Comparison of Bayesian Joint Model and Likelihood Joint Model for Longitudinal and Survival Data, with Application to Oral Cancer Data

Jintong Yan

February 27, 2023

Abstract

Oral cancer comprises 2%–4% of all cancer cases worldwide. To investigate the relationship between time-dependent and time-independent predictors and the development of oral malignant transformation at the end of a study, joint modeling of longitudinal and survival analysis is required. This report focuses on the application and comparison of joint models estimated by 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 by utilizing the *JM* package and the *JMbayes* package, corresponding to the likelihood and Bayesian implementation, respectively. In this dataset, oral lesion area is longitudinal data collected repeatedly for each subject, and the occurrence of malignant transformation is survival data. Several link structures are used to explain the association between longitudinal and survival analyses. Model diagnosis is conducted using Q-Q plots and residual plots for joint models with likelihood inferences. For joint models with Bayesian inferences, sensitivity analyses, trace plots, and other common diagnostic diagrams are used. After comparing the joint model results based on likelihood and Bayesian inferences, it is concluded that the estimates of parameters are similar in the 'current value' association but different in the 'current value plus slope' association. The residual standard errors of joint models estimated by the Bayesian and likelihood methods are very similar. The 'current value' association is deemed the most appropriate structure for this data, although the parameter that explains the strength of the correlation between the longitudinal and survival processes needs more iterations to converge well.

Keywords: joint model, longitudinal, survival, association structures, *JMbayes* package, *JM* package, application.

Contents

List of Figures

List of Tables

1 Introduction

In biostatistics and medical research, it is very common to record repeated observations for each subject, along with dichotomous indicators to mark the time at which an event of interest occurs. The data collected repeatedly over time for each individual are called longitudinal data, in which the repeated measurements of a variable on each individual are likely to be correlated. Hence, one major consideration for analyzing longitudinal data is to incorporate the correlation of the repeated measurements to explain the within-subject correlation. Some common methods that can be used to deal with this include linear mixed-effects (LME) models, nonlinear mixed effects (NLME) models, and generalized linear mixed models (GLMMs). The data recorded by binary indicators to mark at which time the event of interest occurs are called survival data, in which not only the binary outcome but also the time that one subject experiences such an event are collected. Hence, when analyzing such cases, neither logistic analysis for fitting the binary indicators nor linear model for fitting the time is appropriate. Some common survival analyses can be carried out for such questions, like the Kaplan-Meier model, the Exponential model, the Weibull model, and the Cox Proportional-Hazards model.

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, clinical risk indicators like size, appearance, and site are important to determine the cancer risk of OPMLs.

The data we have are recorded in two separate datasets, namely, *cancer_old* and *cancer_old.id*. The former is in a long format with several observations of oral lesion area for each individual, while the latter contains a single row per patient and mainly saves the time-independent variables and the time to experience cancer. Specifically, *cancer_old* contains 408 observations for 41 patients. Among these 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 used in this report were collected 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.

lesion_area is a multiple of the lesion width and length. 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

Table 1.1: Name, Type and Description of The Variables in These Two Datasets

Name	**Type**	**Description**
studyID	Factor	Patient identifier in the study
ID	Factor	Another patient identifier, ranging from 1 to 41

bxdiagnosis	Character	Sequential stage of OPMLs coded as the following:
		H = hyperkeratosis;
		D1 = mild dysplasia;
		D2 = moderate dysplsia;
		D3 = severe dysplasia;
		SCC = squamous cell carcinoma;
		VC = verruous carcinoma.
		(Note that D3, SSC, and VC means progressors,
		while others means non-progressors.)

months	Numerical	Observation period in months start from the study to the measurements
Age	Numerical	The 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	Observation period in years start from the study to the measurements
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): Blank = no data and 0 = lesion left
aneuploid	Numerical	The number of aneuploid cells
Cycling %	Numerical	The percentage of cycling cells
Tetraploid %	Numerical	The percentage of tetraploid cells
Proliferation %	Numerical	The percentage of cancer cell proliferation
type	Character	The diagnostic results coded as progressors and non-progressors
progression	Factor	The diagnostic results: 1 = progressors and 2 = non-progressors

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.

Based on the given background, this report mainly explores the factors that affect the occurrence of oral cancer based on the OSCC datasets. To be specific, both time-varying and time-independent variables are seen as predictors to model the occurrence of oral cancer. Due to the small sample size and the presence of both longitudinal and survival data, appropriate methods must be employed to address this question.

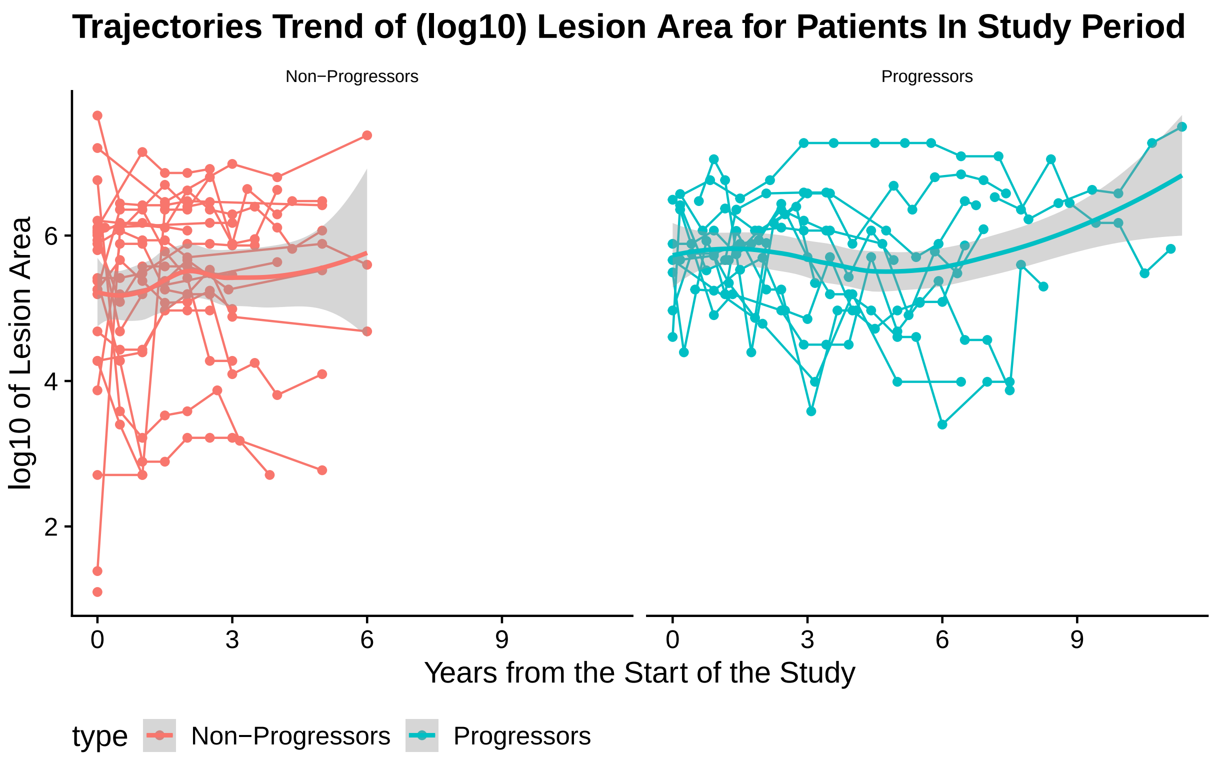

Figure 1.1: Trajectories Trend of (log10) Lesion Area for Patients During Study Period

1.2 Literature Review

According to previous literature, separately analyzing longitudinal data [4] and survival data [5] are well applied. Nevertheless, when they appear together, the longitudinal process and the survival process for one subject is often connected. Hence, the measurement value at a given time point for one subject is informative about whether the subject will experience the event of interest or not in the future. Likewise, the longitudinal trajectory may also be influenced by the occurrence of the event. Therefore, separate analysis is unable to account for the association between these two processes and may produce some inefficient and biased results. Two-step approach is an approach to consider the link between these two data, in which the unknown parameters estimated from the mixed-effects models are used as known covariates in the survival model, like the application in nephrology research [6]. Yet, this approach will produce biased estimation especially when longitudinal process and the survival process are strongly associated, and under-estimated standard errors because of ignoring the uncertainty of estimation in the first step in the second step.

1.3 Joint Modelings of Longitudinal and Survival Data

Joint modeling of longitudinal and survival data is increasingly popular for incorporating all information simultaneously, providing valid and efficient inferences. Several cases are suitable for carrying out joint models of longitudinal and survival data, such as 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 are implemented. Three popular association structures, the ”shared parameters” association, the ”current value” association, and the ”current value plus slope” association, are commonly used. Furthermore, two popular inference methods can be used for estimating parameters: likelihood inference and Bayesian inference. Theoretically, although they are based on different principles, the results should be similar. For this specific dataset we used in this report, with such a small sample size, it would be better to use Bayesian methods for model inference in order to incorporate prior information and obtain more meaningful posteriors. Estimating parameters in joint models under these two methods is often challenging. This report introduces some computational methods corresponding to these two inference methods, such as the Monte Carlo EM (MCEM) algorithm [7] for the likelihood method and Markov Chain Monte Carlo (MCMC) [8] for the Bayesian method.

In this report, we mainly discuss the joint modeling of longitudinal and survival data. We elucidate some aspects, such as how different association structures work, and how likelihood and Bayesian methods are used to estimate parameters. Among the five main joint models fitted, two are estimated by likelihood methods with ’current value’ association and ’current value plus slope’ association. The other three are estimated by 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. In this report, we focus on the package *JM* for likelihood inference and *JMBayes* for Bayesian inference in R. Additionally, after model fitting, we refer to several types of diagnostics, such as residual plots, sensitivity analysis, and trace plots, to evaluate the parameter estimates approximated from the models.

1.4 Outline

This report is organized as follows. In **Chapter 2**, we review the details about conducting longitudinal and survival analysis. In **Chapter 3**, we give an introduction about joint modelings of longitudinal and survival data from a theoretical level, including but not limited to different association structures, different inference methods and the corresponding computational methods. In **Chapter 4**, we focus on an example based on oral cancer data to clarify the practice of the joint model. In this section, five joint models are built with two inference methods and three association structures. After that, the parameters estimated by likelihood and Bayesian method are compared to detect the difference. The materials mainly discussed in this report are shown in **Chapter 5**, with some details about the similarity and difference between likelihood and Bayesian inferences.

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, and sometimes one study will take a long period. 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 [9] 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. The variables in the example collected for each person multiple times are longitudinal data.

In linear regression, all the variables for all the 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

The linear mixed-effects(LME) model is an extension of the simple linear regression model. Specifically, the linear regression model works for independent data; the 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 the 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 the results of using a linear regression model to predict y_{ij} will cause bias as y_{ij} is a response variable measured multiple times on the same individual. It is needed to incorporate the within-individual correlation. To simplify, if we assume the intercept β_0 varies in different individuals, we will introduce a random effect b_{0i} for β_0. Then the intercept is individual-specific. The 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 matrices usually contain covariates. $\boldsymbol{e_i}$ 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 will 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 matrices usually containing covariates. $\boldsymbol{e_i}$ 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, the 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 vectors. $\boldsymbol{\beta} = (1, \beta_1, \ldots, \beta_p)^T$ is the fixed effects vector and $\boldsymbol{b}_i = (1, b_1, \ldots, b_q)^T$ is the random effects vector. $\boldsymbol{X_i}$ and $\boldsymbol{Z}_i$ are design matrices often containing covariates. $\boldsymbol{e}_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})^T$ 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.
D is a $(q+1) \times (q+1)$ covariance matrix of random effects. The variance(the diagonal elements of D) of the $\boldsymbol{b}_i$ helps to explain the variability of the longitudinal measurements between individuals that are unexplained by covariates. D is unstructured, but sometimes it can be structured as a diagonal matrix. R_i is a $n_i \times n_i$ matrix of random errors for within-individuals. The variance (the diagonal elements of R_i) of e_i helps to explain the variability of the repeated measurements within each individual. In practice, we often assume $R_i = \sigma I^2$, 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 [10]. 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. The LME model is capable of working for the longitudinal data with missing data in the response. And we postulate that all of the missing data are missing at random.

The maximum likelihood method(MLE) or the restricted maximum likelihood method are popular for estimating unknown parameters in LME models. Iterative algorithms, like the EM algorithm or the Newton-Raphson method, are used to compute the maximum likelihood elements.

2.1.3 Other Types of Mixed-Effects Models

Nonlinear Mixed-Effects (NLME) Models

The Nonlinear mixed-effects (NLME) model is extended from the linear mixed-effects (LME) model. 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 covariance for each individual i, $\boldsymbol{\beta}$ is the fixed effect parameters, $\boldsymbol{b_i}$ is the random effect. The assumptions of NMLE models are the same as the MLE models, $b_i \sim N(0, D)$, $e_i \sim N(0, R_i)$ and b_i and e_i are independent. D and R_i are covariance matrices for the random effects and repeated measurements within each individual respectively.

Statistical inference for NLME models is commonly based on the likelihood method. However, compared to LME models, the marginal distribution and the likelihood function 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 the MLE models and NMLE models mentioned above are used when the response is continuous and approximately normally distributed. However, if the distribution of the response is from the 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. The GLMM can be written as

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

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

Statistical inference for a GLMM also usually uses the likelihood method. However, similar to NMLE models, GLMMs are also hard to compute because of the nonlinear random effects. Numerical methods or Monte Carlo EM algorithms are introduced to deal with this.

2.2 Models for Survival Data

2.2.1 Introduction

Survival analysis, also called time-to-event analysis, is a commonly used term for the analysis of outcomes that are timed to an event. The clinical field prefers 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, the 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 [11]. 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 samples experience the event before they are lost to follow-up, have a competing event that precludes the possibility of experiencing the event of interest, or the study ends. In these cases, if the individuals were to eventually experience the event, their survival time would be greater than the time 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. To be specific, **right-censored** denotes the event has not yet occurred as of time t, and may never occur; **left-censored** denotes the event occurred before time t; and **interval-censored** denotes the event occurred between time t_1 and t_2.

The 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 the cumulative probability after t, which can be written as

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

Hazard function

In survival analysis, the risk of an event is the probability that an event will occur within a given time period, while the 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, the 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 the event as of time t, the probability of the event occurring in a time period after t per unit time. Then shrink the future time period to zero. With this function, we can explore the individual's hazard of experiencing the event, the change of hazard rate over time, and how the hazard rate differs between groups or depends on other risk factors.

The hazard function is related to the rate of change of the survival function S(t). When the 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 the 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, the Kaplan-Meier model, the Exponential model, the Weibull model, and the Cox Proportional-Hazards model are commonly used. This section gives a brief review of the first three models, and more details about the Cox Proportional-Hazards model are shown in the next section.

The Kaplan-Meier model is classified as a non-parametric model, without any assumption of the distribution of data, hence we can only use the provided information to generate the survival function. The benefit of the Kaplan-Meier model is that it is intuitive and easy to interpret. But due to the minimal complexity, it is hard to draw meaningful insights from it.

The Exponential model is a parametric model, so an assumption needs to be made about the distribution of the data before building the model. The Exponential model assumes that the hazard rate is constant, which means the risk of the event occurring remains the same throughout the period of observation. The benefit of the Exponential model is that it provides substantial information on the survival function and the hazard function. Moreover, it can be used to compare the hazard rates of different groups. However, the strong assumption that the hazard rate is constant at any given time may not match well to all the data of interest.

Another parametric model is the Weibull model. Unlike the Exponential model which assumes that the hazard rate is constant, the Weibull model assumes the change in hazard rate is linear. The hazard rate can always increase, always decrease, or always stay the same, while the hazard rate cannot fluctuate. Similar to the Exponential model, the Weibull model is capable of computing many of the relevant metrics in survival analysis. But it also depends on the 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 introduced by D. R. Cox in 1972 [12]. It is called 'semi-parametric' because no assumption is made about the distribution of the 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. The

Cox model can conduct a survival analysis that examines survival data with respect to multiple variables at once; however, the Exponential model and the Weibull model can only examine each covariate individually. The Cox model, also known as the hazard rate formula, 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 as $h_0(t)$ does not depend on any parameter and drops out completely when estimating the parameters of the 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, the 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 Models of Longitudinal and Survival Data

3.1 Introduction

Longitudinal and survival data are commonly associated with the medical field. It is common to collect medical measurements and markers of an event of interest for patients. For example, in a study with multiple patients, researchers might be interested in the time it takes for the patients to recover. In this case, they might collect the patients' medical information (such as CD4 cell counts) multiple times over a period of time. The longitudinal trajectories of the patients might also be available to explain the recovery time. In such situations, it is important to consider both the longitudinal data (e.g., CD4 cell counts) and the survival data (time to recover) simultaneously. Since longitudinal models and survival models share some unobserved variables and parameters, it is possible to estimate the shared parameters in one model first using observed data and then use them as known parameters in another model. However, this may lead to some problems. When longitudinal data and time-to-effect data have a strong association, the parameter estimates may be biased and inefficient. The standard errors of the parameter estimates in the second model may be underestimated since the uncertainty of estimates in the first model is not considered in the second one, as they are seen as known parameters.

Joint models, such as those proposed by Faucett and Thomas [13] and Wulfsohn and Tsiatis [14], can be used to combine all the available information and account for it simultaneously to produce valid and efficient estimates. Additionally, the focus of a joint model for longitudinal and survival data can vary depending on the objectives of the study. It may be on the longitudinal model, the survival model, or both. Different structures for combining these two models have also been introduced, 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**.

To summarise, this chapter gives an overview of the basic idea behind the joint modeling for longitudinal and survival data. In **Section 3.2**, we illustrate the statistical framework about joint models. Then, in **Section 3.3** and **Section 3.4**, we provide more details on the likelihood method and Bayesian method, respectively, for estimating the parameters in joint models.

3.2 Statistical Framework of Joint Models

As mentioned previously, we primarily focus on right-censored survival data. In this section, we prioritize the survival model and consider the longitudinal model as secondary. This situation often arises when the time-dependent covariates in the survival model have measurement errors or missing data. Therefore, the secondary longitudinal model should be introduced to address these measurement errors and missing data.

For individual $i(i = 1, 2, \dots, N)$, let s_i be the event time, and c_i the censoring time. Since we assume the censors are only right-censored and all of them are randomly censored, the observed time value 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. In the following, we will discuss the longitudinal submodel, the survival submodel and the association structures of the joint models.

