REGRESSION ANALYSIS OF CORRELATED INTERVAL-CENSORED FAILURE TIME DATA WITH A CURED SUBGROUP

A Dissertation presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

DIAN YANG

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

MAY 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

REGRESSION ANALYSIS OF CORRELATED INTERVAL-CENSORED FAILURE TIME DATA WITH A CURED SUBGROUP

presented by Dian Yang,

a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. (Tony) Jianguo Sun

Dr. Shih-Kang Chao

Dr. Shawn Ni

Dr. Yushu Shi

ACKNOWLEDGMENTS

Foremost, I would like to express my most heartfelt gratitude to my advisor Dr. (Tony) Jianguo Sun for his patience and support of my PhD study. Without his continuous encouragement and guidance in the past year, it's impossible for me to get through the tough time and complete this dissertation. It is my greatest honor to be one of Dr. Sun's students.

Besides, I extend my gratitude to my advisory committee members: Dr. Shih-Kang Chao, Dr. Shawn Ni and Dr. Yushu Shi for their encouragement, suggestions and feedbacks.

I also owe a debt to all faculty in our department for teaching me and assisting me. I am deeply grateful to our great staff Judy, Kathleen and Abbie for their help.

Furthermore, I would like to thank my friends for their liveliness, enthusiasm and wisdom. We learn from each other and help each other. I will never forget the funny, wonderful and emotional time we had.

Finally, I am deeply indebted to my parents for their love, patience and understanding. They are not only who gave birth to me, but also who encouraged me to move forward. Without their endless support, this dissertation would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

Regression Analysis of Correlated Interval-censored Failure Time Data with a Cured Subgroup

Dian Yang

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

Interval-censored failure time data commonly occur in many periodic follow-up studies such as epidemiological experiments, medical studies and clinical trials. By interval-censored data, we usually mean that one cannot observe the failure time of interest and instead we know that it belongs to a time interval. Correlated failure time data commonly occur when there are multiple events on one individual or when the study subjects are clustered into some small groups. In this situation, study subjects from same subgroup or the failure events from same individuals are usually regarded as dependent, but the subjects in different clusters or failure events from different individuals are assumed to be independent. Besides the correlation between the cluster, sometimes the cluster size may be informative or carry some information about the failure time of interest. Cured subgroup is another interesting topic that has been discussed by many authors. For this situation, unlike the assumptions in traditional survival model that all study subjects would experience the failure event of interest eventually if the follow-up time is long enough, some subjects may never experience or not be susceptible to the event. Such subjects are treated as cured and assumed to belong to a cured subgroup in a study population.

The research in this dissertation focuses on regression analysis of correlated interval-censored data with a cured subgroup via different approaches based on different data structures. In the first part of this dissertation, we discuss clustered interval-censored data with a cured subgroup and informative cluster size. To address this, we present a within-cluster resampling method and in the approach, the multiple imputation procedure is applied for estimation of unknown parameters. To assess the performance of the proposed method, a simulation study is conducted and suggests that it works well in practical situations. Also, the method is applied to a set of real data that motivated this study.

In the second part of this dissertation, we consider the clustered interval-censored data with a cured subgroup via a non-mixture cure model. We present a maximum likelihood estimation procedure under the semiparametric transformation non-mixture cure model. To estimate the unknown parameters, an expectation maximization (EM) algorithm based on an augmentation of Poisson variable is developed. To assess the performance of the proposed method, a simulation study is conducted and suggests that it works well in practical situations. An application to a study conducted by the National Aeronautics and Space Administration that motivated this study is also provided.

In the third part of this dissertation, we investigate the bivariate interval-censored data with a cured subgroup. A sieve maximum likelihood estimation procedure under the semiparametric transformation non-mixture cure model based on Bernstein polynomials is presented. A simulation study is conducted to assess the finite sample performance of the proposed method and suggests that the proposed model works well. Also, a real data application from the study of AIDS Clinical Trial Group 181

is provided.

Chapter 1

Introduction

1.1 Regression Analysis of Univariate Interval-Censored Data

The analysis of failure time event data with different censoring has been widely studied by many authors. Right censoring, the most common type of censoring, occurs frequently in many areas such as epidemiological studies, psychological experiments and clinical trials. For such situation, the failure time of interest is either observed exactly or known to be greater than the censoring time. Many authors have investigated different analyses of right-censored data. Kaplan and Meier (1958) presented a non-parametric estimator, product limit estimator or Kaplan-Meier estimator, to estimate the survival function of failure event of interest. Mantel (1966) developed a logrank test to compare the survival distributions of two samples. Cox (1972) proposed the proportional hazards model with partial likelihood approach for estimation

of covariate effects. Kalbfleisch and Prentice (2002) and Lawless (2011) provided extensive illustrations and examples and discussed various statistical models and estimation approaches for right-censored data.

Interval-censored data commonly occur in many areas including demo-graphical, epidemiological, medical and sociological studies, and have received increasing attention in the literature. By interval censoring, we mean that the failure time of interest is not under continuous observation. Let T denote the failure time, the exact failure time T cannot be observed exactly and is only known to belong to a time interval $(L, R]$, where $L \leq R$. Also, right-censored data and left-censored data could be treated as special cases of interval censoring, where $R = +\infty$ and $L = 0$, respectively. Meanwhile, when only one observation is observed, the failure time is either left- or right-censored, and the data is referred to as Case I interval-censored data or current status data. The case with more than 2 examination times is usually called case-2 or case-k interval censored data. The analysis of interval-censored data is more challenge than that of right-censored data, since the data structure of the former is more complicated than the latter. As a consequence, the typical methodology for right-censored data including counting process and martingale theory cannot be applied to interval-censored data directly. Also, for regression analysis under the proportional hazards model, the partial-likelihood-based inference cannot be used for interval-censored data.

Many methods have been developed on the analysis of univariate interval-censored failure time data. For nonparametric maximum likelihood estimator(NPMLE) of sur-

vival function with case-2 interval-censored data, Turnbull (1976) and Groeneboom and Wellner (1992) proposed a self-consistency algorithm and an iterative convex minorant algorithm, respectively. Later, Wellner and Zhan (1997) developed a hybrid algorithm combining the self-consistency and expectation maximization (EM) algorithm together, which showed a more rapid convergence speed. For current status data, the NPMLE could be obtained by the isotonic regression via max-min algorithm or pool adjacent violators algorithm (Sun, 2006). Another class of methods to estimate survival function or hazard functions is smoothing estimation approach. The commonly used methods are kernel-based approach using kernel smoothing functions and spline-based approach via maximizing log-likelihood or penalized log likelihood function with different spline basis (Sun, 2006). For the comparison of survival functions among different groups, such as groups with different treatments in clinical studies, many authors discussed different methods including the rank-based test and the testing based on survival function. Zhao and Sun (2004) extended the log-rank test for right-censored data to a generalized log-rank test for a mixed interval-censored data. Petroni and Wolfe (1994) considered a class of asymptotically nonparametric tests for two sample discrete case-2 interval-censored data based on survival functions. Fang *et al.* (2002) developed a class of test statistics for general types of case-2 interval-censored data based on integrated weighted differences between two estimated survival functions. For current status data, the common test procedure is also either based on rank statistics modified from log-rank test or a generalization of the weighted Kaplan-Meier procedure under the assumption that observation time follows the same distribution (Sun, 2006). Sun (1999) considered the test procedure focusing on the situation where the censoring distributions may be different for sub-

jects in different treatment groups.

Regression analysis is of most interest in survival analysis. Many authors have discussed regression analysis of interval-censored data under the proportional hazards model

$$S(t) = exp(-\Lambda(t)e^{X^T\beta}).$$

In the above, $S(t)$ denotes the survival function at time t, $\Lambda(t)$ is a non-decreasing baseline cumulative hazard function, and X and β represents the covariates and unknown parameters, respectively. Finkelstein (1986) discussed regression analysis of case-2 interval-censored data and developed estimation procedure based on full likelihood. This approach requires estimation for both regression parameter and baseline hazards function with Newton-Raphson algorithm, which is computationally intensive when the sample size increases. Satten (1996) and Goggins *et al.* (1998) both proposed the estimation procedures for case-2 interval-censored data without estimating baseline hazard function. The former considered a Gibbs sampling procedure for generating rankings and used a stochastic approximation for solving the score functions, while the latter developed a Monte Carlo EM algorithm to implement the estimation procedure of regression parameters. Another type of approach is based on smoothing functions, which transfers the infinite-dimensional parameters to finite-dimensional parameters via different smoothing functions like splines (Cai and Betensky, 2003). For current status data, Huang (1996) consider a maximum likelihood estimation procedure based on the convex minorant algorithm. However, for large data sets, the required computation could become intensive.

Proportional odds model is another commonly used semiparametric model, which has the form

$$logit\big[S(t)\big] = logit\big[S_0(t)\big] - X^T\beta,$$

where $S_0(t)$ is the baseline survival function, and $logit(x) = log(\frac{x}{1-x})$. Huang and Wellner (1997) discussed the maximum likelihood approach for case-2 interval-censored data and provided the efficient score function for β. Huang and Rossini (1997a) and Shen (1998) proposed the sieve maximum likelihood estimation procedure for case-2 interval-censored data with piece-wise linear function and spline functions, respectively. In terms of current status data, Huang (1995) and Rossini and Tsiatis (1996) studied the nonparametric maximum likelihood estimation and sieve maximum likelihood estimation respectively. Rabinowitz *et al.* (2000) used the conditional likelihood approach based on a conditional logistic regression without estimating baseline function for both case-1 and case-2 interval-censored data.

Since the late 20th century, the transformation model has been introduced for survival analysis (Chen and Little, 2001; Chen *et al.*, 2002). The survival function under such model takes the form

$$S(t) = \exp\left\{ - G\left[\Lambda(t)\, e^{X^T\beta}\right]\right\}. \tag{1.1}$$

In the above, $G(t)$ is a specific transformation function that is strictly increasing. The choices of $G(x) = x$ and $G(x) = log(1+x)$ yields the proportional hazards model and proportional odds model, respectively. One should notice that model (1.1) could be

written as the linear transformation model

$$log[\Lambda(T)] = -X^T\beta + \epsilon, \tag{1.2}$$

where ϵ follows a specific distribution function $F_\epsilon = 1 - exp[-G\{exp(x)\}]$. Many authors have investigated the transformation model for interval-censored data. Zhang *et al.* (2005) and Sun and Sun (2005) proposed the estimation procedures based on estimation equation for case-2 interval-censored data and current status data, respectively. Zhang and Zhao (2013) developed the empirical likelihood (EL) inference approaches for the regression parameters based on the generalized estimating equations. Gu *et al.* (2006) considered the rank-based approach via Markov Chain Monte Carlo stochastic approximation. However, all the methods mentioned above are computationally demanding or statistically inefficient. Zeng *et al.* (2016) proposed the nonparametric maximum likelihood estimation procedure with Poisson random variable data augmentation via an EM algorithm, which could also handle time-dependent covariates for interval-censored data. One advantage of this method is that it gives a closed-form solution for estimating baseline hazard function and reduces the computation cost.

Besides the approaches directly applied to interval-censored data, imputation-based method is another popular approach. Multiple imputation is a general method to handle missing data (Rubin, 1987). By multiple imputation, we mean that the imputation procedure should be carried out for multiple times. Therefore, multiple imputation leads to a multiple imputed data sets, which could be analyzed by a standard model. If we treat interval-censored data as missing data and impute a possible

survival time from the time interval, then we could apply the typical statistical analysis for right-censored data to analyze the imputed data. Pan (2000) discussed the multiple imputation approach for interval-censored data under the proportional hazards model and provided two types of data augmentation. The estimation procedure is easy to implement, and as a consequence, multiple imputation approach has been widely applied to some more complex problem with interval-censored data.

1.2 Regression Analysis of Correlated Interval-censored Failure Time Data

Correlated failure time data commonly occur when there are multiple events on one individual or when the study subjects are clustered into some small groups. In this situation, the study subjects from same subgroup or the failure events from same individuals are usually regarded as dependent, but the subjects in different clusters or failure events from different individuals are assumed to be independent. When dealing with such data, the typical survival models assuming independence among all the subjects are inappropriate. Hougaard (2012) discussed the different types of multiple-event or clustered failure time data and gave a comprehensive review of most commonly used models for such problem. One of the most commonly used methods is marginal approach, which estimates covariate effects based on the marginal distributions with working independence assumption (Lin, 1994; Liang *et al.*, 1995; Clegg *et al.*, 1999). Jianwen and Prentice (1995) considered the weighted partial likelihood estimating equations which gains important efficiency when there exists a strong dependency among the failure times events. The working independence model may loss

some efficiency when driving estimates under an incorrect model.

Some authors suggested a combination of marginal approach and copula models, which connect the two marginal distributions through a copula function to construct the joint distribution and use the copula parameter to determine the dependence (Genest and Rivest, 1993; Glidden and Self, 1999). As a consequence, we could model the margins separably from the copula function, which leads to a good way to interpret both covariate effects and dependence relationship.

Another popular approach for correlated failure time data is the frailty model, which considers a random effects with a specified distribution to stably estimate the dependency (Clayton, 1978; Hougaard, 1986). The conditional cumulative hazard function $\Lambda(t)$ under a commonly used frailty model has the form

$$\Lambda(t) = \eta \times G[\Lambda_0(t)e^{X^T\beta}].$$

In the above, $G(.)$ is a prespecified strictly increasing transformation function, $\Lambda_0(.)$ is an unknown baseline cumulative hazard function and η is a frailty variable following a specific distribution with unknown parameters such as gamma distribution or log-normal distribution. This model could estimate the dependency and covariate effects simultaneously, but may require a larger computation. Another interesting method for clustered data is the approach based on within cluster resampling procedure, which repeats to create independent resampled data by choosing one subject from each cluster to estimate the parameters (Williamson *et al.*, 2008; Cong *et al.*, 2007). It's easy to see that this method could avoid modeling the dependence among

subjects and be easy to implement.

Multivariate or bivariate interval-censored data have also been widely studied in literature. For the analysis of multivariate interval-censored data, one should face to the challenge for analyzing interval-censored data as well as considering the correlation structure between correlated failure times. Goggins and Finkelstein (2000), Chen *et al.* (2007), Tong *et al.* (2008), Chen *et al.* (2013) considered the marginal approach for multivariate interval-censored data based on the working independence assumption under the proportional hazards model, proportional odds model, additive hazards model and linear transformation model. These approaches are commonly based on estimation equation, which leads to a direct estimation procedure. However, as mentioned before, the estimates under an incorrect working correlation matrix may be inaccurate. One of the popular methods to consider the estimation of dependence is copula-based approach. Wang *et al.* (2008) and Hu and Xiang (2013) proposed the sieve maximum likelihood estimation under proportional hazards model and semiparametric transformation model with different copula functions for bivariate current status data. Sun and Ding (2019) developed a computationally efficient sieve maximum likelihood estimation procedure for the unknown parameters with a generalized score test for the regression parameter.

Another popular method for estimating both covariate effects and dependence is the frailty model. Chen *et al.* (2009), Wang *et al.* (2015) and Wen and Chen (2013) fitted the frailty proportional hazards model to multivariate or bivariate current status data and interval-censored data. Zhou *et al.* (2017) and Zeng *et al.* (2017) developed the

frailty-based transformation models for bivariate and multivariate interval-censored data. The former applied the sieve maximum likelihood estimation for regression analysis, while the latter considered a nonparametric maximum likelihood estimation procedure.

For clustered interval-censored data, frailty model is the most commonly used method. Chang *et al.* (2007); Wen and Chen (2011) proposed the profile likelihood inference and nonparametric maximum likelihood estimation for clustered current status data, respectively, under the gamma-frailty proportional hazards model. Li *et al.* (2012) considered the frailty additive hazards model with estimation-equation based estimation procedure for clustered interval-censored data. Besides frailty model, copula-based marginal approach and multiple imputation method have also been studied by many authors. Cook and Tolusso (2009) and Kor *et al.* (2013) considered the copula proportional hazards model with a piecewise-constant baseline hazard function for clustered current-status and interval-censored data, respectively. Lam *et al.* (2010) proposed a multiple imputation approach with EM algorithm under the gamma frailty proportional hazards model.

1.3 Regression Analysis of Failure Time Data with a Cured Subgroup

In traditional survival analysis, the typical assumption is that all the study subjects will experience the failure events of interest eventually if the follow-up time is long enough. However, this assumption may not hold when some of the subjects may not

experience or be susceptible to the failure event(Pierce *et al.*, 1979). Farewell (1982) first considered such subjects as a long-term survival and proposed a mixture model combining logistic model and parametric survival model with Weibull distribution. The survival function under mixture model usually takes the form

$$S(t) = \pi + (1 - \pi)\, S_u(t)\,. \tag{1.3}$$

In the above, π denotes the probability for a subject to be cured, and $S_u(t)$ is the survival function for uncured population which could follow a traditional survival model involving parametric model and semiparametric model.

