2016

Semiparametric Estimation Methods For Complex Accelerated Failure Time Model

Yinding Wang
University of South Carolina

SEMIPARAMETRIC ESTIMATION METHODS FOR COMPLEX ACCELERATED FAILURE TIME MODEL

by

Yinding Wang

Bachelor of Engineering
Hefei University of Technology, 2005

Master of Engineering
Chongqing University, 2008

Master of Science in Public Health
University of South Carolina, 2012

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2016

Accepted by:

Jiajia Zhang, Major Professor

Suzanne McDermott, Committee Member

Bo Cai, Committee Member

Xiaoyan Lin, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

This dissertation would not have gotten this far without the guidance of my committee members, and support from my family.

My first most sincere appreciation must go to my advisor, Dr. Jiajia Zhang, for her support, encouragement, keen insights, selfless help and patient guidance, as well as the enormous amount of time she has spent helping me with the preparation of this dissertation.

Sincere gratitude is also given to my committee members, Dr. Suzanne McDermott, Dr. Bo Cai, and Dr. Xiaoyan Lin, for their precious time, helpful suggestions, and encouragements. Their contributions substantially improve the quality of this dissertation.

Especially, I would like to express my sincere gratitude to Dr. Suzanne McDermott for her selfless help, financial support and advices during my graduate studies in USC. My gratitude also goes to Dr. Bo Cai for his selfless help and patient guidance during my research as a research assistant in USC.

Finally, I am very grateful for the unconditional love and full support from my family. Without their love and support, I would not have the chance to make my dream into reality.

Abstract

The proportional hazards (PH) model and the accelerated failure time (AFT) model are the two most popular survival models in fitting the right-censored data. The AFT model is a useful alternative to the PH model, particularly when the PH assumption is not satisfied. Usually, the linear association is assumed with logarithm of survival time in the AFT model. However, the nonlinear association may exist in practice. The first project aims to handle the nonlinear component in the AFT model, which is called the semiparametric additive partial accelerated failure time (AP-AFT) model. Two estimation methods based on the rank-smooth method and the profile likelihood method are proposed, along with the variance estimation.

The other interest situation in practice is heterogeneity among subjects, which may lead to the different baseline distribution of patients with different characteristics. The AFT mixture model with latent subgroup is investigated in the second project. The semiparametric estimation method is improved by the expectation-maximization (EM) algorithm with the profile likelihood estimation method.

In practice, there exists the cases where either the PH model or the AFT model is appropriate to capture the data characteristic. The extended hazards (EH) model is developed to capture more general forms in survival data, which includes the PH and AFT models as its special cases. With the development of medical research, more and more diseases can be cured. Thus, patients may not die from the disease even with enough follow up time. Mixture cure model is developed to handle the survival data with possible cure fraction. The concepts of mixture model have been adapted to the PH and AFT models. However, there are limited studies on its extension to

the EH model.

The third and fourth projects aim to estimate the EH and EH mixture cure models with the monotone splines. The advantage of the monotone spline is that it can capture any shape of the survival function with the appropriate knots and degrees. The estimated survival curve is parametric, and the inference is easy.

All the above projects are studied through the comprehensive simulation studies. The appropriate data are used for illustration purposes. For example , Mayo primary biliary cirrhosis (PBC) data is used in the AP-AFT model, pregnancy data is applied in the AFT mixture model, Stanford heart transplant data is used in the EH model, the melanoma data from the ECOG phase III clinical trial E1684, and the leukemia data from bone marrow transplant study are used in the EH mixture cure model.

Table of Contents

Acknowledgments iii

Abstract iv

List of Tables ix

List of Figures xi

Chapter 1 Introduction 1

1.1 Standard Survival Model 1

1.2 Standard Cure Model 6

1.3 Outline of Dissertation 9

Chapter 2 Semiparametric Estimations for Additive Partial Accelerated Failure Time Models 11

2.1 Abstract 11

2.2 Introduction 11

2.3 Additive Partial Accelerated Failure Time Model 13

2.4 Estimation Procedure 14

2.5 Simulation Study 19

2.6 Real Data Analysis 24

2.7 Discussion and Conclusion 28

Chapter 3 Profile Likelihood based Estimation Method for the Accelerated Failure Time Mixture Model with Latent Subgroups 29

3.1 Abstract 29

3.2 Introduction 30

3.3 Semiparametric AFT mixture model 32

3.4 Estimation Procedure 33

3.5 Simulation Study 39

3.6 Real Data Analysis 42

3.7 Discussion and Conclusion 45

Chapter 4 Estimation Method for Extended Hazards Model . . 46

4.1 Abstract 46

4.2 Introduction 46

4.3 Extended Hazards Model 48

4.4 Estimation Procedure 48

4.5 Simulation Study 50

4.6 Real Data Analysis 52

4.7 Discussion and Conclusion 56

Chapter 5 Estimation Method for Extended Hazards Mixture Cure Model 57

5.1 Abstract 57

5.2 Introduction 57

5.3 Extended Hazards Mixture Cure Model 61

5.4 Estimation Procedure 61

5.5 Simulation Study 63

5.6 Real Data Analysis 67

5.7 Discussion and Conclusion 70

Chapter 6 Conclusions and Future Work 72

Bibliography 74

List of Tables

Table 2.1 Bias, EMPSD, ESTSD and CP of $\hat{\boldsymbol{\beta}}$, and IMSE1, IMSE2 for $g_1(Z_1)$, $g_2(Z_2)$ from the 500 simulations data set under normal distribution: N(0,0.5) 20

Table 2.2 Bias, EMPSD, ESTSD and CP of $\hat{\boldsymbol{\beta}}$, and IMSE1, IMSE2 for $g_1(Z_1)$, $g_2(Z_2)$ from the 500 simulations data set under mixture normal distribution: 0.5N(0,0.5)+0.5N(0,5) 21

Table 2.3 Estimates, SD and 95% CI of estimated parameters for the PBC data from AP-AFT model, under rank-like estimation method, rank-smooth estimation method and profile likelihood based estimation method 26

Table 3.1 Bias and SE of $\hat{\psi}_1$ of 500 simulated data sets with a sample size of 200 and 400 from the E-BJ algorithm 40

Table 3.2 Bias, SE, SD and CP of $\hat{\psi}_1$ of 500 simulated data sets with a sample size of 200 and 400 from the EM algorithm 41

Table 3.3 Demographic characteristics of data of pregnancy mothers (n = 1342) 44

Table 3.4 Semiparametric AFT mixture model results: estimate and 95% confidence interval for biological efficacy ψ_1 45

Table 4.1 Bias, ESD and BSD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the EH model 51

Table 4.2 Bias, ESD and BSD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the RPH model 53

Table 4.3 Bias, ESD and BSD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the RAFT model 54

Table 4.4 Estimates, standard errors (SE), and 95% confidence intervals of estimated parameters for the Stanford heart transplant data under the EH model 55

Table 5.1 Bias, SE, SD and CP of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the EH mixture cure model 64

Table 5.2 Bias, SE, SD and CP of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the RPH mixture cure model 65

Table 5.3 Bias, SE, SD and CP of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the RAFT mixture cure model 66

Table 5.4 Estimates, SE and P values of estimated parameters for the melanoma data from the ECOG phase III clinical trial E1684 under the PH mixture cure model 68

Table 5.5 Estimates, SE and 95% confidence intervals of estimated parameters for the melanoma data from the ECOG phase III clinical trial E1684 under the EH mixture cure model 68

Table 5.6 Estimates, SE and P values of estimated parameters for the bone marrow transplant data under the AFT mixture cure model 69

Table 5.7 Estimates, SE and 95% confidence intervals of estimated parameters for the bone marrow transplant data under the EH mixture cure model 69

List of Figures

Figure 2.1 Estimated $g_1(Z_1)$ and $g_2(Z_2)$ obtained from the rank-smooth estimation method with sample size 800 and 15 % censoring rate when the error term comes from a normal distribution 23

Figure 2.2 Estimated $g_1(Z_1)$ and $g_2(Z_2)$ obtained from the profile likelihood based estimation method with sample size 800 and 15 % censoring rate when the error term comes from a normal distribution 23

Figure 2.3 Computational time patterns for rank-like estimation method, rank-smooth estimation method and profile likelihood based estimation method, under different sample size of 100, 200, 400 and 800 24

Figure 2.4 Estimated nonlinear terms g_1(bilirubin) and g_2(albumin) obtained from rank-smooth estimation method and profile likelihood based estimation method along with their 95% confidence intervals 27

Chapter 1

Introduction

1.1 Standard Survival Model

Survival data is commonly seen in many areas, such as epidemiological studies, clinical trails and biomedical science. Survival models are developed to handle survival data, and the most popular survival models are the proportional hazards (PH) model [13] and the accelerated failure time (AFT) model [7,60]. One of the special characteristics in survival data is censoring, which happens when the information about the survival time is insufficient. For example, during the study, the patients do not experience the interested event, such as death, or they are loss to follow up. Therefore, the survival time of these patients is not complete, and what we observe is the last observed time of patients, which is referred to as censoring. The right censoring is most commonly seen in survival analysis. The presence of censoring time in the survival data leads to the investigations on the estimation methods for survival models. We will review the PH model and the AFT model in this section.

PH Model

The PH model aims to directly evaluate the covariates effects on the hazard function and assumes the regression structure on the logarithm of hazard function. The PH model can be expressed as

$$\lambda_{T|\boldsymbol{Z}}(t) = \lambda_0(t)e^{\boldsymbol{\alpha}'\boldsymbol{Z}} \tag{1.1}$$

where $\lambda_0(t)$ is the baseline hazard function which is unspecified, $\lambda_{T|\boldsymbol{Z}}(t)$ is the hazard function, $\boldsymbol{Z}$ is the covariates, and $\boldsymbol{\alpha}$ is a vector of unknown parameters. The corresponding survival function is as follows:

$$S_{T|\boldsymbol{Z}}(t) = S_0(t)^{\exp(\boldsymbol{\alpha}'\boldsymbol{Z})} \tag{1.2}$$

where $S_0(t) = e^{-\int_0^t \lambda_0(u)du}$ is the survival function of baseline distribution.

The most important indicator to evaluate the effects of the covariates is the hazard ratio (HR), which can be defined as the ratio of the hazard rates related to the two levels of covariates. The estimated HR of two individuals with different covariates $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$ can be described as

$$\widehat{HR} = \frac{\lambda_{T|\boldsymbol{Z}^*}(t)}{\lambda_{T|\boldsymbol{Z}}(t)} = \frac{\lambda_0(t)e^{\hat{\boldsymbol{\alpha}}'\boldsymbol{Z}^*}}{\lambda_0(t)e^{\hat{\boldsymbol{\alpha}}'\boldsymbol{Z}}} = e^{\hat{\boldsymbol{\alpha}}'(\boldsymbol{Z}^*-\boldsymbol{Z})} \tag{1.3}$$

Based on the estimated parameters $\hat{\boldsymbol{\alpha}}$, and the known covariates: $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$, the estimated HR is a constant independent on time. If $\widehat{HR} > 1$, it means a group under condition of $\boldsymbol{Z}^*$ has a higher chance to experience event than another group under condition of $\boldsymbol{Z}$. If $\widehat{HR} = 1$, it means a group under condition of $\boldsymbol{Z}^*$ has an equivalent chance to experience event, compared to another group under condition of $\boldsymbol{Z}$. If $\widehat{HR} < 1$, it means a group under condition of $\boldsymbol{Z}^*$ has a lower chance to experience event than another group under condition of $\boldsymbol{Z}$. For example, we define the survival time T_i as the time to death of liver cancer patients, and we want to evaluate whether the surgery treatment $\boldsymbol{Z}_i$ has the significant effects on the death of patients. We assume that the estimated value of parameter $\boldsymbol{\alpha}$ is -0.6, and its corresponding P value is less than 0.05. $\boldsymbol{Z}_i = 1$ means the patients received the surgery treatment, and $\boldsymbol{Z}_i = 0$ means they did not receive any surgery treatment. Based on these conditions, $\widehat{HR}$ related to surgery treatment can be calculated as $e^{-0.6\times(1-0)} = e^{-0.6} = 0.5488 < 1$, which means that patients receiving surgery treatment have lower risk of death than those without surgery treatment.

Estimation methods for the PH model have been widely developed. Lin and Wei [32] derived the maximum partial likelihood estimators, proposed robust covariance matrix estimators, and performed the robust score tests for the PH model. Schemper and Smith [50] proposed the probability imputation technique to handle the missing values. Gray [22] developed flexible methods by using splines and applied penalized partial likelihood to estimate the parameters. Most statistical software packages, such as "coxph" in R and "Proc Phreg" in SAS, have been developed for the PH model, and they have been widely used in dealing with survival data.

AFT Model

Serving as an alternative survival model to the PH model, the AFT model is proposed to measure the covariates effects on the survival time directly. Let T be the failure time, the AFT model can be written as

$$\log(T) = \boldsymbol{\beta}'\mathbf{Z} + \varepsilon \tag{1.4}$$

where $\boldsymbol{\beta}$ is p-dimensional unknown parameters, $\boldsymbol{Z}$ denotes the $p\times 1$ possible covariates, and ε is a random error independent of $\boldsymbol{Z}$. Without the distribution assumption of ε, model (1.4) is called as the semiparametric AFT model. The corresponding survival function of failure time T can be written as

$$S(t|\boldsymbol{Z}) = S_0(te^{\boldsymbol{\beta}'\boldsymbol{Z}}) \tag{1.5}$$

where $S_0(t)$ is the baseline survival function of t.

Different from HR in the PH model, the time ratio (TR) of the two groups are often used to evaluate the effects of the covariates in the AFT model. After exponentiation transformation of equation (1.4), we can obtain the following equation

$$T = e^{\boldsymbol{\beta}'\boldsymbol{Z}}e^{\varepsilon} \tag{1.6}$$

Given $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$, the estimated TR of the two groups can be expressed as

$$\widehat{TR} = \frac{T^*}{T} = \frac{e^{\hat{\boldsymbol{\beta}}'\boldsymbol{Z}^*}e^{\varepsilon}}{e^{\hat{\boldsymbol{\beta}}'\boldsymbol{Z}}e^{\varepsilon}} = e^{\hat{\boldsymbol{\beta}}'(\boldsymbol{Z}^*-\boldsymbol{Z})} \tag{1.7}$$

Similar to $\widehat{HR}$, we can also compare the effects of two covariates on the survival time based on the value of $\widehat{TR}$. If $\widehat{TR} > 1$, it means a group under condition of $\boldsymbol{Z}^*$ has a longer survival time to interested event than another group under condition of $\boldsymbol{Z}$. If $\widehat{TR} = 1$, it means a group under condition of $\boldsymbol{Z}^*$ has an equivalent survival time to interested event, compared to another group under condition of $\boldsymbol{Z}$. If $\widehat{TR} < 1$, it means a group under condition of $\boldsymbol{Z}^*$ has a shorter survival time to interested event than another group under condition of $\boldsymbol{Z}$. For example, we define the survival time T_i as the time to death of liver cancer patients, and we want to evaluate whether the surgery treatment $\boldsymbol{Z}_i$ has the significant effects on prolonging the life of patients. We assume that the estimated value of parameter $\boldsymbol{\beta}$ is 0.8, and its corresponding P value is less than 0.05. $\boldsymbol{Z}_i = 1$ means the patients received the surgery treatment, and $\boldsymbol{Z}_i = 0$ means they did not receive any surgery treatment. Based on these conditions, $\widehat{TR}$ related to surgery treatment can be calculated as $e^{0.8\times(1-0)} = e^{0.8} = 2.2255 > 1$, which means that patients receiving surgery treatment have longer survival time than those without surgery treatment. That is to say, the surgery treatment has successfully prolonged the life of liver cancer patients.

The estimation methods for the semiparametric AFT model have been widely discussed in literature, which includes the least square method, the rank estimation method, the induced smooth estimation method, and the profile likelihood estimation method.

The least square method was first proposed by Buckley and James [7]. The least square method using the Kaplan-Meier weights was further investigated by Stute and Wang [51]. On the basis of the least square method, Huang, Ma and Xie [24] used the least absolute shrinkage and selection operator method, as well as the threshold-gradient-directed regularization method to estimate the parameters. They

also applied a bootstrap approach to estimate the variance of estimated parameters. Jin, Lin and Ying [26] utilized the Gehan rank estimator as its initial values to improve the least square method.

The rank estimation method is another important estimation method in the AFT model. Prentice [45] developed linear rank statistics for testing the regression coefficients, and proved that the log rank test was the asymptotically fully efficient rank test. Tsiatis [56] proposed the rank estimation method. The equivalence of the rank estimation method and the least square method were given by Ritov [48]. In order to overcome the difficulties of variance estimation when censoring information existed, Wei et al. [61] proposed a simple rank estimation method through considering nuisance parameters. Without assuming any parametric form for the distribution of the error terms in the AFT model, Lai et al. [30] proposed a rank estimation method for the regression analysis by use of martingale theory and a tightness lemma for stochastic integrals of multiparameter empirical processes. Yang et al. [67] developed weighted integrals of the log-rank estimating function for estimating the parameters in the AFT model, and their asymptotic covariance matrices of estimators could be estimated reliably and efficiently. Jin, Lin, Wei and Ying [25] simplified the estimation procedure of the rank estimation method based on the Gehan-type estimator, and extended it to other weight functions. Additionally, they introduced the resampling technique to estimate the variance of estimators.

In order to handle the estimation difficulties caused by the non-smoothness, Brown et al. [5, 6] proposed the induced smoothing to the Gehan-Wilcoxon weighted rank regression, which could obtain the variance directly. Zeng and Lin [70] proposed a profile likelihood method by approximating the profile likelihood function with a kernel function. The variance of estimated parameters can be easily obtained through the inverse of the second derivative of the kernel-smoothed profile likelihood function. In recent years, statistical software packages, such as "lss" in R, have also been

developed for the AFT model. These software packages have provided effective help for researchers to deal with survival data.

1.2 Standard Cure Model

In the PH model and the AFT model, we often assume that given long enough follow-up time, all patients in the studies will finally experience the interesting event, such as death or relapse of certain disease. However, along with the medical research development, not all patients will experience the event, since these patients may be cured. Since diseases can be potentially cured, this motivates researchers to develop more specific models to evaluate these interesting problems: what is the proportion of patients who may be cured? What are the risk factors which can influence the cure rate of cured patients and the failure time of uncured patients? Fortunately, Boag [4], and Berkson and Gage [3] have successfully developed the mixture cure model, which can be used to evaluate the cure rate for cured patients and the failure time for uncured patients. We will review the mixture cure model, the semiparametric PH mixture cure model and the semiparametric AFT mixture cure model in this section.

Mixture Cure Model

Let T be the survival time, $\boldsymbol{Z}$ be p-dimensional vector of covariates, and $\boldsymbol{X}$ be another p-dimensional vector of covariates independent from $\boldsymbol{Z}$, and $f(t|\boldsymbol{X},\boldsymbol{Z})$ and $S(t|\boldsymbol{X},\boldsymbol{Z})$ be the probability density function and the survival function of failure time T, respectively. Then the mixture cure model proposed by Boag [4], and Berkson and Gage [3] can be expressed as

$$S(t|\boldsymbol{X},\boldsymbol{Z}) = 1 - \pi(\boldsymbol{X}) + \pi(\boldsymbol{X})S_u(t|\boldsymbol{Z}) \tag{1.8}$$

where $\pi(\boldsymbol{X})$, which is called "incidence", is the proportion of uncured patients depending on $\boldsymbol{X}$; $S_u(t|\boldsymbol{Z})$, which is called "latency", is the survival functions of failure

time distribution of uncured patients depending on $\boldsymbol{Z}$.

Denote the density function of failure time distribution of uncured patients as $f_u(t|\boldsymbol{Z})$. Let δ_i be an indicator of censoring with $\delta_i = 1$ for the uncensored time and $\delta_i = 0$ for the censored time. Given the observed value $(t_i, \delta_i, \boldsymbol{Z}_i, \boldsymbol{X}_i)$ for the ith subject, $i = 1, 2, 3, ..., n$, the likelihood function can be written as

$$l_o \propto \prod_{i=1}^{n} \Big\{\pi(\boldsymbol{X}_i) f_u(\boldsymbol{t}_i|\boldsymbol{Z}_i)\Big\}^{\delta_i} \Big\{1 - \pi(\boldsymbol{X}_i) + \pi(\boldsymbol{X}_i) S_u(\boldsymbol{t}_i|\boldsymbol{Z}_i)\Big\}^{1-\delta_i} \tag{1.9}$$

Therefore, specifying the distribution of either $f_u(t|\boldsymbol{Z})$ or $S_u(t|\boldsymbol{Z})$ in equation (1.9) can lead to the parametric mixture model or the semiparametric mixture model.

Many estimation methods have been developed for the parametric mixture cure model. Farewell [19] employed the Weibull regression for the latency part and the logistic regression for the incidence part. Yamaguchi [66] utilized the extended family of generalized Gamma distribution for the latency, and applied the logistic function for the regression model of the surviving fraction. Peng [44] applied the generalized F distribution to a mixture model for cure rate estimation, since the generalized F mixture model could provide great flexibility to model the survival time distribution of uncured patients, as well as covariate effects on the cure rate.

The disadvantage of the parametric mixture cure model is that unsuitably strong distributional assumptions are involved. In order to improve this disadvantage, more and more researchers have committed to develop semiparametric mixture cure models and their estimation methods. Most of them have been focusing on the two important semiparametric mixture cure models: the PH mixture cure model and the AFT mixture cure model.

Semiparametric PH Mixture Cure Model

If the latency part of the mixture cure model (1.8) is modelled with PH model, the mixture cure model is called PH mixture cure model. The incidence part of

PH mixture cure model is usually modelled with logit link function, which can be expressed as

$$\pi(\boldsymbol{X}) = \frac{e^{d'\boldsymbol{X}}}{1 + e^{d'\boldsymbol{X}}} \tag{1.10}$$

where d is a row vector of unknown parameters, $\boldsymbol{X}$ is the vector of covariates. Other link functions are also used for the incidence part of the PH mixture cure model, including probit link function, which can be expressed as

$$\pi(\boldsymbol{X}) = \Phi(d'\boldsymbol{X}) \tag{1.11}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution and complementary log-log link function, which is

$$\pi(\boldsymbol{X}) = 1 - \exp(-e^{d'\boldsymbol{X}}) \tag{1.12}$$

The latency component can be described with PH model as follows:

$$\lambda_u(t|\boldsymbol{Z}) = \lambda_0(t)e^{\boldsymbol{\alpha}'\boldsymbol{Z}} \tag{1.13}$$

where $\lambda_0(t)$ is the baseline hazard function of uncured patients. Based on the PH assumption of the latency part, the survival function of uncured patients can be expressed

$$S_u(t|\boldsymbol{Z}) = S_0(t)^{\exp(\boldsymbol{\alpha}'\boldsymbol{Z})} \tag{1.14}$$

where $S_0(t)$ is the baseline survival function of uncured patients. Until now, many discussions have been focused on developing the estimation methods for the semi-parametric PH mixture cure model [18, 29, 43, 52, 53].

