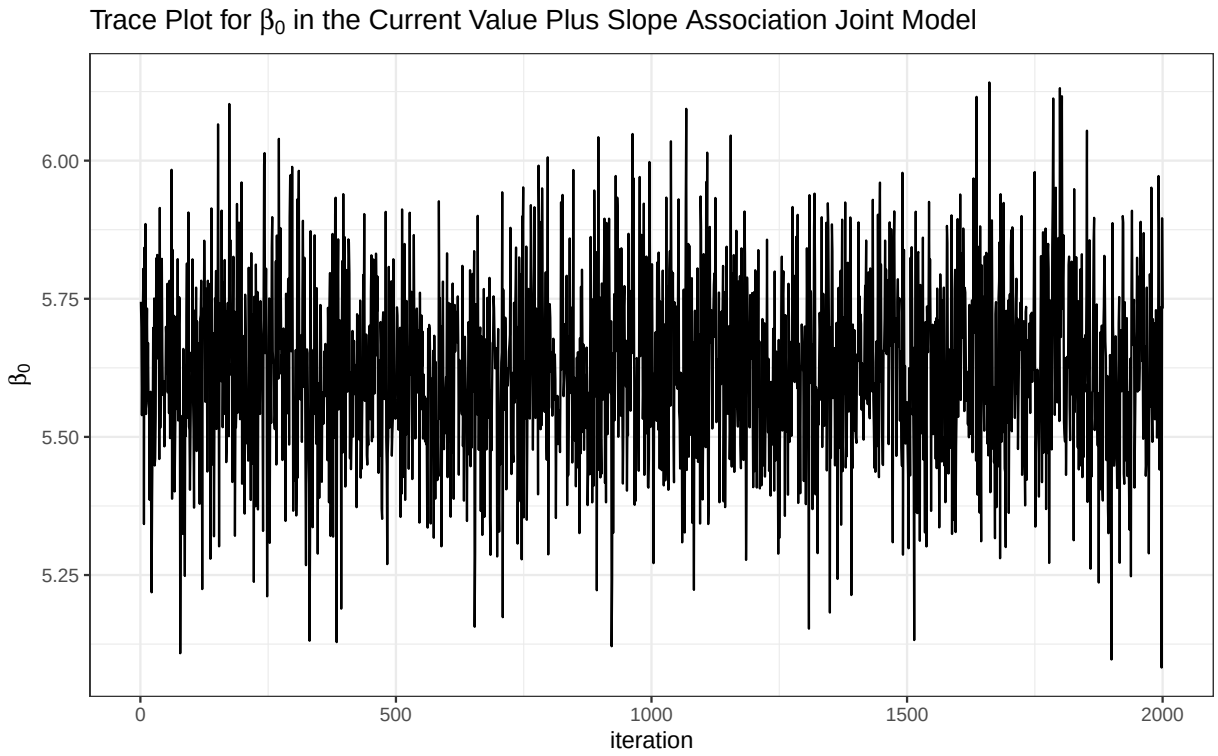

Figure 6.9: Trace Plot for β_1 of the Longitudinal Model in "Current Value Plus Slope" Association Joint Model

