Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

SHULING LIU DATE

Joint Modeling Approaches for Clustered Survival Data with Random Cluster Size

By

Shuling Liu
B.E., Renmin University of China, 2007
M.A., State University of New York at Buffalo, 2009

Dissertation

Doctor of Philosophy

Biostatistics

Emory University

Amita K. Manatunga, Ph.D.
Advisor

Robert H. Lyles, Ph.D.
Committee Member

Michele Marcus, Ph.D.
Committee Member

Limin Peng, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

2015

Joint Modeling Approaches for Clustered Survival Data with Random Cluster Size

By

Shuling Liu
B.E., Renmin University of China, 2007
M.A., State University of New York at Buffalo, 2009

Advisor: Amita K. Manatunga, PhD

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2015

Abstract

Joint Modeling Approaches for Clustered Survival Data with Random Cluster Size

By Shuling Liu

The first part of this dissertation focuses on the development of copula based joint modeling approaches for the clustered survival data with a random cluster size. We propose to adopt Clayton-Oakes model (Clayton, 1978; Oakes, 1989) for measurements within a cluster and the cluster size is modeled via a discrete survival model. The methods are motivated by the Mount Sinai Study of Women Office Workers (MSSWOW) where women were prospectively followed for one year for studying fertility. For each woman, menstrual cycle lengths (MCLs) are recorded until time-to-pregnancy (TTP) or the end of study.

We first consider specifying a parametric distribution as the marginal survival distribution in the Clayton-Oakes model and TTP is modeled using a grouped version of the usual continuous time Cox regression model (Scheike and Jensen, 1997). Second, we consider a semiparametric linear transformation model (Cheng et al., 1995) for the marginal distribution of the Clayton-Oakes model. We develop an EM algorithm to derive an approximate generalized maximum likelihood estimator. We also provide a computationally simple estimation procedure known as the two-stage approach. Asymptotic theory for the two-stage estimators is established. Simulation studies are conducted to evaluate the performance of the proposed joint model and estimation procedures. The proposed methods are also applied to the MSSWOW data.

In the second part of this dissertation, we consider the problem of testing whether a repeatedly measured quantitative biomarker is associated with a subsequent time-to-event process. We propose a nonparametric testing procedure to evaluate the null hypothesis by adopting a linear mixed model for repeated measures, but without imposing modeling assumptions on the time to event. The proposed test can utilize all the information provided by the random effects and is not sensitive to the model misspecification of the time-to-event process. We show that the proposed test statistic is asymptotically consistent and normally distributed under both null and alternative hypotheses. We demonstrate the validity of the new nonparametric test using simulation studies and compare the proposed method to a model-based score test. We finally apply the proposed method to a real data from epidemiological study to illustrate its practical utility.

Joint Modeling Approaches for Clustered Survival Data with Random Cluster Size

By

Shuling Liu
B.E., Renmin University of China, 2007
M.A., State University of New York at Buffalo, 2009

Advisor: Amita K. Manatunga, PhD

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2015

Chapter 1

Introduction

1.1 Background

In the past few decades, reproductive health issues such as conception and infertility, complications of pregnancy and adverse pregnancy outcomes have raised concerns in public health research. Infertility has become more and more common in the United States, about 10% of women (6.1 million) aged 15-44 years suffering difficulties getting pregnant or staying pregnant (Centers for Disease Control and Prevention, 2012). Therefore, how to improve pregnancy health has become very important. In most cases, infertility is due to problems with a woman's ovulation which occurs in each menstrual cycle. Time-to-pregnancy (TTP) is a widely used outcome to study fertility since data on TTP are usually easy and inexpensive to derive by surveys and questionnaires (Bonde et al., 2006). Menstrual cycle length is also a key indicator of women's reproductive health (Harlow and Matanoski, 1991; Small et al., 2006; Guo et al., 2006). Many subject-specific characteristics and environmental factors can affect women's ability to conceive (Scheike and Jensen, 1997; Scheike and Keiding, 2006), which include age, smoking, alcohol assumption, being overweight or underweight and sexual behavior etc. Hence, it is of scientific interest to understand the effects of risk factors on both TTP and menstrual cycle lengths. Recognition of these risks can provide appropriate advice to women who try to get pregnant and protect their

reproductive health. For example, some work-related risk factors have been found to have adverse effects on women's fertility and legislation with regards to the protection of the reproductive health of women have been developed (Liu et al., 2004; Burdorf et al., 2011). Although many statistical tools have been developed for reproductive health and fertility studies in recent years (Harlow and Zeger, 1991; Scheike and Jensen, 1997; Harlow et al., 2000; Guo et al., 2006; Scheike and Keiding, 2006), statistical methods are limited in studying both TTP and menstrual cycle lengths simultaneously. Particularly, challenging and complex features of TTP and menstrual cycle lengths data indicate that efficient and flexible statistical methods are needed.

The first objective of this dissertation is to develop statistical models and their inference for studying risk factors for both menstrual length data and pregnancy outcomes. Despite the fact that the methods are motivated by reproductive studies, they can be generalized to apply for studying joint modeling methodology for longitudinal and survival outcomes in general settings. Another objective is to develop a nonparametric test statistic for the association between a repeatedly measured continuous outcome and a time-to-event process. Our goal is to develop a testing procedure that is more robust to the semiparametric assumption of the survival model.

In this chapter, we provide an introduction and the background. First, we provide scientific reasoning to motivate the statistical problems using a reproductive health study called the Mount Sinai Study of Women Office Workers (MSSWOW). We will describe the details of MSSWOW data including data collection methods, variables and scientific hypotheses that are of interest to epidemiologists. Second, we will summarize current statistical methods for studying TTP and menstrual cycle lengths in the literature. In order to motivate our proposed joint modeling procedure, we provide a background of joint modeling methodology for longitudinal processes and time-to-event data. Since our model is developed based on the Clayton-Oakes models (Clayton, 1978; Oakes, 1989), we give a brief introduction to copula models and the

Clayton-Oakes models. Third, we will briefly summarize the prior literature that were developed under joint modeling framework. We also provide literature reviews on discussion of validity of these methods under model misspecification, which motivates us to propose a robust nonparametric testing procedure without imposing semiparametric model assumptions. Finally, we provide a summary of the methods that are developed in this dissertation.

1.2 The Mount Sinai Study of Women Office Workers

The Mount Sinai Study of Women Office Workers (MSSWOW) was a prospective cohort study that was conducted from 1991 to 1994. The Principle Investigator of this study is Dr. Michele Marcus from the Department of Epidemiology, Rollins School of Public Health, Emory University. Women from 14 companies and government agencies in New York, New Jersey and Massachusetts were enrolled in the study. The aim of the original study was to explore the relationships between Video Display Terminal (VDT) use and rates of spontaneous abortion. Women between the age of 18 and 40 who were at risk of pregnancy were recruited in the study. Three exclusion criteria were implemented prior to the follow-up study, including if the couple had been trying to conceive a child unsuccessfully for more than 1 year, if the woman had a hysterectomy, and if her partner had a vasectomy. A total of 4640 female office workers completed a cross-sectional questionnaire. 563 women who had finished the questionnaire agreed to participate in the study were interviewed to assess possible confounding factors, baseline information and demographics. 79 women who did not collect any urine samples and 14 women who were found to be ineligible after the entry interview were excluded from the study. Finally, 470 women were recruited. After the baseline visit, these women were followed for one year or until the end of

a clinical pregnancy. During the study, the participants were required to complete diaries to record information such as hours of VDT use, exercise performed, stress level, frequency of sexual intercourse, birth control use, and when menstrual bleeding occurred.

As far as now, there is no firm evidence indicating that VDT use would notably increase the risk of infertility of women (Marcus, 1990). Although this study has been designed to investigate the of effect of VDT on spontaneous abortion, the study provided the opportunity to explore the possible roles of risk factors on many reproductive health outcomes such as menstrual cycle length, TTP, spontaneous abortion etc. There are several publications that focused on analyzing reproductive health outcomes resulting from this study (Guo et al., 2006; Small et al., 2006; Small et al., 2007). Menstrual cycle characteristics, including cycle length and bleeding length, are found to be associated with a woman's fertility (Small et al., 2006). Women with 30- to 31-day menstrual cycles and 5-day bleeding lengths were found to have high risk of conception and low risk of spontaneous abortion. Low probability of getting pregnant is also associated with high variability in menstrual cycle lengths (Small et al., 2006). In addition, Guo et al. (2006) found that women's age has a quadratic effect on the variability of menstrual cycle lengths. Chen et al. (unpublished, 2013)'s findings suggested that women's age has significant influence on TTP adjusting for unprotected intercourse.

1.3 Discrete Survival Models for TTP

Since each menstrual cycle indicates a single ovulatory opportunity for getting pregnant, time-to-pregnancy (TTP) is defined as the number menstrual cycles taken to conceive and including a conception cycle and TTP is naturally considered as a discrete random variable (Rothman and Greenland, 1998). Due to this discrete time feature, TTP can be modeled by a grouped version of the continuous survival time

model (Scheike and Jensen, 1997; Kalbfleisch and Prentice, 2002). In this section, we give a brief review of discrete time survival models for TTP.

Let $\tilde{T}$ be the number of cycles to get pregnant for an individual. We assume that $\tilde{T}$ is a non-negative discrete random variable that follows an unspecified distribution and common for all subjects, i.e., $\tilde{T} \sim F_T$. The survival function $S_T(t)$ of $\tilde{T}$ is defined as

$$S_T(t) = \Pr\{\tilde{T} > t\} = \sum_{s>t} \Pr\{\tilde{T} = s\} \tag{1.1}$$

where t takes on possible values of $\tilde{T}$. Given that a woman has not been pregnant in previous cycles, the conditional probability of she getting pregnant at the t-th cycle (i.e., the discrete hazard function) is defined as

$$\lambda_T(t) = \Pr\{\tilde{T} = t | \tilde{T} \geq t\} = \frac{\Pr\{\tilde{T} = t\}}{\sum_{s \geq t} \Pr\{\tilde{T} = s\}}. \tag{1.2}$$

Based on the definition, the following relationship between $\lambda_T(t)$ and $S_T(t)$ can be derived

$$S_T(t) = \prod_{j=1}^{t} (1 - \lambda_T(j)) \tag{1.3}$$

Like all survival data, truncation and censoring issues are very common in TTP studies. To accommodate these problems as well as discreteness in the context of TTP study, several authors proposed grouped version or discrete version survival models that resembles those in the continuous case for TTP data. Scheike and Jensen (1997) proposed a complementary log-log link model for TTP data as following

$$\log(-\log(1 - \lambda_T(t|\mathbf{X}))) = \alpha_t + \boldsymbol{\eta}\mathbf{X} \tag{1.4}$$

where $\mathbf{X}$ is a p-dimensional covariates vector and $\alpha_t = \log(-\log(1 - \lambda_{T_0}(t)))$ is the complementary log-log transformation of the baseline hazard function denoted by $\lambda_{T_0}(t)$. Scheike and Jensen (1997) included a random effect in model (4) to account for potential unknown risk factors and heterogeneity among different women as well

as the within-subject correlation when multiple TTP data are observed for the same subject. A nice feature of this complementary log-log model is that it can be obtained by grouping time in the regular Cox (1972) proportional hazard model in continuous survival time case (Kalbfleisch and Prentice, 2002). To see this, let us assume that Z is a continuous survival time which is modeled by the Cox proportional hazard model

$$\lambda_Z(z|\mathbf{X}) = \lambda_{Z_0}(z)\exp(\boldsymbol{\eta}\mathbf{X}).$$

where $\lambda_{Z_0}(z)$ is the baseline hazard function for time z. Suppose that time Z is grouped into intervals $[a_0, a_1), ..., [a_k, a_{k+1}), ..., [a_m, \infty)$ by partitioning the continuous time space $[0, \infty)$. Assume that the event or censoring is only observed in each of the intervals. If Z falls within the t-th interval $[a_{t-1}, a_t)$, then a grouped observation of Z is $Z_d = t$. Then it can be proved that the grouped survival time Z_d follows the following hazard model

$$\lambda_{Z_d}(t|\mathbf{X}) = 1 - \exp(-\exp(\sigma_t + \boldsymbol{\eta}\mathbf{X}))$$

where $\sigma_t = \log(\int_{a_{t-1}}^{a_t} \lambda_{Z_0}(z)dz)$. Therefore, the complementary log-log model is appropriate for discrete survival data when the underlying event time follows a Cox proportional hazards model (Kalbfleish and Prentice, 2002). The interpretation of the fixed effect of risk factors is straightforward as the logarithm of subject-specific risk ratios. In addition, time-dependent covariates can be easily introduced to the model (Weinberg and Gladen, 1986; Scheike and Jensen, 1997). Based on the model, Ecochard and Clayton (2000) extended the method to more general cases such as multivariate waiting times by including more flexible random effect distributions.

Other methods for TTP include the logistic regression model for grouped failure times suggested by Thompson (1977), which is the explicit use of Cox's (1972) method. Weinberg and Gladen (1986) proposed a generalization of beta-geometric model(Henry, 1953) for TTP data. Under this beta-geometric assumption, conditional on fecundability (which is defined as the probability to conceive during a

month), TTP has a geometric distribution while fecundability itself follows a beta distribution. Thus, the heterogeneity of different subjects is captured by the beta distribution. Although this model can be fitted using general software and be able to handle fixed covariates, it is not easy to incorporate time-dependent covariates. Later Weinberg et al. (1994) developed a semiparametric regression model with a log link for the hazard rates instead of the complementary log-log link function. However, one problem for this method is that the hazard (which is a conditional probability) may be out of boundary of $[0, 1]$ since the covariate coefficients can go from negative infinity to positive infinity. Alternatively, Dunson and Zhou (2000) proposed a Bayesian inference on the fecundability and sterility. Prior information about heterogeneity in fecundability and a mixture model were used to assess the effects of risk factors such as smoking and age while taking into account both sterile couples and heterogeneity among fecund couples. Other Bayesian methods for characterizing covariate effects and heterogeneity among couples include: 1) the more flexible Bayesian semiparametric modeling approach proposed Dunson (2001) based on Dunson and Zhou's model (2000), 2) Bayesian methodology to incorporate known order restrictions to improve efficiency in assessing covariate effects on fecundability (Dunson and Neelon, 2003), 3) Bayesian multiprocess fecundability model developed by Dunson (2003), and 4) hierarchical Bayesian logistic-survival model studied by Hanson et al. (2003) and Thurmond et al. (2005), etc.

1.4 Modeling Menstrual Cycle Lengths

Menstrual cycle length is measured as the number of days from the first day of menstrual bleeding to the first day of next time bleeding. In general, the normal length of a menstrual cycle for a healthy woman is considered to be centered around 28 days but it is possible to have a shorter or longer length than 28 days. Menstrual cycles that are either shorter than 21 days, or longer than 36 days are considered as an irregular

period. Some women may experience irregular cycles occasionally caused by stress, anxiety and illness etc. However, frequent irregular periods may indicate potential menstrual dysfunction and make getting pregnant difficult (Treloar et al., 1967; Bullen et al., 1985; Belsey et al., 1987; Harlow and Matanoski, 1991). With the development of techniques that enable evaluation of menstrual cycle characteristics and assessment of ovulation, the variability and various patterns of menstrual cycles on women's fecundity is of primary interest in recent reproductive studies. Particularly, the effects of exposures such as smoking, alcohol consumption, age and body mass index (BMI) on menstrual function and how these effects vary among different women have been investigated (Treloar et al., 1967; Chiazze et al., 1968; Lenton et al., 1984; Bullen et al., 1985; Harlow and Matanoski, 1991; Harlow and Zeger, 1991; Murphy et al., 1995; Harlow et al., 2000; Guo et al., 2006; Small et al., 2006).

Statistical analysis for menstrual cycle data is often complicated due to several reasons which include: 1) the distribution of menstrual cycle lengths is difficult to describe since it has a mixture of symmetric part with a long right tail; 2) there is generally a sampling bias in follow-up studies like MSSWOW since women with a short cycle would contribute more observations than those with a few long cycles and 3) measurement errors and censoring issues are present in self-reported menstrual cycles.

Normal distribution is a common underlying assumption in early studies for menstrual cycle lengths(Treloar et al., 1967; Chiazze et al., 1968; Lenton et al., 1984). However, it was later found that menstrual cycle length data features a long right tail and therefore a mixture distribution is considered to describe the symmetric part as well as the long right tail (Harlow and Zeger, 1991; Murphy et al., 1995; Harlow et al., 2000; Guo et al., 2006). Suppose the observed menstrual cycle lengths for a subject are denoted by a n-dimensional vector $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)'$ where n is the number of repeated cycle lengths on the same subject. Harlow and Zeger (1991) suggests using

a mixture of symmetric distribution that is centered at 28 days and a stochastically larger component to accommodate the long right tail. Two separate models were fitted for the standard cycle lengths from the symmetric distribution and probability of having a non-standard cycle from the long tail. Lin et al. (1997) extended standard linear mixed models to account for heterogeneous within-cluster variances. To further solve the challenging issues of sampling bias and measurement errors, Murphy et al. (1995) proposed a sequential model approach for the mean cycle lengths conditional on the past cycle lengths and time-varying covariates. This model handles within-woman and between-woman variability of menstrual cycles and identify potential risk factors that affect these variabilities as well as accommodate the sampling bias and censoring issues in the data. Following the argument that the menstrual cycle lengths follow a mixed distribution, Guo et al. (2006) proposed the mixture density for the cycle length Y, $g(y) = pg_1(y) + (1-p)g_2(y)$, where $g_1(y)$ and $g_2(y)$ are the densities for the symmetric distribution and tail distribution, respectively, and p and $1-p$ are the probability weights. Guo et al. (2006) chose a normal distribution for $g_1(\cdot)$ to model the standard cycle lengths and a shifted Weibull distribution for $g_2(\cdot)$ to fit the nonstandard cycle lengths. Modeling of covariates was introduced via two models, one for the normal density and the other for the Weibull density. In addition, the cutoff point that distinguishes the standard and nonstandard cycle lengths were estimated.

Our approach in this dissertation is to consider a general copula model for the joint distribution of menstrual cycle lengths. Specifically, we propose to use the Clayton-Oakes model for the menstrual length data, which is described in details in the next section.

1.5 Copula Models

In recent years, copula models has become a popular statistical tool in modeling dependency between random variables, especially in such fields as finance, actuarial science and survival analysis. A copula is defined as a multivariate distribution function with certain properties that expresses the cumulative distribution function in terms of its one-dimensional marginal distribution functions. For simplicity, we use bivariate distribution for our representation and all definitions and theories can be easily generalized to higher dimensional cases. Specifically, if a two-dimensional vector of random variables $\mathbf{U} = (U_1, U_2)$ follows a copula, denoted by C: $[0,1]^2 \to [0,1]$, then the function C satisfies the following properties:

(i) for $\boldsymbol{\mu} = (\mu_1, \mu_2) \in [0,1]^2$, the realization of $\mathbf{U}$, $C(\boldsymbol{\mu}) = 0$ if at least one $u_i, i = 1, 2$ equals to 0;

(ii) for every $(\mu_{11}, \mu_{12}), (\mu_{21}, \mu_{22}) \in [0,1]^2$ and $\mu_{11} < \mu_{21}, \mu_{12} < \mu_{22}$, $C(\mu_{21}, \mu_{22}) - C(\mu_{21}, \mu_{12}) - C(\mu_{11}, \mu_{22}) + C(\mu_{11}, \mu_{12}) \geq 0$, and

(iii) $C(1, \mu_2) = \mu_2$ and $C(\mu_1, 1) = \mu_1$.

In the framework of statistical modeling, a copula is a function that can be used to link univariate marginals with their full multivariate distribution. Assume that we have a vector of a multivariate random variable vector $\mathbf{X} = (X_1, ..., X_n)$ with a joint distribution function F_J and marginal distribution functions $\mathbf{U} = (F_1(x_1), ..., F_n(x_n))$. The joint distribution F_J can be written as a copula function as (Sklar, 1959)

$$F_J(X_1, ..., X_n) = C(F_1(x_1), ..., F_n(x_n); \rho)$$

where ρ is a dependence parameter. Many copula families are available for constructing statistical models such as Gaussian copulas and Archimedean copulas. Particularly, we focus on Archimedean copulas, which has been explored by statisticians for analyzing the clustered survival data (Genest and Mackay, 1986). Archimedean

copulas are defined as

$$C_A(u_1, ..., u_n) = \phi^{-1}\{\phi(u_1) + \cdots + \phi(u_n)\}$$

where $\phi : [0, 1] \to [0, +\infty]$ is called the generator of the copula. ϕ is a decreasing and convex function such that $\phi(0) = \infty, \phi(1) = 0$. Examples of Archimedean copulas include Clayton copula, Gumbel copula, Frank copula, Gumbel-Hougaard copula etc. In this dissertation, we consider the Clayton copula (Clayton, 1978; Cook and Johnson, 1981; Oakes, 1982). Following Clayton (1978), for the bivariate case, Oakes (1982) assumed that the generator ϕ in the above definition as the inverse of the Laplace transformation (the Laplace transformation of a function $f(x), x \geq 0$ is defined as $\int_0^\infty e^{-sx} f(x)dx$). Then the Clayton copula is obtained by taking the Laplace transformation of gamma distribution with mean 1 and variance θ, as

$$C_\theta(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta})^{-\frac{1}{\theta}}.$$

Under the framework of survival analysis, let (Y_1, Y_2) be two random variables with marginal survival functions $S_1(y_1)$ and $S_2(y_2)$, respectively. The joint survival function of (Y_1, Y_2), $S(y_1, y_2)$, can be expressed via the Clayton copula known as the Clayton-Oakes model which has a form of

$$S(y_1, y_2) = \left[\{S_1(y_1)\}^{-\theta} + \{S_2(y_2)\}^{-\theta} - 1\right]^{-\frac{1}{\theta}}.$$

where $\theta > 0$ is the dependence parameter. As θ approaches to 0, it indicates that Y_1 and Y_2 are independent and thus the joint survival function $S(y_1, y_2)$ is simply the product of the marginal survival functions of $S_1(y_1)$ and $S_2(y_2)$. As θ goes to ∞, the joint survival function converges to its upper Frechet bound and $S(y_1, y_2) = \min\{S_1(y_1), S_2(y_2)\}$. One of the important properties of Archimedean copula is that there exists a general relationship between its generator and Kendall's τ (Kendall, 1938). For the Clayton-Oakes model, the relationship of Kendall's τ and the dependence parameter θ is given by $\theta = 2\tau/(1 - \tau)$. Kendall's τ is a rank-based

measure of correlation, ranging from -1 to $+1$ and hence it is invariant under strictly increasing transformations of the underlying random variables.

Other copulas such as the Gaussian copula are often utilized to construct multivariate dispersion models as well. Some examples of marginal dispersion distributions for Gaussian copulas are given by Song et al. (2000, 2009) such as binary model, Poisson model, Gamma model etc. The parameter estimation based on maximum likelihood procedures for these models have been developed in the literature as well (Oakes, 1982, 1989; Joe, 1994, 1997, 2005; Song, 2005).

The advantages of copula models include their relatively simple and concise mathematical formulation and the introduction of dependency structures without placing restrictions on the marginal distributions. That is, copulas can be utilized in a wide range of parametric, nonparametric or semiparametric frameworks. First, parametric models can be postulated for both copulas and marginal distributions as illustrated in the Clayton-Oakes model. Second, one can consider fitting nonparametric models for both copulas and marginal distributions. Dehuvels (1979) used a multivariate empirical distribution approach. Gijbels and Mielniczuk (1990) proposed a kernel estimator for a bivariate copula. Chen and Huang (2007) later developed a kernel approach based on local linear kernels. A third possibility to work with copulas is the semiparametric approach, which means either a combination of a parametric model and a nonparametric model or semiparametric assumptions for the copula and marginal distributions.

In terms of parameter estimation, Shih and Louis (1995) developed a two-stage semiparametric estimation procedure in the copula models for bivariate survival data where the marginal survival function can be obtained as Kaplan-Mier estimators. Song et al. (2009) extended the multivariate dispersion models generated by Gaussian copulas to more generalized linear models which allows for continuous, discrete and mixed correlated outcomes. For the Clayton-Oakes model, semiparametric marginal

survival distributions were initially incorporated into the model by Oakes(1986). By utilizing the established relationship between the Clayton-Oakes model and the gamma frailty model (Vaupel et al., 1979), Glidden and Self (1999) proposed a semi-parametric likelihood estimation approach for the Clayton-Oakes model which fits into the framework of Nielsen et al. (1992) estimating method for gamma frailty model. They assumed that the marginal hazard function follows a Cox proportional hazard model in the Clayton-Oakes model and an approximate EM-algorithm was developed to obtain the generalized maximum likelihood estimators. Glidden (2000) later proposed a two-stage estimation procedure for the same model.

1.6 Joint Modeling of Longitudinal and Survival Data

In many epidemiological studies and clinical trials, longitudinal (or clustered) measurements of a response and time-to-event data are collected. The longitudinal data, such as CD4 counts, tumor cells, or a health biomarker, are often important predictors or surrogates for a time to event which can be disease-free survival or overall survival. Classical methods for analyzing only the longitudinal data include the linear mixed-effects model or generalized estimating equations (GEE) approach and the Cox proportional hazards model or accelerated failure time model are used when only time-to-event data is available. A more powerful method that considers the relationship and association between longitudinal and time-to-event data is to jointly model the two outcomes (Tsiatis and Davidian, 2004; Hsieh et al., 2006; Ibrahim et al., 2010). That is, the longitudinal and time-to-event data are modeled simultaneously so that one can assess the covariate effects on both outcomes and the association between them. In recent years, this type of joint modeling method has gained popularity because it reduces the bias and improves efficiency in terms of parameter estimation

and can provide valid inference for the effects of risk factors (Wulfsohn and Tsiatis, 1997; Tsiatis and Davidian, 2001; Song et al., 2002b; Tsiatis and Davidian, 2004; Zeng and Cai, 2005; Chen et al., 2011). In addition, the joint models can handle the complicated features of the observed data such as missing and measurement errors in the longitudinal outcome and censoring and truncation in time-to-event data (Lin et al., 2002; Tsiatis and Davidian, 2004; Rizopoulos, 2010; Rizopoulos et al., 2011; Sousa, 2011).

Typically, a joint model consists of two components: a submodel for the longitudinal observations denoted by $Y(t)$ ($t \geq 0$ represents all measurement times) and the other submodel for the time-to-event data denoted by T, where $Y(t)$ is the observation of the longitudinal response for an individual at time t. In addition, censoring and truncation are quite common phenomena in time-to-event data. Let L denote the left truncation variable and C be the right censoring variable. Conditioning on $T \geq L$, $N = \min(T, C)$ is the observed time-to-event data and the censoring indicator is $\Delta = I(T \leq C)$. $Y(t)$ are observed periodically at times $t \leq T$. Hence the observed outcome data for a subject can be denoted by the vector $O = \{N, \Delta, y(t); t \geq 0\}$. By making assumptions on the joint distribution of O and specification for the two submodels, different joint models can be constructed.

In literature, there are usually two ways to construct a joint model for longitudinal and the survival data, including (i) assuming shared random effects for both outcomes (Self and Pawitan, 1992; DeGruttola and Tu, 1994; Hogan and Laird, 1997; Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Wang and Taylor, 2001; Xu and Zeger, 2001; Ibrahim et al., 2004; Zeng and Cai, 2005; Chi and Ibrahim, 2006; Vonesh et al., 2006; Diggle et al., 2008; Rizopoulos and Ghosh, 2011) and (ii) utilizing different factorization of the joint distribution of $(Y(t), T)$ (Little, 1993; Tsiatis et al., 1995; Hu et al., 1998; Huang et al., 2001; Xu and Zeger, 2001a, 2001b).

1.6.1 Shared Random Effects Joint Models

The shared random effects joint models assume that the longitudinal and time-to-event processes are independent conditional on some latent random effects. That is, event time and longitudinal biomarker are assumed to be associated via an underlying progression, defined by random effects, rather than directly dependent on each other. Typically, the model for the repeatedly measured longitudinal data with random effects has a form of

$$Y(t) = \alpha_0 + \alpha_1 t + \boldsymbol{\beta}\mathbf{X}(t) + \varepsilon(t)$$

where $\alpha = \{\alpha_0, \alpha_1\}$ is the vector of random effects that follow a bivariate normal distribution $MVN_2(\mathbf{0}, \boldsymbol{\Sigma}_a)$ and represents the subject-specific random intercept and random slope, respectively; $\mathbf{X}(t)$ is a vector of covariates and $\boldsymbol{\beta}$ is the corresponding covariate coefficients; and $\varepsilon(t)$ are mutually independent normal random errors. For the survival submodel, a parametric distribution such as exponential or Weibull for the response can be assumed. More commonly, the semiparametric Cox proportional hazards models and accelerated failure time models have been widely used (DeGruttola and Tu, 1994; Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Tsiatis and Davidian, 2004; Vonesh et al., 2006; Rizopoulos et al., 2008). The two submodels are typically joined via shared random effects. Given the random random effects, the longitudinal response and the survival time are conditionally independent. For instance, Wulfson and Tsiatis (1997) proposed the Cox proportional hazards model with shared random effects given in the mixed model as

$$\lambda_T(t|\alpha, Y(t)) = \lambda_T(t|\alpha) = \lambda_0(t)\exp(\gamma(\alpha_0 + \alpha_1 t + \boldsymbol{\beta}\mathbf{X}(t)))$$

where γ evaluates the dependency between the longitudinal biomarker and the time to event. This type of joint models have been studied in literature including Self and Pawitan (1992), DeGruttola and Tu (1994), Tsiatis et al. (1995), Faucett and Thomas (1996), Bycott and Taylor (1998), Dafni and Tsiatis (1998) and recently by

Rizopoulos et al. (2009). More complex random effects structure can be specified. For example, a polynomial function of the random effects can be used instead of a linear form. Tsiatis and Davidian (2001) and Song et al. (2002b) generalized this model by relaxing the normality assumption on the random effects. Song et al. (2002a) incorporated multiple time-dependent covariates into the model. Zeng and Cai (2005) proposed a joint model where random effects affect the two outcomes differently. Instead of imposing common random effects for both outcomes, they assumed the random effects in the longitudinal model are (α_0, α_1) and the random effects in the survival model are assumed to have a form of $\tilde{\alpha}_i = \gamma_i \alpha_i + \omega_i, i = 0, 1$ where ω_i is a subject-specific effects that only affect the survival response and γ is a scale parameter that evaluates different intensity of random effects on both outcomes.

To allow a more flexible and feasible structure for the within-subject correlation, some authors suggested adding common stochastic process components instead of random effects into both the submodels and the association between both responses can be captured by the stochastic process (Henderson et al., 2000; Wang and Taylor, 2001; Xu and Zeger, 2001a). Likewise, a similar approach has been proposed by Lin et al. (2002a) where the joint model contained common covariates that have influence on both outcomes. Other issues such as missingness and measurement errors have also been closely investigated (Wulfsohn and Tsiatis, 1997; Wu et al., 2008). In the presence of clustered data, Ratcliffe et al. (2004) proposed a mixed effects type model for the clustered data and a frailty model for the survival outcome along with a common cluster-level random effect that accommodated both within-subject and between-cluster heterogeneity.

Usually, the parameter estimators for these shared random effects joint models can be obtained by EM algorithm. The performance of the EM algorithm was examined by many authors(Wulfson and Tsiatis, 1997; Zeng and Cai, 2005; Hsieh et al., 2006) and it has been demonstrated that these estimators are robust and efficient

under certain assumptions. Hsieh et al. (2006) recommended to use the bootstrap estimators for estimating the standard errors of the parameter estimators unless reliable standard error estimates can be established theoretically.

1.6.2 Mixture and Selection Joint Models

The second approach to construct a joint model is based on a conditional distribution factorization of the joint distribution of $Y(t)$ and T. Specifically, one can assume that

$$f_{Y(t),T}(y(t),t) = f_{T|Y(t)}(t|y(t))f_{Y(t)}(y(t)) \text{ or } f_{Y(t),T}(y(t),t) = f_{Y(t)|T}(y(t)|t)f_T(t)$$

These two types of models are also known as selection and mixture models, respectively (Little, 1993). A linear mixed-effects model is commonly assumed for the distribution of repeated measurements of the longitudinal outcome. Typical choices for the conditional time-to-event distribution include Cox proportional hazard model, accelerated life model, logistic linear regression, probit regression, etc.

The first type of conditional and marginal distribution factorization, i.e., the selection model, is proposed to study the informative drop-out where the conditional distribution of drop-out time T given the longitudinal response is modeled via generalized linear models, with observed longitudinal data as a covariate and the longitudinal response is modeled by linear mixed models (Diggle and Kenward, 1994; Diggle, 1998; Scharfstein et al., 2003). Hogan and Laird (1997) proposed a mixture model for a longitudinal outcome conditioning on a time-to-event which missing values in the longitudinal measures and censoring survival time are accommodated. For longitudinal outcome, the conditional distribution of $Y(t)$ given T is assumed to follow a multivariate normal distribution and a linear mixed effects model is utilized and the event time T is a covariate for the conditional mean of $Y(t)$. No parametric form is assumed for the time-to-event data and Kaplan-Meier product-limit estimators are obtained to replace the cumulative distribution function $F(t)$. Hogan et al. (2004) extended this

mixture model to more general semiparametric framework. They assumed that the longitudinal response follows a varying coefficient random effects model conditional on drop-out time while impact of drop-out time on the longitudinal data is modeled via unspecified nonparametric functions that can be estimated using step functions when drop-out time is discrete and using smoothing splines if drop-out time is continuous.

A joint model for longitudinal measurements and competing risks survival data was studied by Elashoff et al. (2008) where a flexible modeling approach is proposed to handle potential informative missingness in the longitudinal measurements due to dropout and a possible way to incorporate informatively censored events as a competing risk. Linear mixed-effects models for the longitudinal process and proportional cause-specific hazard frailty models for the competing risks survival data are linked via some latent random effects to construct the joint model. The maximum likelihood estimators for mixture models are generally available by using EM algorithms based on the joint distribution $f_{Y(t),T}(y(t), t)$ (Hogan and Laird, 1997; Elashoff et al., 2008).

1.6.3 Other Joint Models

In addition to shared parameter models and mixture and selection models, many authors based their work with joint models within the framework of Bayesian analysis where hierarchical models were generally used (Brown and Ibrahim, 2003ab; Ibrahim et al., 2004; Chi and Ibrahim, 2007; Rizopoulos and Ghosh, 2011) and both longitudinal and survival outcomes are allowed to be multidimensional.

In a different context but related to joint modeling framework, some authors have investigated clustered data that involve clustered outcomes with a random cluster size. The measurements of outcomes within a cluster is often associated with the cluster size and the primary interest is to investigate the dependency of the outcomes per subunit and cluster size. In certain situations, when multiple measurements on the subjects are observed, the number of the measurements is a random variable.

Multivariate random length data are observed when multiple measurements (categorical or continuous) are obtained and at the same time the length of the vector of measurements is also recorded as an outcome for each individual (Barnhart and Sampson, 1995; Barnhart et al., 1999). Barnhart and Sampson (1995) developed a general joint modeling setting for multivariate random length data where the random length is assumed to follow a generalized linear model and the conditional distribution of the observed multiple measurements vector has a form of multivariate normal distribution. Specifically, the random length variable K has a discrete distribution as

$$\Pr\{K = k\} = g_k(\delta + \gamma\mu_i)$$

where δ is the intercept, μ_i represents the i-th population mean and γ is the regression coefficient that evaluates the correlation between the multivariate data and the random length. Conditional on $K = k$, the multivariate vector, denoted by $\mathbf{X}$, has a distribution of

$$\mathbf{X}|K = k \sim MVN\left[\mu_i \mathbf{e}_k, \sigma^2 R_k(\rho)\right]$$

where $\mathbf{e}_k$ is an $k \times 1$ vector with every element equal to one and $R_k(\rho)$ is the $k \times k$ correlation matrix. Maximum likelihood estimators and asymptotic properties are derived by likelihood inferences. Later, Barnhart et al., (1999) extended this population model by including covariates in both generalized linear model and the multivariate normal model. Maximum likelihood estimators as well as asymptotic efficiency are provided. Some technical issues related to the estimation procedures are discussed. Dunson et al. (2003) considered a similar problem to address the association between the outcomes on subunits in a litter and the litter size. They constructed the joint model under a Bayesian framework by assuming an underlying normal model for the subunit-level outcomes and a generalized continuation ratio probit model for the cluster size. The dependency structure between the subunit-level outcomes and cluster size was described by a multivariate normal covariance structure and shared latent

variables. In this dissertation, we consider joint modeling approach in the survival analysis settings where outcomes per subject are time-to-event data with a random cluster size. The basic concept behind the development of our longitudinal and survival models is similar to that suggested by Barnhart et al. (1995, 1999).

1.6.4 Testing Whether Repeated Measured Biomarker Associated with Time To Event

As described in the previous sections, a typical setup of the joint model is that the longitudinal model and the survival model are linked via the shared random effects. Based on the joint models, test statistic such as score test and Wald test are available to investigate the relationship between the two outcomes. For example, Jacqmin-Gadda et al. (2010) proposed a score test based on a joint model with latent classes and shared random effects for testing the null hypothesis that the risk of an event depends on the random effects from the longitudinal marker model in addition to the latent classes.