Semiparametric AFT Mixture Cure Model

If the latency part of the mixture cure model is modelled with AFT model, the mixture cure model is called AFT mixture cure model. AFT mixture cure model is an important alternative mixture cure model for the PH mixture cure model. Similar

to the PH mixture cure model, the incidence part of AFT mixture cure model is also modelled with logit link function, probit link function, and complementary log-log link function. The latency component can be described with AFT model as follows:

$$\log(T) = \boldsymbol{\beta}'\mathbf{Z} + \boldsymbol{\varepsilon} \tag{1.15}$$

where $\boldsymbol{\beta}$ is a p-dimensional unknown parameters, and the distribution of $\boldsymbol{\varepsilon}$ is unknown. Then the survival function of uncured patients can be expressed

$$S_u(t|\boldsymbol{Z}) = S_0(te^{\boldsymbol{\beta}'\boldsymbol{Z}}) \tag{1.16}$$

Similar to the PH mixture cure model, many estimation methods have been discussed and developed for the semiparametric AFT mixture cure model [31, 35, 64, 65, 71].

1.3 Outline of Dissertation

In Chapter 2, we focus on the AP-AFT model, which incorporates multiple nonlinear structures of covariates. We propose both rank-smooth estimation method and the profile likelihood estimation method for the AP-AFT model. We also conduct several simulation studies to evaluate the performance of our two proposed estimation methods. An example is given to illustrate the usage of our two proposed estimation methods.

In Chapter 3, we propose a profile likelihood based estimation method for the AFT mixture model with latent subgroups. That is, given the observed subgroup information for the subjects, we develop an E-step to evaluate the conditional probability of subgroup membership in the control set. Then we incorporate the profile likelihood estimation method in the maximization step to maximize the derived likelihood functions for the observed data. We also provide simulation studies results and apply our proposed estimation method to the pregnancy data.

In Chapter 4, we develop an alternative estimation method for the EH model, which has the merits of both the PH model and the AFT model. The proposed esti-

mation method aims to use monotone splines of Ramsay to approximate the baseline hazard functions in the EH model, and apply resampling techniques to evaluate the variance of parameters. Simulation studies are conducted to investigate the effectiveness of our proposed estimation method, and a real data analysis is also provided for illustration.

In Chapter 5, we propose an EH mixture cure model, which incorporates a logistic regression for the incidence part and an EH model for the latency part of mixture cure model. Based on monotone splines of Ramsay, we also propose an efficient estimation method for the EH mixture cure model. Simulation studies based on the proposed estimation method and application of proposed estimation method to the real data will be discussed.

We summarize our conclusions of this thesis and discuss the future work in Chapter 6.

Chapter 2

Semiparametric Estimations for Additive Partial Accelerated Failure Time Models

2.1 Abstract

The semiparametric additive partial accelerated failure time model is more flexible in use than the semiparametric accelerated failure time partial linear model, since it incorporates multiple nonlinear structures of covariates. Two estimation methods based on the rank-smooth method or the profile likelihood method are proposed. In the rank smooth method, the induced smooth technique is used to estimate the variance of parameters; while in the profile likelihood, the variance is approximately calculated by the secondary derivative. The simulation study shows that both approaches can produce the valid estimations. The proposed estimation methods are illustrated by the study on primary biliary cirrhosis of the liver.

2.2 Introduction

The AFT model, which regresses the logarithm of the survival time, has been popularly applied in survival analysis. In order to make the AFT model be easily used in practice, there are many discussions in its estimation procedures. Tsiatis [56] utilized the linear rank estimate technique to estimate the parameters of the AFT model, and showed that the estimates are approximately fully efficient with the appropriate weight function. Since then, many discussions are focused on the improvement of the accuracy of the rank estimation such as [30, 61, 67]. Jin et al. [25] developed rank-

based monotone estimating functions for the AFT model, and used the resampling technique to estimate the covariance matrix of parameters. However, the resampling technique is time consuming and Brown et al. [5, 6] applied an induced smoothing technique to handle the estimation difficulties due to the non-smoothness. At the same time, the profile likelihood estimation was proposed by Zeng et al. [70], which is easy to estimate the variance of parameters through the inverse of the second derivative of the kernel-smoothed profile likelihood function.

Partial linear models are widely used in regression in order to model the non-linear association between the covariate and response variable, especially when the dependence of the response on one of the covariates is not certain. The AFT partial linear model (AFT-PLM), which incorporated the nonlinear component into the AFT model was discussed in [10, 72]. Chen et al. [10] used the rank estimation method for the AFT-PLM based on stratifying the nonlinear covaraite, which ignored the nonlinear structure in estimation; Zou et al. [72] developed a rank-like estimation method for the AFT-PLM based on the penalized method which can estimate the linear and nonlinear effects simultaneously. The resampling technique was adapted in its variance estimation. Both discussions were limited to one nonlinear component, therefore, when there is more than one nonlinear component, the potential extension and performances need to be investigated.

In this chapter, we extend the partial linear model to the model with more than one nonlinear component, which is called the additive partial accelerated failure time (AP-AFT) model. In order to overcome time-consuming issues in the resampling approach, the induced smoothing technique is applied to the rank estimation approach. At the same time, we also extend the profile likelihood estimation method to the AP-AFT model. The comparison between these two approaches is evaluated through comprehensive simulation studies. The remaining sections in this chapter are organized as follows: Section 2.3 describes the AP-AFT model. Section 2.4 outlines

the rank-smooth estimation method and the profile likelihood method. Simulation studies are conducted in Section 2.5 to investigate the performance of the proposed methods. Real data analysis about primary biliary cirrhosis of the liver is discussed in Section 2.6. Finally, discussion and conclusion are made in Section 2.7.

2.3 Additive Partial Accelerated Failure Time Model

Let T be the survival time, $\boldsymbol{X}$ be the p-dimensional vector of covariates and $Z_1, \ldots, Z_L$ be L one-dimensional covariates. The AP-AFT model can be described as:

$$\log(T) = \beta_0 + \boldsymbol{X}'\boldsymbol{\beta} + \sum_{l=1}^{L} g_l(Z_l) + \varepsilon, \tag{2.1}$$

where $\boldsymbol{\beta}$ is the p-dimensional vector of regression coefficient. The AP-AFT model assumes that the covariate Z_l is related with $\log(T)$ by a centered nonparametric function $g_l(\cdot)$, $l = 1, \ldots, L$, and ε's are independent error terms with a common distribution.

Similar to the definition in Yu and Ruppert [69], we approximate $g_l(\cdot)$ by centered r_lth-degree spline function with l_S fixed knots $k_{l_1}, \ldots, k_{l_S}$ for $l = 1, \ldots, L$ under the working assumption. Then we have

$$g_l(z_l) = \boldsymbol{\pi}_l'(z_l)\boldsymbol{\alpha}_l, l = 1, \ldots, L$$

where $\boldsymbol{\pi}_l(z_l) = (B_1(z_l), \ldots, B_{N_l}(z_l))'$ is a vector of r_lth-degree B-spline basis functions and $\boldsymbol{\alpha}_l \in \boldsymbol{R}^{N_l}$ is the spline coefficient vector.

Replacing the nonparametric function $g_l(z_l)$ by $\boldsymbol{\pi}_l'(z_l)\boldsymbol{\alpha}_l$, $l = 1, \ldots, L$, the AP-AFT model (2.1) can be rewritten as

$$\log(T_i) = \boldsymbol{X}_i'\boldsymbol{\beta} + \sum_{l=1}^{L} \boldsymbol{\pi}_l'(z_{li})\boldsymbol{\alpha}_l + \varepsilon_i = \boldsymbol{D}_i'\boldsymbol{\theta} + \varepsilon_i \tag{2.2}$$

where $\boldsymbol{D}_i = (\boldsymbol{X}_i', \boldsymbol{\pi}_1'(z_{1i}), \ldots, \boldsymbol{\pi}_L'(z_{Li}))'$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_L')'$.

Let $(T_i, \delta_i, \boldsymbol{X}_i, Z_{1i}, \ldots, Z_{Li})$ denote the observed dataset for the ith individual, $i = 1, \ldots, n$, where T_i is the observed survival time, δ_i is a censoring indicator with $\delta_i = 1$ for the uncensored time and $\delta_i = 0$ for the censored one. It is common to assume the censoring is independent and noninformative about the parameters of interest.

Rank-smooth estimation method

The Gehan-rank estimation method proposed by Jin et al. [25] can be expressed as:

$$\boldsymbol{U}_G(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i(\boldsymbol{D}_i - \boldsymbol{D}_j)I(e_j(\boldsymbol{\theta}) \geq e_i(\boldsymbol{\theta})),$$

where $e_i(\boldsymbol{\theta}) = \log(T_i) - \boldsymbol{D}_i'\boldsymbol{\theta}$ and $I(\cdot)$ is an indicator function. The estimating equation is the gradient of the convex function

$$L_G(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i(e_i(\boldsymbol{\theta}) - e_j(\boldsymbol{\theta}))I(e_j(\boldsymbol{\theta}) \geq e_i(\boldsymbol{\theta})), \tag{2.3}$$

The minimization of $\boldsymbol{L}_G(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ can be carried out by the linear programming method.

To achieve a smooth fit, a penalty term is added into the loss function (2.3). The penalized loss function can be defined as

$$L_G^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i(e_i(\boldsymbol{\theta}) - e_j(\boldsymbol{\theta}))I(e_j(\boldsymbol{\theta}) \geq e_i(\boldsymbol{\theta})) + \frac{1}{2}\sum_{l=1}^{L}\lambda_l\boldsymbol{\alpha}_l'\boldsymbol{\Psi}_l\boldsymbol{\alpha}_l \tag{2.4}$$

where, for $l = 1, \ldots, L$, λ_l is the smoothing parameter of the function $g_l(\cdot)$, and $\boldsymbol{\Psi}_l$ is

a $N_l \times N_l$ matrix. According to Eilers and Marx [17], $\boldsymbol{\Psi}_l$ is a belt-shaped matrix as

$$\boldsymbol{\Psi}_l = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ . & . & . & \cdots & . & . & . \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}_{N_l \times N_l}$$

.

The corresponding penalized Gehan-rank estimating equation is

$$\boldsymbol{U}_G^*(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i(\boldsymbol{D}_i - \boldsymbol{D}_j)I(e_j(\boldsymbol{\theta}) \geq e_i(\boldsymbol{\theta})) + (\mathbf{0}, (\lambda_1\boldsymbol{\Psi}_1\boldsymbol{\alpha}_1)', \cdots, (\lambda_L\boldsymbol{\Psi}_L\boldsymbol{\alpha}_L)')'$$

where $\mathbf{0}$ is a $p \times p$ zero matrix.

The estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ can be obtained by either minimizing $L_G^*(\boldsymbol{\theta})$ or solving $\boldsymbol{U}_G^*(\boldsymbol{\theta}) = 0$, equivalently. We utilize the Nelder-Mead algorithm in obtaining the estimator, which is an option in "optim" function in R. The initial value is specified by the linear regression with respect to $\boldsymbol{D}_i$. Then, $\hat{g}_l(z_l)$ can be estimated by $\boldsymbol{\pi}_l'(z_l)\hat{\boldsymbol{\alpha}}_l$. Assuming $\lambda_l = o_p(\frac{1}{\sqrt{n}})$, $l = 1, \ldots, L$, according to the general asymptotic theory for the rank estimator, the random vector $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically distributed as a zero-mean normal.

Choice of smoothing parameters

Selecting suitable values for the smoothing parameters λ_l is crucial to good curve fitting. We define the generalized cross-validation (GCV) score [27, 46] as

$$\text{GCV}(\lambda_1, \ldots, \lambda_L) = \frac{L_G(\boldsymbol{\theta})}{(1 - df/n)^2}, \tag{2.5}$$

where $df = \text{trace}\{H(\lambda_1, \ldots, \lambda_L)\}$ is the effective degree of freedom, and $H(\lambda_1, \ldots, \lambda_L)$ $= \left(\frac{\partial^2 L_G(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \boldsymbol{\Psi}\right)^{-1}\frac{\partial^2 L_G(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}$. Here, $\boldsymbol{\Psi}$ is a $(p+\sum_{l=1}^{L} N_l)\times(p+\sum_{l=1}^{L} N_l)$ penalized matrix

defined as

$$\begin{pmatrix} \mathbf{0} & & & \\ & \lambda_1 \boldsymbol{\Psi}_1 & & \\ & & \ddots & \\ & & & \lambda_L \boldsymbol{\Psi}_L \end{pmatrix}$$

and all other elements are zeros. The best combination of smoothing parameters $\lambda_1, \ldots, \lambda_L$ will be the minimizer of the GCV score, which is

$$(\hat{\lambda}_1, \ldots, \hat{\lambda}_L) = \underset{(\lambda_1, \ldots, \lambda_L)}{\operatorname{argmin}} \mathrm{GCV}(\lambda_1, \ldots, \lambda_L).$$

In practice, the minimization can be carried out by grid search over a sequence of possible $(\lambda_1, \ldots, \lambda_L)$ values. Similar to Yu and Ruppert [69], we select λ_l over 11 grid points ranging equally from 10^{-6} to 10^7 in our study.

Variance estimation

Based on Brown et al. [5, 6], we add a perturbation $\sqrt{\frac{\boldsymbol{\Sigma}_R}{n}}\boldsymbol{Z}$ to $\boldsymbol{\theta}$, where $\boldsymbol{Z}$ is a continuous, mean zero normal random vector independent of all of the data. The smoothing rank estimating function of $\boldsymbol{\theta}$ can be naturally defined as $\tilde{U}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_R) = E\{U_G(\boldsymbol{\theta} + \sqrt{\frac{\boldsymbol{\Sigma}_R}{n}}\boldsymbol{Z})\}$, which is the expectation of the nonsmoothed estimating function with respect to $\boldsymbol{Z}$. Then, the smoothing rank estimating function reduces to

$$\tilde{U}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_R) = E_Z\{U_G(\boldsymbol{\theta} + \sqrt{\frac{\boldsymbol{\Sigma}_R}{n}}\boldsymbol{Z})\} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i(\boldsymbol{D}_i - \boldsymbol{D}_j)\Phi\left(\frac{e_j(\boldsymbol{\theta}) - e_i(\boldsymbol{\theta})}{r_{ij}}\right) \quad (2.6)$$

where $r_{ij}^2 = \frac{1}{n}(\boldsymbol{D}_i - \boldsymbol{D}_j)'\boldsymbol{\Sigma}_R(\boldsymbol{D}_i - \boldsymbol{D}_j)$ and $\Phi(\cdot)$ is the cumulative density function of standard normal distribution. When $\boldsymbol{\Sigma}_R$ is given, the smoothing rank estimating equation (2.6) is convex and continuously differentiable. The asymptotic variance of estimated parameter can be consistently estimated via $\tilde{H}^{-1}\hat{B}\tilde{H}^{-1}$ where

$$\tilde{H}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_R) = \frac{\partial \tilde{U}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_R)}{\partial \boldsymbol{\theta}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i(\boldsymbol{D}_i - \boldsymbol{D}_j)\phi(d_{ij})(\boldsymbol{D}_i - \boldsymbol{D}_j)'/r_{ij}$$

and

$$\hat{B}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_R) = \frac{1}{n^2} \sum_{i=1}^{n} \left[\delta_i \sum_{j=1}^{n} (\boldsymbol{D}_i - \boldsymbol{D}_j) \Phi(d_{ij}) \right]^{\otimes 2},$$

where $d_{ij} = \frac{e_j(\boldsymbol{\theta}) - e_i(\boldsymbol{\theta})}{r_{ij}}$, $\phi(\cdot)$ is the density function of standard normal distribution.

As suggested in Pang et al. [42], an iterative procedure can be used to simultaneously estimate the covariance, which also avoids computational challenges. Let initial $\tilde{\boldsymbol{\Sigma}}_R$ as I_p, repeatedly update $\tilde{\boldsymbol{\Sigma}}_R$ as $\tilde{H}^{-1}\hat{B}\tilde{H}^{-1}$ until convergence of $\tilde{\boldsymbol{\Sigma}}_R$ is achieved to a specified tolerance. $\tilde{\boldsymbol{\Sigma}}_R$ is the variance estimation of $\boldsymbol{\theta}$.

Algorithm

We summarize the algorithm process for the rank-smooth estimation method as follows:

Step 1: Apply the linear regression model to the data by "glm" in R to obtain the initial value of $\boldsymbol{\theta}$.

Step 2: For all the combinations of grid points $(\lambda_1, \ldots, \lambda_L)$, minimize (2.4) to obtain the estimates of $\boldsymbol{\theta}$. Calculate the GCV score through equation (2.5) based on optimized $\boldsymbol{\theta}$ at the same time. The combination of $(\lambda_1, \ldots, \lambda_L)$ which gives the minimum GCV score can be considered the best choice of $(\lambda_1, \ldots, \lambda_L)$, and the corresponding $\hat{\boldsymbol{\theta}}$ can be considered the estimates of $\boldsymbol{\theta}$.

Step 3: The approximation of variance of $\hat{\boldsymbol{\theta}}$ can be obtained through the iterative procedure mentioned in Section 2.4.

Profile Likelihood Based Estimation Method

Based on the theory of Zeng et al. [70], the kernel-smoothed profile likelihood function of (2.2) can be written as

$$\begin{aligned} L_z(\boldsymbol{\theta}) = & -\frac{1}{n}\sum_{i=1}^{n}\delta_i(D_i'\boldsymbol{\theta}+\beta_0) - \frac{1}{n}\sum_{i=1}^{n}\delta_i R_i(\boldsymbol{\theta}) \\ & +\frac{1}{n}\sum_{i=1}^{n}\delta_i \log\left\{\frac{1}{na_n}\sum_{j=1}^{n}\delta_j K\Big(\frac{R_j(\boldsymbol{\theta})-R_i(\boldsymbol{\theta})}{a_n}\Big)\right\} \\ & -\frac{1}{n}\sum_{i=1}^{n}\delta_i \log\left\{\frac{1}{n}\sum_{j=1}^{n}\int_{-\infty}^{\frac{R_j(\boldsymbol{\theta})-R_i(\boldsymbol{\theta})}{a_n}} K(s)ds\right\} \end{aligned} \tag{2.7}$$

where $R_i(\boldsymbol{\theta}) = \log(T_i) - D_i'\boldsymbol{\theta} - \beta_0$, $K(\cdot)$ is a kernel function, and a_n is the bandwidth.

In order to obtain the smoothing estimators, we also take into account incorporating a penalty term into (2.7); then the full penalized profile likelihood function can be expressed as

$$\begin{aligned} PL_z(\boldsymbol{\theta}) = L_z(\boldsymbol{\theta}) + \frac{1}{2}\sum_{l=1}^{L}\lambda_l\boldsymbol{\alpha}_l'\boldsymbol{\Psi}_l\boldsymbol{\alpha}_l = & -\frac{1}{n}\sum_{i=1}^{n}\delta_i(D_i'\boldsymbol{\theta}+\beta_0) - \frac{1}{n}\sum_{i=1}^{n}\delta_i R_i(\boldsymbol{\theta}) \\ & +\frac{1}{n}\sum_{i=1}^{n}\delta_i \log\left\{\frac{1}{na_n}\sum_{j=1}^{n}\delta_j K\Big(\frac{R_j(\boldsymbol{\theta})-R_i(\boldsymbol{\theta})}{a_n}\Big)\right\} \\ & -\frac{1}{n}\sum_{i=1}^{n}\delta_i \log\left\{\frac{1}{n}\sum_{j=1}^{n}\int_{-\infty}^{\frac{R_j(\boldsymbol{\theta})-R_i(\boldsymbol{\theta})}{a_n}} K(s)ds\right\} + \frac{1}{2}\sum_{l=1}^{L}\lambda_l\boldsymbol{\alpha}_l'\boldsymbol{\Psi}_l\boldsymbol{\alpha}_l \end{aligned} \tag{2.8}$$

Similar to the definition of penalty term in the penalized loss function (2.4), λ_l is the smoothing parameter of the function $g_l(\cdot)$, where $l = 1, \ldots, L$. $\boldsymbol{\Psi}_l$ is a belt-shaped matrix with dimension $N_l \times N_l$ [17]. The unknown parameter θ can be obtained by maximizing the penalized profile likelihood function (2.8).

Similarly to the smoothing parameters selection of the rank-smooth estimation method, the smoothing parameter λ_l can be selected through the cross-validation method [27, 46, 72]. The GCV score can be defined as

$$\mathrm{GCV}(\lambda_1, \ldots, \lambda_L) = \frac{L_z(\boldsymbol{\theta})}{(1 - df/n)^2} \tag{2.9}$$

where $df = \text{trace}\{H(\lambda_1, \ldots, \lambda_L)\}$ and $H(\lambda_1, \ldots, \lambda_L) = \left(\frac{\partial^2 L_z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} + \boldsymbol{\Psi}\right)^{-1} \frac{\partial^2 L_z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$. Here, $\boldsymbol{\Psi}$ is a $(p + \sum_{l=1}^{L} N_l) \times (p + \sum_{l=1}^{L} N_l)$ penalized matrix. After minimizing equation (2.9), we can obtain the best smooth parameter λ_l.

Compared with the rank-smooth estimation method, the profile likelihood method can be used to estimate variance of parameters directly. After selecting the optimized smooth parameter $\hat{\lambda}_l$ through the minimizing equation (2.9), we plug $\hat{\lambda}_l$ back into the kernel-smoothed profile likelihood function (2.8), and the variance of $\hat{\boldsymbol{\theta}}$ can be estimated through the inverse of the second derivative of equation (2.8), which is

$$\boldsymbol{\Sigma}_Z = \frac{1}{M} \sum_{h=1}^{M} \left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} PL_z(\boldsymbol{\theta}) \right)^{-1} \Bigg|_{\hat{\boldsymbol{\theta}}}$$

The algorithm is similar to that of the rank-smooth estimation method.

2.5 Simulation Study

We evaluate the performance of the proposed methods through simulations. The model considered is

$$\log(T_i) = 1 + x_i + z_{1i}^3 + sin(\pi z_{2i}) + \varepsilon_i, i = 1, \ldots, n,$$

where x_i is from the standard normal distribution $N(0, 1)$, and z_{1i}, z_{2i} are generated from the uniform distribution $U(0, 1)$. Two types of error distributions are considered: one is from the normal distribution with mean 0 and standard deviation 0.5; the other is from the mixture normal distribution $0.5N(0, 0.5) + 0.5N(0, 5)$. The censoring time is from the exponential distribution, which achieves 15% (light censoring) and 30% (moderate censoring) censoring rate. 500 simulation data sets are generated under sample sizes $n = 200$, $n = 400$ and $n = 800$.