3.2.1 Survival and Longitudinal Submodels

The first is the primary submodel, the survival model. In this section, we use the Cox PH model for survival data analysis, which is a semi-parametric model with no need for any assumption about the distribution of the event times. This 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 exogenous, possibly time-varying and $\boldsymbol{\gamma}$ is the corresponding regression parameters.

For the secondary model, linear mixed-effects models are considered. What we are interested in is the longitudinal data $y_i(t) = (y_{i1}(t), y_{i2}(t), \ldots, y_{in_i}(t))^T$, which denotes the observed measurements for individual i at time $t(t = 1, 2, \ldots, T_i^*)$. But measurement errors may exist in the true observed data. Hence, what we actually focus on is the unobserved variable $\mu_i(t) = (\mu_{i1}(t), \mu_{i2}(t), \ldots, \mu_{in_i}(t))^T$ which contains the repeated measurements for each individual over time without measurement errors. The relationship between $y_i(t)$ and $\mu_i(t)$ can be written as

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

The linear mixed-effects model of true measurements $\mu_i(t)$ can be written as

$$\begin{aligned} \mu_i(t) &= X_i(t)\boldsymbol{\beta} + Z_i(t)\boldsymbol{b_i} \\ &= f(X_i(t)) + f_i(Z_i(t)), \quad i = 1, 2, \ldots, n, \end{aligned} \tag{4}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{b_i}$ are fixed and random effect parameters, respectively. The design matrices $X_i(t)$ and $Z_i(t)$ contain the covariate information for the fixed and random effects, respectively. The second equation in the second line works while using a spline approach to make the model non-parametrically. $f(\cdot)$ and $f_i(\cdot)$ are the spline function for the fixed and random components, respectively.

3.2.2 Joint Models

The joint models are then presented to associate the mixed effect submodel and the survival submodel together. Then we have

$$\begin{aligned} h_i(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T_i^* + \Delta t | T_i^* \geq t, M_i(t), w_i)}{\Delta t} \\ &= h_0(t) exp(\boldsymbol{\gamma}^T \boldsymbol{w_i} + f\{\mu_i(t), \boldsymbol{b}_i, \alpha\}), \end{aligned} \tag{5}$$

where $M_i(t) = \{\mu_i(l), 0 \leq l < t\}$ represents the history of the true longitudinal data $\mu_i(l)$ up to time point t, where $l < t$. In the second line of the function, $f\{\mu_i(t), b_i, \alpha\}$ denotes the longitudinal data, α as 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 $f(\cdot)$ function lead to different forms of joint models. The following three association functions are frequently used in joint modeling.

The 'Current Value' Association

As the name implies, the 'current value' association suggests that the actual 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. The 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 association between time-dependent longitudinal variables μ_i at time t and the survival data at the same time. To be specific, For individual i at time t, if the current value $\mu_i(t)$ of longitudinal part increase one-unit, the risk of this event occurred will increase an $exp(\alpha_1)$-fold, given the event has not occurred before time t. Pay attention that α_1 does not change across all the 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.

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

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 survival data at the same time. To be specific, 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

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

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

where $\boldsymbol{b}_i = (b_{i1}, b_{i2}, \ldots, b_{im})$ is from the longitudinal submodel represents the random intercept and slope effects for individual i. $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ expresses the association between the random effects from longitudinal trajectory and the hazard ratio. As the random effects represent the individual deviations from the sample average intercept and the slope values, one-unit change in these deviations lead to an exponential change of the hazard ratio.

Compared with the aforementioned two associations, the 'shared parameters' association has simpler computation because the associative part $\boldsymbol{\alpha}\boldsymbol{b}_i$ is time-independent. To be specific, for each individual i, random effects b_i is independent of time t, while the current value $\mu_i(t)$ and the current slope $\mu_i'(t)$ depend on time t.

3.3 Statistical Inferences

3.3.1 Likelihood Methods

In this section, we primarily explain one of the most widely used methods for statistical inference in joint modeling of longitudinal and survival data. The likelihood approach estimates all parameters in both the longitudinal and survival models simultaneously, which helps to avoid bias that may result from building the models in separate steps. This results in accurate and trustworthy estimations. Additionally, the joint likelihood method can also be extended to incorporate more than two models.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, D, \alpha)$ denotes all the unknown parameters in the longitudinal and survival submodels. The likelihood method for the joint model can be written as

$$\text{Ł}(\boldsymbol{\theta}) = \prod_{i=1}^{\infty}[\int f(t_i, \delta_i | \boldsymbol{b}_i, \boldsymbol{\theta}, \alpha) f(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\beta}, R_i) f(\boldsymbol{b}_i | D) d\boldsymbol{b}_i]. \tag{9}$$

The joint likelihood of our aforementioned example can be expressed as

$$\begin{aligned} \text{Ł}(\boldsymbol{\theta}) = \prod_{i=1}^{\infty}\{\int & [h_0(t_i)\exp\{\boldsymbol{\gamma}^T \boldsymbol{w_i} + f\{\mu_i(t), \boldsymbol{b}_i, \alpha\}]^{\delta_i} \\ & \times \exp[-\int_0^{t_i} h_0(u)\exp\ \boldsymbol{\gamma}^T \boldsymbol{w_i} + f\{\mu_i(t), \boldsymbol{b}_i, \alpha\}\}du] \\ & \times f(y_i | \boldsymbol{b}_i, \boldsymbol{\beta}, \alpha) f(\boldsymbol{b}_i | D) d\boldsymbol{b}_i\}. \end{aligned} \tag{10}$$

As indicated in Equation 10, the joint likelihood often has a complex form due to the presence of unobserved random effects, censoring, and semi-parametric models. Two methods have been proposed to tackle the computational difficulties: the Monte Carlo EM (MCEM) algorithm and an approximate method for statistical inference. The Monte Carlo simulation aims to estimate the probability of outcomes in the presence of random variables. It is effective in capturing the effects of risk and uncertainty in predictions and forecasts. The Monte Carlo EM algorithm is a variant of the EM algorithm, where the expectation in the E-step is calculated through Monte Carlo simulations. The following part provides further information about the EM algorithm.

EM Algorithm

The 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. For example, some entries in a data table may be missing. 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 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.

To be specific, let $\mathbf{Y}$ denotes incomplete data (real data), $\mathbf{X}$ denotes augmented data (data we constructed), $(\mathbf{Y}, \mathbf{X})$ denotes complete data, and $\boldsymbol{\theta}$ denotes 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:

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

Parameter Estimation in Joint Likelihood Models using the MCEM Algorithm

For the joint likelihood

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

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

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

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 generating $m^{(t)}$ samples of $\boldsymbol{b_i}$ from $f(\boldsymbol{b_i}|\boldsymbol{y_i}, \boldsymbol{w_i}, \boldsymbol{\theta^{(t)}})$ and using the empirical mean to approximate the integral. The $(t+1)$th iteration of the EM algorithm can then be written as follows:

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

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 getting a convergent result.

In conclusion, the EM algorithm is a widely used method for computing maximum likelihood estimates (MLEs) in situations with missing data and ensures that the likelihood increases with each iteration. However, when dealing with high-dimensional data, the integral can become intractable. In such cases, the Monte Carlo approximation in the E-step can provide a solution. However, a large Monte Carlo sample size is required to estimate the integrals accurately as the algorithm approaches convergence, which can be computationally expensive. Additionally, the convergence of the EM algorithm can be slow and may only reach a local optimum.

3.3.2 Bayesian Methods

The joint modeling of longitudinal data and survival data discussed in the previous section involves many parameters, which can lead to poor estimation and identifiability issues. Bayesian methods are well-suited for handling these types of problems because they allow for the incorporation of prior information about population parameters and can reduce identifiability issues. However, choosing an appropriate prior distribution is crucial as it can significantly impact the final results.

Both Bayesian and likelihood inferences have similar computational challenges. With advancements in modern computing and computational tools, various methods have been developed to address these challenges, such as the Markov Chain Monte Carlo (MCMC) methods.

General Concepts

The basic idea for Bayesian inference 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 $f(\boldsymbol{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$. $\boldsymbol{\theta}$ denote the unknown parameters, which are random variables with probability density function(prior distribution) $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$. $\boldsymbol{\theta}_0$ are the hyper-parameters which are frequently obtained from similar studies or expert options. As we discussed before, the Bayesian inference can be written as

$$\begin{aligned} f(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \\ &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}), \end{aligned} \tag{11}$$

where $f(\boldsymbol{y}|\boldsymbol{\theta})$ is the likeliood part, $f(\boldsymbol{\theta})$ is the prior part, and the Bayesian inference of $\boldsymbol{\theta}$ is based on the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$. Thus, under Bayesian inference, 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, like the Gibbs sampler, 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

According to the basic theory of Bayesian inferences, 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 distributions are preferred. In such cases, non-informative conjugate priors should be considered, in which a prior distribution is chosen when the resulting posterior distribution also belongs to the same family of 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

$$p(s) = \left(\begin{array}{c} n \\ s \end{array} \right) q^s (1-q)^{n-s}.$$

Consider a common conjugate prior the 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

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

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

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

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

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

Markov Chain Monte Carlo(MCMC) methods

While facing missing values in longitudinal data and high-dimensional data which lead to an intractable integral, Monte Carlo EM algorithms are introduced for the likelihood inference. Similarly, the posterior distributions in Bayesian methods are as well often highly complicated. Markov chain Monte Carlo (MCMC) is an increasingly popular method for estimating posterior distributions in Bayesian inference. A particular strength of MCMC 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 [15].

MCMC methods consist of two parts, Monte Carlo method and Markov chain. Monte Carlo is 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 sample mean will be calculated as the expected value. Hence, the Monte Carlo approach leads to easier calculation, like for the example above, 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 shown 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 the 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, 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 [15].

MCMC is used to approximate posterior distributions that cannot be directly calculated by randomly drawing a sequence of samples from the posterior and examining their mean, range, and other properties. Estimating the posterior distribution, $f(\boldsymbol{\theta}|\boldsymbol{y})$, is usually difficult. 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 evaluate dimensions. MCMC method can evaluate $f(\boldsymbol{\theta}|\boldsymbol{y})$ by drawing samples from it and avoiding the need to evaluate $f(\boldsymbol{y})$ simultaneously.

Model Selection

Various structures are used to understand the interrelationships between longitudinal and survival outcomes. Among these structures, one important question is how to select the most appropriate structure for a given dataset when using Bayesian inference, since more than one association structure may be potentially used at the same time. One popular approach for model selection is the deviance information criterion (DIC) [16].

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 are 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 formulas can be used for calculating the effective number of parameters in 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

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

As we know, more effective parameters will make it easier to fit the model of the data, 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

Bayesian method can be conducted into joint modeling of different types of survival and longitudinal models. In this part, we consider a joint model of a primary survival model and a secondary longitudinal model which is used to measure errors in a time-dependent covariate. Like the sample aforementioned, 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 value t_i can be written as $t_i = min\{s_i, c_i\}$. The censoring indicator δ_i can be defined as $\delta_i = I(s_i \leq c_i)$. $\delta_i = 0$ 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 over time without measurement errors. The Cox proportional hazards model can be written as

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

where w_i are baseline covariates without measurement errors. $\boldsymbol{\gamma}$ is the regression parameters. The mixed effect model of longitudinal trajectory can be shown 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. D and R_i are covariance matrices for the random effects and repeated measurements within each individual, respectively. The joint model 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. In *JMBayes* package which is a common package for joint models with Bayesian inference, to model the baseline hazard function $h_0(t_i)$ in the survival data, while ensuring the flexibility, it uses the B-spline approach. Mathematically, the logarithm of the baseline hazard function can be written as

$$logh_0(t) = \gamma_{h_0,0} + \Sigma_{q=1}^{Q} \gamma_{h_0,q} B_q(t, \boldsymbol{v}), \tag{12}$$

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

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

where

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

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

and

$$R_i = \sigma^2.$$

For Bayesian inference, the first step is to assume the prior distribution of the parameters. For all of the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{b}_i, \boldsymbol{\beta}, \alpha)$, we take standard prior distributions.

To be specific, the vector of fixed effect parameters of the longitudinal submodel $\boldsymbol{\beta}$, the regression parameters of the survival model $\boldsymbol{\gamma}$, the association parameter α are used independent univariate diffuse normal priors. We assume

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

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

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

and

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

The mean value of the normal distribution is centered at zero meaning that the prior assumes that 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 $\boldsymbol{b}_i \sim N(\mathbf{0}, D)$. And when fitting 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 the prior distributions of the parameters have been defined, the posterior distribution of the parameters given the observed data can be expressed. To address the challenging task of evaluating the intractable integral, various computational techniques such as Markov Chain Monte Carlo (MCMC) methods can be employed.

3.3.3 Summary

In general, both likelihood methods and Bayesian methods are popular approaches for statistical inference in joint modelings for the survival data and longitudinal data. Likelihood methods have high correlation with Bayesian methods. To be specific, if we use a non-informative (uniform distribution) priors to get maximum likelihood estimates in Bayesian methods, they are equivalent to likelihood methods. Mathematical speaking, for a prior distribution $f(\boldsymbol{\theta}) \propto 1$, the Bayesian inference 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. Due to the similar form, Bayesian inference and likelihood inference encounter similar computational challenges and use similar computational tools while facing the intractable high-dimensional data.

While comparing with the likelihood methods, the major difference of Bayesian methods is that the model parameters are treated as random variables. Bayesian inference actually provides 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 [17]. 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. Typically, a normal cell has twice the basic set of 23 chromosomes and is referred to as diploid while cells that do not possess an integer multiple of the basic set of chromosomes are referred to as aneuploid. In each nucleus, the DNA content is measured by a normalized scale referred to as DNA Index(DI). Samples typically show a dominant peak of normal diploid cells (DI=1) and a smaller peak of tetraploid cells (D1=2) stalled at the G2/M phase of cell division. If cell proliferation is high, cycling cells are seen between the diploid and tetraploid peak while cells showing $DI > 2.3$ are considered aneuploid. Percentage of non-diploid cells (cycling, tetraploid, aneuploid) and number of aneuploid cells are common DNA-ICM features which are able to frequently monitor OPMLs without the unnecessary biopsies.

Based on the background, an analysis on OSCC datasets with patient demographic, clinical features and the corresponding diagnosis results during the study time is discussed in this section. To be specific, 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. Joint modeling for these two processes are then conducted with likelihood inference and Bayesian inference and different association structures to get different estimates. Finally, we focus on the comparison of these methods.

For the longitudinal part, we fit mixed-effects models using the **lme(·)** function from the *nlme* package. For the survival part, we fit the Cox proportional hazards model using the **coxph(·)** function from the *survival* package. For joint modeling of these two parts, we use the basic model fitting function called **jointModel(·)** in the *JM* package for likelihood inference, and **jointModelBayes(·)** in the *JMbayes* package for Bayesian inference.

4.2 Data Description

Two data frames, *cancer_old* and *cancer_old.id*, are available and contain longitudinal and survival information, respectively. There are 408 observations for 41 patients in dataset *cancer_old* 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.

As shown before, **lesion_area** is a numeric variable that has been measured several times on each individual at different time points with measurement errors. Therefore, longitudinal analysis, such as a linear mixed-effects model, is needed to fit the longitudinal trajectories. After the initial analysis, patients with **study_id** of *2094* and *2046* have no **lesion_area** data, and therefore, they should be dropped. Patients *1888* and *1913* should also be dropped as they have missing demographic data. Furthermore, when building joint models, no observations should be collected after the diagnosis time. 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 left for future analysis. The longitudinal trajectories of **lesion_area** for patients with different diagnosis are shown separately in Figure 4.1. As can be seen from the trajectory plot, compared to the *Progressors* whose LGD ended up malignant transformation, *Non-Progressors* who did not develop malignant transformation have relatively low lesion area. For the lesion area, individual trajectories of different patients seem to vary and the value across the time within each individual varies. Thus, a longitudinal model with fixed effects and random effects should be conducted to explain both the between-individual differences and the within-individual differences.

Based on the dataset, this section presents an analysis of oral squamous cell carcinoma (OSCC) to investigate how time-dependent variables with measurement errors and time-independent variables affect the dichotomous indicators for the diagnosis results. Specifically, we build a longitudinal model for the time-dependent variable **lesion_area** and a survival model to explore the correlation between other time-independent predictors and the time of diagnostic results. Then, we use joint modeling for longitudinal and survival data, evaluated by the likelihood method and Bayesian method, respectively, to provide a more accurate fit.

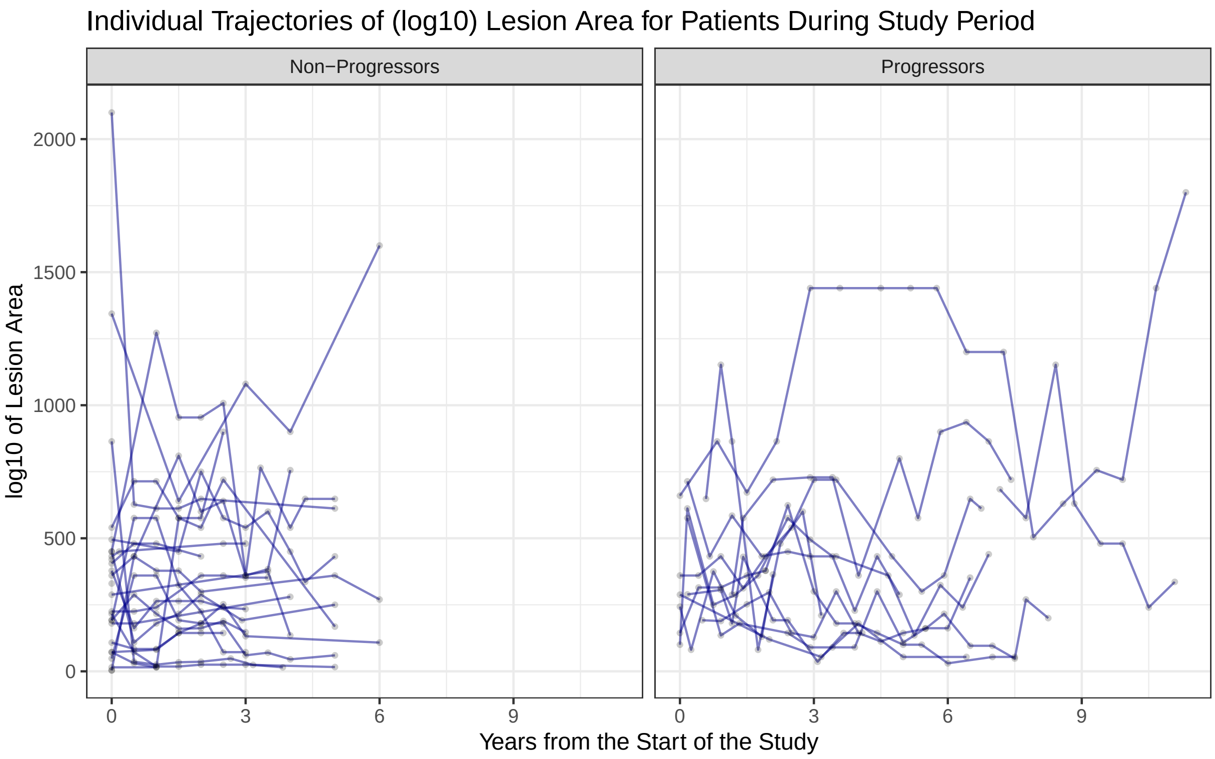

Figure 4.1: Individual Trajectories of (log10) Lesion Area for Patients During Study Period

4.3 Longitudinal and Survival Analysis

To explore the association between the change in lesion area over time until the end of the study and the final diagnosis (progression or not), a longitudinal model is used to model the longitudinal data of the oral lesion area, and a survival model is used to model the event time data.