However, the model-based testing procedures often raise questions regarding the robustness against the model misspecifications. Either or both assumptions for the longitudinal process and time to event may fail. In this dissertation, another objective is to develop a nonparametric test statistic to determine whether a biomarker and a time to event are associated without imposing parametric or semiparametric model assumptions on the time to event.

1.7 Outline

The first two chapters of this dissertation is focused on developing a joint model approach to describe the relation of multiple menstrual lengths and time-to-pregnancy (TTP). We assume the Clayton-Oakes model for the clustered menstrual cycle lengths

and a complementary log-log link for the time-to-pregnancy. In the next chapter, we develop a joint model where a parametric model is postulated in the Clayton-Oakes model for menstrual cycle lengths and a complementary log-log link model is used for TTP and a full likelihood specification is constructed to derive maximum likelihood estimators. The proposed joint modeling method provides meaningful insights on the reproductive health of women and demonstrates flexibility to accommodate truncation and censoring issues in the data. Standard errors are obtained via both likelihood function and bootstrap procedure and the analytical form of the estimating equations are also provided. Simulation studies are conducted to examine the estimators and the performance of the proposed method.

Chapter 3 focuses on a more flexible modeling framework for both outcomes, menstrual lengths and TTP. That is, we generalize the joint model approach proposed in Chapter 1 to semiparametric models. Specifically, the marginal distributions in the Clayton-Oakes model are left semiparametrically specified. An approximate EM algorithm is developed to obtain the maximum likelihood estimators. Monte Carlo simulations illustrate that the estimators from the joint model perform well for finite samples. The proposed joint model and likelihood estimation approach are applied to the MSSWOW data.

Although the likelihood method, i.e., EM algorithm, is often used and has been demonstrated to perform well, it is also shown that this estimation procedure is computationally demanding and uneasy to implement using existing software packages in practice. In Chapter 4, we propose a two-stage method to obtain the parameter estimates, which is computationally simpler. Simulation studies are conducted to evaluate the performance of the two-stage method as well as to compare the relative efficiency of EM algorithm and two-stage approach. The results demonstrate that the two-stage estimation method provide unbiased estimators and the loss of efficiency is relatively small and offset by its simplicity. Under certain regularity con-

ditions, we show that the estimators based on two-stage method are consistent and asymptotically normal. Finally, the proposed approach is applied to the MSSWOW data.

In Chapter 5, we develop a nonparametric test statistic to determine the relationship between a repeatedly measured quantitative biomarker and a subsequent time-to-event process. A linear mixed model for repeated measures is used, but no modeling assumptions such as proportional hazards on the time to event are imposed. The proposed test statistic is shown to be asymptotically consistent and normally distributed under both null and alternative hypotheses. The simulation studies show that the nonparametric test statistic can perform well. We also apply the proposed method to a real data from epidemiological study.

Chapter 6 provides a summary of the dissertation and plans for future research work.

Chapter 2

Joint Models with Marginal Parametric Assumptions

2.1 Introduction

In many medical studies that involve clustered or longitudinal data, outcomes measured within a cluster (subject) along with cluster size (the number of the measurements) are collected. For example, if outcomes per subject (cluster) generate multiple measurements and the number of measurements, then a random cluster size is observed. In this setting, the covariates (e.g., treatment or exposure to some risk) may have influence on both outcomes from the subunits as well as the cluster size. This type of data has been referred to as multivariate random length data by Barnhart and Sampson (1995). These authors presented the National Heart, Lung and Blood Institute (NHLBI) Type II Coronary Intervention study (Brensike et al., 1982, 1984) where patients with Type II hyperlipoproteinemia and coronary heart disease were randomly assigned to a treatment or a placebo group. The outcome measurements of a patient's angiograms consisted of vascular lesion sizes and the number of lesions. The treatment may affect both the lesion sizes and the number of lesions through some underlying mechanism. Another example of such data is described by Dunson et al. (2003) in a rodent teratology study where fetal outcomes (e.g., fetal weight) were measured for each subunit in a litter and the litter size may be associated with

the fetal outcomes. If the dam is exposed to some developmental toxicant prior to mating, the correlation between litter size and fetal weight will be affected by the mechanism of action of the toxicant and random features of the dam and pregnancy. In both of these examples, the outcomes measured on the subunits are continuous. However, in some studies, the outcomes may take the form of survival times. For example, one can consider a study where the outcome, time-to-death, is obtained from subunits in a litter as in Dunson et al.'s manuscript. Challenging features of survival data such as censoring and truncation can pose difficulties in directly applying the models that have been developed for continuous data. In this dissertation, we consider a reproductive study called the Mount Sinai Study of Women Office Workers (MSSWOW) to motivate our research in a joint modeling framework for clustered survival data with a random cluster size.

In the presence of random cluster size, it is important to jointly model the outcomes and the cluster size to accommodate association between them rather than simply assume independence or incorporate cluster size as a covariate. Barnhart and Sampson (1995) proposed a general joint model for multivariate random length data to depict the relationship between the quantitative variable observed from the subjects and the random length of the vector. The distribution of observations within each subject was assumed to follow a multivariate normal distribution with an exchangeable correlations structure while the length of the multivariate data was modeled via a generalized linear model. Both distributions depend on underlying parameters some of which are common. Maximum likelihood estimation procedure was developed for statistical inference. Later, Barnhart et al. (1999) extended this model to include additional covariates in the model. Dunson et al. (2003) proposed a Bayesian framework to jointly model the multivariate outcomes measured on each subunit and the random cluster size. Their method utilized an underlying normal model for the subunit-level outcomes and a generalized continuation ratio probit model for the cluster size. The

relationship between the subunit-level outcomes and cluster size was captured by a multivariate normal covariance structure and shared latent variables. For correlated survival data, we postulate a similar structure as Barnhart and Simpson (1995) model by modeling the survival data in terms of a Copula model. As previously mentioned, our analysis is motivated by MSSWOW data which we describe below.

In the MSSWOW study, women were followed prospectively for one year for studying fertility during which repeated measures of menstrual cycle length for each subject was collected. Time-to-pregnancy (TTP) was recorded as the number of menstrual cycles taken to conceive including the conception cycle. In the literature, statistical models have been developed to investigate the potential covariate effects on menstrual cycle lengths (e.g., Harlow and Zeger, 1991; Guo et al, 2006) without paying much attention to the fact that subjects may get pregnant during the study period. Similarly, the models have been proposed to investigate covariate effects on TTP (e.g., Scheike and Jensen, 1997; Keiding et al., 2002) to study different risk factors for women's fertility without considerations of repeated menstrual lengths that occurred prior to pregnancy. In many reproductive studies, measurements of menstrual lengths are recorded until time to pregnancy or the end of the study. Since both menstrual lengths and TTP are good indicators of reproductive health (Baird et al., 1986; Weinberg et al., 1989; Harlow and Zeger, 1991; Florack et al., 1994; Jensen et al., 1999; Dunson et al., 2002; Scheike and Keiding, 2006), it is of interest to evaluate the effects of covariates on both outcomes (menstrual lengths and TTP) as well as the relationship between the two outcomes. We treat the repeated continuous measurements of menstrual lengths and the end time outcome as clustered outcomes where the cluster size is a random variable. Specifically, when a women gets pregnant and hence is no longer at risk of menstrual bleeding, the menstrual cycle length at the conception cycle cannot be observed. If a subject does not conceive at the end of the study, i.e., the subject's TTP is censored, the remaining menstrual cycle lengths

until pregnancy are missing. Likewise, due to some subjects have been trying to get pregnant before entering the study, TTP is left-truncated and menstrual lengths before entry are also missing. To accommodate these complications, we propose a joint model method, under which, repeated menstrual cycle lengths are assumed to follow the Clayton-Oakes model (Clayton, 1978; Oakes, 1989) of a size of TTP, denoted by T, while T is modeled via a discrete time hazard model (Scheike and Jensen, 1997).

The Clayton-Oakes model is a very flexible type of copula models for multivariate failure time data because it allows to specify arbitrary marginal distributions whiles incorporating the intracluster dependence. Previous work (Harlow and Zeger, 1991; Guo et al., 2006), as well as the histogram of our data suggest the need to address skewness of the distribution of menstrual lengths. By recognizing the log-linear interpretation of Weibull distribution, we assume the marginal distribution of the menstrual lengths as Weibull distribution. One approach for modeling discrete failure time is to use a grouped version of the usual continuous time proportional hazard model (Kalbfleisch and Prentice, 2002). This discrete survival model is convenient to specify and it retains an interpretation in terms of the proportional hazards assumptions with the underlying continuous time.

This chapter is organized as follows. First, we present the joint models for the clustered survival data with a random cluster size. Model is parameterized based on the motivation of MSSWOW study. A maximum likelihood-based procedure is developed to estimate the covariate effects on both menstrual cycle lengths and TTP as well as within-subject association of menstrual cycle lengths. In addition, our method appropriately handles missing and censoring menstrual lengths and also well accommodates censoring and left truncation that occurs in the pregnancy outcome. Simulation studies are conducted to evaluate the performance of the proposed method. Finally we apply our method to MSSWOW data.

2.2 The Model

2.2.1 Notation

Suppose that we have m subjects (or clusters). Let T_i represent the underlying random cluster size for the i-th subject. In this dissertation, we focus on the case where T_i is a discrete survival time, but the modeling approach can be extended to other discrete distributions. Conditional on $T_i = t_i$, the corresponding observable clustered outcome for the subject i is a t_i-dimensional vector denoted by $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{it_i})'$. Note that in the MSSWOW data, since the subjects are not at risk for menstrual bleeding if conception occurs, which results in the menstrual length is inherently missing at the conception cycle, the clustered response vector has a dimension of time-to-pregnancy minus 1 due to the inherent missingness of menstrual cycle lengths. Therefore, $T_i = t_i$ denotes the time-to-pregnancy after t_i observed menstrual cycles.

As a time-to-event variable, the cluster size T_i involves truncation and censoring issues. Let L_i denote the left truncation variable and C_i be the right censoring time, which is assumed to be independent of T_i. Conditioning on that $T_i \geq L_i$, the observed data for the i-th subject on the discrete time-to-event process T_i consists of $(\tilde{T}_i, \Delta_i)$, where $\tilde{T}_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$ is the censoring indicator. In addition, a p-dimensional covariate vector $\mathbf{X}_i$ is collected for each subject, which can affect both the clustered outcome and the cluster size. Thus, the observed data comprises the set $\{\mathbf{Y}_i, \tilde{T}_i, \Delta_i, \mathbf{X}_i; i = 1, ..., m\}$. Throughout this chapter, the upper-case letters represent random variables and we use lower-case letters for their realizations.

To construct a joint model, the joint distribution $f_{\mathbf{Y},T}$ of $(\mathbf{Y}_i, T_i)$ given covariates $\mathbf{X}_i$ is factorized by the conditional distribution of $\mathbf{Y}_i|T_i$ and the marginal distribution of T_i as

$$f_{\mathbf{Y},T}(\mathbf{y}_i, t_i|\mathbf{X}_i; \boldsymbol{\pi}) = f_{\mathbf{Y}|T}(\mathbf{y}_i|t_i, \mathbf{X}_i; \boldsymbol{\omega}) f_T(t_i|\mathbf{X}_i; \boldsymbol{\upsilon})$$

where $\boldsymbol{\pi} = (\boldsymbol{\omega}, \boldsymbol{\upsilon})$ represents the parameter vector in the joint distribution and $\boldsymbol{\omega}$ and

$\boldsymbol{\epsilon}$ are the parameter vectors for $f_{\mathbf{Y}|T}$ and f_T, respectively.

2.2.2 General Framework

First, we define the marginal model for the cluster size T_i. We assume that the hazard rate of the discrete time-to-event T_i $\lambda(t_i|\mathbf{X}_i)$ has a discrete hazard model as

$$\lambda(t_i|\mathbf{X}_i;\boldsymbol{\xi}) = Pr\{T_i = t_i|T_i \geq t_i, \mathbf{X}_i\} = \nu(\alpha(t_i) + \boldsymbol{\xi}\mathbf{X}_i) \tag{2.1}$$

where $\nu(\cdot)$ is a known function, $\boldsymbol{\xi}$ is a set of regression coefficients associated with covariates $\mathbf{X}_i$ and the scalar parameter $\alpha(t_i)$ represents the baseline hazard rate associated with t_i. A particular link function of $\nu(\cdot)$ is the complementary log-log (CLL) function which is equivalent to the Cox regression model for continuous failure time. Other link functions include the logistic model and log linear model etc.

The clustered outcome is assumed to follow some copula model. Copula families are often used to construct statistical models. An important family of copula functions is known as the Archimedean copulas. Conditional on $T_i = t_i$, $\mathbf{Y}_i$ is assumed to follow an Archimedean copula model (Nelsen, 1999):

$$\begin{aligned}\mathbf{S}_J(\mathbf{y}_i|T_i = t_i, \mathbf{X}_i;\boldsymbol{\beta}) &= \Pr\{Y_{i1} > y_{i1}, ..., Y_{it_i} > y_{it_i}|T_i = t_i, \mathbf{X}_i;\boldsymbol{\beta}\} \\ &= \mathbf{C}_A\left(S(y_{i1}|\mathbf{X}_i;\boldsymbol{\beta}), ..., S(y_{it_i}|\mathbf{X}_i;\boldsymbol{\beta})\right)\end{aligned} \tag{2.2}$$

where the subscript J indicates that this is a joint survival function and $\mathbf{C}_A$ is the Archimedean copula function that can be written in the form of

$$\mathbf{C}_A(u_1, ..., u_{t_i}) = \varphi(\varphi^{-1}(u_1) + \cdots + \varphi^{-1}(u_{t_i})), u_j \in [0,1], j = 1, ..., t_i$$

for some generator function φ and its generalized inverse φ^{-1} that satisfy:

1. $\varphi(0) = 1$ and $\lim_{x\to\infty}\varphi(x) = 0$.
2. φ is continuous and strictly decreasing on $[0, \varphi^{-1}(0)]$.
3. φ^{-1} is given by $\varphi^{-1}(x) = \inf\{u : \varphi(u) \leq x\}$.

$\boldsymbol{\beta}$ is a vector of regression parameters for covariates $\mathbf{X}_i$ and $S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ is either a

parametric or semi-parametric marginal survival distribution of the j-th cycle for the i-th subject.

In addition to investigating the covariate effects on both outcomes, it is of importance to understand the association between the processes. For example, in the MSSWOW study, the outcome data consist of repeated menstrual cycle lengths with cluster size measured by time-to-pregnancy (TTP), i.e., the number of menstrual cycles taken to get pregnant. It is well known both menstrual lengths and TTP are key indicators of women's reproductive health. Risk factors such as age and smoking were assessed for both responses. At the same time, the interest also lies in understanding the association between the process of occurrence of menstrual lengths and consequent pregnancy outcome. We introduce shared parameters are imposed in model (2.1) and (2.2) to quantitatively describe the relationship between the clustered outcome and the cluster size (Barnhart and Sampson, 1995; Dunson et al., 2003). In particular, we assume that $\boldsymbol{\xi} = \gamma\boldsymbol{\beta}$ where the regression parameters $\boldsymbol{\beta}$ are common in both models which indicate that the covariates may affect both the clustered outcome and its cluster size and γ is a scalar that evaluates the effects of the clustered response on the cluster size as well as acts as a scaling parameter providing different influences of the covariates on the distribution of both responses.

2.2.3 The Model Specification

Following Scheike and Jensen's work (1997), we adopt the complementary log-log link for $\nu(\cdot)$ in model (2.1), so the hazard rate of the cluster size variable $\lambda(t_i)$ has a time-to-event submodel as

$$\lambda(t_i|\mathbf{X}_i;\gamma,\boldsymbol{\beta}) = 1 - \exp[-\exp(\alpha(t_i) + \gamma\mathbf{X}_i\boldsymbol{\beta})]. \tag{2.3}$$

Although logit link is more common in practice, this model is more appropriate for discrete survival data in terms of retaining the continuous proportional hazard

model interpretation (Kalbfleish and Prentice, 2002).

To understand the within-cluster association, we assume that the menstrual cycle lengths denoted by $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{it_i})'$ follow a special type of Archimedean copula known as the Clayton-Oakes model (Clayton 1978, Oakes 1989):

$$S_J(\mathbf{y}_i|T_i = t_i, \mathbf{X}_i; \boldsymbol{\beta}) = Pr\{Y_{i1} > y_{i1}, ..., Y_{it_i} > y_{it_i}|T_i = t_i, \mathbf{X}_i; \boldsymbol{\beta}\} = \left[\sum_{j=1}^{t_i} S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta})^{-\theta} - t_i + 1\right]^{-\frac{1}{\theta}} \tag{2.4}$$

where $S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta})$ is the marginal survival model with covariates $\mathbf{X}_i$ and regression parameters $\boldsymbol{\beta}$ and $\theta > 0$ depicts the association between any two clustered responses. The association between any two clustered measurements is assumed to be the same. When θ approaches to 0, the observations within the same subject are independent and thus the joint survival function is simply the product of the marginal survival functions. As $\theta \to +\infty$, the joint survival function converges to its upper Frechet bound and $S_J(\mathbf{y}_i|T_i = t_i, \mathbf{X}_i) = \min\{S(y_{ij}|\mathbf{X}_i), j = 1, ..., t_i\}$. The dependence parameter θ is related to Kendall's (1962) coefficient of concordance known as the Kendall's τ which can be expressed as $\tau = \theta/(\theta + 2)$.

Furthermore, we assume that the marginal distribution of the clustered survival data follows a Weibull distribution. Compared to other parametric models, it is very flexible with regards to the assumptions on the hazard rate as well as the description of the tail shape. This is particulary useful for the menstrual cycle lengths because it is well known that the distribution of menstrual lengths is skewed with a long right tail. In addition, Weibull distribution is the only parametric survival model which can be represented as both a proportional hazard model and an accelerated failure time model. Specifically, the survival function of the Weibull distribution for Y_{ij} is given by

$$S_Y(y_{ij}) = \exp(-\mu_{ij}^k y_{ij}^k)$$

where k and μ represent the shape parameter and the scale parameter, respectively.

Taking the logarithm of time Y_{ij}, the $V_{ij} = \log Y_{ij}$ has a survival function as

$$S_V(v_{ij}) = \exp(-\mu_{ij}^k \exp(kv_{ij})).$$

To incorporate the covariates $\mathbf{X}_i$, we define the scale parameter as $\mu = \exp(-(\beta_0 + \mathbf{X}_i\boldsymbol{\beta}))$. Then it follows that log-transformed survival time has a linear relationship with the covariates as

$$V_{ij} = \log Y_{ij} = -\frac{1}{k}\log \mu_{ij}^k + \frac{1}{k}e_{ij} = \beta_0 + \mathbf{X}_i\boldsymbol{\beta} + \frac{1}{k}e_{ij}$$

where e_{ij} has the extreme value distribution with probability density function $f(w) = \exp(w - \exp(w)), -\infty < w < \infty$. Similar to linear models, this model has the interpretation that with one unit increase in the covariate $X_{ib}, b = 1, ..., p$, the average logarithm of Y_{ij} will increase or decrease by absolute value of β_{ib}.

Then we rewrite the Clayton-Oakes model with the marginal survival model as above as

$$S_J(\mathbf{y}_i|T_i = t_i, \mathbf{X}_i; \boldsymbol{\beta}) = \left[\sum_{j=1}^{t_i} \exp\left(\theta y_{ij}^k \exp(-k(\beta_0 + \mathbf{X}_i\boldsymbol{\beta}))\right) - t_i + 1\right]^{-\frac{1}{\theta}}. \qquad (2.5)$$

Combining (2.3) and (2.4), the joint model for the clustered measures $\mathbf{Y}_i$ and the random cluster size T_i is constructed as

$$\Pr(Y_{i1} > y_{i1}, ..., Y_{it_i} > y_{it_i}, T_i = t_i|\mathbf{X}_i; \boldsymbol{\pi}) = \left[\sum_{j=1}^{t_i} S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta})^{-\theta} - t_i + 1\right]^{-\frac{1}{\theta}} \Pr\{T_i = t_i|\mathbf{X}_i; \gamma, \boldsymbol{\beta}\}, \qquad (2.6)$$

where $\boldsymbol{\pi} = (\theta, \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})$ is the parameter space for the joint model and $S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}) = \exp\left(-y_{ij}^k \exp(-k(\beta_0 + \mathbf{X}_i\boldsymbol{\beta}))\right)$ is the marginal Weibull model and $\Pr\{T_i = t_i|\mathbf{X}_i; \gamma, \boldsymbol{\beta}\}$ is written in terms of a discrete hazard with complementary log-log link. Based on joint model (2.6), the joint density function $f_{\mathbf{Y},T}(\mathbf{y}_i, t_i|\mathbf{X}_i; \boldsymbol{\pi})$ can be written as:

$$f_{\mathbf{Y},T}(\mathbf{y}_i, t_i|\mathbf{X}_i;\boldsymbol{\pi})$$
$$= \prod_{i=1}^{t_i}((j-1)\theta+1)f(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})^{-\theta} \cdot \left[\sum_{j=1}^{t_i} S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})^{-\theta} - t_i + 1\right]^{-\frac{1}{\theta}-t_i+1} \Pr\{T_i = t_i|\mathbf{X}_i;\gamma,\boldsymbol{\beta}\}$$

where $f(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ and $S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ are the probability density function and survival function of Weibull distribution. Note that the above models has some good properties. First, zero cluster size is allowed for some subjects. That is, no quantitative measurements of clustered outcome are observed for a subject. In our example of MSSWOW data, some women may get pregnant at the first menstrual cycle. In this case, no measurements of menstrual cycle lengths are observed, which results in a cluster size of zero. Second, the random cluster size can take either any non-negative integer value or non-negative integer up to some known value $T_{max} < \infty$. Third, a stochastic ordering property has been imposed in the joint model by specifying the shared parameters, which can describe the association between the clustered measurements and the cluster size response. Specifically, the discrete hazard function $\lambda(z)$ with the complementary log-log link in model (2.3) is stochastically increasing in z. This implies that for a covariate $X_{ib}, b = 1, ..., p$, when the scaling parameter $\gamma > 0$, the larger the value of $\beta_b X_{ib}$ is, the more likely a larger TTP would be observed as well as a longer menstrual cycle length, and vice versa for $\gamma < 0$. In other words, γ depicts the relationship between the clustered response and the cluster size. If $\mathbf{X}_i\boldsymbol{\beta}$ is seen as the estimate of the centered mean of clustered measurements, $\gamma > 0$ implies that with greater measure of clustered outcome, we will observe a greater risk of having an event in the survival cluster size and vice versa. When γ equals to zero, TTP and menstrual cycle lengths are not associated. In addition, the multivariate distribution of the clustered measurements is invariant under permutation of its components because of the summation form of the Clayton-Oakes model. Another important feature of the model is that the submodel of the Clayton-Oakes model maintains the same form as the full model assuming that the marginal distributions

are correctly specified. This is particularly attractive if some repeated measurements are missing. Note that it is possible to have different covariate effects in the models by simply adding covariates to either or both models without imposing the shared regression coefficients.

2.3 Parameter Estimation

2.3.1 Maximum Likelihood Estimators

Suppose the observed data for the i-th subject of $\{\mathbf{Y}_i, T_i, \mathbf{X}_i, L_i, \eta_{ij}\}$ is $\{\mathbf{y}_i, t_i, \mathbf{x}_i, l_i, \eta_{ij}\}$. The likelihood based on the joint model (2.5) has two components including a part consisting discrete survival model (2.3) and a part from the Clayton-Oakes model (2.4). The likelihood contribution from model (2.3) is straightforward. Specifically, conditioning on the left truncation $L_i = l_i$, the likelihood contribution from the observed cluster size data based on model (2.3) is (see details in Appendix I)

$$\prod_{j=l_i+1}^{t_i} \left[\frac{\lambda(j|\mathbf{X}_i)}{1-\lambda(j|\mathbf{X}_i)}\right]^{\eta_{ij}} (1-\lambda(j|\mathbf{X}_i)) = \prod_{j=l_i+1}^{t_i} \left[\frac{1-\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i\boldsymbol{\beta}))}{\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i\boldsymbol{\beta}))}\right]^{\eta_{ij}} \exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i\boldsymbol{\beta})).$$

where $\eta_{ij} = 1$ if the i-th subject experienced an event at the j-th discrete time point and 0 otherwise. However, in order to write the likelihood contribution from model (2.4), we need to consider several different scenarios. First, for an individual who has an event at time t_i during the study period, the observed clustered response is complete, i.e., $\mathbf{Y}_i = (Y_{i1}, ..., Y_{it_i})'$ given that $T_i = t_i$. Second, if a subject is censored during the study and the underlying time-to-event is $T_i > t_i$, we can only observe $\mathbf{Y}_i = (Y_{i1}, ..., Y_{it_i})'$ but the future observations until the actual time-to-event are missing. Additionally, some subjects may have delayed entry to the study. That is, the time-to-event for those subjects are left-truncated at some point l_i in this case. The clustered response before entry is also missing and the observable outcome becomes $\mathbf{Y}_i = (Y_{i,l_i+1}, Y_{i,l_i+2}, ..., Y_{it_i})'$. Under the assumption of missing at random (MAR) and the marginal distributions are correctly specified, we can integrate out those missing

data to derive the submodel of the observed data. Let $f_J(y_{i1}, ..., y_{it_i})$ denote the full joint distribution of the clustered outcome contributed by some woman. Without loss of generality, we assume that the j-th observation y_j is missing from the vector $\mathbf{y}_i = (y_{i1}, ..., y_{it_i})$ due to left truncation or right censoring. It can be easily shown that the joint distribution for the observed data $\mathbf{y}_i^* = (y_{i1}, ..., y_{i,j-1}, y_{i,j+1}, ..., y_{it_i})$ has the same form of the full data model. Further, we can show that if we assume that the clustered outcome is missing at random, the estimating functions of the parameters based on the observed data are unbiased (See Appendix II). Therefore, if some cluster size values are censored or truncated, the likelihood contribution from the Clayton-Oakes model can still be expressed as:

$$\prod_{j=l_i+1}^{t_i-1} \left[((j-l_i-1)\theta+1)\cdot k\cdot \exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))\cdot y_{ij}^{k-1}\cdot \exp(\theta y_{ij}^k \exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})))\right]$$
$$\times \left[\sum_{j=l_i+1}^{t_i-1} \exp(\theta y_{ij}^k \exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))) - (t_i-l_i-2)\right]^{-\frac{1}{\theta}-(t_i-l_i-1)}$$

where l_i is the left truncation point and t_i is the observed time-to-event. Consequently, the log-likelihood from the i-th subject based on the joint model (2.5) can be expressed as

$$\begin{aligned} l(\boldsymbol{\pi}; \mathbf{y}_i, t_i | T_i > l_i) = &\sum_{j=l_i+1}^{t_i-1} \log((j-l_i-1)\theta+1) + (k-1)\log y_{ij} + \log k - k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}) \\ &+ \theta \exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))y_{ij}^k \\ &- \left(\frac{1}{\theta}+t_i-l_i-1\right)\log\left[\sum_{j=l_i+1}^{t_i-1}\exp(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))) - (t_i-l_i-2)\right] \\ &+ \sum_{j=l_i+1}^{t_i}\left\{\eta_{ij}\log\left[1-\exp(-\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta}))\right] - (1-\eta_{ij})\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta})\right\}. \quad (2.7) \end{aligned}$$

The maximum likelihood estimators can be obtained by solving the score functions $U(\boldsymbol{\pi} = \partial l(\boldsymbol{\pi})/\partial\boldsymbol{\pi}$ based on the log likelihood function (2.6) (Details can be found in Appendix III).

2.3.2 Estimation of Standard Errors

The variance-covariance matrix of the estimator $\hat{\boldsymbol{\pi}} = (\hat{\theta}, \hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\alpha}})$ can be obtained by taking the second derivative of the log likelihood function (2.6), i.e., the Hessian matrix. However, the analytical expression for the variance-covariance matrix is very complex due to the complicated form of the likelihood function. Hence, we use a bootstrap method to obtain the standard errors of our parameter estimators. Specifically, random samples are selected with replacement from the observed data $\{\mathbf{y}_i, t_i, \mathbf{x}_i, l_i, \eta_{ij}\}, i = 1, ..., m; j = 1, ..., n_i$ (Efron and Tibshirani, 1993). For each random sample, estimates of parameters are calculated and the standard errors of these estimates are calculated as the bootstrap standard error estimators. The confidence intervals for the parameters can be constructed using the asymptotic normality property of the estimators.

2.4 Simulation Studies

To evaluate the performance of the joint modeling procedure, we conduct simulation studies in different settings. Five hundred replicates are performed for each setting of parameters. Similar to MSSWOW study, we first generate the random cluster size with a maximum cluster size of 12. That is, an administrative censoring time of 12 is imposed on the cluster size. For simplicity, a common baseline hazard of having an event at each cycle is assumed. In terms of covariates, we consider a categorical predictor with two groups and a continuous factor in the model, denoted by X_1 and X_2, respectively. Therefore, the cluster size for the i-th subject follows the model

$$\lambda(t_i|\mathbf{X}_i) = 1 - \exp[-\exp(\alpha + \gamma(\beta_1 X_{1i} + \beta_2 X_{2i}))]$$

where $\gamma\beta_1$ and $\gamma\beta_2$ are the regression coefficients for the random cluster size after adjusting for the repeatedly measured clustered data and different values of α will give different percentages of censoring. If a subject has an event at time t, the

corresponding clustered data will have a cluster size of $t-1$. If a subject is censored at time t, the vector of clustered observations will have a cluster size of t. Given the cluster size of t_i, the clustered observations are simulated as

$$y_{ij} = \left[-\frac{\log(S_{ij})}{\exp(-k(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}))} \right]^{\frac{1}{k}}$$

where k is the shape parameter, β_0 is the intercept, β_1 and β_2 are the covariate effects, and $S_{ij} \in (0,1)$ is a marginal survival function from the Clayton-Oakes model with a dependence parameter θ, which can be obtained from the following algorithm. The generator function φ and its inverse φ^{-1} of the Archimedean copulas is known to be the Laplace transform of some positive random variable (Marshall and Olkin 1988; Frees and Valdez 1998), which is often referred to as frailty. For the Clayton-Oakes model, this frailty follows a gamma distribution. Let ζ be a realization of the frailty and $\upsilon_1, ..., \upsilon_t$ is a vector of independent observations from uniform distribution $[0,1]$. Then $\mu_i = \varphi^{-1}(-\zeta^{-1} \log \upsilon_i), i = 1, ..., t$ is a realization form the Archimedean copula with generator φ and frailty ζ.

Each dataset is analyzed using the maximum likelihood procedure described in Section 3. The results from the simulation study are summarized in Table 1. The simulation study illustrates that the maximum likelihood estimators perform very well even when clustered outcomes are highly correlated or the percentage of censoring increases.

2.5 MSSWOW Data

In the MSSWOW study, 470 women, aged from 19 to 41, from 41 companies were followed for up to one year until a clinical pregnancy and a total of 3689 menstrual cycles were obtained. Of those participants, 179 (38.1%) of them got pregnant at the end of the study. Several risk factors are identified from MSSWOW study including age, smoking, BMI and unsafe sex etc. Previous findings have shown that menstrual

Table 2.1. Simulation Studies with 500 Replicates, Sample Size $m = 400$

Scenarios	**Association**[1]		**True**	**Bias**	**SE**[2]	**SE**[3]	**95% CP**[3]
Right Censoring:	$\tau = 0.2$	β_1	0.3	-0.0007	0.0202	0.0325	0.94
20% of subjects		β_2	0.5	0.0005	0.0255	0.0354	0.93
are censored.		γ	4.0	-0.0281	0.3729	0.4253	0.94
		θ	0.5	0.0056	0.0479	0.0694	0.94
	$\tau = 0.5$	β_1	0.3	-0.0002	0.0191	0.0299	0.96
		β_2	0.5	0.0003	0.0247	0.0364	0.96
		γ	4.0	-0.0392	0.3720	0.4245	0.92
		θ	2.0	0.0037	0.1574	0.1687	0.94
	$\tau = 0.8$	β_1	0.3	-0.0007	0.0182	0.0248	0.95
		β_2	0.5	0.0019	0.0233	0.0311	0.95
		γ	4.0	-0.0425	0.3639	0.3947	0.93
		θ	8.0	0.0022	0.3594	0.5802	0.92

[1] The association among clustered data is represented by Kendall's τ.

[2] This is the Monte Carlo standard error based on the simulations.

[3] The standard error is based on the bootstrap sampling.

[4] CP stands for the coverage probability.

cycle lengths varies in different age groups as well the risk of getting pregnant (Guo et al., 2006; Small et al., 2006). In MSSWOW data, four age groups, 19-25, 26-30, 31-35 and 36-41, are considered and the last group is set as the reference group. Descriptive statistics of the data are summarized in Table 2. The menstrual cycle lengths decrease when women get older. Women between age 31 and 35 had the highest conception rate. Figure 1 (Small et al., 2006) provides the unadjusted fertility in relation to cycle length of the previous menstrual cycle. On average, a moderate cycle length between 30 and 31 is associated with the highest probability of conception. Either shorter or longer cycle lengths may reduce the chance of getting pregnant. Figure 2 shows hazard rate of getting pregnant for each age group at each cycle.

Table 2.2. Summary of Descriptive Statistics of MSSWOW Data (m=470)

Age Group	**N Obs**	**Mean MCLs**	**Mean TTP**	**Pregnancy (%)**
Age group 19-25	65	29.40	8.48	35.38
Age group 26-30	157	28.85	7.78	41.40
Age group 31-35	157	28.62	9.00	43.31
Age group 36-41	91	26.41	10.78	25.27

The results shown in Table 3 indicate that age group has significant impacts on both TTP and menstrual cycle lengths. Women aged between 31 and 35 had the highest probability of conception. In general, menstrual cycle lengths increase with the increasing of age at first but decrease at age group between 36 and 41. This result is consistent with that from Guo et al. (2006). A significant correlation is found among menstrual cycle lengths with a Kendall's tau of 0.208. Note that the number of unsafe sex is not added to the menstrual cycle length model but only considered in TTP model due to the meaningful interpretation. As expected, the results show that the number of unsafe sex has significant impacts on TTP ($p < .001$). To further

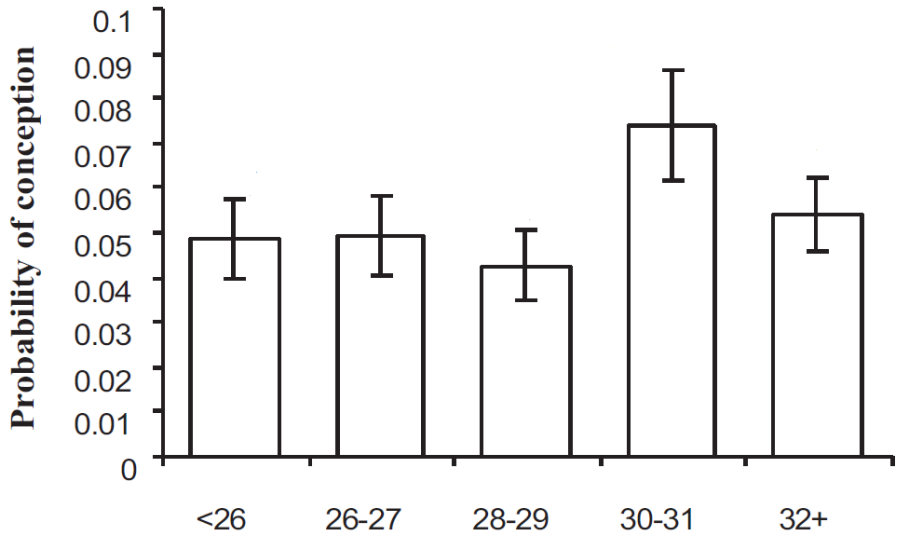

Figure 2.1. Unadjusted relationship of cycle length and risk of getting pregnant. Vertical bar represents the proportion of cycles within each cycle length category prior to pregnancy standard error bars are also given in the plot. (*Source: Small et al., 2006, Epidemiology*)

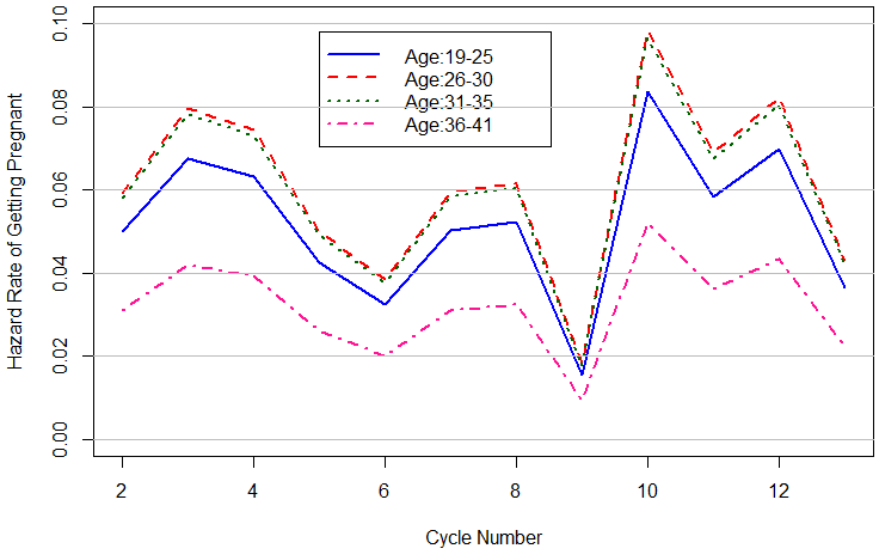

Figure 2.2. Estimated Hazard Rate of Pregnancy for Each Age Group

investigate the relationship between menstrual cycle lengths and TTP, we assume a common scaling parameter γ in the model and only consider the four age groups as the covaraites so that the regression coefficients $\beta_{1b}, b = 1, 2, 3$ represents the estimated difference of menstrual cycle lengths between different age groups. Table 3 shows that menstrual cycle length has a significant influence on TTP ($p = 0.025$). Specifically, the estimated γ is positive, indicating that the longer the menstrual cycle lengths are, the risk of getting pregnant gets higher. Figure 1 displays the pregnancy rate for each group accounting for the impact of baseline hazard and menstrual cycle lengths.