We fit the data sets by the proposed methods: rank-smooth estimation method (rank-smooth) and profile likelihood based estimation method (profile). For the purpose of comparison, we also present the results from the rank-like estimation method

(rank-like) proposed by Zou et al. [72] . The bias (Bias), empirical standard deviation (EMPSD) (average of 500 simulations), estimated standard deviation (ESTSD) and the 95% coverage probability (CP) are recorded for the estimated parameter $\hat{\boldsymbol{\beta}}$. For the nonparametric functions z_1^3 denoted by $g_1(z_1)$, and $sin(\pi z_2)$ denoted by $g_2(z_2)$, the estimated integrated mean square errors (IMSE1 and IMSE2) are reported, where

$$\text{IMSE}_l = \frac{1}{n}\sum_{i=1}^{n}(\hat{g}_l(z_{li}) - g_l(z_{li}))^2, l = 1, 2.$$

The results reported in Table 2.1 and Table 2.2 are for error distributions: normal distribution $N(0, 0.5)$ and mixture normal distribution $0.5N(0, 0.5) + 0.5N(0, 5)$, respectively.

Table 2.1 Bias, EMPSD, ESTSD and CP of $\hat{\boldsymbol{\beta}}$, and IMSE1, IMSE2 for $g_1(Z_1)$, $g_2(Z_2)$ from the 500 simulations data set under normal distribution: N(0,0.5)

n	Rate (%)	Method	$\hat{\boldsymbol{\beta}}$				$g_1(Z_1)$	$g_2(Z_2)$
			Bias	EMPSD	ESTSD	CP	IMSE1	IMSE2
200	15	rank-like	-0.0006	0.0428	0.0318	0.840	0.0376	0.0459
		rank-smooth	-0.0006	0.0428	0.0572	0.962	0.0376	0.0459
		profile	0.0066	0.0459	0.0485	0.964	0.0444	0.0586
	30	rank-like	-0.0076	0.0446	0.0333	0.844	0.0558	0.0522
		rank-smooth	-0.0076	0.0446	0.0621	0.952	0.0558	0.0522
		profile	0.0073	0.0484	0.0527	0.966	0.0592	0.0679
400	15	rank-like	0.0001	0.0290	0.0287	0.952	0.0224	0.0206
		rank-smooth	0.0001	0.0290	0.0338	0.950	0.0224	0.0206
		profile	0.0064	0.0325	0.0327	0.956	0.0285	0.0407
	30	rank-like	-0.0033	0.0314	0.0307	0.936	0.0247	0.0273
		rank-smooth	-0.0033	0.0314	0.0379	0.958	0.0247	0.0273
		profile	0.0085	0.0337	0.0358	0.970	0.0309	0.0404
800	15	rank-like	-0.0005	0.0209	0.0254	0.980	0.0105	0.0121
		rank-smooth	-0.0005	0.0209	0.0226	0.954	0.0105	0.0121
		profile	0.0059	0.0215	0.0219	0.950	0.0110	0.0131
	30	rank-like	-0.0021	0.0223	0.0276	0.974	0.0140	0.0132
		rank-smooth	-0.0021	0.0223	0.0242	0.941	0.0140	0.0132
		profile	0.0089	0.0227	0.0241	0.949	0.0148	0.0130

From Table 2.1-2.2, we can see that the biases, empirical standard deviation, estimated standard deviation, IMSE1 and IMSE2 from the rank-smooth estimation

Table 2.2 Bias, EMPSD, ESTSD and CP of $\hat{\boldsymbol{\beta}}$, and IMSE1, IMSE2 for $g_1(Z_1)$, $g_2(Z_2)$ from the 500 simulations data set under mixture normal distribution: 0.5N(0,0.5)+0.5N(0,5)

n	Rate (%)	Method	$\hat{\boldsymbol{\beta}}$				$g_1(Z_1)$	$g_2(Z_2)$
			Bias	EMPSD	ESTSD	CP	IMSE1	IMSE2
200	15	rank-like	-0.0017	0.0455	0.0384	0.874	0.0753	0.1046
		rank-smooth	-0.0017	0.0455	0.0640	0.954	0.0753	0.1046
		profile	0.0070	0.0484	0.0531	0.956	0.1286	0.1618
	30	rank-like	0.0000	0.0541	0.0410	0.850	0.0984	0.0997
		rank-smooth	0.0000	0.0541	0.0752	0.968	0.0984	0.0997
		profile	0.0208	0.0587	0.0599	0.942	0.1459	0.1449
400	15	rank-like	-0.0006	0.0300	0.0354	0.970	0.0324	0.0430
		rank-smooth	-0.0006	0.0300	0.0383	0.962	0.0324	0.0430
		profile	0.0053	0.0325	0.0360	0.970	0.0486	0.0651
	30	rank-like	-0.0002	0.0371	0.0382	0.942	0.0468	0.0574
		rank-smooth	-0.0002	0.0371	0.0429	0.948	0.0468	0.0574
		profile	0.0146	0.0385	0.0391	0.950	0.0527	0.0727
800	15	rank-like	0.0000	0.0223	0.0322	0.990	0.0210	0.0206
		rank-smooth	0.0000	0.0223	0.0246	0.962	0.0210	0.0206
		profile	0.0060	0.0223	0.0235	0.964	0.0203	0.0224
	30	rank-like	0.0000	0.0246	0.0356	0.988	0.0272	0.0256
		rank-smooth	0.0000	0.0246	0.0278	0.938	0.0272	0.0256
		profile	0.0145	0.0239	0.0263	0.952	0.0263	0.0245

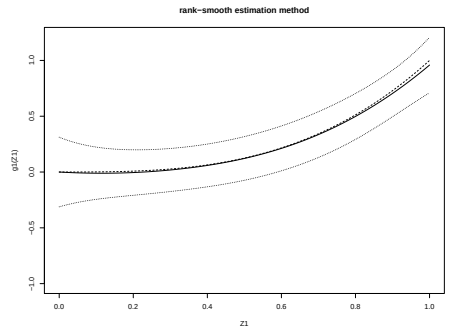
method, the profile likelihood based estimation method, and the rank-like estimation method are comparable under the normal and mixture normal distributions. It also demonstrates that the empirical standard deviation and the estimated standard deviation from both the rank-smooth estimation method and the profile likelihood based estimation method are very similar, which shows that the estimated standard deviations from both the induced smoothing technique and the inverse of the second derivative of the kernel-smoothed profile likelihood function work well. The coverage probabilities are most stable with our two proposed methods, and all their coverage probabilities are close to 95%. However, the coverage probability of the rank-like estimation method is far below from 95% when the sample size is 200, which shows that the rank-like estimation method tends to underestimate the variance when the

sample size is small.

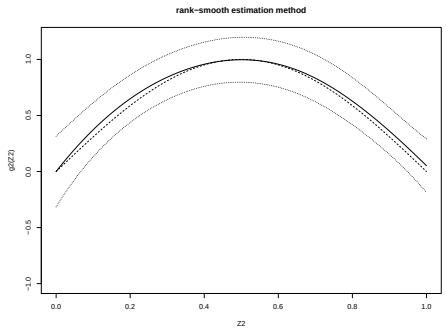
When the sample size increases, empirical standard deviation, estimated standard deviation, IMSE1 and IMSE2 tend to decrease. For example, under the normal distribution, 15% censoring rate, and 200 sample size, the empirical standard deviation, estimated standard deviation, IMSE1 and IMSE2 are (0.0428, 0.0572, 0.0376, 0.0459) from the rank-smooth estimation method, (0.0428, 0.0318, 0.0376, 0.0459) from the rank-like estimation method, and (0.0459, 0.0485, 0.0444, 0.0586) from the profile likelihood based estimation method. When the sample size increases to 400, the empirical standard deviation, estimated standard deviation, IMSE1 and IMSE2 are (0.0290, 0.0338, 0.0224, 0.0206) from the rank-smooth estimation method, (0.0290, 0.0287, 0.0224, 0.0206) from the rank-like estimation method, and (0.0325, 0.0327, 0.0285, 0.0407) from the profile likelihood based estimation method.

We further investigate the proposed methods by comparing the estimated nonparametric functions $g_1(z_1)$ and $g_2(z_2)$ along with their 95% confidence intervals with their true functions. The estimated confidence interval is obtained from the normal approximation using the empirical standard error of $\hat{g}_1(z_1)$ and $\hat{g}_2(z_2)$. For illustration purposes, we only illustrate the curve from the normal distribution with sample size 800 under 15% censoring rate. From Figure 2.1 to 2.2, we can see that the estimated curve $\hat{g}_1(z_1)$ is very close to the true curve z_1^3, $\hat{g}_2(z_2)$ is very close to $\sin(\pi z_2)$ as well, and both the estimated curves and true curves lie in the 95% confidence intervals. All the above results show that the two proposed methods are valid.

In order to compare the computational time, we model 10 sets of data with a sample size of 100, 200, 400, and 800 through the rank-like estimation method, the rank-smooth estimation method and the profile likelihood based estimation method. We record the computational time by hours, so under a sample size of 100, 200, 400, and 800, the computational time of the rank-like estimation method, rank-smooth estimation method, and profile likelihood based estimation method are (0.3953, 0.8345,

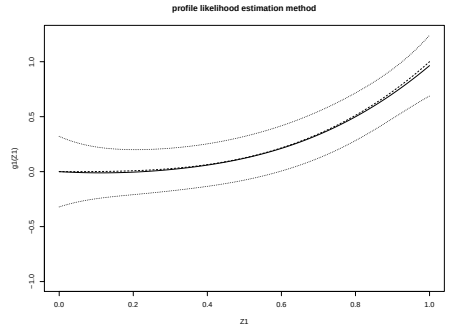


(a) 95% confidence interval from empirical standard error of $g_1(Z_1)$

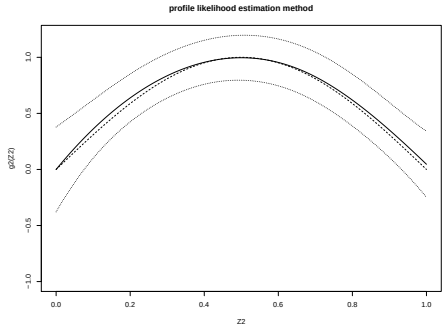


(b) 95% confidence interval from empirical standard error of $g_2(Z_2)$

Figure 2.1 Estimated $g_1(Z_1)$ and $g_2(Z_2)$ obtained from the rank-smooth estimation method with sample size 800 and 15 % censoring rate when the error term comes from a normal distribution



(a) 95% confidence interval from empirical standard error of $g_1(Z_1)$



(b) 95% confidence interval from empirical standard error of $g_2(Z_2)$

Figure 2.2 Estimated $g_1(Z_1)$ and $g_2(Z_2)$ obtained from the profile likelihood based estimation method with sample size 800 and 15 % censoring rate when the error term comes from a normal distribution

2.4297, 5.4071), (0.1678, 0.1862, 0.3297, 0.6885), (0.1962, 0.6888, 2.3972, 6.6362), respectively. The time consuming patterns are shown in Figure 2.3. From this figure, we can see that the rank-smooth estimation method saves more running time than both the profile likelihood based estimation method and the rank-like estimation method. In comparison with the rank-like estimation method, the profile likelihood based estimation method saves lots of time when the sample size is below 400. However, when the sample size is above 400, especially when the sample size is 800, the rank-like estimation method seems to save more time than the profile likelihood based

estimation method.

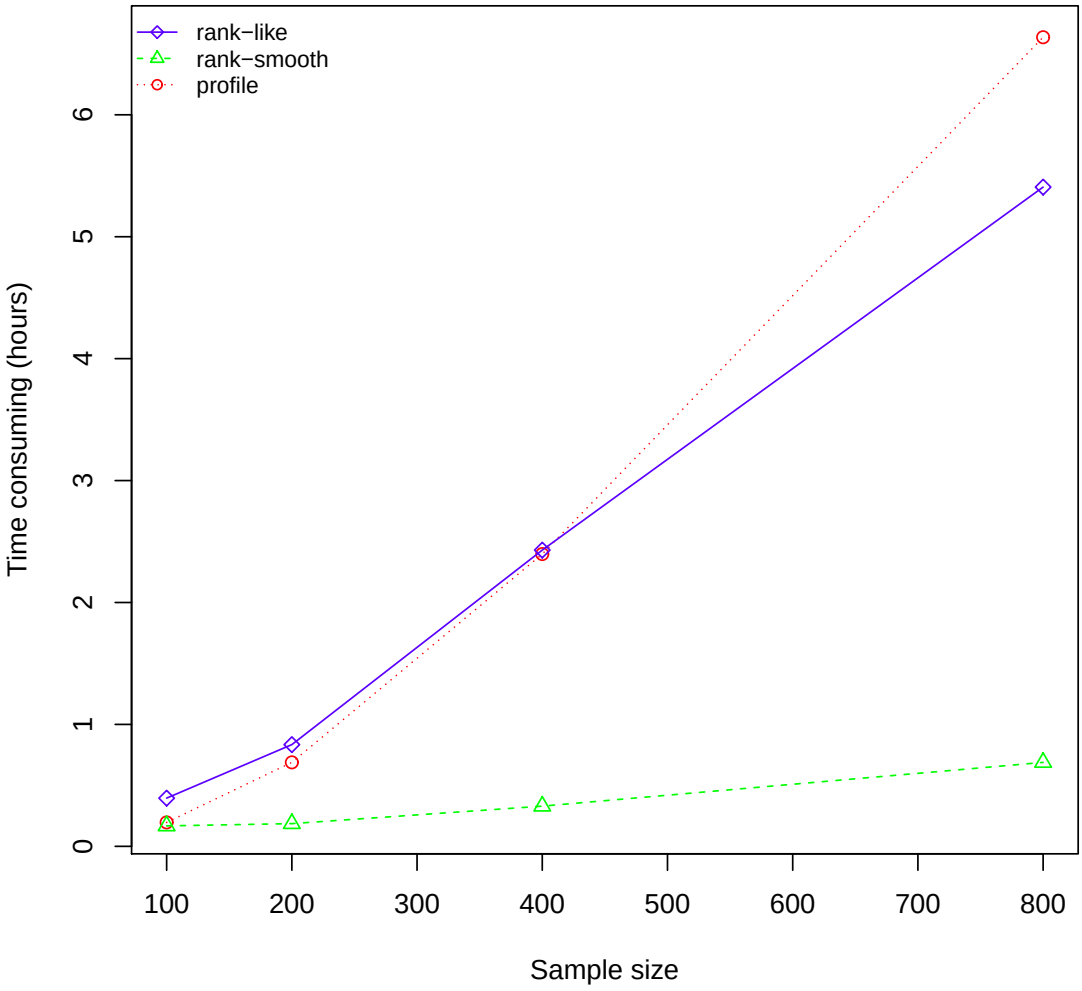


Figure 2.3 Computational time patterns for rank-like estimation method, rank-smooth estimation method and profile likelihood based estimation method, under different sample size of 100, 200, 400 and 800

2.6 Real Data Analysis

The Mayo primary biliary cirrhosis (PBC) data reported by Fleming and Harrington [20] is a double-blinded randomized trial conducted by Mayo Clinic between 1974

and 1984. Similar to [10], we define the survival time as follows: if the patients were dead, the survival time is the number of days between registration and the earlier of death; if the patients did the liver transplantation, the survival time is the number of days between registration and the liver transplantation; if the patients were alive, the survival time is the number of days between registration and the last time the patients are found in the study. If the patients were alive or did liver transplantation, we treat their status as "censored or no event"; If the patients were dead, we treat their status as "event". Other risk factors include sex (female vs. male), presence of edema (0 means no edema and no diuretic therapy for edema; 0.5 means edema present without diuretics, or edema resolved by diuretics; 1 means edema present despite diuretic therapy), level of bilirubin (mg/dl) which is a liver bile pigment, and level of albumin (mg/dl) which is a protein found in the blood.

It is hypothesized that both level of bilirubin and level of albumin would be nonlinearly related with the time to death of patients. To assess the nonlinear hypotheses, we utilized the generalized additive model (GAM), which is a well-established generalized linear model allowing for nonlinear association [23, 63]. When we model the logarithm of the time to death of patients against the smooth term of bilirubin by "gam" in R, the result shows that there is a nonlinear relationship between the logarithm of the time to death and bilirubin (estimated degree of freedom: 3.977, p-value < 0.0001). Similarly, when we model the logarithm of the time to death against the smooth term of albumin, the result also shows that albumin has a nonlinear association with the logarithm of time to death (estimated degree of freedom: 2.424, p-value < 0.0001). Therefore, we consider both bilirubin and albumin with nonlinear effects in the proposed AP-AFT model.

The AP-AFT model we consider for the PBC data is as follows:

$$\log(\text{Time}) = \beta_1 \times \text{sex} + \beta_2 \times \text{edema} + g_1(\text{bilirubin}) + g_2(\text{albumin}) + \varepsilon \qquad (2.10)$$

Table 2.3 displays the estimates, standard deviations (SD) and 95% confidence

intervals (CI) of parameters for the PBC data through fitting our proposed AP-AFT model under three estimation methods. From this table, we can see that estimates and their standard deviations are similar across the rank-like estimation method, the rank-smooth estimation method and the profile likelihood based estimation method. There are consistent conclusions for both sex and edema from the three estimation methods, that is, both sex and edema have the significant effects on the time to death of patients, since the 95% confidence intervals for either sex or edema do not include zero. The results also show that under either the rank-like estimation method or the rank-smooth estimation method, the time to death of male patients is estimated to reduce to $e^{-0.3225} = 0.7243$ of those female patients, and if edema is present among the patients, the time to death is estimated to reduce to $e^{-0.3817} = 0.6827$ of those without edema. Under the profile likelihood based estimation method, the time to death of male patients is estimated to reduce to $e^{-0.5241} = 0.5921$ of those female patients, and the time to death of patients with edema is estimated to reduce to $e^{-0.4228} = 0.6552$ of those without edema.

Table 2.3 Estimates, SD and 95% CI of estimated parameters for the PBC data from AP-AFT model, under rank-like estimation method, rank-smooth estimation method and profile likelihood based estimation method

Methods	Parameters	Estimate	Standard Deviation	95% CI
rank-like	sex	0.3225	0.1300	(0.0678, 0.5773)
	edema	-0.3817	0.0864	(-0.5510, -0.2124)
rank-smooth	sex	0.3225	0.1311	(0.0655, 0.5795)
	edema	-0.3817	0.0637	(-0.5066, -0.2568)
profile	sex	0.5241	0.1487	(0.2327, 0.8155)
	edema	-0.4228	0.0946	(-0.6082, -0.2374)

Figure 2.4 shows the nonlinear effects of the risk factors on the time to death of patients. Both bilirubin and albumin have the nonliner significant effects on the time to death of patients. Along with the increase of bilirubin, the time to death of patients nonlinearly decreases. That is to say, patients with high level of bilirubin

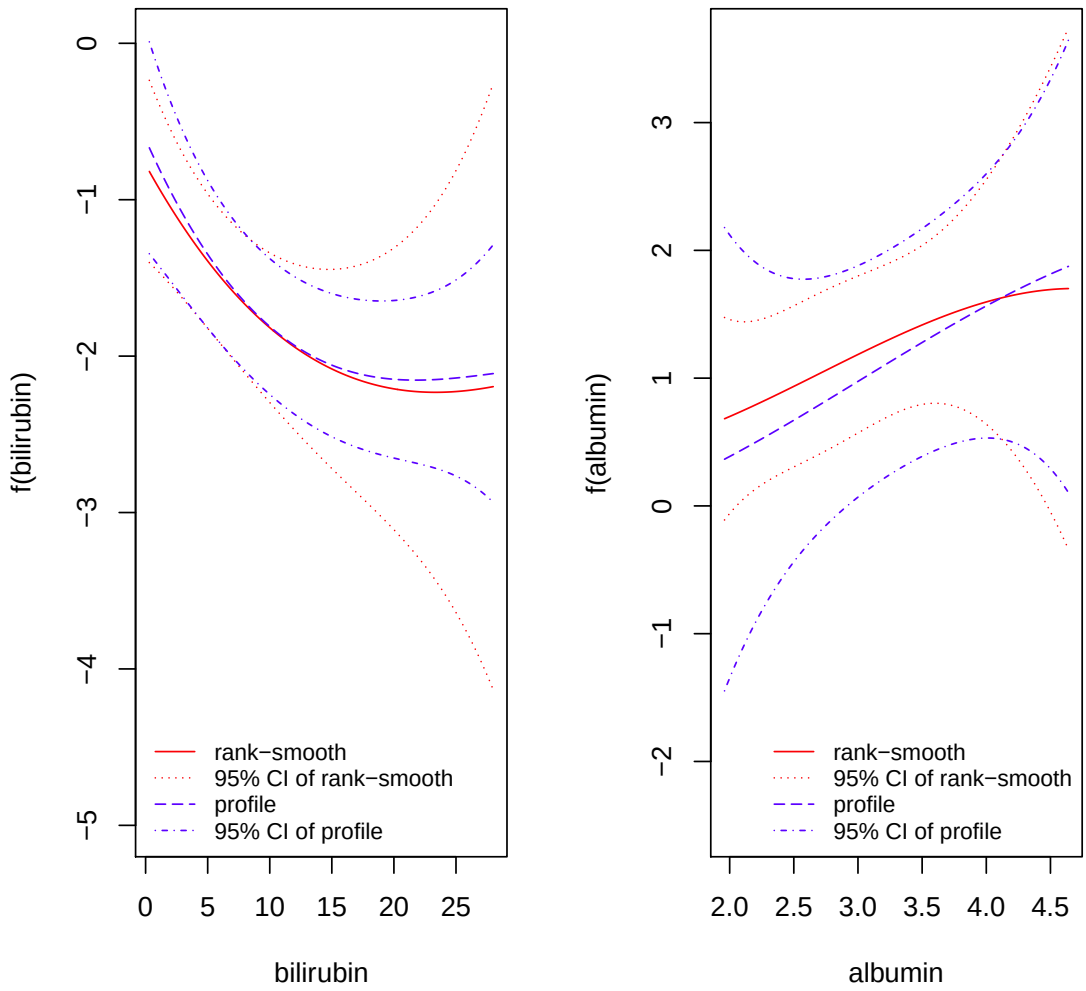


Figure 2.4 Estimated nonlinear terms g_1(bilirubin) and g_2(albumin) obtained from rank-smooth estimation method and profile likelihood based estimation method along with their 95% confidence intervals

tend to have higher death probability than those with low level of bilirubin. We also find that when the level of albumin increases, the time to death nonlinearly increases.