For right-censored data, Kuk and Chen (1992) combined logistic regression with the proportional hazards model and estimated regression parameters based on marginal likelihood with Monte Carlo approximation. Lu and Ying (2004) considered the semiparametric transformation cure model and constructed generalized estimating equations for estimating regression parameters. Ma (2009) and Lam and Xue (2005) proposed the semiparametric AFT model and proportional hazards model for current status data, respectively. For interval-censored data, Ma (2010) and Zhou *et al.* (2016) considered the mixture proportional hazards cure model. The former proposed the a maximum likelihood estimation procedure, while the latter used multiple imputation approach to obtain parameter and variance estimates. The mixture model models the effects of covariate on the cure rate and the failure time of interest separately. As a consequence, we could assume different covariate effects for cure rate and failure risk. When analyzing correlated survival data, one needs to consider both correlation between some subjects and the cure rate. It's natural to consider the approaches

mentioned in Section 1.2 with a mixture cure model. Peng *et al.* (2007), Niu *et al.* (2018), Chen and Lu (2012) and Yu and Peng (2008) discussed estimating equation approaches under mixture cure model for multivariate right-censored and interval-censored data. Su and Lin (2019) considered a copula-based approach under the Cox and ACT mixture cure model for clustered right-censored data. Peng and Taylor (2011), Li and Ma (2010) and Lam and Wong (2014) proposed the maximum likelihood estimation under the mixture cure model by using a frailty variable to measure the correlation for multivariate right-censored data and clustered interval-censored data.

The non-mixture cure model is another approach commonly used for cure subgroup. Chen *et al.* (1999) pointed that the mixture cure model has some drawbacks from both a frequentist and Bayesian perspective. For example, mixture cure model does not appear to have the traditional survival model structure and describe the underlying biological process generating the failure time. Also, the mixture cure model can yield improper posterior distributions for many non-informative improper priors. To overcome these drawbacks, a non-mixture cure model was studied by many authors (Tsodikov, 1998; Chen *et al.*, 1999; Tsodikov *et al.*, 2003). The ideal behind the non-mixture cure model is that we assume that the cumulative hazard function is bounded and usually replaced by a cumulative distribution function. The survival function under the non-mixture proportional hazards cure model takes the form

$$S(t) = \exp\left\{ - F(t)\, e^{X^T \beta} \right\}. \tag{1.4}$$

where $F(t)$ is a prespecified distribution function and β denotes the unknown parameter. The probability for regarding a subject as cured is the asymptotic value

of survival function when $t \to \infty$, which is $\exp\{-e^{X^T\beta}\}$ under model (1.4). Since the model only involves a single uniform model for covariates, the estimation procedure is usually easy to implement. The study of non-mixture cure rate models for interval-censored data has also been investigated by several authors. Hu and Xiang (2013) and Li *et al.* (2019) considered the semiparametric transformation non-mixture cure model for interval-censored data, while Diao and Yuan (2019) used the same model for current status data. For multivariate survival data, the combination of frailty variable and non-mixture cure model is of most interest for the reason that there is only one frailty variable in the model. Diao and Yin (2012) and Yin (2008) discussed a semiparametric transformation non-mixture cure frailty model for multivariate right-censored data. The former proposed the nonparametric maximum likelihood estimation procedure, while the latter considered the estimation procedure in Bayesian paradigm. In terms of interval-censored data with repeated measurements, Thompson and Chhikara (2003) considered a non-mixture cure frailty model by assuming parametric models for both survival function and cumulative distribution function.

1.4 Outline of the Dissertation

The remainder of this dissertation is organized as follows.

In Chapter 2, we will discuss clustered interval-censored data with a cured subgroup and informative cluster size. To address this, we present a within-cluster resampling method and in the approach, the multiple imputation procedure is applied for estimation of unknown parameters. To assess the performance of the proposed

method, a simulation study is conducted and suggests that it works well in practical situations. Also, the method is applied to a set of real data that motivated this study.

In Chapter 3, we will consider the clustered interval-censored data with a cured subgroup via a non-mixture cure model. We present a maximum likelihood estimation procedure under the semiparametric transformation non-mixture cure model. To estimate the unknown parameters, an expectation maximization (EM) algorithm based on an augmentation of Poisson variable is developed. To assess the performance of the proposed method, a simulation study is conducted and suggests that it works well in practical situations. An application to a set of real data that motivated this study is also provided.

In Chapter 4, we will investigate the bivariate interval-censored data with a cured subgroup. We present a sieve maximum likelihood estimation procedure under the semiparametric transformation non-mixture cure model based on Bernstein polynomials. A simulation study is conducted to assess the finite sample performance of the proposed method, and the result suggests that the proposed model works well. Also, a real data application is provided.

Several directions for future research will be discussed in Chapter 5.

Chapter 2

Regression Analysis of Clustered Interval-censored Failure Time Data under the Proportional Hazards Mixture Cure Model

2.1 Introduction

This chapter discusses regression analysis of clustered interval-censored failure time data with cure fraction and informative cluster size. Clustered interval-censored failure time data occur in many areas and many methods have been proposed for their analysis. For the situation, one needs to deal with two general issues and they are correlated failure time variables and interval censoring. In practice, in addition to them, one may also have to deal with two other issues, cure fraction and informative cluster size. In the following, we will discuss regression analysis of failure time data for which all of these issues may exist and present a within-cluster resampling

estimation approach.

The failure time data with interval censoring naturally occurs in many areas such as epidemiological and medical studies as well as animal carcinogenicity experiments or more generally periodic follow-up studies (Finkelstein (1986); Sun (2006)). For the situation, one cannot observe the exact value of the failure time of interest and instead, can only know that it belongs to an interval. It is easy to see that interval censoring includes right and left censoring as special cases. As we mentioned in Section 1.1, Many authors have investigated the analysis of interval-censored data.

The existence of a cure fraction or cured subgroup in a study population has been discussed by many authors under different set-ups (Kuk and Chen, 1992; Ma, 2010; Zhou *et al.*, 2016). For the situation, unlike the standard or traditional failure time study where it is assumed that all study subjects would experience the failure event of interest eventually if the follow-up time is long enough, some subjects may never experience or not be susceptible to the event. It is well-known that with a curd subgroup, the regular or usual failure time model and approach will not be appropriate, and one type of commonly used methods is the two-component mixture model approach that models the cure rate and the failure risk separately. For example, one such approach is the combination of the logistic and Cox models for the cure rate and the failure risk, respectively.

Clustered failure time data arise in a failure time study when some failure times of interest are correlated due to some common features such as genetic traits or shared environmental factors. For the situation, the study population consists of a number of clusters where the subjects in the same cluster may be related but the subjects in different clusters can be treated to be independent. In addition to the correlation,

another complicated issue that may occur is that the cluster size may be informative or carry some information about the failure time of interest. One such example was given by Williamson *et al.* (2003) about a study of factors associated with the periodontal disease. In the data, the cluster size, the number of teeth of a subject, and the disease status of teeth may be related with each other even given covariates. The references that discussed this issue include Williamson *et al.* (2008) and Cong *et al.* (2007), who proposed a weighted estimating equation (WEE) approach and a within-cluster resampling (WCR) procedure, respectively, for right-censored data. Following them, Chen *et al.* (2016), Zhang and Sun (2010) and Zhao *et al.* (2018) generalized the WEE and WCR methods to interval-censored data under different models. In this chapter, we will develop a WCR approach for such data in the presence of a cured subgroup.

The remainder of the chapter is organized as follows. In Section 2.2, we will first introduce some notation and the assumptions that will be used throughout the chapter and then discuss the resulting likelihood functions. Section 2.3 will present the proposed WCR approach (Hoffman *et al.* (2001)) and in the method, by following Zhou *et al.* (2016), the multiple imputation approach will be employed for the estimation of parameters and their covariance. One advantage of the method is that it can be easily implemented. Some results from a simulation study conducted to assess the finite sample properties of the presented method will be provided in Section 2.4 and suggest that the proposed method works well in practical situations. Section 2.5 applies the approach to a set of real data that motivated this study, and Section 2.6 includes some discussion and concluding remarks.

2.2 Notation, Assumptions and Likelihood Function

Consider a failure time study that consists of N independent clusters with n_i subjects in the ith cluster. Let T_{ij} denote the failure time of interest for the jth individual in the ith cluster, $j = 1, \ldots, n_i$, $i = 1, \ldots, N$, and suppose that for T_{ij}, only an interval $(L_{ij}, R_{ij}]$ is observed such that $L_{ij} < T_{ij} \leq R_{ij}$. Also suppose that for subject (i, j), there exist two vectors of covariates X_{ij} and Z_{ij} to be described below, and define the indicators $\delta_{L_{ij}} = \mathcal{I}(L_{ij} = 0)$, $\delta_{R_{ij}} = \mathcal{I}(R_{ij} = \infty)$ and $\delta_{I_{ij}} = \mathcal{I}(0 < L_{ij} < R_{ij} < \infty)$ with $\mathcal{I}$ denoting the indicator function. Then we have that $\delta_{L_{ij}} + \delta_{R_{ij}} + \delta_{I_{ij}} = 1$ and the observed data have the form $\mathbf{O}_F = \{ (L_{ij}, R_{ij}, \delta_{L_{ij}}, \delta_{R_{ij}}, \delta_{I_{ij}}, X_{ij}, Z_{ij}, n_i); j = 1, \ldots, n_i, i = 1, \ldots, N \}$.

In the following, we will assume that there may exist a cured subgroup in the study population. Let u_{ij} denote the cured indicator for the jth individual in the ith cluster such that $u_{ij} = 0$ if a subject belongs to the cured subgroup and 1 otherwise. Also let $\pi(Z_{ij}) = P(u_{ij} = 1|Z_{ij})$ and assume that $\pi(Z_{ij})$ satisfies

$$logit(\pi(Z_{ij})) = \gamma^T Z_{ij} \, .$$

Furthermore, let $S_u(t|X_{ij})$ denote the survival function of an uncured subject with covariates X_{ij} and assume that $S_u(t|X_{ij})$ can be characterized by the proportional hazards model as

$$S_u(t|X_{ij}) = S_{u0}(t)^{\exp(\beta^T X_{ij})} \, ,$$

where $S_{u0}(t)$ denotes the baseline survival function and β a vector of regression pa-

rameters. Then the survival function of a study subject can be written as

$$S(t|X_{ij}, Z_{ij}) = \{1 - \pi(Z_{ij})\} + \pi(Z_{ij})\, S_u(t|X_{ij})\,, \tag{2.1}$$

That is, we assume that the cure rate can be described by the logistic model and the failure time of interest can be described by the two component mixture cure model. Also X and Z represent the covariates that may affect the cure rate and the failure risk, respectively, and it will be assumed that they can be completely different, share some components or be the same.

For inference about β, γ and $S_{u0}(t)$, first assume that there is only one subject in each cluster or $n_i = 1$ for all i. In this case, it is apparent that a natural approach would be to base the inference on the full likelihood function $L(\beta, \gamma, S_{u0}|\mathbf{O}_F)$ given by

$$\prod_{i=1}^{N} \pi(Z_i)^{1-\delta_{L_i}} \left[1 - S_{u0}(R_i)^{exp(\beta^T X_i)}\right]^{\delta_{L_i}} \left[S_{u0}(L_i)^{exp(\beta^T X_i)} - S_{u0}(R_i)^{exp(\beta^T X_i)}\right]^{\delta_{I_i}}$$

$$\times \left[1 - \pi(Z_i) + \pi(Z_i) \times S_{u0}(L_i)^{exp(\beta^T X_i)}\right]^{\delta_{R_i}}\,.$$

On the other hand, the maximization of the likelihood function above may not be straightforward and to deal with this, one can notice that if the u_i's were known, we would have the pseudo likelihood function $L(\beta, \gamma, S_{u0}|u_i's, \mathbf{O}_F)$ given by

$$\prod_{i=1}^{N} \Bigg\{ \pi(Z_i)^{1-\delta_{L_i}} \left[1 - S_{u0}(R_i)^{exp(\beta^T X_i)}\right]^{\delta_{L_i}} \left[S_{u0}(L_i)^{exp(\beta^T X_i)} - S_{u0}(R_i)^{exp(\beta^T X_i)}\right]^{\delta_{I_i}}$$

$$\times \left[1 - \pi(Z_i)\right]^{(1-u_i)\times\delta_{R_i}} \left[\pi(Z_i)\, S_{u0}(L_i)^{exp(\beta^T X_i)}\right]^{u_i \times \delta_{R_i}} \Bigg\}. \tag{2.2}$$

Furthermore one can decompose it into two parts as

$$L(\beta, \gamma, S_{u0}|u_i\text{'}s, \mathbf{O}_F) = L_1(\gamma) \times L_2(\beta, S_{u0}) \,,$$

where

$$L_1(\gamma) = \prod_{i=1}^{N} \pi(Z_i)^{u_i}(1 - \pi(Z_i))^{1-u_i} \,,$$

and

$$L_2(\beta, S_{u0}) = \prod_{u_i=1} \left[1 - S_{u0}(R_i)^{exp(\beta^T X_i)}\right]^{\delta_{L_i}} \left[S_{u0}(L_i)^{exp(\beta^T x_i)}\right]^{\delta_{R_i}}$$

$$\left[S_{u0}(L_i)^{exp(\beta^T X_i)} - S_{u0}(R_i)^{exp(\beta^T X_i)}\right]^{\delta_{I_i}} \,.$$

This suggests that if the u_i's were known, one could maximize L by maximizing L_1 and L_2 separately and some imputation methods could be used (Zhou *et al.*, 2016).

In general, for interval-censored data with informative cluster size, the main difficulty is not the specification of the joint distribution of the failure times but the joint distribution of failure times and the cluster size. We will develop a within-cluster resampling procedure that does not need this in the next section.

2.3 Within-cluster Resampling Estimation

Note that in the second step, we will iteratively update the parameter estimators and their variance estimators under the scenario that there is only one subject in each cluster. Specifically, in each iteration, we will do multiple imputation m times to obtain the new estimators based on the estimators in the previous iteration. The details of algorithm are described as follows.

Step 1. For the selection of initial estimates, first we set the initial cured indicator $u_i^{(0)}$ to be $1-\delta_{R_i}$ and the failure time $T_i^{(0)}$ to be the midpoint of (L_i, R_i) for uncured subjects and L_i for cured subjects. Then we generate initial estimates of regression parameters by fitting the logistics model with $u_i^{(0)}$ and Z_i to get $\hat{\gamma}^{(0)}$ and $\hat{\Sigma}_\gamma^{(0)}$ and fitting the Cox model for the uncured patients with $\{T_i^{(0)}, 1-\delta_{R_i}, X_i\}$ to get $\hat{\beta}^{(0)}$, $\hat{\Sigma}_\beta^{(0)}$ and $\hat{S}_{u0}^{(0)}$.

Step 2. For $(l+1)$th iteration, where $l = 0, 1, \ldots,$ we impute the cure indicator u_i and the survival time T_i m times based on the results from the lth iteration $(\hat{\gamma}^{(l)}, \hat{\Sigma}_\gamma^{(l)}, \hat{\beta}^{(l)}, \hat{\Sigma}_\beta^{(l)}, \hat{S}_{u0}^{(l)})$. Specifically, in the kth imputation $(k = 1, \ldots, m)$,

Step 2.1. Sample β and γ from the normal distributions $N(\hat{\beta}^{(l)}, \hat{\Sigma}_\beta^{(l)})$ and $N(\hat{\gamma}^{(l)}, \hat{\Sigma}_\gamma^{(l)})$ and denote them as $\tilde{\beta}_{(k)}^{(l+1)}$ and $\tilde{\gamma}_{(k)}^{(l+1)}$.

Step 2.2. Update the conditional probability of being uncured as

$$w_{(k),i}^{(l+1)} = E(u_i | \tilde{\beta}_{(k)}^{(l+1)}, \tilde{\gamma}_{(k)}^{(l+1)}, \hat{S}_{u0}^{(l)}, O_{(k)}^{(l)})$$

$$= 1 - \delta_{R_i} + \delta_{R_i} \frac{\hat{\pi}(Z_i)(\hat{S}_{u0}^{(l)}(T_{(k),i}^{(l)}))^{exp(\tilde{\beta}_{(k)}^{(l+1)T} X_i)}}{1 - \hat{\pi(Z_i)} + \hat{\pi}(Z_i)(\hat{S}_{u0}^{(l)}(T_{(k),i}^{(l)}))^{exp(\tilde{\beta}_{(k)}^{(l+1)T} X_i)}},$$

where $\hat{\pi}(Z_i) = \exp(\tilde{\gamma}_{(k)}^{(l+1)T} Z_i)/[1+\exp(\tilde{\gamma}_{(k)}^{(l+1)T} Z_i)]$, and $T_{(k),i}^{(0)} = T_i^{(0)}$, for $k = 1, \ldots, m$.

Step 2.3. Based on $w_{(k),i}^{(l+1)}$, sample $u_{(k),i}^{(l+1)} \sim Ber(w_{(k),i}^{(l+1)})$.