Table 2.3. Analysis of MSSWOW Data (m=470)

Model	**Effects**		**Estimates**	**SE**	**P-value**
Joint Model	Intercept	β_0	3.339	0.013	<.001
	Age group 19-25	β_{11}	0.117	0.019	<.001
	Age group 26-30	β_{12}	0.126	0.006	<.001
	Age group 31-35	β_{13}	0.181	0.014	<.001
	Age group 36-41	-	-	-	-
	Baseline for TTP:$\leq$2	α_1	-3.406	0.268	<.001
	Baseline for TTP:3 $\sim$ 8	α_2	-3.533	0.213	<.001
	Baseline for TTP:=9	α_3	-4.750	0.600	<.001
	Baseline for TTP:10 $\sim$ 12	α_4	-3.180	0.243	<.001
	Baseline for TTP:$\geq$13	α_5	-3.934	0.222	<.001
	Scaling parameter	γ	2.850	1.272	0.025
	Unsafe Sex	β_4	0.078	0.010	<.001
Association among	Association	θ	0.526	0.032	<.001
Cycle Lengths	Kendall's tau	τ	0.208	-	-
	Shape	κ	2.521	0.053	<.001

2.6 Remarks

In this chapter, we have proposed a joint modeling framework for the analysis of clustered data with a random cluster size where the clustered data was modeled via a Clayton-Oakes model and the random cluster size was treated as a discrete survival time that was assumed to follow a complementary log-log hazard function. Particularly, we imposed the parametric Weibull assumption for the Clayton-Oakes model as the marginal model. Maximum likelihood estimators are obtained based on the fully parametric specification of the model.

In the application to MSSWOW data, we found that the association between women's menstrual cycle lengths is significant. The menstrual cycle lengths have significant influence on the risk of getting pregnant, i.e., time-to-pregnancy (TTP). However, we can see that the inference based on the joint model is slightly different from what we observed in the data. This might imply that the Weibull model does not fit the data very well. In next chapter, we will relax this parametric assumption and use more generalized semi-parametric models.

Appendix I: Probability Mass Function of TTP

For the i-th subject, the probability of observed TTP $T_i = t_i$ conditioning on $T_i > l_i$ is

$$
\begin{aligned}
\Pr\{T_i = t_i | T_i > l_i, \mathbf{X}_i\} &= \left[Pr\{\tilde{T}_i = t_i | \tilde{T}_i > l_i, \mathbf{X}_i\}\right]^{\eta_i} \left[Pr\{\tilde{T}_i > t_i | \tilde{T}_i > l_i, \mathbf{X}_i\}\right]^{1-\eta_i} \\
&= \left[\frac{Pr\{\tilde{T}_i = t_i + l_i | \mathbf{X}_i\}}{Pr\{\tilde{T}_i > l_i | \mathbf{X}_i\}}\right]^{\eta_i} \left[\frac{Pr\{\tilde{T}_i > t_i + l_i | \mathbf{X}_i\}}{Pr\{\tilde{T}_i > l_i | \mathbf{X}_i\}}\right]^{1-\eta_i} \\
&= \left[\frac{\lambda(t_i|\mathbf{X}_i) \prod_{j=1}^{t_i-1}(1-\lambda(j|\mathbf{X}_i))}{\prod_{j=1}^{l_i}(1-\lambda(j|\mathbf{X}_i))}\right]^{\eta_i} \left[\frac{\prod_{j=1}^{t_i}(1-\lambda(j|\mathbf{X}_i))}{\prod_{j=1}^{l_i}(1-\lambda(j|\mathbf{X}_i))}\right]^{1-\eta_i} \\
&= \left[\frac{\lambda(t_i + l_i|\mathbf{X}_i)}{1-\lambda(t_i + l_i|\mathbf{X}_i)}\right]^{\eta_i} \prod_{j=l_i+1}^{t_i+l_i} (1-\lambda(j|\mathbf{X}_i)) \\
&= \prod_{j=l_i+1}^{t_i} \left[\frac{\lambda(j|\mathbf{X}_i)}{1-\lambda(j|\mathbf{X}_i)}\right]^{\eta_{ij}} (1-\lambda(j|\mathbf{X}_i)) \\
&= \prod_{j=l_i+1}^{t_i} \left[\frac{1-\exp(-\exp(\alpha(j) + \gamma\mathbf{X}_i\boldsymbol{\beta}))}{\exp(-\exp(\alpha(j) + \gamma\mathbf{X}_i\boldsymbol{\beta}))}\right]^{\eta_{ij}} \exp(-\exp(\alpha(j) + \gamma\mathbf{X}_i\boldsymbol{\beta})).
\end{aligned}
$$

Appendix II: Unbiasness of the Estimating Equation

Since we assume the longitudinal survival outcome suffers from the missing issue, we only need to consider the likelihood contribution from the clustered survival data. Given cluster size $\tilde{T}_i = N$, the complete clustered survival data for the i-th subject is $\mathbf{Y}_{i(all)} = (y_{i1}, ..., y_{i,N-1})$ where $N-1$ is the largest possible cluster size. Denote the corresponding missing data indicator vector by $\mathbf{M}_i = (m_{i1}, ..., m_{i,N-1})$ where $m_{ij} = 1, j = 1, ..., N-1$ if y_{ij} is observed and 0 if missing. The observed data is denoted by $\mathbf{Y}_{i(obs)}$ and $\mathbf{Y}_{i(mis)}$ represents the missing data. Under the assumption of missing at random (MAR), the joint distribution of $\mathbf{Y}_{i(obs)}$ and $\mathbf{M}_i$ can be written as

$$f(\mathbf{y}_{i(obs)}, \mathbf{M}_i|\boldsymbol{\pi}, \boldsymbol{\phi}) = f(\mathbf{y}_{i(obs)}|\boldsymbol{\pi}) f(\mathbf{M}_i|\boldsymbol{\phi}).$$

For our proposed model, we make the following assumptions with regards to the missing mechanism of the clustered survival time:

Assumption 1. The cluster survival times are missing at random. That is, the probability of missingness may be related to the observed data but not depend on the missing values.

Assumption 2. The missingness does not depend on covariates so that $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ are distinct.

Let $U_i(\boldsymbol{\pi}; \mathbf{y}_{i(obs)})$ denote the score function of the observed clustered survival data contributed by the i-th subject. If the marginal models are correctly specified, then we have the following results under Assumptions 1 and 2:

$$\begin{aligned}
E\left(U_i(\boldsymbol{\pi}; \mathbf{y}_{i(obs)})\right) &= \int_{\mathbf{y}_i} \int_{\mathbf{M}_i} U_i(\boldsymbol{\pi}; \mathbf{y}_{i(obs)}) f(\mathbf{y}_{i(obs)}, \mathbf{M}_i|\boldsymbol{\pi}, \boldsymbol{\phi}) d\mathbf{M}_i d\mathbf{y}_i \\
&= \int_{\mathbf{y}_{i(obs)}} U_i(\boldsymbol{\pi}; \mathbf{y}_{i(obs)}) \left\{ \int_{\mathbf{y}_{i(mis)}} \int_{\mathbf{M}_i} f(\mathbf{y}_{i(obs)}, \mathbf{M}_i|\boldsymbol{\pi}, \boldsymbol{\phi}) d\mathbf{M}_i d\mathbf{y}_{i(mis)} \right\} d\mathbf{y}_{i(obs)} \\
&= \int_{\mathbf{y}_{i(obs)}} U_i(\boldsymbol{\pi}; \mathbf{y}_{i(obs)}) f(\mathbf{y}_{i(obs)}|\boldsymbol{\pi}) d\mathbf{y}_{i(obs)} \\
&= 0.
\end{aligned}$$

Therefore, we have proved that the score function is unbiased as long as MAR holds and the marginal density is correct. Moreover, inference for $\boldsymbol{\pi}$ based on the

score function $U_i(\boldsymbol{\pi}; \mathbf{y}_{i(obs)})$ is the same as that from the full data score function $U_i(\boldsymbol{\pi}; \mathbf{y}_{i(all)})$. In other words, the maximum likelihood estimators by ignoring the missing data are efficient. If assumption 2 is not true, i.e., distinctness of $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ does not hold, we can still obtain unbiasedness of the estimating function. However, the estimators are not fully efficient since we ignore the missingness which has contribution to the estimation of $\boldsymbol{\pi}$.

Appendix III: Estimating Equations

$$
\begin{aligned}
\theta: & \sum_{i=1}^{m}\sum_{j=l_i+1}^{t_i-1}\left[\frac{j-l_i-1}{(j-l_i-1)\theta+1}+\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))y_{ij}^k\right] \\
& +\sum_{i=1}^{m}\frac{1}{\theta^2}\log\left[\sum_{j=l_i+1}^{t_i}\exp\left(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))\right)-(t_i-l_i-2)\right] \\
& -\sum_{i=1}^{m}\left(\frac{1}{\theta}+t_i-l_i-1\right)\frac{\sum_{j=l_i+1}^{t_i-1}y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))\exp(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})))}{\sum_{j=l_i+1}^{t_i-1}\exp(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})))-(t_i-l_i-2)}=0 \\
k: & \sum_{i=1}^{m}\left(\log y_{ij}+\frac{1}{k}-\mathbf{x}_i\boldsymbol{\beta}\right)+\sum_{i=1}^{m}\sum_{j=l_i+1}^{t_i-1}\theta\mathbf{x}_i\boldsymbol{\beta}\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))y_{ij}^k+\theta y_{ij}^k\log y_{ij}\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})) \\
& -\sum_{i=1}^{m}\left(\frac{1}{\theta}+t_i-l_i-1\right)\frac{\sum_{j=l_i+1}^{t_i-1}\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))(\log y_{ij}-\mathbf{x}_i\boldsymbol{\beta})\exp(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})))}{\sum_{j=l_i+1}^{t_i-1}\exp(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})))-(t_i-l_i-2)}=0 \\
\alpha_j: & \sum_{i=1}^{m}\eta_{ij}\frac{\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta})\exp(-\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta}))}{1-\exp(-\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta}))}-\sum_{i=1}^{m}(1-\eta_{ij})\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta})=0 \\
\gamma: & \sum_{i=1}^{m}\sum_{j=l_i+1}^{t_i}\eta_{ij}\frac{\mathbf{x}_i\boldsymbol{\beta}\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta})\exp(-\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta}))}{1-\exp(-\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta}))} \\
& -\sum_{i=1}^{m}\sum_{j=l_i+1}^{t_i}(1-\eta_{ij})\mathbf{x}_i\boldsymbol{\beta}\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta})=0 \\
\beta_0: & -\sum_{i=1}^{m}\sum_{j=l_i+1}^{t_i-1}\left\{k+k\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))\right\} \\
& +\sum_{i=1}^{m}\left(\frac{1}{\theta}+t_i-l_i-1\right)\frac{\sum_{j=l_i+1}^{t_i-1}k\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))\exp(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})))}{\sum_{j=l_i+1}^{t_i-1}\exp(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})))-(t_i-l_i-2)} \\
\beta_b: & -\sum_{i=1}^{m}\sum_{j=l_i+1}^{t_i-1}\left\{kx_{ib}+k\theta x_{ib}y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))\right\} \\
& +\sum_{i=1}^{m}\left(\frac{1}{\theta}+t_i-l_i-1\right)\frac{\sum_{j=l_i+1}^{t_i-1}kx_{ib}\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta}))\exp(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})))}{\sum_{j=l_i+1}^{t_i-1}\exp(\theta y_{ij}^k\exp(-k(\beta_0+\mathbf{X}_i\boldsymbol{\beta})))-(t_i-l_i-2)} \\
& +\sum_{i=1}^{m}\sum_{j=l_i+1}^{t_i}\left[\eta_{ij}\frac{\gamma_b x_{ib}\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta})\exp(-\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta}))}{1-\exp(-\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta}))}-(1-\eta_{ij})x_{ib}\gamma_b\exp(\alpha_j+\gamma\mathbf{x}_i\boldsymbol{\beta})\right]=0
\end{aligned}
$$

where $b=1,\ldots,p$ is the b-th parameter and x_{ib} is the b-th covariate for subject i.

Chapter 3

Semiparametric Joint Models

3.1 Introduction

There has been extensive literature for the analysis of correlated data with different types of outcomes including continuous, ordinal and survival outcomes etc. (Laird and Ware, 1982; Cai and Prentice, 1995; Liang and Zeger, 1995). In such literature, multiple measurements from each subject are correlated and the number of these measurements is treated as fixed. When the number of measurements is a random variable and may be an outcome of interest, joint models with random effects have been considered to model longitudinal measurements and the number of measurements. These joint models aim to determine the influence of covariates on both outcomes, the within-subject correlation as well as the association between the two outcomes (Hogan and Laird, 1997; Dunson et al., 2003). Alternatively, Barnhart and Sampson (1995) described this type of data as random length data where a vector of observations is collected for each subject and the length of the vector is also a random variable. They considered a multivariate normal distribution for the correlated outcomes and the marginal distributions are parameterized in terms of treatment effects. In addition, the treatment effects are modified by a scalar to reflect the influence of the treatment on the distribution of the random length, which is modeled via a generalized linear model.

In this Chapter, we study the modeling approach as in Barnhart and Sampson (1995), but we consider the copula model for repeated measures to relax the multivariate normality assumption and the distribution of the length of the vector is specified via a generalized linear model. Our work is motivated by a reproductive study called the Mount Sinai Study of Women Office Workers (MSSWOW). In this study, women were followed prospectively for one year in order to study fertility until a clinical pregnancy or the end of the study. Multiple measures of menstrual cycle lengths (MCLs) for each subject were observed. Time-to-pregnancy (TTP) was defined as the number of menstrual cycles taken to conceive excluding the conception cycle (when a woman gets pregnant and hence is no longer at risk of menstrual bleeding, MCL at the conception cycle cannot be observed). Therefore, it is natural to view MCL collected from the first conception attempt to conception as a vector containing multiple measurements of MCL with random vector length equal to TTP. As in many other epidemiologic studies, several complicated issues have been raised by MSSWOW study. First, there is evidence showing that the MCL data have a long right tail and therefore a normal distribution is not adequate (Harlow and Zeger, 1991; Murphy et al., 1995; Harlow et al., 2000; Guo et al., 2006). Second, when a subject does not conceive at the end of the study, the subject's TTP is censored and the remaining MCLs until pregnancy are missing. Third, since some subjects were trying to get pregnant before entering the study, TTP is left-truncated and MCLs before entry are also missing.

Recently, McLain et al. (2012) considered a Bayesian framework for modeling MCL and TTP simultaneously. They used a mixture of normal and Weibull (or extreme-value) distribution to handle the skewness of the distribution of MCLs. However, with skewed continuous outcomes, parametric method with mixture distributions imposes computational difficulties and raises questions about the robustness of inference to its assumptions; therefore, more flexible semiparametric modeling is

desired. This motivates us to specify the marginal distribution of MCLs via a general class of semiparametric transformation model (Zeng and Lin, 2007; Chen and Yu, 2012), which includes proportional hazards model and proportional odds model as two special cases.

We consider a joint model where the repeated measurements are assumed to follow a special type of copula model known as the Clayton-Oakes model (Clayton, 1978; Oakes, 1989) and the marginal distribution of the Clayton-Oakes model follows the semiparametric transformation model. Furthermore, the length of the vector, which corresponds to TTP, is modeled via a discrete time hazard model (Scheike and Jensen, 1997). This discrete survival model can be expressed as a complementary log-log link model and retains an interpretation in terms of the underlying proportional hazards for grouped data (Kalbfleisch and Prentice, 2002). To understand the relationship between multiple measurements and the random length (e.g., MCL and TTP), shared parameters are imposed on both the Clayton-Oakes model and the discrete survival time model. Glidden and Self (1999) proposed a semiparametric estimation approach for the Clayton-Oakes model, when the vector length is fixed and the marginal distribution follows a proportional hazards model. We adopt a similar estimation method, but in our case we propose a shared-parameter joint model, assuming a more general specification for the marginal distributions and a generalized linear model with a complementary log-log link function for the random length.

In Section 2, we describe our joint modeling framework and marginal model specification as well as properties of the model. Due to the semiparametric nature of the model specification, direct maximum likelihood estimators are not available. Section 3 proposes an approximate EM-algorithm to derive generalized maximum likelihood estimators for the parameters in the joint model. In Section 4, we conduct general simulation studies to evaluate the performance of the proposed method. We apply our method to the MSSWOW study in Section 5.

3.2 The Models

Suppose that we have m subjects. Let i index the subject and j index the measurement for each subject. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iN_i})'$ denote a vector of multiple measurements on a quantitative variable with a random length N_i for the i-th subject, $i = 1, ..., m$. In addition, a p-dimensional covariate vector $\mathbf{X}_i$ is collected for each subject which may affect both the repeatedly measured outcome and the random length. The joint model is based on the factorization of the joint distribution of $(\mathbf{Y}_i, N_i)$.

First, we define the model for the vector of multiple measurements $\mathbf{Y}_i$ given N_i. Conditional on $N_i = n_i$, we assume that $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ has a multivariate survival distribution that follows the Clayton-Oakes model (Clayton, 1978; Oakes, 1989) as

$$S_J(\mathbf{y}_i|N_i = n_i, \mathbf{X}_i; \boldsymbol{\beta}) = \Pr\{Y_{i1} > y_{i1}, ..., Y_{in_i} > y_{in_i}|N_i = n_i, \mathbf{X}_i; \boldsymbol{\beta}\} = \left[\sum_{j=1}^{n_i} S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta})^{-\theta} - n_i + 1\right]^{-\frac{1}{\theta}} \tag{3.1}$$

where $S_J(\cdot)$ indicates a joint distribution of the multiple measurements, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters associated with $\mathbf{X}_i$, $S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta})$ represents the marginal survivor distribution of the j-th observation for the i-th subject given $\mathbf{X}_i$, and θ depicts the within-subject dependence. As θ approaches 0, the observations within the same subject become independent, and the joint survival function is simply the product of the marginal survival functions. When θ goes to $+\infty$, the joint survival function converges to its upper Frechet bound and $S_J(\mathbf{y}_i|N_i = n_i, \mathbf{X}_i; \boldsymbol{\beta}) = \min\{S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}), j = 1, ..., n_i\}$. In addition, θ is related to Kendall's (1962) coefficient of concordance known as the Kendall's τ which can be expressed as $\tau = \theta/(\theta + 2)$. The model implies that the association between any two measurements from the same subject is constant, which is a reasonable assumption for the MSSWOW study.

For the marginal distribution of each Y_{ij}, we consider a class of semiparametric linear transformation models. That is, we specify $S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ in model (3.1) where Y_{ij} depends on the covariates via an unknown function $q(\cdot)$ as

$$q(Y_{ij}) = -\mathbf{X}_i^T\boldsymbol{\beta} + \varepsilon_{ij}, \tag{3.2}$$

and $q(\cdot)$ is a completely unspecified and strictly increasing function and ε_{ij} is a random error with a known distribution function denoted by F_ε. Let $S_\varepsilon = 1 - F_\varepsilon$ be the survivor function for ε, then the marginal survival function of Y_{ij} given $\mathbf{X}_i$ can be written as $S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = S_\varepsilon(q(y_{ij}) + \mathbf{X}_i^T\boldsymbol{\beta})$. Then the hazard function of Y_{ij} can be written as $h(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = \partial q(y_{ij})/\partial y_{ij} \cdot h_\varepsilon(q(y_{ij}) + \mathbf{X}_i^T\boldsymbol{\beta})$. If we reparameterize the transformation model as $\Phi(y_{ij}) = \exp(q(y_{ij})) = \exp(-\mathbf{X}_i^T\boldsymbol{\beta})\exp(\varepsilon_{ij})$ where $\Phi(\cdot)$ is a strictly increasing positive function with $\Phi(0) = 0$ and $\lim_{y\to\infty}\Phi(y) = \infty$, we can write the hazard function for Y_{ij} given $\mathbf{X}_i$ as

$$h(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = \phi(y_{ij})\exp(\mathbf{X}_i^T\boldsymbol{\beta})h_0(\Phi(y_{ij})\exp(\mathbf{X}_i^T\boldsymbol{\beta})) \tag{3.3}$$

where $\phi(y_{ij}) = \Phi'(y_{ij})$ and $h_0(\cdot)$ is the hazard function associated with $\exp(\varepsilon)$.

Various choices for S_ε will generate different marginal models. For example, if S_ε follows the extreme value distribution as $S_\varepsilon(s) = \exp(-\exp(s))$, model (3.2) becomes the familiar proportional hazards model where the hazard function $h(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = \phi(y_{ij})\exp(\mathbf{X}_i^T\boldsymbol{\beta})$ and $\phi(y_{ij})$ is the unspecified baseline hazard function in this case. If ε_{ij} has a standard logistic distribution with $S_\varepsilon(s) = \exp(s)/(1+\exp(s))$, model (3.2) yields the proportional odds model that has a form of $logit(1 - S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})) = \log(\Phi(y_{ij})) + \mathbf{X}_i^T\boldsymbol{\beta}$ in which case $\Phi(y_{ij})$ is the baseline odds. A more general form for S_ε can be expressed as the class of logarithmic transformations $S_\varepsilon(s) = [1 + r\exp(s)]^{-\frac{1}{r}}$, $r \geq 0$ (Dabrowska and Doksum, 1988; Chen and Yu, 2012), where $r = 0$ corresponds to the proportional hazards model and $r = 1$ yields the proportional odds model.

For the random length N_i, which for example may capture TTP, a discrete survival

outcome, we assume that N_i follows a discrete distribution with a general form of

$$\Pr\{N_i = n_i|\mathbf{X}_i^T\boldsymbol{\beta};\gamma,\boldsymbol{\alpha}\} = \nu(\alpha(n_i) + \gamma\mathbf{X}_i^T\boldsymbol{\beta}),\ n_i = 0, 1, ..., M, \tag{3.4}$$

where $\nu(\cdot)$ is a probability density function, $M > 0$ is a known positive integer, and $\nu(\alpha(n_i))$ denotes the baseline probability density. The parameter $\boldsymbol{\beta}$ is shared by the Clayton-Oakes model (3.1) and the discrete model (3.4), and γ is a scaling parameter that evaluates the impact of shifts or changes in the distribution of $\mathbf{Y}_i$ with respect to covariates $\mathbf{X}_i$ on the distribution of N_i. In other words, the association between $\mathbf{Y}_i$ and N_i is induced by the covariates $\mathbf{X}_i$. The parameterization also allows the two models contain different covariates. For example, if a predictor X_{ib} is included in the generalized linear model but not in the Clayton-Oakes model, the b-th regression coefficient β_b in the vector $\boldsymbol{\beta}$ in the Clayton-Oakes is set to be zero.

In this dissertation, we focus on the case where the random length N_i is a discrete survival time, though it is straightforward to extend the modeling approach to other discrete distributions. In particular, we use the complementary log-log (CLL) function for modeling the hazard rate of N_i (Scheike and Jensen, 1997; Kalbfleish and Prentice, 2002). That is, we write the hazard rate of the random length variable N_i given $\mathbf{X}_i$ as

$$\lambda\left(n_i|\mathbf{X}_i^T\boldsymbol{\beta};\gamma,\boldsymbol{\alpha}\right) = 1 - \exp\left(-\exp(\alpha(n_i) + \gamma\mathbf{X}_i^T\boldsymbol{\beta})\right), n_i = 0, 1, ..., M, \tag{3.5}$$

where the parameters $\alpha(n_i)$, $\boldsymbol{\beta}$ and γ are as defined as before. Under the assumption of model (3.5), the probability density of the discrete distribution in (3.4) has a form of

$$\Pr\{N_i = n_i|\mathbf{X}_i^T\boldsymbol{\beta};\gamma,\boldsymbol{\alpha}\} = \lambda(n_i|\mathbf{X}_i^T\boldsymbol{\beta};\gamma,\boldsymbol{\alpha}) \prod_{j=1}^{n_i-1}\left(1 - \lambda(j|\mathbf{X}_i^T\boldsymbol{\beta};\gamma,\boldsymbol{\alpha})\right)$$

Combining the Clayton-Oakes model (3.1) and the discrete model (3.4), a general form of the joint model for the multiple measurements and the random length $(\mathbf{Y}_i, N_i)$

is derived as

$$\begin{aligned}
&\Pr\{Y_{i1} > y_{i1}, ..., Y_{in_i} > y_{iN_i}, N_i = n_i | \mathbf{X}_i; \boldsymbol{\pi}\} \\
&= \left[\sum_{j=1}^{n_i} S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})^{-\theta} - n_i + 1\right]^{-\frac{1}{\theta}} \Pr\{N_i = n_i | \mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha}\} \qquad (3.6)
\end{aligned}$$

where $\boldsymbol{\pi} = (\theta, \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})$ is the vector of parameters to estimate.

Model (3.6) has some interesting properties (Barnhart and Sampson, 1995). First, zero random length is allowed for some subjects. In MSSWOW data, some women may get pregnant at the first menstrual cycle. In this case, no measurements of MCLs are observed, which results in a random length of zero. Second, the random length can take either any non-negative integer value or non-negative integer up to some known value $M < \infty$. Third, a stochastic ordering property has been imposed by specifying the shared parameters and the scaling parameter, which can describe the association between the multiple measurements and the random length. Specifically, if the probability density function stochastically increases with the increasing of n_i, it implies that for a covariate $X_{ib}, b = 1, ..., p$, when the scaling parameter $\gamma > 0$, the larger the value of $\beta_b X_{ib}$ is, the more likely a larger random length as well as a larger value of each component in the vector of multiple measurements would be observed, and vice versa for $\gamma < 0$. In addition, the multivariate distribution of the multiple measurements is invariant under permutation of its components because of the summation form of the Clayton-Oakes model. Another important feature of the model is that the sub-model of the Clayton-Oakes model maintains the same Clayton-Oakes form as the full model, which becomes convenient to handle the missingness in the multiple measurements.

3.3 Parameter Estimation

We consider the estimation of the joint model (3.6) where $S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ and $\Pr\{N_i = n_i|\mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha}\}$ are specified via models (3.2) and (3.5), respectively. Due to the semi-

parametric specification in the Clayton-Oakes model part of the joint model, the likelihood of $(\mathbf{Y}_i, N_i)$ involves the unknown function $\Phi(\cdot)$. We adopt an approximate EM algorithm for parameter estimation, which is similar to the approach suggested by Glidden and Self (1999) for the Clayton-Oakes model with a marginal proportional hazards model where the length of the vector is fixed. We start with constructing the joint likelihood by exploiting the equivalence of Clayton-Oakes model and gamma frailty model (Nielsen et al., 1992; Klein, 1992). Under the gamma frailty model, dependence among the repeated measures from the same subject is captured by an unobservable frailty. We propose to use a two-level EM algorithm for the joint model (3.6). At the first level, the association parameter θ is fixed and an EM algorithm is used to derive generalized maximum likelihood estimators (GMLEs) of other parameters in the joint model. With θ fixed, in the E-step, the expected values of the latent frailties are calculated with respect to the observed data and the M-step involves the maximization of the full joint likelihood function and the unknown baseline function is estimated by a Breslow-type estimator. At the second level of the iteration, the profile likelihood of θ is maximized to obtain GMLE of θ. The steps are iterated until convergence is achieved.

3.3.1 Likelihood Construction

The form of joint model (3.6) implies that the likelihood from the joint density contains two components, denoted by LF_1 and LF_2, respectively. The first component LF_1 is the likelihood contribution from the vector of multiple measurements $\mathbf{Y}_i$ conditioning on N_i, and LF_2 denotes the likelihood function for the random length N_i. The full joint likelihood function for $\boldsymbol{\pi} = (\boldsymbol{\beta}, \theta, \gamma, \boldsymbol{\alpha})$ is given by

$$LF(\boldsymbol{\pi}|\mathbf{Y}, N) = LF_1(\boldsymbol{\beta}, \theta|\mathbf{Y}, N) \cdot LF_2(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}|N).$$

First, we consider the likelihood contribution from the random length N_i, i.e., LF_2. As a time-to-event variable, the random length N_i involves truncation and censoring issues. Let L_i denote the left truncation variable and C_i be the right censoring time, which is assumed to be independent of N_i. Conditioning on that $N_i \geq L_i$, the observed data for the i-th subject on the discrete time-to-event N_i consists of $(\tilde{N}_i, \Delta_i)$, where $\tilde{N}_i = \min(N_i, C_i)$ and $\Delta_i = I(N_i \leq C_i)$ is the censoring indicator. Assuming that a subject enters the study after time l_i, i.e., the left truncation $L_i = l_i$, and an event occurs at time $N_i = n_i$, the likelihood contribution of N_i from the i-th subject is

$$\Pr\{N_i = n_i | N_i > l_i, \mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha}\} = \lambda(n_i|\mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha}) \prod_{j=l_i+1}^{n_i-1} \left(1 - \lambda(j|\mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha})\right)$$

where $\lambda(j|\mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha}) = 1 - \exp(-\exp(\alpha(j) + \gamma\mathbf{X}_i^T\boldsymbol{\beta}))$ is as defined in model (3.5). In the scenario where a subject enters the study after time l_i, but the event time N_i is censored at n_i, the likelihood function for N_i is

$$Pr\{N_i > n_i | N_i > l_i, \mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha}\} = \prod_{j=l_i+1}^{n_i} \left(1 - \lambda(j|\mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha})\right).$$

Therefore, given $L_i = l_i$ and $\mathbf{X}_i$, the likelihood contribution from N_i taking into account the left truncation and right censoring is given by

$$LF_2(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}|N) = \prod_{i=1}^{m} \prod_{j=l_i+1}^{n_i} \left(\frac{\lambda(j|\mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha})}{1 - \lambda(j|\mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha})}\right)^{\eta_{ij}} (1 - \lambda(j|\mathbf{X}_i^T\boldsymbol{\beta}; \gamma, \boldsymbol{\alpha})) \quad (3.7)$$

where η_{ij} is the longitudinal censoring indicator for N_i. $\eta_{ij} = 1$ if an event occurs at the the j-th time for the i-th individual, and $\eta_{ij} = 0$ otherwise.

In order to evaluate the likelihood contribution $LF_1(\boldsymbol{\beta}, \theta|\mathbf{Y}, N)$ from the Clayton-Oakes model part, we utilize a key feature that the Clayton-Oakes model can be obtained as a gamma frailty model (Clayton, 1978; Oakes, 1982; Glidden and Self, 1999). Assuming the observations from the i-th subject are independent conditional on a latent frailty, denoted by μ_i, we write the hazard rate for Y_{ij} as

$$\lim_{s \to 0} \frac{\Pr(y_{ij} \leq Y_{ij} < y_{ij} + s | Y_{ij} \geq y_{ij}, \mu_i, \mathbf{X}_i; \boldsymbol{\beta}, \theta)}{s} = \mu_i h^*(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}) \quad (3.8)$$

where $h^*(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ is called the basic hazard functions for gamma frailty model, and μ_i has a gamma distribution with mean one and variance θ. The representation of (3.8) provides a gamma frailty model (Vaupel et al., 1979; Glidden and Self, 1999). It has been shown that a gamma frailty model has a joint survival function in the form of the Clayton-Oakes model if the basic hazard function is written as

$$h^*(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = h(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) \exp\{\theta H(y_{ij} - |\mathbf{X}_i;\boldsymbol{\beta})\}$$

where $h(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ is the hazard function of Y_{ij} associated with the marginal distribution $S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ in the Clayton-Oakes model and $H(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = \int_0^{y_{ij}} h(s|\mathbf{X}_i;\boldsymbol{\beta})ds$ (Clayton, 1978; Oakes, 1982). Based on the hazard function $h(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ defined in (3.3), the basic hazard function in the gamma frailty model can be written as

$$h^*(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = \phi(y_{ij}) \exp(\mathbf{X}_i^T\boldsymbol{\beta}) h_0(\Phi(y_{ij}) \exp(\mathbf{X}_i^T\boldsymbol{\beta})) \exp\{\theta H_0(\Phi(y_{ij}-) \exp(\mathbf{X}_i^T\boldsymbol{\beta}))\} \tag{3.9}$$

where $h_0(\cdot)$ is the hazard function associated with $\exp(\varepsilon)$ and $H_0(y) = \int_0^y h_0(s)ds$.

We take equation (3.9) as our basic model for the intensity of the associated counting processes for Y_{ij} where $\phi(y_{ij})$ is treated as the unknown baseline function. Define the counting processes as $K_{ij}(y) = I(Y_{ij} \geq y)$ and $N_{ij}(y) = I(Y_{ij} \leq y)$, $y \in [0, \tau]$ where τ is the upper bound for Y_{ij}. Given $L_i = l_i$ and $N_i = n_i$, the "complete data" for $\mathbf{Y}_i$ is defined as a filtration of $\mathcal{G} = \{\mu_i, N_{ij}(s), K_{ij}(s+), Y_{ij}, \mathbf{X}_i, 0 \leq s \leq y; j = l_i + 1, ..., n_i, i = 1, ..., m\}$ which contains the unobservable frailty μ_i for each subject. The "incomplete data" is observations of the filtration of $\mathcal{F} = \{N_{ij}(s), K_{ij}(s+), Y_{ij}, \mathbf{X}_i, 0 \leq s \leq y; j = l_i + 1, ..., n_i, i = 1, ..., m\}$. The (partial) likelihood of "complete data" for the Clayton-Oakes model part can be obtained as

$$LF_1^{\mathcal{G}}(\boldsymbol{\beta}, \theta|\mathbf{Y}, N) =$$
$$\prod_{i=1}^{m} \prod_{j=l_i+1}^{n_i} g(\mu_i;\theta) \exp\left(-\mu_i \int_0^{\tau} K_{ij}(s) dH_{ij}^*(s|\mathbf{X}_i;\boldsymbol{\beta})\right) \prod_{y\in[0,\tau]} \left(\mu_i K_{ij}(y) h_{ij}^*(y|\mathbf{X}_i;\boldsymbol{\beta})\right)^{dN_{ij}(y)}$$

where $g(\cdot;\theta)$ is the density for gamma random variable with mean one and variance θ and $H^*(y) = \int_0^y h^*(s)ds$. Under regularity conditions, integrating over the frailties

gives the (partial) likelihood for the "incomplete data" of $\mathcal{F}$

$$LF_1^{\mathcal{F}}(\boldsymbol{\beta},\theta|\mathbf{Y},N) =$$
$$\prod_{i=1}^{m}\frac{\theta^{-\theta^{-1}}\Gamma(\theta^{-1}+N_{i\cdot}(\tau))}{\Gamma(\theta^{-1})\left(\theta^{-1}+\sum_{j=1}^{n_i}\int_0^\tau K_{ij}(y)dH_{ij}^*(y|\mathbf{X}_i;\boldsymbol{\beta})\right)^{\theta^{-1}+N_{i\cdot}(\tau)}}\prod_{j=l_i+1}^{n_i}\prod_{y\in[0,\tau]}\left(K_{ij}(y)h_{ij}^*(y|\mathbf{X}_i;\boldsymbol{\beta})\right)^{dN_{ij}(y)}$$

where "$\cdot$" in the subscript indicates a sum over the corresponding index and Γ is the gamma function. Therefore, the (partial) likelihood of the joint model for the "complete data" is written as

$$\begin{aligned}
LF^{\mathcal{G}}(\boldsymbol{\pi}) &= LF_1(\boldsymbol{\beta},\theta|\mathbf{Y},N)\cdot LF_2(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}|N)\\
&\propto\prod_{i=1}^{m}g(\mu_i;\theta)\prod_{j=l_i+1}^{n_i}\exp\left(-\mu_i\int_0^\tau K_{ij}(y)dH_{ij}^*(y|\mathbf{X}_i;\boldsymbol{\beta})\right)\\
&\cdot\prod_{y\in[0,\tau]}\left(\mu_iK_{ij}(y)h_{ij}^*(y|\mathbf{X}_i;\boldsymbol{\beta})\right)^{dN_{ij}(y)}\\
&\cdot\prod_{i=1}^{m}\prod_{j=l_i+1}^{n_i}\left[\frac{1-\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i^T\boldsymbol{\beta}))}{\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i^T\boldsymbol{\beta}))}\right]^{\eta_{ij}}\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i^T\boldsymbol{\beta}))
\end{aligned}\tag{3.10}$$

The (partial) likelihood for the "incomplete data" is given by

$$\begin{aligned}
LF^{\mathcal{F}}(\boldsymbol{\pi}) &= LF_1(\boldsymbol{\beta},\theta|\mathbf{Y},N)\cdot LF_2(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}|N)\\
&\propto\prod_{i=1}^{m}\frac{\theta^{-\theta^{-1}}\Gamma(\theta^{-1}+N_{i\cdot}(\tau))}{\Gamma(\theta^{-1})\left(\theta^{-1}+\sum_{j=1}^{n_i}\int_0^\tau K_{ij}(y)dH_{ij}^*(y|\mathbf{X}_i;\boldsymbol{\beta})\right)^{\theta^{-1}+N_{i\cdot}(\tau)}}\\
&\cdot\prod_{j=l_i+1}^{n_i}\prod_{y\in[0,\tau]}\left(K_{ij}(y)h_{ij}^*(y|\mathbf{X}_i;\boldsymbol{\beta})\right)^{dN_{ij}(y)}\\
&\cdot\prod_{i=1}^{m}\prod_{j=l_i+1}^{n_i}\left[\frac{1-\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i^T\boldsymbol{\beta}))}{\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i^T\boldsymbol{\beta}))}\right]^{\eta_{ij}}\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i^T\boldsymbol{\beta}))
\end{aligned}\tag{3.11}$$

Note that if the hazard function $h(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ is fully parametric, the likelihood of (3.11) can be directly maximized. However, for semiparametric basic hazard models, maximization of (3.11) can be computationally prohibitive, while the maximization of (3.10) based on the "complete data" are feasible. Since the structure of the "complete data" likelihood (3.10) contains the "missing data" of frailties μ_i, an EM-algorithm approach is proposed.