2.7 Discussion and Conclusion

In this chapter, we have proposed an additive partial AFT model, and its corresponding estimation methods based on either the rank-smooth method or the profile likelihood method. The simulation studies show the good performance of the proposed methods. The main difference between the rank-smooth estimation method and the rank-like estimation method is that the variance of the estimated parameters are different. The rank-like estimation method used the the resampling techniques while the other one used the induced smooth techniques. Furthermore, comparing with the rank-smooth estimation method, the variance estimation of parameters based on the profile likelihood based method is easy and straightforward, since the variance of estimated parameters can be obtained through the inverse of the second derivative of the kernel smoothed profile likelihood function. However, with the increase of sample size, it may need more time than the rank-smooth method. In summary, we suggest to using either the rank-smooth method or the profile likelihood method in practice when sample size is moderate, and otherwise, the rank-smooth estimation method is recommended.

Chapter 3

Profile Likelihood based Estimation Method for the Accelerated Failure Time Mixture Model with Latent Subgroups

3.1 Abstract

In randomized clinical trials, subgroup analysis is very important and popular. Latent subgroups arise since we can only know the subgroup membership in the treatment set, and we know nothing about subgroup membership in the control set. Biological efficacy, related to the subgroup of patients who benefit from the treatment, is an important index for researchers to evaluate the treatment effects in the treatment set. In this chapter, we develop a new estimation method for the semiparametric accelerated failure time mixture model with a latent subgroup based on the expectation-maximization algorithm and the profile likelihood estimation method. Simulation studies show that the proposed method is comparable to the existing E-BJ algorithm, which incorporates a weighted Buckley-James optimization in the maximization step. For illustration, we apply the proposed method to pregnancy data obtained from Medicaid billing records, birth certificates, Department of Education, and Department of Disabilities and Special Needs in South Carolina.

3.2 Introduction

The PH model and the AFT model are the two most popular survival models in fitting the right censored data. There are broad investigations into the estimation methods for the AFT model. Miller [40] proposed the least square estimation method to estimate the parameters in the AFT model. Buckley and James [7] developed a least square estimation method based on the modified normal equations. Jin, Lin and Ying [26] proposed one least square estimation method based on the Buckley-James estimating equation. Tsiatis [56] used the linear rank test technique to estimate the parameters of the AFT model. Wei, Ying and Lin [61] also used the linear rank statistics as estimating functions for the regression parameters. Zeng et al. [70] proposed an approximate nonparametric maximum likelihood method for the AFT model. This method can be easily used to estimate the variance of parameters through the inverse of the second derivative of the kernel-smoothed profile likelihood function.

In randomized clinical trials, subgroup analysis is very important and popular, since it pertains to the assessment of treatment effects for a specific end point in subgroups of patients defined by baseline characteristics [58]. We focus on the situation where patients who are treatable or not treatable do not receive any allocated treatment in the control set. In comparison, patients in the treatment set may or may not receive the treatment based on the baseline characteristics. For example, based on race (African American and White), participants are randomized into two sets: treatment set (all are African American) and control set (all are White) to test prostate cancer. The participants with positive prostate biopsy are classified as treatable subjects, and untreatable subjects are those without positive prostate biopsy. Whether the participants are treatable subjects or not is not influenced by the randomization. Treatable subjects in the treatment set belong to treatable subgroup A. The untreatable subjects in the treatment set belong to untreatable subgroup B. Whether the participants in the control set are treatable subjects or not is unknown, so the sub-

group information in the control set is unknown. Treatable subjects in the treatment set will receive treatment of prostate surgery; in comparison, there is no treatment for the untreatable subjects in the treatment set, and no treatment for any subjects in the control set. The latent subgroup arises since we can only know the information of prostate biopsy in the treatment set, and we know nothing about prostate biopsy in the control set. Biological efficacy, related to the subgroup of patients who benefit from the treatment, is an important index for researchers to evaluate the treatment effects in the treatment set.

When we incorporate the biological efficacy into the PH model, the PH model becomes to the PH mixture model. Follmann [21] proposed a PH model and propensity score approach to examine the effect of treatment, and used the nonparametric bootstrap to calculate standard errors of estimated parameters. Loeys and Goetghebeur [34] developed a PH model with treatment effect on the treated subgroup, derived an estimating equation, and applied a jackknife to estimate the variance of parameters. Cuzick et al. [14] developed a PH model with non-compliance and contamination, and applied the Mantel-Haenszel approach to estimate the treatment effects.

In comparison, when we incorporate the biological efficacy into the AFT model, the AFT model becomes to the AFT mixture model. Robins and Tsiatis [49] proposed correcting for non-compliance and used rank estimators to estimate the parameters. Altstein et al. [2] proposed a parametric estimation method for evaluating the biological efficacy based on the expectation-maximization (EM) algorithm. The results showed that their methods perform well when the parametric assumption is right for the real situation. However, when the parametric assumption of the AFT mixture model is violated from the real situation, the paramatric estimation method is unstable. Therefore, Altstein et al. [1] proposed a semiparametric estimation method for the AFT mixture model without the parametric assumption. Based on the Buckley-James estimator, they derived an EM algorithm, called E-BJ algorithm, to estimate

the the biological efficacy.

Motivated by the profile likelihood estimation method proposed by Zeng et al. [70], which is an approximate nonparametric maximum likelihood method that estimates parameters and is very convenient to estimate variance of estimated parameters, we propose an alternative EM algorithm to estimate the biological efficacy. That is, given the observed subgroup information for the subjects, we develop an E-step to evaluate the conditional probability of subgroup membership in the control set. Then we incorporate the profile likelihood estimation method in the maximization step to maximize the derived likelihood functions for the observed data.

The remainder of this chapter is organized as follows: Section 3.3 describes the semiparametric AFT mixture model. Section 3.4 outlines the profile likelihood based estimation method. Simulation studies are conducted in Section 3.5 to investigate the performance of the proposed EM algorithm. A real data analysis about pregnancy mothers is discussed in Section 3.6. Finally, discussion of the results and conclusions will be made in Section 3.7.

3.3 Semiparametric AFT mixture model

Let $Y_i = \min(T_i, C_i)$ denote the observed time with $i = 1, 2, ..., n$, where T_i and C_i are the failure time and censoring time for subject i, respectively. Let δ_i be the censoring indicator with 1 if $T_i \leq C_i$ and 0 otherwise, R_i be the randomization assignment ($R_i = 1$ for treatment set and $R_i = 0$ for control set) and G_i be the subgroup indicator. We assume only two subgroups ($G_i = g, g = 0, 1$) in our study: one is treatable subgroup of patients (g = 1), and another is untreatable subgroup of patients (g = 0). Given the observed covariates $\mathbf{Z}_i = (Z_{1i}, Z_{2i}, ..., Z_{qi})'$, and the randomization assignment R_i, we assume that T_i and C_i are independent. Let $\mathbf{O}$ denote the observed data, including $(y_i, \delta_i, \mathbf{z}_i, g_i; i : R_i = 1)$ for the treatment set, and $(y_i, \delta_i, \mathbf{z}_i; i : R_i = 0)$ for the control set. Then the semiparametric accelerated failure

time (AFT) mixture model [1] is given by

$$\log(T_i|(G_i = g)) = \psi_g R_i + \mathbf{Z}_i'\boldsymbol{\beta}_g + \epsilon_{gi} \tag{3.1}$$

where ψ_g is the biological efficacy, ψ_1 is unknown, and $\psi_0 \equiv 0$, that means treatment only has effect on the treatable subgroup of patients (g = 1) and no effect on the untreatable subgroup (g = 0). $\boldsymbol{\beta}_g$ is a vector of unknown regression parameters, and the distributions of the error terms ϵ_g at $g = 0$ or 1 are unknown here.

Let $f_g(\cdot)$ be the density function of $t_i e^{-\psi_g R_i - \mathbf{Z}_i'\boldsymbol{\beta}_g}$, and $S_g(\cdot)$ be the corresponding survival function. Then the completed likelihood function for the observed data $\mathbf{O}$ is

$$\begin{aligned}
&L(p, \psi_1, \beta_1, \beta_0, S_1, S_0|\mathbf{O}) = \\
&\prod_{i:R_i=1} \{p(e^{-\psi_1 R_i - \mathbf{Z}_i'\boldsymbol{\beta}_1} f_1(t_i e^{-\psi_1 R_i - \mathbf{Z}_i'\boldsymbol{\beta}_1}))^{\delta_i} S_1(t_i e^{-\psi_1 R_i - \mathbf{Z}_i'\boldsymbol{\beta}_1})^{1-\delta_i}\}^{g_i} \\
&\times \{(1-p)(e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0} f_0(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0}))^{\delta_i} S_0(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0})^{1-\delta_i}\}^{1-g_i} \\
&\times \prod_{i:R_i=0} \{p e^{-\mathbf{Z}_i'\boldsymbol{\beta}_1} f_1(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_1}) + (1-p) e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0} f_0(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0})\}^{\delta_i} \\
&\times \{p S_1(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_1}) + (1-p) S_0(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0})\}^{1-\delta_i}
\end{aligned} \tag{3.2}$$

where p is the population proportion of the treatable subgroup, that is, $p = P(G_i = 1)$. Since subgroup status g_i is unkonwn in the control set ($R_i = 0$), the likelihood function (3.2) cannot be maximized directly.

3.4 Estimation Procedure

We assume the subgroup indicator g_i in the control set is observed. Based on the data of $\mathbf{O}^p = (y_i, \delta_i, \mathbf{z}_i, g_i; i : R_i = 0)$, the completed likelihood function can be rewritten as

$$\begin{aligned}
&L(p, \psi_1, \beta_1, \beta_0, S_1, S_0|\mathbf{O}^p) = \\
&\prod_{i=1}^{n} p^{g_i}(1-p)^{1-g_i} \{(e^{-\psi_1 R_i - \mathbf{Z}_i'\boldsymbol{\beta}_1} h_1(t_i e^{-\psi_1 R_i - \mathbf{Z}_i'\boldsymbol{\beta}_1}))^{\delta_i} \exp(-H_1(t_i e^{-\psi_1 R_i - \mathbf{Z}_i'\boldsymbol{\beta}_1}))\}^{g_i} \\
&\times \{(e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0} h_0(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0}))^{\delta_i} \exp(-H_0(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0}))\}^{1-g_i}
\end{aligned} \tag{3.3}$$

where $h_g(\cdot)$ is the baseline hazard function, and $H_g(\cdot)$ is the cumulative hazard function for subgroup g corresponding to $f_g(\cdot)$. The logarithm of equation (3.3) can be written as $l(p, \psi_1, \beta_1, \beta_0, S_1, S_0|\mathbf{O}^p) = l(p|\mathbf{O}^p) + l(\psi_1, \beta_1, h_1, H_1|\mathbf{O}^p) + l(\beta_0, h_0, H_0|\mathbf{O}^p)$, where

$$l(p|\mathbf{O}^p) = \sum_{i=1}^{n} \{g_i \log(p) + (1 - g_i) \log(1 - p)\} \tag{3.4}$$

$$\begin{aligned} l(\psi_1, \beta_1, h_1, H_1|\mathbf{O}^p) = \sum_{i=1}^{n} \{ & -g_i\delta_i(\psi_1 R_i + \mathbf{Z}_i'\boldsymbol{\beta}_1) + g_i\delta_i \log(h_1(t_i e^{-\psi_1 R_i - \mathbf{Z}_i'\beta_1})) \\ & - g_i H_1(t_i e^{-\psi_1 R_i - \mathbf{Z}_i'\beta_1})\} \end{aligned} \tag{3.5}$$

$$\begin{aligned} l(\beta_0, h_0, H_0|\mathbf{O}^p) = \sum_{i=1}^{n} \{ & -(1 - g_i)\delta_i(\mathbf{Z}_i'\boldsymbol{\beta}_0) + (1 - g_i)\delta_i \log(h_0(t_i e^{-\mathbf{Z}_i'\beta_0})) \\ & - (1 - g_i) H_0(t_i e^{-\mathbf{Z}_i'\beta_0})\} \end{aligned} \tag{3.6}$$

EM Algorithm

E-step

Since the subgroup indicator G_i is unobserved in the control set ($R_i = 0$), the conditional expectation of G_i is computed by using E-step based on the observed data $\mathbf{O}$, and the estimated parameters $\boldsymbol{\Theta}^{(m)} = (p^{(m)}, \psi_1^{(m)}, \beta_1^{(m)}, \beta_0^{(m)}, S_1^{(m)}, S_0^{(m)})$ at the mth iteration of the M-step. For the treatment set ($R_i = 1$), since the subgroup indicator is observed, then $E(G_i|\mathbf{O}, \boldsymbol{\Theta}^{(m)}) = g_i$. The conditional expectation of G_i can be expressed as

$$\begin{aligned} w_i^{(m)} &= E(G_i|\mathbf{O}, \boldsymbol{\Theta}^{(m)}) \\ &= \begin{cases} \delta_i \frac{p^{(m)} h_1^{(m)}(\epsilon_{1i}^{(m)}) S_1^{(m)}(\epsilon_{1i}^{(m)})}{p^{(m)} h_1^{(m)}(\epsilon_{1i}^{(m)}) S_1^{(m)}(\epsilon_{1i}) + (1 - p^{(m)}) h_0^{(m)}(\epsilon_{0i}^{(m)}) S_0^{(m)}(\epsilon_{0i}^{(m)})} \\ +(1 - \delta_i) \frac{p^{(m)} S_1^{(m)}(\epsilon_{1i}^{(m)})}{p^{(m)} S_1^{(m)}(\epsilon_{1i}^{(m)}) + (1 - p^{(m)}) S_0^{(m)}(\epsilon_{0i}^{(m)})}, R_i = 0 \\ g_i, R_i = 1 \end{cases} \end{aligned} \tag{3.7}$$

where $\epsilon_{0i}^{(m)} = y_i - \mathbf{Z}_i'\boldsymbol{\beta}_0^{(m)}$, and $\epsilon_{1i}^{(m)} = y_i - \psi_1^{(m)} R_i - \mathbf{Z}_i'\boldsymbol{\beta}_1^{(m)}$. Replace g_i with $w_i^{(m)}$ in the logarithm of completed likelihood functions (3.4), (3.5) and (3.6), we obtain

$$E(l(p^{(m+1)}|\mathbf{O}^p)) = \sum_{i=1}^{n}\{w_i^{(m)}\log(p) + (1-w_i^{(m)})\log(1-p)\} \tag{3.8}$$

$$\begin{aligned} E(l(\psi_1^{(m+1)}, \beta_1^{(m+1)}, h_1^{(m+1)}, H_1^{(m+1)}|\mathbf{O}^p)) &= \sum_{i=1}^{n}\{-w_i^{(m)}\delta_i(\psi_1^{(m)}R_i + \mathbf{Z}_i'\boldsymbol{\beta}_1^{(m)}) \\ &+ w_i^{(m)}\delta_i \log(h_1^{(m)}(t_i e^{-\psi_1^{(m)}R_i - \mathbf{Z}_i'\boldsymbol{\beta}_1^{(m)}})) - w_i^{(m)} H_1^{(m)}(t_i e^{-\psi_1^{(m)}R_i - \mathbf{Z}_i'\boldsymbol{\beta}_1^{(m)}})\} \end{aligned} \tag{3.9}$$

$$\begin{aligned} E(l(\beta_0^{(m+1)}, h_0^{(m+1)}, H_0^{(m+1)}|\mathbf{O}^p)) &= \sum_{i=1}^{n}\{-(1-w_i^{(m)})\delta_i(\mathbf{Z}_i'\boldsymbol{\beta}_0^{(m)}) \\ &+ (1-w_i^{(m)}) \times \delta_i \log(h_0^{(m)}(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0^{(m)}})) - (1-w_i^{(m)}) H_0^{(m)}(t_i e^{-\mathbf{Z}_i'\boldsymbol{\beta}_0^{(m)}})\} \end{aligned} \tag{3.10}$$

Therefore, the purpose of M-step is to maximize the completed likelihood function (3.3), which is equivalent to maximize the likelihood functions (3.8), (3.9) and (3.10) separately.

Maximization-step

Similar to the E-BJ algorithm [1], the estimation of p can be obtained by maximizing equation (3.8), that is,

$$p^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(m)} \tag{3.11}$$

Since the baseline hazard function $h_1^{(m)}(\cdot)$ in equation (3.9) and the baseline hazard function $h_0^{(m)}(\cdot)$ in equation (3.10) are not specified, it is hard for us to maximize the likelihood functions (3.9) and (3.10) directly. Motivated by the kernel-smoothed profile likelihood estimation method proposed by Zeng and Lin [70], we assume the piecewise hazard functions to deal with the challenge of unspecified baseline hazard functions and apply the kernel-smoothed functions to approximate the likelihood functions.

Let $M_g^{(m)}$ denote an upper bound for the $t_i e^{-\psi_g^{(m)}R_i - \mathbf{Z}_i'\boldsymbol{\beta}_g^{(m)}}$ over all possible $\psi_g^{(m)\prime}s$ and $\boldsymbol{\beta}_g^{(m)\prime}s$ in a bounded set, $g = 0, 1$. An interval $[0, M_g^{(m)}]$ can be partitioned into

$J_n^{(m)}$ equally spaced internals, $0 \equiv t_{g0} < t_{g1} < ... < t_{gJ_n^{(m)}} \equiv M_g^{(m)}$. A piecewise assumption of the hazard function, denoted by $h_g^{(m)}(t)$, can be written as

$$h_g^{(m)}(t) = \sum_{k=1}^{J_n^{(m)}} C_{gk}^{(m)} I(t \in [t_{g(k-1)}, t_{gk})), g = 0, 1$$

The corresponding cumulative hazard function, denoted by $H_g^{(m)}(t)$, is

$$H_g^{(m)}(t) = \sum_{k=1}^{J_n^{(m)}} (t - t_{gk}) C_{gk}^{(m)} I(t_{k-1} \leq t < t_k) + \frac{M_g^{(m)}}{J_n^{(m)}} \sum_{k=1}^{J_n^{(m)}} C_{gk}^{(m)} I(t \geq t_{gk}), g = 0, 1$$

For the treatable subgroup (g = 1), the logarithm of the likelihood function of (3.9) can be written as

$$\begin{aligned}
&E(l(\psi_1^{(m+1)}, \beta_1^{(m+1)}, h_1^{(m+1)}, H_1^{(m+1)} | \mathbf{O}^p)) = \sum_{i=1}^{n} (-w_i^{(m)} \delta_i (\psi_1^{(m)} R_i + \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}) \\
&+ \sum_{k=1}^{J_n^{(m)}} \log C_{1k}^{(m)} \times \Big\{ \sum_{i=1}^{n} \delta_i w_i^{(m)} I(t_i e^{-\psi_1^{(m)} R_i - \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}} \in [t_{1(k-1)}, t_{1k})) \Big\} \\
&- \sum_{k=1}^{J_n} C_{1k}^{(m)} \Big\{ \sum_{i=1}^{n} w_i^{(m)} (t_i e^{-\psi_1^{(m)} R_i - \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}} - t_{1k}) \\
&\times I(t_{1(k-1)} \leq t_i e^{-\psi_1^{(m)} R_i - \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}} < t_{1k}) + \frac{M_1^{(m)}}{J_n^{(m)}} \sum_{i=1}^{n} w_i^{(m)} I(t_i e^{-\psi_1^{(m)} R_i - \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}} \geq t_{1k}) \Big\}
\end{aligned}$$

By differentiating with respect to $C_{1k}^{(m)}$, and solving the score equation of $C_{1k}^{(m)}$, we obtain

$$\begin{aligned}
\hat{C}_{1k}^{(m)} &= \sum_{i=1}^{n} \delta_i w_i^{(m)} I(t_i e^{-\psi_1^{(m)} R_i - \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}} \in [t_{1(k-1)}, t_{1k})) \\
&\times \{ \sum_{i=1}^{n} w_i^{(m)} (t_i e^{-\psi_1^{(m)} R_i - \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}} - t_{1k}) I(t_{1(k-1)} \leq t_i e^{-\psi_1^{(m)} R_i - \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}} < t_{1k}) \\
&+ \frac{M_1^{(m)}}{J_n^{(m)}} \sum_{i=1}^{n} w_i^{(m)} I(t_i e^{-\psi_1^{(m)} R_i - \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}} \geq t_{1k}) \}^{-1}
\end{aligned}$$

After replacing $C_{1k}^{(m)}$ with $\hat{C}_{1k}^{(m)}$, and discarding the irrelevant components to $\psi_1^{(m)}$ and $\boldsymbol{\beta}_1^{(m)}$, the approximating likelihood function for the treatable subgroup (g = 1)

can be written as

$$
\begin{aligned}
l_{1n}(\psi_1^{(m+1)}, \boldsymbol{\beta}_1^{(m+1)}) &= -\sum_{i=1}^{n} w_i^{(m)} \delta_i (\psi_1^{(m)} R_i + \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}) - \sum_{i=1}^{n} w_i^{(m)} \delta_i \epsilon_{1i}^{(m)}(\psi_1^{(m)}, \boldsymbol{\beta}_1^{(m)}) \\
&+ \frac{1}{n} \sum_{i=1}^{n} \delta_i \log \Big[\frac{1}{n a_n} \sum_{j=1}^{n} \delta_j w_j^{(m)} K\Big(\frac{\epsilon_{1j}^{(m)}(\psi_1^{(m)}, \boldsymbol{\beta}_1^{(m)}) - \epsilon_{1i}^{(m)}(\psi_1^{(m)}, \boldsymbol{\beta}_1^{(m)})}{a_n} \Big) \Big] \\
&- \frac{1}{n} \sum_{i=1}^{n} \delta_i \log \Big[\frac{1}{n} \sum_{j=1}^{n} w_i^{(m)} \int_{-\infty}^{\frac{\epsilon_{1j}^{(m)}(\psi_1^{(m)}, \boldsymbol{\beta}_1^{(m)}) - \epsilon_{1i}^{(m)}(\psi_1^{(m)}, \boldsymbol{\beta}_1^{(m)})}{a_n}} K(s) ds \Big]
\end{aligned}
\tag{3.12}
$$

where $\epsilon_{1i}^{(m)}(\psi_1^{(m)}, \boldsymbol{\beta}_1^{(m)}) = y_i - \psi_1^{(m)} R_i - \mathbf{Z}_i' \boldsymbol{\beta}_1^{(m)}$, $K(\cdot)$ is the kernel function, and a_n is the bandwidth. Therefore, we will maximize equation (3.12) to replace maximizing equation (3.9) to obtain $\psi_1^{(m+1)}$ and $\boldsymbol{\beta}_1^{(m+1)}$.