4.3.1 Longitudinal Analysis

In this section, we focus on the dataset *cancer_old*, which includes repeated measurements of **lesion_area** for each study participant over a given period of time. The original values of **lesion_area** in *cancer_old* have a wide range, and the distribution is not normal. To normalize the distribution, we took the logarithm of **lesion_area**. As shown in Figure 4.2, the distribution of the log of **lesion_area** is normal.

Based on prior research, it is known that repeated observations for the same individual are often correlated. Therefore, a mixed-effects model is used to analyze the longitudinal part of the data. This model accounts for between-person variability by using invariant fixed effects and 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.

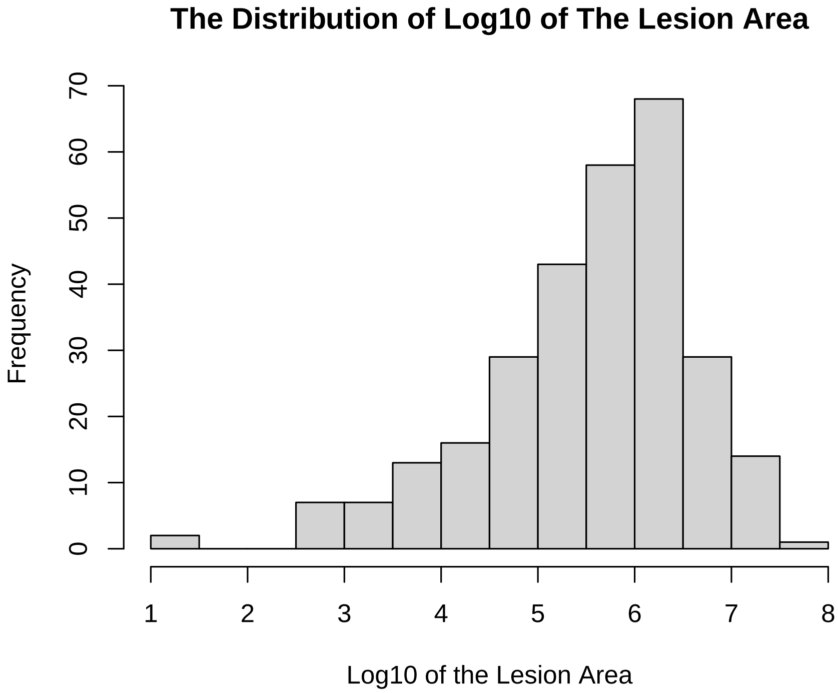

Figure 4.2: The Distribution of Log10 of The Lesion Area

Moreover, previous research has suggested a high prevalence of oral mucosal lesions and disease in the geriatric population, with the risk increasing with age [18]. Hence, we also considered the age of patients at the start of the study, as it may affect the rate of change in lesion area over time.

We finally tested seven mixed-effects submodels: the first with fixed effects for intercept and the linear slope and random effects for only the intercept, the second adding to the first a random effects 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 effects for linear slope (to describe linear change over time), the fifth adding 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. $obstime_i(t)$ means the period from the start of the study to that measurement time. Age_i is the age of the patient at the beginning of the study. Mathematically, these seven submodels can be expressed as

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

Among the seven submodels, we select the one that provides the most satisfying description of the lesion area trajectories based on the Bayesian Information Criterion(BIC) [19]. BIC is a criterion for model selection, and the models with lower BIC are generally preferred. When fitting models, adding parameters tend to increase the likelihood, but uncontrolled adding might cause overfitting. This problem can be resolved by introducing a penalty term for the number of parameters in the model, like BIC and Akaike information criterion (AIC) [20]. To be specific, the BIC is formally defined as

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

where $\hat{L}$ is the maximized value of the likelihood function of the model M, n the sample size, k 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 different random effects should be performed using maximum likelihood estimation (ML). However, we fit the longitudinal submodel with the *lme(·)* function from the **nlme** package [21], 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 values which are computed by ML.

The results of the statistical criteria are displayed in Table 4.1. Hence, we choose the second model as the best fitting model, which is expressed as

$$y_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) obstime_i(t) + \epsilon_i(t). \tag{13}$$

Then using the function *lme(·)* from **nlme** package to fit Equation 13. The results and the diagnosis are shown in the following.

Model Diagnosis

There are four general assumptions for a linear mixed-effects model [10]:
1. The explanatory variables are related linearly to the response.

Table 4.1: BIC Values for Mixed Effect Model Comparisons

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

2. The errors have constant variance.
3. The errors are independent.
4. The random effects and the errors are Normally distributed.

To assess the first assumption, 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 the second assumption, a plot of the residuals in order can reveal any seasonal patterns or autocorrelation that may exist. To evaluate the third assumption, 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 for error independence. To assess the fourth assumption, 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 for evaluating whether a data set is approximately normally distributed. The points should form a roughly straight line if the data are from a normal distribution.

According to the previous discussion, several issues need to be diagnosed. First, we need to check for 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 helps to detect 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". In conclusion, we observe some outliers in the data frame, but no significant violations of the assumptions are detected.

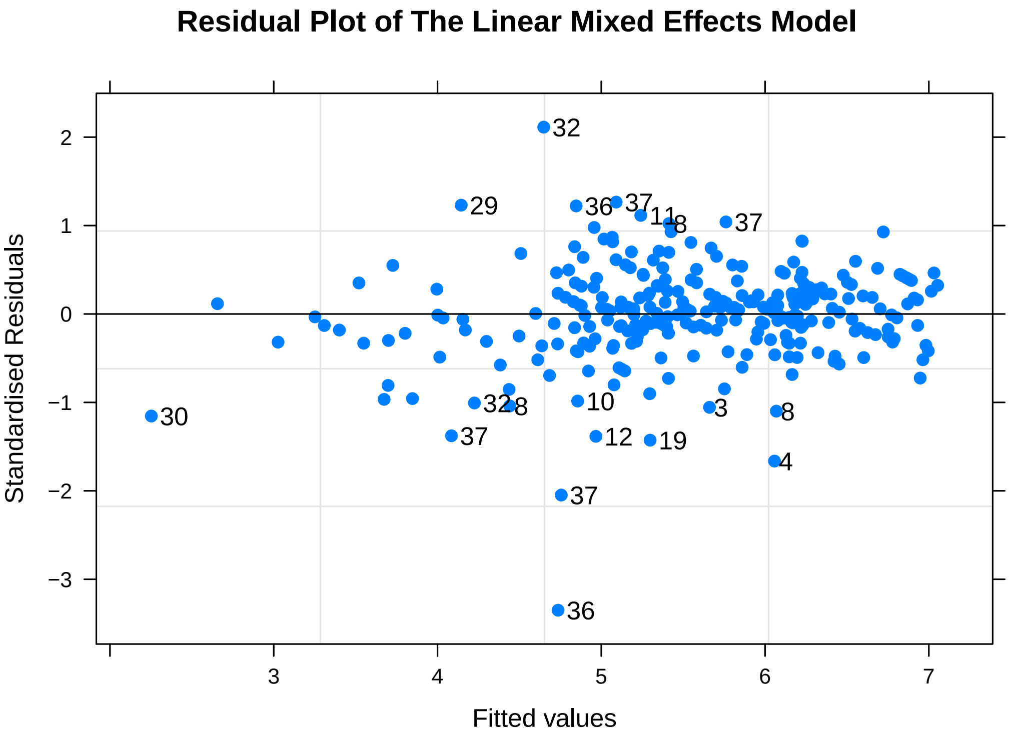

Figure 4.3: Residuals Plot of The Linear Mixed Effects Model

Results

All of the results are shown in Table 4.2. In the summary table, $\boldsymbol{\beta} = (\beta_0, \beta_1)$ represents the fixed effect parameters 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 for the test of this coefficient being 0 was 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 reject the null hypothesis that the global intercept is zero. Besides the global intercept, each group has its own intercept in LME models, which will be an offset from the global intercept. 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

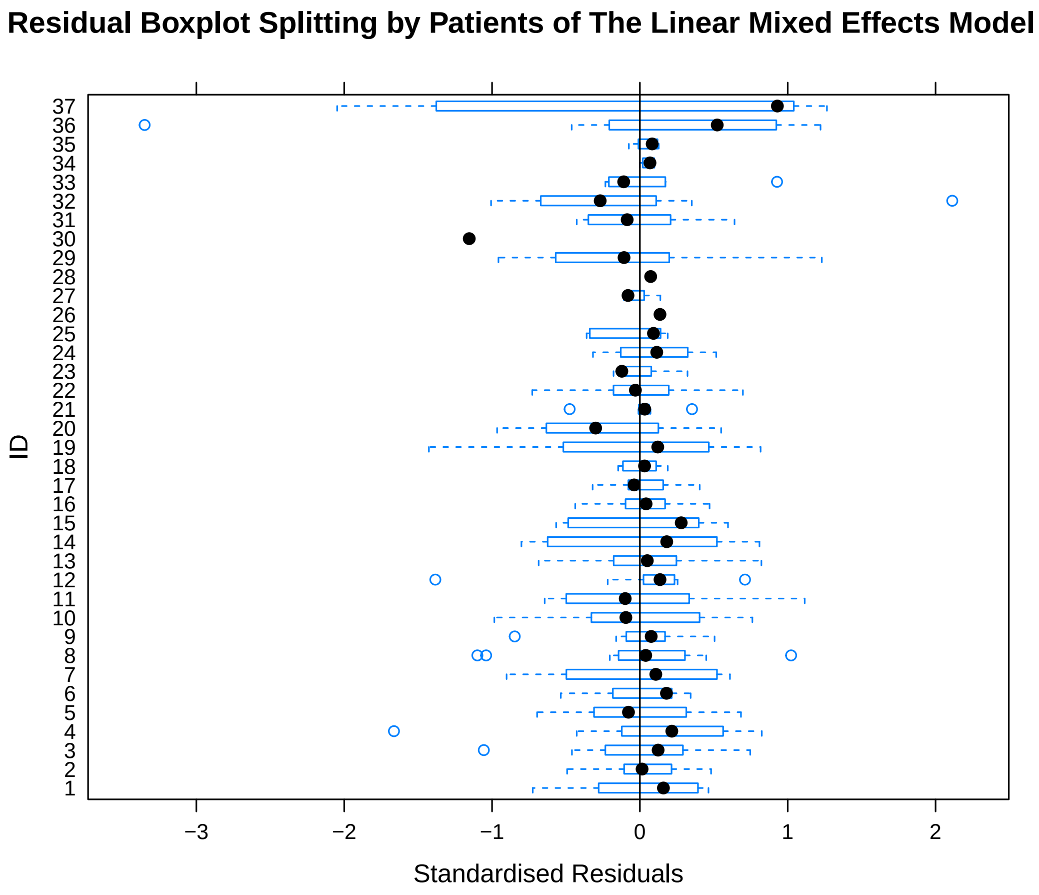

Figure 4.4: Residual Boxplot Splitting by Patients of The Linear Mixed Effects Model

of **lesion_area**. The p-values for the tests of these coefficients being 0 were larger than 0.05. However, since we only have 286 observations for a small sample size, which is not large enough to 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 effect parameters 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

Besides the summary of fixed effects, an additional part of the LME model is the random effects. In this model Equation 13, there are two random effects for the intercept and

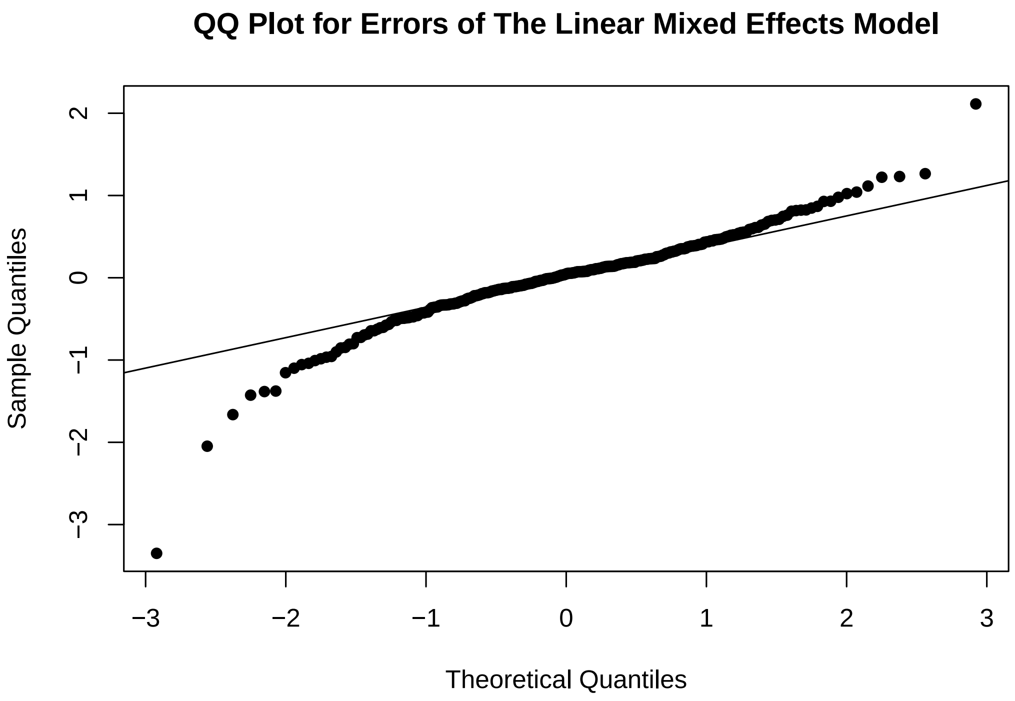

Figure 4.5: QQ Plot for Errors of The Linear Mixed Effects Model

obstime. As we know, the standard errors of variance components in a LME model can provide valuable information about the contribution of the random effects to the model. However, the reported parameters of random effects in R are only the standard deviation of the random intercepts or random slopes. We need to use another function to get the standard errors. The standard errors of the intercept random effect b_{0i} is 0.1799 and the slope random effect b_{1i} is 0.0160. Since both of these values are small, hence a relatively high certainty that the different subjects differ in their intercepts and slopes.

To visually assess the goodness of the model fits, nine patients were randomly selected to visualize the results. 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 the survival submodel. We perform a Cox PH model to model the hazard of developing oral squamous cell carcinoma at the end of the study. According to previous research, heavy alcohol consumption, especially when combined with cigarette smoking, increases the risk of oral squamous cell carcinoma [22]. Hence, in the COX PH model, we consider **cycling_mean**, **Smoke** and **Alcohol** as explanatory variables and the time to diagnosis as response.

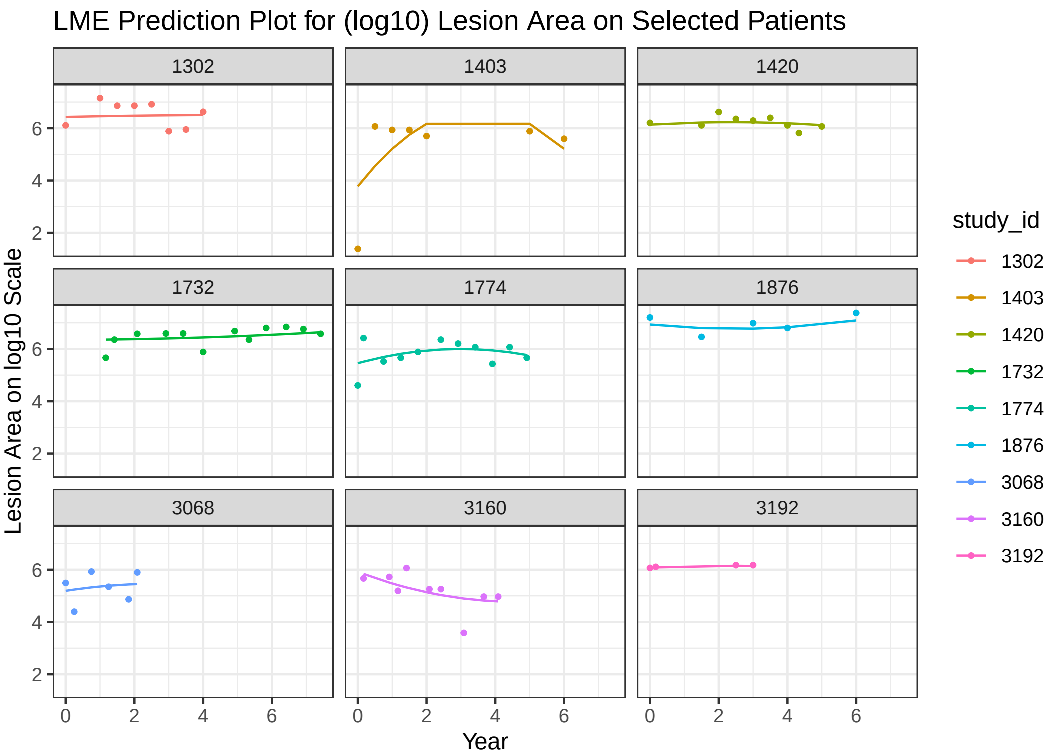

Figure 4.6: LME Prediction Plot for (log10) Lesion Area on Selected Patients

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

$$h_i(t) = h_0(t)exp(\gamma_1 w_1 + \gamma_2 w_2 + \gamma_3 w_3 + \gamma_4 w_4), \tag{14}$$

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 at the end of the study. According to Figure 4.7, some differences exist between these two diagnostic 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 Kaplan-Meier estimate 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 the Cox PH model is that the baseline hazard functions for model predictors are proportional [23]. In other words, the survival curves for different