3.3.2 EM algorithm

The EM algorithm involves two levels of iteration. At the first level, θ is treated as a fixed value at $\tilde{\theta}$, and an EM algorithm is iterated to convergence to maximize (3.10) with respect to $(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})$ which leads to the GMLEs $(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\alpha}})|\tilde{\theta}$. The estimated unknown baseline function $\hat{\Phi}(\cdot)|\tilde{\theta}$ is obtained by a Breslow-type estimator. Substituting these estimators in the likelihood (3.10) gives us the profile likelihood for θ. Repeated evaluations of this profile likelihood provide us the GMLE for θ. Therefore, the parameter estimation procedure involves repeated assessment of the profile likelihood for θ and each assessment requires iteration of an EM algorithm to obtain GMLEs for $(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}, \Phi(\cdot))$.

For E-step of the EM algorithm, we take the logarithm of likelihood function (3.10) which results in a function of the latent frailty μ_i that still has a form of gamma distribution conditional on the observed data (Nielsen et al., 1992). Then the expectation of the latent frailties given the observed data are obtained as

$$E\left(\mu_i|\mathcal{F}\right) = \frac{1 + \theta N_{i\cdot}(y)}{1 + \theta \sum_{j=l_i+1}^{n_i} \int_0^y K_{ij}(s) dH_{ij}^*(s|\mathbf{X}_i; \boldsymbol{\beta})} \tag{3.12}$$

This is the key basis for the E-step where the frailties in the complete data log-likelihood are replaced by this conditional expectation. In particular, the basic cumulative hazard function $H_{ij}^*(s|\mathbf{X}_i; \boldsymbol{\beta})$ in this expectation involves with the unknown baseline function $\Phi(\cdot)$. Therefore, $\Phi(\cdot)$ is treated as nuisance parameters and needs to be estimated in the M-step as if the frailties μ_i were observed.

Assuming that θ is fixed at the value $\tilde{\theta}$ and given the initial values $(\hat{\boldsymbol{\beta}}^{(0)}, \hat{\gamma}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}, \hat{\Phi}(\cdot)^{(0)})$, the $(l+1)$-th iteration of the EM algorithm has a structure as follows.

Step0: Calculation of Basic Cumulative Hazard.

To obtain the expectation of the frailties, the basic cumulative hazard function needs to be calculated. Based on equation (3.9) and given the values of $(\hat{\boldsymbol{\beta}}^{(l)}, \hat{\gamma}^{(l)}, \hat{\Phi}(\cdot)^{(l)})$,

the estimator of $H^*_{ij}(s|\mathbf{X}_i;\boldsymbol{\beta})$ in the $l+1$-th iteration can be expressed as

$$\hat{H}^{*(l)}_{ij}(y|\mathbf{X}_i;\hat{\boldsymbol{\beta}}^{(l)}) = \int_0^y \exp\left\{\tilde{\theta}\hat{H}^{(l)}_{0,ij}(\hat{\Phi}^{(l)}_{ij}(s-)\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(l)}))\right\} d\hat{H}^{(l)}_{0,ij}(\hat{\Phi}^{(l)}_{ij}(s-)\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(l)}))$$

where $\hat{\Phi}(\cdot)$ is given in the M-step.

E-step: Posterior Expectation of the latent frailty μ_i.

In the E-step, equation (3.12) is evaluated under current parameter estimates as

$$\hat{\mu}_i^{(l+1)} = \frac{1+\tilde{\theta}N_{i\cdot}(\tau)}{1+\tilde{\theta}\sum_{j=l_i+1}^{n_i}\int_0^\tau K_{ij}(s)d\hat{H}^{*(l)}_{ij}(s|\mathbf{X}_i;\boldsymbol{\beta})}$$

M-step: Estimation of $(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})$ and $\Phi(\cdot)$.
The M-step is involved with maximization of the likelihood function (3.10) with respect to $(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})$ and at the same time the unknown baseline function $\Phi(\cdot)$ is estimated by a Breslow-type estimator by keeping $\boldsymbol{\beta}$ fixed.
M1: M-step for $(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})$

$$(\hat{\boldsymbol{\beta}},\hat{\gamma},\hat{\boldsymbol{\alpha}})^{(l+1)} = \arg\max_{(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})} \prod_{i=1}^m \prod_{j=l_i+1}^{n_i}\left[\frac{1-\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i^T\boldsymbol{\beta}))}{\exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i^T\boldsymbol{\beta}))}\right]^{\eta_{ij}} \exp(-\exp(\alpha(j)+\gamma\mathbf{X}_i^T\boldsymbol{\beta}))$$
$$\cdot \prod_{y\in[0,\tau]}\left(\frac{\hat{\mu}_i^{(l+1)}\exp\{\tilde{\theta}\hat{H}_{0,ij}(\hat{\Phi}_{ij}(y-)\exp(\mathbf{X}_i^T\boldsymbol{\beta}))\}d\hat{H}_{0,ij}(\hat{\Phi}_{ij}(y)\exp(\mathbf{X}_i^T\boldsymbol{\beta}))\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{\sum_{k=1}^m\sum_{l=1}^{t_k}\hat{\mu}_k^{(l+1)}\exp\{\tilde{\theta}\hat{H}_{0,kl}(\hat{\Phi}_{kl}(y-)\exp(\mathbf{X}_k\boldsymbol{\beta}))\}d\hat{H}_{0,kl}(\hat{\Phi}_{kl}(y)\exp(\mathbf{X}_k\boldsymbol{\beta}))\exp(\mathbf{X}_k\boldsymbol{\beta})K_{kl}(y)}\right)^{dN_{ij}(y)}$$

M2: Approximate M-step for $\hat{\Phi}(\cdot)$ given $\left(\hat{\boldsymbol{\beta}},\hat{\gamma},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\mu}}\right)^{(l+1)}$.

The unknown baseline function $\Phi(\cdot)$ can be obtained by solving the following step function inductively with starting value of $\Phi(0)=0$.

$$\hat{\Phi}^{(l+1)}(y) = \int_0^y \left[\sum_{i=1}^m\sum_{j=1}^{n_i}\hat{\mu}_i^{(l+1)}R\left(\hat{\Phi}^{(l+1)}(s-),\hat{\boldsymbol{\beta}}^{(l+1)}|\tilde{\theta},\mathbf{X}_i\right)K_{ij}(s)\right]^{-1} dN_{\cdot\cdot}(s)$$

where

$$R\left(\hat{\Phi}^{(l+1)}(s-),\hat{\boldsymbol{\beta}}^{(l+1)}|\tilde{\theta},\mathbf{X}_i\right) =$$
$$\exp\{\tilde{\theta}\hat{H}_{0,ij}(\hat{\Phi}^{(l+1)}(s-)\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(l+1)}))\}\hat{h}_{0,ij}(\hat{\Phi}^{(l+1)}(s)\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(l+1)}))\exp(\mathbf{X}_i\hat{\boldsymbol{\beta}}^{(l+1)}).$$

Note that the solution for $\hat{\Phi}(\cdot)$ in this M-step is an approximation. In the E-step, the expectation is estimated by replacing $dH^*_{ij}(\cdot)$ with its estimator $d\hat{H}^*_{ij}(\cdot)$ and by fixing θ at $\tilde{\theta}$. For the M-step, we first obtain the GMLEs for the parameters $(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})$ by maximizing the likelihood of (3.10) where the frailties μ_i's are substituted by

their estimates from the E-step and $H^*_{ij}(\cdot)$ and $dH^*_{ij}(\cdot)$ are replaced with $\hat{H}^*_{ij}(\cdot)$ and $d\hat{H}^*_{ij}(\cdot)$, respectively. Using the estimators for $(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})$, the estimator for the baseline function $\Phi(\cdot)$ is also updated. Therefore, the M-step is involved with maximization of likelihood function of (3.10) with respect to $(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})$ as well as the baseline function $\Phi(\cdot)$.

Step0, E-step, M-step1 and M-step2 are repeated until convergence is obtained. The values of $\boldsymbol{\beta}$ and $\Phi(\cdot)$ at convergence, denoted by $\hat{\boldsymbol{\beta}}|\tilde{\theta}$ and $\hat{\Phi}(\cdot)|\tilde{\theta}$, are used to maximize the profile likelihood of θ using one dimensional optimization method. Due to the complexity of the estimators $(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\alpha}}, \hat{\Phi}(\cdot))$, we propose to use a bootstrap procedure to derive the estimates of variance of the estimators as well as the confidence intervals.

3.4 Simulation Studies

To evaluate the performance of the joint modeling procedure, we conducted simulation studies in different settings. One thousand replicates were performed for each set. Each simulation sample consisted of m subjects ($m = 200$ and 400). First, random lengths were generated as a time-to-event process from the complementary log-log model with a constant baseline hazard $\lambda(n_i|\mathbf{X}_i^T\boldsymbol{\beta}; \gamma) = 1 - \exp(-\exp(\alpha + \gamma(\beta_1 X_{1i} + \beta_2 X_{2i})))$ where X_1 and X_2 represent dummy variables of a categorical variable with three levels. In addition, N_i's were subject to censorship by means of an independent censoring time that was simulated from a complementary log-log model with intercept only.

The vector of multiple measures $\mathbf{Y}_i$ for each subject was generated from the Clayton-Oakes model with a marginal transformation model, and the length of vector $\mathbf{Y}_i$ was equal to the time-to-event N_i. For the marginal model, we considered the class of logarithmic transformations of the form of $S(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta}) =$
$[1 + r\exp(q(y_{ij}) + \beta_1 X_{1i} + \beta_2 X_{2i})]^{-\frac{1}{r}}$ with $r = 0$ corresponding to the marginal pro-

portional hazards model and $r = 1$ corresponding to the marginal proportional odds model (Dabrowska and Doksum, 1988; Chen and Yu, 2012).

The parameters were chosen to be similar to these estimated in the MSSWOW study presented in Section 5. Specifically, the following values were set for the vector of parameters $(\beta_1, \beta_2, \gamma, \alpha) = (0.3, 0.5, 4, -4)$. Approximately, these values provided the mean of the observed random vector length equal to 5 with 65% of censoring, and the mean of multiple measurements equal to 24. The dependence parameter θ was set to be 0.5 and 3.0, corresponding to Kendall's tau of 0.2 and 0.6, respectively. Each data set was analyzed using the joint modeling procedure and the EM algorithm described in Section 2 and 3. Biases for β_1, β_2, γ and θ were evaluated from the simulations. Bootstrapping standard deviations of the parameters, average simulation standard errors, and 95% coverage probabilities were also calculated. The simulation results are summarized in Table 1.

The results show that the biases for estimating β_1, β_2 and γ are small, particularly when sample size is large. The largest biases (between 2% and 4%) are seen in the estimators of β_1, β_2 and γ in the scenario where the sample size is small with $m = 200$ and at the same time the dependence parameter is large for the multiple measurements with $\theta = 3.0$. There exists a slight negative bias for estimation of θ, which becomes smaller with decreasing of association and increasing of sample size. This negative bias has been observed previously with maximizing the profile likelihood (Nielsen et al., 1992; Glidden and Self, 1999). The bootstrap standard deviations agree well with the Monte Carlo simulation standard errors. When dependence parameter decreases and sample size increases, the difference between the two standard errors becomes negligible. The 95% coverage probability for the regression parameters (β_1, β_2) maintains near the nominal level in all cases. Most of 95% confidence intervals for γ have reasonable coverage probabilities except for a few cases when $r = 0$ and $\theta = 3.0$ (coverage probability for γ: 91%-93%). The coverage probability

for θ is slightly lower than 95% in the case where $r = 0$ and $\theta = 3.0$ (coverage probability for θ: 92%). A similar phenomenon has been reported in previous literature for transformation model (Chen and Yu, 2012). In addition, as sample size increases, the bias and standard errors of all the parameters decrease. Overall, the simulation study illustrates that the generalized maximum likelihood estimators based on the EM algorithm perform reasonably well even when within-subject correlation is relatively high and sample size is moderate.

3.5 Application to MSSWOW Study

MSSWOW was a prospective cohort study conducted from 1991 to 1994. Women who were between the ages of 19 and 40 and at risk for pregnancy (sexually active and not consistently using birth control) were eligible for the study and a total of 470 women were finally enrolled and the participants were followed with menstrual diaries and urine samples for up to one year until a clinical pregnancy or the end of the study. Women kept a daily record of menstrual bleeding and unprotected intercourse during follow-up and collected urine samples which were tested for pregnancy. Time-to-pregnancy (TTP) was recorded as the number of cycles for a woman taken to conceive. Menstrual cycle lengths (MCLs) were calculated from the first day of menstrual bleeding until the day before the next onset of menses. A total of 3689 MCLs were recorded, and 179 (38.1%) of the participants got pregnant by the end of the study.

This study has been originally designed to investigate the effect of Video Display Terminal on spontaneous abortion (Marcus, 1990), but it provided the opportunity to explore the possible roles of risk factors on reproductive health outcomes of MCLs and TTP. In previous work, menstrual cycle characteristics, including cycle length and bleeding length, were found to be associated with a woman's likelihood of becoming pregnant as well as pregnancy outcome (Small et al., 2006). Guo et al. (2006)

Table 3.1. Simulation Results for Fitting Joint Models and Estimating the Parameters Using the EM Algorithm with 1000 Replicates[1]

		$\theta = 0.5$				$\theta = 3.0$			
Scenarios	**True**	**Bias**	**SD**[2]	**SE**[3]	**CP**[4]	**Bias**	**SD**	**SE**	**CP**
$m = 200$									
r=0	$\beta_1(0.3)$	0.0083	0.1353	0.1399	0.95	0.0084	0.1906	0.2028	0.94
	$\beta_2(0.5)$	0.0038	0.1384	0.1431	0.95	0.0081	0.1942	0.2025	0.94
	$\gamma(4.0)$	-0.0175	0.8772	0.8278	0.91	-0.0284	0.6688	0.6457	0.96
	$\theta(0.5/3.0)$	-0.0162	0.0650	0.0687	0.92	-0.0708	0.2404	0.2918	0.93
r=0.5	$\beta_1(0.3)$	0.0097	0.0776	0.0718	0.97	-0.0075	0.0892	0.0862	0.95
	$\beta_2(0.5)$	0.0102	0.0778	0.0851	0.97	-0.0084	0.0828	0.0824	0.94
	$\gamma(4.0)$	-0.0091	0.6433	0.6822	0.94	-0.0324	0.5269	0.5148	0.93
	$\theta(0.5/3.0)$	-0.0061	0.0323	0.0354	0.91	-0.0208	0.0728	0.0802	0.94
r=1	$\beta_1(0.3)$	-0.0066	0.0747	0.0759	0.97	0.0105	0.3088	0.3151	0.94
	$\beta_2(0.5)$	-0.0097	0.0736	0.0724	0.96	0.0101	0.3144	0.3048	0.95
	$\gamma(4.0)$	0.0577	0.6441	0.6577	0.94	0.0365	0.8568	0.8294	0.94
	$\theta(0.5/3.0)$	0.0021	0.0447	0.0545	0.94	-0.0241	0.2460	0.2732	0.97
$m = 400$									
r=0	$\beta_1(0.3)$	0.0039	0.0945	0.0939	0.95	-0.0023	0.1329	0.1412	0.93
	$\beta_2(0.5)$	-0.0016	0.0984	0.1005	0.95	0.0015	0.1366	0.1342	0.95
	$\gamma(4.0)$	0.0276	0.5581	0.5360	0.93	0.0475	0.7970	0.7845	0.91
	$\theta(0.5/3.0)$	-0.0107	0.0447	0.0446	0.94	-0.0978	0.1993	0.2119	0.93
r=0.5	$\beta_1(0.3)$	-0.0017	0.0701	0.0775	0.97	-0.0056	0.0583	0.0616	0.96
	$\beta_2(0.5)$	0.0053	0.0770	0.0721	0.96	-0.0019	0.0764	0.0741	0.95
	$\gamma(4.0)$	-0.0359	0.6011	0.5882	0.95	0.0443	0.5334	0.5283	0.94
	$\theta(0.5/3.0)$	-0.0051	0.2808	0.2543	0.94	-0.0137	0.1560	0.1639	0.92
r=1	$\beta_1(0.3)$	-0.0033	0.2092	0.2121	0.95	-0.0059	0.2157	0.2321	0.93
	$\beta_2(0.5)$	0.0011	0.2160	0.2110	0.95	0.0061	0.2198	0.2137	0.95
	$\gamma(4.0)$	0.0584	0.8225	0.8393	0.97	0.0259	0.7110	0.6514	0.96
	$\theta(0.5/3.0)$	-0.0047	0.0650	0.0730	0.96	-0.0159	0.1865	0.2003	0.93

[1] Censoring percentage is about 65%.

[2] The standard deviation is based on the bootstrap sampling.

[3] This is the Monte Carlo standard error based on the simulations.

[4] CP stands for the 95% coverage probability.

demonstrated the inadequacy of normal distribution for MCLs and showed that MCLs were distributed differently among various sub-populations of different age groups, but the impact of age on TTP was not investigated. Similar to previous work, we categorize age into four groups (19-25, 26-30, 31-35 and 36-41) and aim to evaluate the influence of aging effects on both MCLs and TTP as well as to determine the relationship between MCLs and TTP.

Table 2 presents the results from the MSSWOW data analysis including the parameter estimates as well as their bootstrapped standard errors. With different marginal distributions for menstrual lengths (proportional hazards model and proportional odds model etc.), the results indicate that there exists a significant age effect on the distribution of MCLs. Specifically, the first three groups are not significantly different in terms of the distributions of MCLs and they are significantly different from the older age group 36-41. The size of the effects gets larger as women get younger in reference to the oldest age group (36-41). The estimated association parameter $\hat{\theta}$ (Kendall's tau $\hat{\tau}$=0.136,0.152, and 0.232 in different models, respectively) was found to be significantly greater than zero, indicating a modest correlation was observed among MCLs from the same woman.

The estimated scaling parameter $\hat{\gamma}$ is negative and is significantly different from zero in all three marginal models (r=0,0.5,1), implying the effect of MCLs is significantly associated with TTP through underlying aging effects. In other words, the risk of getting pregnant stochastically decreases with the increasing of the MCLs. This means that with the increasing of a woman's age, we would observe longer MCLs and lower probability of getting pregnant. Adjusting for the impact of MCLs, women between 19 and 25 have the highest chance of getting pregnant, followed by those between 26 and 30 and those between 31 and 35. The women in age group 36 to 41 have the lowest pregnancy rate. The status of unprotected sex is found to have significant influence on the higher risk of getting pregnant. (Although all eligible

women reported having unprotected sex in the previous three months at entry to the study, not all women had unprotected sex during each menstrual cycle.)

To evaluate the fitted models, we compare the Akaike Information Criterion (AIC) across the models, which shows that the joint model based on the Clayton-Oakes model with a marginal proportional odds model (AIC=23477.50) is preferable to the other two models (AIC=24484.71 for proportional hazards model and AIC=23850.31 for transformation model with $r = 0.5$).

In particular, one can expect that the impact of age on MCLs reduces over time when women reach a certain age. For the proportional odds model, the covariate effects are specified as a multiplicative factor on the baseline odds function (Bennett, 1983), which indicates that the difference in hazards by covariates diminishes over time. That is, the covariate effects diminish with time. This property of the proportional odds model seems intuitively reasonable compared to proportional hazards model for interpreting the age effects on menstrual lengths. Furthermore, we plotted the estimated odds of survival function of the log menstrual lengths based on the Kaplan-Meier estimator (See Figure 1) and it is shown that the curves of the odds of survival functions are approximately parallel to each other. This also demonstrates the appropriateness of the proportional odds model for MCLs, while other plots (not presented here) also agree with this conclusion.

3.6 Discussion

In this Chapter, we proposed a joint modeling framework for the analysis of multivariate random length data where the multiple measurements were modeled via a Clayton-Oakes model and the random length was a discrete survival time with a complementary log-log link hazard function. Particularly, we specified a general class of semiparametric transformational models for the marginal models in the Clayton-Oakes part of the joint model. Under the joint modeling assumptions, an approximate

Table 3.2. Estimation of Joint Model Based on the Clayton-Oakes Model and Complementary Log-log Model for MSSWOW Data (m=470)

Model	**Effects**[1]		**Estimates**	**SE**	**P-value**
Joint Models: $r = 0$	Age group 19-25	β_{11}	-0.386	0.040	<.001
	Age group 26-30	β_{12}	-0.349	0.034	<.001
	Age group 31-35	β_{13}	-0.244	0.041	0.011
	Age group 36-41	-	-	-	-
	Scaling parameter	γ	-1.414	0.282	<.001
	Unsafe Sex	β_2	0.079	0.011	<.001
Association among	Association	θ	0.315	0.015	<.001
MCL	Kendall's tau	τ	0.136	-	-
Joint Models: $r = 0.5$	Age group 19-25	β_{11}	-0.483	0.067	<.001
	Age group 26-30	β_{12}	-0.302	0.011	<.001
	Age group 31-35	β_{13}	-0.194	0.066	0.003
	Age group 36-41	-	-	-	-
	Scaling parameter	γ	-1.033	0.249	0.036
	Unsafe Sex	β_2	0.079	0.010	<.001
Association among	Association	θ	0.359	0.013	<.001
MCL	Kendall's tau	τ	0.152	-	-
Joint Models: $r = 1$	Age group 19-25	β_{11}	-0.979	0.112	<.001
	Age group 26-30	β_{12}	-0.674	0.088	<.001
	Age group 31-35	β_{13}	-0.377	0.086	<.001
	Age group 36-41	-	-	-	-
	Scaling parameter	γ	-0.496	0.220	0.024
	Unsafe Sex	β_2	0.079	0.010	<.001
Association among	Association	θ	0.623	0.055	<.001
MCL	Kendall's tau	τ	0.232	-	-

[1] The estimates for parameters associated with baseline hazard of TTP are given as 1) for model $r = 0$, ≤ 2: $\alpha_1 = -3.440(0.107)$, $3 \sim 8$: $\alpha_2 = -3.568(0.091)$, $= 9$: $\alpha_3 = -4.798(0.425)$, $10 \sim 12$: $\alpha_4 = -3.195(0.102)$, and ≥ 13: $\alpha_5 = -3.948(0.094)$; 2) for model $r = 0.5$, ≤ 2: $\alpha_1 = -3.610(0.161)$, $3 \sim 8$: $\alpha_2 = -3.733(0.158)$, $= 9$: $\alpha_3 = -4.955(0.268)$, $10 \sim 12$: $\alpha_4 = -3.370(0.155)$, and ≥ 13: $\alpha_5 = -4.104(0.130)$; and 3) for model $r = 1$, ≤ 2: $\alpha_1 = -3.305(0.548)$, $3 \sim 8$: $\alpha_2 = -3.435(0.539)$, $= 9$: $\alpha_3 = -4.674(0.730)$, $10 \sim 12$: $\alpha_4 = -3.065(0.538)$, and ≥ 13: $\alpha_5 = -3.835(0.494)$.

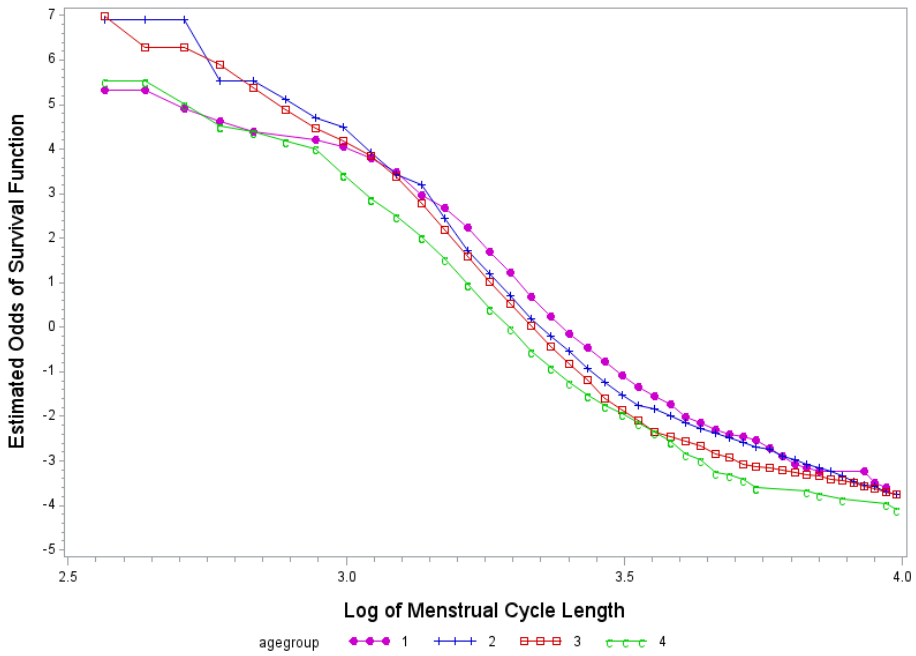

Figure 3.1. Plot of Estimated Log Odds of Survival Function of log(MCL) vs. log(MCL) (Survival functions are estimated by Kaplan-Meier estimators.)

EM algorithm based on gamma frailty model was developed to derive the likelihood inferences for parameters.

Our proposed method provide a flexible modeling framework to analyze two related outcomes jointly. First, our joint model can appropriately handle missing outcomes and censoring and truncation issues of the random vector length. Second, we use semiparametric transformation models for MCLs which include the commonly used proportional hazards model and the proportional odds model as special cases. Third, the estimation procedure has been developed to obtain generalized maximum likelihood estimators of all the model parameters including the regression coefficients in the joint model, dependence parameter and the unknown baseline function in the transformation model. AIC is used to assess the global goodness-of-fit for the joint model, however, more work is needed in determining the goodness-of-fit of these models.

MSSWOW study was conducted in a prospective manner and hence has several

advantages with regards to the analysis of TTP (Scheike and Keiding, 2006). Retrospectively collected data may be biased by the outcome, which is known by the woman when reporting TTP. In addition, it is more difficult to recall accurately whether unprotected intercourse occurred during each of the menstrual cycles. Because approximately 50% of pregnancies in the United States are unintended, the MSSWOW population is more likely to represent the population at risk of pregnancy than a retrospective study where women who may have terminated a pregnancy are excluded. For prospectively collected TTP data, TTP is observed as a waiting time that can be any positive number. Therefore, conditional on initiation time and covariates, TTP can be modeled via the conditional distribution as $\Pr\{N = n | N \geq n\}$. That is, the hazard of TTP given covariates can be directly observed. For retrospective study, we only observe the sample from a certain period $[0, T]$ instead of the entire timeline. In this case, TTP is both right and left-truncated conditional on initiation time. TTP from retrospective sample has a conditional distribution of $\Pr\{N = n | 0 \leq N \leq T\}$. Extensive details about the analysis of TTP using prospective and retrospective methods can be found in Scheike and Keiding's paper (2006). Our joint model proposed here is designed for the analysis of prospectively collected TTP data.

Chapter 4

A Two-Stage Estimation Approach

4.1 Introduction

Many medical and epidemiological studies involve with multivariate failure time data when an event can occur multiple times for a subject or the event times in a cluster are correlated. When the number of multiple measurements or the cluster size is an outcome of interest (a random variable), we observe multivariate random length data (Barnhart and Sampson, 1995; Barnhart et al., 1999). For these types of studies, there are two outcomes of interest: the multiple measurements and the random length. In addition, both of the outcomes may depend on common underlying covariates. For example, in the Mount Sinai Study of Women Office Workers (MSSWOW), multiple measurements of menstrual cycle lengths (MCLs) for each subject were collected and time-to-pregnancy (TTP) was calculated as the number of menstrual cycles taken to conceive (not including the conception cycle). It is natural to view MSSWOW data as a vector of multiple measures of MCLs with the random vector length equal to TTP. Both MCLs and TTP are important indicators of reproductive health. There is an extensive literature devoted to analyzing the separately but little literature on analyzing both outcomes simultaneously (Harlow and Zeger, 1991; Scheike and Jensen, 1997; Dunson et al., 2002; Guo et al., 2006; Small et al., 2006).

Barnhart et al. (1995; 1999) introduced a multiple population model to analyze

multivariate random length data where normality was assumed for the multiple measures. Motivated by MSSWOW study, Liu et al. (unpublished, 2014) proposed to use a semiparametric joint modeling approach where the multiple measurements are modeled via a special type of copula model known as the Clayton-Oakes model and the random length is assumed to follow a semiparametric discrete survival model. Their method provides a flexible modeling framework to analyze two correlated outcomes simultaneously and at the same time can handle missing, censoring and truncation issues. They specified semiparametric transformation models as the marginal distributions of the Clayton-Oakes model, which appropriately addresses the issue of right skewness of the distribution of menstrual cycle lengths. For parameter estimation of the joint model, Liu et al. (unpublished, 2014) developed an EM algorithm in the spirit of the estimation for gamma frailty model and the standard errors of the estimates were obtained via a bootstrapping procedure. This method is computationally intensive and sensitive to the form of the copula model.

In this manuscript, our goal is to provide a computationally simple approach for obtaining parameter estimates in the model proposed by Liu et al. (unpublished, 2014). The method can be implemented by using existing statistical software packages, and therefore it is feasible for epidemiologists to apply in practice. To this end, we propose a "two-stage" method for statistical inference for the joint model. In the first stage, estimators of the marginal parameters for the joint model are derived under the working-independence assumption, i.e., under working assumption of that repeated measurements are independent. In the second stage, the marginal parameters in the copula model are replaced by the estimators obtained in the first stage. Then a pseudo-likelihood approach is used for estimating the dependence parameter. This method is applicable to general copula models such as Clayton-Oakes models and positive stable models. Asymptotic theory for the estimators can be established as well.

Specifically, we provide a comprehensive analysis of MSSWOW data by considering a general form of copula models. Marginal models of the copula model for the repeated measurements (e.g., MCLs) are modeled via a semiparametric transformation model. In addition, we model the random length (e.g., TTP) using two different discrete models, including (a) complementary log-log link model and (b) proportional odds model. The dissertation is organized as follows. In Section 2, we describe the joint model for the multivariate random length data. Section 3 provides the "two-stage" estimation procedure for both marginal parameters and the association parameter. In Section 4, the performance of the proposed method is evaluated by a wide range of simulation studies. In Section 5, the proposed model and estimation approach are applied to the MSSWOW data.

4.2 Model Specifications

Suppose that we have a total of m subjects. Let i index the subject and j index the measurement for each subject. Let $\mathbf{Y}_i$ represent a vector of multiple measurements on a response where the vector has a random length of N_i, $i = 1, ..., m$. In addition, a p-dimensional covariate vector $\mathbf{X}_i$, such as age group, BMI, smoking status, and number of unprotected sex in MSSWOW study, is collected for each subject which can affect both the repeatedly measured response as well as the random length.

We propose a joint model for the multiple measurements and the random length, denoted by a vector as $(\mathbf{Y}_i, N_i)$, that has a form of

$$\begin{aligned} &\Pr(Y_{i1} > y_{i1}, ..., Y_{in_i} > y_{in_i}, N_i = n_i | \mathbf{X}_i) \\ &= C\left(S(y_{i1}|\mathbf{X}_i), ..., S(y_{in_i}|\mathbf{X}_i) | N_i = n_i\right) \Pr\{N_i = n_i | \mathbf{X}_i\}. \end{aligned} \tag{4.1}$$

where $\boldsymbol{\pi}$ denotes the vector of parameters to estimate. In the first part of the joint model, $C(\cdot)$ is a copula distribution function conditional on that $N_i = n_i$ and $S(y_{ij}|\mathbf{X}_i)$ denotes the marginal distribution of Y_{ij} given $\mathbf{X}_i$. The second part of the

model is a probability density function of the random length N_i.

4.2.1 Marginal models for repeated measurements

We consider a class of semiparametric transformation models as the marginal distributions for each Y_{ij} in the copula model. Under the transformation model assumption, Y_{ij} depends on the covariates $\mathbf{X}_i$ via an unknown function $q(\cdot)$ as

$$q(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = -\mathbf{X}_i\boldsymbol{\beta} + \varepsilon_{ij}, \tag{4.2}$$

where $q(\cdot)$ is a completely unspecified and strictly increasing function and ε_{ij} is a random error with a known distribution function denoted by F_ε. Let $S_\varepsilon = 1 - F_\varepsilon$ be the survivor function for ε. Then the marginal survival function of Y_{ij} given $\mathbf{X}_i$ can be written as

$$S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = \Pr(Y_{ij} > y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = S_\varepsilon(q(y_{ij}) + \mathbf{X}_i\boldsymbol{\beta}). \tag{4.3}$$

There are different choices for S_ε to provide different marginal models. A general form for S_ε can be written as the class of logarithmic transformations $S_\varepsilon(s) = [1 + r\exp(s)]^{-\frac{1}{r}}, r \geq 0$ (Dabrowska and Doksum, 1988; Chen and Yu, 2012), which include proportional hazards model and proportional odds model as two special cases. When $r = 0$, S_ε follows the extreme value distribution as $S_\varepsilon(s) = \exp(-\exp(s))$ and model (4.2) becomes the proportional hazards model. If $r = 1$, ε_{ij} has a standard logistic distribution of $S_\varepsilon(s) = 1/(1 + \exp(s))$, model (4.2) is the proportional odds model.

4.2.2 Different copula models

The joint distribution of the vector of multiple measurements $\mathbf{Y}_i$ is modeled via copula models. Conditional on $N_i = n_i$, we assume that $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ has a multivariate survival distribution that follows a copula model as

$$\Pr\{Y_{i1} > y_{i1}, ..., Y_{in_i} > y_{in_i}|N_i = n_i, \mathbf{X}_i;\boldsymbol{\beta},\theta\} = C\left(S(y_{i1}|\mathbf{X}_i;\boldsymbol{\beta}), ..., S(y_{in_i}|\mathbf{X}_i;\boldsymbol{\beta})|N_i = n_i,;\theta\right)$$

where $C(\cdot)$ and $S(Y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$ are defined as before. $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression parameters associated with $\mathbf{X}_i$; and θ depicts the within-subject dependence. Many copula families are available for constructing statistical models such as Gaussian copulas and Archimedean copulas. Particularly, we focus on Archimedean copulas, which has been explored by statisticians for analyzing multivariate survival data (Genest and Mackay, 1986). Archimedean copulas have a form of

$$\begin{aligned} &C\left(S(y_{i1}|\mathbf{X}_i;\boldsymbol{\beta}),...,S(y_{in_i}|\mathbf{X}_i;\boldsymbol{\beta})|N_i=n_i,;\theta\right) \\ &=\varphi^{-1}\{\varphi(S(y_{i1}|\mathbf{X}_i;\boldsymbol{\beta}))+\cdots+\varphi(S(y_{in_i}|\mathbf{X}_i;\boldsymbol{\beta}))|N_i=n_i,;\theta\} \end{aligned} \tag{4.4}$$

where $\varphi:[0,1]\to[0,+\infty]$ is called the generator of the copula which is a decreasing and convex function such that $\varphi(0)=\infty$ and $\varphi(1)=0$.

Examples of Archimedean copulas include Clayton copula, Gumbel copula, Frank copula, Gumbel-Hougaard copula etc. In this Chapter, we consider the Clayton-Oakes model from the Clayton copula family (Clayton, 1978; Oakes, 1982) and positive stable frailty model belonging to the Gumbel-Hougaard copula family.