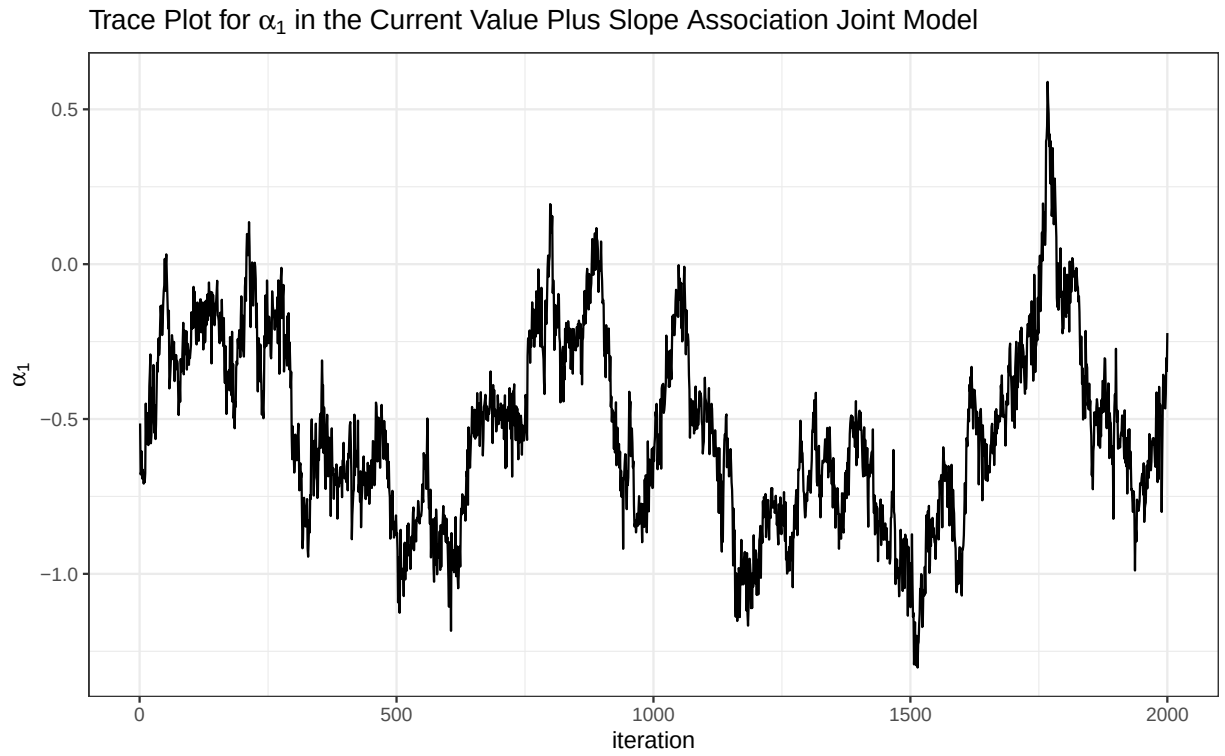

Figure 6.10: Trace Plot for α_1 of the Survival Model in "Current Value Plus Slope" Association Joint Model