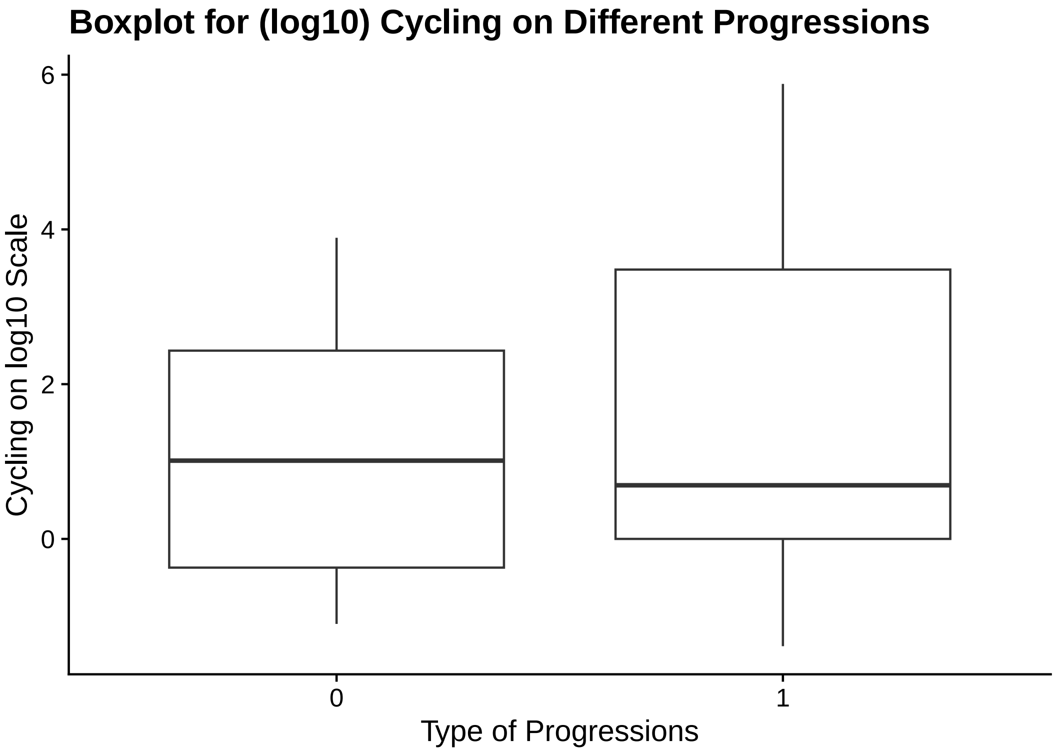

Figure 4.7: The Boxplot for (log10) Cycling on Different Progressions

strata of a given explanatory variable have hazard functions that are proportional over time. For example, in our model, where we specify **Alcohol** (habitual drinker vs. non-habitual drinker) as a mortality predictor, it is assumed that the hazard functions for habitual drinkers and non-habitual drinkers are proportional at the same time point. Hence, after fitting the model, it is necessary to check the PH assumption for each covariate. This can be done using the Schoenfeld Residuals Test (SRT) [24], which evaluates the independence between model residuals and the time variable **obstime**. The test results are shown in Table 4.3. The results indicate that the PH assumption holds for all of the predictors (with $p - values > 0.05$) and for the whole survival submodel (with a $p - value > 0.05$).

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

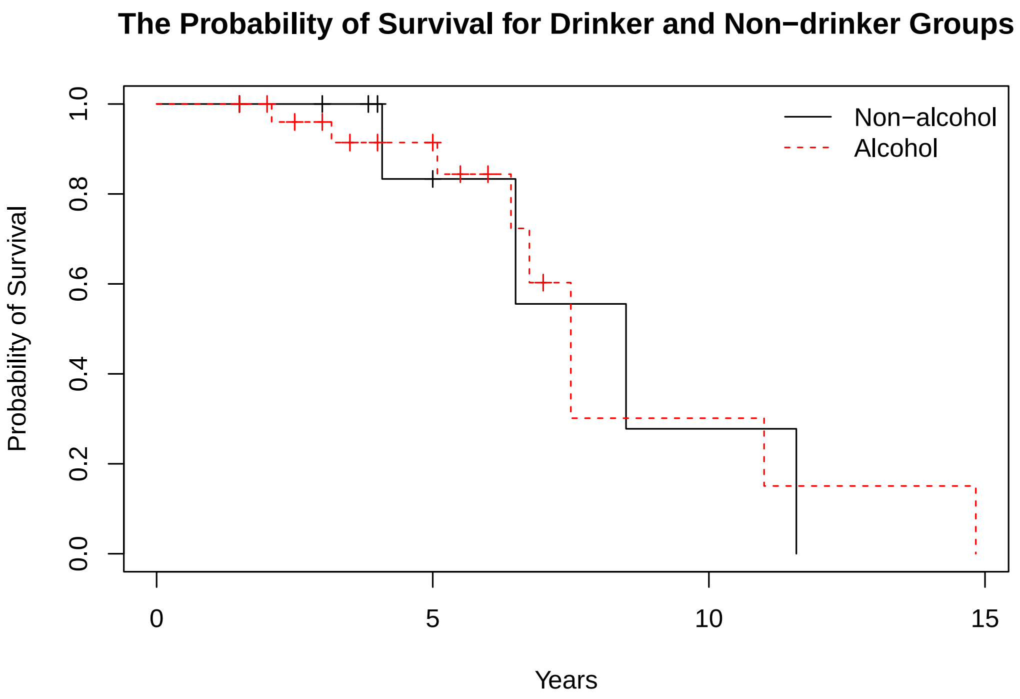

Figure 4.8: Kaplan-Meier Estimates of the Probability of Survival for Drinker and Non-drinker Groups

Results

After fitting the survival submodel Equation 14, the results of the parameters can be seen in Table 4.4. It should be noted that the p-value for all three overall tests (likelihood, Wald, and score) are not significant, suggesting that the model is not statistically significant.

Table 4.4: Parameter Estimates From 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 $HR = \exp(1.2132) = 3.36$, with a 95% confidence interval of 0.3998 to 28.306. Because the confidence interval for HR includes 1, which indicates that habitual drinker or not makes a smaller contribution to the difference in the HR after adjusting for other predictors. Hence, it is not a significant contribution.

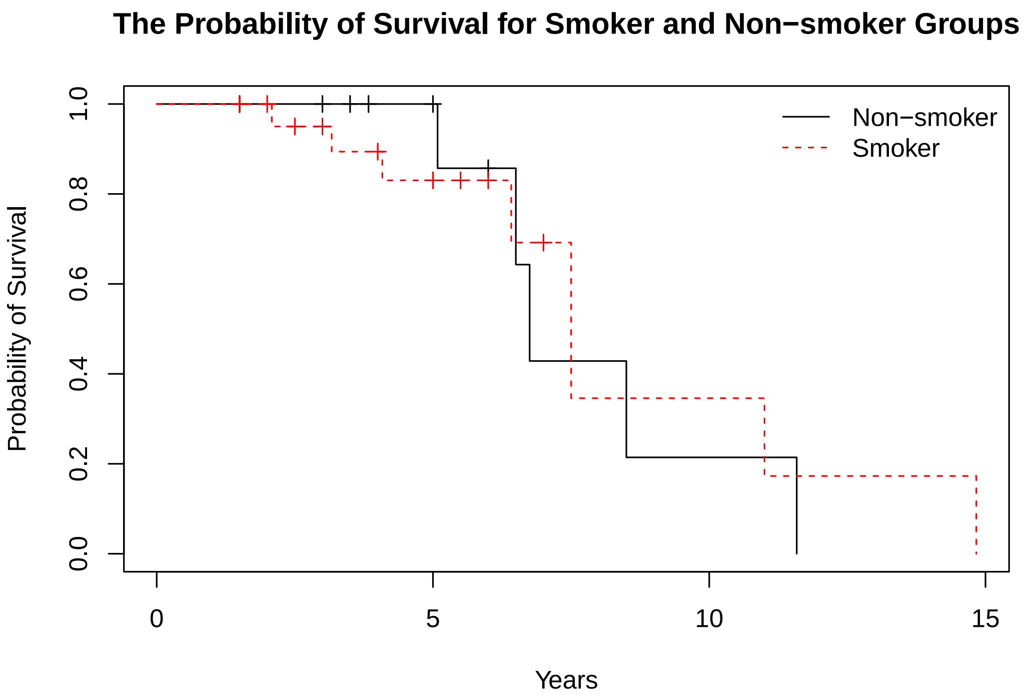

Figure 4.9: Kaplan-Meier Estimates of the Probability of Survival for Smoker and Non-smoker Groups

The p-values of **Smoke** and the interaction between **Smoke** and **Alcohol** are around 0.06, which is relatively low. The hazard ratio of **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 will lead to lower risk of experiencing progression, which do not make sense in practice. 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, thus we should build a Null Cox PH Model with no explanatory variables to compare nested survival models. The Null COX PH Model can be written as

$$h_i(t) = h_0(t). \tag{15}$$

Then, 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 in order to identify the model that best fits the population from which the data were sampled [25]. The results can be found 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. As a result, the Null Cox PH model should be selected.

Table 4.5: Comparing 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 Investigating the Joint Models using Likelihood Methods

The Cox PH model is used to model the time it takes for a patient's condition to progress (transform into something malignant). However, some research elucidates that higher **lesion_area**, especially higher 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 postulates the linear mixed-effects model Equation 13 for the **lesion_area** and Null Cox PH model Equation 15 for the survival data.

In this section, we demonstrate an analysis of joint models with likelihood inference which is implemented by the **jointModel(·)** function in R package *JM*. As we elaborated before, the joint model can be shown as

$$h_i(t) = h_0(t)exp(f\{\mu_i(t), \boldsymbol{b_i}, \boldsymbol{\alpha}\}), \tag{16}$$

where $h_i(t)$ denotes the hazard of the event for individual i at time point t, and $h_0(t)$ 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. Attention that $\mu_i(t)$ is different from $y_i(t)$, with the latter being contaminated with measurement errors at time t. $\boldsymbol{b}_i$ is the random effect parameters, and $\boldsymbol{\alpha}$ measures the association strength of the **lesion_area** to the risk for developing malignant transformation. Joint models for such joint distributions are of the following form:

$$p(T_i, \delta_i, y_i) = \int p(y_i|\boldsymbol{b}_i)\{h(T_i|\boldsymbol{b}_i)^{\delta_i} S(T_i|\boldsymbol{b}_i)\}p(\boldsymbol{b}_i)d\boldsymbol{b}_i, \tag{17}$$

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

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

4.4.1 Exploring Joint Models with Varying Associations

Maximum likelihood estimation for Equation 16 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

$$p(T_i, \delta_i, y_i | \boldsymbol{b}_i; \boldsymbol{\theta}) = p(T_i, \delta_i | \boldsymbol{b}_i, \boldsymbol{\theta}) p(y_i | \boldsymbol{b}_i; \boldsymbol{\theta}), \tag{18}$$

$$p(y_i | \boldsymbol{b}_i; \boldsymbol{\theta}) = \prod_j p\{y_i(t_{ij}) | \boldsymbol{b}_i; \boldsymbol{\theta}\}, \tag{19}$$

where $\boldsymbol{\theta} = (\theta_t, \theta_y, \theta_b)$ denotes all of the parameters. θ_t contains the parameters for the event time outcome, θ_y contains the parameters for longitudinal outcome and θ_b denotes the unique parameters of the random-effects covariance matrix. $p(\cdot)$ denotes an appropriate probability density function of the linear mixed-effects model. The joint log-likelihood contribution for the i-th subject can be expressed as

$$log p(T_i, \sigma_i, y_i; \boldsymbol{\theta}) = log \int p(T_i, \sigma_i | \boldsymbol{b}_i; \theta_t, \boldsymbol{\beta}) [\prod_j^{n_i} p\{y_i(t_{ij}) | \boldsymbol{b}_i; \theta_y\}] p(\boldsymbol{b}_i; \theta_b) d\boldsymbol{b}_i, \tag{20}$$

where $p(T_i, \sigma_i | \boldsymbol{b}_i; \theta_t, \boldsymbol{\beta})$ is the likelihood of the survival part, which can be formulated as

$$p(T_i, \sigma_i | \boldsymbol{b}_i; \theta_t, \boldsymbol{\beta}) = h_0(t_i) \exp\{f\{\mu_i(t), \boldsymbol{b}_{\boldsymbol{i}}, \boldsymbol{\alpha}\}\}]^{\delta_i} \times \exp[-\int_0^{t_i} h_0(u) \exp\{f\{\mu_i(t), \boldsymbol{b}_{\boldsymbol{i}}, \boldsymbol{\alpha}\}\} du. \tag{21}$$

As aforementioned, there are mainly three structures of the association between longitudinal trajectories $f\{\mu_i(t), \boldsymbol{b}_{\boldsymbol{i}}, \boldsymbol{\alpha}\}$ in processing and risk of the event interested of, the ’current value’ association, the ’current value plus slope’ association and the ’shared parameters’ association. For the **jointModel(·)** function in R package *JM* with likelihood inference, it can only conduct the first two structures of the association. However, for the **jointModelBayes(·)** function in R package *JMbayes* with Bayesian inference, all of these three structures are able to achieve. If we assume that the longitudinal submodel and survival submodel associate with current value, it can be written as

$$f\{\mu_i(t), \boldsymbol{b}_{\boldsymbol{i}}, \boldsymbol{\alpha}\} = \alpha \mu_i(t). \tag{22}$$

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

$$f\{\mu_i(t), \boldsymbol{b}_{\boldsymbol{i}}, \boldsymbol{\alpha}\} = \alpha_1 \mu_i(t) + \alpha_2 \mu_i'(t). \tag{23}$$

The method argument of **jointModel(·)** specifies several types of the baseline hazard function $h_0(t)$ of the survival submodel which can be used and several types of numerical integration method. One common method is $\boldsymbol{method = ''piecewise - PH - GH''}$, which means the relative risk model Equation 16 has a piecewise-constant baseline risk function. In particular, the function can be wirtten as

$$h_0(t) = \Sigma_{q=1}^{Q} \xi_q I(v_{q-1} < t \leq v_q), \tag{24}$$

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

$\boldsymbol{method = "piecewise - PH - GH"}$, and assume the baseline risk function is assumed piecewise constant with six knots placed at equally spaced percentiles of the observed event time, and use the Gauss-Hermite integration rule to approximate Equation 20.

Maximization of the log-likelihood function corresponding to Equation 20. The maximum likelihood estimation usually involves intractable integrals that may be caused by high dimensions. To resolve such a problem, one effective method is to reduce the high dimensional data. Two Monte Carlo methods are proposed for approximating the integrals involved in the reduced score equation, one based on direct Monte Carlo integration and the other based on a stochastic version of the Gauss–Hermite quadrature. Monte Carlo EM algorithms are then conducted to solve such a challenging computation. One of the most important advantages of such algorithms is that it only simulates random effects in the first iteration; hence, the computation burden can be reduced and the likelihood that the increasing property of the original EM algorithm can be preserved [26].

4.4.2 Results based on Joint Models with Likelihood Methods

The results of the two joint models with likelihood methods under different association structures are shown in this section. One is to fit Equation 16 with 'current value' association, which formally can be be expressed as

$$h_i(t) = h_0(t)exp(\alpha_1 \mu_i(t)). \tag{25}$$

The other is to fit Equation 16 with 'current and slope value' association, which formally can be written as

$$h_i(t) = h_0(t)exp(\alpha_1 \mu_i(t) + \alpha_2 \mu_i'(t)), \tag{26}$$

where

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

The results of the maximum likelihood estimates for equations Equation 25 and Equation 26 are summarized in Table 4.6. The coefficient α_1 represents the strength of the correlation between the "current value" and the risk of developing the disease, while α_2 represents the strength of the correlation between the rate of change and the risk of experiencing the disease. $\alpha_1 = -0.3943$ for the "current value" association can be interpreted as a one unit difference in the current true value of $\mu_i(t)$ causing a 67.42% decrease in the risk of progression, calculated as $(1 - exp(-0.3943)) \times 100$. $\alpha_2 = 3.9933$ for the "current value plus slope" association can be interpreted as a one unit difference in the rate of change of the current value of $\mu_i(t)$ causing a 5325.35% increase in the risk of progression, given the constant current value. It is important to note that the p-values for all of these estimates are less than 0.05, indicating that it is significant and not likely that the estimates are zero.

To further illustrate the advantages of using joint modeling, we conduct a comparison between a standard time-dependent Cox proportional hazards (PH) model with the log of **lesion_area** and a joint model incorporating the "current value" association. The results

Table 4.6: Joint Models with Likelihood Inference 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	-

show that the joint model has a larger standard error, indicating a stronger bias in the estimation of the effect of lesion area, with estimated regression coefficients of -0.3427 for the time-dependent Cox model and -0.3943 for the joint model.

Akaike Information Criterion(AIC) and Bayesian Information Criterion(BIC) are good approaches for comparing the performance and accuracy of these two likelihood joint models. 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. Generally, adding parameters when fitting models will increase the likelihood, while uncontrolled adding parameters may result in overfitting. Both BIC and AIC resolve such a problem by including a penalty term for the number of parameters in the model. Typically, the penalty term is larger in BIC than in AIC. The results are summarized in Table 4.7. After comparing the AIC and BIC values, it is evident that the 'current value' association joint model has a better fit, as indicated by its lower AIC and BIC values.

Table 4.7: AIC and BIC of the Joint Models with Likelihood Inference

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

4.4.3 Diagnosing the Joint Models with Likelihood Methods

The following diagnostics are based on the joint model with the 'current value' association. Since, after comparing the joint models Equation 25 and Equation 26 using *anova*$(\cdot)$ function, no significant difference shows out. Figure 4.10 shows the diagnostic plots for the fitted joint model. The top two shows the diagnostics of the longitudinal analysis. To be specific, the top-left panel describes the within-subject residuals for the longitudinal analysis versus their corresponding fitted values. The top-right panel describes the Q-Q plot of the standardized subject specific residuals for the longitudinal analysis. The bottom-left estimates of the marginal survival function for the survival analysis. The bottom-right describes an estimate of the marginal cumulative risk function for the survival analysis.