Step 2.4. Let the censoring indicator $\delta_{(k),i}^{(l+1)} = 1 - \delta_{R_i}$, and generate the failure time $T_{(k),i}^{(l+1)}$ for each subject i as follows: if subject i is right-censored $(R_i = \infty)$, let $T_{(k),i}^{(l+1)} = L_i$; otherwise sample $T_{(k),i}^{(l+1)}$ from $\hat{S}_{u0}^{(l)}(t)^{exp(\tilde{\beta}_{(k)}^{(l+1)} X_i)}$ conditional on $T_{(k),i}^{(l+1)} \in (L_i, R_i]$.

Step 2.5. Fit the logistic model with $u_{(k),i}^{(l+1)}$ in (Step 2.3) and Z_i to get estimators $\hat{\gamma}_{(k)}^{(l+1)}$ and $\hat{\Sigma}_{\gamma_{(k)}}^{(l+1)}$, where $\hat{\gamma}_{(k)}^{(l+1)}$ denotes the maximum likelihood estimate and $\hat{\Sigma}_{\gamma(k)}^{(l+1)}$ the inverse of the Fisher information matrix calculated based on the second order

partial derivatives of the log-likelihood function.

Step 2.6. Fit the Cox model for the uncured patients ($u_{(k),i}^{(l+1)} = 1$) with $\{T_{(k),i}^{(l+1)}, \delta_{(k),i}^{(l+1)}, X_i\}$ and get $\hat{\beta}_{(k)}^{(l+1)}$, $\hat{\Sigma}_{\beta_{(k)}}^{(l+1)}$ and $\hat{S}_{u0(k)}^{(l+1)}$, where $\hat{\beta}_{(k)}^{(l+1)}$ denotes the partial likelihood estimates, $\hat{\Sigma}_{\beta(k)}^{(l+1)}$ the inverse of Fisher information matrix calculated based on the second order partial derivatives of the logarithm of the partial likelihood function, and $\hat{S}_{u0(k)}^{(l+1)}$ the Breslow estimator.

Step 3. At the end of each iteration, update the estimates as follows

1. $\hat{\beta}^{(l+1)} = \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_{(k)}^{(l+1)}$,
2. $\hat{\gamma}^{(l+1)} = \frac{1}{m}\sum_{k=1}^{m} \hat{\gamma}_{(k)}^{(l+1)}$,
3. $\hat{\Sigma}_{\beta}^{(l+1)} = \frac{1}{m}\sum_{k=1}^{m} \hat{\Sigma}_{\beta_{(k)}}^{(l+1)} + (1+\frac{1}{m})\frac{\sum_{k=1}^{m}(\hat{\beta}_{(k)}^{(l+1)} - \hat{\beta}^{(l+1)})(\hat{\beta}_{(k)}^{(l+1)} - \hat{\beta}^{(l+1)})^T}{m-1}$,
4. $\hat{\Sigma}_{\gamma}^{(l+1)} = \frac{1}{m}\sum_{k=1}^{m} \hat{\Sigma}_{\gamma_{(k)}}^{(l+1)} + (1+\frac{1}{m})\frac{\sum_{k=1}^{m}(\hat{\gamma}_{(k)}^{(l+1)} - \hat{\gamma}^{(l+1)})(\hat{\gamma}_{(k)}^{(l+1)} - \hat{\gamma}^{(l+1)})^T}{m-1}$,
5. $\hat{S}_{u0}^{(l+1)} = \frac{1}{m}\sum_{k=1}^{m} \hat{S}_{u0(k)}^{(l+1)}$.

Step 4. Repeat Steps 2 and 3 until $max\{((\hat{\beta}^{(l)} - \hat{\beta}^{(l-1)}) \odot (\hat{\beta}^{(l)} - \hat{\beta}^{(l-1)}), (\hat{\gamma}^{(l)} - \hat{\gamma}^{(l-1)}) \odot (\hat{\gamma}^{(l)} - \hat{\gamma}^{(l-1)}))\} < 0.001$, where $\odot$ denotes the element-wise multiplication, or $l > 500$.

Note that in the above, to accomplish the sampling procedure in Step 2.4, we can first denote the ordered and distinct time points of all finite L_i's and R_i's as $t_{(1)} < t_{(2)} < \cdots < t_{(v)}$. Then there will be v_i time points from $\{t_{(.)}\}$ between $(L_i, R_i]$ for each subject i and $L_i = t_0^* < t_1^* < \cdots < t_{v_i}^* = R_i$. Next we calculate the corresponding probability mass $\{p_1, \ldots, p_{n_i}\}$, where $p_s = \hat{S}_{u0}^{(l)}(t_{s-1}^*)^{exp(\tilde{\beta}_{(k)}^{(l+1)} X_i)} - \hat{S}_{u0}^{(l)}(t_s^*)^{exp(\tilde{\beta}_{(k)}^{(l+1)} X_i)}$.

At last we sample $T_{(k),i}^{(l+1)}$ from $\{t_1^*, \ldots, t_{n_i}^*\}$ based on the probabilities $\{p_1, \ldots, p_{v_i}\}$. If there is no other time points from $\{t_{(.)}\}$ between $(L_i, R_i]$ $(v_i = 1)$, we sample $T_{(k),i}^{(l+1)}$ from $U(L_i, R_i)$. For the estimation of the baseline cumulative hazard function, the Breslow's method can be used, and the regression parameters are estimated by the average of the estimates from the m imputations. For the variance estimation, one can employ the weighted average of within imputation variance and between imputation variance with an additional weight $1/m$ used to take into account of the finite number of imputations. Note that for the size of multiple imputations m, it does not need to be very large and by following the suggestion of Zhou *et al.* (2016), we use $m = 10$ in the numerical study below.

Let Q be an integer, representing the number of the resampling processes in the first step. In other words, we repeat the resampling process Q times to generate Q resamples. More specifically, in the qth resampling $(1 \leq q \leq Q)$, we draw a subject randomly from each cluster, giving a sample of independent observations $\mathbf{O}^q = \{ (L_i^q, R_i^q, \delta_{L_i}^q, \delta_{R_i}^q, \delta_{I_i}^q, X_i^q, Z_i^q),\ i = 1, \ldots, N \}$. Then one applies the estimation procedure described in the second step to $\mathbf{O}^q$ and obtain the estimators of β and γ, denoted by $\hat{\beta}^q$ and $\hat{\gamma}^q$, along with the associated covariance estimators $\hat{\Sigma}_\beta^q$ and $\hat{\Sigma}_\gamma^q$, respectively. The final WCR estimators are given by

$$\hat{\beta}_{WCR} = \frac{1}{Q} \times \sum_{q=1}^{Q} \hat{\beta}^q, \qquad \hat{\gamma}_{WCR} = \frac{1}{Q} \times \sum_{q=1}^{Q} \hat{\gamma}^q \,,$$

with their covariance matrices estimated by

$$\hat{\Sigma}_{WCR}^{\beta} = \frac{1}{Q} \times \sum_{q=1}^{Q} \hat{\Sigma}_\beta^q - \frac{1}{Q-1} \times \sum_{q=1}^{Q} (\hat{\beta}^q - \hat{\beta}_{WCR})(\hat{\beta}^q - \hat{\beta}_{WCR})^T \,,$$

and

$$\hat{\Sigma}^{\gamma}_{WCR} = \frac{1}{Q} \times \sum_{q=1}^{Q} \hat{\Sigma}^{q}_{\gamma} - \frac{1}{Q-1} \times \sum_{q=1}^{Q} (\hat{\gamma}^{q} - \hat{\gamma}_{WCR})(\hat{\gamma}^{q} - \hat{\gamma}_{WCR})^{T} ,$$

respectively. One can expect that the proposed estimators are consistent and their distributions can be asymptotically approximated by the normal distributions.

2.4 A Simulation Study

In this section, we present some results obtained from a simulation study conducted to assess the finite sample performance of the estimation procedure proposed in the previous sections with the focus on the estimation of the regression parameters β and γ. First we assumed that the covariates X_i's and Z_i's are either one- or two-dimensional and generated them in two different ways for each situation. More specifically, for the one-dimensional situation, we assumed that $X = x_1$ and $Z = (z_0, z_1)$ with setting $z_0 = 1$ and either generating both x_1 and z_1 independently from the Bernoulli distribution with the probability of success 0.5 (Case 1) or generating x_1 in the same way as above but taking $z_1 = x_1$ (Case 2). For the two-dimensional situation with the covariates $(X, Z) = (x_1, x_2, z_0, z_1, z_2)$, we also set $z_0 = 1$ and either generated (x_1, x_2) and (z_1, z_2) independently from the uniform distribution over $U(0, 2)$ and the Bernoulli distribution with the probability of success 0.5 (Case 1) or only generated (x_1, x_2) in the same way as above but taking $(z_1, z_2) = (x_1, x_2)$ (Case 2).

To generate the underlying data, given covariates, we first generated the cured indicators u_i's based on the logistic model

$$\pi(Z) = p(u = 1) = \frac{\exp(Z'\gamma)}{1 + \exp(Z'\gamma)} .$$

Then for the generation of the cluster size n_i and the true failure time T_{ij} for uncured subjects, we first generated the latent variables $\{w_i = (\delta_i/E_i)^{(1-\alpha)/\alpha}, i = 1, .., N\}$. Here E_i denotes a random number following the exponential distribution with mean 1, α is a positive constant between 0 and 1, and

$$\delta_i = \left[\frac{sin(\alpha U_i)}{sin(U_i)}\right]^{\frac{1}{1-\alpha}} \frac{sin[(1-\alpha)U_i]}{sin(\alpha U_i)}$$

with U_i being generated from the uniform distribution over $(0, \pi)$. Given w_i, we generated n_i from either the Binomial distribution $B(5, 0.75)$ if w_i is less than or equal to the median of the positive stable distribution or the Binomial distribution $B(5, 0.25)$ otherwise. If a zero was generated, a new number was generated.

For the generation of the observed data, we assumed that the marginal survival function of T_{ij} has the form

$$S_u(t_{ij}|w_i, X, Z) = S_{u0}(t_{ij})^{w_i \times \exp(\beta^{*T} X)}$$

given w_i, X_i and Z_i, which gives

$$S_u(t_{ij}|X_i, Z_i) = S_{u0}(t_{ij})^{\exp(\beta^T X_i)}$$

with $\beta = \beta^* \times \alpha$ after integrating out w_i. In the above, we took $S_{u0}(t)$ to be the Weibull survival distribution with the shape parameter of 2 and the scale parameter of 1. To generate the censoring interval for T_{ij}, we considered the sequence of the observation times $0 < Y_{ij} < Y_{ij} + \tau < ... < Y_{ij} + p \times \tau < \infty$ with Y_{ij} generated from the uniform distribution over $(0, v)$ and v, τ and p being some constants chosen to give the desired

censoring percentages. If $u_{ij} = 0$, we set $(L_{ij}, R_{ij}) = (Y_{ij} + p \times \tau, \infty)$. For the subject with $u_{ij} = 1$, we set $(L_{ij}, R_{ij}) = (0, Y_{ij})$, $(Y_{ij} + p \times \tau, \infty)$, or $(Y_{ij} + t \times \tau, Y_{ij} + (t+1) \times \tau)$ if $T_{ij} < Y_{ij}$, $T_{ij} > Y_{ij} + p \times \tau$, or $Y_{ij} + t \times \tau < T_{ij} < Y_{ij} + (t+1) \times \tau$ for $t \in (1, .., p-1)$, corresponding to the left-censored, right-censored or interval-censored observation, respectively. The results given below are based on $N = 400$ and $Q = 100$ with 1000 replications.

Table 2.1 presents the results obtained based on the simulated data for the estimation of β and γ for the one-dimensional situation with the true values of the parameters being $\beta_1^* = 1.25$, $\alpha = 0.8$ and $\gamma = (0, 1)$ along with $\tau = 0.2$ and $p = 10$, In the table, we calculated the estimated bias given by the average of the estimates minus the true value (Bias), the sample standard error (SSE), the average of the estimated standard errors (ASE) and the 95% empirical coverage probability (CP). One can see that the proposed estimators seem to be unbiased and the estimated standard error appears to be reasonable. Also, the results on the coverage probabilities indicate that the normal approximation to the distribution of the proposed estimators seems to be appropriate.

The simulation results on the estimation of β and γ for the two-dimensional situation are given in Table 2.2 with the true values of the parameters β^*, α and γ being $(1.25, 1.25)$, 0.8 and $(0, 1, 1)$, respectively. Also, here we set $\tau = 0.2$ and $p = 5$. It seems that they gave similar conclusions as above and again suggest that the proposed estimation approach appears to work reasonably well for the situations considered. For the assessment of the normal approximation to the distributions of the proposed estimators of the regression parameters, we studied the quantile plots of the standardized estimates against the standard normal random variables and Figure 2.1, Figure

2.2, Figure 2.3 display them corresponding to the results given in Table 2.1 and Table 2.2 under Case 2, respectively. They suggest that the approximation seems to be reasonable. We also considered other set-ups and obtained similar results.

2.5 An Application

Now we apply the estimation procedure proposed above to a set of clustered interval-censored failure time data, the hypobaric decompression sickness data (HDSD) arising from a study conducted by the National Aeronautics and Space Administration (Conkin *et al.* (1992)). In the study, the volunteers were recruited and measured at different time points for their times to onset of the grade IV VGE, a high level of venous gas emboli that can be treated as a marker of the decompression sickness. Furthermore, each volunteer could take part in the study more than once and thus we have clustered interval-censored data with the cluster size being the number of experiments in which each volunteer participated in, which ranges from 1 to 13. The air bubble in venous blood could cause the decompression sickness in hypobaric environments.

The data set includes 548 records in total arising from 238 volunteers with ages between 20 and 54. In addition, there exists some information on three other covariates, Genders with 1 denoting males and 0 for females, TR360 measuring the decompression stress, and NOADYN indicating by 1 if the individual was ambulatory. To see if there may exist a cured subgroup, we obtained the Kaplan-Meier estimator by assuming that all observations were independent and present it in Figure 2.4. One can see from the figure that the right-censoring rate is over 70%, indicating the possible exis-

tence of a cured fraction. Figure 2.5 gives the percentage or the rate of the observed grade IV VGE within each cluster against the cluster size and indicates that it seems that the two were negatively correlated. In other words, we may have informative cluster size and it is better to apply the proposed estimation approach to determine the covariate effects on the time to onset of the grade IV VGE.

For the analysis, we considered different set-ups in terms of selecting X and Z among the four covariates and also tried different numbers of resamples and multiple imputations. Based on the similarities of the results, we present in Table 2.3 the estimation results obtained from three set-up with $m = 10$ and $Q = 100$. In the first set-up, all of the four variables were assumed to have some possible effects on the failure time, and only Age and Gender were considered possibly to have some effects on the cure risk. In the second set-up, based on the results from the first set-up, we removed Age and Gender from the failure time model and kept the other the same but added the interaction between TR360 and MOADYN by following the suggestion of Lam and Wong (2014) to the failure time model. The third set-up is the same as the first one except that the covariate TR360 was added to the cure risk model. We considered other values for m and Q and obtained similar conclusions.

The analysis results indicate that Age seems to have some mild effects on the cure risk and so does Gender if not considering TR360. When TR360 was considered for the cure risk, Gender's effect disappeared and the results suggest that the environment with more decompression stress may lead to a high probability for the onset of grade IV VGE. With respect to the onset risk of grade IV VGE for uncured subjects, among all covariates, only NOADYN seems to have some significant effects, and the subject who was ambulatory had higher risk of developing grade IV VGE, which are similar

to those obtained by Lam and Wong (2014). Unlike Lam and Wong (2014) who assumed the independent cluster size, however, the results indicate that it seems that there was no significant interaction effect between TR360 and NOADYN.

2.6 Concluding Remarks

In this chapter, we discussed regression analysis of clustered interval-censored failure time data in the presence of a cured fraction and informative cluster size, and for the problem, a WCR-based multiple imputation approach was developed and investigated. Unlike the existing methods, the presented approach can deal with all of four issues together and can be easily implemented. The simulation study was performed and suggested that the approach seems to work well for practical situations. Also, the method was applied to a real set of clustered interval-censored data.

Note that for the multiple imputation, we employed the asymptotic normal data augmentation (ANDA) in the proposed approach. Instead, one may use the Poor man's data augmentation (PMDA). On the other hand, Pan (2000) pointed out that the result from the PMDA is asymptotically equivalent to that based on the ANDA but the PMDA can underestimate the variance when data have a relatively large proportion of right-censored observations, which is usually the case when there exists a cured subgroup.

Table 2.1: Simulation results with one-dimensional covariates based on N=400, Q=100 and replication=1000.

		β_1	γ_0	γ_1
Case 1:	True	1.25*0.8	0	1
	Bias	0.0034	-0.0014	0.0082
	SSE	0.0916	0.0912	0.1463
	ASE	0.0896	0.0946	0.1432
	CP	0.925	0.947	0.943
		β_1	γ_0	γ_1
Case 2:	True	1.25*0.8	0	1
	Bias	0.0009	-0.0037	0.0115
	SSE	0.0924	0.0976	0.1437
	ASE	0.0917	0.0926	0.1414
	CP	0.926	0.940	0.938

Table 2.2: Simulation results with two-dimensional covariates based on N=400, Q=100 and replication=1000.