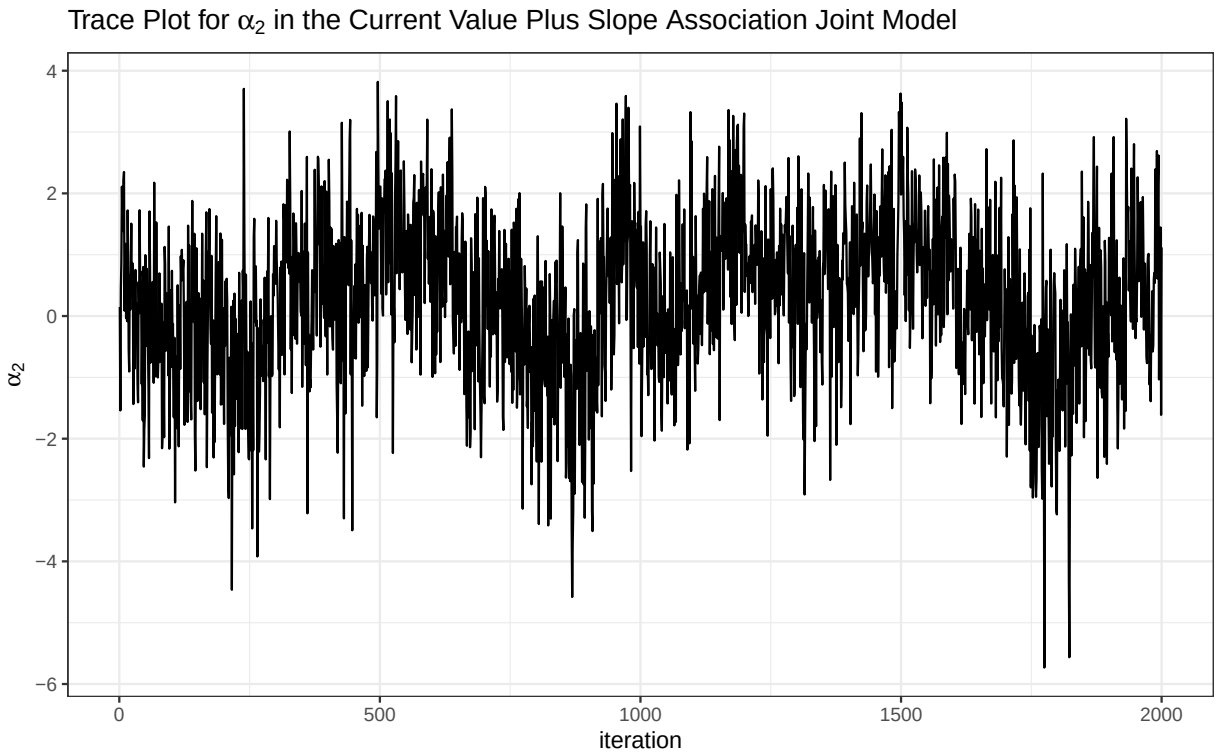

Figure 6.11: Trace Plot for α_2 of the Survival Model in "Current Value Plus Slope" Association Joint Model