Clayton-Oakes model:

For the Clayton-Oakes model, the generator φ is defined as the Laplace transformation of a gamma distribution (the Laplace transformation for a function $f(x)$ is defined as $\varphi(s)=\int_0^\infty e^{-sx}f(x)dx$, $x\geq 0$). Then the Clayton-Oakes model is obtained as

$$\begin{aligned} &C\left(S(y_{i1}|\mathbf{X}_i;\boldsymbol{\beta}),...,S(y_{in_i}|\mathbf{X}_i;\boldsymbol{\beta})|N_i=n_i,;\theta\right) \\ &=\left[S(y_{i1}|\mathbf{X}_i;\boldsymbol{\beta})^{-\theta}+\cdots+S(y_{in_i}|\mathbf{X}_i;\boldsymbol{\beta})^{-\theta}|N_i=n_i,;\theta\right]^{-\frac{1}{\theta}} \end{aligned}$$

where $\theta>0$ is the dependence parameter that depicts the correlation among Y_{ij}'s for the same subject. As θ approaches 0, the observations within the same subject become independent, and the joint survival function is simply the product of the marginal survival functions. When θ goes to $+\infty$, the joint survival function converges to its

upper Frechet bound and $S_J(\mathbf{y}_i|N_i = n_i, \mathbf{X}_i) = \min\{S(y_{ij}|\mathbf{X}_i), j = 1, ..., n_i\}$. The dependence parameter θ is related to Kendall's (1962) coefficient of concordance known as the Kendall's τ which can be expressed as $\tau = \theta/(\theta + 2)$.

Positive stable model:

When $\varphi(\cdot)$ is defined as the Laplace transformation of a positive stable distribution, the joint survival function takes the form of a positive stable frailty model as

$$\begin{aligned} &C\left(S(y_{i1}|\mathbf{X}_i;\boldsymbol{\beta}), ..., S(y_{in_i}|\mathbf{X}_i;\boldsymbol{\beta})|N_i = n_i, ;\theta\right) \\ &= \exp\left\{-\left[(-\log S(y_{i1}|\mathbf{X}_i;\boldsymbol{\beta}))^{\frac{1}{\theta}} + \cdots + (-\log S(y_{in_i}|\mathbf{X}_i;\boldsymbol{\beta}))^{\frac{1}{\theta}}\right]^{\theta}\right\} \end{aligned}$$

where θ is the dependence parameter and is also associated with Kendall's τ where $\tau = 1 - \theta$.

4.2.3 Discrete model for the random length

For random length N_i, we focus on the case where N_i is a discrete survival time in this dissertation, but the modeling approach can be easily extended to other discrete distributions. We assume that N_i follows a discrete distribution with a general form of

$$\Pr\{N_i = n_i|\mathbf{X}_i;\gamma,\boldsymbol{\beta},\boldsymbol{\alpha}\} = \nu(\alpha_{n_i} + \gamma\mathbf{X}_i\boldsymbol{\beta}),\ n_i = 0, 1, ..., M,$$

where $\nu(\cdot)$ is a probability density function, $M > 0$ is a known positive integer, and α_{n_i} is associated the baseline probability density. $\boldsymbol{\beta}$ is a vector of parameters that shared by the copula model and the discrete survival model. γ is a scaling parameter that evaluates the effects on the distribution of N_i due to the shift or change in the distribution of $\mathbf{Y}_i$ with respect to covariates $\mathbf{X}_i$. This means that the association between $\mathbf{Y}_i$ and N_i is induced by the covariates $\mathbf{X}_i$. In other words, the covariates $\mathbf{X}_i$ can affect the distribution of the repeated measurements $\mathbf{Y}_i$ and these effects are

modified by the scaling parameter γ to influence the distribution of the random length N_i. The parameterization also allows the two models contain different covariates.

Hazard rate is commonly used to describe the distribution of a survival time in literature. In particular, we use a complementary log-log (CLL) link function to model the hazard rate of N_i, denoted by $\lambda(j|\mathbf{X}_i;\gamma,\boldsymbol{\beta},\boldsymbol{\alpha})$ (Scheike and Jensen, 1997; Kalbfleish and Prentice, 2002). This CLL link model has an interpretation of a discrete version of the continuous time proportional hazards model which has the following form of

$$\lambda(j|\mathbf{X}_i;\gamma,\boldsymbol{\beta},\boldsymbol{\alpha}) = 1-\exp(-\exp(\alpha_j+\gamma\mathbf{X}_i\boldsymbol{\beta})). \tag{4.5}$$

In this model, α_j is associated with the baseline hazard when all $\mathbf{X}_i = 0$. The parameters $\boldsymbol{\beta}$ and γ are as defined as before.

An alternative method of modeling N_i is to use the proportional odds model (Chen et al., unpublished, 2014), which models the discrete log odds of the cumulative probability instead of discrete hazards. Compared to the traditional discrete relative risk model, the proportional odds model provides a different interpretation with respect to assessing the impact of risk factors on the survival time N_i. For MSSWOW data, the proportional hazards model is used to investigate the probability of getting pregnant at a certain cycle given that the subject has not been pregnant before then. Alternatively, the proportional odds model describes the time-to-pregnancy in terms of the probability of getting pregnant within a certain number of menstrual cycles. Under the proportional odds model assumption, the ratio of the hazard rates in different sub-populations or with different risk factors converges to unity as time goes to infinity (Bennett, 1983; Yang and Prentice, 1999; Chen et al., unpublished, 2014). We specify the discrete proportional odds model in following way

$$\text{logit}\left(\Pr\{N_i \le j|\mathbf{X}_i;\gamma,\boldsymbol{\beta},\boldsymbol{\xi}\}\right) = \xi_j+\gamma\mathbf{X}_i\boldsymbol{\beta}, j=1,...,J, \tag{4.6}$$

where ξ_j represents the baseline log odds function with regards to a subject having

an event less than or equal to j, and J is the maximum value of observed N_i. γ and $\boldsymbol{\beta}$ are defined similarly in the CLL model.

4.3 Parameter Estimation Procedure

We consider a two-stage approach to derive the parameter estimates for the joint model (4.1). In the first stage, we estimate the marginal parameters $(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}(\boldsymbol{\xi}))$ under the independence working assumption that the multiple observations from the same subjects are assumed to be independent. In the second stage, a pseudo likelihood for the dependence parameter θ is constructed by replacing the marginal parameters in the full likelihood by the estimators obtained in the first stage. Large sample properties for the two-stage estimators can be established by following the previous work (Spiekerman and Lin, 1998; Glidden, 1999; Chen et al., 2002; Lu, 2005).

4.3.1 First stage: estimation parameters under working independence assumption

Based on the form of the joint model, the likelihood is composed of two parts including the likelihood contribution from the multiple measurements given the random length and the other part from the random length.

First, let us consider the likelihood contribution from the multiple measurements $\mathbf{Y}_i$ given $N_i = n_i$. Under the assumption of working independence, the joint density function of $\mathbf{Y}_i$ for the i-th subject is simply the product of the marginal density functions. Therefore, we write the hazard of Y_{ij} given $\mathbf{X}_i$, denoted by $h(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$, as

$$h(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta}) = \phi(y_{ij})\exp(\mathbf{X}_i\boldsymbol{\beta})h_0(\exp(\mathbf{X}_i\boldsymbol{\beta})\Phi(y_{ij})), j = 1, ..., n_i, i = 1, ..., m \quad (4.7)$$

where $\Phi(y) = \exp(q(y))$ is a strictly increasing positive function with $\Phi(0) = 0$ and $\lim_{y\to\infty}\Phi(y) = \infty$ and $\phi(y) = \Phi'(y)$. $h_0(\cdot)$ is the hazard associated with $\exp(\varepsilon)$. Model (4.7) is taken as the basic model for the intensity of the associated counting

process. Let C_{ij}^Y be the censoring time and $\tilde{Y}_{ij}$ represent the observed value of Y_{ij} where $\tilde{Y}_{ij} = \min(Y_{ij}, C_{ij}^Y)$. Denote the censoring indicator by $\delta_{ij} = I(Y_{ij} \leq C_{ij}^Y)$. Assuming the working independence of Y_{ij} on the same subject, a total of $M(M = \sum_{i=1}^m n_i)$ i.i.d. counting processes, which represent the multiple measurements of survival times with independent censoring, are observed. In other words, we have an M-dimensional counting process of all subjects, denoted by $N_{ij}(y) = \delta_{ij} I(\tilde{Y}_{ij} \leq y), j = 1, ..., n_i; i = 1, ..., m$, with intensity $K_{ij}(y)h(y_{ij}|\mathbf{X}_i; \boldsymbol{\beta})$ where $K_{ij}(y) = I(\tilde{Y}_{ij} \geq y)$ is the at-risk process. Let $M_{ij}(y)$ denote the associated martingale process, then the martingale decomposition of $dN_{\cdot}(y)$ is

$$dN_{\cdot}(y) = \Gamma_0(y, \boldsymbol{\beta}, \Phi) d\Phi(y) + dM_{\cdot}(y),$$

where $\Gamma_0(y, \boldsymbol{\beta}, \Phi) = \sum_{i=1}^m \sum_{j=1}^{n_i} K_{ij}(y) \exp(\mathbf{X}_i\boldsymbol{\beta}) h_0(\exp(\mathbf{X}_i\boldsymbol{\beta})\Phi(y-))$. Based on this decomposition and keeping $\boldsymbol{\beta}$ fixed, $\Phi(y)$ can be estimated by a Breslow-type estimator

$$\hat{\Phi}(y, \boldsymbol{\beta}) = \int_0^y \frac{1}{\Gamma_0(u, \boldsymbol{\beta}, \hat{\Phi})} dN_{\cdot}(u) \tag{4.8}$$

where $\hat{\Phi}(y, \boldsymbol{\beta})$ can be solved iteratively by using the fact that $\hat{\Phi}(0, \boldsymbol{\beta}) = 0$.

By replacing $\Phi(y)$ with $\hat{\Phi}(y, \boldsymbol{\beta})$ and $d\Phi(y)$ with $d\hat{\Phi}(y, \boldsymbol{\beta})$, the likelihood contribution from $\mathbf{Y}_i$ for estimating $\boldsymbol{\beta}$ can be written as

$$\prod_{i=1}^m \prod_{j=1}^{n_i} \prod_{y \geq 0} \left[K_{ij}(y) \exp(\mathbf{X}_i\boldsymbol{\beta}) d\hat{\Phi}(y, \boldsymbol{\beta}) h_0(\exp(\mathbf{X}_i\boldsymbol{\beta})\hat{\Phi}(y-, \boldsymbol{\beta})) \right]^{dN_{ij}(y)}. \tag{4.9}$$

Second, we derive the likelihood function of the random length N_i. As a discrete time-to-event variable, the random length N_i involves left truncation and right censoring issues. Let L_i denote the left truncation variable and C_i^N be the right censoring time, which is assumed to be independent of N_i. Conditioning on that $N_i \geq L_i$, the observed data on the discrete time-to-event process N_i for the i-th subject consists of $(\tilde{N}_i, \Delta_i)$, where $\tilde{N}_i = \min(N_i, C_i^N)$, and $\Delta_i = I(N_i \leq C_i^N)$ is the censoring indicator.

Assuming that a subject enters the study after time l_i, i.e., the left truncation $L_i = l_i$, the likelihood contribution from the i-th subject for N_i is

$$\prod_{i=1}^{m} \Pr\{N_i = n_i | N_i > l_i, \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha}(\boldsymbol{\xi})\}]^{\Delta_i} \Pr\{N_i > n_i | N_i > l_i, \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha}(\boldsymbol{\xi})\}^{1-\Delta_i} \tag{4.10}$$

When N_i is assumed to follow a CLL (4.4), it is easy to calculate that

$$\Pr\{N_i = n_i | N_i > l_i, \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha}\} = \lambda(n_i | \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha}) \prod_{j=l_i+1}^{n_i-1} (1 - \lambda(j | \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha}));$$

$$\Pr\{N_i > n_i | N_i > l_i, \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha}\} = \prod_{j=l_i+1}^{n_i} (1 - \lambda(j | \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha})).$$

Given the left truncation $L_i = l_i$ and covariates $\mathbf{X}_i$, the likelihood contribution from N_i taking into account the left truncation and right censoring can be written as

$$\prod_{i=1}^{m} \prod_{j=l_i+1}^{n_i} \left(\frac{\lambda(j | \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha})}{1 - \lambda(j | \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha})} \right)^{\eta_{ij}} (1 - \lambda(j | \mathbf{X}_i; \gamma, \boldsymbol{\beta}, \boldsymbol{\alpha}))$$

where η_{ij} is the longitudinal censoring indicator for N_i. That is, $\eta_{ij} = 1$ if an event occurs at the the j-th time for the i-th individual, and $\eta_{ij} = 0$ otherwise.

In the scenario where the discrete model for N_i is defined as a proportional odds model (4.5), we can obtain that

$$\Pr\{N_i = n_i | \mathbf{X}_i; \gamma, \boldsymbol{\beta}\} = 1/(1 + \exp(-\xi_{n_i} - \gamma \mathbf{X}_i \boldsymbol{\beta})) - 1/(1 + \exp(-\xi_{n_i-1} - \gamma \mathbf{X}_i \boldsymbol{\beta}));$$

$$\Pr\{N_i > n_i | \mathbf{X}_i; \gamma, \boldsymbol{\beta}\} = 1 - 1/(1 + \exp(-\xi_{n_i-1} - \gamma \mathbf{X}_i \boldsymbol{\beta})).$$

where ξ_0 is defined as $-\infty$ and ξ_J is assumed to be ∞. Based on the two quantities above, the likelihood function of N_i for the discrete proportional odds model has a form of

$$\prod_{i=1}^{m} \left[\frac{1}{1 + \exp(-\xi_{n_i} - \gamma \mathbf{X}_i \boldsymbol{\beta})} - \frac{1}{1 + \exp(-\xi_{n_i-1} - \gamma \mathbf{X}_i \boldsymbol{\beta})} \right]^{\Delta_i} \left[1 - \frac{1}{1 + \exp(-\xi_{n_i-1} - \gamma \mathbf{X}_i \boldsymbol{\beta})} \right]^{1-\Delta_i}$$

Combining the likelihood contribution (4.9) from $\mathbf{Y}_i$ given $N_i = n_i$ and the like-

lihood contribution (4.10) from N_i yields the likelihood for the joint model

$$\prod_{i=1}^{m}\prod_{j=1}^{n_i}\prod_{y\geq 0}\left[K_{ij}(y)\exp(\mathbf{X}_i\boldsymbol{\beta})d\hat{\Phi}(y,\boldsymbol{\beta})h_0(\exp(\mathbf{X}_i\boldsymbol{\beta})\hat{\Phi}(y-,\boldsymbol{\beta}))\right]^{dN_{ij}(y)}$$
$$\cdot\prod_{i=1}^{m}\Pr\{N_i=n_i|N_i>l_i,\mathbf{X}_i;\gamma,\boldsymbol{\beta}\}^{\Delta_i}\Pr\{N_i>n_i|N_i>l_i,\mathbf{X}_i;\gamma,\boldsymbol{\beta}\}^{1-\Delta_i} \tag{4.11}$$

Estimating equations for the marginal parameters $(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}(\boldsymbol{\xi}))$ are derived by taking the first derivative of the logarithm of the likelihood function (4.11). Under the regularity conditions, the consistency and asymptotic normality of $(\hat{\boldsymbol{\beta}},\hat{\gamma},\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\xi}}))$ as well as the weak convergence of $\hat{\Phi}(y,\boldsymbol{\beta})$ can be established (See Appendix I).

4.3.2 Second stage: estimation of association parameter

In the second stage, $(\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}(\boldsymbol{\xi}))$ and $\phi(y_{ij})$ are replaced by $(\hat{\boldsymbol{\beta}},\hat{\gamma},\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\xi}}))$ and $\hat{\phi}(y_{ij},\hat{\boldsymbol{\beta}})$ obtained in the first stage in the full log-likelihood function, which produces the pseudo log-likelihood for estimating θ as

$$l_m(\theta)=c\left(\hat{S}(y_{i1}|\mathbf{X}_i;\hat{\boldsymbol{\beta}}),...,\hat{S}(y_{in_i}|\mathbf{X}_i;\hat{\boldsymbol{\beta}})|N_i=n_i,;\theta\right)$$

where $c(\cdot)$ is the density function associated with the copula function $C(\cdot)$ in model (4.1) and $\hat{S}(y_{ij}|\mathbf{X}_i;\hat{\boldsymbol{\beta}})$ is the estimated marginal survival function. For example, if we have a Clayton-Oakes model, the pseudo-likelihood for θ is

$$\begin{aligned}l_m(\theta)=&\sum_{i=1}^{m}\sum_{j=1}^{D_i}\log\left((j-1)\theta+1\right)-\sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta\delta_{ij}\log\left(\hat{S}(y_{ij}|\mathbf{X}_i;\hat{\boldsymbol{\beta}})\right)\\&-\sum_{i=1}^{m}\left(\frac{1}{\theta}+D_i\right)\log\left(\sum_{j=1}^{n_i}\hat{S}(y_{ij}|\mathbf{X}_i;\hat{\boldsymbol{\beta}})^{-\theta}-n_i+1\right)\end{aligned} \tag{4.12}$$

where $D_i=\sum_{j=1}^{n_i}\delta_{ij}$. When the multivariate survival model has a form of positive stable model, the pseudo-likelihood for estimating θ becomes

$$l_m(\theta)\propto\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}\left(\frac{1}{\theta}-1\right)\log\left(-\log S(y_{ij})\right)+\sum_{i=1}^{m}\left[D_i(\theta-1)\log S_i+\log J\left(D_i,S_i\right)-S_i^{\theta}\right] \tag{4.13}$$

where $D_i = \sum_{j=1}^{n_i} \delta_{ij}$, $S_i = \sum_{j=1}^{n_i} (-\log S(y_{ij}))^{\frac{1}{\theta}}$, and $J(D_i, S_i)$ is a function of D_i and S_i that has a form of

$$J(D_i, S_i) = \sum_{k=0}^{D_i-1} \Xi_{D_i,k} S_i^{-k\theta}.$$

Here $\Xi_{D_i,k}$ is a polynomial of degree k and can be calculated recursively by the following equations $\Xi_{D_i,0} = 1$, $\Xi_{D_i,k} = \Xi_{D_i-1,k} + \Xi_{D_i-1,k-1}[(D_i - 1)/\theta - (D_i - k)]; k = 1, ..., D_i - 2$, and $\Xi_{D_i,D_i-1} = \theta^{1-D_i}\Gamma(D_i - \theta)/\Gamma(1 - \theta)$.

The pseudo log-likelihood for $\theta$is continuous in θ and is defined at zero and for negative values close to zero. The estimator of θ is the root of the pseudo score equation, i.e., the derivative of (4.12) or (4.13) with respective to $\boldsymbol{\theta}$, which can be solved by Newton-Raphson method. The large sample properties of $\hat{\theta}$ can be derived based on the asymptotic results for $\hat{\boldsymbol{\beta}}$ (See Appendix II).

4.4 Simulation Studies

Simulation studies are carried out to evaluate the performance of the two-stage estimators. 500 simulations are conducted with different sample sizes (200,400,600) under joint model (4.6). We first generate the discrete random length variable N_i from the complementary-log-log model

$$\lambda(j|\mathbf{X}_i; \gamma, \boldsymbol{\beta}) = 1 - \exp[-\exp(\alpha_j + \gamma(\beta_1 X_{i1} + \beta_2 X_{i2}))] \tag{4.14}$$

where α_j represents the baseline hazard rate associated with j and X_{i1} and X_{i2} correspond to the first two levels of a categorical covariate that has three groups (the third group is set as the reference level). According to this model, N_i is assumed to follow a multinomial distribution that take values from 1 to J^* with probability $\mathbf{p} = (p_1, ..., p_{12}, p_{J^*})^T$ and $\sum_{j=1}^{J^*} p_j = 1$ where $J^* > 0$ is an unknown positive integer.

The p_j's are obtained as

$$\begin{aligned}
&\Pr\{N_i = 1|\mathbf{X}_i;\gamma,\boldsymbol{\beta}\} = \lambda(1|\mathbf{X}_i;\gamma,\boldsymbol{\beta});\\
&\Pr\{N_i = j|\mathbf{X}_i;\gamma,\boldsymbol{\beta}\} = \lambda(j|\mathbf{X}_i;\gamma,\boldsymbol{\beta})\prod_{k=1}^{j-1}(1-\lambda(j|\mathbf{X}_i;\gamma,\boldsymbol{\beta})), j = 2,...,J;\\
&\Pr\{N_i > J|\mathbf{X}_i;\gamma,\boldsymbol{\beta}\} = \prod_{j=1}^{J}\left(1-\lambda(j|\mathbf{X}_i;\gamma,\boldsymbol{\beta},\boldsymbol{\alpha})\right).
\end{aligned}$$

where J is the maximum value of observed N_i and $\lambda(j|\mathbf{X}_i)$ is defined in model (4.14). To reduce the dimension of parameters, we only assume four baseline parameters for the random length denoted by $(\alpha_1,\alpha_2,\alpha_3,\alpha_4)$. In other words, restrictions are posed for baseline parameters.

Alternatively, we generate the discrete random length N_i from the proportional odds model

$$\text{logit}\left(\Pr\{N_i \le j|\mathbf{X}_i;\gamma,\boldsymbol{\beta}\}\right) = \xi_j + \gamma(\beta_1 X_{i1} + \beta_2 X_{i2}), j = 1,...,J, \tag{4.15}$$

where J is the maximum value of observed N_i, ξ_j is associated with baseline log odds for random length equal to j, and X_{i1} and X_{i2} are defined as in model (4.14). Similarly, we assume that N_i is generated from a multinomial distribution that take values from 1 to $J^* \ge J$ with probability $\mathbf{p} = (p_1,...,p_J,...,p_{J^*})^T$ and $\sum_{j=1}^{J^*} p_j = 1$ where p_j the is probability of having an event at the j the length defined as

$$\begin{aligned}
&\Pr(N_i = 1|\mathbf{X}_i;\gamma,\boldsymbol{\beta}) = \frac{1}{1+\exp(-\xi_1-\gamma(\beta_1 X_{i1}+\beta_2 X_{i2}))};\\
&\Pr(N_i = j|\mathbf{X}_i;\gamma,\boldsymbol{\beta}) = \frac{1}{1+\exp(-\xi_j-\gamma(\beta_1 X_{i1}+\beta_2 X_{i2}))} - \frac{1}{1+\exp(-\xi_{j-1}-\gamma(\beta_1 X_{i1}+\beta_2 X_{i2}))}, j = 2,...,J-1;\\
&\Pr(N_i \ge J|\mathbf{X}_i;\gamma,\boldsymbol{\beta}) = 1 - \frac{1}{1+\exp(-\xi_{J-1}-\gamma(\beta_1 X_{i1}+\beta_2 X_{i2}))}.
\end{aligned}$$

As in the complementary-log-log model (4.14), we pose restrictions to the baseline parameters δ_j's so that only four baseline parameters need to be estimated. In addition, the vectors $(\alpha_1,\alpha_2,\alpha_3,\alpha_4)$ and $(\xi_1,\xi_2,\xi_3,\xi_4)$ are carefully chosen so that the underlying probability of having an event at each j in model (4.14) and model

(4.15) are close. In this way, we can legitimately compare the estimators from the two models.

Since the random length is also a discrete survival time, we assume an independent censoring random variable C_i for N_i, where C_i follows a discrete uniform distribution as Unif[a, b], where a and b are chosen to give censoring percentages for N_i. Here, we choose $a = 0$ and $b = 12$. That is, the maximum observation of N_i is 12 (i.e., $J = 12$).

Given $N_i = n_i$, the observations of the vector of multiple measurements $\mathbf{Y}_i$ are simulated from the copula model (4.1) with a marginal transformation model (4.2). First, we generate a vector of observed survival function $(S(y_{i1}|\mathbf{X}_i;\boldsymbol{\beta}), ..., S(y_{in_i}|\mathbf{X}_i;\boldsymbol{\beta}))^T$ from a copula model for the i-th subject and Y_{ij}'s from the same subject i will have a common association parameter θ. From model (4.2), the baseline survival function $S_0(y_{ij})$ can be expressed as a function of $S(y_{ij}|\mathbf{X}_i;\boldsymbol{\beta})$. If we assume the baseline survival follows a Weibull distribution, we can solve the equation to obtain the observed menstrual cycle lengths Y_{ij}.

Table 1-Table 3 summarize the results from the simulation for each case. Bias along with the associated percentage of bias, standard deviation and average standard error, as well as the 95% coverage probability for each parameter are presented. Each table corresponds to different sample sizes and two different marginal models (complementary-log-log model and proportional odds model) for random length N_i are compared.

The results show that biases of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(\boldsymbol{\xi})$ are relatively small. The bias of the scaling parameter γ and association parameter θ is slightly larger in the scenario where sample size is small ($m = 200$) and the association parameter is large ($\theta = 3.0$). When sample size increases, the estimates for all the parameters are very close to the true values with quite small percentages of bias. It can be seen that the estimated standard deviations agree well with the average standard errors calculated based on the simulations. In addition, the standard errors of the parameter estimators decrease

with the increasing of sample size. The 95% coverage probabilities also demonstrate that the two-stage method performs very well. Most of the coverage probabilities are controlled and close to 95% except for a few cases due to the relatively large bias in the parameter estimates. Also, it is worth noticing that the estimates of $\boldsymbol{\beta}$, γ and θ are similar under different model specifications. That is, when we compare different transformation model for $\mathbf{Y}_i$ or different discrete survival model for N_i, the estimation procedure can perform equally well.

To evaluate the efficiency loss due to using a working independence assumption, the parameter estimators from the two-stage method are compared with those from the EM algorithm proposed by Liu et al. (unpublished, 2014). Specifically, standard deviations are obtained for each method, denoted by SD_{EM} and SD_{TS}, and the relative efficiency is calculated as SD_{EM}/SD_{TS}. The results are displayed in Table 4. From Table 4, it can be seen that the estimates based on EM algorithm have smaller standard deviations than those from the two-stage approach, but the efficiency loss of the two-stage estimators is less than 5% compared to the EM algorithm. A similar phenomenon has been observed in previous literature (Glidden, 2000) where the random length is treated as a fixed number. Therefore, although we lose a small amount of efficiency, the computational burden is notably reduced by using the two-stage estimation procedure.

4.5 Application to MSSWOW Data

We apply the two-stage approach for the joint model with a marginal transformation model to the MSSWOW data. In this joint model, the repeatedly measured menstrual cycle lengths for each woman are assumed to follow a copula model (the Clayton-Oakes or the positive stable model) and TTP has a discrete distribution in terms of the complementary log-log model or the proportional odds model. For the copula model part, the marginal survival function has a form of model with the logarithmic

Table 4.1. Simulation Results for Two-Stage Methods-Clayton-Oakes model and $r = 0$

		$r = 0$ and CLL Model				$r = 0$ and PO Model			
True		**Bias(%)**	**SD**	**SE**	**CP**	**Bias(%)**	**SD**	**SE**	**CP**
$\theta = 0.5$ and $n = 200$									
γ	4.0	0.219(5.5)	0.724	0.719	0.92	0.155(3.9)	0.794	0.804	0.93
β_1	0.3	-0.001(0.4)	0.104	0.105	0.95	0.001(0.3)	0.104	0.106	0.96
β_2	0.5	0.006(1.2)	0.134	0.132	0.95	0.005(0.9)	0.125	0.131	0.94
θ	0.5	-0.023(4.6)	0.067	0.065	0.93	-0.018(3.7)	0.066	0.067	0.94
$\theta = 0.5$ and $n = 400$									
γ	4.0	0.086(2.2)	0.534	0.536	0.93	0.109(2.7)	0.537	0.534	0.93
β_1	0.3	0.002(0.6)	0.073	0.074	0.94	-0.004(1.4)	0.077	0.078	0.94
β_2	0.5	0.007(1.3)	0.093	0.096	0.96	0.004(0.8)	0.095	0.096	0.95
θ	0.5	-0.013(2.6)	0.049	0.051	0.94	-0.012(2.3)	0.049	0.050	0.94
$\theta = 0.5$ and $n = 600$									
γ	4.0	0.064(1.6)	0.404	0.418	0.95	0.038(1.0)	0.405	0.407	0.94
β_1	0.3	0.002(0.6)	0.061	0.060	0.96	0.003(0.8)	0.061	0.062	0.96
β_2	0.5	0.005(1.0)	0.079	0.079	0.95	0.006(1.1)	0.077	0.080	0.94
θ	0.5	-0.008(1.6)	0.040	0.042	0.94	-0.009(1.8)	0.042	0.042	0.93
$\theta = 3.0$ and $n = 200$									
γ	4.0	0.226(5.7)	0.733	0.738	0.94	0.207(5.2)	0.789	0.790	0.93
β_1	0.3	0.008(2.8)	0.153	0.148	0.95	0.011(3.7)	0.165	0.161	0.94
β_2	0.5	0.024(4.9)	0.186	0.185	0.94	0.026(5.1)	0.188	0.185	0.93
θ	3.0	-0.158(5.3)	0.227	0.249	0.93	-0.175(5.8)	0.275	0.275	0.91
$\theta = 3.0$ and $n = 400$									
γ	4.0	0.099(2.5)	0.531	0.528	0.94	0.099(2.5)	0.577	0.597	0.94
β_1	0.3	0.007(2.3)	0.101	0.105	0.94	0.002(0.7)	0.113	0.111	0.95
β_2	0.5	0.014(2.8)	0.132	0.135	0.95	0.011(2.2)	0.133	0.132	0.94
θ	3.0	-0.089(3.0)	0.206	0.202	0.93	-0.093(3.1)	0.195	0.194	0.93
$\theta = 3.0$ and $n = 600$									
γ	4.0	0.057(1.4)	0.495	0.475	0.94	0.054(1.3)	0.494	0.487	0.93
β_1	0.3	0.002(0.8)	0.081	0.083	0.95	-0.003(0.9)	0.090	0.094	0.95
β_2	0.5	-0.004(0.9)	0.107	0.109	0.95	0.005(1.0)	0.113	0.119	0.94
θ	3.0	-0.062(2.1)	0.171	0.168	0.94	-0.064(2.1)	0.165	0.169	0.93

Table 4.2. Simulation Results for Two-Stage Methods-Clayton-Oakes model $r = 0.5$

		$r = 0.5$ and CLL Model				$r = 0.5$ and PO Model			
True		**Bias(%)**	**SD**	**SE**	**CP**	**Bias(%)**	**SD**	**SE**	**CP**
$\theta = 0.5$ and $n = 200$									
γ	4.0	-0.085(2.1)	0.674	0.693	0.95	-0.141(3.5)	0.869	0.839	0.95
β_1	0.3	0.011(3.5)	0.108	0.111	0.96	0.009(2.9)	0.127	0.123	0.95
β_2	0.5	0.027(5.5)	0.134	0.135	0.94	0.025(4.9)	0.151	0.139	0.95
θ	0.5	-0.007(1.4)	0.074	0.072	0.94	-0.009(1.7)	0.070	0.068	0.96
$\theta = 0.5$ and $n = 400$									
γ	4.0	-0.087(2.2)	0.524	0.489	0.96	-0.126(3.2)	0.589	0.547	0.96
β_1	0.3	0.004(1.2)	0.076	0.071	0.96	0.006(2.1)	0.089	0.081	0.96
β_2	0.5	0.016(3.2)	0.093	0.088	0.95	0.012(2.4)	0.100	0.112	0.94
θ	0.5	0.004(0.7)	0.056	0.053	0.96	-0.011(2.1)	0.050	0.052	0.96
$\theta = 0.5$ and $n = 600$									
γ	4.0	-0.065(1.6)	0.421	0.443	0.95	-0.053(1.3)	0.505	0.510	0.95
β_1	0.3	0.006(1.8)	0.058	0.065	0.96	0.005(1.7)	0.065	0.070	0.96
β_2	0.5	0.013(2.5)	0.075	0.078	0.96	0.006(1.2)	0.083	0.084	0.96
θ	0.5	-0.003(0.6)	0.045	0.044	0.95	-0.006(1.2)	0.033	0.029	0.95
$\theta = 3.0$ and $n = 200$									
γ	4.0	-0.238(6.0)	0.852	0.853	0.94	-0.168(4.2)	1.062	1.074	0.95
β_1	0.3	0.023(7.7)	0.144	0.140	0.93	0.028(9.4)	0.179	0.183	0.96
β_2	0.5	0.047(9.3)	0.172	0.177	0.93	0.047(9.4)	0.210	0.243	0.93
θ	3.0	-0.204(6.8)	0.277	0.275	0.95	-0.133(4.4)	0.276	0.220	0.93
$\theta = 3.0$ and $n = 400$									
γ	4.0	-0.087(2.2)	0.639	0.648	0.94	-0.150(3.7)	0.756	0.757	0.94
β_1	0.3	0.011(3.8)	0.099	0.103	0.94	0.017(5.7)	0.121	0.124	0.93
β_2	0.5	0.024(4.9)	0.123	0.127	0.95	0.026(5.1)	0.143	0.136	0.94
θ	3.0	-0.120(3.9)	0.194	0.205	0.95	-0.139(4.6)	0.198	0.183	0.94
$\theta = 3.0$ and $n = 600$									
γ	4.0	-0.054(1.3)	0.567	0.508	0.93	-0.092(2.3)	0.563	0.558	0.94
β_1	0.3	0.008(2.5)	0.080	0.078	0.95	0.007(2.3)	0.096	0.097	0.93
β_2	0.5	0.015(3.0)	0.103	0.096	0.94	0.012(2.4)	0.112	0.105	0.94
θ	3.0	-0.083(2.8)	0.167	0.165	0.95	-0.098(3.3)	0.141	0.145	0.95

Table 4.3. Simulation Results for Two-Stage Methods-Clayton-Oakes model and $r = 1$

		$r = 1$ and CLL Model				$r = 1$ and PO Model			
True		**Bias(%)**	**SD**	**SE**	**CP**	**Bias(%)**	**SD**	**SE**	**CP**
$\theta = 0.5$ and $n = 200$									
γ	4.0	-0.080(2.0)	0.551	0.540	0.95	-0.170(4.2)	0.806	0.880	0.94
β_1	0.3	0.015(4.9)	0.110	0.107	0.95	0.023(7.7)	0.127	0.132	0.95
β_2	0.5	0.025(5.0)	0.123	0.124	0.94	0.041(8.3)	0.157	0.160	0.94
θ	0.5	-0.006(1.2)	0.074	0.078	0.96	-0.005(0.9)	0.077	0.072	0.95
$\theta = 0.5$ and $n = 400$									
γ	4.0	-0.069(1.7)	0.391	0.424	0.95	-0.122(3.1)	0.618	0.606	0.95
β_1	0.3	0.011(3.6)	0.069	0.079	0.96	0.013(4.4)	0.083	0.088	0.95
β_2	0.5	0.019(3.7)	0.082	0.094	0.95	0.027(5.3)	0.109	0.109	0.95
θ	0.5	-0.005(0.9)	0.054	0.056	0.95	-0.001(0.2)	0.054	0.054	0.96
$\theta = 0.5$ and $n = 600$									
γ	4.0	-0.021(0.5)	0.240	0.227	0.95	-0.071(1.8)	0.505	0.528	0.96
β_1	0.3	0.007(2.3)	0.036	0.038	0.96	0.008(2.6)	0.065	0.069	0.95
β_2	0.5	0.006(1.2)	0.063	0.065	0.95	0.014(2.8)	0.084	0.092	0.96
θ	0.5	-0.003(0.6)	0.049	0.044	0.96	-0.005(1.0)	0.042	0.044	0.94
$\theta = 3.0$ and $n = 200$									
γ	4.0	-0.176(4.4)	0.705	0.719	0.94	-0.282(7.0)	1.056	0.975	0.93
β_1	0.3	0.028(9.2)	0.148	0.156	0.94	0.026(8.6)	0.168	0.179	0.93
β_2	0.5	0.041(8.2)	0.177	0.186	0.96	0.033(6.7)	0.225	0.224	0.95
θ	3.0	-0.197(6.6)	0.289	0.284	0.93	-0.183(6.1)	0.284	0.275	0.92
$\theta = 3.0$ and $n = 400$									
γ	4.0	-0.093(2.3)	0.529	0.519	0.93	-0.209(5.2)	0.843	0.858	0.94
β_1	0.3	0.010(3.2)	0.099	0.092	0.95	0.023(7.7)	0.123	0.142	0.94
β_2	0.5	0.024(4.8)	0.121	0.115	0.95	0.028(5.6)	0.167	0.183	0.94
θ	3.0	-0.098(3.3)	0.210	0.199	0.95	-0.102(3.4)	0.205	0.207	0.95
$\theta = 3.0$ and $n = 600$									
γ	4.0	-0.044(1.1)	0.395	0.461	0.94	-0.146(3.7)	0.472	0.499	0.95
β_1	0.3	0.003(0.8)	0.075	0.085	0.95	0.010(3.3)	0.042	0.039	0.95
β_2	0.5	0.015(3.1)	0.090	0.101	0.96	0.013(2.6)	0.055	0.059	0.93
θ	3.0	-0.074(2.5)	0.174	0.164	0.95	-0.007(1.4)	0.162	0.174	0.95

Table 4.4. Relative Efficiency of the EM Algorithm and Two-Stage Method($n = 400$)

Models	**Parameters**	$\mathbf{SD}_{EM}$	$\mathbf{SD}_{TS}$	**RE**
r=0	$\gamma(4.0)$	0.523	0.546	0.958
	$\beta_1(0.3)$	0.095	0.097	0.979
	$\beta_2(0.5)$	0.098	0.102	0.961
	$\theta(0.5)$	0.051	0.053	0.962
r=0.5	$\gamma(4.0)$	0.578	0.591	0.978
	$\beta_1(0.3)$	0.077	0.079	0.975
	$\beta_2(0.5)$	0.084	0.088	0.954
	$\theta(0.5)$	0.052	0.054	0.963
r=1	$\gamma(4.0)$	0.842	0.866	0.972
	$\beta_1(0.3)$	0.197	0.207	0.952
	$\beta_2(0.5)$	0.210	0.218	0.963
	$\theta(0.5)$	0.062	0.064	0.969

transformations as $S_\varepsilon(s) = [1 + r\exp(s)]^{-\frac{1}{r}}, r \geq 0$ and we consider $r = 0, 0.5$ and 1, respectively.