		β_1	β_2	γ_0	γ_1	γ_2
Case 1	True	1.25*0.8	1.25*0.8	0	1	1
	Bias	0.0029	0.0051	0.0019	0.0194	0.0097
	SSE	0.0867	0.0983	0.1806	0.1649	0.1961
	ASE	0.1101	0.1134	0.1766	0.1610	0.1837
	CP	0.976	0.969	0.936	0.940	0.933
		β_1	β_2	γ_0	γ_1	γ_2
Case 2	True	1.25*0.8	1.25*0.8	0	1	1
	Bias	0.0071	0.0035	0.0207	0.0100	0.0133
	SSE	0.0944	0.0960	0.1919	0.1604	0.1902
	ASE	0.1123	0.1173	0.1603	0.1567	0.1783
	CP	0.968	0.972	0.854	0.936	0.910

Table 2.3: Estimated treatment effects for the HDSD.

Set-up	Model	Covariate	Estimated effects	SD	p-value
1	Cure	Age	0.0451	0.0249	0.0701
		Gender	1.1746	0.4608	0.0108
	Survival	Age	-0.0085	0.0233	0.7140
		Gender	-0.2869	0.4689	0.5406
		TR360	0.3609	0.3053	0.2372
		NOADYN	1.4307	0.5926	0.0157
2	Cure	Age	0.3374	0.1751	0.0540
		Gender	1.1533	0.4616	0.0124
	Survival	TR360	-0.0747	0.5390	0.8897
		NOADYN	1.2555	0.6274	0.0354
		TR360NOADYN	0.6005	0.6448	0.3516
3	Cure	Age	0.3767	0.2130	0.0769
		Gender	0.9074	0.6056	0.1340
		TR360	0.9912	0.3041	0.0011
	Survival	Age	-0.0428	0.1771	0.8089
		Gender	0.0109	0.5652	0.9845
		TR360	0.0413	0.3321	0.9010
		NOADYN	1.7504	0.6150	0.0044

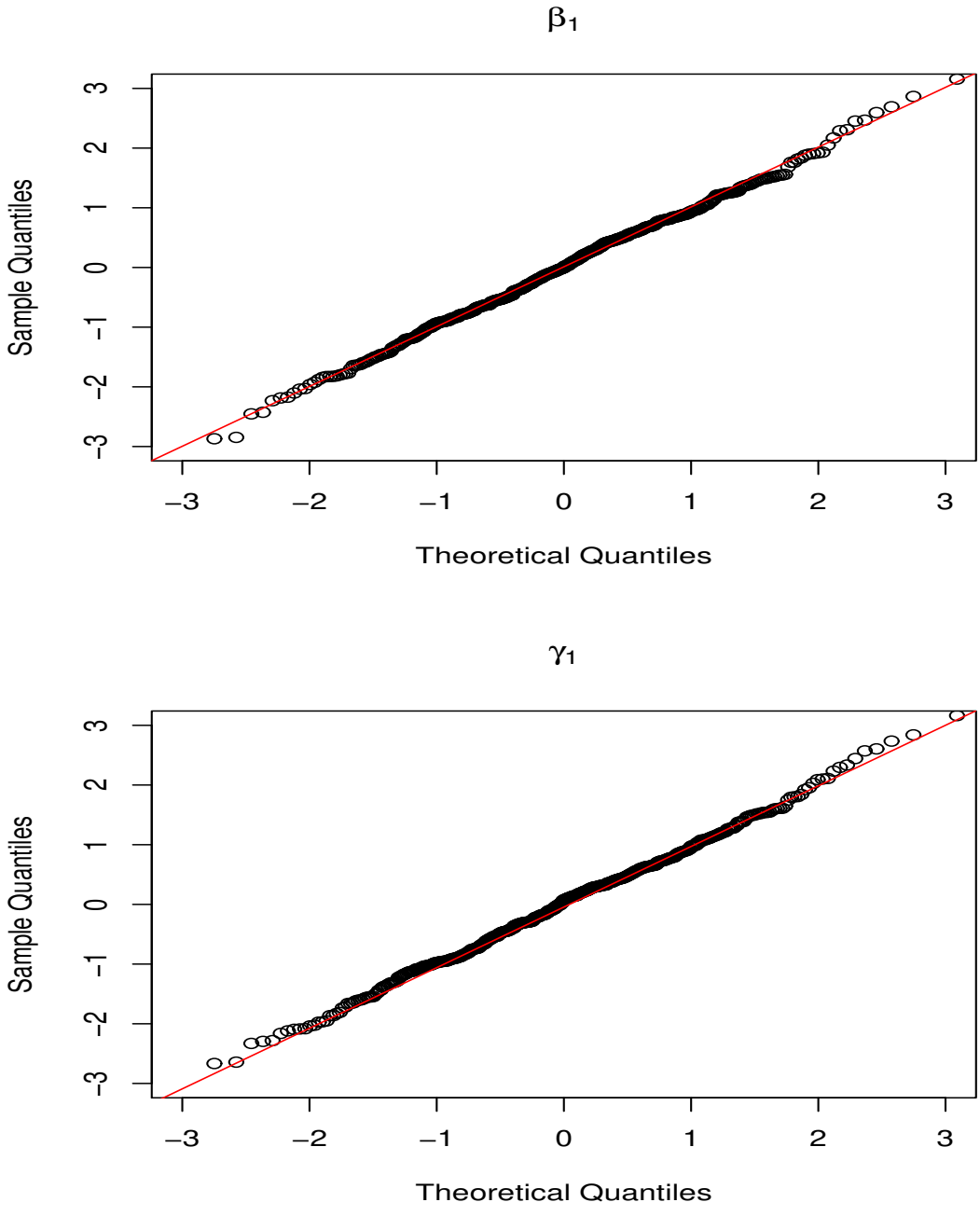

Figure 2.1: Quantile plots of the standardized $\hat{\beta}$ and $\hat{\gamma}$ for the one-dimensional covariate situation

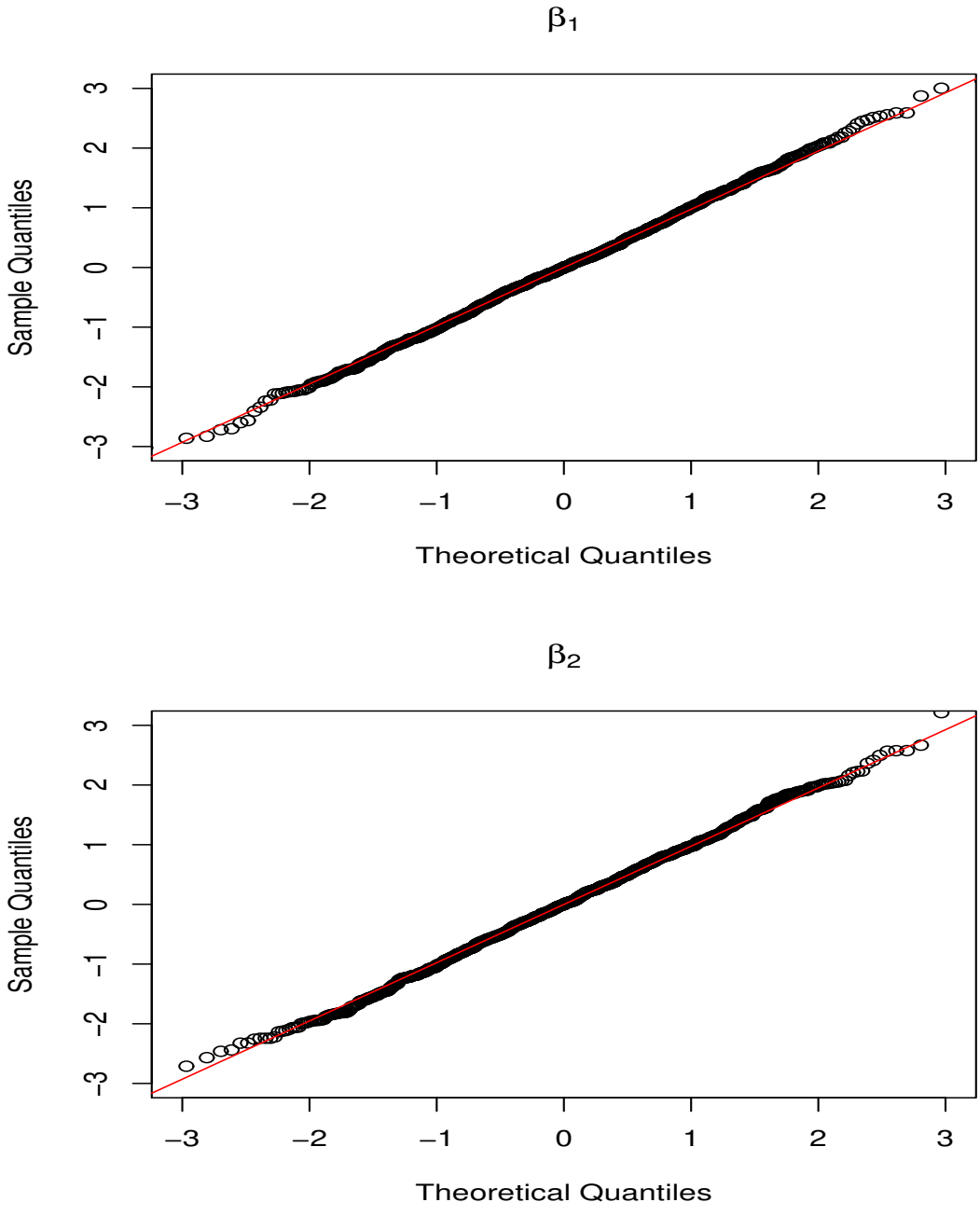

Figure 2.2: Quantile plots of the standardized $\hat{\beta}$ for the two-dimensional covariate situation

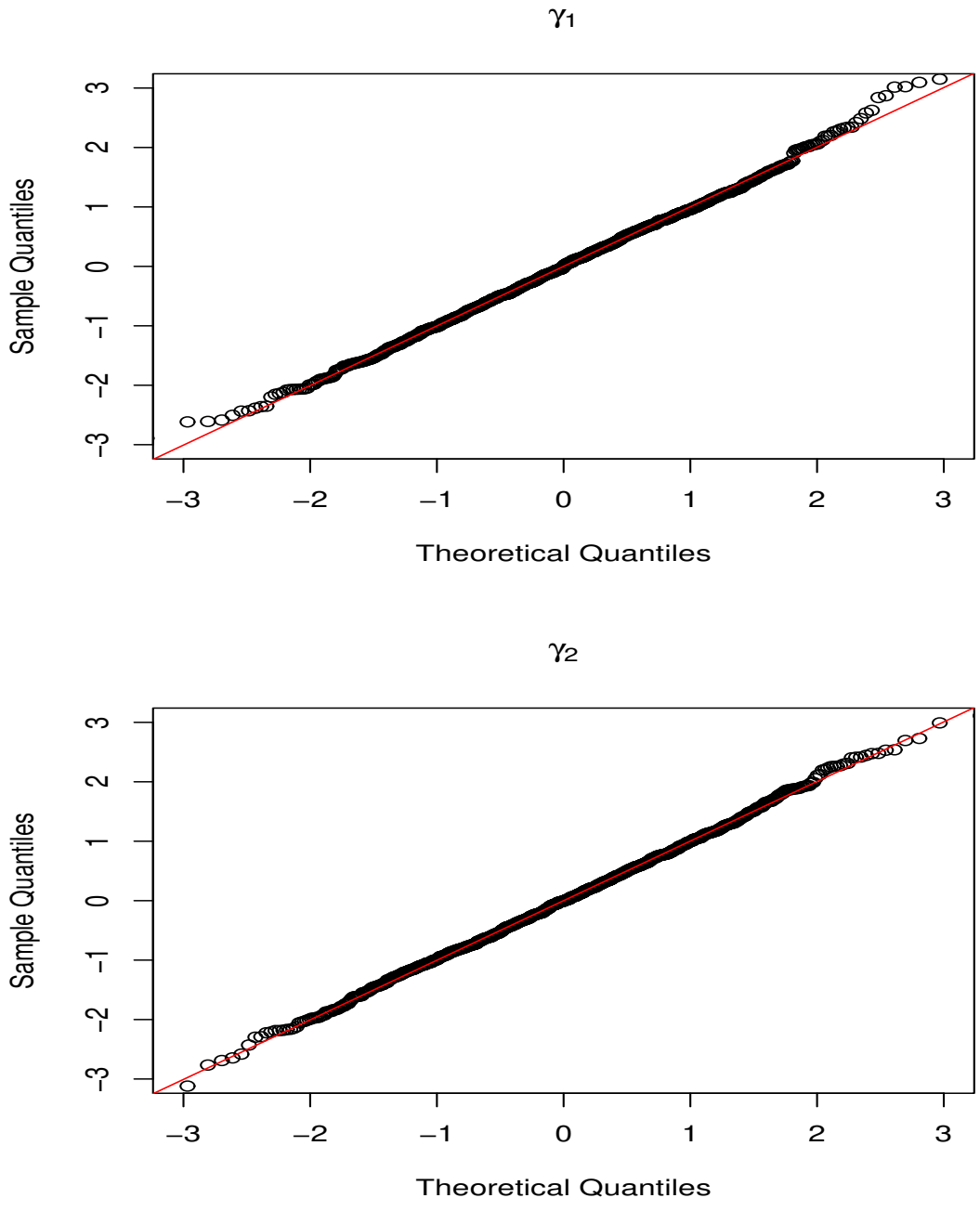

Figure 2.3: Quantile plots of the standardized $\hat{\gamma}$ for the two-dimensional covariate situation

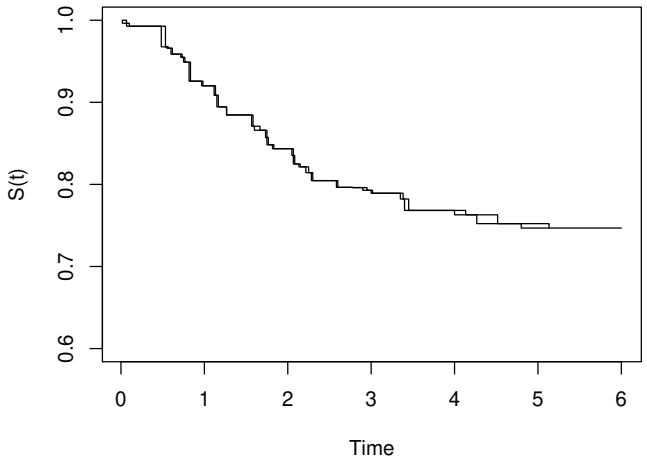

Figure 2.4: KM estimate of the survival function for the HDSD

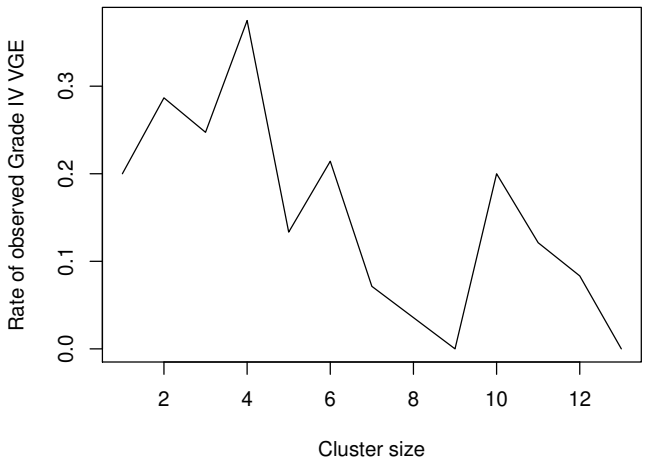

Figure 2.5: The observed onset rate of grade IV VGE against the cluster size

Chapter 3

Regression Analysis of Clustered Interval-censored Failure Time Data under the Semiparametric Transformation Non-mixture Cure Model

3.1 Introduction

In this chapter, we focus on regression analysis of clustered interval-censored failure time data in the presence of a cured subgroup or fraction under a semiparametric transformation non-mixture cure model. In the following, we will discuss the proposed model and present a maximum likelihood estimation procedure with a developed EM algorithm.

Clustered interval-censored failure time data arise when some failure times of inter-

est are correlated and clustered into small groups due to some common characteristics such as clinical sites or environmental factors. In this situation, the subjects from the same cluster are usually related, but the subjects in different clusters can be treated as being independent. Among others, Kor *et al.* (2013) and Lam *et al.* (2010) discussed regression analysis of clustered interval-censored data under the framework of the proportional hazards model. Chen *et al.* (2016) and Zeng *et al.* (2017) also considered the same problem but under the additive hazards model and the semiparametric transformation model, respectively.