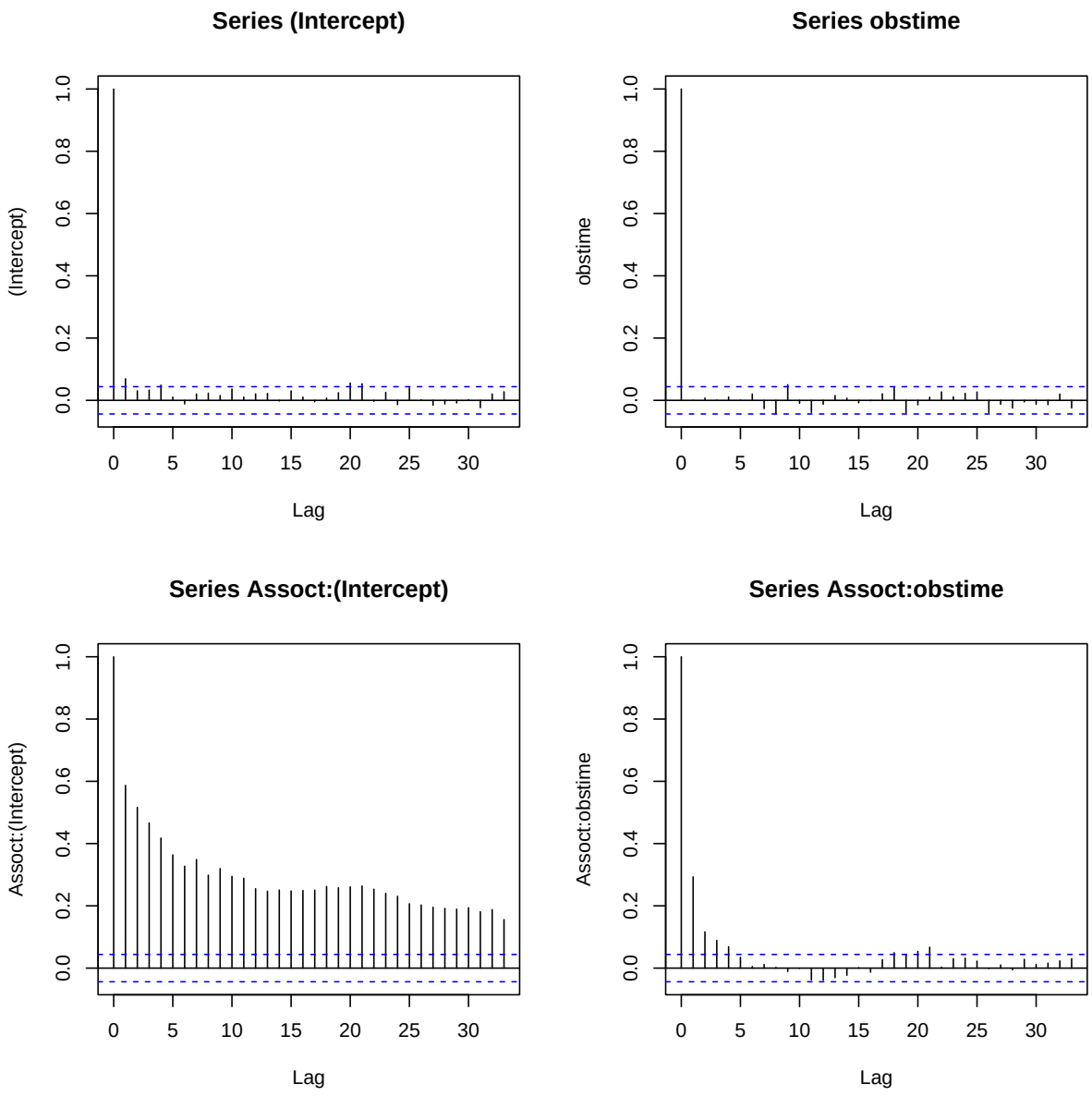

Figure 6.12: Autocorrelation Plots for the Parameters in "Shared Parameters" Joint Model

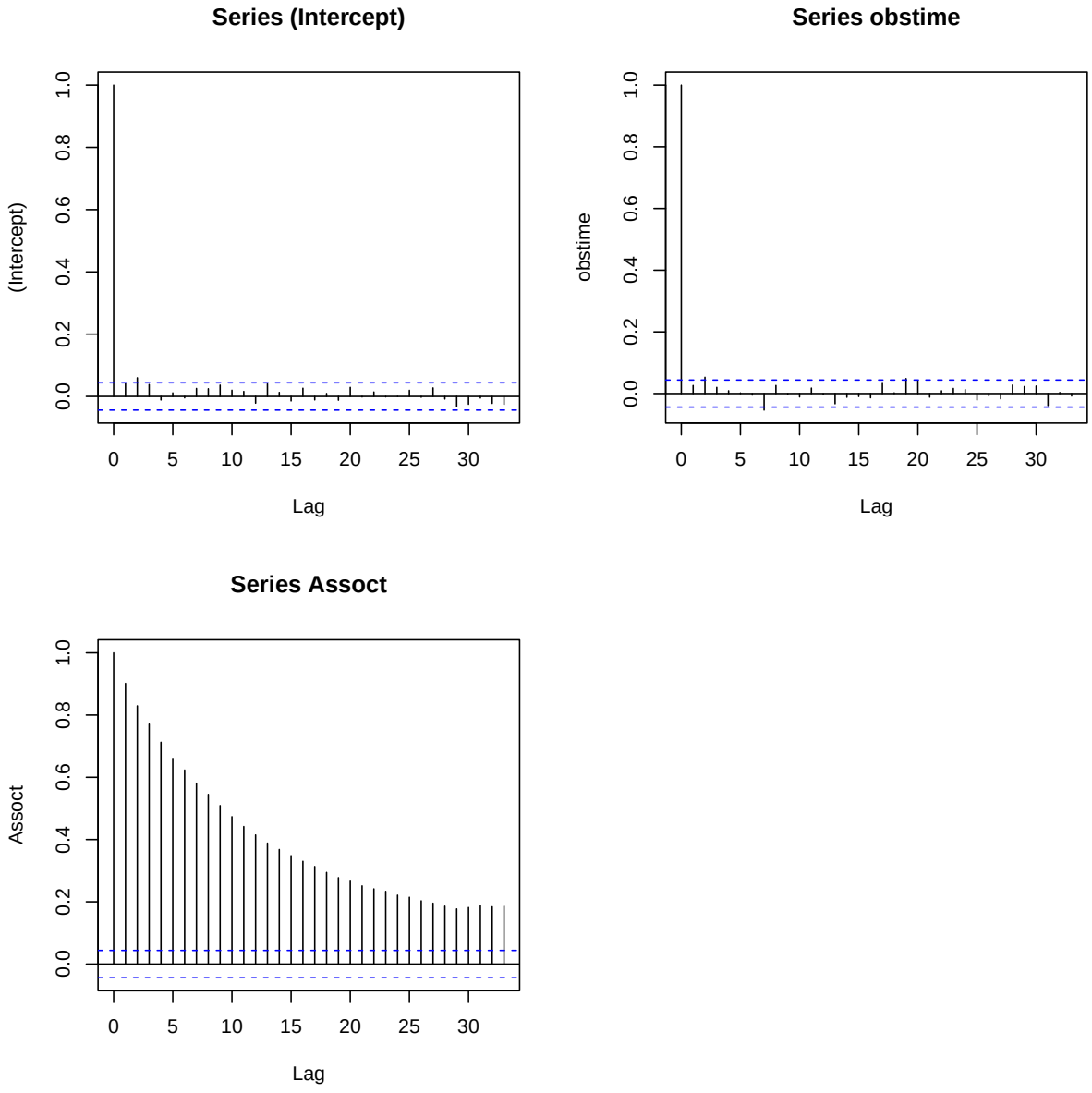

Figure 6.13: Autocorrelation Plots for the Parameters in the "Current Value" Association Joint Model