The model fitting results are summarized in Table 5, Table 6 and Table 7. Each model indicates that there exists a significant age effect on the distribution of menstrual cycle lengths. The size of the effect gets larger as women get younger compared to the oldest age group (36-41). All of the covariates are significant with regards to TTP. The scaling parameter $\hat{\gamma}$ is significantly lower than zero in all three models, indicating that menstrual cycle lengths and TTP are significantly associated through underlying aging effects. This implies that with the increasing of a woman's age, we

observe longer menstrual cycle and lower probability of getting pregnant. Adjusting for the impact of menstrual cycle lengths, women between 19 and 25 have the highest risk of getting pregnant, followed by those between 26 and 30 and those between 31 and 35. The women in age group 36 to 41 have the lowest pregnancy rate. Smoking was also a significant factor associated with menstrual lengths and TTP. Specifically, being a smoker is likely to increase the menstrual cycle length and reduce the probability of getting pregnant. There is no evidence showing a significant impact of BMI on both outcomes. In addition, the frequency of unprotected sex during each cycle is quite influential on TTP. The estimated association parameter $\hat{\theta}$ was found to be significantly greater than zero in both Clayton-Oakes model and the positive stable model, implying a modest correlation was found among menstrual cycle lengths from the same woman.

By comparing the model-based and the Kaplan-Meier survival curves of Y_{ij} (see Appendix III), it can be shown that the joint model where the copula model part has a marginal proportional odds model fits the data better than other marginal models.

4.6 Discussion

This chapter proposed a flexible joint modeling approach for multivariate random length data and applied the two-stage estimation method to obtain parameter estimators. Our proposed method can appropriately handle data complications such censoring, truncation and missingness and at the same time can allow specifying different marginal models as well as copula models to accommodate different situations. Furthermore, we showed that the two-stage estimators are consistent and asymptotically normal and obtained the variance of the parameter estimators.

First, we specified a semiparametric linear transformation model for the multiple measures, which relaxed the parametric assumption for analyzing menstrual cycle lengths in MSSWOW study. Second, for the discrete survival time, a relative risk

Table 4.5. Analysis of MSSWOW Data-$r = 0$ (m=470)

Model	**Effects**		**Estimates**	**SE**	**P-Value**
CLL model for TTP	Age group 19-25	β_{11}	-0.502	0.104	$< .001$
	Age group 26-30	β_{12}	-0.361	0.094	$< .001$
	Age group 31-35	β_{13}	-0.263	0.091	0.004
	Age group 36-41	-	-	-	-
	BMI	β_2	-0.007	0.006	0.243
	Smoking: Yes	β_3	0.174	0.068	0.011
	Scaling parameter	γ	-1.130	0.516	0.028
	Unsafe Sex	β_4	0.079	0.011	$< .001$
Association among	Association-CO[3]	θ	0.338	0.034	$< .001$
cycle length	Kendall's tau-CO	τ	0.145	0.012	$< .001$
	Association-PS[3]	θ_1	0.797	0.058	$< .001$
	Kendall's tau-PS	τ_1	0.203	0.058	$< .001$
PO model for TTP	Age group 19-25	β_{11}	-0.495	0.105	$< .001$
	Age group 26-30	β_{12}	-0.358	0.098	$< .001$
	Age group 31-35	β_{13}	-0.267	0.091	0.003
	Age group 36-41	-	-	-	-
	BMI	β_2	-0.005	0.006	0.405
	Smoking Status	β_3	0.177	0.071	0.013
	Scaling parameter	γ	-0.774	0.305	0.011
	Unsafe Sex	β_4	0.191	0.011	$< .001$
Association among	Association-CO	θ	0.337	0.033	$< .001$
cycle length	Kendall's tau-CO	τ	0.144	0.012	$< .001$
	Association-PS	θ_1	0.793	0.058	$< .001$
	Kendall's tau-PS	τ_1	0.207	0.058	$< .001$

[1] The baseline parameters for the CLL model are α_1=-3.424(SE:0.185), α_2=-3.557(SE:0.137), α_3=-4.787(SE:0.521), α_4=-3.197(SE:0.181), α_5=-3.943(SE:0.119).

[2] The baseline parameters for the proportional odds model are ξ_1=-4.075(SE:0.485), ξ_2=-1.873(SE:0.355), ξ_3=-1.702(SE:0.363), ξ_4=-1.170(SE:0.361), ξ_5=-1.015(SE:0.339).

[3] CO stands for Clayton-Oakes model and PS stands for Positive Stable copula model.

Table 4.6. Analysis of MSSWOW Data-$r = 0.5$ (m=470)

Model	**Effects**		**Estimates**	**SE**	**P-Value**
Joint Model:	Age group 19-25	β_{11}	-0.651	0.106	$< .001$
CLL model for TTP	Age group 26-30	β_{12}	-0.411	0.066	$< .001$
	Age group 31-35	β_{13}	-0.278	0.075	$< .001$
	Age group 36-41	-	-	-	-
	BMI	β_2	-0.004	0.003	0.182
	Smoking: Yes	β_3	0.347	0.094	$< .001$
	Scaling parameter	γ	-0.792	0.293	0.007
	Unsafe Sex	β_4	0.078	0.011	$< .001$
Association among	Association-CO[3]	θ	0.357	0.073	$< .001$
cycle length	Kendall's tau-CO	τ	0.151	0.026	$< .001$
	Association-PS[3]	θ_1	0.798	0.056	$< .001$
	Kendall's tau-PS	τ_1	0.202	0.056	$< .001$
Joint model:	Age group 19-25	β_{11}	-0.640	0.108	$< .001$
PO model for TTP	Age group 26-30	β_{12}	-0.407	0.067	$< .001$
	Age group 31-35	β_{13}	-0.261	0.072	$< .001$
	Age group 36-41	-	-	-	-
	BMI	β_2	-0.003	0.003	0.317
	Smoking Status	β_3	0.339	0.098	0.001
	Scaling parameter	γ	-0.619	0.237	0.009
	Unsafe Sex	β_4	0.193	0.010	$< .001$
Association among	Association-CO	θ	0.357	0.067	$< .001$
cycle length	Kendall's tau-CO	τ	0.151	0.024	$< .001$
	Association-PS	θ_1	0.798	0.057	$< .001$
	Kendall's tau-PS	τ_1	0.202	0.057	$< .001$

[1] The baseline parameters for the CLL model are α_1=-3.281(SE:0.181), α_2=-3.415(SE:0.143), α_3=-4.645(SE:0.591), α_4=-3.051(SE:0.186), α_5=-3.787(SE:0.120).

[2] The baseline parameters for the proportional odds model are ξ_1=-3.922(SE:0.489), ξ_2=-1.796(SE:0.362), ξ_3=-1.642(SE:0.349), ξ_4=-1.152(SE:0.341), ξ_5=-0.913(SE:0.315).

[3] CO stands for Clayton-Oakes model and PS stands for Positive Stable copula model.

Table 4.7. Analysis of MSSWOW Data-$r = 1$ (m=470)

Model	**Effects**		**Estimates**	**SE**	**P-Value**
Joint model:	Age group 19-25	β_{11}	-0.989	0.133	$< .001$
CLL model for TTP	Age group 26-30	β_{12}	-0.671	0.100	$< .001$
	Age group 31-35	β_{13}	-0.501	0.092	$< .001$
	Age group 36-41	-	-	-	-
	BMI	β_2	-0.003	0.004	0.453
	Smoking: Yes	β_3	0.334	0.080	$< .001$
	Scaling parameter	γ	-0.482	0.209	0.021
	Unsafe Sex	β_4	0.079	0.011	$< .001$
Association among	Association-CO[3]	θ	0.367	0.091	$< .001$
cycle length	Kendall's tau-CO	τ	0.155	0.032	$< .001$
	Association-PS[3]	θ_1	0.801	0.056	$< .001$
	Kendall's tau-PS	τ_1	0.199	0.056	$< .001$
Joint model:	Age group 19-25	β_{11}	-0.953	0.137	$< .001$
PO model for TTP	Age group 26-30	β_{12}	-0.616	0.104	$< .001$
	Age group 31-35	β_{13}	-0.430	0.096	$< .001$
	Age group 36-41	-	-	-	-
	BMI	β_2	-0.007	0.006	0.243
	Smoking Status	β_3	0.344	0.079	$< .001$
	Scaling parameter	γ	-0.363	0.159	0.022
	Unsafe Sex	β_4	0.194	0.010	$< .001$
Association among	Association-CO	θ	0.368	0.073	$< .001$
cycle length	Kendall's tau-CO	τ	0.155	0.026	$< .001$
	Association-PS	θ_1	0.800	0.055	$< .001$
	Kendall's tau-PS	τ_1	0.200	0.055	$< .001$

[1] The baseline parameters for the CLL model are α_1=-3.357(SE:0.174), α_2=-3.491(SE:0.188), α_3=-4.719(SE:0.595), α_4=-3.121(SE:0.196), α_5=-3.853(SE:0.123).

[2] The baseline parameters for the proportional odds model are ξ_1=-3.972(SE:0.485), ξ_2=-1.727(SE:0.355), ξ_3=-1.657(SE:0.363), ξ_4=-1.221(SE:0.361), ξ_5=-0.936(SE:0.339).

[3] CO stands for Clayton-Oakes model and PS stands for Positive Stable copula model.

model and a proportional odds model were proposed, which illustrates that our joint model can be applied to different scenarios where the covariate effects on the hazard rates of the discrete time may vary. Third, the two-stage estimation procedure was developed to obtain the parameter estimators. Compared to the EM algorithm (Liu et al., unpublished, 2014), the two-stage approach is computationally simple and efficiency loss is small. In particular, a nice feature of the two-stage method is that the procedure can be easily implemented using existing software packages such as SAS and R. This feature would be attractive to many epidemiologists and medical scientists especially when one is interested in performing data analysis in cases where the data structure is similar to MSSWOW study.

Appendix I

Theorem 1: Under some regularity conditions, the limiting distribution of the parameter estimators $\hat{\boldsymbol{\pi}} = (\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\xi}}))$ is

$$\sqrt{(n)}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}_0) \xrightarrow{D} N\left(\mathbf{0}, \mathbf{A}^{-1}\boldsymbol{\Sigma}(\mathbf{A}^{-1})^T\right)$$

as $n \to +\infty$. $\boldsymbol{\pi}_0$ is the vector of true parameters. The matrices $\mathbf{A}$ and $\boldsymbol{\Sigma}$ are defined as below.

$$\begin{aligned}\mathbf{A} &= n^{-1} \cdot \mathbf{I}(\boldsymbol{\pi}) = n^{-1} \cdot \left[-\frac{\partial \mathbf{U}(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}}\right] \\ &= n^{-1} \cdot - \begin{bmatrix} \frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} & \frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \gamma} & \frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} \\ \frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \gamma} & \frac{\partial \mathbf{U}(\gamma)}{\partial \gamma} & \frac{\partial \mathbf{U}(\gamma)}{\partial \boldsymbol{\alpha}} \\ \frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} & \frac{\partial \mathbf{U}(\gamma)}{\partial \boldsymbol{\alpha}} & \frac{\partial \mathbf{U}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \end{bmatrix}\end{aligned}$$

where $\mathbf{U}(\boldsymbol{\beta})$, $\mathbf{U}(\gamma)$, and $\mathbf{U}(\boldsymbol{\alpha})$ are the score functions based on the likelihood (4.12) for $\boldsymbol{\beta}$, γ, and $\boldsymbol{\alpha}$, respectively. That is, $\hat{\boldsymbol{\beta}}$, $\hat{\gamma}$, and $\hat{\boldsymbol{\alpha}}$ are the roots of those score functions that satisfy $\mathbf{U}(\boldsymbol{\beta}) = 0$, $\mathbf{U}(\gamma) = 0$, and $\mathbf{U}(\boldsymbol{\alpha}) = 0$.

Let $\mathbf{A}_{11} = \partial \mathbf{U}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, $\mathbf{A}_{12} = \mathbf{A}_{21} = \partial \mathbf{U}(\boldsymbol{\beta})/\partial \gamma$, $\mathbf{A}_{13} = \mathbf{A}_{31} = \partial \mathbf{U}(\boldsymbol{\beta})/\partial \boldsymbol{\alpha}$, $\mathbf{A}_{22} = \partial \mathbf{U}(\gamma)/\partial \gamma$, $\mathbf{A}_{23} = \mathbf{A}_{32} = \partial \mathbf{U}(\gamma)/\partial \boldsymbol{\alpha}$, and $\mathbf{A}_{33} = \partial \mathbf{U}(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$. Then we can express these quantities as

$$\begin{aligned}\mathbf{A}_{11} = &\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \int_0^{\omega} [\mathbf{X}_i - u(y)]\, \mathbf{X}_i^T \partial h_{\varepsilon}(q(y) + \mathbf{X}_i\boldsymbol{\beta})/\partial \boldsymbol{\beta} K_{ij}(y) dq(y) \\ &+ \eta_{ij} \frac{\partial^2 \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta}^2 \cdot \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}) - (\partial \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta})^2}{\lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})^2} \\ &- (1 - \eta_{ij}) \frac{\partial^2 \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta}^2)(1 - \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})) + (\partial \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta})^2}{(1 - \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}))^2}\end{aligned}$$

$$\begin{aligned}&\mathbf{A}_{12} = \mathbf{A}_{21} = \\ &\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \eta_{ij} \frac{\partial^2 \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta} \partial \gamma \cdot \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}) - \partial \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta} \cdot \partial \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \gamma}{\lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})^2} \\ &- (1 - \eta_{ij}) \frac{\partial^2 \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta} \partial \gamma \cdot (1 - \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})) + \partial \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta} \cdot \partial \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \gamma}{(1 - \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}))^2}\end{aligned}$$

$$\mathbf{A}_{13} = \mathbf{A}_{31} =$$

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \eta_{ij} \frac{\partial^2 \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha} \cdot \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}) - \partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta} \cdot \partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})^2}$$

$$- (1-\eta_{ij}) \frac{\partial^2 \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}\partial\boldsymbol{\alpha} \cdot (1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})) + \partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta} \cdot \partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}}{(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}))^2}$$

$$\mathbf{A}_{22} = \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \eta_{ij}\eta_{ij} \frac{\partial^2 \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}^2 \cdot \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}) - (\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta})^2}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})^2}$$

$$- (1-\eta_{ij}) \frac{\partial^2 \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}^2)(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})) + (\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta})^2}{(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}))^2}$$

$$\mathbf{A}_{23} = \mathbf{A}_{32} =$$

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \eta_{ij} \frac{\partial^2 \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\gamma\partial\boldsymbol{\alpha} \cdot \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}) - \partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\gamma \cdot \partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})^2}$$

$$- (1-\eta_{ij}) \frac{\partial^2 \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\gamma\partial\boldsymbol{\alpha} \cdot (1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})) + \partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\gamma \cdot \partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}}{(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}))^2}$$

$$\mathbf{A}_{33} = \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \eta_{ij}\eta_{ij} \frac{\partial^2 \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}^2 \cdot \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}) - (\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha})^2}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})^2}$$

$$- (1-\eta_{ij}) \frac{\partial^2 \lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}^2)(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})) + (\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha})^2}{(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}))^2}$$

$$\boldsymbol{\Sigma} = \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \begin{bmatrix} \sum_{j=1}^{n_i} \int_0^{\omega} [\mathbf{X}_i - u(y)]\, dM_{ij}(y) + \eta_{ij} \frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})} - (1-\eta_{ij}) \frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}}{1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})} \\ \sum_{j=1}^{n_i} \eta_{ij} \frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\gamma}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})} - (1-\eta_{ij}) \frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\gamma}{1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})} \\ \sum_{j=1}^{n_i} \eta_{ij} \frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})} - (1-\eta_{ij}) \frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}}{1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})} \end{bmatrix}^{\otimes 2}$$

where

$$u(y) = \frac{\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{X}_i h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta} + q(y_{ij})\} K_{ij}(y) B(y, \tilde{y}_{ij})}{B_2(y)}$$

$$B_2(y) = \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta} + q(y_{ij})\} K_{ij}(y)$$

$$B(y, \tilde{y}_{ij}) = \exp\left\{ \int_{\tilde{y}_{ij}}^{y} \frac{\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta} + q(u)\}/\partial\boldsymbol{\beta} \cdot K_{ij}(u)}{B_2(u)} dq(u) \right\}$$

$\mathbf{A}$ and $\mathbf{\Sigma}$ can be consistently estimated by substituting $(\boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}, q(\cdot))$ with their estimators $(\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\boldsymbol{\alpha}}, \hat{q(\cdot)})$.

Proof:) Let $\hat{q}(\cdot) = \hat{q}(\cdot; \boldsymbol{\beta}_0)$ where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. We first need to show that $\hat{q}(\cdot)$ is consistent. Define $h^*(q(y)) = B(y, a)$ and $H^*(q(y)) = \int_b^y h^*(q(u))du$ where $a > 0$ and b are fixed finite numbers such that h^* and H^* are finite for all y. Following Lu (2005), we have

$$H^*\{\hat{q}(y)\} - H^*\{q_0(y)\} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\int_0^y \frac{h^*(q(u))}{B_2(u)} dM_{ij}(u; \boldsymbol{\beta}_0, q_0) + o_p(n^{-\frac{1}{2}})$$

$$\left.\frac{\partial \hat{q}(y, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = B(y) + o_p(1)$$

By Law of Large Numbers, we can derive that

$$\begin{aligned}
&\frac{1}{n}\cdot -\left.\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0, \gamma=\gamma_0, \boldsymbol{\alpha}=\boldsymbol{\alpha}_0, q=q_0} \\
&= -\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\int_0^{\omega} \mathbf{X}_i \frac{\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta} + q(y_{ij})\}/\partial \boldsymbol{\beta}}{h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta} + q(y_{ij})\}} K_{ij}(u) \frac{\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta} + q(y_{ij})\}}{\partial \boldsymbol{\beta}} \mathbf{X}_i^T dq_0(u) \\
&- \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\int_0^{\omega} C_1(u) K_{ij}(u) \frac{\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta} + q(y_{ij})\}}{\partial \boldsymbol{\beta}} \mathbf{X}_i^T dq_0(u) \\
&+ \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\int_0^{\omega} C_2(u) K_{ij}(u) \frac{\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta} + q(y_{ij})\}}{\partial \boldsymbol{\beta}} \mathbf{X}_i^T dq_0(u) \\
&+ \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \eta_{ij} \frac{\partial^2 \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta}^2 \cdot \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}) - (\partial \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta})^2}{\lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})^2} \\
&- \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} (1 - \eta_{ij}) \frac{\partial^2 \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta}^2 \cdot (1 - \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})) + (\partial \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha})/\partial \boldsymbol{\beta})^2}{(1 - \lambda(j|\mathbf{X}_i; \boldsymbol{\beta}, \gamma, \boldsymbol{\alpha}))^2} \\
&= \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i} C_{3ij} + o_p(1)
\end{aligned}$$

where

$$C_1(u) = \frac{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}\mathbf{X}_i Q\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}K_{ij}(u)B(u,\tilde{y}_{ij})}{B_2(u)}$$

$$C_2(u) = \frac{\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\delta_{ij}\mathbf{X}_i\cdot\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}/\partial\boldsymbol{\beta}\cdot K_{ij}(u)B(u,\tilde{y}_{ij})}{B_2(u)}$$

$$Q\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\} = \frac{h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}\cdot\partial h_\varepsilon^2\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}/\partial\boldsymbol{\beta}^2-(\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}/\partial\boldsymbol{\beta})^2}{h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}^2}$$

$$C_{3ij} = \int_0^\omega\left[\mathbf{X}_i\frac{\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}/\partial\boldsymbol{\beta}}{h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}}-(C_2(y)-C_1(y))\right]K_{ij}(y)\frac{\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}}{\partial\boldsymbol{\beta}}\mathbf{X}_i dq_0(u)$$
$$+\eta_{ij}\frac{\partial^2\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}^2\cdot\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})-(\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta})^2}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})^2}$$
$$-(1-\eta_{ij})\frac{\partial^2\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}^2\cdot(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}))+(\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta})^2}{(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}))^2}$$

Thus, using the fact that $u(y) = C_2(y) - C_1(y)$ and applying the Law of Large Numbers, we have that

$$\frac{1}{n}\cdot-\frac{\partial\mathbf{U}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\gamma=\gamma_0,\boldsymbol{\alpha}=\boldsymbol{\alpha}_0,q=q_0} = \mathbf{A}_{11}+o_p(1)$$

By straightforward calculation, we can obtain the following equation as

$$\frac{1}{n}\mathbf{U}(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}C_{4ij}+o_p(1)$$

where

$$\begin{aligned}C_{4ij} = &\int_0^\omega\mathbf{X}_i\frac{\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}/\partial\boldsymbol{\beta}}{h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q(y_{ij})\}}dM_{ij}(y,\boldsymbol{\beta}_0,q_0)\\&+\delta_{ij}\mathbf{X}_i\left[\frac{\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+\hat{q}(y_{ij})\}/\partial\boldsymbol{\beta}}{h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+\hat{q}(y_{ij})\}}-\frac{\partial h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q_0(y_{ij})\}/\partial\boldsymbol{\beta}}{h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q_0(y_{ij})\}}\right]\\&-\mathbf{X}_i\left[h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+\hat{q}(y_{ij})\}-h_\varepsilon\{\mathbf{X}_i\boldsymbol{\beta}+q_0(y_{ij})\}\right]\\&+\eta_{ij}\frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})}-(1-\eta_{ij})\frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}}{(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}))}\end{aligned}$$

Again, by Law of Large Numbers, we can show that

$$\frac{1}{n}\mathbf{U}(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{m}W_i+o_p(1)$$

where $W_i = \sum_{j=1}^{n_i}\int_0^\omega(\mathbf{X}_i-u(y))dM_{ij}(y)+\eta_{ij}\frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}}{\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})}-(1-\eta_{ij})\frac{\partial\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha})/\partial\boldsymbol{\beta}}{(1-\lambda(j|\mathbf{X}_i;\boldsymbol{\beta},\gamma,\boldsymbol{\alpha}))}$.

Similarly, we can show that the following equations are true.

$$\frac{1}{n}\cdot -\frac{\partial \mathbf{U}(\gamma)}{\partial \gamma}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\gamma=\gamma_0,\boldsymbol{\alpha}=\boldsymbol{\alpha}_0,q=q_0} = \mathbf{A}_{22}+o_p(1)$$

$$\frac{1}{n}\cdot -\frac{\partial \mathbf{U}(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\gamma=\gamma_0,\boldsymbol{\alpha}=\boldsymbol{\alpha}_0,q=q_0} = \mathbf{A}_{33}+o_p(1)$$

$$\frac{1}{n}\cdot -\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \gamma}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\gamma=\gamma_0,\boldsymbol{\alpha}=\boldsymbol{\alpha}_0,q=q_0} = \mathbf{A}_{12}(\mathbf{A}_{21})+o_p(1)$$

$$\frac{1}{n}\cdot -\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\alpha}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\gamma=\gamma_0,\boldsymbol{\alpha}=\boldsymbol{\alpha}_0,q=q_0} = \mathbf{A}_{13}(\mathbf{A}_{31})+o_p(1)$$

$$\frac{1}{n}\cdot -\frac{\partial \mathbf{U}(\gamma)}{\partial \boldsymbol{\alpha}}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0,\gamma=\gamma_0,\boldsymbol{\alpha}=\boldsymbol{\alpha}_0,q=q_0} = \mathbf{A}_{23}(\mathbf{A}_{32})+o_p(1)$$

Hence, if we apply the Central Limit Theorem for those summations of i.i.d. random vectors with zero mean and finite variance, it follows that

$$\sqrt{n}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbf{A}^{-1}\nu_{ij}$$

with a Taylor approximation. That is,

$$\sqrt{n}(\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}_0)\xrightarrow{D} N\left(\mathbf{0},\mathbf{A}^{-1}\boldsymbol{\Sigma}(\mathbf{A}^{-1})^T\right)$$

where $\hat{\boldsymbol{\pi}}=(\hat{\boldsymbol{\beta}},\hat{\gamma},\hat{\boldsymbol{\alpha}})$ and $\boldsymbol{\pi}_0=(\boldsymbol{\beta}_0,\gamma_0,\boldsymbol{\alpha}_0)$. The consistency of $\hat{\boldsymbol{\pi}}$ is therefore straightforward.

Appendix II

Theorem 2: Under some regularity conditions, the estimator $\hat{\theta}$ obtained by the 2nd stage is consistent and

$$\sqrt{n}\hat{\theta} - \theta_0) \xrightarrow{D} N\left(0, \sigma_\theta^2\right) \text{ as } n \to +\infty,$$

where $\sigma_\theta^2 = I^{-1}(\theta_0)\sigma_\psi^2 I^{-1}(\theta_0)$, which is defined in details below.

$$I(\theta_0) = \lim_{n\to+\infty} -\frac{1}{n}\frac{\partial^2 l(\theta, \boldsymbol{\beta}_0, q_0)}{\partial\theta^2}\bigg|_{\theta=\theta_0}$$

$$\begin{aligned}
\tilde{\Omega}(y) = \frac{1}{n}\Bigg\{ & \sum_{i=1}^{m}\sum_{j=1}^{n_i} \theta_0^{-1} \frac{\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\}}{\sum_j \exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\} - n_i + 1} \\
& - (\theta_0^{-1} + N_{i\cdot}(\omega)) \frac{\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\}(1 + \theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij})))}{\sum_j \exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\} - n_i + 1} \\
& + (1 + \theta_0 N_{i\cdot}(\omega)) \frac{\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\}\left[\sum_j H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta} + q_0(y_{ij}))\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\}\right]}{\left[\sum_j \exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\} - n_i + 1\right]^2} \\
& + N_{ij}(\omega)\Bigg\} h_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))K_{ij}(y)
\end{aligned}$$

$$\begin{aligned}
\psi_i(\theta_0) = & \sum_{j=1}^{n_i} N_{ij}(\omega) H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta} + q_0(y_{ij})) + \int_0^\omega \frac{N_{i\cdot}(u-)}{\theta_0 N_{i\cdot}(u-) + 1} dN_{i\cdot}(u) \\
& + \theta_0^{-2}\log\left[\sum_{j=1}^{n_i}\exp(\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta} + q_0(y_{ij}))) - n_i + 1\right] \\
& - (\theta_0^{-1} + N_{i\cdot}(\omega)) \frac{\sum_{j=1}^{n_i} H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\}}{\sum_j \exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0 + q(y_{ij}))\} - n_i + 1}
\end{aligned}$$

$$\tilde{F} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\Bigg\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\theta_0^{-1}\frac{\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0+q(y_{ij}))\}}{\sum_j \exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0+q(y_{ij}))\}-n_i+1}$$
$$-(\theta_0^{-1}+N_{i\cdot}(\omega))\frac{\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0+q(y_{ij}))\}(1+\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0+q(y_{ij})))}{\sum_j \exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0+q(y_{ij}))\}-n_i+1}$$
$$+(1+\theta_0 N_{i\cdot}(\omega))\frac{\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0+q(y_{ij}))\}\left[\sum_j H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}+q_0(y_{ij}))\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0+q(y_{ij}))\}\right]}{\left[\sum_j \exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0+q(y_{ij}))\}-n_i+1\right]^2}$$
$$+N_{ij}(\omega)\Bigg\}h_\varepsilon(\mathbf{X}_i\boldsymbol{\beta}_0+q(y_{ij}))\mathbf{X}_i$$

$$\sigma_\psi^2 = E(\Psi_1^2) = E\left(\psi_1(\theta_0)+\int_0^\omega \tilde{\Omega}(y)d\zeta_1(y)+\tilde{F}\mathbf{A}^{-1}W_i\right)$$

where $\Psi_i(\theta)=\psi_i(\theta)+\int_0^\omega \tilde{\Omega}(y)d\zeta_i(y)+\tilde{F}\mathbf{A}^{-1}W_i$ and $\zeta_i(y) = B(y)\mathbf{A}^{-1}W_i+\sum_{j=1}^{n_i}\frac{B(u,y)}{B_2(u)}dM_{ij}(y;\hat{\boldsymbol{\beta}},\hat{q})$. The quantities $B(y)$, $B_2(y)$, $\mathbf{A}^{-1}$ and W_i are as defined in Appendix I but with $(\boldsymbol{\beta},q(\cdot))$ being substituted by their estimators $(\hat{\boldsymbol{\beta}},\hat{q}(\cdot))$.

Proof:) Based on the pseudo log-likelihood function in (4.12), we can obtain the pseudo score function as

$$\hat{S}_n(\theta) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}N_{ij}(\omega)H_\varepsilon(\mathbf{X}_i\hat{\boldsymbol{\beta}}+q(y_{ij}))+\sum_{i=1}^{m}\int_0^\omega\frac{N_{i\cdot}(u-)}{\theta N_{i\cdot}(u-)+1}dN_{i\cdot}(u)$$
$$+\sum_{i=1}^{m}\theta^{-2}\log\left[\sum_{j=1}^{n_i}\exp\{\theta H_\varepsilon(\mathbf{X}_i\hat{\boldsymbol{\beta}}+q(y_{ij}))\}-n_i+1\right]$$
$$-\sum_{i=1}^{m}(\theta^{-1}+N_{i\cdot}(\omega))\frac{\sum_j H_\varepsilon(\mathbf{X}_i\hat{\boldsymbol{\beta}}+q(y_{ij}))\exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\hat{\boldsymbol{\beta}}+q(y_{ij}))\}}{\sum_j \exp\{\theta_0 H_\varepsilon(\mathbf{X}_i\hat{\boldsymbol{\beta}}+q(y_{ij}))\}-n_i+1}$$

By appling a von Mises expansion of $\frac{1}{n}\hat{S}_n(\theta)$ around the true values $(\boldsymbol{\beta}_0,q_0(\cdot))$, we have

$$\frac{1}{\sqrt{n}}\hat{S}_n(\theta_0) = \frac{1}{\sqrt{n}}S_n(\theta_0)+\int_0^\omega\tilde{\Omega}(y)d\sqrt{n}\left[\hat{q}(y,\hat{\boldsymbol{\beta}})-q_0(y)\right]+\tilde{F}\cdot\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)+o_p(1)$$

In Appendix I, we have shown that $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{m}\mathbf{A}^{-1}W_i+o_p(1)$ and that $\sqrt{n}\left(\hat{q}(y;\hat{\boldsymbol{\beta}})-q_0(y)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{m}\zeta_i(y)+o_p(1)$. Therefore,

$$\frac{1}{\sqrt{n}}\hat{S}_n(\theta_0)=\frac{1}{n}\sum_{i=1}^{m}\Psi_i(\theta_0)+o_p(1)$$

By Central Limit Theorem, the above result implies that

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right) = I^{-1}(\theta_0)\cdot\frac{1}{\sqrt{n}}\hat{S}_n(\theta_0) + o_p(1)$$

That is, $\sqrt{n}\left(\hat{\theta}-\theta_0\right) \xrightarrow{D} N\left(0,\sigma_\theta^2\right)$ as $n \to +\infty$ where $\sigma_\theta^2 = I^{-1}(\theta_0)\sigma_\psi^2 I^{-1}(\theta_0)$. The consistency of $\hat{\theta}$ follows.

Appendix III

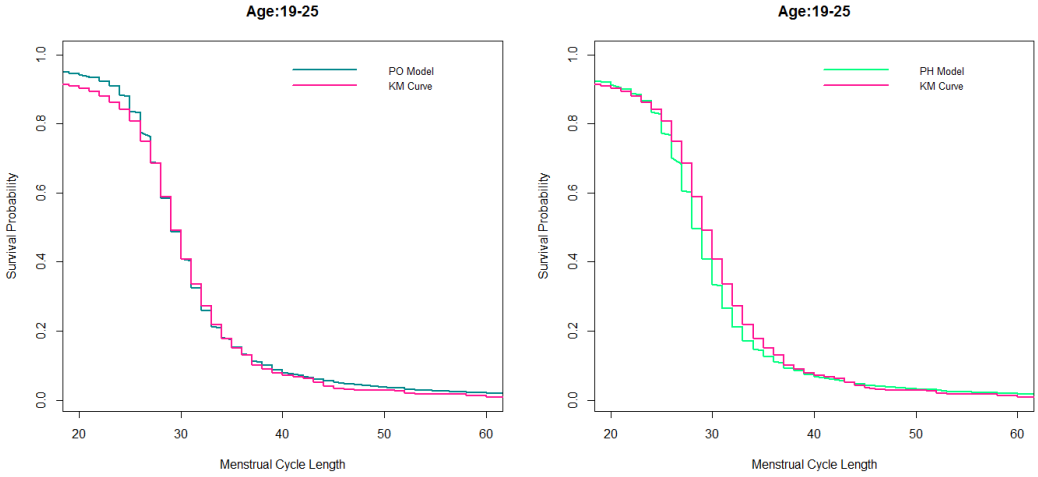

Figure 4.1. Plot of Estimated Log Odds of Survival Function vs. Log(Y)(Survival functions are estimated by Kaplan-Meier estimators.)

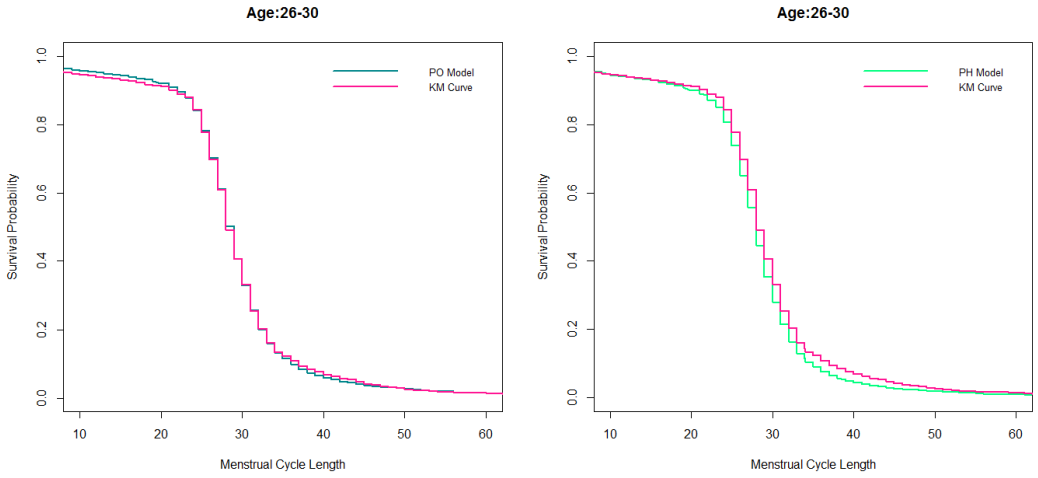

Figure 4.2. Plot of Estimated Log Odds of Survival Function vs. Log(Y)(Survival functions are estimated by Kaplan-Meier estimators.)

Chapter 5

Nonparametric Test for the Conditional Independence between a Biomarker and Time-to-Event Data

5.1 Introduction

Investigation of biomarkers that predict the onset of disease has become an important topic in medical research. Often, multiple measurements of a quantitative biomarker are available and testing the association between the biomarker and the onset or progression of disease is of interest. A typical biomarker analysis involves modeling the repeatedly measured biomarker in terms of a linear mixed-effects model (Laird and Ware, 1982; Lindstrom and Bates, 1990; Pearson et al., 1994; Morrell et al., 1995; Slate and Turnbull, 2000) and conditional on the random effects, a survival model for the time to event is assumed. The most commonly used model for the time-to-event process is the Cox proportional hazard model (DeGruttola and Tu, 1994; Hogan and Laird, 1997; Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Wang and Taylor, 2001; Xu and Zeger, 2001; Ibrahim et al., 2004; Chi and Ibrahim, 2006; Diggle et al., 2008; Rizopoulos and Ghosh, 2011). In some cases accelerated failure time model is also considered (Tseng et al., 2005). This type of analysis is generally referred as

a joint modeling of longitudinal and survival outcomes (Tsiatis and Davidian, 2004; Ibrahim et al., 2010; Rizopoulos, 2013; Sousa, 2011).

Model-based inference via an Expectation-Maximization (Dempster et al., 1977) algorithm has been proposed by many authors (Wulfsohn and Tsiatis, 1997; Henderson et al., 2000; Lin et al., 2002; Hsieh et al., 2006; Elashoff et al., 2008), and the corresponding score test is used to test the null hypothesis that there is no association between the biomarker and the onset of the disease. Jacqmin-Gadda et al. (2010) developed a score test based on a joint model with latent classes and shared random effects for testing the conditional independence of a longitudinal quantitative outcome and a time to event. The null hypothesis assumes that the biomarker and the time to event are independent given the latent classes while under the alternative hypothesis, random effects from the mixed model have significant influence on the time to event process. Concerns of the score test may arise with regards to the sensitivity to model misspecifications (Lagakos and Schoenfeld, 1984; Andersen et al., 1993; Li et al., 1996; DiRienzo and Lagakos, 2001). For example, Lagakos and Schoenfeld (1984) and Lagakos (1988) investigated the loss of efficiency of the score test for testing the regression coefficient in the proportional hazards model when the model is misspecified. Li et al. (1996) showed that the score test may perform poorly by inflating the size and power when the model assumption is violated and the sample size is small. DiRienzo and Lagakos (2001) suggested that the asymptotic distribution of the score test under null hypothesis is not centered at zero when the proportional hazard model is not correctly specified and the tests of treatment effects can be severely biased.