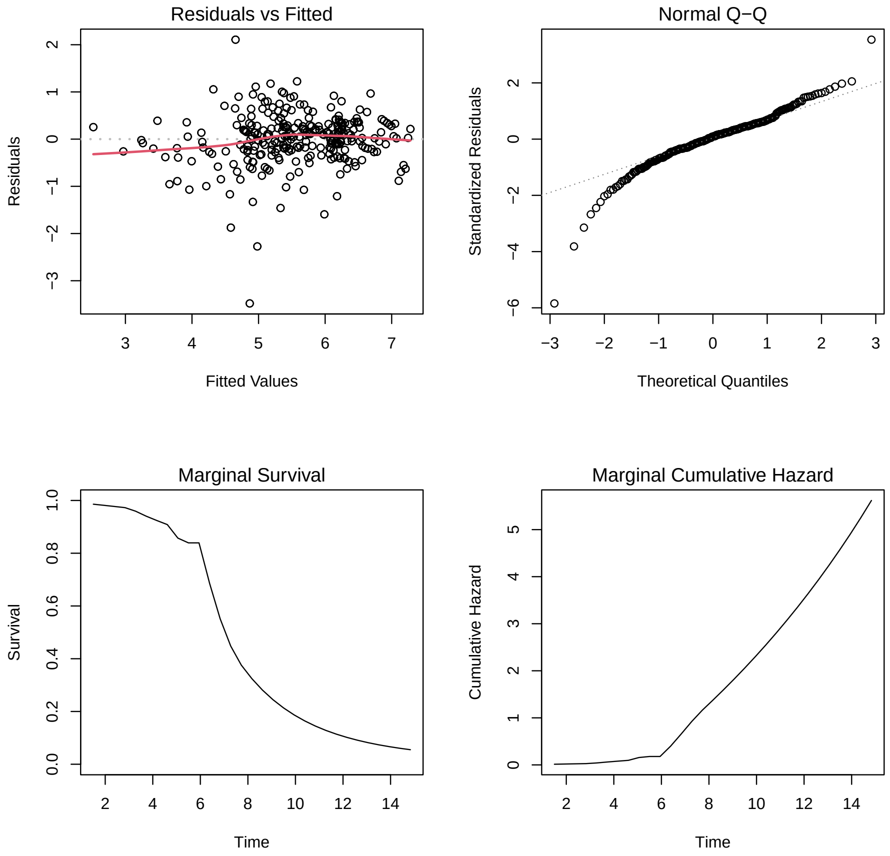

Figure 4.10: Diagnostic Plots for the Fitted Joint Model with Likelihood Method

4.5 Investigating the Joint Models using the Bayesian Methods

In this section, we demonstrate an analysis of joint models with Bayesian inference which is implemented by the **jointModelBayes(·)** function in R package *JMbayes*. The joint model can be shown as

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

As aforementioned, Bayesian methods provide good estimations by borrowing information from similar studies or experts, which means they are capable of incorporating the prior information into the current studies. The general concept for Bayesian method to do estimation is to assume the prior distributions $f(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$ based on the hyper-parameter $\boldsymbol{\theta}_0$ for the unknown parameters $\boldsymbol{\theta}$ in the models. Then make inference based on the posterior distributions $f(\boldsymbol{\theta}|\boldsymbol{y})$ of the parameters given the observed data $\boldsymbol{y}$. This can be mathematically written as

$$\begin{aligned} f(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}). \end{aligned} \tag{27}$$

Two parts should be solved to achieve the inference in Equation 27, one is the prior distribution $f(\boldsymbol{\theta})$ and the other is the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$.

4.5.1 Joint Models with 'Shared Parameters' Association

It is important to note that there are three structures of association available for joint modeling with Bayesian inference: the 'current value' association, the 'current value plus slope' association, and the 'shared parameters' association. In **Section 4.4**, we focus on the likelihood of the first two structures. This section will switch gears to the 'shared parameters' association, also known as the 'shared random effects' parameters. As the name suggests, the vector of time-independent random effects $\boldsymbol{b}_i$ exists in both the longitudinal and survival analysis. These random effects capture both the association between the longitudinal and event outcomes and the correlation between repeated measurements in the longitudinal analysis. In the following, we will provide more details about the inference of the joint model with Bayesian inference under the assumption of the 'shared random effects' parameters. The joint model with the 'shared parameters' association can be written as

$$h_i(t) = h_0(t)exp(\boldsymbol{\alpha}\boldsymbol{b}_i) = h_0(t)exp(\alpha_0 b_{0i} + \alpha_1 b_{1i}), \tag{28}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ denotes the strength of association between the random effects from longitudinal analysis and the survival analysis, and $\boldsymbol{b}_i = (b_{0i}, b_{1i})$ denotes the random effects. This association is computationally simpler than the "current value" and the "current value plus slope" associations, because the associative part in Equation 28 is time-independent. To be specific, for each specific subject i, the random effect $\boldsymbol{b}_i$ is unique. In this study, we assume the random effects $\boldsymbol{b}_i = (b_{0i}, b_{1i}) \sim N(0, D)$, the error in longitudinal analysis $e_i \sim N(0, R_i)$, and b_i and e_i are independent. D and R_i are covariance matrices for the random effects and repeated measurements within each individual respectively.

For the joint model with '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 analyses are independent, and the longitudinal trajectories for different subjects are also independent.

For the expression of the likelihood of Equation 28, it can be written as

$$p(T_i, \delta_i, y_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = p(T_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{\theta})p(y_i|\boldsymbol{b}_i; \boldsymbol{\theta}), \tag{29}$$

$$p(y_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = \prod_j p\{y_i(t_{ij})|\boldsymbol{b}_i; \boldsymbol{\theta}\}, \tag{30}$$

where T_i is the observed failure time for the subject $i(i = 1, 2, \dots, n)$, δ_i is the indicator of censoring, and y_i is the observed longitudinal outcome. $\boldsymbol{\theta} = (\theta_t, \theta_y, \theta_b)$ contains all of the unknown parameters. $\theta_t = \alpha$ denotes the parameters for the event time outcome, $\theta_y = (\boldsymbol{b}_i, \boldsymbol{\beta}, D, R_i)$ contains the parameters for longitudinal outcome and $\theta_b = (\boldsymbol{b}_i, D)$ denotes the unique parameters of the random-effects covariance matrix. After adding the prior distribution $p(\boldsymbol{\theta})$, the posterior distribution can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{b}) \propto \prod_i^n \prod_j^{n_i} p(y_i(t_{ij})|\boldsymbol{b}_i; \boldsymbol{\theta})p(T_i, \delta_i|\boldsymbol{b}_i, \boldsymbol{\theta})p(\boldsymbol{b}_i|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{31}$$

where

$$p(y_i(t_{ij})|\boldsymbol{b}_i; \boldsymbol{\theta}) = exp\{[y_{ij}\psi_{ij}(\boldsymbol{b}_i) - c\psi_{ij}(\boldsymbol{b}_i)]/a(\psi) - d(y_{ij}, \psi)\}. \tag{32}$$

$y_{ij} = y_i(t_{ij})$ means the the longitudinal observed value for individual $i(i = 1, 2, \dots, n)$ at time $j = (1, 2, \dots, n_i)$. $\psi_{ij}(\boldsymbol{b}_i)$ denotes the natural and ψ denotes the dispersion parameters in the exponential family. $c(\cdot)$, $a(\cdot)$, and $d(\cdot)$ are known functions specifying the member of the exponential family. 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

$$p(T_i, \delta_i|\boldsymbol{b}_i; \theta_t) = [h_0(t_i)\exp\{\boldsymbol{\alpha}\boldsymbol{b}_i\}]^{\delta_i} \times \exp[-\int_0^{t_i} h_0(u)\exp\{\boldsymbol{\alpha}\boldsymbol{b}_i\}du. \tag{33}$$

The function **jointModelBayes(·)** in *JMbayes* package is used to do the basic model-fitting.

4.5.2 Results based on the Joint Models with Shared Parameter Association

After fitting the model Equation 28 using the **jointModelBayes(·)** function in the *JMbayes* package, the results 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^2. According to the parameters of $\boldsymbol{\alpha}$, patients with lower baseline levels of the longitudinal outcome (i.e. intercept) and steeper decreases in their longitudinal trajectories (i.e. slope) are more likely to experience the event. However, the results suggest that both the baseline levels of the underlying lesion area (represented by α_0) and the longitudinal evolution (represented by α_1) 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 Inference 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 other summary statistics 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. LPML is a measure of model fit, while DIC is a popular information criterion used for model selection.

4.5.3 Analysing Joint Models with Other Associations

To compare the parameter estimates computed using different methods under different associations, two more models were conducted using Bayesian Methods, which is completed by the function **jointModelBayes(·)** in *JMbayes* package in R. The results can be found in Section 4.4.2. In Table 4.9, α_1 quantifies the association between the current features of the marker process up to time t and the hazard of progression at the same time point, and α_2 describes the strength of the correlation between the slope of features of the marker process up to 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, as are the values in the 'shared parameters' association joint model.
The following relies on the DIC to compare the three models with different associations based on Bayesian inference, with smaller values indicating better model adjustments to the data. The results show in Table 4.10. According to the DIC value, these three models show very similar performance, albeit the joint model with 'current value' association has the best performance. Recall that this model includes the current true value of lesion area as predictors of the risk of progression.

Table 4.9: Comparing Joint Models using Bayesian Inference with 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	-

Table 4.10: DIC of Joint Models with Bayesian Inference on Different Associations

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

4.5.4 Diagnosing the Joint Models with Bayesian Methods

Different diagnostic approaches are conducted to assess the performance of the three models estimated through Bayesian inference. A number of parameters play a role in this section. β_0 represents the fixed effect parameter for the intercept, while β_1 represents the fixed effect parameter for the variable **obstime**. In the joint model with ”shared parameters” association, α_0 and α_1 denote the correlation between the random effect of the intercept and **obstime** and the survival analysis, respectively. In the joint model with ”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 the joint model with ”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

Recall that the baseline hazard function $h_0(t_i)$ in the survival model uses the B-spline approach in the *JMBayes* package, and the number of knots will influence the flexibility of the model. To conduct a concrete investigation about the impact the number of knots for the baseline hazard can have on the final results, 12 models are built, with each association building 4 models with 11 knots, 13 knots, 15 knots, and 17 knots. Table 4.11 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 'shared parameters' and 'current value plus slope' joint models are stable but show erratic behavior in the 'current value' joint models, indicating that they may not have converged well.

Table 4.11: Posterior Means for the Fixed Effects of the Longitudinal Submodel

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 shows the posterior means for the strength of association of the survival submodel for the 12 models. The parameters in the joint models with 'shared parameters' and 'current value' associations are consistent. However, in the joint models with 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 the Survival Submodel

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

To summarize, the joint models with "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.

The following analyzes the sensitivity of the parameter estimates to 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, the Bayesian method will be comparable to the likelihood method. In other words, the larger the variance of the prior, the closer the results from these two methods will be. However, if the results from the Bayesian method are credible, they should not be affected by the prior variance defined. To conduct a concrete investigation about the impact of the prior assumption, 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 shows the posterior means for the fixed effects of the survival submodel for these 12 models. According to Table 4.14, the choice of prior variance does not significantly affect the parameter estimates for 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 the Survival Submodel Under 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 the Survival Submodel Under 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

The trace plot shows the history of a parameter value across the iteration of the chain. It can be utilized to make sure that the prior distribution is well calibrated, which is indicated by the parameters having sufficient state changes as the 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. The following figures, from Figure 6.1 to Figure 6.11, show the trace plots for the three different joint models with 'shared parameters', 'current value', and 'current value plus slope' associations estimated by Bayesian inferences, respectively. All the trace plots can be found in the Appendix.

The trace plots in Figure 6.1 to Figure 6.4 for the 'shared parameters' association model show that the parameters intercept (β_0), obstime (β_1), and obstime (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 the intercept (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 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 the strength of the association (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 the strength of the current association (α_1) and the slope association (α_2) in the survival submodel show long-term trends, indicating the need for more iterations.

In conclusion, one or more unstable parameters are present in all three models estimated using Bayesian methods. Therefore, future analysis should consider using more iterations.

Autocorrelation Plots

Autocorrelation plots [27] are a commonly used tool for checking randomness in a data set. They are 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 values should be near zero for all time-lag separations. This is important for Bayesian inferences 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 different models estimated by Bayesian inferences with 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 the intercept (β_0) in the longitudinal analysis. The upper right panel shows the autocorrelation of obstime (β_1) in the longitudinal analysis. The lower left panel shows the autocorrelation of the intercept (α_0) in the survival analysis, and the lower right panel shows the autocorrelation of obstime (α_1) in the survival analysis. 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 the intercept (β_0) and obstime (β_1) in the longitudinal analysis, respectively. The lower left panel shows the autocorrelation of the parameter α_1, which measures the strength of correlation between the longitudinal and survival analysis. The top two plots express randomness of the parameters, 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 the parameter α_2. The top two plots express randomness of the parameters, 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 following is a description of the KDE plots for three different associations estimated by Bayesian inferences. All of the plots can be found in the Appendix as well.

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 the intercept (β_0) in the longitudinal analysis, which follows a normal distribution. Similarly, the upper right panel displays the PDF of **obstime** (β_1) in the longitudinal analysis, also following a normal distribution. The lower left panel shows the PDF of the intercept (α_0) in the survival analysis, which also follows a normal distribution. The lower right panel displays the PDF of **obstime** (α_1) in the survival analysis, 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 among various inference 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.

This section compares joint models estimated by both likelihood and Bayesian inference, considering both the 'current value' association and the 'current value plus slope' association. The results are summarized in Table 4.15.

The estimates of the parameters are similar, with the exception of the parameter α_2 in the 'current value plus slope' association. As discussed in **Section 4.5.4**, α_2 in this association joint model is sensitive to the choice of prior and the number of knots, and its computation fails to converge well. The standard errors of the parameter estimates in 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 'current value' association Bayesian joint model.

Table 4.15: Likelihood Joint Models and Bayesian Joint Models with Different Associations

		Likelihood Method			**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

In this report, we provide an overview of joint modeling for longitudinal and survival data. We first review the longitudinal and survival models to clarify the framework of the joint model. In general, we assume that the longitudinal submodel provides certain characteristics that are used in combination with the survival submodel to form various joint models. There are three commonly used association structures in practice: the ”current value” association, the ”current value plus slope” association, and the ”shared parameters” association. The ”current value” association assumes that the true (unobserved) value of the longitudinal analysis at one time point predicts the risk of the event occurring at that same time. The ”current value plus slope” association assumes that both the true value and the rate of change (slope) of the true value of the longitudinal analysis at one time point predict the risk of the event occurring at that same time. The ”shared parameters” association includes only the random effects from the longitudinal submodel as predictors in the hazard risk submodel, and is computationally simpler than the other two associations because the association part is time-independent, meaning that the random effects explaining individual specificities do not depend on time.

Two inference methods are popular for estimating the joint modeling of longitudinal and survival data: the likelihood method and the Bayesian method. The likelihood method involves finding the parameter values that maximize the likelihood of the model given the observed data. On the other hand, the Bayesian method combines prior information with the observed data to guide statistical inference. This approach treats 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 inference provides more information than the point estimators produced by maximum likelihood estimation (MLE) and can lead to more complex calculations. MLE produces a single fixed value, while Bayesian inference returns 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.

The R package *JM* can be used to fit a variety of joint models for normal longitudinal responses and survival data under maximum likelihood. The function **jointModel(·)** accepts a linear mixed effects object fitted by the function **lme(·)** from the package *nlme* and a survival object fitted by either function **coxph(·)** or function **survreg(·)** from the package survival as main arguments. The type of the survival submodel to be used and the type of numerical integration method to be conducted are able to be clarified as well, and Monte Carlo EM algorithms based on Gauss-Hermite integration rule is one of the most popular methods to approximate the integral. The R package *JMBayes* can be used to fit a variety of joint models for longitudinal responses and survival data under Bayesian methods. The function **jointModelBayes(·)** accepts a linear mixed effects object fitted by the function **lme(·)** from package *nlme* or by function **glmmPQL(·)** from package *MASS* and a survival object fitted by the function **coxph(·)** as main arguments. The baseline hazard is by default approximated using penalized B-splines. The **jointModelBayes($\cdot$)** function uses Markov chain Monte Carlo (MCMC) estimation to yield very precise estimates. Many effective diagnostic tools are available for evaluating MCMC convergence and model-fitting in the package *JMBayes*.

A practical example is provided to illustrate the implementation of two inference methods, specifically with different association structures. Likelihood joint models are fitted using the function **jointModel($\cdot$)**, with ”current value” and ”current value plus slope” associations, while Bayesian joint models are fitted using the function **jointModelBayes($\cdot$)**, with ”current value,” ”current value plus slope,” and ”shared parameters” associations. Model comparison is performed using 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 ”current value” association joint model outperforms the others with both the likelihood and Bayesian methods. However, as mentioned earlier, comparing the performance and accuracy of likelihood and Bayesian inference is challenging. Therefore, we focus on comparing the parameter values, diagnosing the models, and evaluating the residual standard errors. The results reveal that under the ”current value” association, these two inferences provide similar estimates, while under the ”current value plus slope” association, the estimates of α_2 differ due to convergence issues. Nevertheless, these two inference 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] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied longitudinal analysis*. John Wiley & Sons, 2012.

[5] S. P. Jenkins, “Survival analysis,” *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, vol. 42, pp. 54–56, 2005.

[6] I. Guler, C. Faes, C. Cadarso-Suárez, L. Teixeira, A. Rodrigues, and D. Mendonca, “Two-stage model for multivariate longitudinal and survival data with application to nephrology research,” *Biometrical Journal*, vol. 59, no. 6, pp. 1204–1220, 2017.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

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

[24] 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.

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

[26] 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.