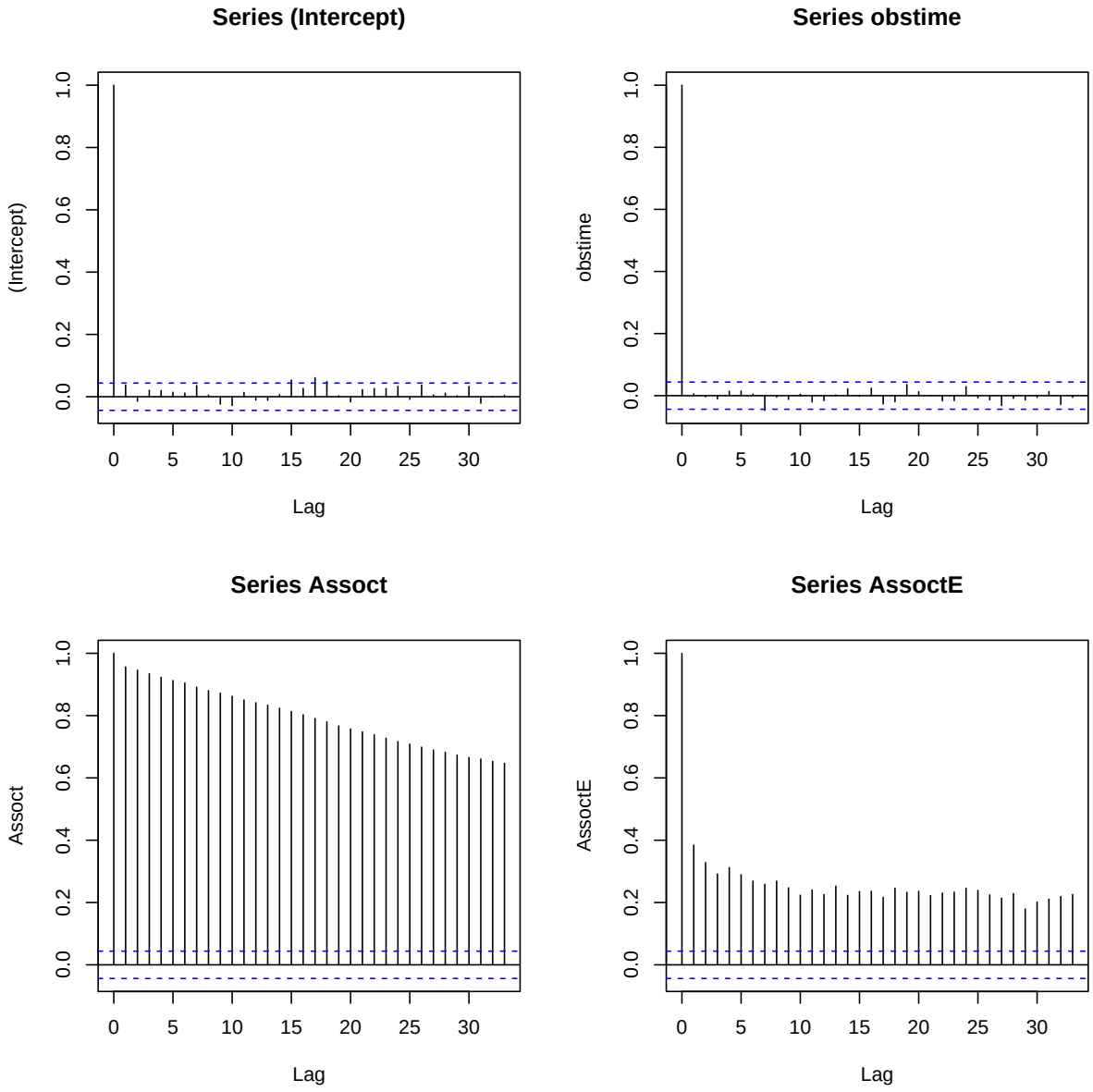

Figure 6.14: Autocorrelation Plots for the Parameters in the "Current Value Plus Slope" Association Joint Model