Given the kernel-smoothed estimators $\psi_1^{(m+1)}$ and $\boldsymbol{\beta}_1^{(m+1)}$, the baseline hazard function $h_1^{(m+1)}(t)$, the cumulative hazard function $H_1^{(m+1)}(t)$, and the survival function $S_1^{(m+1)}(t)$ can be estimated by

$$
h_1^{(m+1)}(t) = \frac{\frac{1}{n a_n t} \sum_{i=1}^{n} \delta_i w_i^{(m)} K\Big(\frac{\epsilon_{1i}^{(m+1)}(\psi_1^{(m+1)}, \hat{\boldsymbol{\beta}}_1^{(m+1)}) - \log t}{a_n} \Big)}{\frac{1}{n} \sum_{i=1}^{n} w_i^{(m)} \int_{-\infty}^{\frac{\epsilon_{1i}^{(m+1)}(\psi_1^{(m+1)}, \boldsymbol{\beta}_1^{(m+1)}) - \log t}{a_n}} K(s) ds} \tag{3.13}
$$

$$
H_1^{(m+1)}(t) = \int_{-\infty}^{\log t} \frac{\frac{1}{n a_n} \sum_{i=1}^{n} \delta_i w_i^{(m)} K\Big(\frac{\epsilon_{1i}^{(m+1)}(\hat{\psi}_1^{(m+1)}, \hat{\boldsymbol{\beta}}_1^{(m+1)}) - s}{a_n} \Big)}{\frac{1}{n} \sum_{i=1}^{n} w_i^{(m)} \int_{-\infty}^{\frac{\epsilon_{1i}^{(m+1)}(\psi_1^{(m+1)}, \boldsymbol{\beta}_1^{(m+1)}) - s}{a_n}} K(u) du} ds \tag{3.14}
$$

$$
S_1^{(m+1)}(t) = \exp(-H_1^{(m+1)}(t)) \tag{3.15}
$$

Similarly, for the untreatable subgroup (g = 0), the logarithm of the likelihood function of (3.10) can be written as

$$
\begin{aligned}
E(l(\beta_0^{(m+1)}, h_0^{(m+1)}, H_0^{(m+1)} | \mathbf{O}^p)) &= \sum_{i=1}^{n} (-(1 - w_i^{(m)}) \delta_i (\mathbf{Z}_i' \boldsymbol{\beta}_0^{(m)}) + \sum_{k=1}^{J_n^{(m)}} \log C_{0k}^{(m)} \\
&\times \Big\{ \sum_{i=1}^{n} \delta_i (1 - w_i^{(m)}) I(t_i e^{-\mathbf{Z}_i' \beta_0^{(m)}} \in [t_{0(k-1)}, t_{0k})) \Big\} - \sum_{k=1}^{J_n} C_{0k}^{(m)} \Big\{ \sum_{i=1}^{n} (1 - w_i^{(m)}) \\
&\times (t_i e^{-\mathbf{Z}_i' \beta_0^{(m)}} - t_{0k}) \times I(t_{0(k-1)} \le t_i e^{-\mathbf{Z}_i' \beta_0^{(m)}} < t_{0k}) \\
&+ \frac{M_0^{(m)}}{J_n^{(m)}} \sum_{i=1}^{n} (1 - w_i^{(m)}) I(t_i e^{-\mathbf{Z}_i' \beta_0^{(m)}} \ge t_{0k}) \Big\}
\end{aligned}
$$

By differentiating with respect to $C_{0k}^{(m)}$, and solving the score equation of $C_{0k}^{(m)}$, we obtain

$$\begin{aligned}\hat{C}_{0k}^{(m)} &= \sum_{i=1}^{n} \delta_i (1 - w_i^{(m)}) I(t_i e^{-\mathbf{Z}_i' \beta_0^{(m)}} \in [t_{0(k-1)}, t_{0k})) \\ &\times \{ \sum_{i=1}^{n} (1 - w_i^{(m)}) (t_i e^{-\mathbf{Z}_i' \beta_0^{(m)}} - t_{0k}) I(t_{0(k-1)} \leq t_i e^{-\mathbf{Z}_i' \beta_0^{(m)}} < t_{0k}) \\ &+ \frac{M_0^{(m)}}{J_n^{(m)}} \sum_{i=1}^{n} (1 - w_i^{(m)}) I(t_i e^{-\mathbf{Z}_i' \beta_0^{(m)}} \geq t_{0k}) \}^{-1}\end{aligned}$$

After replacing $C_{0k}^{(m)}$ with $\hat{C}_{0k}^{(m)}$, and discarding the irrelevant components to $\boldsymbol{\beta}_0^{(m)}$, the approximating likelihood function for the untreatable subgroup (g = 0) can be written as

$$\begin{aligned}l_{0n}(\boldsymbol{\beta}_0^{(m+1)}) = &- \sum_{i=1}^{n} (1 - w_i^{(m)}) \delta_i (\mathbf{Z}_i' \boldsymbol{\beta}_0^{(m)}) - \sum_{i=1}^{n} (1 - w_i^{(m)}) \delta_i \epsilon_{0i}^{(m)} (\boldsymbol{\beta}_0^{(m)}) \\ &+ \frac{1}{n} \sum_{i=1}^{n} \delta_i \log \Big[\frac{1}{n a_n} \sum_{j=1}^{n} \delta_j (1 - w_j^{(m)}) K\Big(\frac{\epsilon_{0j}^{(m)}(\boldsymbol{\beta}_0^{(m)}) - \epsilon_{0i}^{(m)}(\boldsymbol{\beta}_0^{(m)})}{a_n} \Big) \Big] \\ &- \frac{1}{n} \sum_{i=1}^{n} \delta_i \log \Big[\frac{1}{n} \sum_{j=1}^{n} (1 - w_i^{(m)}) \int_{-\infty}^{\frac{\epsilon_{0j}^{(m)}(\boldsymbol{\beta}_0^{(m)}) - \epsilon_{0i}^{(m)}(\boldsymbol{\beta}_0^{(m)})}{a_n}} K(s) ds \Big]\end{aligned} \tag{3.16}$$

where $\epsilon_{0i}^{(m)}(\boldsymbol{\beta}_0^{(m)}) = y_i - \mathbf{Z}_i' \boldsymbol{\beta}_0^{(m)}$. After replacing equation (3.10) by equation (3.16), we obtain $\boldsymbol{\beta}_0^{(m+1)}$ through maximizing equation (3.16). Given $\boldsymbol{\beta}_0^{(m+1)}$, we estimate $h_0^{(m+1)}(t)$, $H_0^{(m+1)}(t)$ and $S_0^{(m+1)}(t)$ by the following equations

$$h_0^{(m+1)}(t) = \frac{\frac{1}{n a_n t} \sum_{i=1}^{n} \delta_i (1 - w_i^{(m)}) K\big(\frac{\epsilon_{0i}^{(m+1)}(\boldsymbol{\beta}_0^{(m+1)}) - \log t}{a_n} \big)}{\frac{1}{n} \sum_{i=1}^{n} (1 - w_i^{(m)}) \int_{-\infty}^{\frac{\epsilon_{0i}^{(m+1)}(\boldsymbol{\beta}_0^{(m+1)}) - \log t}{a_n}} K(s) ds} \tag{3.17}$$

$$H_0^{(m+1)}(t) = \int_{-\infty}^{\log t} \frac{\frac{1}{n a_n} \sum_{i=1}^{n} \delta_i (1 - w_i^{(m)}) K\big(\frac{\epsilon_{0i}^{(m+1)}(\boldsymbol{\beta}_0^{(m+1)}) - s}{a_n} \big)}{\frac{1}{n} \sum_{i=1}^{n} (1 - w_i^{(m)}) \int_{-\infty}^{\frac{\epsilon_{0i}^{(m+1)}(\boldsymbol{\beta}_0^{(m+1)}) - s}{a_n}} K(u) du} ds \tag{3.18}$$

$$S_0^{(m+1)}(t) = \exp(-H_0^{(m+1)}(t)) \tag{3.19}$$

We summarize the proposed EM algorithm as follows:

Step 1: Given initial values: $p^{(0)}, h_0^{(0)}, h_1^{(0)}, S_0^{(0)}, S_1^{(0)}$, we obtain the initial values $w_i^{(0)}$ by equation (3.7) for $i = 1, 2, ..., n$.

Step 2: Maximize equation (3.12) and equation (3.16) to update the estimates $\boldsymbol{\beta}_1^{(1)}, \psi_1^{(1)}$, and $\boldsymbol{\beta}_0^{(1)}$, respectively. Update $p^{(1)}$ by equation (3.11), $h_1^{(1)}(t), S_1^{(1)}(t)$ by equation (3.13)

and (3.15), and $h_0^{(1)}(t), S_0^{(1)}(t)$ by equation (3.17) and (3.19). And then update $w_i^{(1)}$ by equation (3.7).

Step 3: At the jth step, given $p^{(j-1)}, h_0^{(j-1)}(t), S_0^{(j-1)}(t), h_1^{(j-1)}(t), S_1^{(j-1)}(t)$, calculate the conditional expectation of G_i through equation (3.7).

Step 4: Maximize equation (3.12) and equation (3.16) to update $\boldsymbol{\beta}_1^{(j)}, \psi_1^{(j)}$, and $\boldsymbol{\beta}_0^{(j)}$, respectively. Update $p^{(j)}$ by equation (3.11), $h_1^{(j)}(t), S_1^{(j)}(t)$ by equation (3.13) and (3.15), and $h_0^{(j)}(t), S_0^{(j)}(t)$ by equation (3.17) and (3.19).

Step 5: Repeat Step 2 and Step 4 until convergence occurs. The convergence criterion is set to 0.001 for the sum of the square error of the estimates from the jth step and $(j+1)$th step.

Because of the existence of non-differentiable subgroups, it is hard to evaluate the variance based on the traditional variance estimation technique. Similar to E-BJ algorithm, we also use the bootstrap technique to estimate the standard errors of parameters for our proposed EM algorithm.

3.5 Simulation Study

In order to examine the performance of the proposed method, we conducted simulation studies under several settings. We will compare the results obtained from our proposed method with those from the E-BJ algorithm.

We generate 500 simulation data sets with a sample size of 200 and 400 from the AFT mixture model:

$$\log(T_i|(G_i = g)) = \psi_g R_i + \epsilon_{gi}$$

Where g is drawn from a Bernoulli distribution with success probability of population proportion of treatable subjects. Randomization assignment R is drawn from a binary variable taking 0 and 1 with equal probability (1 for the treatment set and 0 for the control set). We consider two sets of error distributions: normal distributions and extreme-value distributions. For the normal distributions, ϵ_{1i} follows the normal distribution with mean 1.5 and standard deviation 0.16, and ϵ_{0i} follows the normal distribution with mean 1.5 and

Table 3.1 Bias and SE of $\hat{\psi}_1$ of 500 simulated data sets with a sample size of 200 and 400 from the E-BJ algorithm

n	F_1	F_0	p	Rate (%)	Bias	SE
200	EV(2,0.1)	EV(2,0.3)	0.3	30	0.0346	0.1478
				50	0.0474	0.1624
			0.5	30	0.0139	0.0507
				50	0.0108	0.0713
			0.8	30	0.0001	0.0289
				50	-0.0019	0.0448
	N(1.5,0.16)	N(1.5,0.2)	0.3	30	0.0162	0.0944
				50	0.0264	0.1173
			0.5	30	0.0053	0.0557
				50	0.0029	0.0694
			0.8	30	0.0019	0.0344
				50	-0.0028	0.0488
400	EV(2,0.1)	EV(2,0.3)	0.3	30	0.0407	0.0806
				50	0.0399	0.0842
			0.5	30	0.0191	0.0309
				50	0.0194	0.0505
			0.8	30	0.0041	0.0192
				50	0.0023	0.0311
	N(1.5,0.16)	N(1.5,0.2)	0.3	30	0.0070	0.0699
				50	0.0149	0.0908
			0.5	30	0.0040	0.0414
				50	-0.0006	0.0556
			0.8	30	0.0014	0.0260
				50	-0.0030	0.0341

standard deviation 0.2. For the extreme-value distributions, ϵ_{1i} follows the extreme-value distribution with location parameter 2 and scale parameter 0.1, and ϵ_{0i} follows the extreme-value distribution with location parameter 2 and scale parameter 0.3. The censoring time is generated from the exponential distribution to achieve 30% and 50% censoring rate.

For comparisons, we fit the data sets using both E-BJ algorithm and the proposed EM algorithm. We choose the true value of the biological efficacy ψ_1 as 0.6 in our simulation study, and we also select the treatable proportion (p) as 0.3, 0.5 and 0.8 to test the robustness of our proposed method. The bias and empirical standard error (SE) of estimated biological efficacy ψ_1, under the condition of several treatable proportions, are obtained

Table 3.2 Bias, SE, SD and CP of $\hat{\psi}_1$ of 500 simulated data sets with a sample size of 200 and 400 from the EM algorithm

n	F_1	F_0	p	Rate (%)	Bias	SE	SD	CP
200	EV(2,0.1)	EV(2,0.3)	0.3	30	0.0062	0.0450	0.0457	0.958
				50	0.0294	0.0690	0.0689	0.972
			0.5	30	0.0003	0.0296	0.0293	0.962
				50	0.0186	0.0423	0.0428	0.960
			0.8	30	0.0000	0.0203	0.0205	0.972
				50	0.0114	0.0260	0.0263	0.944
	N(1.5,0.16)	N(1.5,0.2)	0.3	30	0.0029	0.0624	0.0609	0.948
				50	0.0330	0.1364	0.1287	0.978
			0.5	30	-0.0133	0.0423	0.0420	0.954
				50	0.0015	0.0681	0.0698	0.958
			0.8	30	-0.0053	0.0339	0.0335	0.950
				50	0.0081	0.0433	0.0437	0.936
400	EV(2,0.1)	EV(2,0.3)	0.3	30	0.0055	0.0305	0.0289	0.980
				50	0.0268	0.0464	0.0444	0.914
			0.5	30	0.0002	0.0199	0.0197	0.934
				50	0.0143	0.0256	0.0251	0.934
			0.8	30	0.0002	0.0134	0.0135	0.970
				50	0.0089	0.0174	0.0176	0.924
	N(1.5,0.16)	N(1.5,0.2)	0.3	30	-0.0055	0.0439	0.0437	0.962
				50	0.0269	0.0741	0.0736	0.960
			0.5	30	-0.0127	0.0306	0.0305	0.938
				50	-0.0018	0.0417	0.0415	0.954
			0.8	30	-0.0060	0.0221	0.0219	0.960
				50	0.0040	0.0255	0.0255	0.972

from E-BJ algorithm. Furthermore, we also obtain the bias, the empirical standard error (SE), the estimated standard error (SD) based on 200 bootstrap samplings for each simulation, and the coverage probability (CP) of estimated biological efficacy ψ_1 through EM algorithm. All the results are recorded in Table 3.1 and Table 3.2.

From Table 3.1-3.2, we can see that the bias from EM algorithm tends to be smaller than those from E-BJ algorithm in most cases, when the error terms are from extreme-value distributions. In comparison, the bias from EM algorithm tends to be greater than those from E-BJ algorithm in most cases, when the error terms are from normal distributions. The coverage probability obtained from our proposed EM algorithm is stable, since all of

them are close to 95%. The variance estimation of EM algorithm works well, since the SE and SD of estimated biological efficacy ψ_1 are very close to each other. The EM algorithm performs better when estimating the SE than E-BJ algorithm, since E-BJ algorithm tends to overestimate the variance of estimated biological efficacy ψ_1. Our proposed method performs better when the treatable proportion p is small, compared with the E-BJ algorithm. For example, when the sample size is 200, the error terms are from extreme-value distributions, p is 0.3, censoring rate is 30%, the Bias and SE of E-BJ algorithm and EM algorithm are (0.0346, 0.1478) and (0.0062, 0.0450), respectively.

Along with the increase of sample size from 200 to 400, both SE and SD of estimated biological efficacy ψ_1 tend to decrease. For example, under the normal distributions, when the treatable proportion p is 0.8, censoring rate is 30%, and sample size is 200, the SE of estimated biological efficacy ψ_1 is 0.0344 from E-BJ algorithm, and the SE and SD of estimated biological efficacy ψ_1 are (0.0339, 0.0335) from EM algorithm. However, when the sample size increases to 400, the SE of estimated biological efficacy ψ_1 decreases to 0.0260 from E-BJ algorithm, and the SE and SD of estimated biological efficacy ψ_1 decrease to (0.0221, 0.0219) from EM algorithm.

With the increase of the treatable proportion p, the biases or variances tend to decrease. For example, when the extreme-value distributions are assumed, the treatable proportion p is 0.3, the sample size is 200, and the censoring rate is 50%, the bias and SE for estimated biological efficacy ψ_1 are (0.0474, 0.1624) from E-BJ algorithm, and (0.0294, 0.0690) from EM algorithm. When the treatable proportion p increases to 0.5, the bias and SE for estimated biological efficacy ψ_1 are (0.0108, 0.0713) from E-BJ algorithm, and (0.0186, 0.0423) from EM algorithm.

3.6 Real Data Analysis

For illustration, we apply our proposed profile likelihood based estimation method to the pregnancy data obtained from Medicaid billing records, birth certificates, Department of Education (DOE), and Department of Disabilities and Special Needs (DDSN) in South Carolina [59]. In this cohort study of maternal and child pairs born between 2004 and

2010 in South Carolina, the total number of original subjects is 210,176. After we exclude 5,894 non singleton births from the data, the number of observation is reduced to 204,282. Similar to the inclusion criteria and exclusion criteria mentioned by Wang et al. [59], we limit the data sample size to 123,922.

Preeclampsia, a common complication of pregnancy, is defined as hypertension with proteinuria during pregnancy [15, 16, 36]. Incidences of preeclampsia are increasing and are becoming to a growing health concern. Pregnancy diabetes, another common complication of pregnancy, is defined as any degree of glucose intolerance with onset or first recognition during pregnancy [38]. A number of studies have identified that pregnancy diabetes significantly increased the risk of preeclampsia [41, 68]. Because of improved prenatal care, the risk of preeclampsia of pregnancy mothers may be decreased, even though pregnancy mothers have pregnancy diabetes. Therefore, we are interested in building one semiparametric AFT mixture model without covariates to estimate the biological efficacy of prenatal care, and evaluate its impact on the survival time to preeclampsia of mothers who have pregnancy diabetes or not. We hypothesis that prenatal care can prolong the survival time of mothers with pregnancy diabetes to be diagnosed with preeclampsia. We define the survival time in our study as follows: if mothers have preeclampsia, the survival time is time to the diagnosis of preeclampsia; if mothers do not have preeclampsia, the survival time is time to the last time the mothers are found in the study or the end of the study. Observations with missing values on race, starting time of prenatal care, age, BMI, pregnancy diabetes, and survival time to diagnosis of preeclampsia are excluded from this study. Since we are interested in whether prenatal care of mothers has significant effects on the survival time to preeclampsia of mothers, we exclude those mothers with preeclampsia before the starting of prenatal care, and also exclude those mothers whose pregnancy diabetes happened after the diagnosis of preeclampsia. Therefore, based on the above exclusion, the total number of subjects in our study is 80,930.

Mothers in our study are randomized into two groups based on prenatal care, that is, if mothers receive prenatal care, then they will be included into treatment group, otherwise, they will be in the control group. In the treatment group (with prenatal care), whether

mothers have pregnancy diabetes or not is known, but in the control group (without prenatal care), we assume that we know nothing about the pregnancy diabetes of mothers. In order to capture the association between the prenatal care and the risk to preeclampsia, the pregnancy diabetes status of mothers should be considered as the latent group. Based on the above experimental design, we have 642 mothers without prenatal care and 80,288 mothers with prenatal care in our study. In order to assure the efficiency of experimental design, we randomly select 700 observations without replacement from 80,288 mothers by "sample" in R, and combine them with 642 mothers without prenatal care for our final study. That is, the final total number of mothers used in our study is 1,342.

Table 3.3 summarizes demographic characteristics of data of pregnancy mothers. Among 70 mothers with preeclampsia, 34 of them have the prenatal care, 36 of them do not have prenatal care. Among 1272 mothers without preeclampsia, 666 of them have the prenatal care, 606 of them do not have prenatal care. 9 (12.86%) mothers with pregnancy diabetes and 25 (35.71%) mothers without pregnancy diabetes in the treatment group (with prenatal care) have been diagnosed with preeclampsia. In comparison, 36 (51.43%) mothers with or without pregnancy diabetes in the control group (without prenatal care) have been diagnosed with preeclampsia. 74 (5.82%) mothers with pregnancy diabetes and 592 (46.54%) mothers without pregnancy diabetes in the treatment group (with prenatal care), and 606 (47.64%) mothers who are in the control group (without prenatal care) do not have preeclampsia. The mean range of time to develop preeclampsia for mothers in both treatment group and control group is greater than 260 days.

Table 3.3 Demographic characteristics of data of pregnancy mothers (n = 1342)

Summary Statistics	Preeclampsia		Survival Time
	Yes (N=70)	No (N=1272)	Mean (SD)
Treatment (Prenatal Care)			
Pregnancy Diabetes	9 (12.86%)	74 (5.82%)	269.94 (9.56)
Without Pregnancy Diabetes	25 (35.71%)	592 (46.54%)	270.46 (9.74)
Control (Without Prenatal Care)	36 (51.43%)	606 (47.64%)	266.04 (14.53)

Based on the above study design, the semiparametric AFT mixture model without covariates we consider for the data of pregnancy mothers is as follows:

$$\log(T_i|(G_i = g)) = \psi_g R_i + \epsilon_{gi}$$

Where $\psi_0 \equiv 0$, g = 1 if mothers have pregnancy diabetes, g = 0 if mothers do not have pregnancy diabetes, $R_i = 1$ if mothers receive prenatal care, $R_i = 0$ if mothers do not receive prenatal care.

Table 3.4 Semiparametric AFT mixture model results: estimate and 95% confidence interval for biological efficacy ψ_1

Biological Efficacy	Estimate	Standard Deviation	95% Confidence Interval
ψ_1	0.0522	0.4120	(-0.7554, 0.8597)

Table 3.4 shows estimate, standard deviation, and its corresponding 95% confidence intervals of biological efficacy ψ_1 for testing a treatment difference. It can be seen that prenatal care does not significantly prolong the mean days to develop preeclampsia of mothers with pregnancy diabetes, since 95% confidence interval for the biological efficacy ψ_1 includes zero.

3.7 Discussion and Conclusion

In this chapter, we develop an EM algorithm based estimation method for the AFT mixture model. Our simulation study shows that our proposed method is comparable to the E-BJ algorithm based estimation method. In particular, our proposed method outperforms the existing method in estimating variance of parameters, as well as fitting the small treatable proportion survival data. The pregnancy data results show that our proposed estimation method is valid to handle the real data.

Chapter 4

Estimation Method for Extended Hazards Model

4.1 Abstract

The extended hazards model is more flexible than either the proportional hazards model or the accelerated failure time model, since it has the merits of both. We proposed an alternative estimation method for the extended hazards model by modeling baseline cumulative hazard function with monotone splines of Ramsay. Simulation studies show that the proposed estimation method performs as well as the existing profile likelihood estimation method in estimating regression parameters. The proposed monotone splines estimation method is illustrated through Stanford heart transplant data.