An underlying assumption behind all of the methods described above is that all study subjects are supposed to eventually experience the failure event of interest if the follow-up time is long enough. However, as discussed by many authors, this assumption may not hold sometimes since some study subjects may never experience or not be susceptible to the failure event of interest, and the methods that do not take this into account would not be valid (Hu and Xiang, 2013; Kuk and Chen, 1992). These subjects are usually considered as cured and to belong to a cured fraction or subgroup. An example of such situations is given by a study conducted by the National Aeronautics and Space Administration on the time to onset of the grade IV venous gas embolism. Due to the nature of the study, only clustered interval-censored data are available with 77% right-censored observations, indicating the possible presence of a cured subgroup. To further see this, Figure 1 presents the estimated survival functions and again suggests that there may exist a cured friction. More details and discussion on this will be given below.

To deal with the existence of a cured subgroup, two types of approaches are commonly used. One is the two-component mixture cure model-based approach and the

other is the non-mixture cure model-based approach. The former models the effects of covariate on the cure rate and the failure time of interest separately and thus does not have the typical structure of traditional survival models like the proportional hazards or proportional odds model (Kuk and Chen, 1992; Ma, 2010; Zhou *et al.*, 2016). In contrast, the latter employs a single, uniform model for covariate effects (Chen *et al.*, 1999; Hu and Xiang, 2013; Li *et al.*, 2019). Several authors have investigated regression analysis of clustered interval-censored data with a cured fraction (Lam *et al.*, 2010; Lam and Wong, 2014; Xiang *et al.*, 2011) but all under two-component mixture cure models. In the following, we will propose a maximum likelihood estimation approach under a class of semiparametric transformation non-mixture cure models.

The remainder of this chapter is organized as follows. In Section 3.2, we will first introduce some notation and assumptions used throughout the chapter and then discuss the resulting likelihood function. The proposed maximum likelihood estimation approach is described in Section 3.3 along with a novel EM algorithm with the use of the Poisson variable-based data augmentation. Some results from a simulation study conducted to assess the finite sample properties of the proposed method are presented in Section 3.4 and they indicate that the approach works well in practice. Section 3.5 applies the approach to the National Aeronautics and Space Administration study described above that motivated this investigation, and Section 3.6 provides some discussion and concluding remarks.

3.2 Notation and Assumptions

Consider a failure time study that consists of N independent clusters with n_i subjects in the ith cluster. For $j = 1, \ldots, n_i$, $i = 1, \ldots, N$, let T_{ij} denote the failure time of interest for the jth individual in the ith cluster and X_{ij} a vector of associated covariates. Suppose that for T_{ij}, only an interval $(L_{ij}, R_{ij}]$ is observed such that $L_{ij} < T_{ij} \leq R_{ij}$ and the T_i's within the same cluster may be correlated. Then the observed data have the form $\mathbf{O}_F = \{ (L_{ij}, R_{ij}, \delta_{ij}, X_{ij}, n_i); j = 1, \ldots, n_i, i = 1, \ldots, N \}$. In the following, we will assume that the censoring is independent (Sun, 2006).

To describe the covariate effects, suppose that there exists a vector of latent variables with mean zero denoted by b_i and given X_{ij} and b_i, the survival function of T_{ij} has form

$$S(t|X_{ij}, b_i) = \exp\left\{ - G\left[F(t)\, e^{X_{ij}^T\beta + Z_{ij}^T b_i}\right] \right\}. \tag{3.1}$$

In the above, $G(\cdot)$ denotes a prespecified increasing transformation function, Z_{ij} could be same as or part of X_{ij}, $F(\cdot)$ is an unknown cumulative distribution function, and β is a vector of regression parameters. Also, suppose that given b_i, the T_{ij}'s within the ith cluster are independent. Then under the independent censoring assumption, the observed data likelihood function has the form

$$\begin{aligned} L(\beta, F) = \prod_{i=1}^{n} \Bigg\{ \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} \Bigg[& \exp\left\{-G[F(L_{ij})\, e^{X_{ij}^T\beta + Z_{ij}^T b_i}]\right\} - \\ & \delta_{ij} * \exp\left\{-G[F(R_{ij})\, e^{X_{ij}^T\beta + Z_{ij}^T b_i}]\right\} \Bigg] \times f_b(b_i)\, \mathrm{d}b_i \Bigg\}, \end{aligned} \tag{3.2}$$

where $\delta_{ij} = \mathcal{I}(R_{ij} < \infty)$ and $f_b(b_i)$ denotes the density function of the b_i's. In the following, we will assume that f_b is the multivariate normal density function with the

covariance matrix $\Sigma(\eta)$ depending on the unknown parameter η and some comments on this will be given below.

As mentioned above, both Lam *et al.* (2010) and Chen *et al.* (2016) discussed the problem considered here under the two-component mixture cure model. The former proposed a multiple imputation procedure and the latter gave a resampling method. Both methods simply the problem by removing either interval censoring or clustering, respectively. In the following, we will develop a maximum likelihood approach that directly maximizes the likelihood function (3.2). In particular, for the maximization, an EM algorithm will be developed that employs the Poisson variable-based data augmentation and gives a closed-form solution for estimation of the cumulative distribution function F.

3.3 Maximum Likelihood Estimation

Before discussing the maximization of the likelihood function $L(\beta, F)$ given in (3.2), first note that one can rewrite the transformation function $G(\cdot)$ as

$$G(t) = -\log \int_0^\infty \exp(-\xi t)\, f_\xi(\xi)\, \mathrm{d}\xi$$

with respect to the density function $f_\xi(\cdot)$ of a frailty variable ξ with the support on $[0, +\infty)$. In particular, by letting $f_\xi(\cdot)$ be the gamma density function with mean 1 and variance r, we obtain the class of logarithmic transformations $G(x) = \log(1 + rx)/r$, which gives the proportional odds model and the proportional hazards model with $r = 1$ and $r = 0$, respectively. In consequence, the likelihood function $L(\beta, F)$

can be rewritten as

$$L(\beta, F) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{J_i} \int_{0}^{\infty} \left\{ \exp\left[-F(L_{ij}) e^{X_{ij}^T \beta + Z_{ij}^T b_i} \times \xi_{ij} \right] - \right.\right.$$
$$\left.\left. \delta_{ij} * \exp\left[-F(R_{ij}) e^{X_{ij}^T \beta + Z_{ij}^T b_i} \xi_{ij} \right] \right\} * f_\xi(\xi_{ij}) \mathrm{d}\xi_{ij} \right\} \times f_b(b_i)\ \mathrm{d}b_i\,. \qquad (3.3)$$

For the maximization, we will take the nonparametric approach with respect to F by treating it as a step function with jump p_k at the time point t_k. Here $t_1 < ... < t_K$ denote the ordered distinct observation times of the L_{ij}'s and R_{ij}'s with $R_{ij} < \infty$ and it is assumed that $\sum_{k=1}^{K} p_k = 1$. Then we can rewrite the likelihood function above as

$$L(\beta, F) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{J_i} \int_{0}^{\infty} \left[\exp\left(- \sum_{t_k \le L_{ij}} p_k e^{X_{ij}^T \beta + Z_{ij}^T b_i} \times \xi_{ij} \right) - \right.\right.$$
$$\left.\left. \delta_{ij} * \exp\left(- \sum_{t_k \le R_{ij}} p_k\, e^{X_{ij}^T \beta + Z_{ij}^T b_i}\, \xi_{ij} \right) \right] * f_\xi(\xi_{ij})\, \mathrm{d}\xi_{ij} \right\} \times f_b(b_i)\ \mathrm{d}b_i\,. \qquad (3.4)$$

In the following, we will develop an EM algorithm for maximizing the likelihood function above.

To describe the EM algorithm, let W_{ijk} denote the Poisson variable with mean $p_k \exp(X_{ij}^T \beta + Z_{ij}^T b_i)\, \xi_{ij}$ and define $A_{ij} = \sum_{t_k \le L_{ij}} W_{ijk}$ and $B_{ij} = \delta_{ij} \sum_{L_{ij} \le t_k \le R_{ij}} W_{ijk}$. Then one can easily show that given ξ_{ij} and b_i, the joint probability of $A_{ij} = 0$ and $B_{ij} > 0$ is given by

$$\exp\left(- \sum_{t_k \le L_{ij}} p_k\, e^{X_{ij}^T \beta + Z_{ij}^T b_i} \times \xi_{ij} \right) - \delta_{ij} * \exp\left(- \sum_{t_k \le R_{ij}} p_k\, e^{X_{ij}^T \beta + Z_{ij}^T b_i} \xi_{ij} \right),$$

the term inside the likelihood function given in (3.4). In other words, the likelihood

function given in (3.4) can be written as the function of the joint probability of $(A_{ij} = 0, B_{ij} > 0)$. This suggests that for the EM algorithm, we can treat the W_{ijk}'s, ξ_{ij}'s and b_i's as the complete data and the resulting complete-data log likelihood has the form

$$\begin{aligned} l_c(\beta, F) = & \sum_{i=1}^{n} \Bigg\{ \sum_{j=1}^{J_i} \Bigg[\sum_{k=1}^{K} \Big\{ W_{ijk} * (\log(p_k) + X_{ij}^T\beta + Z_{ij}^T b_i + \log(\xi_{ij})) \\ & - p_k \, e^{X_{ij}^T\beta + Z_{ij}^T b_i} \xi_{ij} - \log(W_{ijk}!) \Big\} + \log(f_\xi(\xi_{ij})) \Bigg] \\ & + \Bigg\{ -\frac{d}{2} - \frac{1}{2}\log(|\Sigma|) - \frac{b_i^T \Sigma^{-1} b_i}{2} \Bigg\} \Bigg\} \end{aligned} \tag{3.5}$$

with $A_{ij} = 0$, $B_{ij} > 0$ and $\sum_{k=1}^{K} p_k = 1$.

In the M-step of the EM algorithm, we need to maximize $l_c(\beta, F) - \lambda\,(\sum_{k=1}^{K}\, p_k - 1)$, where λ is the Lagrange multiplier. For this, one can derive and solve the following score equations

$$S_\beta = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \Bigg\{ -\hat{E}(e^{X_{ij}^T\beta + Z_{ij}^T b_i} \xi_{ij}) X_{ij} + \sum_{k=1}^{K} \hat{E}(W_{ijk}) X_{ij} \Bigg\} = 0, \tag{3.6}$$

$$S_{p_k} = \sum_{i=1}^{n} \sum_{j=1}^{J_i} \Bigg\{ -\hat{E}(e^{X_{ij}^T\beta + Z_{ij}^T b_i} \xi_{ij}) + \hat{E}(W_{ijk}) \frac{1}{p_k} \Bigg\} - \lambda = 0, \tag{3.7}$$

$$S_\lambda = \sum_{k=1}^{K} p_k - 1 = 0, \tag{3.8}$$

with respect to β, p_k and λ, where $\hat{E}(\cdot)$ denotes the conditional expectation given

the observed data. In particular, by combining the equations (3.7) and (3.8), we can obtain a closed-form solution for p_k as

$$\hat{p}_k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{J_i}\hat{E}(W_{ijk})}{\sum_{i=1}^{n}\sum_{j=1}^{J_i}\sum_{k=1}^{K}\hat{E}(W_{ijk})}.$$

For the covariance matrix Σ, one can simply use the nonparametric estimator given by $\hat{\Sigma} = n^{-1}\sum_{i=1}^{n}\hat{E}\left(b_i b_i^T\right)$.

In the E-step of the EM algorithm, one needs to calculate the conditional expectations $\hat{E}[\xi_{ij}e^{X_{ij}^T\beta+Z_{ij}^T b_i}]$, $\hat{E}[W_{ijk}]$ and $\hat{E}[b_i^T b_i]$ given the observed data. For this purpose, one can employ the following two facts. One is that the joint density function of ξ_{ij} and b_i given the observed data is proportional to

$$\left\{\prod_{j=1}^{J_i}\left\{\exp[-\sum_{t_k\le L_{ij}} p_k e^{X_{ij}^T\beta+Z_{ij}^T b_i}\times\xi_{ij}] - \delta_{ij} * \exp[-\sum_{t_k\le R_{ij}} p_k e^{X_{ij}^T\beta+Z_{ij}^T b_i}\xi_{ij}]\right\} * f_\xi(\xi_{ij})\right\}$$
$$\times (2\pi)^{-d/2}\left|\Sigma\right|^{-1/2}\exp\left\{-\frac{b_i^{\mathrm{T}}\Sigma^{-1}b_i}{2}\right\}. \tag{3.9}$$

The other is the conditional expectation of W_{ijk} given b_i, ξ_{ij} and observed data

$$E(W_{ijk}|b_i,\xi_{ij}) = \delta_{ij}I(L_{ij} < t_k \le R_{ij})\frac{p_k e^{X_{ij}^T\beta+Z_{ij}^T b_i}\xi_{ij}}{1-exp\{-\sum_{L_{ij}<t_l<R_{ij}} p_k e^{X_{ij}^T\beta+Z_{ij}^T b_i}\xi_{ij}\}}$$
$$+\delta_{ij} * I(t_k > R_{ij})p_k e^{X_{ij}^T\beta+Z_{ij}^T b_i}\xi_{ij} + (1-\delta_{ij}) * I(t_k > L_{ij})p_k e^{X_{ij}^T\beta+Z_{ij}^T b_i}\xi_{ij}\,. \tag{3.10}$$

By combining the steps above, the EM algorithm can be summaried as follows.

- Step 1: Choose initial values for β, p_k and Σ,
- Step 2: Calculate $\hat{E}[\xi_{ij}e^{X_{ij}^T\beta+Z_{ij}^T b_i}]$, $\hat{E}[W_{ijk}]$ and $\hat{E}[b_i^T b_i]$ with the Gaussian quadrature method.

- Step 3: Update β, p_k and Σ by using the closed-form estimator given above for the p_k's and one-step Newton-Raphson method for β and Σ, respectively.

- Step 4: Repeat Steps 2–3 until the convergence is achieved.

Let θ denote all parameters in β and Σ and $\hat{\theta}$ the estimator of θ given by the approach proposed above. Then one can expect that asymptotically, $\hat{\theta}$ is unbiased and its distribution can be approximated by the normal distribution. For estimation of the covariance matrix of $\hat{\theta}$, by following Zeng *et al.* (2017), we suggest to use the profile likelihood approach. More specifically, define $\hat{\mathcal{F}}_\theta = \arg\max_{\mathcal{F}} \log L(\theta, \mathcal{F})$, which could be determined by the EM algorithm above with only updating $\mathcal{F}$ in the M-step. Also define

$$\hat{V}_n = n^{-1} \sum_{i=1}^{n} \left[\{ \frac{\partial}{\partial \theta} l_i(\theta, \hat{\mathcal{F}}_\theta)|_{\theta=\hat{\theta}} \}^{\otimes 2} \right],$$

an estimator of the information matrix of θ, where $l_i(\theta, \hat{\mathcal{F}}_\theta)$ denotes the part of the log-likelihood function $\log L(\theta, \mathcal{F})$ corresponding to the ith cluster. Here $\partial l_i(\theta, \hat{\mathcal{F}}_\theta)/\partial\theta$ could be estimated by the first-order numerical difference with a perturbation constant h_n, which could be set to be a constant of order $n^{-1/2}$ by following the suggestion of Zeng *et al.* (2017). Then one can estimate the covariance matrix of $\hat{\theta}$ by $(n\hat{V}_n)^{-1}$.

3.4 A Simulation Study

In this section, we present some results obtained from a simulation study conducted to assess the finite sample performance of the proposed estimation procedure with the focus on the estimation of regression parameters. In the study, we considered the situation with two covariates X_{1ij} and X_{2ij} generated independently from the uniform

distribution over $U(0,1)$ and the Bernoulli distribution with the probability of success 0.5, respectively. For the cluster size n_i, two set-ups were used with the first one being to generate the n_i's from the set of $\{1,2,3\}$ with the probability $\{0.1, 0.7, 0.2\}$ and the second being to generate them from the uniform distribution over the set of $\{1,..,k\}$.

To generate the true failure times, we first generated the latent variables b_i's from the normal distribution with mean zero and variance $\sigma^2 = 0.5$. Given the X_{ij} and b_i's, the failure times T_i's were assumed to follow model (1) with the logarithmic transformation $G(x) = \log(1+rx)/r$, $F(t) = 1-\exp(-t)$, and setting the Z_{ij} to be 1. For the generation of interval-censored observations, it was supposed that each subject potentially had 10 observation times with the first being generated from the uniform distribution over $(0, 0.3)$ and the gap time between any two successive observation times being generated from the same distribution plus 0.1. For all subjects, the length of study was assumed to be 3. The results given below are based on the number of clusters $N = 200$ or 400 with 1000 replications.