[27] 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

6.2 Plots

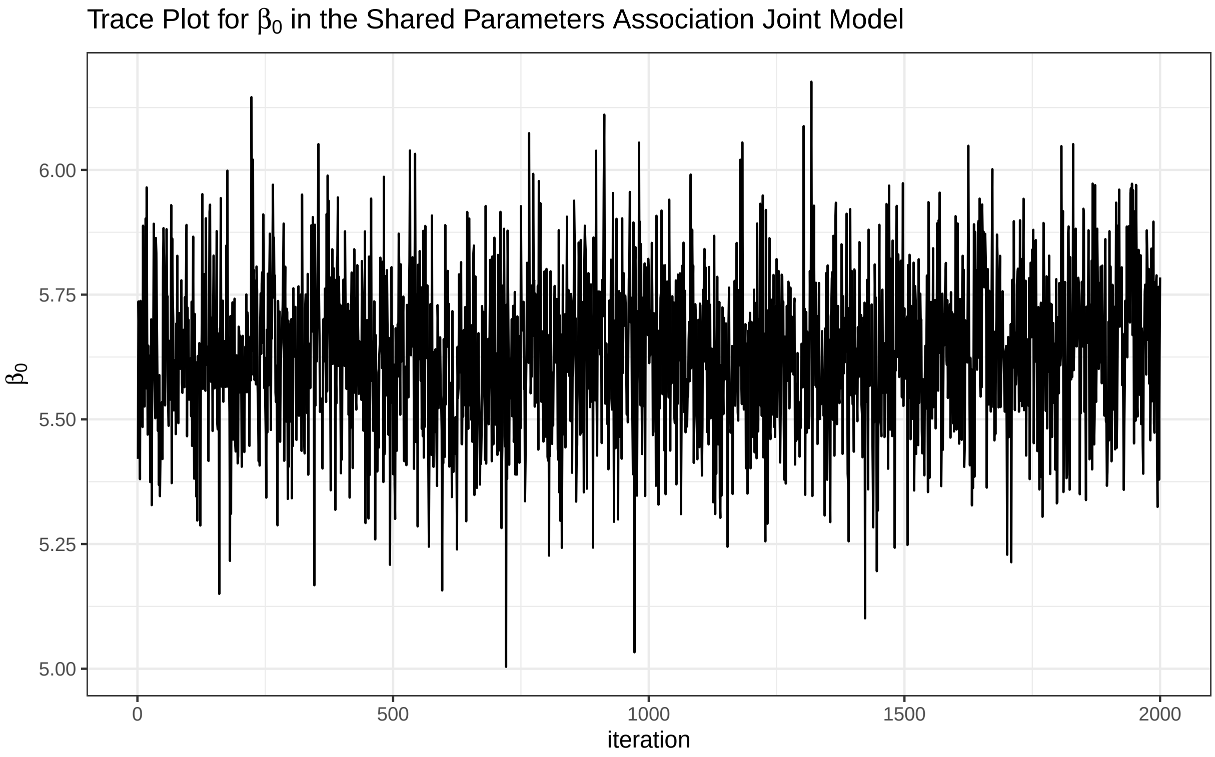

Figure 6.1: Trace Plot for β_0 of the Longitudinal Model in Shared Parameters Association Joint Model

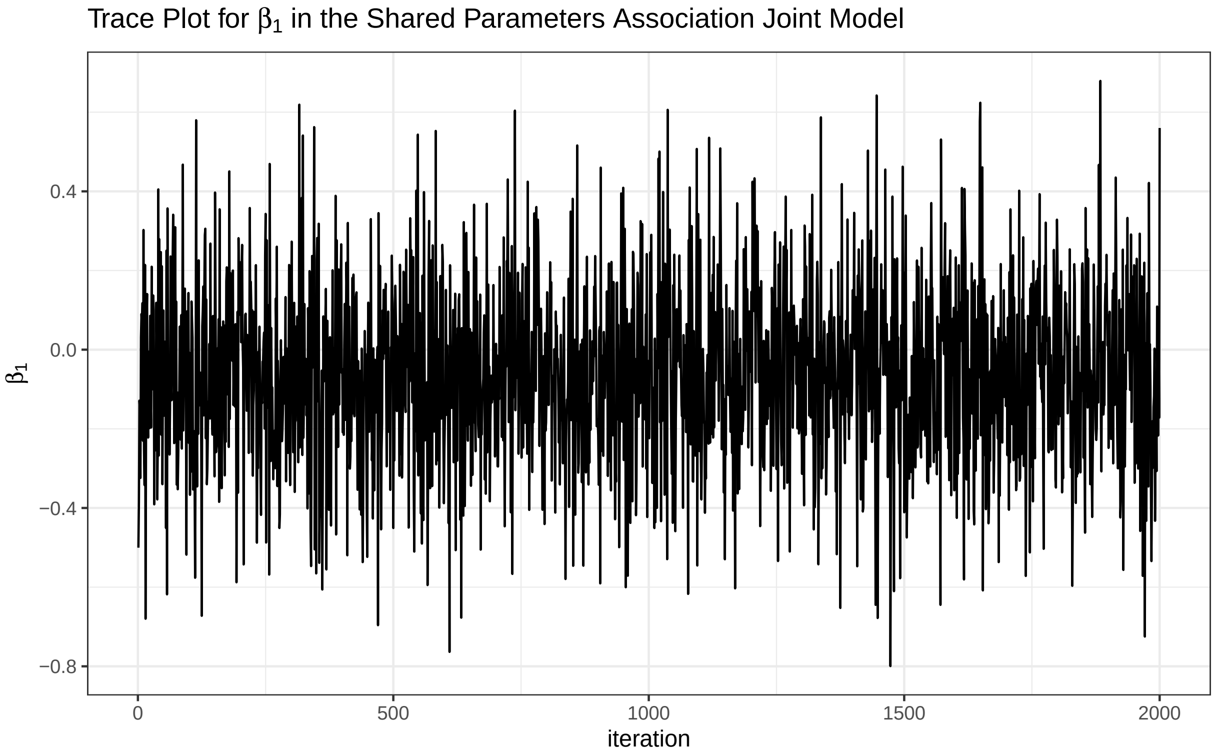

Figure 6.2: Trace Plot for β_1 of the Longitudinal Model in Shared Parameters Association Joint Model

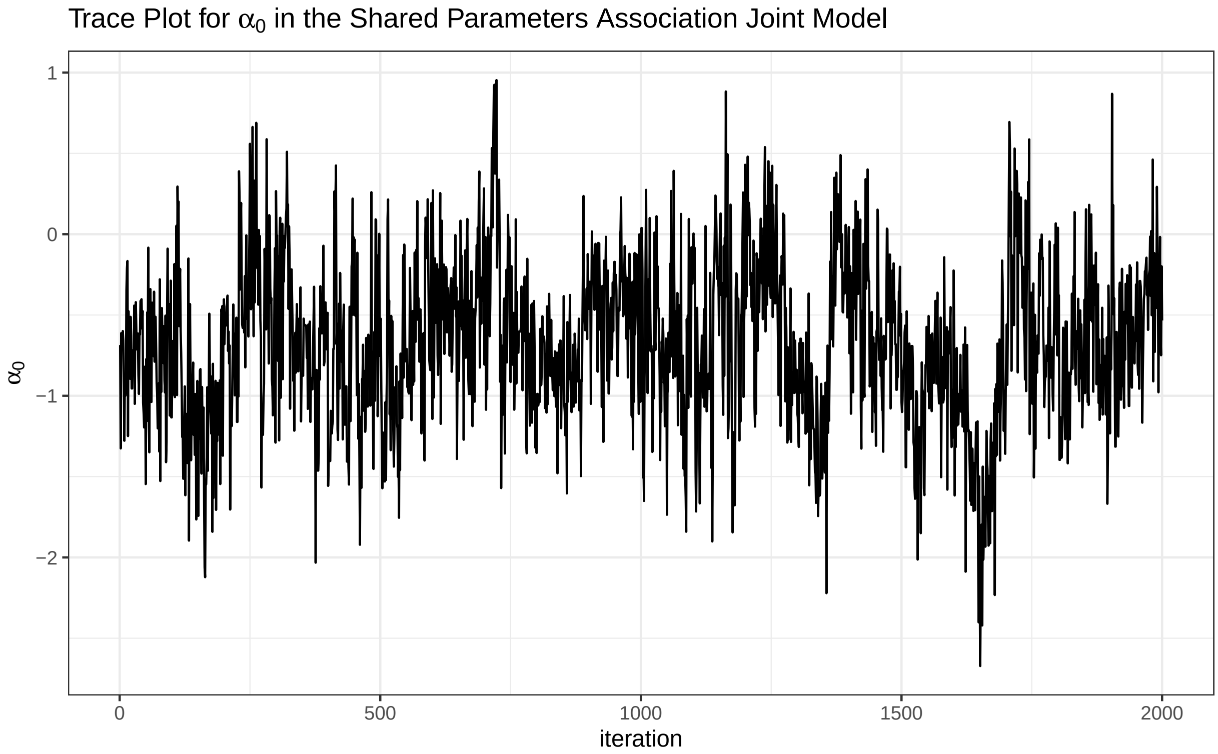

Figure 6.3: Trace Plot for α_0 of the Survival Model in Shared Parameters Association Joint Model

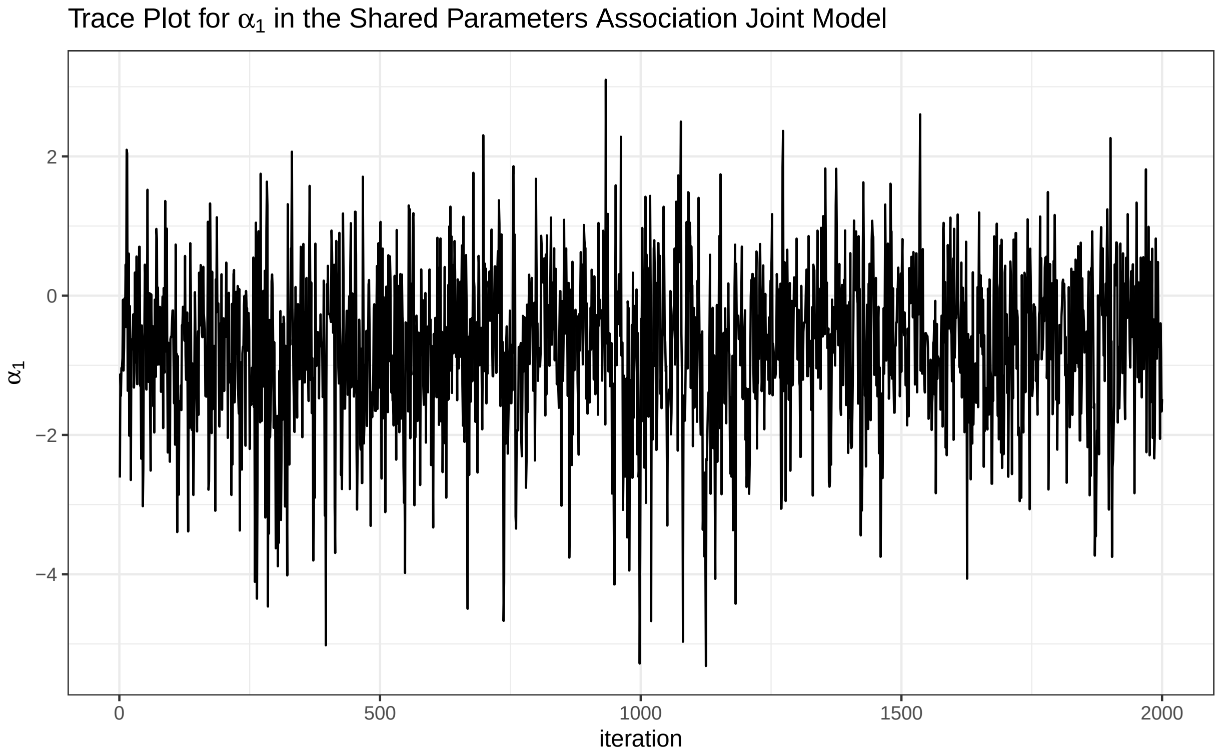

Figure 6.4: Trace Plot for α_1 of the Survival Model in Shared Parameters Association Joint Model

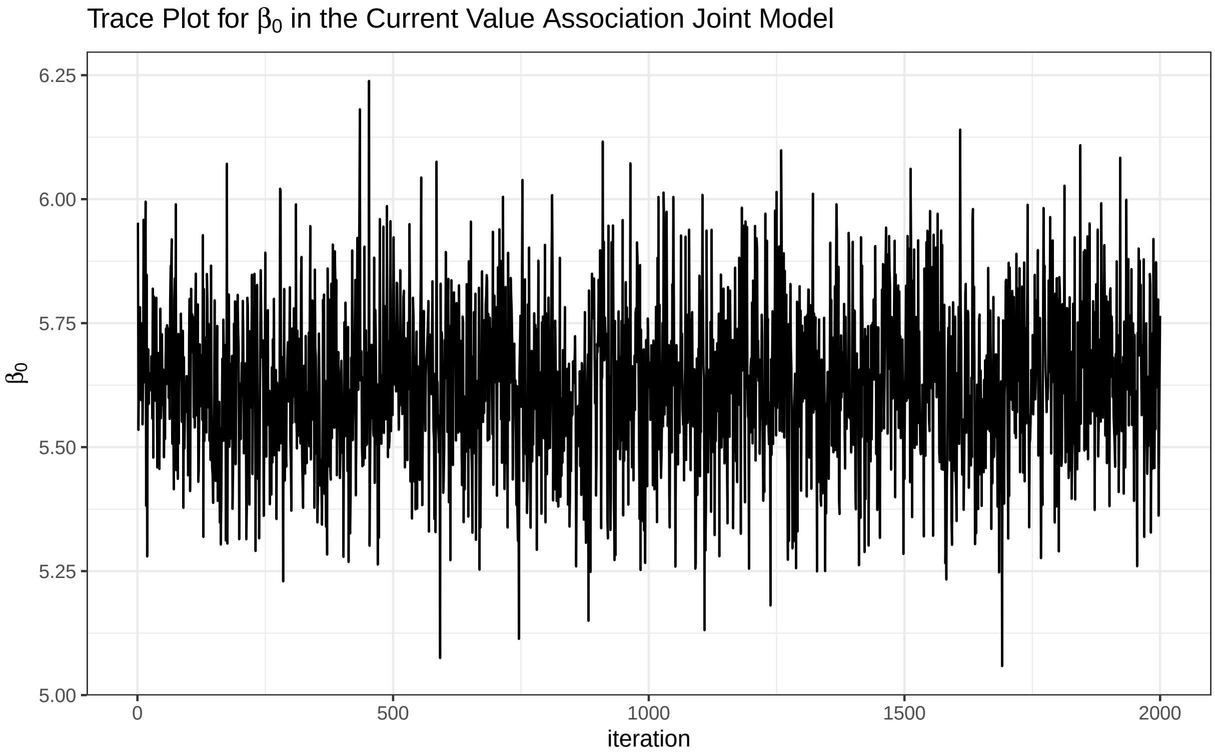

Figure 6.5: Trace Plot for β_0 of the Longitudinal Model in Current Value Association Joint Model

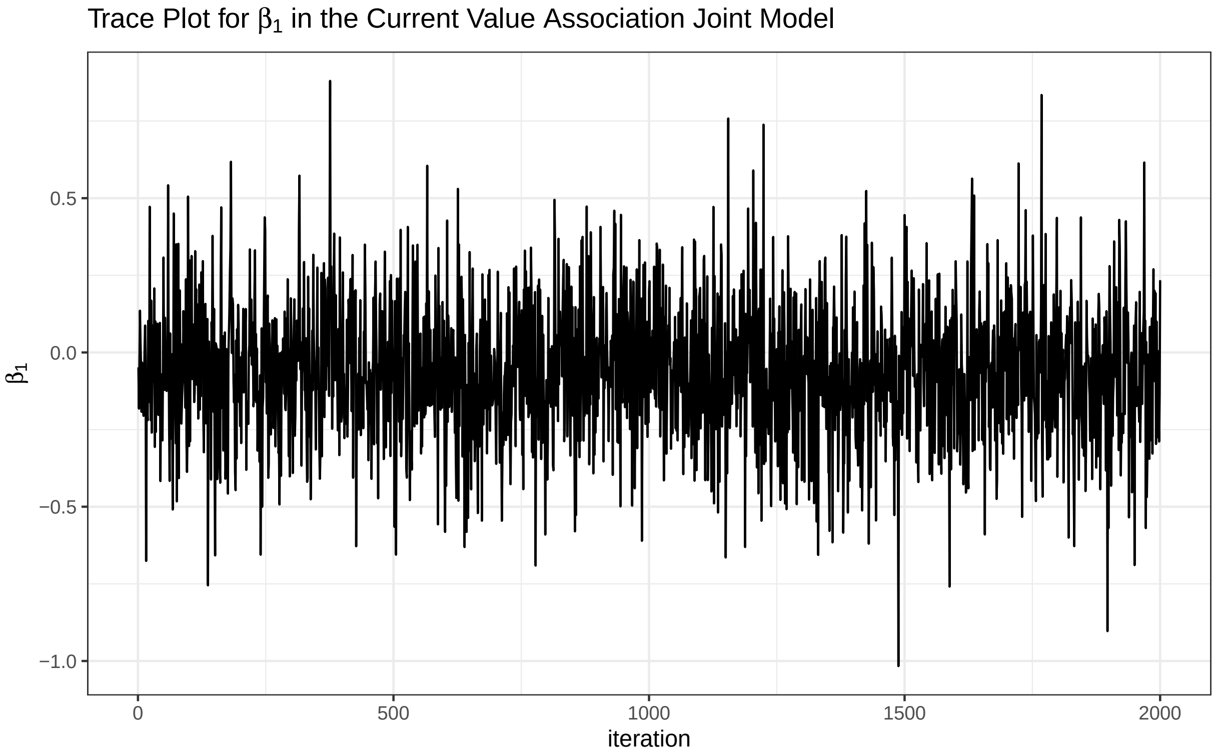

Figure 6.6: Trace Plot for β_1 of the Longitudinal Model in Current Value Association Joint Model