4.2 Introduction

Survival data is very commonly seen in disease-related studies, and survival analysis is a major tool to help researchers to identify the potential risk factors for the interesting disease, such as cancer. The main issue in accurately predicting the survival probability is how to correctly model the survival data. Survival data has its special characteristics, such as right censoring, which cannot be handled by the traditional regression models. The PH model [13] and the AFT model [7, 60] are the most popular survival models in survival analysis. The PH model assumes the regression structure on the logarithm of hazard function, while the AFT model assumes the regression structure on the time scale. Before we use the PH model to model the survival data, we often need to check the PH assumption. However, it is hard for us to check the PH assumption when the sample size is finite. When the PH

assumption is not satisfied, the AFT model becomes an alternative tool, because the AFT model has quite direct physical interpretation. Due to the PH assumption in the PH model and the linear regression form in the AFT model, there are cases where neither the PH model nor the AFT model can be applied in practice directly. Therefore, we consider an extended hazards (EH) model proposed by Ciampi et al. [12], which not only includes a nested structure of both the PH model and the AFT model, but also possesses the merits of both models.

The unspecific baseline hazard function in the EH model increases the challenge for statistical inference. Therefore, many approaches have been proposed for handling this challenge. Ciampi et al. [12] used a polynomial function to approximate the baseline hazard function, and built a likelihood function to estimate the parameters. However, this method is not always efficient. Chen and Jewell [11] utilized the counting process approach to estimate the parameters of the EH model without restricting the baseline hazard function. The non-smoothness of the estimating equation in the counting process approach may lead to some problems in statistical inference. Both Tseng et al. [55] and Tong et al. [54] developed a similar estimation method assuming the piecewise of baseline hazard function, based on the kernel-smoothed profile likelihood estimation method proposed by Zeng et al. [70]. Their methods may generate excessively smooth estimations and therefore induce bias of estimation. Furthermore, the estimation of parameters may be sensitive to the choice of bandwidths. Therefore, we propose to use monotone splines of Ramsay [47] to approximate the baseline hazard functions in the EH model, and apply resampling techniques to evaluate the variance of parameters. Our new method is easy to implement, as well as more efficient and stable in estimating the unknown parameters. Furthermore, the use of monotone splines for modeling the baseline hazard function ensures smoothness in the estimated survival function.

The organization of this chapter is as follows: Section 4.3 describes the EH model. Section 4.4 provides the details to use monotone splines of Ramsay and outlines our proposed estimation method. Simulation studies are conducted in Section 4.5 to evaluate the performance of the proposed estimation method. A real data analysis about Stanford heart

transplant data is discussed in Section 4.6. Finally, we give our discussion and conclusion in Section 4.7.

4.3 Extended Hazards Model

Let T be the survival time, the EH model proposed by Ciampi et al. [12] can be described as:

$$\lambda_{T|\boldsymbol{Z}}(t) = \lambda_0(te^{\boldsymbol{\beta}'\boldsymbol{Z}})e^{\boldsymbol{\alpha}'\boldsymbol{Z}} \tag{4.1}$$

where $\lambda_0(\cdot)$ is the unspecified baseline hazard function, $\boldsymbol{Z}$ is the p-dimensional vector of covariates, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the two $p \times 1$ vectors of unknown parameters. When $\boldsymbol{\beta} = \mathbf{0}$ then the EH model becomes the PH model, that is:

$$\lambda_{T|\boldsymbol{Z}}(t) = \lambda_0(t)e^{\boldsymbol{\alpha}'\boldsymbol{Z}} \tag{4.2}$$

When $\boldsymbol{\alpha} = \boldsymbol{\beta}$ then the EH model becomes the AFT model, that is:

$$\log(T) = -\boldsymbol{\beta}'\boldsymbol{Z} + \varepsilon \tag{4.3}$$

where ε are independent error terms with a common distribution, which is independent of $\boldsymbol{Z}$.

4.4 Estimation Procedure

Let T_i denote the failure time for the subject i, C_i be the censoring time for the subject i, $\boldsymbol{Z}_i$ be the $p \times 1$ vector of covariates, $Y_i = min(T_i, C_i)$ be the observed failure time, and $\delta_i = I(T_i \leq C_i)$ be the censoring indicator, where $i = 1, \ldots, n$. Conditional on covariates $\boldsymbol{Z}_i$, T_i is assumed to be independent of C_i. Given the observed data $\boldsymbol{O}_i = (Y_i, \delta_i, \boldsymbol{Z}_i)$, the likelihood function can be written as

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda_0|\boldsymbol{O}_i) = \prod_{i=1}^{n} \left\{\lambda_0(Y_i e^{\boldsymbol{\beta}'\boldsymbol{Z}_i})e^{\boldsymbol{\alpha}'\boldsymbol{Z}_i}\right\}^{\delta_i} \exp(-\Lambda_0(Y_i e^{\boldsymbol{\beta}'\boldsymbol{Z}_i})e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})'\boldsymbol{Z}_i}) \tag{4.4}$$

where $\lambda_0(\cdot)$ is the baseline hazard function, and $\Lambda_0(\cdot)$ is the cumulative hazard function.

Then the logarithm of equation (4.4) can be rewritten as

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda_0|\boldsymbol{O}_i) = \sum_{i=1}^{n} \delta_i \boldsymbol{\alpha}'\boldsymbol{Z}_i + \sum_{i=1}^{n} \delta_i \log\left\{\lambda_0(Y_i e^{\boldsymbol{\beta}'\boldsymbol{Z}_i})\right\} - \sum_{i=1}^{n} \left\{\Lambda_0(Y_i e^{\boldsymbol{\beta}'\boldsymbol{Z}_i})e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})'\boldsymbol{Z}_i}\right\} \tag{4.5}$$

Since the log-likelihood function (4.5) is a function of unknown parameters of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\Lambda_0(\cdot)$, we aim to estimate all of them. However, we may face challenges when we maximize the log-likelihood function (4.5) to obtain $\Lambda_0(\cdot)$, since $\Lambda_0(\cdot)$ is an unspecified, nondecreasing function with infinite dimension. Motivated by the inference of the PH model through modeling baseline cumulative hazard function with monotone splines [8, 37], we propose to use monotone splines of Ramsay [47] to model $\Lambda_0(Y_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i})$ in the equation (4.5).

The monotone splines of Ramsay [47] can be described as

$$\Lambda_0(t) = \sum_{l=1}^{k} \gamma_l I_l(t) \tag{4.6}$$

where $I_l(\cdot)$ is the integrated spline basis function, which is nondecreasing from 0 to 1, and γ_l are chosen from nonnegative value in order to keep the monotonicity of $\Lambda_0(t)$. Since $k = knots + degree - 2$, we can fully determine the k spline basis functions if the knots and degree are specified. Knots, which determine the shape of the monotone splines, can be specified through a sequence of increasing points within a finite interval of the minimum and maximum of the censoring times. Degree, which determines the smoothness of the monotone splines, can be specified with linear (degree = 1), quadratic (degree = 2) and cubic functions (degree = 3), respectively. The number of knots are very important when we use monotone splines of Ramsay [47] to model $\Lambda_0(\cdot)$, since a large number of knots may lead to the over-fitting of the $\Lambda_0(\cdot)$, and a small number of knots may cause the ill-fitting of the $\Lambda_0(\cdot)$. Recommended by Lin and Wang [33], Wang and Dunson [57], Cai et al. [8], and McMahan et al. [37], a moderate number (10-30) of equally-spaced knots will be used in our study.

After modeling $\Lambda_0(Y_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i})$ through monotone splines of Ramsay [47], the log-likelihood function (4.5) can be rewritten as

$$\begin{aligned} &l_s(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} | \boldsymbol{O}_i) = \\ &\sum_{i=1}^{n} \delta_i \boldsymbol{\alpha}' \boldsymbol{Z}_i + \sum_{i=1}^{n} \delta_i \log \{ \sum_{l=1}^{k} \gamma_l M_l(Y_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i}) \} - \sum_{i=1}^{n} \{ \sum_{l=1}^{k} \gamma_l I_l(Y_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i}) e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})' \boldsymbol{Z}_i} \} \end{aligned} \tag{4.7}$$

where $M_l(\cdot)$ is a set of basis splines. Then the unknown parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ can be obtained directly by maximizing equation (4.7). We estimate the variance of the estimated parameters based on the resampling technique.

4.5 Simulation Study

The performance of the proposed method in our study will be investigated through simulation studies. For the purpose of comparison, we also present the results from the kernel-smoothed profile likelihood estimation method proposed by Tseng et al. [55]. Similar to the study of Tseng et al. [55], we will generate 500 simulation datasets with a sample size of 200 and 400, the baseline hazard function is from a log-logistic distribution:

$$\lambda_0(t) = \frac{\frac{b}{a}(\frac{t}{a})^{b-1}}{1+(\frac{t}{a})^b} \tag{4.8}$$

where $a = 120$ and $b = 4$. In our simulation study, we choose the true value of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as $(-1,-1)'$ and $(-0.5,-0.5)'$ for the EH model, $(-1,-1)'$ and $(0,0)'$ for the EH model, called restricted PH (RPH) model, $(-1,-1)'$ and $(-1,-1)'$ for the EH model, called restricted AFT (RAFT) model, respectively. The covariates $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)'$, where $\boldsymbol{Z}_1$ is generated from Bernoulli distribution with success probability of 0.5, $\boldsymbol{Z}_2$ is generated from standard normal distribution. Censoring times are generated from an exponential distribution with different means to obtain different censoring rates. Recommended by Lin and Wang [33], Wang and Dunson [57], Cai et al. [8], and McMahan et al. [37], we will choose knots as 30, and degree as 2 to control the smoothness of the splines. $\boldsymbol{\gamma}$ will be given some equal nonnegative values. Then the unknown parameters in the EH model can be estimated based on the maximum likelihood estimation (MLE) method. The bootstrapped standard deviations (BSD) for our proposed monotone splines method is obtained by 200 repetitions.

The biases, empirical standard deviations (ESD), and BSD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the EH model, RPH model, and RAFT model are computed, and the results are shown in Tables 4.1-4.3. The results from the EH model, seen in Table 4.1, show that the performance of our proposed monotone splines method and the profile likelihood method proposed by Tseng et al. are quite similar for different censoring rates and sample sizes. The ESD and BSD of our proposed method are close, which shows that the proposed estimation method performs well in simulation settings. With respect to the ESD, the ESD from our proposed method is less than that from the profile likelihood method, which shows that our proposed method has significantly

improved the standard deviation of estimated parameters. For example, with the censoring rate of 15%, the ESD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are (0.1797, 0.1076, 0.1468, 0.0816) for the profile likelihood method, and (0.0677, 0.0402, 0.0668, 0.0578) for our proposed estimation method.

Table 4.1 Bias, ESD and BSD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the EH model

Methods	n	CR (%)	Parameters	Bias	ESD	BSD
Monotone Splines	200	15	α_1	0.0573	0.0677	0.0589
			α_2	0.0165	0.0402	0.0409
			β_1	-0.0406	0.0668	0.0530
			β_2	-0.0100	0.0578	0.0471
		35	α_1	0.0039	0.0596	0.0577
			α_2	0.0447	0.0553	0.0534
			β_1	0.0401	0.0827	0.0772
			β_2	0.0525	0.0897	0.0752
	400	15	α_1	0.0619	0.0731	0.0609
			α_2	0.0193	0.0355	0.0349
			β_1	-0.0680	0.0675	0.0500
			β_2	-0.0367	0.0548	0.0427
		35	α_1	0.0244	0.0602	0.0564
			α_2	0.0232	0.0448	0.0454
			β_1	0.0116	0.0628	0.0631
			β_2	0.0409	0.0663	0.0607
Profile Likelihood	200	15	α_1	0.0282	0.1797	
			α_2	0.0404	0.1076	-
			β_1	0.0026	0.1468	-
			β_2	-0.0104	0.0816	-
		35	α_1	0.0853	0.1946	-
			α_2	0.0837	0.1204	-
			β_1	-0.0388	0.1508	-
			β_2	-0.0347	0.0955	-
	400	15	α_1	0.0286	0.1309	-
			α_2	0.0229	0.0786	-
			β_1	-0.0093	0.1005	-
			β_2	-0.0035	0.0590	-
		35	α_1	0.0766	0.1583	-
			α_2	0.0657	0.0949	-
			β_1	-0.0305	0.1166	-
			β_2	-0.0274	0.0670	-

Table 4.2 displays the results of bias, ESD, and BSD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the RPH model. The results reveal that when the censoring rate is light (15%) and the sample size is large (n = 400), our proposed estimation method and the profile likelihood method are comparable. For example, when the censoring rate is 15% and the sample size is 400, the bias of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ are (0.0931, 0.0905, -0.0134, -0.0036) for the profile likelihood method, and (-0.0162, 0.0107, -0.0281, -0.0314) for the monotone splines method. However, when the sample size is small (n = 200) or the censoring rate is 35%, the profile likelihood method tends to have larger biases for $\hat{\alpha}_1$ and $\hat{\alpha}_2$ than that from monotone splines method. For example, when the sample size is 400 and the censoring rate is 35%, the biases of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are (0.1273, 0.1220) for the profile likelihood method, and (-0.0432, -0.0061) for our proposed monotone splines method. Similar to the EH model, the ESD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ from our proposed monotone splines method tends to be smaller than that from the profile likelihood method, which indicates that the profile likelihood method may overestimate the variances of estimated parameters. The ESD and BSD from our proposed monotone splines method are also close to each other, which shows that our proposed method performs well.

Compared to the profile likelihood method, the proposed monotone splines method also performs well in the RAFT model (seen in the Table 4.3). The biases of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ are comparable from the proposed monotone splines method and the profile likelihood method. The ESDs from the proposed monotone splines method tend to be smaller than those from the profile likelihood method, except for $\hat{\beta}_2$. The ESD and BSD from our proposed monotone splines method are close to each other, which shows that our proposed estimation method is valid to estimate the parameters in the RAFT model. For example, when the sample size is 400 and the censoring rate is 35%, the ESD and BSD of $\hat{\beta}_2$ from monotone splines method are (0.1174, 0.1135).

4.6 Real Data Analysis

For illustration, we apply our proposed monotone splines estimation method to the famous Stanford heart transplant data [39, 60]. This data includes 184 patients who had received

Table 4.2 Bias, ESD and BSD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the RPH model

Methods	n	CR (%)	Parameters	Bias	ESD	BSD
Monotone Splines	200	15	α_1	-0.0076	0.0890	0.0878
			α_2	0.0072	0.0765	0.0779
			β_1	-0.0446	0.0820	0.0775
			β_2	-0.0395	0.0471	0.0449
		35	α_1	-0.0137	0.0800	0.0759
			α_2	-0.0147	0.0669	0.0622
			β_1	-0.0442	0.0637	0.0661
			β_2	-0.0503	0.0476	0.0446
	400	15	α_1	-0.0162	0.0694	0.0761
			α_2	0.0107	0.0556	0.0654
			β_1	-0.0281	0.0533	0.0640
			β_2	-0.0314	0.0374	0.0368
		35	α_1	-0.0432	0.0907	0.0848
			α_2	-0.0061	0.0680	0.0625
			β_1	-0.0396	0.0591	0.0629
			β_2	-0.0380	0.0352	0.0367
Profile Likelihood	200	15	α_1	0.1332	0.1775	-
			α_2	0.1411	0.0975	-
			β_1	-0.0102	0.1353	-
			β_2	-0.0124	0.0818	-
		35	α_1	0.1727	0.1927	-
			α_2	0.1770	0.1356	-
			β_1	-0.0463	0.1585	-
			β_2	-0.0455	0.1050	-
	400	15	α_1	0.0931	0.1250	-
			α_2	0.0905	0.0780	-
			β_1	-0.0134	0.0899	-
			β_2	-0.0036	0.0567	-
		35	α_1	0.1273	0.1558	-
			α_2	0.1220	0.0979	-
			β_1	-0.0238	0.1201	-
			β_2	-0.0275	0.0681	-

Table 4.3 Bias, ESD and BSD of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the RAFT model

Methods	n	CR (%)	Parameters	Bias	ESD	BSD
Monotone Splines	200	15	α_1	-0.0137	0.0715	0.0654
			α_2	0.0116	0.0508	0.0482
			β_1	0.0388	0.1010	0.0908
			β_2	-0.0093	0.1175	0.1117
		35	α_1	-0.0207	0.0749	0.0668
			α_2	0.0448	0.0607	0.0526
			β_1	0.0456	0.1169	0.1038
			β_2	-0.0634	0.1524	0.1458
	400	15	α_1	0.0071	0.0598	0.0616
			α_2	0.0037	0.0361	0.0392
			β_1	0.0075	0.0819	0.0778
			β_2	-0.0146	0.0949	0.0897
		35	α_1	-0.0087	0.0669	0.0648
			α_2	0.0261	0.0510	0.0482
			β_1	0.0292	0.1003	0.0906
			β_2	-0.0523	0.1174	0.1135
Profile Likelihood	200	15	α_1	-0.0167	0.2185	-
			α_2	-0.0235	0.1145	-
			β_1	0.0139	0.1630	-
			β_2	0.0070	0.0868	-
		35	α_1	-0.0462	0.1763	-
			α_2	-0.0499	0.1045	-
			β_1	0.0173	0.1438	-
			β_2	0.0184	0.0819	-
	400	15	α_1	-0.0163	0.1285	-
			α_2	-0.0103	0.0771	-
			β_1	0.0093	0.1022	-
			β_2	0.0043	0.0607	-
		35	α_1	-0.0530	0.1335	-
			α_2	-0.0352	0.0854	-
			β_1	0.0187	0.1075	-
			β_2	0.0094	0.0658	-

heart transplants, and 27 patients are excluded from our study since they have the missing value of T5 mismatch scores. T5 mismatch scores were used to evaluate the degree of tissue incompatability between the initial donor and recipient hearts with respect to HLA antigens. Age at the time of the first transplant was recorded into the data. The survival time was defined as time to death after transplantation in days. The censoring indicator is defined as 1 if patients died, and as 0 if patients did not die. We are very interested in assessing whether age has significant effects on the survival time of patients. The standard errors of estimates from our proposed monotone splines estimation method are obtained using the bootstrap method based on 200 bootstrap samples. For comparison, we also present the results of Stanford heart transplant data from the profile likelihood method.

The model we consider for the Stanford heart transplant data is as follows:

$$\lambda_{T|age}(t) = \lambda_0(te^{\boldsymbol{\beta}'\times age})e^{\boldsymbol{\alpha}'\times age} \tag{4.9}$$

Table 4.4 Estimates, standard errors (SE), and 95% confidence intervals of estimated parameters for the Stanford heart transplant data under the EH model

Methods	Parameter	Estimate	SE	95% Confidence Interval
Profile Likelihood	$\boldsymbol{\alpha}$			
	age	0.0073	0.0198	(-0.0315, 0.0461)
	$\boldsymbol{\beta}$			
	age	-0.0421	0.0308	(-0.1025, 0.0183)
Monotone Splines	$\boldsymbol{\alpha}$			
	age	0.0197	0.0036	(0.0127, 0.0267)
	$\boldsymbol{\beta}$			
	age	-0.0463	0.0029	(-0.0521, -0.0406)

Table 4.4 displays estimates, standard errors (SE), and 95% confidence intervals of estimated parameters for the Stanford heart transplant data from both the profile likelihood method and the monotone splines method under the EH model. From the results, we can see that the estimates of the parameters from the profile likelihood estimation method are very similar to those from the monotone splines method, but the standard errors of estimated parameters from the profile likelihood estimation method tend to be higher than those from the monotone splines method. The results of our proposed monotone splines method show

that age has both the PH effect and the AFT effect, since all the 95% confidence intervals do not include zero. The PH effect ($\hat{\boldsymbol{\alpha}} = 0.0197$) may mean that, along with the increase of age at the time of the first transplant, the patients had high relative risk to be dead. The AFT effect ($\hat{\boldsymbol{\beta}}$ = -0.0463) may mean that the younger the patients were at the first time transplant, the longer they would survive after heart transplant.

4.7 Discussion and Conclusion

In this chapter, we have proposed an alternative estimation method based on monotone splines of Ramsay. The simulation studies show that our proposed monotone splines estimation method is valid and flexible. It also shows that our proposed monotone splines method is comparable to the profile likelihood estimation method. Especially, our proposed method performs well in the restricted EH model, where $\boldsymbol{\beta}$ of the EH model are set as zero. The main difference between the profile likelihood method and our proposed monotone splines method is that the profile likelihood method may tend to overestimate the variance of estimated parameters. Comparing with our proposed method, the variance estimation of parameters based on the profile likelihood method is easy and straightforward, since the variance of estimated parameters can be obtained through the inverse of the second derivative of the kernel smoothed profile likelihood function.

The real data results show that our conclusion is different from those from the profile likelihood estimation method. One possible reason for this is that the variance of estimated parameters from our proposed method is lower than those from the profile likelihood method. We believe that our proposed estimation method is valid, since our conclusions about the Stanford heart transplant data are very close to the conclusions from either the AFT model or the PH model.

Chapter 5

Estimation Method for Extended Hazards Mixture Cure Model

5.1 Abstract

We propose an extended hazards mixture cure model, which incorporates a logistic regression for the incidence part and an extended hazards model for the latency part of mixture cure model. The extended hazards mixture cure model not only retains the merits of the proportional hazards mixture cure model and the accelerated failure time mixture cure model, but also is more flexible than either the proportional hazards mixture cure model or the accelerated failure time mixture cure model. We also proposed an estimation method by modeling baseline cumulative hazard function with monotone splines of Ramsay in the latency part of the mixture cure model. Simulation studies show that the proposed estimation method performs well in estimating regression parameters. The proposed estimation method and extended hazards mixture cure model are illustrated using melanoma data from the ECOG phase III clinical trial E1684 and leukemia data from a bone marrow transplant study.

5.2 Introduction

In the survival analysis, the PH model [13] assumes the regression structure of the logarithm of hazard function, and the AFT model [7,60] assumes there is a linear relationship between the logarithm of survival time and covariates. We may need to check the PH assumption before we use the PH model. But when the sample size is small, checking PH assumption is not so easy, therefore, the model results may be biased when the PH assumption is violated.

Under this situation, the AFT model may be an alternative for the PH model, because of its direct physical interpretation. However, there must be situations where neither the PH model nor the AFT model is suitable. Therefore, Ciampi et al. [12] proposed a more flexible model, called the extended hazards (EH) model, which can be used as not only the PH model but also the AFT model based on some conditions.

The common assumption hidden in both the PH model and the AFT model is that if the follow-up time is long enough, all the patients in the studies will experience the interesting event. However, due to the development of technology, more and more patients have been cured and fewer have experienced the interesting event. For example, through curative surgical resection of tumors, the cure rate of patients with stage I lung cancer (tumor size greater than 45 mm in diameter) could be up to 43% [62]. Kuflik [28] used deep cryosurgery technique to cure patients with skin cancer, and concluded that the overall 30-year cure rate was 98.6%. That diseases can be cured will potentially motivate researchers to develop mixture cure models [3, 4].