Table 3.1 presents the results obtained on estimation of the regression parameter β as well as σ^2 with the true value $\beta = (\beta_1, \beta_2)^T = (-0.5, 0.5)^T$, $r = 0$, 1 or 1.5 for the logarithmic transformation, and the cluster size generated under the first set-up. They include the estimated empirical bias (Bias) given by the average of the obtained estimates minus the true value, the sample standard error (SSE) of the estimates, the average of the estimated standard errors (ASE) and the 95% empirical coverage probability (CP). One can see from the table that the proposed estimator seems to be unbiased and the estimated standard error appears to be in agreement with the sample standard error. Also, the results on the 95% empirical coverage probabilities indicate that the normal approximation to the distribution of the proposed estimators

seems to be reasonable.

To assess the possible effect of the cluster size on the estimation, we repeated the study above except generating the cluster size by using the second set-up. The obtained results with $k = 3$ or 10, $n = 200$, and the true value $\beta = (\beta_1, \beta_2)^T = (0.5, -0.5)^T$ are presented in Table 3.2 and they seem to give the same conclusions as with Table 3.1. In other words, the proposed estimators seem to be robust with respect to cluster sizes. We also considered other set-ups, including different distributions for the latent variables b_i's, and obtained similar results.

3.5 An Application

Now we apply the maximum likelihood estimation approach proposed in the previous sections to the study conducted by the National Aeronautics and Space Administration (NASA) on the hypobaric decompression sickness data mentioned above (Conkin *et al.*, 1992; Lam and Wong, 2014). In hypobaric environments, a high grade of venous gas embolism (VGE), an abnormal collection of air, may form a bubble in venous blood and cause serious decompression sickness. To assess and measure the decompression sickness, the time to the onset of the grade IV VGE, a high level of venous gas embolism, is often used and was recorded for volunteers in the study. However, the exact onset time is not available and instead, only interval-censored observations were obtained. Furthermore, since each individual could take part in the experiment more than once, we face clustered interval-censored data on the time to the onset of the grade IV VGE.

More specifically, the observed data set consists of 548 records or observations

from 238 volunteers, and among the observations, 124 are interval-censored and the remaining 424 are right-censored. In other words, the right-censored rate is about 77%, indicating that there may exist a cured subgroup. As mentioned above, this can be further seen from Figure 3.1, which gives two Kaplain-Meier (KM) estimates, obtained by assuming all observations being independent and setting T_{ij} to be equal to L_{ij} or R_{ij}, respectively, if $R_{ij} < \infty$ or treating T_{ij} to be right-censored if $R_{ij} = \infty$, of the survival function of the time to the onset of the grade IV VGE. For each subject, there exist four covariates and they are the age ranging from 20 to 54, the gender with 0 denoting females and 1 males, TR360 measuring the decompression stress, and NOADYN with 1 indicating if the individual was ambulatory.

Table 3.3 presents the analysis results given by the application of the estimation approach proposed in the previous sections. Here for the transformation function G and the covariate Z_{ij}, as in the simulation study, we considered the logarithmic transformation with different values for r and set $Z_{ij} = 1$. Since they are quite similar, we only provide the results obtained with $r = 0$ or 1, corresponding to the proportional hazards model and the proportional odds model, respectively. For the analysis, we considered two types of models or two set-ups. One is to include only the four covariates, and the other is first to include both four covariates and their interactions and then perform a backward variable selection based on Akaike information criterion. The latter results in the inclusion of the two interactions of NOADYN with the gender and TR360, respectively. In particular, the age did not seem to have interaction with other covariates.

Without considering any interaction, the analysis indicates that except gender, all of three other covariates had significant effects on the onset of grade IV VGE.

More specifically, it seems that the subjects with younger age, more decompression stress or being in ambulatory tend to have significantly higher risks for the onset of grade IV VGE. By including the interaction, the analysis suggests that one only needs to consider the interaction of NOADYN with gender or TR360 and both seems to be significant. In particular, they indicate that among the individuals who were ambulatory, males had a higher risk for the onset of the grade IV VGE than females, and otherwise, females seem to have a higher risk than males. Also, it is interesting to note that the analysis suggests that TR360 seems to have significant effects on the onset of the grade IV VGE for ambulatory subjects but have no effects otherwise. Lam and Wong (2014) analyzed the same data set under a mixture cure model and suggested that only NOADYN and the interaction between NOADYN and TR360 had signifiant effects on the onset time.

3.6 Discussion and Concluding Remarks

In this chapter, we discussed semiparametric regression analysis of clustered interval-censored failure time data in the presence of a cured fraction or subgroup, and for the problem, a class of semiparametric transformation non-mixture cure models was presented. For estimation, the maximum likelihood estimation procedure was derived and in particular, an EM algorithm that employed the Poisson variable-based data augmentation was developed for the implementation of the approach. To assess the finite sample performance of the proposed method, a simulation study was performed and suggested that the approach seems to work well for practical situations. Also, the method was applied to the hypobaric decompression sickness data that motivated

this study.

Note that for the variance estimation, the proposed method depends on the selection of perturbation constant h_n. In the simulation study, we followed the suggestion from Zeng *et al.* (2017) and tried several different values of h_n of order $n^{-1/2}$, and the results are similar. Also in EM algorithm, we used "mvQuad" Package for the calculation of Gaussian quadrature and chose the number of nodes to be 5. Based on our simulation results by using 10 and 20 as the number of nodes with $n = 200$ in the first set-up, the choice of the number of nodes did not significantly affect the result.

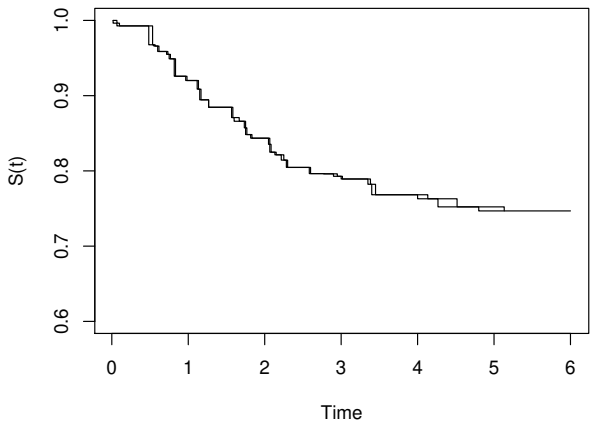

Figure 3.1: KM estimates of the survival function for the time to the onset of the grade IV VGE

Table 3.1: Simulation results on estimation of β and σ^2

r	n		β_1	β_2	σ^2	n		β_1	β_2	σ^2
0	200	True	-0.5	0.5	0.5	400	True	-0.5	0.5	0.5
		Bias	0.0030	0.0091	0.0182		Bias	0.0048	0.0009	0.0503
		SSE	0.1593	0.2780	0.2411		SSE	0.1331	0.2306	0.1783
		ASE	0.1601	0.2747	0.2479		ASE	0.1310	0.2251	0.1676
		CP	0.956	0.949	0.930		CP	0.949	0.950	0.955
1	200	True	-0.5	0.5	0.5	400	True	-0.5	0.5	0.5
		Bias	0.0188	0.0320	0.0407		Bias	0.0054	0.0085	0.0221
		SSE	0.2234	0.3849	0.3247		SSE	0.1582	0.2558	0.2254
		ASE	0.2244	0.3878	0.3571		ASE	0.1519	0.2629	0.2265
		CP	0.944	0.954	0.960		CP	0.946	0.954	0.942
1.5	200	True	-0.5	0.5	0.5	400	True	-0.5	0.5	0.5
		Bias	0.0427	0.0370	0.0124		Bias	0.0414	0.0318	0.0047
		SSE	0.2143	0.3721	0.3151		SSE	0.1997	0.3258	0.2322
		ASE	0.2177	0.3763	0.3312		ASE	0.1897	0.3278	0.2465
		CP	0.957	0.951	0.940		CP	0.933	0.958	0.958

Table 3.2: Simulation results on estimation of β and σ^2 based on n=200 and replication=1000 with different cluster size k.

r	k		β_1	β_2	σ^2	k		β_1	β_2	σ^2
0	3	True	0.5	-0.5	0.5	10	True	0.5	-0.5	0.5
		Bias	0.0076	-0.0034	0.0408		Bias	0.0001	0.0027	0.0106
		SSE	0.1921	0.3275	0.2150		SSE	0.0901	0.1563	0.1053
		ASE	0.1987	0.3456	0.2272		ASE	0.0876	0.1534	0.1333
		CP	0.949	0.951	0.959		CP	0.946	0.942	0.976
1	3	True	0.5	-0.5	0.5	10	True	0.5	-0.5	0.5
		Bias	0.0025	0.0036	0.0431		Bias	0.0175	-0.0185	0.0057
		SSE	0.2676	0.4883	0.3003		SSE	0.1346	0.2218	0.1325
		ASE	0.2740	0.4784	0.3017		ASE	0.1295	0.2236	0.1404
		CP	0.934	0.941	0.962		CP	0.934	0.946	0.937
1.5	3	True	0.5	-0.5	0.5	10	True	0.5	-0.5	0.5
		Bias	0.0370	0.0367	-0.0112		Bias	0.0429	0.0439	0.0243
		SSE	0.2899	0.5051	0.2929		SSE	0.1353	0.2166	0.1235
		ASE	0.2872	0.5001	0.3265		ASE	0.1356	0.2344	0.1471
		CP	0.945	0.937	0.970		CP	0.936	0.958	0.942

Table 3.3: Estimated covariate effects for the NASA study.

Set-up	Model	Covariate	Estimated effects	SD	p-value
1	Proportional Hazards	Age	-0.0699	0.0080	<0.0001
		Gender	-0.0724	0.3403	0.8314
		TR360	0.9705	0.1880	<0.0001
		NOADYN	0.8077	0.3373	0.0166
	Proportional Odds	Age	-0.0651	0.0087	<0.0001
		Gender	0.0684	0.3639	0.8507
		TR360	1.0584	0.2063	<0.001
		NOADYN	0.8730	0.3743	0.0196
2	Proportional Hazards	Age	-0.0221	0.0070	0.0018
		Gender	-1.6785	0.6111	0.0060
		TR360	-0.4460	0.3129	0.1540
		NOADYN	-1.4727	0.4273	0.0005
		Gender*NOADYN	2.4738	0.7543	0.0010
		TR360*NOADYN	1.7618	0.4129	<0.0001
	Proportional Odds	Age	-0.0188	0.0096	0.0500
		Gender	-1.8553	0.7111	0.0009
		TR360	-0.4937	0.3878	0.2030
		NOADYN	-1.5308	0.4998	0.0022
		Gender*NOADYN	2.7473	0.8840	0.0018
		TR360*NOADYN	2.0891	0.4930	<0.0001

Chapter 4

Regression Analysis of Bivariate Interval-censored Failure Time Data under the Semiparametric Transformation Non-mixture Cure Model

4.1 Introduction

In this chapter, we consider a semiparametric transformation non-mixture cure model for bivariate interval-censored failure time data in the presence of a cured subgroup. In the following, we propose a sieve maximum likelihood estimation procedure and conduct a series of numeric studies.

Bivariate interval-censored failure time data commonly occur in clinical trials and biomedical studies. There exist several approaches for regression analysis of bivari-

ate interval-censored failure time data. Goggins and Finkelstein (2000), Chen *et al.* (2007), Tong *et al.* (2008), Chen *et al.* (2013) fitted various marginal models for general multivariate interval-censored data based on the working independence assumption. Wang *et al.* (2008) and Sun and Ding (2019) considered different copula-based models for bivariate current status data and interval-censored data, respectively. Also, the frailty model has been widely studied by many authors (Chen *et al.*, 2009; Wang *et al.*, 2015; Wen and Chen, 2011, 2013; Zhou *et al.*, 2017; Zeng *et al.*, 2017).

The existence of cured subgroup in a study population has been discussed by many authors under different set-ups. To address the existence of cured subgroup, one type of commonly used methods is the two component mixture model approach, which models the cure rate and the failure risk separately (Kuk and Chen, 1992; Ma, 2010; Zhou *et al.*, 2016; Niu *et al.*, 2018; Peng and Taylor, 2011). Another popular approach is based on the non-mixture cure model, which builds a uniform model for covariate effects and therefore, has the typical structure of traditional survival model like proportional hazards model or proportional odds model(Chen *et al.*, 1999; Hu and Xiang, 2013; Li *et al.*, 2019; Yin, 2008; Castro *et al.*, 2014). In terms of multivariate or bivariate interval-censored data with a cured subgroup, several authors discussed the estimation under the mixture cure model (Kim, 2017; Yu and Peng, 2008; Lam *et al.*, 2010; Lam and Wong, 2014; Li and Ma, 2010). For non-mixture cure frailty model, Thompson and Chhikara (2003) assumed the parametric models for both survival function and cumulative distribution function. In the following, we will propose a sieve maximum likelihood estimation approach under a class of semiparametric transformation non-mixture cure models by using a frailty variable to model the correlation within same subjects.

The remainder of the chapter is organized as follows. In Section 4.2, we will introduce some notation, assumptions and models used throughout the chapter and then discuss the resulting likelihood functions. Details of the proposed sieve maximum likelihood estimation approach will be described in Section 4.3. Specifically, Bernstein polynomials are employed to approximate the unknown baseline cumulative distribution functions and it can be relatively easy to implement. Section 4.4 provides some numerical results from a simulation study conducted to assess the finite sample properties of the proposed method. Section 4.5 applies the approach to a real data example and Section 4.6 includes some discussion and concluding remarks.

4.2 Notation and Assumptions

We now consider a bivariate interval-censored data framework. Suppose that there are two related failure times of interest denoted as T_1 and T_2 among n independent subjects. Let T_{ij}, $j = 1, 2; i = 1, ..., n$ denote the failure time for the jth event in the ith subject. Let Z_i denote the p-dimensional covariates which may affect two failure times for ith subject. We assume that there exists a latent variable η with mean 1 and unknown variance $\gamma > 0$, and given Z_i and η_i, the conditional cumulative hazard function of T_{ij} has the form

$$\Lambda(t_{ij}|Z_i, \eta_i) = \eta_i \times G_j[F_j(t)e^{Z_i^T\beta}]. \tag{4.1}$$

In the above, $G_j(.)$ is a prespecified strictly increasing transformation function, and $F_j(.)$ is an unknown cumulative distribution function, so that we allow a cure proportion which is $exp\{-\eta_i \times G[e^{Z_i^T\beta}]\}$. Furthermore, we assume that given Z_i and η_i,

T_{i1} and T_{i2} are independent.

Note that model (4.1) provides a class of flexible models including many commonly used models. For example, by letting $G(x) = x$ and $G(x) = log(1 + x)$, model (4.1) becomes the proportional hazards model and proportional odds model. By using different transformation function G_j, this model allows T_1 and T_2 to follow different survival models. This model has been discussed by Castro *et al.* (2014), which assumed that the covariate effects are different among different failure events of interest. Here, for simplicity, we assume that the covariate effects on two failure times are same in model (4.1). Zhou *et al.* (2017) also used a similar model without assuming a cure subgroup, which replaced $F_j(t)$ with $\Lambda_{0j}(t)$, where $\Lambda_{0j}(t)$ is a baseline cumulative hazard function. One should also notice that model (4.1) is similar but not same to the model used in Yin (2008), which takes the form $\Lambda(t_{ij}|Z_i, w_i) = G[w_i \times F(t)e^{Z_i^T\beta}]$, where w_i is a frailty variable. It's not hard to see that the frailty we used here is outside of the transformation function, which gives a closed form likelihood function after integrating η_i when η_i follows a gamma distribution.

Next we suppose that the failure time T_{ij} cannot be observed exactly, and instead we only know that it fall in the interval $(L_{ij}, R_{ij}]$, where $L_{ij} < R_{ij}$. If $R_{ij} = \infty$, the subject is either cured or experience the failure after last examination time. Define a censoring indicator $\delta_{ij} = \mathbf{I}(R_{ij} < \infty)$, where $\mathbf{I}$ denotes the indicator function. In the following, we assume that the joint distribution of the L_{ij}, R_{ij} and Z_i does not involve unknown parameters in model (4.1) and define $\theta = (\beta, \gamma, F_1, F_2)$. Then the

observed likelihood function could be written as

$$L(\theta) = \prod_{i=1}^{n} \Big\{ \int_{-\infty}^{\infty} \prod_{j=1}^{2} \big[\exp\{-\eta * G[F(L_{ij})e^{Z_i^T\beta}]\} - \delta_{ij} * \exp\{-\eta * G[F(R_{ij})e^{Z_i^T\beta}]\}\big] \times f_\eta(\eta)\ \eta b_i \Big\}, \tag{4.2}$$

where $f_\eta(\eta)$ is the density function of η. If we assume that η follows a gamma distribution as one usually does, we obtain that

$$L(\theta) = \prod_{i=1}^{n} \Big\{ S(L_{i1}, L_{i2}|Z_i) - \delta_{i2}S(L_{i1}, R_{i2}|Z_i) - \delta_{i1}S(R_{i1}, L_{i2}|Z_i) + \delta_{i1} * \delta_{i2}S(R_{i1}, R_{i2}|Z_i) \Big\}, \tag{4.3}$$

where $S(t_1, t_2|Z_i) = [1 + \gamma G_1\{F_1(t_1)e^{Z_i^T\beta}\} + \gamma G_2\{F_2(t_2)e^{Z_i^T\beta}\}]^{-1/\gamma}$. In next section, we will propose a sieve maximum likelihood estimation procedure focusing on the situation that η follows the gamma distribution.