In this Chapter, our goal is to propose a nonparametric testing procedure that is not sensitive to the semiparametric assumption (e.g., the proportional hazards or accelerated failure time assumption) for the time to event model. In order to test the hypothesis whether the biomarker is associated with the time to event, one may not necessarily need to assume a semiparametric model such as the proportional hazards

model for the time-to-event process. Our proposed approach involves two stages. We first estimate the random effects from the biomarker model (Laird and Ware, 1982; Robinson, 1991). Next we propose to consider all possible choices of cutoff points of the predicted random effects by integrating the two-sample nonparametric test statistic over all of the possible dichotomizations. We develop a nonparametric test statistic that does not impose any model assumption on the survival data and at the same time can utilize all the information provided by the random effects. Peng and Fine (2008) introduced a nonparametric test statistic which can be considered a generalization of the log-rank test for testing the regression coefficient associated with a continuous covariate in a survival model. If the random effects are known, our test statistic is the same as the one suggested by Peng and Fine (2008).

To obtain the asymptotic distribution of the proposed test statistic, we write the dichotomized statistic in terms of independent and identically distributed (i.i.d.) quantities. Under certain assumptions, we are able to show the normality of the proposed test statistic under null and alternative hypotheses. In particular, we assume that the random effects that characterize the longitudinal process are normally distributed and this distribution is identical at all event times (Wulfsohn and Tsiatis, 1997). That is, we assume that whether a subject drops out from the study due to an event or censoring does not have influence on the distribution of random effects.

Specifically, this Chapter is organized as follows. In Section 2, we describe the general framework for jointly modeling a longitudinal outcome and a time-to-event process. In Section 3, we first introduce a model-based score test statistic under semiparametric model assumption. We formulate the testing problem of interest and the propose a nonparametric testing procedure in Section 4. The details of asymptotic properties of the test statistic are provided in the Appendix. In Section 5, simulation studies are conducted to assess the performance of the proposed nonparametric test including Type I error and power as well as comparison to model-based methods. We

apply the proposed method to a real data from an epidemiological study in Section 6. Conclusions and remarks are given in Section 6.

5.2 The Joint Modeling Framework

Suppose that we have n subjects. For each subject, two outcomes are collected including repeated measurements of a biomarker and time-to-event data. The repeated measures for the i-th subject are denoted by the vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{im_i})', i = 1, ..., n; j = 1, ..., m_i$, where m_i is the number of repeated measurements for the subject. We assume that Y_{ij} follows a linear mixed-effects model (Laird and Ware, 1982) as

$$Y_{ij} = \beta_0 + \mathbf{X}_{ij}\boldsymbol{\beta} + b_{0i} + \varepsilon_{ij} \tag{5.1}$$

where β_0 is the fixed intercept and $\boldsymbol{\beta}$ is the regression coefficients vector associated with fixed effects $\mathbf{X}_{ij}$. b_{0i} represents the random intercept for the i-th subject. The error term is denoted by ε_{ij}. Covariates in $\mathbf{X}_{ij}$ can be time dependent. For the linear mixed model (5.1), we assume that (i) b_{0i} is normally distribution as $N(0, \sigma_b^2)$; (ii) ε_{ij} has a normal distribution as $N(0, \sigma_e^2)$; (iii) $\varepsilon_{i'j'}$ is independent of each other, denoted by $\varepsilon_{i'j'} \perp \varepsilon_{ij}$ for any i, j, i', j' (hereafter, the '$\perp$' is used to indicate that two variables are independent); and (iv) $b_{0i} \perp \varepsilon_{ij}$, for any i, j.

In addition to the observations for the biomarker, a time-to-event process is also observed for each subject, denoted by T_i, $i = 1, ..., n$. Let $Q(t)$ be the survival function of T_i without conditioning on the random effects. Our objective is to determine whether $\mathbf{Y}_i$ and T_i are independent conditional on the random effects b_{0i}, i.e., whether $Q(t)$ depends on b_{0i}. If the form of the model for T_i and b_{0i} is completely unspecified, the null hypothesis for the independence of $\mathbf{Y}_i$ and T_i conditional on b_{0i} is represented by

$$H_0 : Q(t|b_{0i} \in B) = Q(t), \forall B \in \mathbb{B} \tag{5.2}$$

where $Q(t|b_{0i} \in B)$ denotes the survival function conditional on the random effects and $\mathbb{B}$ is the support of b_{0i}. If H_0 is true, the survival function of T_i is not affected by b_{0i}, and therefore the biomarker and time to event are not associated.

In the literature, a typical assumption for time to event is the proportional hazards model. The relationship between the repeated measures and the time to event can be specified via shared random effects as $\lambda(t_i|b_{0i}, \mathbf{Y}_i) = \lambda_0(t_i)\exp(\eta b_{0i})$ where $\lambda(\cdot)$ is the hazard at t_i, $\lambda_0(\cdot)$ is the baseline hazard function, and b_{0i} is the shared random effects for both repeated measurements of biomarker and time-to-event process. Another commonly used model assumption for time to event is the accelerated failure time model that has a form of $\log T_i = \eta b_{0i} + \epsilon_i$ where ϵ_i is the random error with an unspecified distribution. However, there are many cases where these semiparametric assumptions do not hold and the corresponding model-based score test is not valid. In Section 4, we will use simulation studies to show that our proposed non-parametric testing procedure is more robust than the score test against the model misspecifications.

Note that in this manuscript, our goal is to describe the relationship between the risk of the event and the change of biomarker over time preceding the event, and therefore, we only focus on the case where measurements of the biomarker after the event are excluded. That is, we do not consider the scenarios where T_i is a time to dropout (Jacqmin-Gadda et al., 2010). In addition, we assume that the censoring and the random effects are independent, which is true in many cases of clinical trials.

5.3 Model Based Score Test

As described in the last section, we can assume that the time to event T_i follows some parametric or semiparametric models. For example, a commonly used semiparametric

survival model is Cox proportional hazards model that has a form of

$$\lambda(t_i|b_{0i}) = \lambda_0(t_i)\exp(\eta b_{0i})$$

Under the assumption of shared random effects joint model, the null and alternative hypotheses for the conditional independence of the longitudinal process and the time to event given the random effects are represented by

$$H_0 : \eta = 0 \text{ vs. } H_a : \eta \neq 0.$$

A score test can be derived for testing the above null hypothesis H_0. First, the likelihood function for $(\mathbf{Y}_i, T_i)$ from all the subjects is written as

$$L(\eta, \beta; \mathbf{Y}_i, T_i) = \prod_{i=1}^{n} \int \left[\prod_{j=1}^{m_i} f(y_{ij}|b_{0i})\right] \lambda(t_i|b_{0i})^{\delta_i} S(t_i|b_{0i}) f(b_{0i}) db_{0i}$$

where $\lambda(t_i|b_{0i}) = \lambda_0(t_i)\exp(\eta_0 b_{0i})$ and $S(t_i|b_{0i})$ is the corresponding survival function for T_i conditional on b_{0i} and δ_i is the censoring indicator. Then the score function can be obtained by taking the derivative of the logarithm of the likelihood function, denoted by $U(\eta, \beta) = \frac{\partial \log L(\eta,\beta;\mathbf{Y}_i,T_i)}{\partial \eta} = \sum_{i=1}^{n} \frac{1}{L_i}\frac{\partial L_i}{\partial \eta}$. By simple calculations, we can obtain that

$$\frac{\partial L_i(\eta, \beta; \mathbf{Y}_i, T_i)}{\partial \eta} = \int \left[\prod_{j=1}^{m_i} f(y_{ij}|b_{0i})\right] \lambda(t_i|b_{0i})^{\delta_i} S(t_i|b_{0i}) \left[\delta_i - \Lambda(t_i|b_{0i})\right] b_{0i} f(b_{0i}) db_{0i}$$

where $\Lambda(t_i|b_{0i})$ is the cumulative hazard function associated with T_i conditional on b_{0i}. Finally, the score test statistic is derived by evaluating the score function $U(\eta, \beta)$ at $\eta = 0$ as

$$\begin{aligned} U(0, \beta) &= \sum_{i=1}^{n} \frac{1}{\int [\prod_{j=1}^{m_i} f(y_{ij}|b_{0i})] f(b_{0i}) db_{0i}} [\delta_i - \Lambda(t_i)] \int \left[\prod_{j=1}^{m_i} f(y_{ij}|b_{0i})\right] f(b_{0i}) b_{0i} db_{0i} \\ &= \sum_{i=1}^{n} [\delta_i - \Lambda(t_i)] \cdot E(b_{0i}|\mathbf{Y}_i) \end{aligned}$$

where $\Lambda(t_i)$ is the cumulative hazard function of T_i when $\eta = 0$ and $E(b_{0i}|\mathbf{Y}_i)$ is the posterior expectation of b_{0i} given the observed longitudinal data. The asymptotic

variance of the score test statistic is calculated as the second derivative of the log likelihood function. To perform the testing procedure, the score test statistic and its variance are calculated by replacing β by $\hat{\beta}$ and $\Lambda(t_i)$ by $\hat{\Lambda}(t_i)$ obtained under H_0.

5.4 Nonparametric Testing Procedure

We propose to construct a nonparametric test statistic for the null hypothesis (5.2) without imposing a regression model assumption for $Q(t)$ and b_{0i}. The rationale behind our proposed method is as follows. If the null hypothesis is true, then for any given cutoff point b of the random effects b_{0i}, the survival distribution conditional on $b_{0i} \geq b$ and $b_{0i} < b$ should be equal, denoted by $Q_+(t|b)$ and $Q_-(t|b)$, respectively. Testing the null hypothesis is equivalent to assessing the difference between $Q_+(t|b)$ and $Q_-(t|b)$ uniformly over the support of b.

5.4.1 Derivation of the nonparametric test statistic

Since $Q(t|b)$ and the random effects b_{0i} can not be observed, the test statistic is established based on the estimator of $Q(t|b)$. Let b_{0i}^E denote the best linear unbiased predictor (BLUP) of the random effects that is obtained using the Bayesian approach (Henderson, 1975; Robinson, 1991) where the superscript E indicates that the empirical predicted random effects are used. Define $\bar{Y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} Y_{ij}$ and $\bar{\mathbf{X}}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} \mathbf{X}_{ij}$. An empirical BLUP for b_{0i}^E has a form of

$$b_{0i}^E = \frac{m_i\hat{\sigma}_b^2}{m_i\hat{\sigma}_b^2 + \hat{\sigma}_e^2}(\bar{Y}_i - \hat{\beta}_0 - \bar{\mathbf{X}}_i\hat{\boldsymbol{\beta}}) \tag{5.3}$$

where $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\sigma}_b^2, \hat{\sigma}_e^2)$ are consistent estimators for $(\beta_0, \boldsymbol{\beta}, \sigma_b^2, \sigma_e^2)$.

Let $\hat{Q}_+^E(t|b)$ and $\hat{Q}_-^E(t|b)$ represent the nonparametric estimators (i.e., Kaplan-Meier estimator) for the conditional survival function based on that $b_{0i}^E \geq b$ and $b_{0i}^E < b$. Instead of choosing an arbitrary cutoff point, we propose to integrate across all possible cutoff points of b_{0i} in order to combine all the information provided by

the random effects. Therefore, the test statistic can be expressed as

$$T_n^E = \sqrt{n} \int_{\hat{b}_L}^{\hat{b}_U} \int_{t_L}^{t_U} \hat{w}^E(t,b) \left\{ \hat{Q}_+^E(t|b) - \hat{Q}_-^E(t|b) \right\} dtdb. \tag{5.4}$$

where $\hat{b}_U$ and $\hat{b}_L$ are the empirical upper and lower limits of b_{0i}^E and the quantities in the formula are defined as $\hat{w}^E(t,b) = \frac{\hat{C}_+^E(t|b)\hat{C}_-^E(t|b)}{\hat{p}_+^E\hat{C}_+^E(t|b)+\hat{p}_-^E\hat{C}_-^E(t|b)}$, $\hat{Q}_+^E(t|b) = \hat{Q}(t|b_{0i}^E \geq b)$, $\hat{Q}_-^E(t|b) = \hat{Q}(t|b_{0i}^E < b)$, $\hat{p}_+^E = \Pr\{b_{0i}^E \geq b\}$, and $\hat{p}_-^E = \Pr\{b_{0i}^E < b\}$. In particular, $\hat{p}_+^E$ and $\hat{p}_-^E$ are the empirical proportions of the stratification of $b_{0i}^E \geq b$ and $b_{0i}^E < b$. $\hat{C}_+^E(t|b)$ and $\hat{C}_-^E(t|b)$ are the Kaplan-Meier estimators of the censoring distribution based on the $b_{0i}^E \geq b$ and $b_{0i}^E < b$. Formula (5.4) shows that the test statistic is calculated as the sum of the difference between the survival functions at each stratification of $b_{0i}^E \geq b$ and $b_{0i}^E < b$ weighted by a non-negative weight function. When random effects b_{0i} are known, the test statistic of (5.4) will be reduced to the one proposed by Peng and Fine (2008).

5.4.2 Asymptotic property of the nonparametric test statistic

In this section, we study the asymptotic properties of the proposed test statistic T_n^E defined in (5.4). The main idea is provided as follows. First, it can be shown that the BLUP of b_{0i}^E defined in formula (5.3) is asymptotically equivalent to $\upsilon_i(b_{0i} + \bar{\varepsilon}_i)$ where $\upsilon_i = \frac{m_i\sigma_b^2}{m_i\sigma_b^2+\sigma_e^2}$ is a constant and $\bar{\varepsilon}_i = \sum_{j=1}^{m_i} \varepsilon_{ij}$ is a random variable that is the average of the random errors for each subject in the mixed model (5.1). Let $\tilde{Q}(t|b)$ represent the Kaplan-Meier estimator based on $\upsilon_i(b_{0i} + \bar{\varepsilon}_i)$ and $\hat{Q}^E(t|b)$ denote the plug-in Kaplan-Meier estimator based on b_{0i}^E. We can show that the difference between $\hat{Q}^E(t|b)$ and $\tilde{Q}(t|b)$ is asymptotically equal to zero. Next, we assume that the random error in the mixed-effects model is independent of the time-to-event process, i.e., $T_i \perp \varepsilon_{ij}$. Then testing whether $Q(t)$ is associated with $\upsilon_i(b_{0i} + \bar{\varepsilon}_i)$ is equal to the test for whether $Q(t)$ depends b_{0i}. Peng and Fine (2008) demonstrated that the test statistic based on b_{0i} converges to a normal distribution with mean zero and limited variance under

the null hypothesis. By Slusky's Theorem, we can establish that T_n^E based on the plug-in Kaplan-Meier estimator $\hat{Q}^E(t|b)$ also converges to a normal distribution with a mean zero and some variance ϕ_0^2 under H_0. The results are stated in the following theorem. The details of proof of the theorem can be found in the Appendix.

THEOREM: Given that $\phi_0^2 < \infty$, the test statistic T_n^E converges in distribution to $N(0, \phi_0^2)$ as n goes into infinity.

In order to carry out the proposed nonparametric testing procedure, we need to derive the variance ϕ_0^2 in the limiting distribution described in the above theorem. When b_{0i} are known, the corresponding variance of the test T_n, denoted by σ_0^2, can be estimated by

$$\hat{\sigma}_0^2 = n^{-1} \sum_{i=1}^{n} \left[\int_{b_L}^{b_U} \int_{t_L}^{t_U} \hat{w}(t,b) \left\{ \hat{\iota}_{+,i}(t|b) - \hat{\iota}_{-,i}(t|b) \right\} dtdb \right]^2$$

where $\hat{\iota}_{+,i}(t|b)$ is asymptotically equivalent to $\iota_{+,i}(t|b)$ which is a function of the Kaplan-Meier estimator based on $b_{0i} \geq b$ and can be obtained by using influence function. $\hat{\iota}_{-,i}(t|b)$ is defined similarly for $b_{0i} < b$. However, there are some difficulties due to the latent random effects. It is very complicated to derive the exact and general form for $\hat{\iota}_{+,i}^E(t|b)$ due to that it is conditioning on the estimated b_{0i}^E. Hence, we suggest using a bootstrapping technique (Efron, 1994) to estimate the variance of the test statistic T_n^E.

5.5 Simulation Studies

To evaluate the performance of the nonparametric testing procedure, we conduct simulation studies in different settings. The simulation procedure is provided as follows. First, we generate random observations for the random intercept from a normal distribution with mean 0 and variance σ_b^2 where σ_b^2 is chosen to be 2, i.e., $b_{0i} \sim N(0, 2)$, $i = 1, ..., n$. Conditional on b_{0i}, the time-to-event process is generated from a survival model. To evaluate the robustness of the proposed method against model

misspecification, we consider the proportional hazards model that can be expressed as

$$\lambda(t_i|b_{0i}, \mathbf{Y}_i) = \lambda_0(t_i) \exp(\eta_0 b_{0i})$$

where the baseline hazard $\lambda_0(t_i)$ is assumed to follow a Weibull distribution with a scale of 2 and shape of 20. Another survival model we consider is the accelerated failure time model that has a form of

$$\log T_i = \eta_0 b_{0i} + \epsilon_i$$

where the error term is generated from a logistic distribution with a location parameter equal to 2 and a scale parameter of 0.5. That is, the proportional hazards assumption does not hold under the accelerated failure model. The coefficient parameter η_0 is set to be 0 under the null hypothesis and chosen to be 0.15 and 0.2 to calculate the power of the test.

The censoring time for each subject, denoted by C_i, is generated from a Weibull distribution, where the scale and shape parameters are chosen to given different censoring percentages of 10%, 20%, and 30%. Then the observed time-to-event data are $\tilde{T}_i = \min(T_i, C_i)$.

Based on b_{0i}, repeated measurements for the i-th subject are simulated from the model

$$Y_{ij} = \beta_0 + \beta_1 X_i + b_{0i} + \varepsilon_{ij}, j = 1, ..., m_i$$

where $(\beta_0, \beta_1) = (1, 1)$ and $\mathbf{X}_{1i}$ is a random observation from uniform distribution $U[1, 10]$. The error term ε_{ij} is from a Normal distribution with mean 0 and variance 1. The number of repeated measures $m_i = 5$ is fixed.

In particular, we propose to use a bootstrapping procedure to obtained the variance of the nonparametric test statistics based on the predicted random effects of b_{0i}. Bootstrapped samples are drawn randomly with replacement from $(\mathbf{Y}_i, \tilde{T}_i, \Delta_i), i =$

$1, ..., n$. We calculate T_n^E for each bootstrapped sample and variance estimator is based on all the T_n^E from all bootstrapped samples (number of bootstrapping= 500).

For each setting, we generate 500 data sets with sample size $n = 50$, 100, and 200 and perform four tests, including (i) the score test based on proportional hazards assumption using b_{0i} (i.e., assuming that the underlying random effects are observed), (ii) the score test based on proportional hazards assumption using the estimated posterior expectation b_{0i}^E, (iii) the nonparametric test using b_{0i}, and (iv) the nonparametric test using b_{0i}^E. The size of test under null hypothesis and power of test under alternative hypothesis are calculated as empirical proportions of rejection based on a nominal significance level of 0.05. The simulation results are shown in Table 1.

Table 1 shows that when the survival model assumption is true, all the four tests retain a type I error close to the nominal level of 0.05. Overall, the empirical power increases with the increasing of sample size and effect size and decreasing of censoring percentages. In all scenarios, the tests using b_{0i} has more power compared to the tests based on the estimated posterior expectation b_{0i}^E. When the proportional hazards model is the true underlying survival model, the score test based on the model assumption has larger power than the proposed nonparametric test. When the underlying survival submodel has a nonproportional hazards form, both the score test under the proportional hazards model assumption and nonparametric test are still valid since the type I error is still close to nominal level of 0.05. However, the nonparametric testing procedure is more powerful compared to the model-based score test. In addition, the loss of power for score test is larger than that of nonparametric test with the increasing of censoring percentage in the time to event. In conclusion, the proposed nonparametric testing approach performs reasonably well compared to the model-based score test. Particularly, if the model is not correctly specified, the nonparametric test is more powerful than the score test.

Table 5.1. Simulation Results for Continuous Time-to-Event and Longitudinal Process

			PH Submodel				AFT Submodel			
CP[1]	η_0	n	T_s[2]	T_s^E	T_n	T_n^E	T_s	T_s^E	T_n	T_n^E
10%	0	50	0.062	0.066	0.066	0.072	0.046	0.054	0.076	0.078
		100	0.050	0.054	0.058	0.064	0.052	0.056	0.066	0.076
		200	0.052	0.054	0.058	0.062	0.050	0.052	0.060	0.062
	0.15	50	0.292	0.256	0.270	0.200	0.362	0.306	0.368	0.310
		100	0.496	0.420	0.430	0.414	0.546	0.502	0.592	0.540
		200	0.768	0.716	0.690	0.644	0.704	0.668	0.862	0.808
	0.2	50	0.450	0.400	0.420	0.366	0.498	0.460	0.512	0.472
		100	0.748	0.668	0.628	0.578	0.770	0.694	0.790	0.732
		200	0.958	0.906	0.918	0.876	0.980	0.946	0.982	0.972
20%	0	50	0.052	0.066	0.052	0.064	0.058	0.072	0.064	0.080
		100	0.048	0.050	0.058	0.058	0.048	0.056	0.052	0.060
		200	0.050	0.046	0.054	0.056	0.050	0.054	0.054	0.056
	0.15	50	0.218	0.196	0.218	0.188	0.328	0.292	0.362	0.316
		100	0.476	0.398	0.424	0.388	0.508	0.478	0.584	0.534
		200	0.776	0.716	0.654	0.616	0.702	0.660	0.856	0.804
	0.2	50	0.366	0.292	0.348	0.322	0.496	0.460	0.502	0.464
		100	0.664	0.596	0.604	0.568	0.706	0.664	0.784	0.728
		200	0.952	0.912	0.882	0.852	0.936	0.906	0.974	0.966
30%	0	50	0.054	0.060	0.068	0.076	0.062	0.068	0.070	0.076
		100	0.048	0.052	0.052	0.058	0.054	0.058	0.056	0.056
		200	0.048	0.050	0.046	0.050	0.052	0.054	0.054	0.066
	0.15	50	0.224	0.166	0.206	0.174	0.308	0.272	0.354	0.312
		100	0.416	0.376	0.368	0.346	0.502	0.472	0.584	0.532
		200	0.698	0.650	0.624	0.564	0.696	0.652	0.852	0.790
	0.2	50	0.330	0.290	0.336	0.296	0.484	0.442	0.488	0.426
		100	0.648	0.558	0.548	0.518	0.680	0.626	0.782	0.700
		200	0.910	0.874	0.844	0.804	0.906	0.874	0.970	0.952

[1] CP: censoring percentages

[2] T_s: the score test based on PH assumption using b_{0i}

T_s^E: the score test based on PH assumption using BLUP b_{0i}^E

T_n: the nonparametric test using b_{0i}

T_n^E: the nonparametric test using BLUP b_{0i}^E

5.6 A Real Data Example

We consider a data set from a prospective cohort study known as the Michigan Polybrominated Biphenyl (PBB) Registry. In 1973, PBB, which is a man-made chemical used as fire retardants in plastics, was accidentally mixed into livestock feed and consumed by cattle, pigs, and chickens (Carter, 1976; Fries, 1985). Contaminated farm products were sold throughout the state and residents across the state of Michigan were exposed to PBB. Children who were born years later may also have been exposed in the womb and through breastfeeding. Laboratory animal studies showed that PBB exposure during pregnancy and early infancy could change the hormonal signaling necessary in the developing fetus and neonate. Hence, there has been some suspicion that PBB exposure may disrupt endocrine functions in humans. A study is undertaken to investigate the relationship between PBB in utero and pubertal development in females.

Specifically, if mothers participated in the Michigan PBB Registry, all female offspring who were born during or after the Michigan PBB accident and who were at least 5 years of age in 1997 were invited to take part in the study. Information such as age at first menstrual period, whether breastfed, height, weight, and so on was collected for the daughters. A self-report pubertal development was also included using the Tanner stage schematic drawings of both breast and pubic hair development. For mothers who were in PBB cohort, questions including breastfeeding duration, farm chemical use, age at menarche, smoking status, alcohol consumption etc. were asked. Finally, a total of 327 daughters between 524 years of age were eligible for the menarche study ($N = 327$). Among the participants, 209 (63.9%) of them reached menarche at the time of the study. The average age of the participants was 15 years and the average age at menarche was 12.3 years. In addition, due to that limited breast milk and maternal PBB measurements were available, PBB exposure in utero was estimated using a maternal decay rate model (Blanck et al., 2000).

Blanck et al. (2000) evaluated the effects of the initial PBB measurement of mothers on the age to menarche of daughters. A Cox proportional hazards model was utilized to analyze the pubertal development in daughters and PBB exposure. The aim of our analysis is to evaluate the association between daughters' age at menarche and the repeated measurements of PBB level using the proposed nonparametric test. First, a mixed model for PBB level is defined as

$$PBB_{ij} = \beta_0 + \beta_1 \cdot (Time_{ij} - Time_i) + b_{0i} + \varepsilon_{ij}$$
$$\text{with } b_{0i} \sim N(0, \sigma_b^2) \text{ and } \varepsilon_{ij} \sim N(0, \sigma_e^2)$$

where PBB_{ij} is the PBB level in mothers, $Time_{ij}$ is defined as years before the date of birth of daughters, $Time_i$ is the date of birth of daughters, b_{0i} represents the random intercept, and ε_{ij} is the random error. Under this mixed effects model, the random intercept has an interpretation of PBB level for each mother at time of giving birth to the daughter (i.e., $Time_{ij} = Time_i$). The null hypothesis testing of interest is $H_0 : \Pr\{T > t|b_{0i}\} = \Pr\{T > t\}$, where T stands for daughter's age at menarche. In other words, under null hypothesis, the longitudinal PBB exposure in mothers has no impact on age at menarche in females.

Table 5.2. Testing PBB Exposure in Utero and Age at Menarche in Daughters

Scenarios	**Tests**	**Test Statistic**	**P-Values**
Breastfed	Nonparametric	4.0893	0.0432
& Mom's age at menarche > 12.6	Score (PH)	5.1965	0.0226
Breastfed	Nonparametric	2.5934	0.1073
& Mom's age at menarche < 12.6	Score (PH)	1.6108	0.2044

Table 2 displays the results based on our proposed nonparametric test and model based score tests using Cox proportional hazard (PH) model. For participants who

were breastfed and whose mother's age at menarche was less than 12.6 years, the relationship between PBB exposure at time of birth and risk of earlier age at menarche is not significant using all three tests and p-values are 0.1776 and 0.1800, respectively. If we adjust our analysis to those who were breastfed and whose mother's age at menarche is greater than 12.6 years, PBB level at time of birth is significantly associated with risk of reaching menarche with p-values from all three tests being less than 0.05.

5.7 Discussion

In this chapter, we developed a nonparametric testing procedure to investigate the association between a longitudinal biomarker and a time to event process. For the repeatedly measured biomarker, we assume a linear mixed effects model (with random intercept only). Based on the predicted random effects from the mixed model, we proposed a nonparametric test statistic without imposing parametric or semiparametric model assumptions for the time to event. We also established the asymptotic properties of the proposed test statistic and simulation studies were conducted to evaluate its performance under finite sample size. The simulation results showed that the test statistic has reasonable power to test the null hypothesis under different alternative hypotheses including proportional hazards and accelerated failure time models. Finally, we applied the proposed method to Michigan PBB data. We found that PBB exposure in mothers at time of giving birth and risk of earlier age at menarche in daughters were significantly associated. Our proposed approach can be extended to the case where random slope is also present in the linear mixed effects model. For example, we have both random intercept and random slope as $b_{0i} + b_{1i}t_{ij}$ in the linear mixed model and we may be interested in testing the relationship between time to event and $b_{0i}+b_{1i}t_{ij}$ (Wulfsohn and Tsiatis, 1997). With the assumption that (b_{0i}, b_{1i}) is bivariate normal, the Bayes estimator of $b_{0i} + b_{1i}t_{ij}$ can be used to construct our nonparametric test and the associated theory can be extended.

Appendix

I: Evaluation of asymptotic property of b_{0i}^E

In this manuscript, we consider using the empirical BLUP for b_{0i}^E that has a form of

$$b_{0i}^E = \hat{\upsilon}_i(\bar{Y}_i - \hat{\beta}_0 - \bar{\mathbf{X}}_i\hat{\boldsymbol{\beta}}), i = 1, ..., n$$

where $\hat{\upsilon}_i = \frac{m_i\hat{\sigma}_b^2}{m_i\hat{\sigma}_b^2+\hat{\sigma}_e^2}$. For other estimators for b_{0i}, the property can be derived similarly. Based on mixed model (5.1), we have that $\bar{Y}_i - \beta_0 - \bar{\mathbf{X}}_i\hat{\boldsymbol{\beta}} = b_{0i} + \bar{\varepsilon}_i$ where $\bar{\varepsilon}_i = \frac{1}{m_i}\sum_{j=1}^{m_i}\varepsilon_{ij}$. Then we can write the difference between b_{0i}^E and $\upsilon_i(b_{0i} + \bar{\varepsilon}_i)$ as

$$b_{0i}^E - \upsilon_i(b_{0i} + \bar{\varepsilon}_i) = \frac{\omega_i}{\sqrt{n}} \tag{5.5}$$

where $\frac{\omega_i}{\sqrt{n}} = (\hat{\upsilon}_i - \upsilon_i)(b_{0i} + \bar{\varepsilon}_i) - \hat{\upsilon}_i(\hat{\beta}_0 - \beta_0) - \hat{\upsilon}_i\bar{\mathbf{X}}_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Note that ω_i's are not i.i.d. random variables.

If $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_b^2$, and $\hat{\sigma}_e^2$ are MLE (or REML) from the mixed model, we have $\hat{\beta}_0 \xrightarrow{p} \beta_0$, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, $\hat{\sigma}_b^2 \xrightarrow{p} \sigma_b^2$, and $\hat{\sigma}_e^2 \xrightarrow{p} \sigma_e^2$. By Slusky's Theorem, we can obtain that $\hat{\upsilon}_i \xrightarrow{p} \upsilon_i$, $\hat{\upsilon}_i(\hat{\beta}_0 - \beta_0) \xrightarrow{p} 0$, $\hat{\upsilon}_i\bar{\mathbf{X}}_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{p} 0$, and $(\hat{\upsilon}_i - \upsilon_i)(b_{0i} + \bar{\varepsilon}_i) \xrightarrow{p} 0$. Hence, $\hat{\beta}_0 - \beta_0 = O_p(1/\sqrt{n})$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_p(1/\sqrt{n})$ and $\hat{\upsilon}_i - \upsilon = O_p(1/\sqrt{n})$. Then with probability $1 - \gamma_n$, $(\hat{\beta}_0 - \beta_0) \leq A\sqrt{\log n/n}$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \leq B\sqrt{\log n/n}$ and $(\hat{\upsilon}_i - \upsilon) \leq C\sqrt{\log n/n}$. That is, $\hat{\beta}_0 - \beta_0$, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|$ and $\hat{\upsilon}_i - \upsilon$ are bounded by $\sqrt{\log n/n}$.

We evaluate $n^{-1}\sum_{i=1}^n 1\{|\omega_i/\sqrt{n}| > \alpha_n\}$ for some α_n, such that $\alpha_n \to 0$ and $\alpha_n\sqrt{n}/\log^2 n \to \infty$. Then with probability at least $1 - \gamma_n$,

$$\begin{aligned}
&n^{-1}\sum_{i=1}^n 1\left\{\left|\omega_i/\sqrt{n}\right| > \alpha_n\right\} \\
&= n^{-1}\sum_{i=1}^n 1\left\{\left|(\hat{\upsilon}_i - \upsilon_i)(b_{0i} + \bar{\varepsilon}_i) - \hat{\upsilon}_i(\hat{\beta}_0 - \beta_0) - \hat{\upsilon}_i\bar{\mathbf{X}}_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right| > \alpha_n\right\} \\
&\leq \sup_{\|\boldsymbol{\delta}\|\leq B, |\eta|<C} n^{-1}\sum_{i=1}^n 1\left\{\left|\eta(b_{0i} + \bar{\varepsilon}_i) - (\upsilon_i + \eta\sqrt{\log n/n})\boldsymbol{\delta}^T\bar{\mathbf{X}}_i\right| > \alpha_n\sqrt{n}/\sqrt{\log n}\right\}
\end{aligned}$$

Therefore, by Glivenko-Cantelli theorem, we have

$$\sup_{\|\boldsymbol{\delta}\|\leq B, |\eta|<C} n^{-1}\sum_{i=1}^n 1\left\{\left|\eta(b_{0i} + \bar{\varepsilon}_i) - (\upsilon_i + \eta\sqrt{\log n/n})\boldsymbol{\delta}^T\bar{\mathbf{X}}_i\right| > \alpha_n\sqrt{n}/\sqrt{\log n}\right\} \xrightarrow{p} 0$$

$$n^{-1}\sum_{i=1}^{n} 1\{\left|\omega_i/\sqrt{n}\right| > \alpha_n\} \xrightarrow{p} 0 \tag{5.6}$$

II: Asymptotic equivalency of $\hat{Q}^E(t|b)$ *and* $\tilde{Q}(t|b)$

We use $\tilde{Q}(t|b)$ denote the stratified Kaplan-Meier estimator of $Q^E(t|b) := Pr(T > t|\upsilon_i(b_{0i} + \bar{\varepsilon}_i) > b)$, $i = 1, ..., n$. Following Peng and Fine (2008), we have

$$\sup_{0<t\leq t_U, b_L\leq b\leq b_U} \left|\tilde{Q}(t|b) - Q^E(t|b)\right| \xrightarrow{p} 0$$

Let $\hat{Q}^E(t|b)$ denote the stratified Kaplan-Meier estimator based on $\hat{b}_{0i}$. If we can show

$$\sup_{0<t\leq t_U, b_L\leq b\leq b_U} \left|\hat{Q}^E(t|b) - \tilde{Q}(t|b)\right| \xrightarrow{p} 0, \tag{5.7}$$

then we immediately obtain

$$\sup_{0<t\leq t_U, b_L\leq b\leq b_U} \left|\hat{Q}^E(t|b) - Q^E(t|b)\right| \xrightarrow{p} 0.$$

We denote $\upsilon_i(b_{0i} + \bar{\varepsilon}_i)$ by U_i. Since

$$\begin{aligned}\hat{Q}^E(t|b) - \tilde{Q}(t|b) = &\prod_{i=1}^{n}\left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i + \omega_i/\sqrt{n} > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i + \omega_i/\sqrt{n} > b\} + 1}\right)^{1\{\Delta_i=1, \tilde{T}_i\leq t, U_i+\omega_i/\sqrt{n}>b\}} \\ &- \prod_{i=1}^{n}\left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\} + 1}\right)^{1\{\Delta_i=1, \tilde{T}_i\leq t, U_i>b\}},\end{aligned}$$

we have

$$
\begin{aligned}
&\log\left(\hat{Q}^E(t|b)\right) - \log\left(\tilde{Q}(t|b)\right) \\
&= \sum_{i=1}^{n} 1\{\Delta_i = 1, \tilde{T}_i \le t, U_i + \omega_i/\sqrt{n} > b\} \log\left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i + \omega_i/\sqrt{n} > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i + \omega_i/\sqrt{n} > b\} + 1}\right) \\
&- \sum_{i=1}^{n} 1\{\Delta_i = 1, \tilde{T}_i \le t, U_i > b\} \log\left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\} + 1}\right) \\
&= \sum_{i=1}^{n} 1\{\Delta_i = 1, \tilde{T}_i \le t, U_i + \omega_i/\sqrt{n} > b\} \times \\
&\left(\log\left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i + \omega_i/\sqrt{n} > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i + \omega_i/\sqrt{n} > b\} + 1}\right) - \log\left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\} + 1}\right)\right) \\
&+ \sum_{i=1}^{n} \left(1\{\Delta_i = 1, \tilde{T}_i \le t, U_i + \omega_i/\sqrt{n} > b\} - 1\{\Delta_i = 1, \tilde{T}_i \le t, U_i > b\}\right) \times \\
&\log\left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\} + 1}\right) \\
&:= I_1 + I_2
\end{aligned}
$$

We consider I_2 first,

$$
\begin{aligned}
|I_2| &\le n^{-1} \sum_{i=1}^{n} \left(1\{\Delta_i = 1, \tilde{T}_i \le t, U_i + \omega_i/\sqrt{n} > b\} - 1\{\Delta_i = 1, \tilde{T}_i \le t, U_i > b\}\right) \\
&\qquad \times n \max_{1 \le i \le n} \left| \log\left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\} + 1}\right) 1\{\tilde{T}_i \le t\} \right| \\
&:= I_{21} \times I_{22}
\end{aligned}
$$

For I_{21}, we have

$$
\begin{aligned}
& I_{21} \\
& \leq n^{-1} \sum_{i=1}^{n} 1\{\Delta_i = 1, \tilde{T}_i \leq t, |U_i - b| < |\omega_i/\sqrt{n}|\} \\
& \leq n^{-1} \sum_{i=1}^{n} 1\{\Delta_i = 1, \tilde{T}_i \leq t, |U_i - b| \geq \alpha_n, |U_i - b| < |\omega_i/\sqrt{n}|\} \\
& \quad + n^{-1} \sum_{i=1}^{n} 1\{\Delta_i = 1, \tilde{T}_i \leq t, |U_i - b| < \alpha_n, |U_i - b| < |\omega_i/\sqrt{n}|\} \\
& \leq n^{-1} \sum_{i=1}^{n} 1\{\Delta_i = 1, \tilde{T}_i \leq t, \alpha_n < |\omega_i/\sqrt{n}|\} + n^{-1} \sum_{i=1}^{n} 1\{\Delta_i = 1, \tilde{T}_i \leq t, |U_i - b| < |\alpha_n|\} \\
& \leq n^{-1} \sum_{i=1}^{n} 1\{\alpha_n < |\omega_i/\sqrt{n}|\} + n^{-1} \sum_{i=1}^{n} 1\{|U_i - b| < |\alpha_n|\}
\end{aligned}
$$

By equation (5.6) and the property of the distribution of U_i (absolute continuous), we have

$$
\sup_{t,b} n^{-1} \left| \sum_{i=1}^{n} \left(1\{\Delta_i = 1, \tilde{T}_i \leq t, U_i + \omega_i/\sqrt{n} > b\} - 1\{\Delta_i = 1, \tilde{T}_i \leq t, U_i > b\} \right) \right| \xrightarrow{p} \quad 0 \quad (5.8)
$$

It is straightforward to see that

$$
\begin{aligned}
& n \max_{1 \leq i \leq n} \left| \log \left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\} + 1} \right) 1\{\tilde{T}_i \leq t\} \right| \\
& \leq n \left| \log \left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > t, U_i > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > t, U_i > b\} + 1} \right) \right| = n \log \left(1 - \frac{1}{\sum_{k=1}^{n} 1\{\tilde{T}_k > t, U_i > b\} + 1} \right) \\
& \leq \frac{2n}{\sum_{k=1}^{n} 1\{\tilde{T}_k > t, U_i > b\} + 1},
\end{aligned}
$$

where the last inequality follows from that $|\log(1-t)| \leq 2t$ for $0 < t < 1/2$.