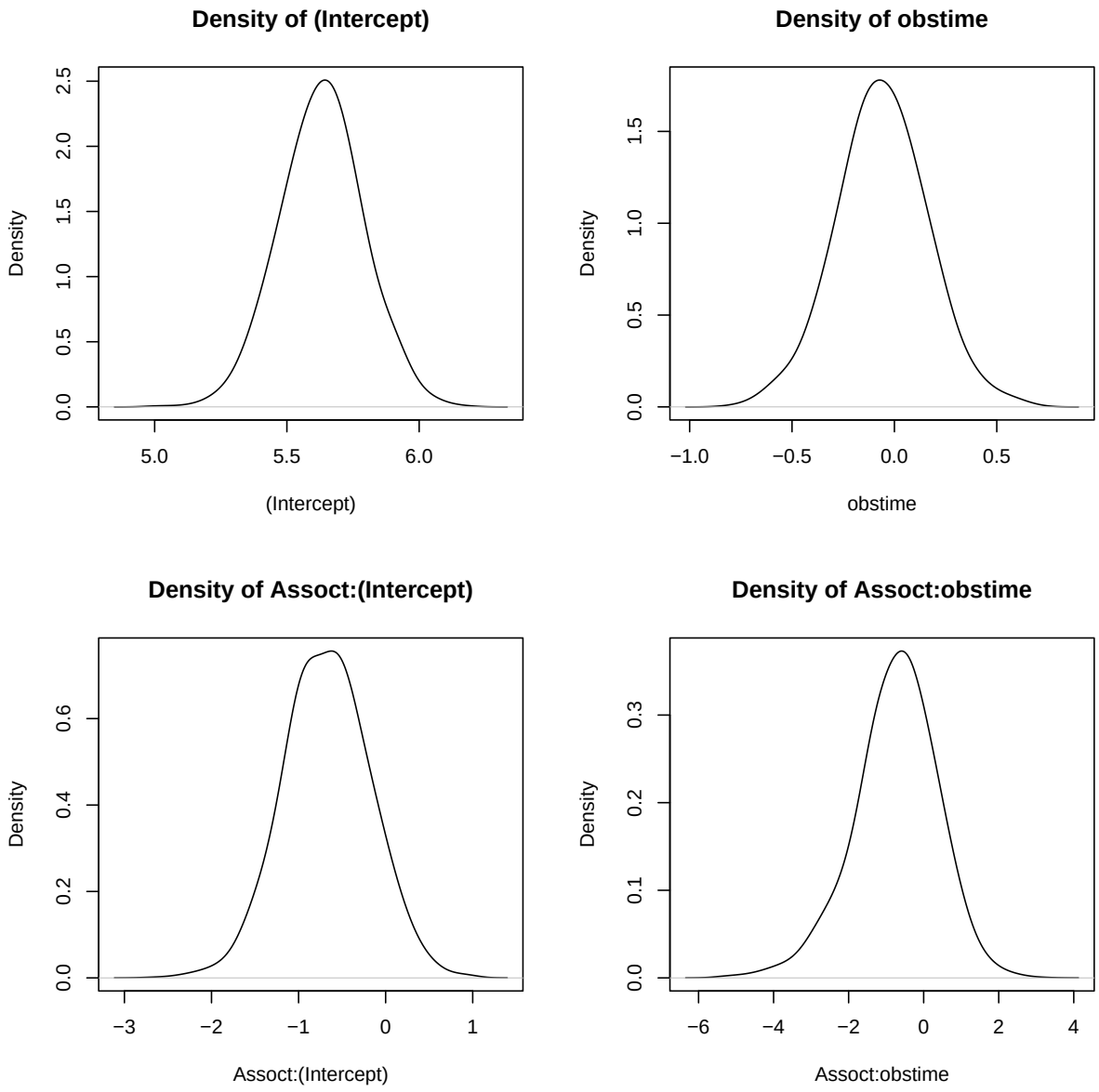

Figure 6.15: Trace Plot for the Parameters of the Longitudinal Model in "Shared Parameters" Joint Model