4.3 Sieve Maximum Likelihood Estimation Procedure

In this section, we consider a sieve maximum likelihood estimation procedure to estimate the unknown parameters $\theta \in \Theta$

$$\Theta = \{\theta = (\beta, \gamma, F_1, F_2) \in \mathcal{B} \otimes \mathcal{M}^1 \otimes \mathcal{M}^2\}.$$

Here $\mathcal{B} = \{(\beta, \gamma) \in R^p \times R^+, ||\beta|| + ||\gamma|| \leq M\}$ with p denoting the dimension of β and M denoting a positive constant, and $\mathcal{M}^j$ is the collection of all distribution functions over the interval $[c_j, u_j]$, where $0 \leq c_j < u_j < \infty$, $j = 1, 2$. In practice, $[c_j, u_j]$ is usually taken as the range of L_{ij}'s and finite R_{ij}'s. Then it's natural to obtain the MLE for θ by maximizing $log(L(\theta))$. Following Huang and Rossini (1997b), we consider the sieve maximum likelihood estimation method and define the sieve space as

$$\Theta_n = \{\theta_n = (\beta, \gamma, F_{1n}, F_{2n}) \in \mathcal{B} \otimes \mathcal{M}_n^1 \otimes \mathcal{M}_n^2\}.$$

In the above,

$$M_n^j = \{F_{nj}(t) = \sum_{k=0}^{m} \phi_{jk} B_k(t, m, c_j, u_j) : \sum_{0 \leq k \leq m} |\phi_{jk}| \leq M_n, 0 \leq \phi_{j0} \leq \phi_{j1} \leq ... \leq \phi_{jm} = 1\}.$$

with $B_k(t, m, c_j, u_j)$ being Bernstein basis polynomial as

$$B_k(t, m, c_j, u_j) = \binom{m}{k} (\frac{t - c_j}{u_j - c_j})^k \times (1 - \frac{t - c_j}{u_j - c_j})^{m-k} \quad , \quad k = 0, \ldots, m,$$

where $m = o(n^v)$ for some $v \in (0, 1), j = 1, 2$. Then the sieve maximum likelihood estimator can be defined as $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n, \hat{F}_{1n}, \hat{F}_{2n})$, the value of θ that maximizes the log-likelihood function $l_n(\theta) = log(L_n(\theta))$ over Θ_n.

By using Bernstein polynomials, we transfer the estimation problem with both finite-dimension and infinite-dimension parameters into an estimation problem only involving finite-dimension parameters. Comparing to other smoothing functions like B-splines, Bernstein polynomial has a benefit that we don't need to specify the interior knots, which makes the estimation based on Bernstein polynomial more flexi-

ble. Meanwhile, the Bernstein polynomial has the optimal shape preserving property among all approximation polynomials (Carnicer and Peña, 1993). To estimate θ, we should first notice that there are some restriction for parameters of ϕ due to the property of F_1 and F_2 including nonnegativity, monotonicity and boundedness on $[0, 1]$. The restriction that $0 \leq \phi_{j0} \leq \phi_{j1} \leq ... \leq \phi_{jm}$ could guarantee the nonnegativity and monotonicity of F_1 and F_2 (Chak *et al.*, 2005). Also, by letting $\phi_{jm} = 1$, F_1 and F_2 will go to 1 when $t \to \infty$. To obtain the maximum likelihood estimation, several existing constrained nonlinear optimization methods including the method of moving asymptotes (MMA) and augmented Lagrangian algorithm could be considered (Svanberg, 2002; Birgin and Martínez, 2008). For the numerical study in Section 4.4 and 4.5, we use the *nloptr* function, which is a built-in function in R package "nloptr".

In addition, for the implementation of the estimation approach proposed above, we still need to consider the selection of m and two transformation function G_1 and G_2. For this, we suggest to consider several different values of m and various transformation functions, and choose the combination of (m, F_1, F_2) that minimizes the Akaike information criterion (AIC),

$$\text{AIC} = -2l_n\left(\hat{\theta}_n\right) + 2(p + 1 + 2(m + 1)).$$

In summary, the estimation procedure could be organized as follows:

1. Choose m, G_1 and G_2 and obtain the maximum likelihood estimator $\hat{\theta}$ by maximizing the likelihood function.

2. Calculate AIC based on $\hat{\theta}$ obtained in step 1.

3. Repeat step 1-2 for all combination of m, G_1 and G_2. Choose the combination

that gives the smallest AIC and record the corresponding estimator.

For the estimation of covariance matrix of parameter $\theta_1 = (\beta, \gamma)$, following Zeng *et al.* (2017), we suggest to use the profile likelihood approach. More specifically, define $\hat{\mathcal{F}}_{\theta_1} = \arg\max_{\mathcal{F}} \log L(\theta_1, \mathcal{F})$, which could be determined by keeping θ_1 fixed and estimate F_1 and F_2 via maximizing the log-likelihood function. Also define

$$\hat{V}_n = n^{-1} \sum_{i=1}^{n} \left[\{ \frac{\partial}{\partial \theta_1} l_i(\theta_1, \hat{\mathcal{F}}_{\theta_1})|_{\theta_1 = \hat{\theta}_1} \}^{\otimes 2} \right],$$

an estimator of the information matrix of θ_1, where $l_i(\theta_1, \hat{\mathcal{F}}_{\theta_1})$ denotes the part of the log-likelihood function $\log L(\theta_1, \mathcal{F})$ corresponding to the ith cluster. Here $\partial l_i(\theta_1, \hat{\mathcal{F}}_{\theta_1})/\partial \theta_1$ could be estimated by the first-order numerical difference with a perturbation constant h_n, which could be set to be a constant of order $n^{-1/2}$ by following the suggestion of Zeng *et al.* (2017). Then one can estimate the covariance matrix of $\hat{\theta}_1$ by $(n\hat{V}_n)^{-1}$.

4.4 A Simulation Study

In this section, we conduct a simulation study to evaluate the finite sample performance of the proposed estimation procedure. We focus on the estimation of regression parameters under different transformation functions. In the study, we only considered a one-dimensional covariate Z_i, which is generated independently from Bernoulli distribution with the probability of success being 0.5. To generate the true failure times, we first generated the latent variables η_i's from a gamma distribution with mean 1 and variance $\gamma = 0.5$. Given the Z_i and η_i's, the failure times T_i's were

assumed to follow model (1) with $F_1(t) = F_2(t) = 1 - \exp(-t)$ and $\beta = 0.5$ or -0.5. For transformation function, we use a class of logarithmic transformation functions with form

$$G(x) = log(1 + rx)/r, r \geq 0, \tag{4.4}$$

which includes the proportional hazards model and proportional odds model with $r = 0$ and $r = 1$, respectively. Specifically, we consider 4 different combinations of transformation functions G_1 and G_2: (1) $G_1(x) = G_2(x) = x$; (2)$G_1(x) = G_2(x) = log(1 + x)$; (3) $G_1(x) = x; G_2(x) = log(1 + x)$; (4)$G_1(x) = log(1 + r_1 * x)/r_1$ $G_2(x) = log(1 + r_2 * x)/r_2$, where $r_1 = 0.5$, $r_2 = 1.5$. In terms of choice of m, we used $m = \lceil n^{1/4} \rceil = 4$, the smallest integer larger than $n^{1/4}$. And we set the end of study to be 3 and used $[0, 3]$ for $[c_j, u_j], j = 1, 2$ in Bernstein polynomial.

For the generation of interval-censored observations, it was supposed that each subject will be examined at 10 equally spaced time points over $(0.01, 2.99)$. At each time point, a subject was examined for the occurrence of the failure events with probability 0.5. Then for each subject, L_{ij} and R_{ij} were defined as the last exact observation time before T_{ij} and the first exact observation time after T_{ij}, $j = 1, 2$ and $i = 1, \ldots, n$. The results showed below are based on $n = 200$ with 500 replications.

The results in Table 4.1 include the estimated empirical bias (Bias) given by the average of the obtained estimates minus the true value, the sample standard error (SSE) of the estimates, the average of the estimated standard errors (ASE) and the 95% empirical coverage probability (CP). One can see that the proposed estimator seems to be unbiased and the estimated standard error appears to be in agreement with the sample standard error. Also, the results on the 95% empirical coverage probabilities indicate that the normal approximation to the distribution of

the proposed estimators seems to be reasonable.

4.5 An Application

Now we apply the proposed sieve maximum likelihood estimation procedure to a real data set from the study of AIDS Clinical Trial Group 181 (ACTG181), which concerns the time to the shedding of opportunistic infection cytomegalovirus in the urine and blood of an HIV-infected individual (Goggins and Finkelstein, 2000). The patients in the study were examined for the presence of CMV shedding at their discrete clinic visits, so that the times to CMV shedding were interval-censored. In particular, some patients had already had the positive results for a blood test or urine test, which leading to a left-censored data. Some patients had not yet started shedding by the end of the study and yielded right-censored observations. Among the total 204 patients, 174 patients have right-censored observations for CMV shedding in blood, and 88 individuals have right-censored records for both events. Therefore, there may exist a subgroup of patients who are not susceptible to the CMV shedding. This can be further seen from Figure 4.1, which provides the Kaplain-Meier (KM) estimates for CMD shedding in blood and urine, respectively, obtained by setting T_{ij} to be equal to L_{ij} or R_{ij} if $R_{ij} < \infty$ or treating T_{ij} to be right-censored if $R_{ij} = \infty$.

The main purpose for this study is to analyze the effect of baseline CD4 cell counts on the presence of CMD shedding in patient's urine and blood. In particular, the patients were classified into two groups based on the CD4 cell counts. Let T_1 and T_2 denote the CMV shedding time in the blood and urine, respectively. Define $Z = 1$ if the baseline CD4 count was less than 75 cells/mul, and $Z = 0$ otherwise. For

$[c_j, u_j]$, we set the $c_j = 0$ and use the end time of study 24 for u_j. For the analysis, we considered several combination of different m and G_1 and G_2 and calculated the AIC defined in Section 4.3 for each combination. In particular, we chose m from $\{3, 4, 5\}$ following Zhou *et al.* (2017). For the transformation functions G_1 and G_2, we considered a series of logarithmic transformation functions with different values of r in model (4.4). We first try a large window by choosing the equally spaced grid points of r ranging from 0 to 20 with increments of 1, and further set a small window for r_1 ranging from 14 to 16 with increments of 0.1.

Table 4.2 presents the estimation results of two models including the one with smallest AIC and the one considering the proportional odds model for T_1 and the proportional hazards model for T_2 under different m. We can see that the combination of a large value for r_1 around 15 and 0 for r_2 gave the smallest AIC value. This result indicates that the probability of CMD shedding occurrence in the urine is higher than that in the blood for the same individual, while the probability that a patient was not susceptible to CMD shedding in blood, which is $(1 + e^{Z_i^T \beta} * 15)^{-\eta_i/15}$, is higher than that in urine, which is $exp\{-\eta_i \times e^{Z_i^T \beta}\}$. Zhou *et al.* (2017) showed that the different choices of G_1 and G_2 didn't affect the estimation results and AIC values. In our situation, a preference for a larger r_1 and smaller r_2 may be caused by the consideration of cure rate. Since the right censored rate of the time to CMD shedding in blood is higher than that in urine, it's natural to choose the models providing a larger cure rate for T_1 and a smaller cure rate for T_2. In addition, the result that the AIC values with different m and same transformation functions are similar shows a robustness to the choice of the degree of Bernstein polynomials m when it's around $n^{-1/4}$.

For testing of covariate effect, the parameter estimates under the model with smallest AIC is 1.8044, which leading to a p-value much smaller than 0.05. This result indicates that the subjects whose baseline CD4 count were lower than 75 had a significantly higher risk for CMV shedding in the blood and urine than those whose baseline CD4 count were above 75. Thus, the baseline CD4 count seems to have the effect of predicting the occurrence of CMV shedding. This result is in agreement with Kim (2017), where they considered a copula-based mixture cure model and showed that the patients with lower CD4 cells tended to have a higher chance to shed both on urine and blood. If we choose m=3, $G_1 = log(1+x)$ and $G_2(x) = x$, the estimate of β is 1.4189 with estimated standard error equal to 0.2682, yielding a score statistic equal to 5.2904, is similar to those obtained by Zhou *et al.* (2017), which gave the estimates $\hat{\beta} = 1.5039$ yielding a score statistic equal to 4.68.

4.6 Discussion and Concluding Remarks

In this chapter, we discussed regression analysis of bivariate interval-censored failure time data in the presence of a cured fraction or subgroup. We presented a class of semiparametric transformation non-mixture cure models with a frailty variable. To obtain the estimates of regression parameters, a sieve maximum likelihood estimation procedure based on Bernstein Polynomial was derived. A simulation study was performed to evaluate the finite sample performance of proposed method and suggested that the approach worked well for practical situation. Also the method was applied to a real set from a AIDS study.

Table 4.1: Simulation results with one covariate with n=200 and replication=500.

Para.	True	Bias	SSE	ASE	CP
$G_1(x) = x$, $G_2(x) = x$					
β	0.5	0.0194	0.1619	0.1607	0.946
γ	0.5	0.0467	0.1686	0.1673	0.944
β	-0.5	0.0351	0.1413	0.1489	0.954
γ	0.5	0.0587	0.1918	0.1912	0.946
$G_1(x) = log(1+x)$, $G_2(x) = log(1+x)$					
β	0.5	0.0200	0.2214	0.2142	0.934
γ	0.5	0.0479	0.1983	0.1916	0.964
β	-0.5	0.0152	0.1822	0.1926	0.950
γ	0.5	0.0573	0.2469	0.2315	0.940
$G_1(x) = x$, $G_2(x) = log(1+x)$					
β	0.5	0.0194	0.1667	0.1701	0.962
γ	0.5	0.0271	0.1726	0.1724	0.954
β	-0.5	0.0415	0.1547	0.2184	0.950
γ	0.5	0.0395	0.1645	0.2050	0.936
$G_1(x) = \frac{log(1+0.5*x)}{0.5}$, $G_2(x) = \frac{log(1+1.5*x)}{1.5}$					
β	0.5	0.0345	0.2037	0.2044	0.934
γ	0.5	0.0341	0.1925	0.1899	0.944
β	-0.5	0.0339	0.1784	0.1880	0.960
γ	0.5	0.0599	0.2308	0.2282	0.956

Table 4.2: Estimated covariate effects and AIC value.

m=3	Para.	Est	SE	P-value	AIC
$G_1(x) = log(1 + x * 15.2)/15.2$	β	1.8062	0.2921	< 0.0001	879.7528
$G_2(x) = x$	γ	1.3685	0.4329		
$G_1(x) = log(1 + x)$	β	1.4189	0.2682	< 0.0001	954.4004
$G_2(x) = x$	γ	0.9596	0.2458		
m=4	Para.	Est	SE	P-value	AIC
$G_1(x) = log(1 + x * 15)/15$	β	1.8044	0.2931	< 0.0001	877.0185
$G_2(x) = x$	γ	1.2994	0.3262		
$G_1(x) = log(1 + x)$	β	1.3393	0.2409	< 0.0001	949.4478
$G_2(x) = x$	γ	0.6321	0.2209		
m=5	Para.	Est	SE	P-value	AIC
$G_1(x) = log(1 + x * 14.9)/14.9$	β	1.7518	0.2899	< 0.0001	877.465
$G_2(x) = x$	γ	1.2608	0.3854		
$G_1(x) = log(1 + x)$	β	1.2662	0.2464	< 0.0001	947.4579
$G_2(x) = x$	γ	0.5881	0.2458		

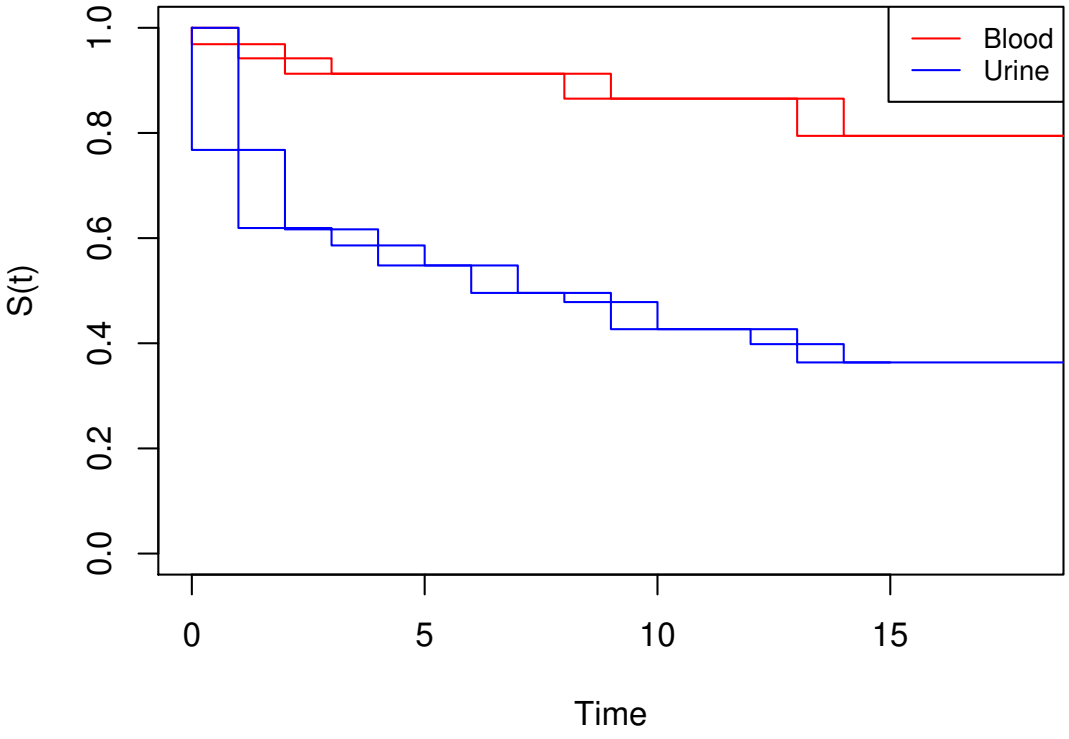

Figure 4.1: KM estimate of the survival function for CMD shedding in blood and urine

Chapter 5

Future Research

For regression analysis of correlated interval-censored failure time data with a cured subgroup, a number of issues remain unsolved and need further investigation. In this chapter, we will briefly discuss and point out several directions for future research.