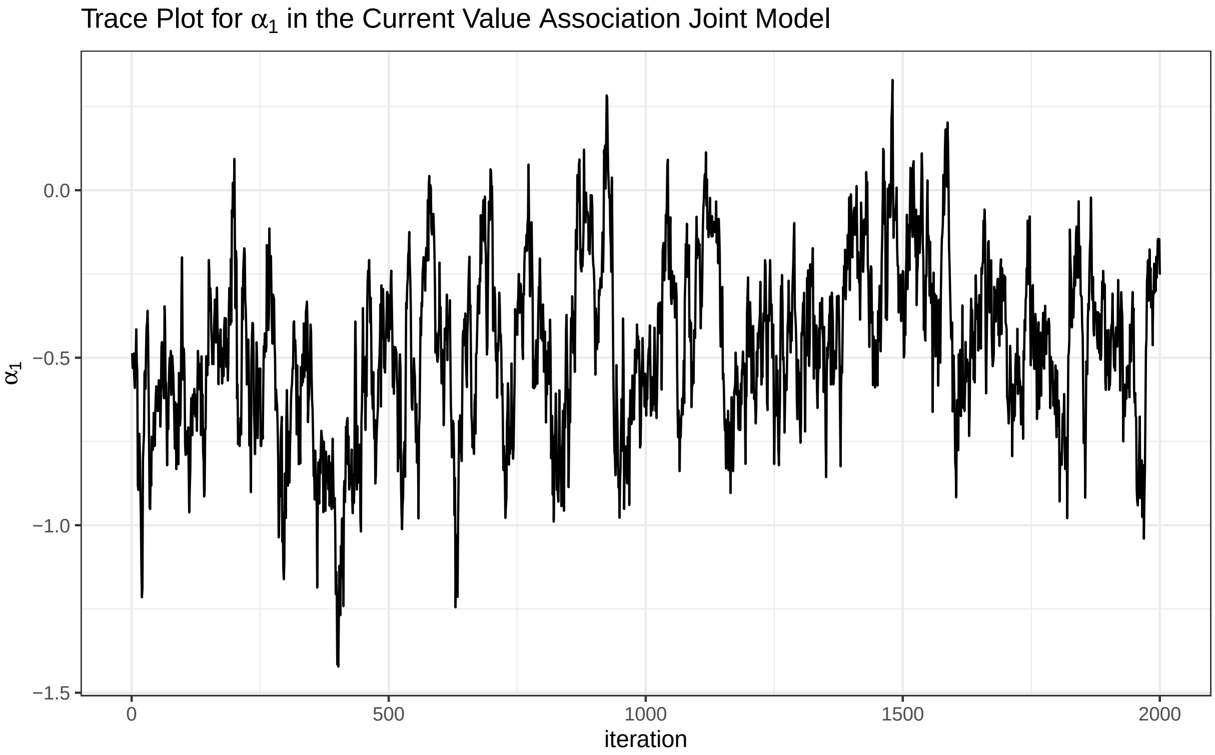

Figure 6.7: Trace Plot for α_1 of the Survival Model in Current Value Association Joint Model

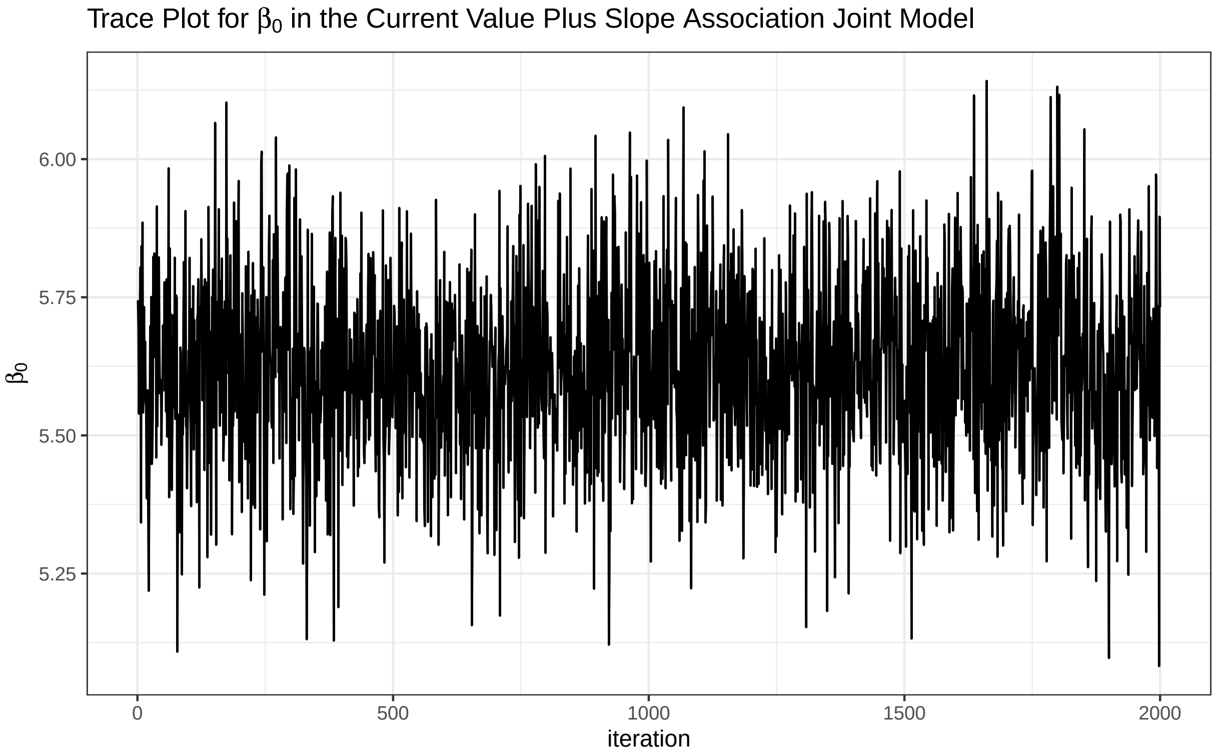

Figure 6.8: Trace Plot for β_0 of the Longitudinal Model in Current Value Plus Slope Association Joint Model

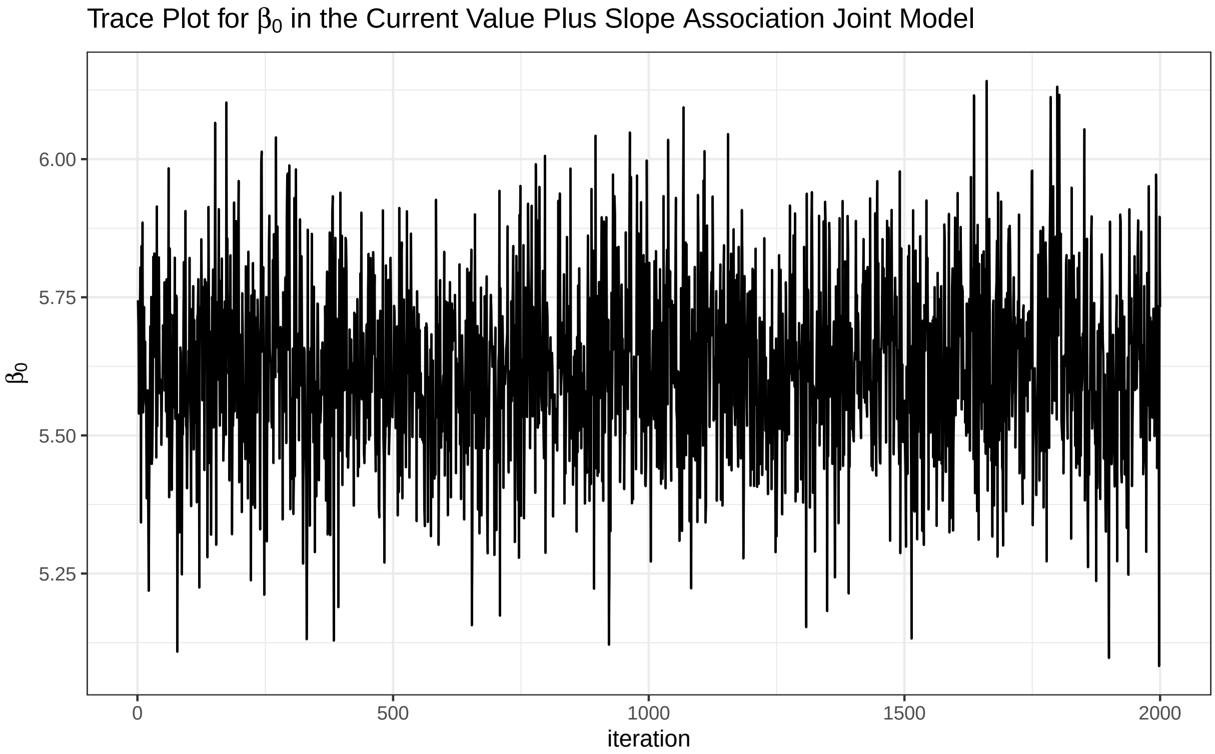

Figure 6.9: Trace Plot for β_1 of the Longitudinal Model in Current Value Plus Slope Association Joint Model

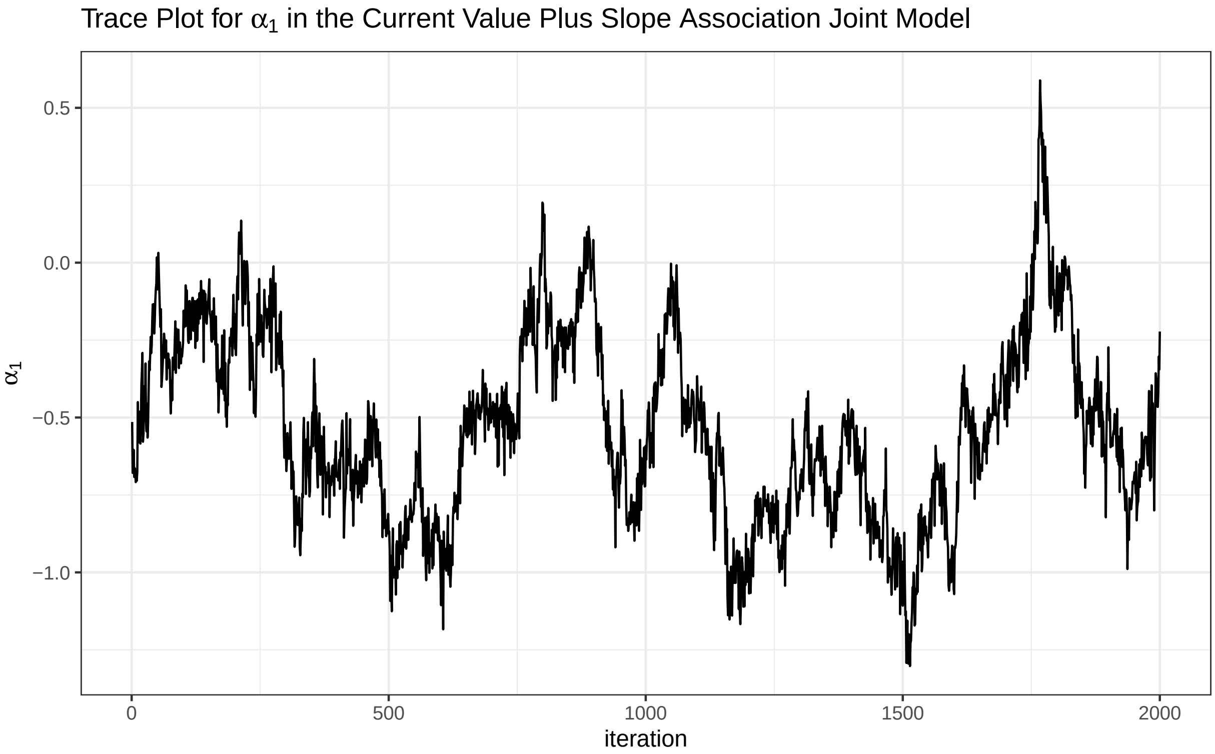

Figure 6.10: Trace Plot for α_1 of the Survival Model in Current Value Plus Slope Association Joint Model

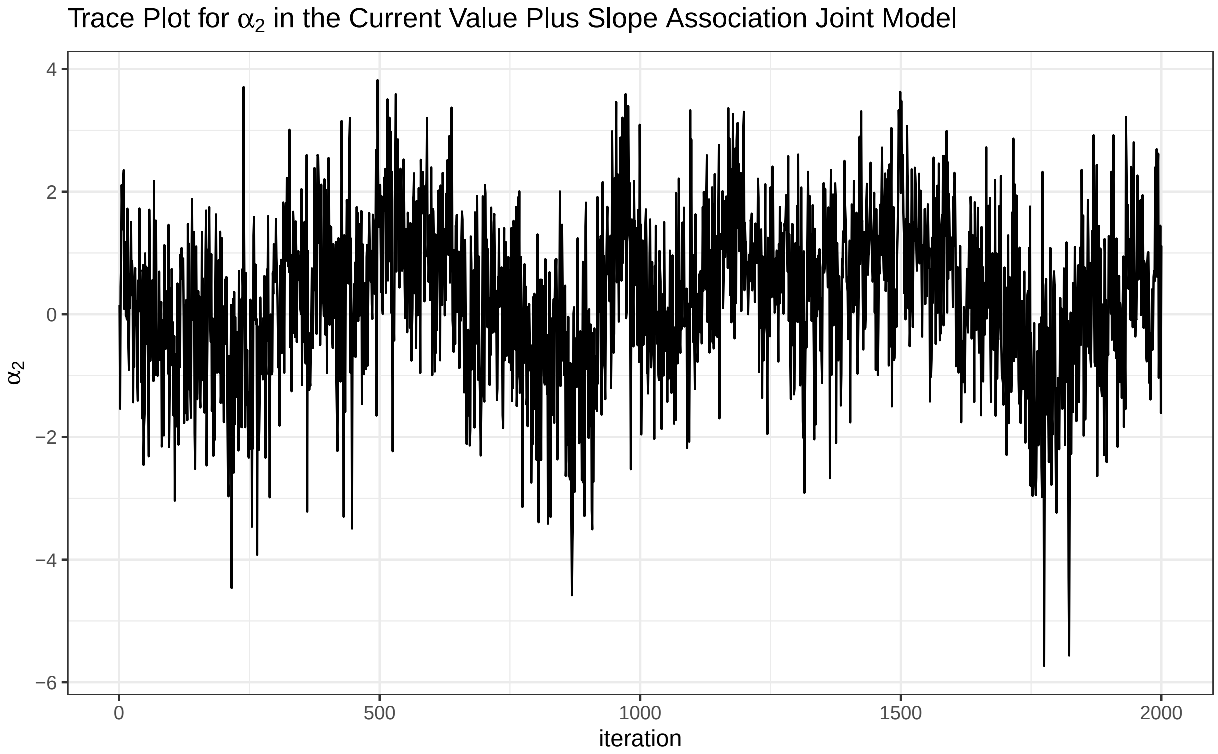

Figure 6.11: Trace Plot for α_2 of the Survival Model in Current Value Plus Slope Association Joint Model

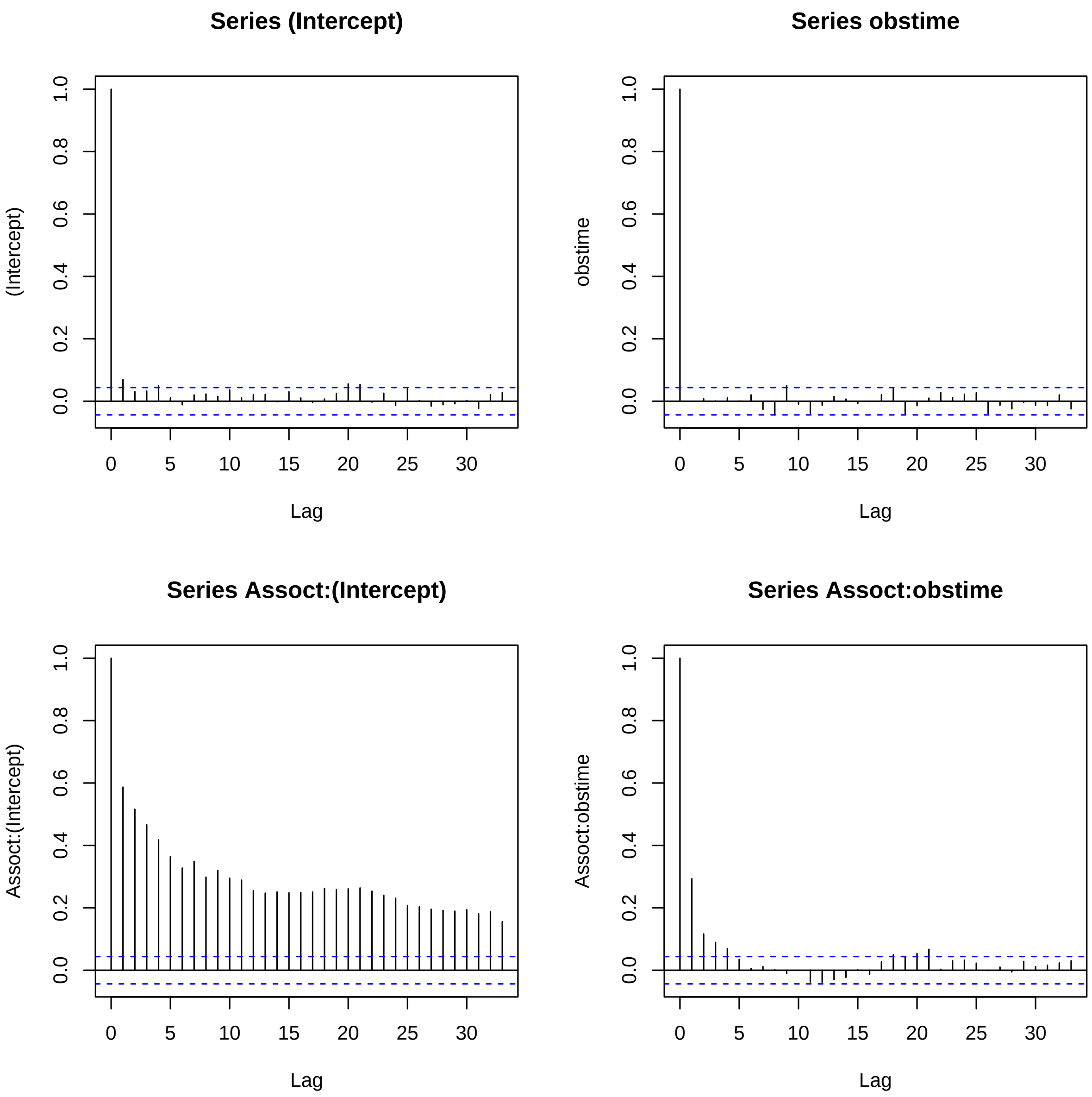

Figure 6.12: Autocorrelation Plots for the Parameters in Shared Parameters Joint Model