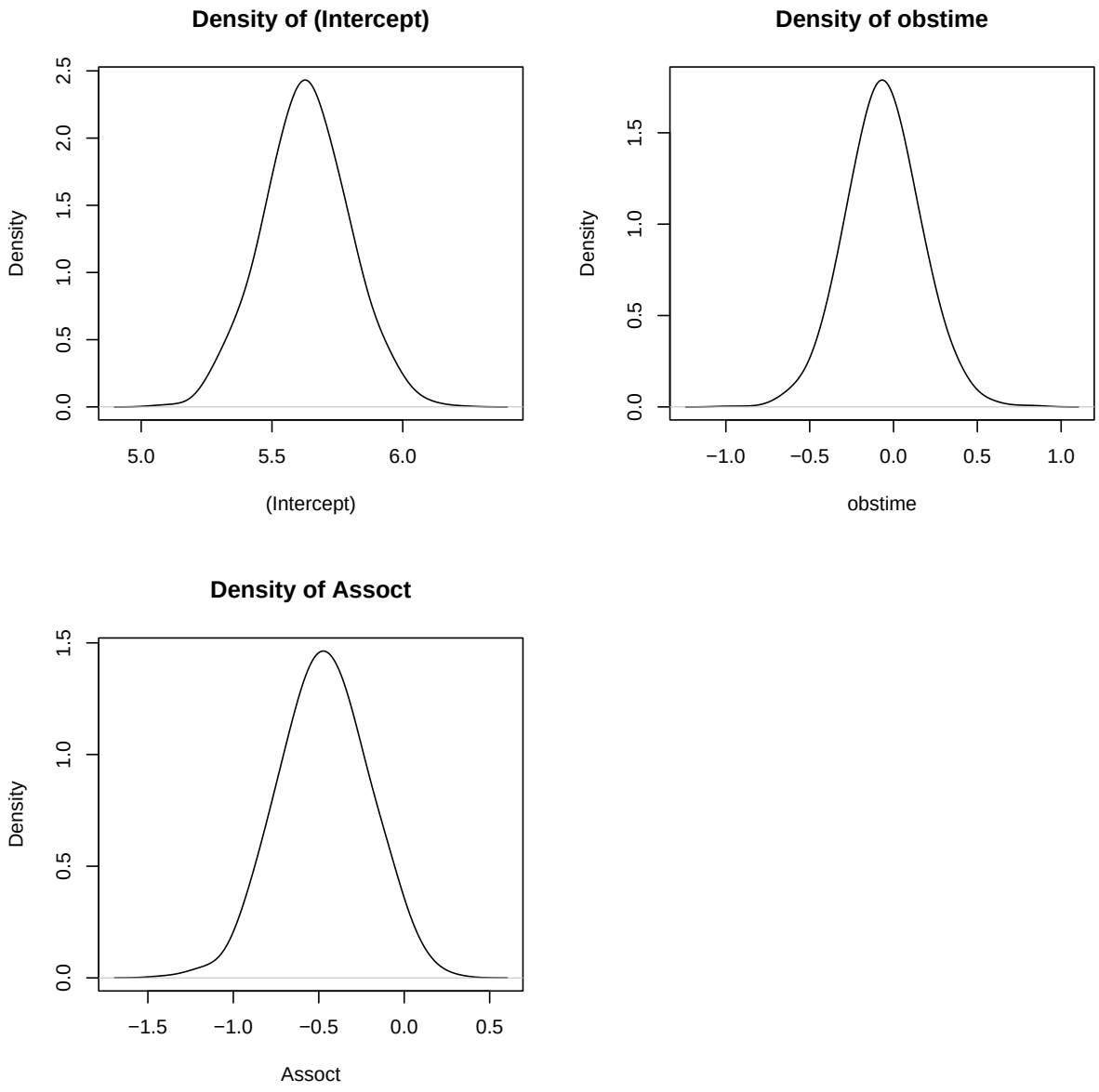

Figure 6.16: Trace Plot for the Parameters of the Longitudinal Model in the "Current Value" Association Joint Model