5.1 Regression Analysis of Clustered Interval-censored Failure Time Data under the Proportional Hazards Mixture Cure Model

For the topic in Chapter 2, one direction for future research is that here we have assumed that it is known which covariates may have the effects on the cure risk or the failure risk and it is apparent that this may not be true. In practice, as in the application discussed above, one could try different set-ups for the selection of X and Z and compare the obtained results. However, it is clear that it would be helpful if a data-driven approach can be developed for this. Another direction is that it would be

useful to provide a rigorous justification for the asymptotic normality of the proposed estimators of the regression parameters as well as other asymptotic properties. Also, for the uncured subjects, instead of the proportional hazards model, one may consider the use of some other models such as the additive hazards model to model the failure time of interest.

5.2 Regression Analysis of Clustered Interval-censored Failure Time Data under the Semiparametric Transformation Non-mixture Cure Model

For the problem in Chapter 3, there exist several directions for future research. One is that the focus in Chapter 3 has been on the situation where both covariates and their effects are time independent. It is apparent that sometimes this may not be true and it would be useful to generalize the proposed method to the situation where either covariates or their effects or both are time-dependent. Another direction is that, for the selection of the parameter r in the logarithmic transformation or the transformation function G in model (3.1), we took the trying and comparison approach. It is clear that it would be helpful if a data-driven approach for this could be developed. Also, it would be useful to provide a rigorous justification for the asymptotic properties of the proposed estimator.

5.3 Regression Analysis of Bivariate Interval-censored Failure Time Data under the Semiparametric Transformation Non-mixture Cure Model

For the topic in Chapter 4, one could extend it in several parts. In the estimation procedure, we only focused on gamma frailty. It's natural to consider a frailty variable following different distribution like log-normal distribution. Also, in model (4.1), we assumed that the covariate effects for both two failure events were same. This assumption may be invalid in some situation, so that one may consider a more general model

$$S(t_{ij}|Z_i, \eta) = exp\{ -\eta \times G_j[F_j(t)e^{Z_i^T \beta_j}]\}.$$

In addition, for the selection of transformation functions G_1 and G_2, we chose them based on AIC criterion. One may put some assumptions on the form of G_j like the logarithmic transformation function and develop some estimation procedures to determine them. Another direction is that our discussions were focused on time-independent covariates and their effects. It will be useful to extend the proposed model and estimation procedure to time-dependent covariates or coefficients.

Bibliography

Birgin, E. and Martínez, J. (2008). Improving ultimate convergence of an augmented lagrangian method. *Optimization Methods and Software*, **23**(2), 177–195.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics*, **59**(3), 570—579.

Carnicer, J. M. and Peña, J. M. (1993). Shape preserving representations and optimality of the bernstein basis. *Advances in Computational Mathematics*, **1**(2), 173–196.

Castro, M. D., Chen, M.-H., Ibrahim, J. G., and Klein, J. P. (2014). Bayesian transformation models for multivariate survival data. *Scandinavian Journal of Statistics*, **41**(1), 187–199.

Chak, P. M., Madras, N., and Smith, B. (2005). Semi-nonparametric estimation with bernstein polynomials. *Economics Letters*, **89**(2), 153–156.

Chang, I.-S., Wen, C.-C., and Wu, Y.-J. (2007). A profile likelihood theory for the correlated gamma-frailty model with current status family data. *Statistica Sinica*, **17**, 1023–1046.

Chen, C.-M. and Lu, T.-F. C. (2012). Marginal analysis of multivariate failure time data with a surviving fraction based on semiparametric transformation cure models. *Computational Statistics and Data Analysis*, **56**(3), 645–655.

Chen, H.-Y. and Little, R. (2001). A profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates. *Lifetime data analysis*, **7**, 207–24.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, **89**(3), 659–668.

Chen, L., Sun, J., and Xiong, C. (2016). A multiple imputation approach to the analysis of clustered interval-censored failure time data with the additive hazards model. *Computational Statistics and Data Analysis*, **103**, 242–249.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**(447), 909–919.

Chen, M.-H., Tong, X., and Sun, J. (2007). The proportional odds model for multivariate interval-censored failure time data. *Statistics in Medicine*, **26**(28), 5147–5161.

Chen, M.-H., Tong, X., and Sun, J. (2009). A frailty model approach for regression analysis of multivariate current status data. *Statistics in Medicine*, **28**(27), 3424–3436.

Chen, M.-H., Tong, X., and Zhu, L. (2013). A linear transformation model for multivariate interval-censored failure time data. *Canadian Journal of Statistics*, **41**(2), 275–290.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**(1), 141–151.

Clegg, L. X., Cai, J., and Sen, P. K. (1999). A marginal mixed baseline hazards model for multivariate failure time data. *Biometrics*, **55**(3), 805–812.

Cong, X. J., Yin, G., and Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics*, **63**(3), 663–672.

Conkin, J., Bedahl, S., and Van Liew, H. (1992). A computerized databank of decompression sickness incidence in altitude chambers. *Aviation, space, and environmental medicine*, **63**, 819–24.

Cook, R. J. and Tolusso, D. (2009). Second-order estimating equations for the analysis of clustered current status data. *Biostatistics (Oxford, England)*, **10**(4), 756—772.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2), 187–202.

Diao, G. and Yin, G. (2012). A general transformation class of semiparametric cure rate frailty models. *Annals of the Institute of Statistical Mathematics*, **64**(5), 959–989.

Diao, G. and Yuan, A. (2019). A class of semiparametric cure models with current status data. *Lifetime Data Analysis*, **25**.

Fang, H.-B., Sun, J., and Lee, M.-L. T. (2002). Nonparametric survival comparisons for interval-censored continuous data. *Statistica Sinica*, **12**(4), 1073–1083.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**(4), 1041–1046.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, **42**(4), 845.

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate archimedean copulas. *Journal of the American Statistical Association*, **88**(423), 1034–1043.

Glidden, D. V. and Self, S. G. (1999). Semiparametric likelihood estimation in the clayton-oakes failure time model. *Scandinavian Journal of Statistics*, **26**(3), 363–372.

Goggins, W. B. and Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics*, **56**(3), 940–943.

Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A., and Zaslavsky, A. M. (1998). A markov chain monte carlo em algorithm for analyzing interval-censored data under the cox proportional hazards model. *Biometrics*, **54**(4), 1498–1507.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. Birkhauser Verlag.

Gu, M., Sun, L., and Zuo, G. (2006). A baseline-free procedure for transformation models under interval censorship. *Lifetime data analysis*, **11**, 473–88.

Hoffman, E. B., Sen, P. K., and Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika*, **88**(4), 1121–1134.

Hougaard, P. (1986). A class of multivanate failure time distributions. *Biometrika*, **73**(3), 671–678.

Hougaard, P. (2012). *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. Springer New York.

Hu, T. and Xiang, L. (2013). Efficient estimation for semiparametric cure models with interval-censored data. *Journal of Multivariate Analysis*, **121**, 139 – 151.

Huang, J. (1995). Maximum likelihood estimation for proportional odds regression model with current status data. *Lecture Notes-Monograph Series*, **27**, 129–145.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, **24**(2), 540 – 568.

Huang, J. and Rossini, A. J. (1997a). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*, **92**(439), 960–967.

Huang, J. and Rossini, A. J. (1997b). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*, **92**(439), 960–967.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 123–169, New York, NY. Springer US.

Jianwen, C. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, **82**(1), 151–164.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2nd edition.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457–481.

Kim, Y.-J. (2017). Cure rate model with bivariate interval censored data. *Communications in Statistics - Simulation and Computation*, **46**(9), 7116–7124.

Kor, C.-T., Cheng, K.-F., and Chen, Y.-H. (2013). A method for analyzing clustered interval-censored data based on cox's model. *Statistics in Medicine*, **32**(5), 822–832.

Kuk, A. Y. C. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**(3), 531–541.

Lam, K. and Wong, K.-Y. (2014). Semiparametric analysis of clustered interval-censored survival data with a cure fraction. *Computational Statistics and Data Analysis*, **79**, 165 – 174.

Lam, K. F. and Xue, H. (2005). A semiparametric regression cure model with current status data. *Biometrika*, **92**(3), 573–586.

Lam, K. F., Xu, Y., and Cheung, T.-L. (2010). A multiple imputation approach for clustered interval-censored survival data. *Statistics in Medicine*, **29**(6), 680–693.

Lawless, J. (2011). *Statistical Models and Methods for Lifetime Data*. Wiley Series in Probability and Statistics. Wiley.

Li, J. and Ma, S. (2010). Interval-censored data with repeated measurements and

a cured subgroup. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**(4), 693–705.

Li, J., Wang, C., and Sun, J. (2012). Regression analysis of clustered interval-censored failure time data with the additive hazards model. *Journal of Nonparametric Statistics*, **24**(4), 1041–1050.

Li, S., Hu, T., Zhao, X., and Sun, J. (2019). A class of semiparametric transformation cure models for interval-censored failure time data. *Computational Statistics and Data Analysis*, **133**, 153 – 165.

Liang, K., Self, S., Bandeen-Roche, K., and Zeger, S. (1995). Some recent developments for regression analysis of multivariate failure time data. *Lifetime data analysis*, **1**(4), 403—415.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine*, **13**(21), 2233–2247.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, **91**(2), 331–343.

Ma, S. (2009). Cure model with current status data. *Statistica Sinica*, **19**, 233–249.

Ma, S. (2010). Mixed case interval censored data with a cured subgroup. *Statistica Sinica*, **20**(3), 1165–1181.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports*, **50**(3), 163—170.

Niu, Y., Song, L., Liu, Y., and Peng, Y. (2018). Modeling clustered long-term survivors using marginal mixture cure model. *Biometrical Journal*, **60**(4), 780–796.

Pan, W. (2000). A multiple imputation approach to cox regression with interval-censored data. *Biometrics*, **56**(1), 199–203.

Peng, Y. and Taylor, J. (2011). Mixture cure model with random effects for the analysis of a multi-center tonsil cancer study. *Statistics in medicine*, **30**, 211–23.

Peng, Y., Taylor, J., and Yu, B. (2007). A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Analysis*, **13**, 351–369.

Petroni, G. R. and Wolfe, R. A. (1994). A two-sample test for stochastic ordering with interval-censored data. *Biometrics*, **50**(1), 77–87.

Pierce, D., Stewart, W., and Kopecky, K. J. (1979). Distribution-free regression analysis of grouped survival data. *Biometrics*, **35 4**, 785–93.

Rabinowitz, D., Betensky, R., and Tsiatis, A. (2000). Using conditional logistic regression to fit proportional odds models to interval censored data. *Biometrics*, **56**(2), 511–518.

Rossini, A. J. and Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, **91**(434), 713–721.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley.

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, **83**(2), 355–370.

Shen, X. (1998). Propotional odds regression and sieve maximum likelihood estimation. *Biometrika*, **85**(1), 165–177.

Su, C.-L. and Lin, F.-C. (2019). Analysis of clustered failure time data with cure fraction using copula. *Statistics in Medicine*, **38**(21), 3961–3973.

Sun, J. (1999). A nonparametric test for current status data with unequal censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(1), 243–250.

Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. Springer New York.

Sun, J. and Sun, L. (2005). Semiparametric linear transformation models for current status data. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, **33**(1), 85–96.

Sun, T. and Ding, Y. (2019). Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics*.

Svanberg, K. (2002). A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM Journal on Optimization*, **12**(2), 555–573.

Thompson, L. A. and Chhikara, R. (2003). A bayesian cure rate model for repeated measurements and interval censoring. *Proceedings of JSM*, **40**, 41.

Tong, X., Chen, M.-H., and Sun, J. (2008). Regression analysis of multivariate interval-censored failure time data with application to tumorigenicity experiments. *Biometrical Journal*, **50**(3), 364–374.

Tsodikov, a. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*.

Tsodikov, A. D., Ibrahim, J. G., and Yakovlev, A. Y. (2003). Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**(464), 1063–1078.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*.

Wang, L., Sun, J., and Tong, X. (2008). Efficient estimation for the proportional hazards model with bivariate current status data. *Lifetime data analysis*, **14**, 134–53.

Wang, N., Wang, L., and McMahan, C. S. (2015). Regression analysis of bivariate current status data under the gamma-frailty proportional hazards model using the em algorithm. *Computational Statistics and Data Analysis*, **83**, 140–150.

Wellner, J. A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association*, **92**(439), 945–959.

Wen, C.-C. and Chen, Y.-H. (2011). Nonparametric maximum likelihood analysis of clustered current status data with the gamma-frailty cox model. *Computational Statistics and Data Analysis*, **55**(2), 1053–1060.

Wen, C.-C. and Chen, Y.-H. (2013). A frailty model approach for regression analysis of bivariate interval-censored survival data. *Statistica Sinica*, **23**(1), 383–408.

Williamson, J. M., Datta, S., and Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics*.

Williamson, J. M., Kim, H.-Y., Manatunga, A., and Addiss, D. G. (2008). Modeling survival data with informative cluster size. *Statistics in Medicine*, **27**(4), 543–555.

Xiang, L., Ma, X., and Yau, K. K. W. (2011). Mixture cure model with random effects for clustered interval-censored survival data. *Statistics in medicine*, **30 9**, 995–1006.

Yin, G. (2008). Bayesian transformation cure frailty models with multivariate failure time data. *Statistics in Medicine*, **27**(28), 5929–5940.

Yu, B. and Peng, Y. (2008). Mixture cure models for multivariate survival data. *Computational Statistics and Data Analysis*, **52**(3), 1524–1532.

Zeng, D., Mao, L., and Lin, D. Y. (2016). Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, **103**(2), 253–271.

Zeng, D., Gao, F., and Lin, D. Y. (2017). Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika*, **104**(3), 505–525.

Zhang, X. and Sun, J. (2010). Regression analysis of clustered interval-censored failure time data with informative cluster size. *Computational Statistics and Data Analysis*, **54**(7), 1817–1823.

Zhang, Z. and Zhao, Y. (2013). Empirical likelihood for linear transformation models with interval-censored failure time data. *Journal of Multivariate Analysis*, **116**, 398–409.

Zhang, Z., Sun, L., Zhao, X., and Sun, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics*, **33**(1), 61–70.

Zhao, H., Ma, C., Li, J., and Sun, J. (2018). Regression analysis of clustered interval-censored failure time data with linear transformation models in the presence of informative cluster size. *Journal of Nonparametric Statistics*, **30**(3), 703–715.

Zhao, Q. and Sun, J. (2004). Generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine*, **23**(10), 1621–1629.

Zhou, J., Zhang, J., McLain, A. C., and Cai, B. (2016). A multiple imputation approach for semiparametric cure model with interval censored data. *Computational Statistics and Data Analysis*, **99**, 105–114.

Zhou, Q., Hu, T., and Sun, J. (2017). A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association*, **112**(518), 664–672.

VITA

Dian Yang was born in 1990 in Maanshan, a beautiful city located in Anhui Province in China. He received his Bachelor's degree in science from the department of statistics of Shanghai University of Finance and Economics in China in 2012. Then he joined the M.A. applied track program in the Department of Statistics at the University of Missouri - Columbia in August 2013. In May 2015, he obtained his Master degree and did an internship in Purdue University as a research assistant. In Jan 2016, he was accepted into the Ph.D. program in the Department of Statistics at the University of Missouri - Columbia. He will start a position of Senior Biostatistician in Harbour BioMed, Shanghai on June 2021.