By Glivenko-Cantelli theorem, we have

$$
\sup_{t,b} n \max_{1 \leq i \leq n} \left| \log \left(\frac{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\}}{\sum_{k=1}^{n} 1\{\tilde{T}_k > \tilde{T}_i, U_i > b\} + 1} \right) 1\{\tilde{T}_i \leq t\} \right| \overset{p}{\leq} 2/Pr(\tilde{T} > t_U, U > b_U). \quad (5.9)
$$

Combining (5.8) and (5.9) together yields $\sup_{t,b} |I_2| \xrightarrow{p} 0$. We can also show $\sup_{t,b} |I_1| \xrightarrow{p} 0$ with similar arguments. Therefore, we can establish (5.7) and

$$
\sup_{0 < t \leq t_U, b_L \leq b \leq b_U} \left| \hat{Q}^E(t|b) - Q^E(t|b) \right| \xrightarrow{p} 0.
$$

follows. That is, the difference between the plug-in Kaplan-Meier estimator and the corresponding true survival function is asymptotically zero. It is easy to see that, given $T_i \perp \varepsilon_{ij}$,

$$T_i \perp b_{0i} \Leftrightarrow T_i \perp \upsilon_i(b_{0i} + \bar{\varepsilon}_i).$$

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this dissertation, we focus our attention to method developments for the analysis of a longitudinal outcome and a time-to-event process with the emphasis on modeling repeatedly measured menstrual lengths and time-to-pregnancy. The first four chapters of this dissertation focus on the joint modeling approach for menstrual cycle lengths and time-to-pregnancy from the MSSWOW data. Specifically, the multiple observations of the menstrual cycle lengths (MCLs) are modeled through a Clayton-Oakes model and a discrete survival model is assumed for time-to-pregnancy (TTP) and a share parameter is introduced to model the association between MCLs and TTP. For marginal distribution in the Clayton-Oakes model, we consider both parametric and semiparametric models.

In Chapter 2, we consider the joint model where marginal distributions are specified as Weibill distributions. We consider the maximum likelihood estimation and the variance of parameters are obtained via deriving the corresponding information matrix. In Chapter 3, we relax the parametric assumptions of Weibull distribution by proposing a semiparametric linear transformation model, which includes the commonly used proportional hazards model and proportional odds model as two special cases. The joint modeling framework is very flexible and is able to handle many

complications raised by MSSWOW data such as truncation, censoring and missingness. When semiparametric assumptions are imposed, EM algorithm is developed by exploiting the equivalence of Clayton-Oakes models and gamma frailty models. In Chapter 4, we propose a computationally simple two-stage estimating procedure with the same joint model. Furthermore, we implement the two-stage method to provide reasonable parameter estimations while allowing flexible, different copula models. We also extend the model to other scenarios where (1) different dependence parameters according to age groups in the Clayton-Oakes model; (2) discrete survival models for TTP such as proportional odds model; and (3) alternative copula models including positive stable models for the repeatedly measured MCLs.

We apply the proposed joint model to MSSWOW data. The data analysis results show that age group has significant impact on the distribution of MCL data. Specifically, the older group is signicantly different and has shorter MCLs compared to the other groups. The association among menstrual length is signicantly lower in the oldest age group compared to other groups. In addition, smoking was a significant risk factor associated with MCLs and frequency of unprotected intercourse has significant influence on TTP. The association between MCL and TTP is significant. Among different transformation models, proportional odds model appears to fit the MCLs data better.

In Chapter 5, we develop nonparametric statistical testing tool to study whether a repeatedly measured biomarker is a good predictor for the onset of disease. We first assume a linear mixed model for the repeated measurements of the biomarker. Based on the estimated random effects from the linear mixed model, we propose a nonparametric test statistic for the null hypothesis that there is no association between the biomarker and time to event. That is, no parametric or semiparametric model assumptions are imposed for the time-to-event process. We also examine the asymptotic properties of the proposed test statistic and evaluate its finite sample

performance using simulation studies. We show that the test statistic has reasonable power to test the null hypothesis under different alternative hypotheses including proportional hazards and accelerated failure time models. Finally, we demonstrate the practical utility of the testing procedure by applying the method to Michigan PBB data.

6.2 Future Work

In this section, we discuss several possible extensions and future work of this dissertation.

In the joint modeling approach, we introduce covariates on marginal distributions of menstrual cycle lengths and the distribution of TTP and a common parameter is introduced to model the dependence between the length and the TTP. How one can select covariates for each distribution is unclear and needs further investigations. Another related topic is the goodness-of-fit of the joint model including the Clayton-Oakes model and the distribution of the TTP. With semiparamteric (transformation model) specification of marginal distributions in the Clayton-Oakes model, obtaining standard errors from the EM algorithm is challenging. Recent work (Xu et al., 2014) may provide insight into obtaining proper standard errors.

In Chapter 5, to develop a nonparametric statistical test to study the association of a repeatedly measured biomarker and the onset of disease, a linear mixed model for the biomarker is assumed. It would be of interest to relax this assumption since the biomarker may not necessarily follow a normal model.

References

[1] Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer.

[2] Andersen, P.K. and Gill, R.D. (1982) Cox's Regression Model for Counting Processes: a Large Sample Study. *The Annals of Statistics* Vol 10, No. 4, 1100-1120.

[3] Barnhart, H.X., Kosinski, A.S. and Sampson, A.R. (1999) A Regression Model for Multivariate Random Length Data. *Statistics in Medicine* 18, 199-211.

[4] Barnhart, H.X. and Sampson, A.R. (1995) Multiple Population Models for Multivariate Random Length Data-With Applications in Clinical Trials. *Biometrics* 51, 195-204.

[5] Baird, D.D., Wilcox, A.J. and Weinberg, C.R. (1986) Use of time to pregnancy to study environmental exposures. *American Journal of Epidemiology* 124, 470-480.

[6] Belsey, M.A. et al. (1987) *WHO Laboratory Manual for Examination of Human Semen and Semen-Cervical Mucus Interaction.* 2nd edition. Cambridge University Press, Cambridge

[7] Bennett, S. (1983) Analysis OF Survival Data by the Proportional Odds Model. *Statistics in Medicine* Vol 2, 273-277.

[8] Blanck, H.M., Marcus, M., Hertzberg, V.S., Tolbert, P.E., Rubin, C., Henderson, A.K., and Zhang, R.H. (2000) Determinants of Polybrominated Biphenyl Serum Decay among Women in the Michigan PBB Cohort. *Environmental Health Perspective* 108(2), 147-152.

[9] Blanck, H.M., Marcus, M., Tolbert, P.E., Rubin, C., Henderson, A.K., Hertzberg, V.S., Zhang, R.H., and Cameron, L. (2000) Age at Menarche and Tanner Stage in Girls Exposed In Utero and Postnatally to Polybrominated Biphenyl. *Epidemiology* 11(6), 641-647.

[10] Bonde, J.P., Joffe, M., Sallmen, M., Kristensen, P., Olsen, J., Roeleveld, N. and Wilcox, A. (2006) Validity Issues Relating to Time-to-Pregnancy Studies of Fertility. *Epidemiology* Vol 17, No. 4, 347-349.

[11] Brensike, J.F., Kelsey, S.F., Passamani, E.R. et al. (1982) NHLBI type II coronary intervention study: design, methods and baseline characteristics. *Controlled Clinical Trials* 3, 91.

[12] Brensike, J.F., Levy, R.I., Kelsey, S.F. et al. (1984) Effect of theray with cholestyranmine on progression of coronary arteriosclerosis: results of the NHLBI type II coronary intervention study. *Circulation* 69, 313-324.

[13] Breslow, N.E. and Clayton, D.G. (1993) Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* Vol 88, No. 421, 9-25.

[14] Brown, E.R. and Ibrahim, J.G. (2003) A Bayesian Semiparametric Joint Hierarchical Model for Longitudinal and Survival Data. *Biometrics* 59, 221-228.

[15] Brown, E.R. and Ibrahim, J.G. (2003) Bayesian Approaches to Joint Cure-Rate and Longitudinal Models with Applications to Cancer Vaccine Trials. *Biometrics* 59, 686-693.

[16] Bullen, B.A., Skrinar, G.S., and Beitins, I.Z. et al. (1985) Induction of menstrual disorders by strenuous exercise in untrained women. *New England Journal of Medicine* 312, 1349-1353.

[17] Burdorf, A., Brand, T. Jaddoe, V.W., Hofman, A., Mackenbach, J.P. and Steegers, E.A.P. (2011) The effects of work-related maternal risk factors on time to pregnancy, preterm birth and birth weight: the Generation R Study. *Occupational and Environmental Medicine* 68, 197-204.

[18] Bycott, P. and Taylor, J. (1998) A Comparison Of Smoothing Techniques For Cd4 Data Measured With Error In A Time-Dependent Cox Proportional Hazards Model. *Statistics in Medicine* 17, 2061-2077.

[19] Cai, J. and Prentice, R.L. (1995) Estimating Equations for Hazard Ratio Parameters Based on Correlated Failure Time Data. *Biometrika* Vol 82, No. 1, 151-164.

[20] Carter, L.J. (1976) Michigan's PBB incident: Chemical Mix-up Leads to Disaster. *Science* 192, 240-243.

[21] Carvalho, M.S. and Knorr-Held, L. (2003) Modelling discrete time survival data with random slopes: evaluating haemodialysis centres. *Statistics in Medicine* 22, 3543-3555.

[22] Chen, X., Fan, Y. and Tsyrennikov, V. (2006) Efficient Estimation of Semiparametric Multivariate Copula Models. *Journal of the American Statistical Association* Vol. 101, No. 475, 1228-1240.

[23] Chen, S.X. and Huang, T.M. (2007) Nonparametric estimation of copula functions for dependence modelling. *Canadian Journal of Statistics* 35, 265282.

[24] Chen, L.M., Ibrahim, J.G. and Chu, H. (2011) Sample size and power determination in joint modeling of longitudinal and survival data. *Statistics in Medicine* 30, 2295-2309.

[25] Chen, M-H., Ibrahim, J.G., and Sinha, D. (2004) A New Joint Model for Longitudinal and Survival Data with a Cure Fraction. *Journal of Multivariate Analysis* Vol 91: 18-34.

[26] Chen, H.Y. and Little, R.J. (1999) Proportional Hazards Regression with Missing Covariates. *Journal of the American Statistical Association* Vol. 94, No. 447, 896-908.

[27] Chen, C.M. and Yu, C.Y. (2012) A Two-Stage Estimation in the Clayton-Oakes Model with Marginal Linear Transformation Models for Multivariate Failure Time Data. *Lifetime Data Analysis* 18, 94-115

[28] Chen, M-H., Ibrahim, J.G., and Sinha, D. (2004) A New Joint Model for Longitudinal and Survival Data with a Cure Fraction. *Journal of Multivariate Analysis* Vol 91: 18-34.

[29] Cheng, S.C., Wei, L.J. and Ying, Z. (1995) Analysis of Transformation Models with Censored Data. *Biometrika* Vol. 82, No. 4, 835-845.

[30] Chi, Y-Y. and Ibrahim, J.G. (2006) Joint Models for Multivariate Longitudinal and Multivariate Survival Data. *Biometrics* Vol 62: 432-445.

[31] Chiazze, L., Brayer, F.T., Macisco, J.J., Parker, M.P. and Duffy, B.J. (1968) The Length and Variability of the Human Menstrual Cycle. *Journal of the American Medical Association* Vol. 203, No. 6, 377-380.

[32] Clayton, D.G. (1978) A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika* Vol. 65, No. 1, 141-151.

[33] Cook, R.D. and Johnson, M.E. (1981) A Family of Distributions for Modelling Non-elliptically Symmetric Multivariate Data. *Journal of the Royal Statistical Society* 43, 210-218.

[34] Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement* Vol 7: 249253.

[35] Cox, D.R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society* Vol. 34, No. 2, 187-220.

[36] Cox, D.R. (1975) Partial Likelihood. *Biometrika* Vol. 62, No. 2, 269-276.

[37] abrowska, D.M. and Doksum, K.A. (1988) Estimation and testing in a two-sample generalized odds-rate model. *Journal of American Statistical Association* 83, 744749.

[38] Dafni, U.G. and Tsiatis, A.A. (1998) Evaluating Surrogate Markers of Clinical Outcome When Measured with Error. *Biometrics* Vol. 54, No. 4, 1445-1462.

[39] DeGruttola, V. and Tu, X.M. (1994) Modeling Progression of CD4-Lymphocyte Count and its Relationship to Survival Time. *Biometrics* Vol 50: 1003-1014.

[40] Deheuvels, P. (1979) La fonction de dependance empirique et ses proprietes Un test non parametrique dindependance.*Academie Royale de Belgique Bulletin de la Classe des Sciences 5e Serie* 65, 274-292.

[41] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society* Series B, 39(1), 1-38.

[42] Diggle, P.J. (1998) Dealing with missing values in longitudinal studies. In Recent Advances in the Statistical Analysis of Medical Data, Everitt, B.S. and Dunn, G. (eds). *Arnold: Paris* 203228.

[43] Diggle, P.D. and Kenward, M.G. (1994) Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics* 43, 4973.

[44] Diggle, P.J., Sousa, I., and Chetwynd, A.G. (2008) Joint Modelling of Repeated Measurements and Time-To-Event Outcomes: The Fourth Armitage Lecture. *Statistics in Medicine* Vol 27: 2981-2998.

[45] DiRienzo, A.G. and Lagakos, S.W. (2001) Effects of Model Misspecification on Tests of No Randomized Treatment Effect Arising from Cox's Proportional Hazards Model. *Journal of the Royal Statistical Society* Series B 63, 745-757.

[46] DiRienzo, A.G. and Lagakos, S.W. (2001) Bias Correction for Score Tests Arising from Mis-specified Proportional Hazards Regression Models. *Biometrika* 88, 421-434.

[47] Donoho, D.L. and Liu, R.C. (1988) The "Automatic" Robustness of Minimum Distance Functionals. *The Annals of Statistics* Vol 16, No. 2, 552-586.

[48] Dunson, D.B. (2001) Bayesian Modeling of the Level and Duration of Fertility in the Menstrual Cycle. *Biometrics* Vol. 57, No. 4, 1067-1073.

[49] Dunson, D.B., Chen, Z. and Harry, J. (2003) A Bayesian Approach for Joint Modeling of Cluster Size and Subunit-Specific Outcomes. *Biometrics* 59, 521-530.

[50] Dunson, D.B., Colombo, B., and Baird, D.D. (2002) Changes With Age in the Level and Duration of Fertility in the Menstrual Cycle. *Human Reproduction* 17, 1399-1403.

[51] Dunson, D.B. and Neelon, B. (2003) Bayesian Inference on Order-Constrained Parameters in Generalized Linear Models. *Biometrics* 59(2), 286-295.

[52] Dunson, D.B. and Zhou, H. (2000) A Bayesian Model for Fecundability and Sterility. *Journal of the American Statistical Association* 95:452, 1054-1062.

[53] Ecochard, R. and Clayton, D.G. (2000) Multivariate Parametric Random Effect Regression Models for Fecundability Studies. *Biometrics* Vol. 56, No. 4, 1023-1029.

[54] Efron, B. (1994) Missing Data, Imputation, and the Bootstrap. *Journal of the American Statistical Association* 89(426), 463-475.

[55] Efron, B. and Tibshrani, R.J. (1993) *An Introduction to the Bootstrap.* Chapman and Hall, New York. Z.

[56] Elashoff, R.M., Li, G. and Li, N. (2008) A Joint Model for Longitudinal Measurements and Survival Data in the Presence of Multiple Failure Types. *Biometrics* 64, 762-771.

[57] Fahrmeir, L. (1994) Dynamic Modelling and Penalized Likelihood Estimation for Discrete time Survival data. *Biometrika* Vol 81, No. 2, 317-330.

[58] Faucett, C. J. and Thomas, D. C. (1996) Simultaneously modeling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine* 15, 1663-1685.

[59] Finkelstein, D.M., Wang, R., Ficociello, L.H., and Schoenfeld, D.A. (2010) A Score Test for Association of a Longitudinal Marker and an Event with Missing Data. *Biostatistics* 66(3), 726-732.

[60] Florack, E.I.M., Zielhuis, G.A. and Rolland, R. (1994) The Influence of Occupational Physical Activity on the Menstrual Cycle and Fecundability. *Epidemiology* Vol 5, No. 1, 14-18.

[61] Frees, E.W. and Valdez, E.A. (1998) Understanding relationships using copulas. *North American Actuarial Journal* 2(1):1-25.

[62] Fries, G.F. (1985) The PBB Episode in Michigan: an Overall Appraisal. *Critical Reviews in Toxicology* 16(2), 105-156.

[63] Genest, C., Ghoudi, K. and Rivest, L.P. (1995) A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions. *Biometrika* Vol. 82, No. 3, 543-552.

[64] Genest, C. and MacKay, J. (1986) The Joy of Copulas: Bivariate Distributions with Uniform Marginals. *The American Statistician* 40, 280-283.

[65] Ghosh, D. and Lin, D.Y. (2003) Semi-parametric Analysis of Recurrent Events Data in the Presence of Dependent Censoring. *Biometrics* 59, 877-885.

[66] Glidden, D.V. and Self, S.G. (1999) Semi-parametric Likelihood Estimation in the Clayton-Oakes Failure Time Model. *Scandinavian Journal of Statistics* Vol 26, No. 3, 363-372.

[67] Glidden, D.V. (2000) A Two-Stage Estimator of the Dependence Parameter for the Clayton-Oakes Model. *Lifetime Data Analysis* 6, 141-156.

[68] Guo, Y., Manatunga, A.K., Chen, S., and Marcus, M. (2006) Modeling menstrual cycle length using a mixture distribution. *Biostatistics* Vol 7, No. 1, 100-114.

[69] Hanson, T., Bedrick, E., Johnson, W., and Thurmond, M. (2003) A mixture model for bovine abortion and fetal survival. *Statistics in Medicine* 22, 1725-1739.

[70] Harlow, S.D and Matanoski, G.M. (1991) The Association between Weight, Physical Activity, and Stress and Variation in the Length of the Menstrual Cycle. *American Journal of Epidemiology* 133:1, 38-49.

[71] Harlow, S.D., Lin, X. and Ho, M.J. (2000) Analysis of menstrual diary data across the reproductive life span Applicability of the bipartite model approach and the importance of within-woman variance. *Journal of Clinical Epidemiology* 53, 722-733.

[72] Harlow, S.D. and Zeger, S.L. (1991) An Application of Longitudinal Methods to the Analysis of Menstrual Diary Data. *Journal of Clinical Epidemiology* 44:10, 1015-1025.

[73] Henderson, R., Diggle, P., and Dobson, A. (2000) Joint Modelling of Longitudinal Measurements and Event Time Data. *Biostatistics* 1, 465-480.

[74] Henry, L. (1953) Fondements theoretiques des measures da la fecondite naturelle. *Perue de 1 Instut International de Statistique* 21, 135.

[75] Hogan, J.W. and Laird, N.M. (1997) Mixture Models For The Joint Distribution Of Repeated Measures And Event Times. *Statistics in Medicine* Vol 16, 239-257.

[76] Hogan, J.W., Lin, X., and Herman, B. (2004) Mixtures of Varying Coefficient Models for Longitudinal Data with Discrete or Continuous Nonignorable Dropout. *Biometrics* 60, 854-864.

[77] Hougaard, P. (1986) A Class of Multivariate Failure Time Distributions. *Biometrika* 73:3, 671-678.

[78] Hsieh, F., Tseng, Y.K. and Wang, J.L. (2006) Joint Modeling of Survival and Longitudinal Data: Likelihood Approach Revisited. *Biometrics* 62, 1037-1043.

[79] Hu, P., Tsiatis, A.A., and Davidian, M. (1998) Estimating the Parameters in the Cox Model When Covariate Variables are Measured with Error. *Biometrics* 54(4), 1407-1419.

[80] Huang, X. and Liu, L. (2007) A Joint Frailty Model for Survival and Gap Times Between Recurrent Events. *Biometrics* 63, 389-397.

[81] Huang, W.H., Zeger, S.L., Anthony, J.C., and Garrett, E. (2001) Latent variable model for joint anlaysis of multiple repeated measures and bivariate event times. *American Statistical Association* 96, 906914.

[82] Huster, W.J., Brookmeyer, R. and Self, S.G. (1989) Modelling Paired Survival Data with Covariates. *Biometrics* Vol 45, No. 1, 145-156.

[83] Ibrahim, J.G., Chen, M-H., and Sinha, D. (2004) Bayesian Methods for Joint Modeling of Longitudinal and Survival Data with Applications to Cancer Vaccine Studies. *Statistica Sinica* Vol 14: 863-883.

[84] Ibrahim, J.G., Chu, H., and Chen, L.M. (2010) Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data. *Journal OF Clinical Oncology* Vol 28: 2796-2801.

[85] Jacqmin-Gadda, H., Proust-Lima, C., Taylor, J.M.G., and Commenges, D. (2010) Score Test for Conditional Independence Between Longitudinal Outcome and Time to Event Given the Classes in the Joint Latent Class Model. *Biometrics* Vol 66: 11-19.

[86] Jensen, T.K., Scheike, T., Keiding, N., Schaumburg, I. and Grandjean, P. (1999) Fecundability in Relation to Body Mass and Menstrual Cycle Patterns. *Epidemiology* Vol. 10, No. 4, 422-428.

[87] Joe, H. (1994) Multivariate extreme value distributions with applications to environmental data. *Canadian Journal of Statistics* 22, 4764.

[88] Joe, H. (1997) Multivariate Models and Dependence Concepts. *Monographs on Statistics and Applied Probability* Chapman & Hall, London.

[89] Joe, H. (2005) Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis* 94(2), 401419.

[90] Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimator from Incomplete Observations. *Journal of the American Statistical Association* Vol 53: 457481.

[91] Kalbfleisch J.D. and Prentice, R.L. (2002) *The Statistical Analysis of Failure Time Data* Wiley, 2nd Edition.

[92] Kendall, M.G. (1938) A New Measure of Rank Correlation. *Biometrika* 30(12), 8189.

[93] Kendall, M.G. (1962) *Rank Correlation Methods.* 3rd Edition. Griffin, London.

[94] Keiding, N. (2006) Event history analysis and the cross-section. *Statistics in Medicine* Vol. 25, 2343-2364.

[95] Keiding, N., Kvist, K., Hartvig, H. and Tvede, M. (2002) Estimating time to pregnancy from current durations in a cross-sectional sample. *Biometrics* 3:4, 565-578.

[96] Klein, J.P. (1992). Semi-parametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* 48, 795-806.

[97] Koul, H.L. (1985) Minimum Distance Estimation in Linear Regression with Unknown Error Distribution. *Statistics and Probability Letters* 3, 1-8.

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

[99] Lagakos, S.W. (1988) Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in Medicine* 7(1-2), 257-274.

[100] Lagakos, S.W. and Schoenfeld, D.A. (1984) Properties of Proportional-Hazards Score Tests under Misspecified Regression Models. *Biometrics* Vol 40: 1037-1048.

[101] Law, N., Taylor, J.M.G and Sandler, H. (2002) The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. *Biostatistics* 3:4, 547-563.

[102] Lenton, E.A., Landgren, B.M., and Sexton, L. (1984) Normal variation in the length of the luteal phase of the menstrual cycle: Identification of the short luteal phase. *British Journal of Obstretrics and Gynecology* 91, 685-689.

[103] Li, Y., Klein, J.P., and Moeschberger, M.L. (1996) Effects of Model Misspecification in Estimating Covariate Effects in Survival Analysis for a Small Sample Size. *Computational Statistics and Data Analysis* 22, 177192.

[104] Liang, K.Y. and Zeger, S.L. (1995) Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science* 10, 158-199.

[105] Lin, H., McCulloch, C.E. and Mayne, S.T. (2002) Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in Medicine* 21, 2369-2382.

[106] Lin, X., Raz, J., and Harlow, S.D. (1997) Linear Mixed Models with Heterogeneous Within-Cluster Variances. *Biometrics* 53, 910-923.

[107] Lindstrom, M.J. and Bates, D.M. (1990) Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics* 46(3), 673-687.

[108] Liu, Y., Gold, E.B., Laslley, B.L. and Johnson, W.O. (2004) Factors Affecting Menstrual Cycle Characteristics. *American Journal of Epidemiology* Vol. 160, No. 2, 131-140.

[109] Little, R.J. (1993) Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association* Vol 88, No. 421, 125-134.

[110] Marcus, M. (1990) Epidemiologic Studies of VDT Use And Pregnancy Outcome. *Reproductive Toxicology* Vol. 4, 51-56.

[111] Marshall, A.W. and Olkin, I. (1988) Families of Multivariate Distributions. *Journal of the American Statistical Association* Vol. 83, No. 403, 834-841.

[112] McLain, A.C., Lum, K.J. and Sundaram, R. (2012) A Joint Mixed Effects Dispersion Model for Menstrual Cycle Length and Time-to-Pregnancy. *Biometrics* 68, 648-656

[113] Morgan, T.M. and Elashoff, R.M. (1986). Effects of Categorizing a Continuous Covariate on the Comparison of Survival Time. *Journal of the American Statistical Association* Vol 81: 917921.

[114] Morrell, C.H., Pearson, J.D., Carter, H.B. and Brant, L.J. (1995) Estimating unknown transition times using a piecewise nonlinear mixed-effects model in men with prostate cancer. *Journal of the American Statistical Association* 90, 45-53.

[115] Murphy, S.A., Rossini, A.J. and van der Vaart, A.W. (1997) Maximum Likelihood Estimation in the Proportional Odds Model. *Journal of the American Statistical Association* 92, 968-976.

[116] Murphy, F.A., Fauquet, C.M., Bishop, D.H.L., Ghabrial, S.A., Jarvis, A.W., Martelli, G.P., Mayo, M.A., and Summers, M.D. (1995) *Virus taxonomy: Sixth report of the International Committee on the Taxonomy of Viruses* Springer-Verlag, New York.

[117] Nielsen, G.G., Gill, R.D., Andersen, P.K. and Sorensen, T.I.A. (1992) A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models. *Scandinavian Journal of Statistics* 19, 25-43.

[118] Oakes, D. (1982) A Model for Association in Bivariate Survival Data. *Journal of the Royal Statistical Society* Vol 44, 414-422.

[119] Oakes, D. (1989) Bivariate Survival Models Induced by Frailties. *Journal of the American Statistical Association* Vol. 84, No. 406, 487-493.

[120] Pearson, J.D., Morrell, C.H., Landis, P.K. and Brant, L.J. (1994) Mixed-effects regression models for studying the natural history of prostate disease. *Statistics in Medicine* 13, 587-601.

[121] Peng, L. and Fine, J.P. (2008) Nonparametric Tests for Continuous Covariate Effects with Multistate Survival Data. *Biometrics* Vol 64: 1080-1089.

[122] Rizopoulos, D. (2013) Joint Modeling of Longitudinal and Time-to-Event Data: Challenges and Future Directions.*Studies in Theoretical and Applied Statistics* 199-209

[123] Rizopoulos, D. (2011) Dynamic Predictions and Prospective Accuracy in Joint Models for Longitudinal and Time-to-Event Data. *Biometrics* Vol 67,819-829.

[124] Rizopoulos, D. and Ghosh, P. (2011) A Bayesian Semiparametric Multivariate Joint Model for Multiple Longitudinal Outcomes and a Time-To-Event. *Statistics in Medicine* Vol 30: 1366-1380.

[125] Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009) Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society* Series B 71, 637654.

[126] Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008) Shared parameter models under random effects misspecification. *Biometrika* 95, 6374.

[127] Robinson, G.K. (1991) That BLUP Is a Good Thing: The Estimation of Random Effects. *Statistical Science* Vol 6: 15-32.

[128] Rothman, K.J. and Greenland, S. (1998) *Modern Epidemiology*. Lippincott-Raven, Philadelphia.

[129] Scharfstein, D.O., Daniels, M.J., and Robins, J.M. Incorporating Prior Beliefs about Selection Bias into the Analysis of Randomized Trials with Missing Outcomes. *Biostatistics* 4(4), 495.

[130] Scheike, T.H. and Jensen, T.K. (1997) A Discrete Survival Model with Random Effects: An Application to Time to Pregnancy. *Biometrics* Vol 53, No. 421, 318-329.

[131] Scheike, T.H. and Keiding, N. (2006) Design and analysis of time-to-pregnancy. *Statistical Methods in Medical Research*. 15:127-140

[132] Self, S. and Pawitan, Y. (1992) Modeling a marker of disease progression and onset of disease. *In AIDS Epidemiology: Methodological Issues* Birkhauser, Boston.

[133] Selvin, S. (1987) Two Issues Concerning the Analysis of Grouped Data. *European Journal of Epidemiology* Vol 3: 284287.

[134] Shih, J.H. and Louis, T.A. (1995) Inferences on the Association Parameter in Copula Models for Bivariate Survival Data. *Biometrics* 51, 1384-1399.

[135] Slate, E.H. and Turnbull, B.W. (2000) Statistical Models for Longitudinal Biomarkers of Disease Onset. *Statistics in Medicine* 19(4), 617-637.

[136] Small C.M., Manatunga A.K., Klein M., Feigelson H.S., Dominguez C.E., McChesney R., et al. (2006) Menstrual cycle characteristics: Associations with fertility and spontaneous abortion. *Epidemiology*. 17:52-60.

[137] Small, C.M., Manatunga, A.K. and Marcus, M. (2007) Validity of self-reported menstrual cycle length. *Annal of Epidemiology* 17:163-170.

[138] Song, P.X.K., Davidian, M. and Tsiatis, A.A. (2002) A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-To-Event Data. *Biometrics* Vol 58: 742753.

[139] Song, P.X.K., Fan, Y., and Kalbfleisch, J.D. (2005) Maximization by Parts in Likelihood Inference. *Journal of the American Statistical Association* 100(472), 1145-1158.

[140] Song, P.X.K., Li, M., and Yuan, Y. (2009) Joint Regression Analysis of Correlated Data Using Gaussian Copulas. *Biometrics* 65, 60-68.

[141] Sousa, I. (2011) A Review on Joint Modelling of Longitudinal Measurements and Time-to-Event. *Revstat* Vol 9: 57-81.

[142] Sweeting, M.J. and Thompson, S.G. (2011) Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal* 53(5), 750-763.

[143] Terrell, M.L., Manatunga, A.K., Small, C.M., Cameron, L.L., Wirth, J., Blanck, H.M., Lyles, R.H., and Marcus, M. (2008) A decay model for assessing polybrominated biphenyl exposure among women in the Michigan Long-Term PBB Study. *Journal of Exposure Science and Environmental Epidemiology* 18, 410-420.

[144] Thompson, W.A. (1977). On the treatment of grouped observations in life studies. *Biometrics* 33, 463-470.

[145] Thurmond, M.C., Branscum, A.J., Johnson, W.O., Bedrick, E.J., and Hanson, T.E. (2005) Predicting the probability of abortion in dairy cows: a hierarchical Bayesian logistic-survival model using sequential pregnancy data. *Preventive Veterinary Medicine* 68, 223-239.

[146] Treloar, A.E., Boynton, R.E., Behn, B., and Brown, B.W. (1967) Variation of the human menstrual cycle through reproductive life. *International Journal of Fertility* 12, 77126.

[147] Tseng, Y-K., Hsieh, F., and Wang, J-L. (2005) Joint Modelling of Accelerated Failure Time and Longitudinal Data. *Biometrika* Vol 92: 587-603.

[148] Tsiatis, A.A., DeGruttola, V., Wulfsohn, M.S. (1995) Modeling the Relationship of Survival to Longitudinal Data Measured with Error: Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association* Vol 90: 27-37.

[149] Tsiatis, A.A. and Davidian, M. (2001) A Semiparametric Estimator for the Proportional Hazards Model with Longitudinal Covariates Measured with Error. *Biometrika* Vol 88: 447-458.

[150] Tsiatis, A.A. and Davidian, M. (2004) Joint Modeling of Longitudinal and Time-to-Event Data: an Overview. *Statistica Sinica* Vol 14: 809-834.

[151] Vaupel, J.W., Manton, K.G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16, 439-454.

[152] Vonesh, E.F., Greene, T. and Schluchter, M.D. (2006) Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine* Vol 25, 143-163.

[153] Wang, Y. and Taylor, J.M.G. (2001) Jointly Modeling Longitudinal and Event Time Data with Application to Acquired Immunodeficiency Syndrome. *Journal of the American Statistical Association* Vol 96: 895-905.

[154] Weinberg, C.S., Baird, D.D., and Wilcox, A.J. (1994) Sources of bias in studies of time to pregnancy. *Statistics in Medicine* 13, 671-681.

[155] Weinberg, C.R. and Gladen, B.C. (1986) The Beta-Geometric Distribution Applied to Comparative Fecundability Studies. *Biometrics* 42, 547-560.

[156] Weinberg, C.R., Wilcox, A.J. and Baird. D.D. (1989) Reduced fecundability in women with prenatal exposure to cigarette smoking. *American Journal of Epidemiology* 129, 1072- 1078.

[157] Wu, M.C. and Carroll, R.J. (1988) Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. *Biometrics* 44, 175-188.

[158] Wu, L., Hu, X.J., and Wu, H. (2008) Joint Inference for Nonlinear Mixed-Effects Models and Time To Event at the Presence of Missing Data. *Biostatistics* 9, 308-320.

[159] Wulfsohn, M.S. and Tsiatis, A.A. (1997) A Joint Model for Survival and Longitudinal Data Measured with Error. *Biometrics* Vol 53: 330-339.

[160] Xu, C., Baines, P.D., and Wang, J.L. (2014) Standard error estimation using the EM algorithm for the joint modeling of survival and longitudinal data. *Biostatistics* 15(4), 731-744.

[161] Xu, J. and Zeger, S.L. (2001a) Joint analysis of longitudinal data comprising repeated measures and times to events. *Applied Statistics* 50, 375-387.

[162] Xu, J. and Zeger, S.L. (2001b) The Evaluation of Multiple Surrogate Endpoints. *Biometrics* Vol 57: 81-87.

[163] Yang, S. and Prentice, R.L. (1999) Semi-parametric Inference in the Proportional Odds Regression Model. *Journal of the American Statistical Association* Vol 94, 125-136.

[164] Zeng, D. and Cai, J. (2005) Asymptotic Results for Maximum Likelihood Estimators in Joint Analysis of Repeated Measurements and Survival Time. *The Annals of Statistics* Vol 33, No. 5, 2132-2163.

[165] Zeng, D. and Lin, D.Y. (2006) Maximum likelihood estimation in semiparametric transformation models for counting processes. *Biometrika* 93(3), 627-640.

[166] Zeng, D. and Lin, D.Y. (2007) Semiparametric Transformation Models With Random Effects for Recurrent Events. *Journal of the American Statistical Association* 102(477), 167-180.

Contents

List of Figures

List of Tables