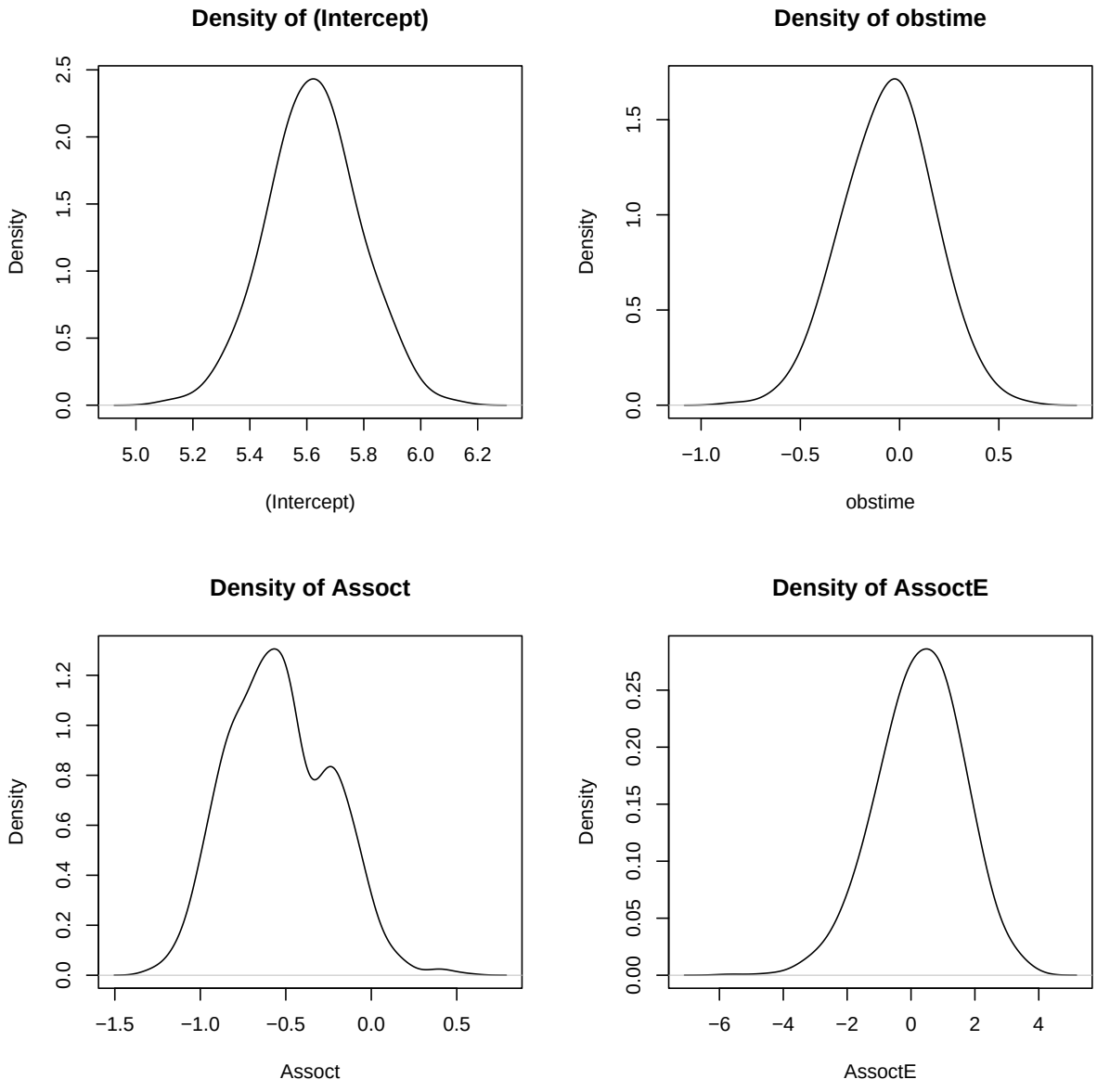

Figure 6.17: Trace Plot for the Parameters of the Longitudinal Model in the "Current Value Plus Slope" Association Joint Model