Let T be the survival time, $\boldsymbol{X}$ be another p-dimensional vector of covariates independent from $\boldsymbol{Z}$, and $f(t|\boldsymbol{X},\boldsymbol{Z})$ and $S(t|\boldsymbol{X},\boldsymbol{Z})$ be the probability density function and the survival function of failure time T, respectively. Then the mixture cure model [3, 4] can be expressed as

$$S(t|\boldsymbol{X},\boldsymbol{Z}) = 1 - \pi(\boldsymbol{X}) + \pi(\boldsymbol{X})S_u(t|\boldsymbol{Z}) \tag{5.1}$$

where $\pi(\boldsymbol{X})$, which is called "incidence", is the proportion of uncured patients depending on $\boldsymbol{X}$; $S_u(t|\boldsymbol{Z})$, which is called "latency", is the survival probability of uncured patients depending on $\boldsymbol{Z}$. If the latency part of the mixture cure model is modelled with the PH model, the mixture cure model is called the PH mixture cure model. If the latency part of the mixture cure model is modelled with the AFT model, the mixture cure model is called the AFT mixture cure model. The PH mixture cure model and the AFT mixture cure model are the two most important survival models for the survival data with cure information.

Estimation methods for the PH mixture cure model have been widely developed by researchers in recent years. Kuk and Chen [29] combined a logistic formulation for the in-

cidence part and used the PH model to model the latency part in the mixture cure model. They maximized a Monte Carlo approximation of a marginal likelihood to estimate the parameters, and applied the expected-maximization (EM) algorithm to estimate the baseline survivor function. Their methods showed reasonable efficiency. Taylor [53] proposed a logistic regression model for the incidence part of the model, and a Kaplan-Meier type approach to estimate the latency part of the model. His methods seem to be less efficient for estimating the latency distribution. Peng and Dear [43] considered the dependence of the probability on the survival function of uncured patients, and utilized the EM algorithm, the marginal likelihood approach, and multiple imputations to estimate the parameters. Sy and Taylor [52] developed maximum likelihood techniques and the EM algorithm for the joint estimation of the incidence and latency regression parameters. They had proved that their methods were generally better than the parametric methods when the censoring rate is higher. Fang et al. [18] investigated large sample inference from the semiparametric PH mixture cure model.

With regard to the AFT mixture cure model, there are also many discussions of estimation methods. Li and Taylor [31] used an AFT model with unspecified error distribution to determine the latency, and developed an EM algorithm estimation method to estimate the unknown parameters. Zhang and Peng [71] proposed a new estimation method for the semiparametric AFT mixture cure model, and their methods employed the EM algorithm and the rank estimator of the AFT model to estimate the parameters of interest. Results showed that their proposed estimation methods improved the identifiability of the parameters, compared to the parametric estimation methods. Xu and Zhang [64] proposed an alternative estimation method by incorporating the profile likelihood into the M-step of the EM algorithm, and their method worked well for the light censoring survival data. Xu and Zhang [65] proposed a multiple imputation method based on the rank estimation method and the profile likelihood method. Lu [35] developed a kernel-smoothing-based EM algorithm for efficient estimation and derived the asymptotic properties for the resulting estimates.

When we fit the survival data with cure information, we should choose whether to use

the PH mixture cure model or the AFT mixture cure model by checking the assumptions. However, not all the survival data with cure information can be modelled by either the PH mixture cure model or the AFT mixture cure model. Therefore, we proposed an EH mixture cure model, which incorporates a logistic regression for the incidence part, and an EH model for the latency part of the mixture cure model, because the unspecific baseline hazard function of the EH model in the latency part of the mixture cure model increases the challenge for statistical inference. Therefore, many approaches have been proposed for handling this challenge. Ciampi et al. [12] used a polynomial function to approximate the baseline hazard function, and built a likelihood function to estimate the parameters. However, this method is not always efficient. Chen and Jewell [11] utilized the counting process approach to estimate the parameters of the EH model without restricting the baseline hazard function. The non-smoothness of the estimating equation in the counting process approach may lead to some problems in statistical inference. Both Tseng et al. [55] and Tong et al. [54] developed a similar estimation method assuming the piecewise of baseline hazard function, based on the kernel-smoothed profile likelihood estimation method proposed by Zeng et al. [70]. Their methods may generate excessive smoothness and therefore induce bias of estimation. Furthermore, the estimation of parameters may be sensitive to the choice of bandwidths. Therefore, the purpose of this study is to develop an EH mixture cure model, use monotone splines of Ramsay to approximate the baseline hazard functions of the EH model in the latency part of the mixture cure model, apply maximum likelihood techniques to estimate the unknown parameters, and utilize resampling techniques to evaluate the variance of estimated parameters.

We organize this chapter as follows: Section 5.3 depicts the EH mixture cure model. Section 5.4 describes the details of our proposed estimation method. Simulation studies are conducted in Section 5.5 to evaluate the performance of the proposed estimation method. Finally, the discussion and conclusion of our study are given in Section 5.7.

5.3 Extended Hazards Mixture Cure Model

The EH mixture cure model we propose has the logistic regression for the incidence and the EH model for the latency. A logit link function used to model the incidence component can be expressed as

$$\pi(\boldsymbol{X}) = \frac{e^{\boldsymbol{d}'\boldsymbol{X}}}{1 + e^{\boldsymbol{d}'\boldsymbol{X}}} \tag{5.2}$$

where $\boldsymbol{d}$ is a row vector of unknown parameters. The latency component can be described as:

$$\lambda_u(t|\boldsymbol{Z}) = \lambda_0(te^{\boldsymbol{\beta}'\boldsymbol{Z}})e^{\boldsymbol{\alpha}'\boldsymbol{Z}} \tag{5.3}$$

where $\lambda_0(\cdot)$ is the unspecified baseline hazard function, $\boldsymbol{Z}$ is the p-dimensional vector of covariates, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the two $p \times 1$ vectors of unknown parameters. For uncured patients, if the hazard function $\lambda_u(t|\boldsymbol{Z})$ satisfies PH assumption, the mixture cure model reduces to the PH mixture cure model; if there is some linear relationship between time and covariates of uncured patients, the mixture cure model reduces to the AFT mixture cure model.

5.4 Estimation Procedure

Let Y_i be the observed failure time, $\boldsymbol{Z}_i$ be $p \times 1$ vector of covariates in the latency part, $\boldsymbol{X}_i$ be $p \times 1$ vector of covariates in the incidence part, and δ_i be the censoring indicator with $\delta_i = 1$ for the uncensored time and $\delta_i = 0$ for the censored time, where $i = 1, \ldots, n$. We assume the censoring time is independent and noninformative. Based on the equation (5.3), we can easily obtain the cumulative hazard function $\Lambda_u(\boldsymbol{Y}_i|\boldsymbol{Z}_i)$ through integration, then

$$\Lambda_u(\boldsymbol{Y}_i|\boldsymbol{Z}_i) = \Lambda_0(\boldsymbol{Y}_i e^{\boldsymbol{\beta}'\boldsymbol{Z}_i})e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})'\boldsymbol{Z}_i} \tag{5.4}$$

where $\Lambda_0(\cdot)$ is the baseline cumulative hazard function.

Let $f_u(\boldsymbol{Y}_i|\boldsymbol{Z}_i)$ and $S_u(\boldsymbol{Y}_i|\boldsymbol{Z}_i)$ be the density probability function of $\boldsymbol{Y}_i$ and the corresponding survival function, and given the observed data $\boldsymbol{O}_i = (Y_i, \delta_i, \boldsymbol{Z}_i, \boldsymbol{X}_i)$, the likelihood function can be written as

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, d, f_u, S_u|\boldsymbol{O}_i) = \prod_{i=1}^{n} \{\pi(\boldsymbol{X}_i) f_u(\boldsymbol{Y}_i|\boldsymbol{Z}_i)\}^{\delta_i} \{1 - \pi(\boldsymbol{X}_i) + \pi(\boldsymbol{X}_i) S_u(\boldsymbol{Y}_i|\boldsymbol{Z}_i)\}^{1-\delta_i} \tag{5.5}$$

After some transformation based on the equation (5.4), then the equation (5.5) can be rewritten as

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, d, \lambda_0, \Lambda_0 | \boldsymbol{O}_i) = & \prod_{i=1}^{n} \left\{ \pi(\boldsymbol{X}_i) e^{-\Lambda_0(\boldsymbol{Y}_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i}) e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})' \boldsymbol{Z}_i}} \lambda_0(\boldsymbol{Y}_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i}) e^{\boldsymbol{\alpha}' \boldsymbol{Z}_i} \right\}^{\delta_i} \\ & \times \left\{ 1 - \pi(\boldsymbol{X}_i) + \pi(\boldsymbol{X}_i) e^{-\Lambda_0(\boldsymbol{Y}_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i}) e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})' \boldsymbol{Z}_i}} \right\}^{1-\delta_i} \end{aligned} \tag{5.6}$$

Then the logarithm of equation (5.6) can be written as

$$\begin{aligned} l(\boldsymbol{\alpha}, \boldsymbol{\beta}, d, \lambda_0, \Lambda_0 | \boldsymbol{O}_i) = & \sum_{i=1}^{n} \{ \delta_i (\log[\pi(\boldsymbol{X}_i)] + \boldsymbol{\alpha}' \boldsymbol{Z}_i) - \delta_i \Lambda_0(\boldsymbol{Y}_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i}) e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})' \boldsymbol{Z}_i} \\ & + \delta_i \log[\lambda_0(\boldsymbol{Y}_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i})] + (1 - \delta_i) \log[1 - \pi(\boldsymbol{X}_i) + \pi(\boldsymbol{X}_i) e^{-\Lambda_0(\boldsymbol{Y}_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i}) e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})' \boldsymbol{Z}_i}}] \} \end{aligned} \tag{5.7}$$

Since the log-likelihood function (5.7) is a function of unknown parameters of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, d, λ_0, Λ_0, we aim to estimate all of them. However, we may face challenges when we maximize the log-likelihood function (5.7) to obtain $\Lambda_0(\cdot)$, since $\Lambda_0(\cdot)$ is an unspecified nondecreasing function with infinite dimension. Motivated by the inference of the PH model through modeling baseline cumulative hazard function with monotone splines [8, 37], we propose to use monotone splines of Ramsay [47] to model $\Lambda_0(Y_i e^{\boldsymbol{\beta}' \boldsymbol{Z}_i})$ in the equation (5.7).

The monotone splines of Ramsay [47] can be described as

$$\Lambda_0(t) = \sum_{l=1}^{k} \gamma_l I_l(t) \tag{5.8}$$

where $I_l(\cdot)$ is the integrated spline basis function, which is nondecreasing from 0 to 1, and γ_l are chosen from nonnegative value in order to keep the monotonicity of $\Lambda_0(t)$. Since $k = knots + degree - 2$, we can fully determine the k spline basis functions if the knots and degree are specified. Knots, which determine the shape of the monotone splines, can be specified through a sequence of increasing points within a finite interval of the minimum and maximum of the censoring times. Degree, which determines the smoothness of the monotone splines, can be specified with linear (degree = 1), quadratic (degree = 2) and cubic functions (degree = 3), respectively. The number of knots are very important when we use monotone splines of Ramsay [47] to model $\Lambda_0(\cdot)$, since a large number of knots may lead to the over-fitting of the $\Lambda_0(\cdot)$, and a small number of knots may cause the ill-fitting of the $\Lambda_0(\cdot)$. Recommended by Lin and Wang [33], Wang and Dunson [57], Cai et al. [8],

and McMahan et al. [37], a moderate number (10-30) of equally-spaced knots will be used in our study.

After modeling $\Lambda_0(Y_i e^{\boldsymbol{\beta}'\boldsymbol{Z}_i})$ through monotone splines of Ramsay [47], the log-likelihood function (5.7) can be rewritten as

$$\begin{aligned} l_s(\boldsymbol{\alpha}, \boldsymbol{\beta}, d, \boldsymbol{\gamma}|\boldsymbol{O}_i) = &\sum_{i=1}^{n} \{\delta_i(\log[\pi(\boldsymbol{X}_i)] + \boldsymbol{\alpha}'\boldsymbol{Z}_i) - \delta_i \sum_{l=1}^{k} \gamma_l I_l(Y_i e^{\boldsymbol{\beta}'\boldsymbol{Z}_i}) e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})'\boldsymbol{Z}_i} \\ &+ \delta_i \log[\sum_{l=1}^{k} \gamma_l M_l(Y_i e^{\boldsymbol{\beta}'\boldsymbol{Z}_i})] + (1-\delta_i)\log[1-\pi(\boldsymbol{X}_i) \\ &+ \pi(\boldsymbol{X}_i) e^{-\sum_{l=1}^{k} \gamma_l I_l(Y_i e^{\boldsymbol{\beta}'\boldsymbol{Z}_i}) e^{(\boldsymbol{\alpha}-\boldsymbol{\beta})'\boldsymbol{Z}_i}}]\} \end{aligned} \tag{5.9}$$

where $M_l(\cdot)$ is a set of basis splines. Then the unknown parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, d, and $\boldsymbol{\gamma}$ can be obtained directly by maximizing equation (5.9). We estimate the variance of the estimated parameters based on the resampling technique.

5.5 Simulation Study

The performance of the proposed method for our proposed EH mixture cure model will be investigated through simulation studies. Similar to the study of Tseng et al. [55], we will generate 1,000 simulation datasets with a sample size of 200 and 400, and the baseline hazard function for the uncured patients is from a log-logistic distribution:

$$\lambda_0(t) = \frac{\frac{b}{a}(\frac{t}{a})^{b-1}}{1+(\frac{t}{a})^b} \tag{5.10}$$

4 where $a = 120$ and $b = 4$. In our simulation study, we choose the true value of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as $(-1,-1)'$ and $(-0.5,-0.5)'$ for the EH mixture cure model, $(-1,-1)'$ and $(0,0)'$ for the EH mixture cure model, called restricted PH (RPH) mixture cure model, $(-1,-1)'$ and $(-1,-1)'$ for the EH mixture cure model, called restricted AFT (RAFT) mixture cure model, respectively. The covariates are $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)'$, where $\boldsymbol{Z}_1$ is generated from standard normal distribution, and $\boldsymbol{Z}_2$ is generated from Bernoulli distribution with success probability of 0.5. In the logistic link function $\pi(\boldsymbol{X})$, the true values of the parameter $\boldsymbol{d}$ is $(2,-1)'$, and the covariates are set as $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)'$, where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are from constant 1 and Bernoulli distribution with success probability of 0.5, respectively. Censoring times

are generated from an exponential distribution with different means to obtain 35% and 55% censoring rates. Recommended by Lin and Wang [33], Wang and Dunson [57], Cai et al. [8], and McMahan et al. [37], we will choose the number of knots as 30, and the degree as 2 to control the smoothness of the splines. $\boldsymbol{\gamma}$ will be given some equal nonnegative values.

Table 5.1 Bias, SE, SD and CP of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the EH mixture cure model

n	Censoring Rate (%)	Parameters	Bias	SE	SD	CP
200	35	α_1	0.0084	0.0624	0.0522	0.965
		α_2	0.0089	0.0589	0.0476	0.961
		β_1	0.0058	0.0363	0.0387	0.963
		β_2	-0.0301	0.0576	0.0610	0.948
		d_1	-0.0129	0.0505	0.0490	0.949
		d_2	-0.0176	0.0584	0.0544	0.951
	55	α_1	-0.0256	0.0797	0.0782	0.940
		α_2	0.0180	0.0653	0.0797	0.948
		β_1	0.0093	0.0391	0.0412	0.956
		β_2	-0.0363	0.0864	0.0863	0.949
		d_1	-0.0024	0.0606	0.0655	0.946
		d_2	-0.0101	0.0654	0.0730	0.958
400	35	α_1	0.0080	0.0525	0.0462	0.954
		α_2	0.0020	0.0432	0.0333	0.974
		β_1	0.0023	0.0284	0.0322	0.957
		β_2	-0.0284	0.0520	0.0526	0.949
		d_1	-0.0100	0.0411	0.0359	0.961
		d_2	-0.0146	0.0493	0.0445	0.968
	55	α_1	-0.0172	0.0711	0.0666	0.943
		α_2	0.0052	0.0606	0.0636	0.955
		β_1	0.0091	0.0335	0.0365	0.961
		β_2	-0.0314	0.0724	0.0696	0.962
		d_1	-0.0106	0.0569	0.0640	0.942
		d_2	-0.0176	0.0641	0.0594	0.959

The bias, standard error (SE), standard deviation (SD) and coverage probability (CP) of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the EH mixture cure model, RPH mixture cure model, and RAFT mixture cure model are computed, and the results are shown in Tables 5.1-5.3. The results from the EH mixture cure model, seen in Table 5.1, show that when the censoring rate is 35%, the

biases, standard errors, and standard deviations of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ tend to decrease along with the increase of sample size from 200 to 400. For example, with the censoring rate of 35%, the bias, standard error, and standard deviation of $\hat{\alpha}_1$ are (0.0084, 0.0624, 0.0522) for the sample size 200, and (0.0080, 0.0525, 0.0462) for the sample size 400. While the censoring rate is 55%, the biases, standard errors, and standard deviations of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, and $\hat{\beta}_2$ tend to decrease with the increase of sample size from 200 to 400. Along with the increase of the censoring rate from 35% to 55%, the biases of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, and $\hat{\beta}_2$, and all the standard errors or standard deviations tend to increase. The standard deviations and standard errors are very similar and the coverage probabilities are close to 95%, which shows that our estimation method works well for the EH mixture cure model.

Table 5.2 Bias, SE, SD and CP of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the RPH mixture cure model

n	Censoring Rate (%)	Parameters	Bias	SE	SD	CP
200	35	α_1	0.0112	0.0473	0.0449	0.962
		α_2	0.0013	0.0426	0.0349	0.969
		β_1	-0.0138	0.0243	0.0247	0.939
		β_2	-0.0246	0.0368	0.0397	0.937
		d_1	-0.0138	0.0404	0.0382	0.952
		d_2	-0.0176	0.0497	0.0484	0.952
	55	α_1	-0.0171	0.0538	0.0597	0.939
		α_2	0.0160	0.0470	0.0443	0.945
		β_1	-0.0330	0.0541	0.0592	0.932
		β_2	-0.0677	0.0923	0.1036	0.902
		d_1	-0.0015	0.0487	0.0471	0.949
		d_2	-0.0101	0.0532	0.0607	0.944
400	35	α_1	0.0110	0.0383	0.0354	0.955
		α_2	-0.0007	0.0307	0.0240	0.972
		β_1	-0.0113	0.0210	0.0239	0.947
		β_2	-0.0205	0.0339	0.0326	0.942
		d_1	-0.0134	0.0347	0.0333	0.942
		d_2	-0.0156	0.0442	0.0366	0.963
	55	α_1	-0.0112	0.0515	0.0485	0.959
		α_2	0.0087	0.0468	0.0506	0.959
		β_1	-0.0232	0.0372	0.0372	0.937
		β_2	-0.0516	0.0824	0.0789	0.930
		d_1	-0.0154	0.0510	0.0620	0.941
		d_2	-0.0202	0.0597	0.0678	0.937

Table 5.3 Bias, SE, SD and CP of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ of 500 simulated data sets with a sample size of 200 and 400 from the RAFT mixture cure model

n	Censoring Rate (%)	Parameters	Bias	SE	SD	CP
200	35	α_1	-0.0087	0.0631	0.0587	0.950
		α_2	0.0145	0.0582	0.0514	0.964
		β_1	0.0346	0.0558	0.0687	0.948
		β_2	-0.0232	0.0635	0.0628	0.954
		d_1	-0.0080	0.0519	0.0483	0.952
		d_2	-0.0138	0.0603	0.0582	0.954
	55	α_1	-0.0629	0.1037	0.1073	0.914
		α_2	0.0171	0.0794	0.0773	0.965
		β_1	0.0277	0.0559	0.0664	0.934
		β_2	-0.0354	0.0882	0.1002	0.953
		d_1	0.0032	0.0782	0.0677	0.967
		d_2	-0.0100	0.0798	0.0770	0.958
400	35	α_1	-0.0068	0.0518	0.0537	0.945
		α_2	0.0132	0.0446	0.0486	0.953
		β_1	0.0286	0.0489	0.0545	0.944
		β_2	-0.0224	0.0569	0.0590	0.952
		d_1	-0.0069	0.0437	0.0433	0.958
		d_2	-0.0119	0.0540	0.0551	0.955
	55	α_1	-0.0437	0.0917	0.0823	0.942
		α_2	0.0037	0.0668	0.0614	0.965
		β_1	0.0258	0.0491	0.0617	0.936
		β_2	-0.0346	0.0779	0.0753	0.957
		d_1	-0.0092	0.0653	0.0592	0.965
		d_2	-0.0168	0.0697	0.0636	0.957

The proposed estimation method also performs well in the RPH mixture cure model, and the results have been shown in Table 5.2. With the increase of sample size from 200 to 400, all the biases, standard errors and standard deviations tend to decrease under the censoring rate of 35%. When the censoring rate is 55%, the biases and standard errors of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, and $\hat{\beta}_2$ tend to decrease with the increase of sample size from 200 to 400. When the censoring rate increases from 35% to 55%, the biases, standard errors and standard deviations of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, and $\hat{\beta}_2$ tend to increase. For example, with the sample size 200, the bias, standard error and standard deviation of $\hat{\alpha}_2$ are (0.0013, 0.0426, 0.0349) for censoring rate 35%, and (0.0160, 0.0470, 0.0443) for censoring rate 55%. Similar to

the results shown for the EH mixture cure model, the standard errors and the standard deviations are comparable, and the coverage probabilities are also close to 95% under the RPH mixture cure model.

Table 5.3 displays the results of the bias, standard error, standard deviation and coverage probability of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ from the RAFT mixture cure model. The results reveal that along with the increase of the sample size from 200 to 400, all the biases, standard errors and standard deviations decrease when the censoring rate is 35%. For example, under the censoring rate 35%, the bias, standard error, and standard deviation of $\hat{\beta}_2$ are (-0.0232, 0.0635, 0.0628) for the sample size 200, and (-0.0224, 0.0569, 0.0590) for the sample size 400. When the censoring rate is 55%, the biases, standard errors and standard deviations of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$ and $\hat{\beta}_2$ decrease along with the increase of sample size from 200 to 400. The patterns for the biases of $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{d}_1$ and $\hat{d}_2$ may not change regularly when the censoring rate changes from 35% to 55%. In comparison, all the standard errors and standard deviations increase when the censoring rate increases from 35% to 55%. The fact that the standard errors are close to the standard deviations, and that the coverage probabilities are close to 95%, show that our proposed estimation method is valid for the RAFT mixture cure model.