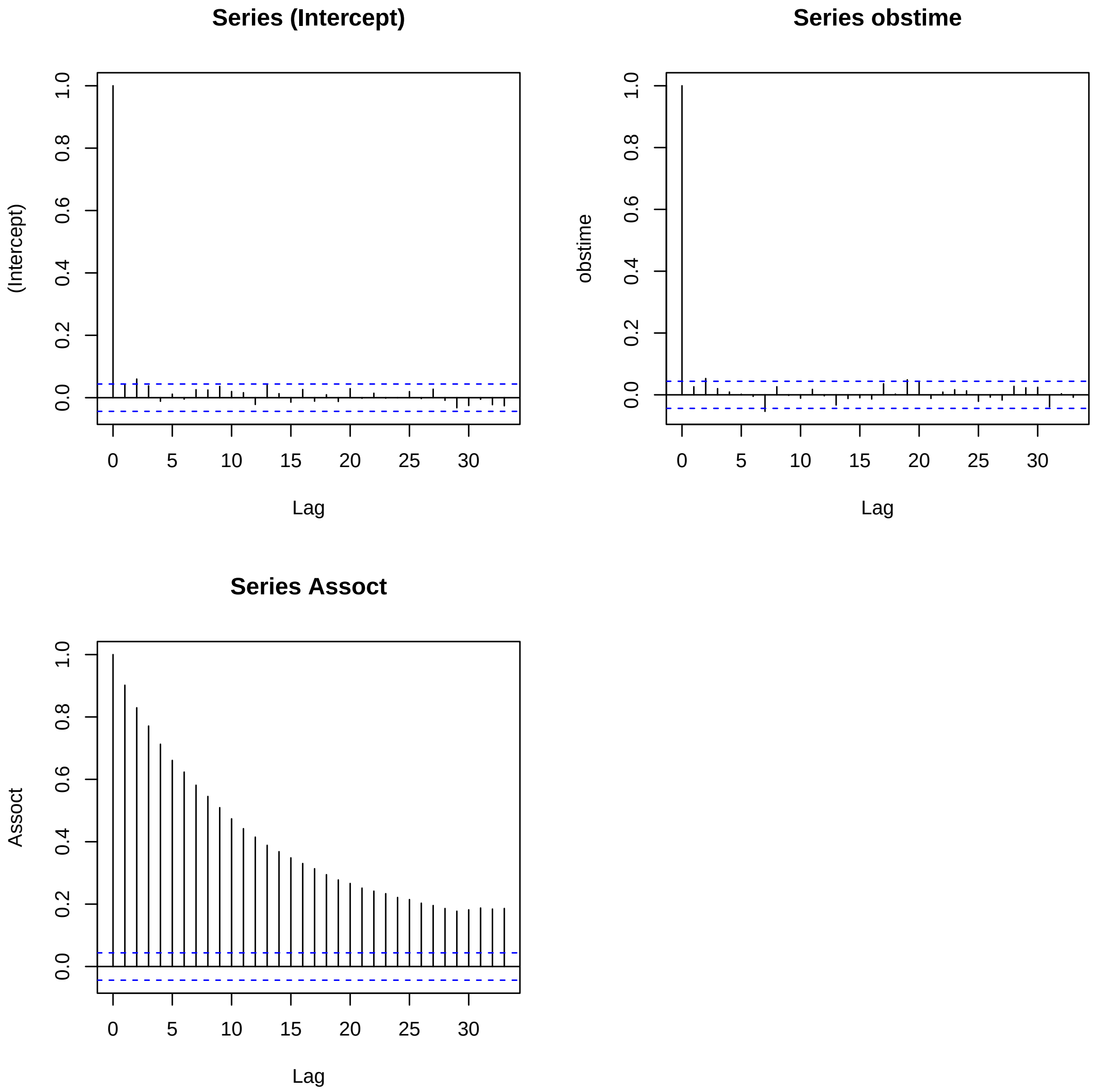

Figure 6.13: Autocorrelation Plots for the Parameters in the Current Value Association Joint Model

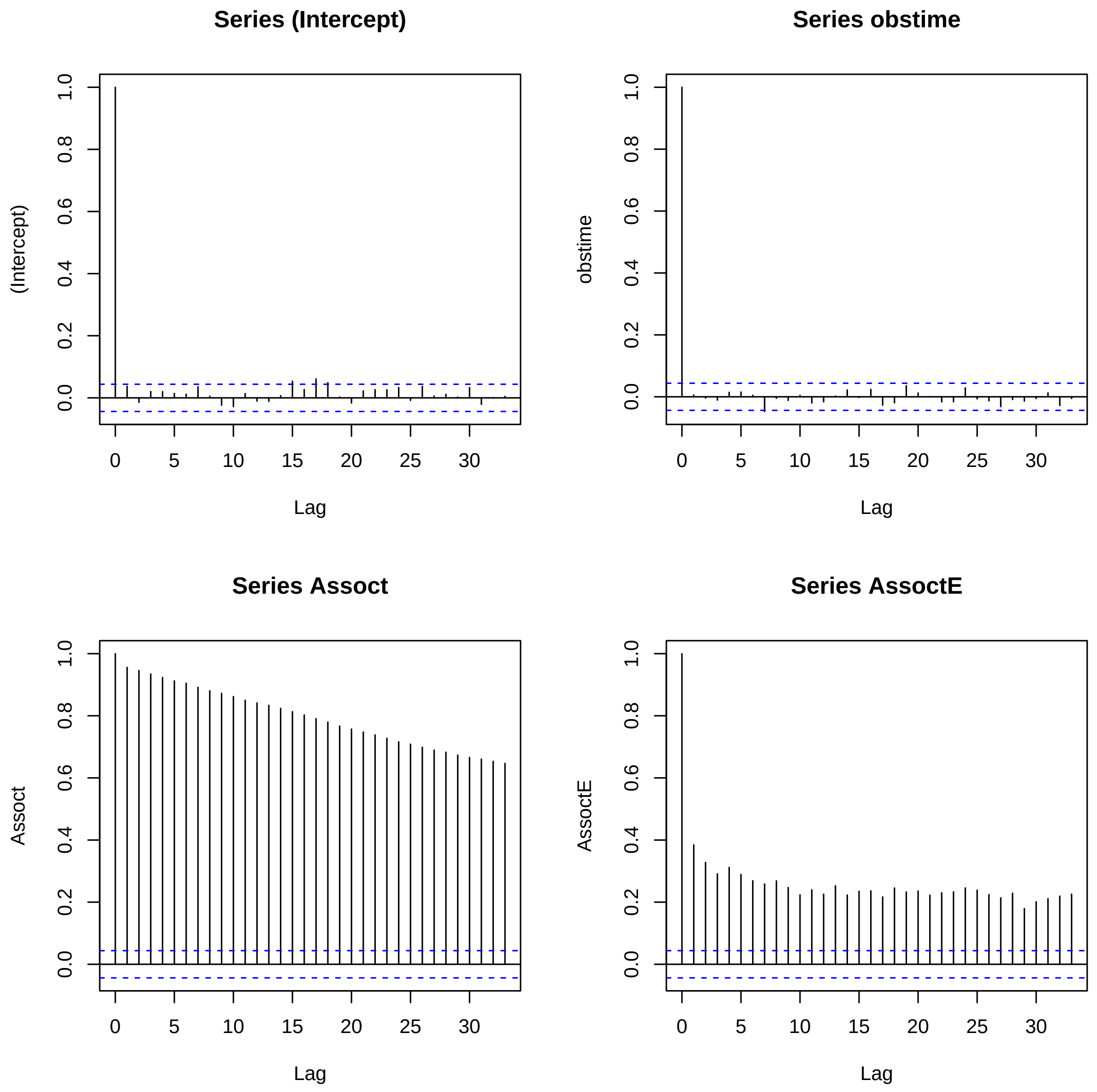

Figure 6.14: Autocorrelation Plots for the Parameters in the Current Value Plus Slope Association Joint Model

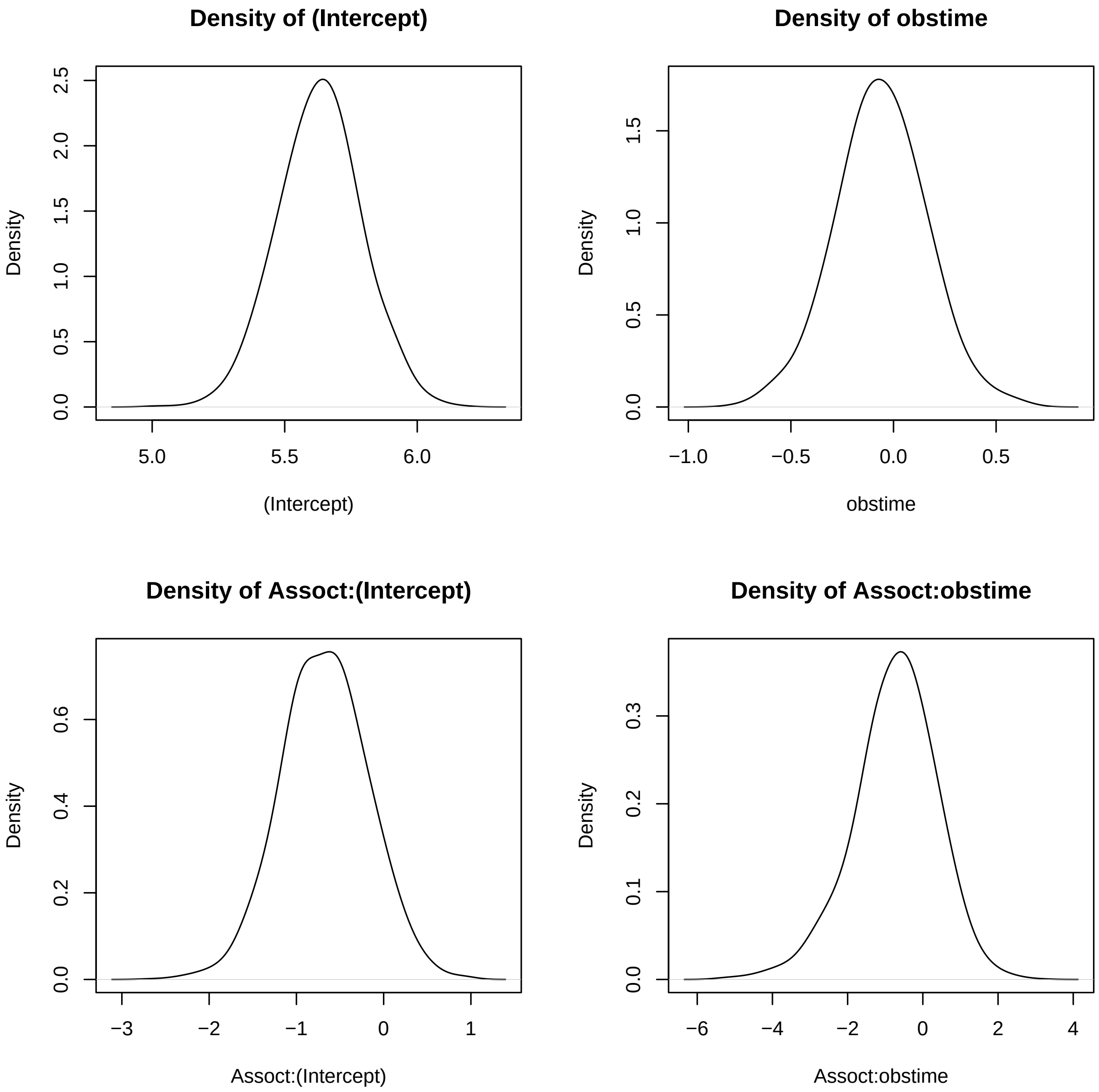

Figure 6.15: Trace Plot for the Parameters of the Longitudinal Model in Shared Parameters Joint Model

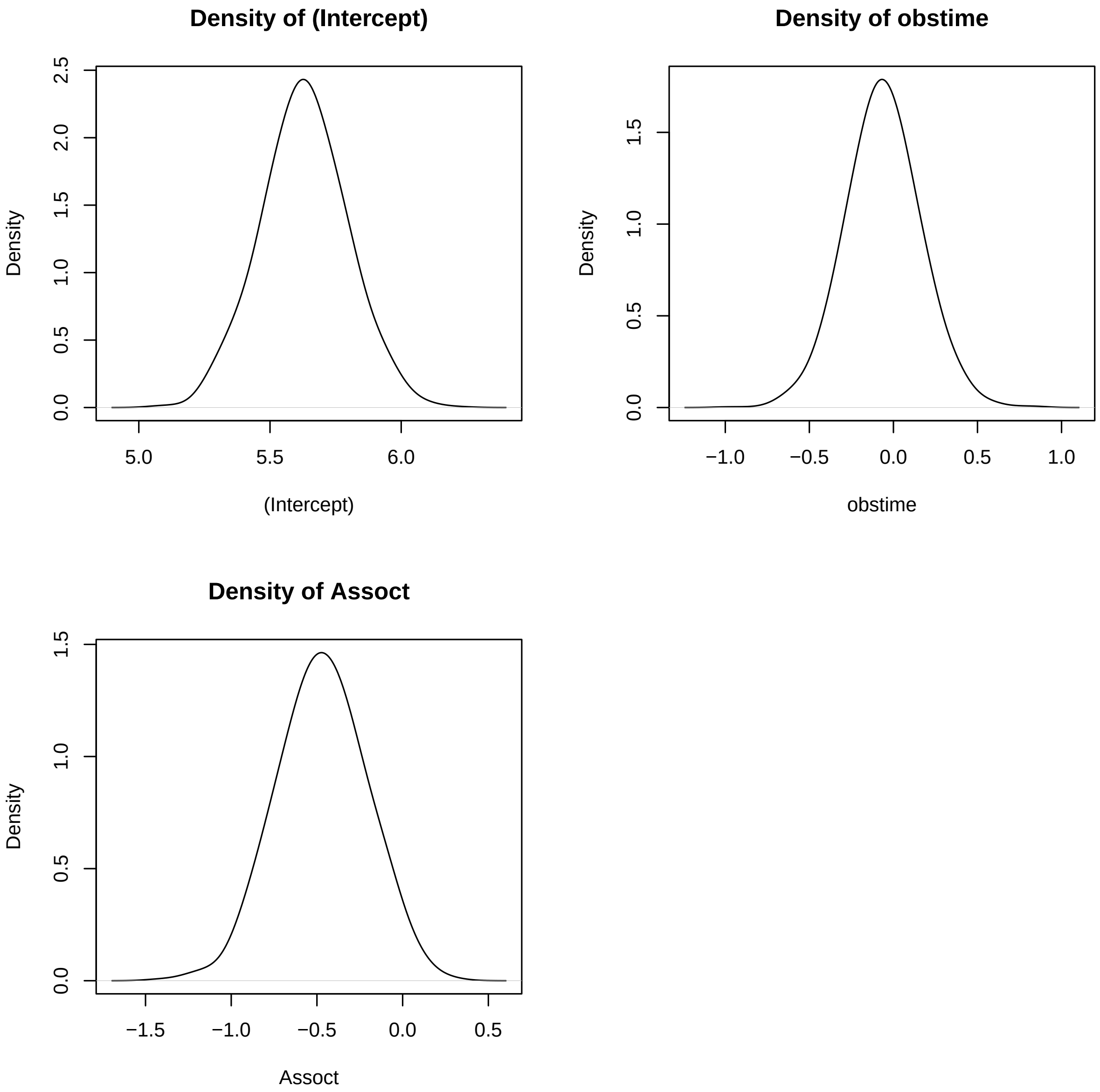

Figure 6.16: Trace Plot for the Parameters of the Longitudinal Model in the Current Value Association Joint Model

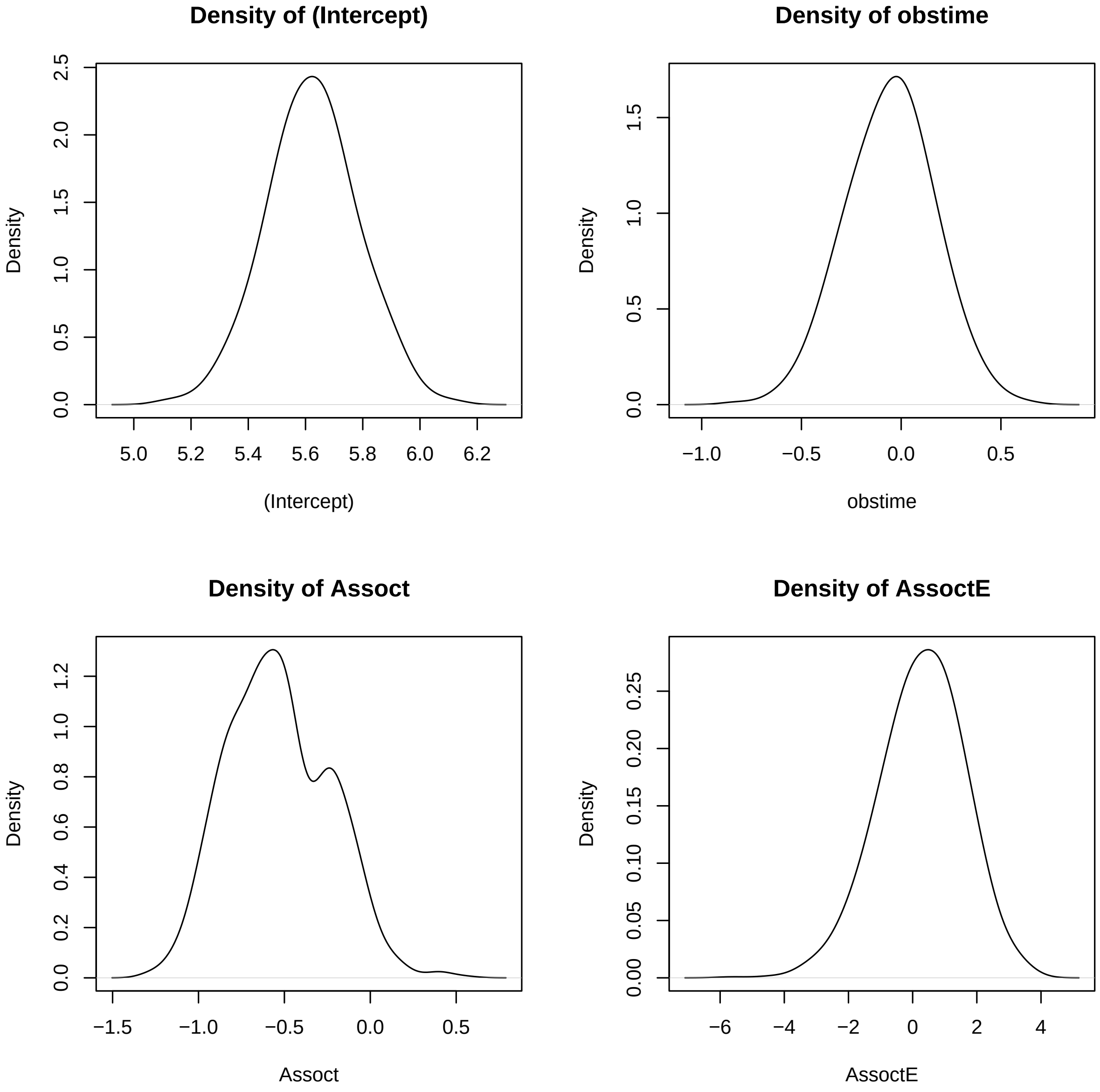

Figure 6.17: Trace Plot for the Parameters of the Longitudinal Model in the Current Value Plus Slope Association Joint Model