5.6 Real Data Analysis

For illustration, we apply our proposed estimation method and proposed EH mixture cure model to the melanoma data from the ECOG phase III clinical trial E1684 and the leukemia data from bone marrow transplant study. There are a total number of 284 observations for the melanoma data after deleting the missing value. This phase III clinical trial E1684 aimed to compare the high dose interferon alpha-2b (IFN) regimen group with the placebo group. The dependent variable in the melanoma data is relapse-free survival in years, which is defined as the time to death or relapse after randomization. Censoring indicator is defined as 1 if death or relapse of patients happen, and as 0 if patients do not die or relapse. Similar to Cai et al. [9], three risk factors are considered for our study, including treatment (0 = control, 1 = treatment), sex (0 = male, 1 = female) and age (continuous variable which is

centered to the mean). We investigate the effects of treatment, sex and age on the cure rate of cured patients as well as the failure time of uncured patients through the PH mixture cure model [9] and our proposed EH mixture cure model. The standard errors of estimates are obtained using the bootstrap method based on 500 bootstrap samples.

Table 5.4 Estimates, SE and P values of estimated parameters for the melanoma data from the ECOG phase III clinical trial E1684 under the PH mixture cure model

PH Mixture Cure Model	Parameter	Estimate	SE	P values
Latency Part	Treatment	-0.1536	0.1721	0.3722
	Sex	0.0995	0.1908	0.6022
	Age	-0.0077	0.0067	0.2523
Incidence Part	Treatment	-0.5885	0.3065	0.0548
	Sex	-0.0870	0.3291	0.7916
	Age	0.0203	0.0145	0.1593

Table 5.5 Estimates, SE and 95% confidence intervals of estimated parameters for the melanoma data from the ECOG phase III clinical trial E1684 under the EH mixture cure model

EH Mixture Cure Model	Parameter	Estimate	SE	95% CI
Latency Part	$\boldsymbol{\alpha}$			
	Treatment	-0.1556	0.1762	(-0.5010, 0.1898)
	Sex	0.0909	0.1935	(-0.2882, 0.4701)
	Age	-0.0063	0.0049	(-0.0160, 0.0033)
	$\boldsymbol{\beta}$			
	Treatment	-0.0577	0.0366	(-0.1293, 0.0140)
	Sex	-0.1243	0.0593	(-0.2405, -0.0082)
	Age	0.0078	0.0024	(0.0031, 0.0124)
Incidence Part	$\boldsymbol{d}$			
	Treatment	-0.5862	0.5082	(-1.5822, 0.4099)
	Sex	-0.1005	0.3954	(-0.8756, 0.6745)
	Age	0.0048	0.0089	(-0.0126, 0.0222)

Table 5.4-5.5 display estimates, standard errors and 95% confidence intervals or P values of estimated parameters for the melanoma data under the PH mixture cure model and the EH mixture cure model. From the results, we can see that treatment, sex and age all have no significant effects on either the cure rate of cured patients or the failure time to

Table 5.6 Estimates, SE and P values of estimated parameters for the bone marrow transplant data under the AFT mixture cure model

AFT Mixture Cure Model	Parameter	Estimate	SE	P values
Latency Part	Treatment	-0.3531	0.2706	0.1919
Incidence Part	Treatment	0.4273	0.4844	0.3776

Table 5.7 Estimates, SE and 95% confidence intervals of estimated parameters for the bone marrow transplant data under the EH mixture cure model

EH Mixture Cure Model	Parameter	Estimate	SE	95% CI
Latency Part	$\boldsymbol{\alpha}$			
	Treatment	-0.2812	0.2947	(-0.8588, 0.2964)
	$\boldsymbol{\beta}$			
	Treatment	-0.4254	0.2955	(-1.0046, 0.1537)
Incidence Part	$\boldsymbol{d}$			
	Treatment	0.3090	0.5281	(-0.7261, 1.3442)

death or relapse of uncured patients under the PH mixture cure model, since all the p values are greater than 0.05. Furthermore, the results from the EH mixture cure model also show that treatment, sex and age all have no significant effects on the cure rate of cured patients, since 95% confidence intervals of all the three covariates in the incidence part contain zero. However, treatment, sex and age do not all have no significant effects on the failure time to death or relapse of uncured patients in the latency part: treatment, sex and age in the relative hazard part have no significant effects on the failure time to death or relapse of uncured patients, since their 95% confidence intervals contain zero; treatment in the baseline hazard part has no significant effect on the failure time to death or relapse of uncured patients, since the 95% confidence interval includes zero, while sex and age in the baseline hazard part have significant effects on the failure time to death or relapse of uncured patients, since their 95% confidence intervals are (-0.2405, -0.0082) and (0.0031, 0.0124), respectively.

The leukemia data from the bone marrow transplant study has been widely investigated through the AFT mixture cure model, since the PH assumption is not appropriate for the latency distribution [9, 43, 65, 71]. Among the 91 patients who had been treated with

high-dose chemoradiotherapy in the bone marrow transplant study, 46 patients were in the allogeneric marrow group and 45 patients were in the autologous marrow group. The response variable in the leukemia data is time to death. Censoring indicator is defined as 1 if patients died and as 0 if patients did not die. Treatment (1 is for autologous treatment group and 0 is for allogeneic treatment group) is the only covariate we consider in the models. Since we are interested in examining whether the treatment has significant effects on the cure rate of cured patients and the failure time to death of uncured patients, we apply both the AFT mixture cure model [9] and our proposed EH mixture cure model to fit the leukemia data. The standard errors of estimates are obtained using the bootstrap method based on 200 bootstrap samples.

Table 5.6-5.7 display estimates, standard errors and 95% confidence intervals or P values of estimated parameters for the leukemia data under the AFT mixture cure model and the EH mixture cure model. The results obtained from the AFT mixture cure model show that treatment has no significant effects on either the cure rate of cured patients or the failure time to death of uncured patients, since all the p values are greater than 0.05. Similarly, the EH mixture cure model results also show that treatment has no significant effects on either the cure rate of cured patients or the failure time to death of uncured patients, since all the 95% confidence intervals include zero. The fact that the estimate of treatment in the relative hazard part is -0.2812 and the estimate of treatment in the baseline hazard part is -0.4254 shows that they are not exactly equivalent, which may violate the assumption of the AFT mixture cure model. Therefore, the AFT mixture cure model may not be a better model to fit the leukemia data, even though the conclusions are similar from either the AFT mixture cure model or the EH mixture cure model.

5.7 Discussion and Conclusion

In this chapter, we have proposed an EH mixture cure model and its corresponding estimation method based on monotone splines of Ramsay. The EH mixture cure model is a very useful extension of both the PH mixture cure model and the AFT mixture cure model. It has been shown that the EH mixture cure model is more flexible than either the PH mixture

cure model or the AFT mixture cure model, since it incorporates a logistic regression for the incidence part and an EH model for the latency part of the mixture cure model. Our simulation studies show that the proposed estimation method performs well when estimating the regression parameters in the EH mixture cure model, the RPH mixture cure model and the RAFT mixture cure model.

The real data results also show that our proposed estimation method performs well, and our proposed EH mixture cure model outperforms either the PH mixture cure model or the AFT mixture cure model. The first reason is that not only the similar conclusions for the incidence part in either the PH mixture cure model or the AFT mixture cure model can be obtained from the EH mixture cure model, but also more details in the latency part can be obtained from the EH mixture cure model than either the PH mixture cure model or the AFT mixture cure model, to reveal the accurate association of interested covariates with the failure time of uncured patients. The second reason is that there is no need to check either the PH assumption or the AFT assumption when we apply our proposed EH mixture cure model to the real data, compared with using either the PH mixture cure model or the AFT mixture cure model.

Therefore, our proposed EH mixture cure model is more efficient and flexible than either the PH mixture cure model or the AFT mixture cure model, and we recommend using our proposed EH mixture cure model to fit the survival data with cure information without knowing either the PH assumption or the AFT assumption.

Chapter 6

Conclusions and Future Work

In this dissertation, we have discussed variety survival models, and their estimation procedures. The details are illustrated through: (1) Semiparametric estimations for the AP-AFT model; (2) Profile likelihood based estimation method for the AFT mixture model with latent subgroups; (3) Spline based estimation method for the EH Model; (4) Spline based estimation method for the EH mixture cure model.

In the first project, we propose an AP-AFT model and its corresponding estimation methods based on either the rank-smooth method or profile likelihood method. Our simulation studies show that the performance of the proposed estimation methods are valid. Comparing to the rank-like estimation method, the proposed rank-smooth method using the smoothing technique for variance calculation instead of resampling technique in the rank-like method, improves the computational time. Comparing to the rank-smooth method, the profile likelihood based estimation method is easy and straightforward to evaluate the variance of estimated parameters. Due to the burden of computing, we suggest using the rank-smooth estimation method when the sample size is huge, and utilizing either the rank-smooth estimation method or the profile likelihood based estimation method when the sample size is moderate.

In the second project, we develop the EM algorithm based estimation method for the AFT mixture model with latent subgroup. Our simulation studies show that the proposed method outperforms the E-BJ algorithm based estimation method regarding to the variance estimation.

The third project aims to develop the estimation method for the EH model based on the monotone splines. The simulation results indicate the spline based estimation method

is valid and flexible. Comparing to the existed profile likelihood estimation method, the proposed estimation method is better in calculating variance of estimated parameters.

We extend the discussion of the EH model to the EH mixture cure model in the fourth project, which is more flexible than either the PH mixture cure model or the AFT mixture cure model. The EH mixture cure model has following advantages in practice: (1) There is no need to check either the PH assumption or the AFT assumption when we apply the EH mixture cure model to the real data; (2) The latency part of the mixture cure model is modelled with the EH model, which has the nested structure of both the PH model and the AFT model. The estimation method for EH mixture cure model is developed based on the monotone splines. The bootstrap method is used to estimate the variance of estimated parameters. The simulation studies illustrate the good performance of the proposed model. Finally, we recommend using our proposed EH mixture cure model to the survival data with cure information without knowing either the PH assumption or the AFT assumption.

All estimation methods discussed in the dissertation are about right censored survival data. However, in the real situation, the survival data may be interval censored. Therefore, one of future direction may aim to extend the proposed models and estimation methods to the interval censored data. Furthermore, in the third and fourth project, we use the resampling technique to evaluate variance of estimated parameters, but this is often time-consuming. Therefore, we would like to investigate more advanced techniques to estimate the variance of parameters in the future. Finally, work in the future may include constructing a goodness-of-fit test for either the EH model or the EH mixture cure model to check the fit of models.

Bibliography

1. Lily L Altstein and Gang Li, *Latent subgroup analysis of a randomized clinical trial through a semiparametric accelerated failure time mixture model*, Biometrics **69** (2013), no. 1, 52–61.

2. Lily L Altstein, Gang Li, and Robert M Elashoff, *A method to estimate treatment efficacy among latent subgroups of a randomized clinical trial*, Statistics in medicine **30** (2011), no. 7, 709–717.

3. Joseph Berkson and Robert P Gage, *Survival curve for cancer patients following treatment*, Journal of the American Statistical Association **47** (1952), no. 259, 501–515.

4. John W Boag, *Maximum likelihood estimates of the proportion of patients cured by cancer therapy*, Journal of the Royal Statistical Society. Series B (Methodological) **11** (1949), no. 1, 15–53.

5. BM Brown and You-Gan Wang, *Standard errors and covariance matrices for smoothed rank estimators*, Biometrika **92** (2005), no. 1, 149–158.

6. _____, *Induced smoothing for rank regression with censored survival times*, Statistics in medicine **26** (2007), no. 4, 828–836.

7. Jonathan Buckley and Ian James, *Linear regression with censored data*, Biometrika **66** (1979), no. 3, 429–436.

8. Bo Cai, Xiaoyan Lin, and Lianming Wang, *Bayesian proportional hazards model for current status data with monotone splines*, Computational Statistics & Data Analysis **55** (2011), no. 9, 2644–2651.

9. Chao Cai, Yubo Zou, Yingwei Peng, and Jiajia Zhang, *smcure: An r-package for estimating semiparametric mixture cure models*, Computer methods and programs in biomedicine **108** (2012), no. 3, 1255–1260.

10. Kani Chen, Jia Shen, and Zhiliang Ying, *Rank estimation in partial linear model with censored data*, Statistica sinica (2005), 767–779.

11. Ying Qing Chen and Nicholas P Jewell, *On a general class of semiparametric hazards regression models*, Biometrika **88** (2001), no. 3, 687–702.

12. A Ciampi and J Etezadi-Amoli, *A general model for testing the proportional hazards and the accelerated failure time hypotheses in the analysis of censored survival data with covariates*, Communications in Statistics-Theory and Methods **14** (1985), no. 3, 651–667.

13. DH Cox, *Regression models and life-tables*, Journal of the Royal Statistical Society (1972), 187–220.

14. Jack Cuzick, Peter Sasieni, Jonathan Myles, and Jonathan Tyrer, *Estimating the effect of treatment in a proportional hazards model in the presence of non-compliance and contamination*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69** (2007), no. 4, 565–588.

15. Lelia Duley, *The global impact of pre-eclampsia and eclampsia*, Seminars in perinatology, 2009, pp. 130–137.

16. Elosha Eiland, Chike Nzerue, and Marquetta Faulkner, *Preeclampsia 2012*, Journal of pregnancy **2012** (2012).

17. Paul HC Eilers and Brian D Marx, *Flexible smoothing with b-splines and penalties*, Statistical science (1996), 89–102.

18. Hongbin Fang, Gang Li, and Jianguo Sun, *Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model*, Scandinavian Journal of Statistics **32** (2005), no. 1, 59–75.

19. Vern T Farewell, *The use of mixture models for the analysis of survival data with long-term survivors*, Biometrics (1982), 1041–1046.

20. Thomas R Fleming and David P Harrington, *Counting processes and survival analysis*, Vol. 169, John Wiley & Sons, 2011.

21. Dean A Follmann, *On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial*, Journal of the American Statistical Association **95** (2000), no. 452, 1101–1109.

22. Robert J Gray, *Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis*, Journal of the American Statistical Association **87** (1992), no. 420, 942–951.

23. Jie Hong, Yinding Wang, Suzanne McDermott, Bo Cai, C Marjorie Aelion, and Jamie Lead, *The use of a physiologically-based extraction test to assess relationships between bioaccessible metals in urban soil and neurodevelopmental conditions in children*, Environmental Pollution **212** (2016), 9–17.

24. Jian Huang, Shuangge Ma, and Huiliang Xie, *Regularized estimation in the accelerated failure time model with high-dimensional covariates*, Biometrics **62** (2006), no. 3, 813–820.

25. Zhezhen Jin, DY Lin, LJ Wei, and Zhiliang Ying, *Rank-based inference for the accelerated failure time model*, Biometrika **90** (2003), no. 2, 341–353.

26. Zhezhen Jin, DY Lin, and Zhiliang Ying, *On least-squares regression with censored data*, Biometrika **93** (2006), no. 1, 147–161.

27. Brent A Johnson, *Variable selection in semiparametric linear regression with censored data*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70** (2008), no. 2, 351–370.

28. Emanuel G Kuflik, *Cryosurgery for skin cancer: 30-year experience and cure rates*, Dermatologic surgery **30** (2004), no. s2, 297–300.

29. Anthony YC Kuk and Chen-Hsin Chen, *A mixture model combining logistic regression with proportional hazards regression*, Biometrika **79** (1992), no. 3, 531–541.

30. Tze Leung Lai and Zhiliang Ying, *Rank regression methods for left-truncated and right-censored data*, The Annals of Statistics (1991), 531–556.

31. Chin-Shang Li and Jeremy MG Taylor, *A semi-parametric accelerated failure time cure model*, Statistics in medicine **21** (2002), no. 21, 3235–3247.

32. Danyu Y Lin and Lee-Jen Wei, *The robust inference for the cox proportional hazards model*, Journal of the American statistical Association **84** (1989), no. 408, 1074–1078.

33. Xiaoyan Lin and Lianming Wang, *A semiparametric probit model for case 2 interval-censored failure time data*, Statistics in medicine **29** (2010), no. 9, 972–981.

34. Tom Loeys and Els Goetghebeur, *A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance*, Biometrics **59** (2003), no. 1, 100–105.

35. Wenbin Lu, *Efficient estimation for an accelerated failure time model with a cure fraction*, Statistica Sinica **20** (2010), 661.

36. Joshua R Mann, Suzanne McDermott, Margaret I Griffith, James Hardin, and Anthony Gregg, *Uncovering the complex relationship between pre-eclampsia, preterm birth and cerebral palsy*, Paediatric and perinatal epidemiology **25** (2011), no. 2, 100–110.

37. Christopher S McMahan, Lianming Wang, and Joshua M Tebbs, *Regression analysis for current status data using the em algorithm*, Statistics in medicine **32** (2013), no. 25, 4452–4466.

38. Boyd E Metzger, Donald R Coustan, Organizing Committee, et al., *Summary and recommendations of the fourth international workshop-conference on gestational diabetes mellitus*, Diabetes care **21** (1998), B161.

39. Rupert Miller and Jerry Halpern, *Regression with censored data*, Biometrika **69** (1982), no. 3, 521–531.

40. Rupert G Miller, *Least squares regression with censored data*, Biometrika **63** (1976), no. 3, 449–464.

41. Ingrid Östlund, Bengt Haglund, and Ulf Hanson, *Gestational diabetes and preeclampsia*, European Journal of Obstetrics & Gynecology and Reproductive Biology **113** (2004), no. 1, 12–16.

42. Lei Pang, Wenbin Lu, and Huixia Judy Wang, *Variance estimation in censored quantile regression via induced smoothing*, Computational statistics & data analysis **56** (2012), no. 4, 785–796.

43. Yingwei Peng and Keith BG Dear, *A nonparametric mixture model for cure rate estimation*, Biometrics **56** (2000), no. 1, 237–243.

44. Yingwei Peng, Keith BG Dear, JW Denham, et al., *A generalized f mixture model for cure rate estimation*, Statistics in medicine **17** (1998), no. 8, 813–830.

45. Ross L Prentice, *Linear rank tests with right censored data*, Biometrika **65** (1978), no. 1, 167–179.

46. Annie Qu and Runze Li, *Quadratic inference functions for varying-coefficient models with longitudinal data*, Biometrics **62** (2006), no. 2, 379–391.

47. JO Ramsay, *Monotone regression splines in action*, Statistical science (1988), 425–441.

48. Y Ritov, *Estimation in a linear regression model with censored data*, The Annals of Statistics (1990), 303–328.

49. James M Robins and Anastasios A Tsiatis, *Correcting for non-compliance in randomized trials using rank preserving structural failure time models*, Communications in Statistics-Theory and Methods **20** (1991), no. 8, 2609–2631.

50. Michael Schemper and Terry L Smith, *Efficient evaluation of treatment effects in the presence of missing covariate values*, Statistics in medicine **9** (1990), no. 7, 777–784.

51. Winfried Stute and J-L Wang, *The strong law under random censorship*, The Annals of Statistics (1993), 1591–1607.

52. Judy P Sy and Jeremy MG Taylor, *Estimation in a cox proportional hazards cure model*, Biometrics **56** (2000), no. 1, 227–236.

53. Jeremy MG Taylor, *Semi-parametric estimation in failure time mixture models*, Biometrics (1995), 899–907.

54. Xingwei Tong, Liang Zhu, Chenlei Leng, Wendy Leisenring, and Leslie L Robison, *A general semiparametric hazards regression model: efficient estimation and structure selection*, Statistics in medicine **32** (2013), no. 28, 4980–4994.

55. Yi-Kuan Tseng and Ken-Ning Shu, *Efficient estimation for a semiparametric extended hazards model*, Communications in Statistics - Simulation and Computation **40** (2011), no. 2, 258–273.

56. Anastasios A Tsiatis, *Estimating regression parameters using linear rank tests for censored data*, The Annals of Statistics (1990), 354–372.

57. Lianming Wang and David B Dunson, *Semiparametric bayes' proportional odds models for current status data with underreporting*, Biometrics **67** (2011), no. 3, 1111–1118.

58. Rui Wang, Stephen W Lagakos, James H Ware, David J Hunter, and Jeffrey M Drazen, *Statistics in medicine - reporting of subgroup analyses in clinical trials*, New England Journal of Medicine **357** (2007), no. 21, 2189–2194.

59. Yinding Wang, Suzanne McDermott, Joshua R Mann, and James W Hardin, *Preventing intellectual disability during pregnancy: what are the potentially high yield targets?*, Journal of perinatal medicine **44** (2016), no. 4, 421–432.

60. LJ Wei, *The accelerated failure time model: a useful alternative to the cox regression model in survival analysis*, Statistics in medicine **11** (1992), no. 14-15, 1871–1879.

61. LJ Wei, Z Ying, and DY Lin, *Linear regression analysis of censored survival data based on rank tests*, Biometrika **77** (1990), no. 4, 845–851.

62. Juan P Wisnivesky, David Yankelevitz, and Claudia I Henschke, *The effect of tumor size on curability of stage i non-small cell lung cancers*, CHEST Journal **126** (2004), no. 3, 761–765.

63. Simon Wood, *Generalized additive models: an introduction with r*, CRC press, 2006.

64. Linzhi Xu and Jiajia Zhang, *An alternative estimation method for the semiparametric accelerated failure time mixture cure model*, Communications in Statistics-Simulation and Computation **38** (2009), no. 9, 1980–1990.

65. ———, *Multiple imputation method for the semiparametric accelerated failure time mixture cure model*, Computational Statistics & Data Analysis **54** (2010), no. 7, 1808–1816.

66. Kazuo Yamaguchi, *Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in japan*, Journal of the American Statistical Association **87** (1992), no. 418, 284–292.

67. Song Yang, *Extended weighted log-rank estimating functions in censored regression*, Journal of the American Statistical Association **92** (1997), no. 439, 977–984.

68. Yariv Yogev, Elly MJ Xenakis, and Oded Langer, *The association between preeclampsia and the severity of gestational diabetes: the impact of glycemic control*, American journal of obstetrics and gynecology **191** (2004), no. 5, 1655–1660.

69. Yan Yu and David Ruppert, *Penalized spline estimation for partially linear single-index models*, Journal of the American Statistical Association **97** (2002), no. 460, 1042–1054.

70. Donglin Zeng and DY Lin, *Efficient estimation for the accelerated failure time model*, Journal of the American Statistical Association **102** (2007), no. 480, 1387–1396.

71. Jiajia Zhang and Yingwei Peng, *A new estimation method for the semiparametric accelerated failure time mixture cure model*, Statistics in medicine **26** (2007), no. 16, 3157–3171.

72. Yubo Zou, Jiajia Zhang, and Guoyou Qin, *A semiparametric accelerated failure time partial linear model and its application to breast cancer*, Computational statistics & data analysis **55** (2011), no. 3, 1479–1487.