MODEL EVALUATION AND VARIABLE SELECTION FOR INTERVAL-CENSORED DATA

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

TYLER COOK

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

MAY 2015

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

MODEL EVALUATION AND VARIABLE SELECTION
FOR INTERVAL-CENSORED DATA

presented by Tyler Cook,
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. Hongyuan Cao

Dr. Subharup Guha

Dr. Xianyang Zhang

Dr. X.H. Wang

ACKNOWLEDGMENTS

First and foremost I would like to express my sincere gratitude to my advisor Dr. (Tony) Jianguo Sun. His patience and guidance were invaluable throughout the entire course of my graduate studies and subsequent research. This work certainly would not have been possible without his inspiration and expertise.

I would also like to thank the members of my advisory committee: Dr. Hongyuan Cao, Dr. Subharup Guha, Dr. Xianyang Zhang, and Dr. X.H. Wang. I am truly grateful for the time and effort they have taken in order to assist with the completion of my dissertation.

In addition, I appreciate the continued support of the entire statistics department at the University of Missouri. I am indebted to all the faculty members that have skillfully lead my education. I am forever grateful for the guidance Dr. Lawrence Ries has provided in making me an effective instructor. I also owe thanks to Judy, Tracy, and Kathleen for their help over the years.

Finally, I would like to acknowledge the support of all of my family and loved ones who have kept me going throughout my entire academic career.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

ABSTRACT

Survival analysis is a popular area of statistics dealing with time-to-event data. This type of data can be seen in many disciplines, but it is perhaps most commonly encountered in medical studies. Doctors, for example, might be testing different treatments developed to prolong the lifetimes of cancer patients. Unfortunately, in practical problems such as clinical trials, there is often incomplete data thanks to patients dropping out of the study. This results in censoring, which is a special characteristic of survival data. There are many different types of censoring. This dissertation focuses on the analysis of interval-censored data, where the failure time is only known to belong to some interval of observation times.

One problem that researchers face when analyzing survival data is how to handle the censoring distribution. It is often assumed that the observation process generating the censoring is independent of the event time of interest. Consequently, the observation process can effectively be ignored. However, this assumption is clearly not always realistic. Unfortunately, one cannot generally test for independent censoring without additional assumptions or information. Therefore, the researcher is faced with a choice between using methods designed for informative or noninformative censoring. Chapters 2 and 3 of this dissertation investigate the effectiveness of different methods developed for the analysis of informative case I and case II interval censored data under both types of censoring. Extensive simulation studies indicate that the methods produce unbiased results in the presence of both informative and noninformative censoring. The efficiency of the informative censoring methods is then compared with approaches created to handle noninformative censoring. The results

of these simulation studies can provide guidelines for deciding between models when facing a practical problem where one is unsure about the dependence of the censoring distribution.

Another important problem seen in survival analysis is determining the set of predictors that are significantly related with the failure time being studied. Variable selection has received substantial attention both in classical linear models as well as survival analysis. This is largely thanks to recent technological advances making it easier for researchers in biology to collect huge amounts of genetic data. For example, a researcher with access to gene expression levels for hundreds of genes is interested in identifying which of those genes can predict tumor development time in cancer patients. One must sift through the large number of genes in order to find the small set of significant genes that influence tumor growth. Several methods using penalized likelihood procedures have been proposed to perform parameter estimation and variable selection simultaneously. A number of these techniques have also been extended to the case of right-censored survival data, but little has been done in the context of interval-censoring. In chapter 4, we propose an imputation approach for variable selection of interval-censored data that utilizes these penalized likelihood procedures. This method uses imputation to create a new dataset of imputed exact failure times and right-censored observations. Variable selection can then be performed on the imputed dataset using any of the popular variable selection techniques created for right-censored data. Comprehensive simulation studies illustrate the effectiveness of this new approach. Also, this method is attractive due to how easy it is to implement, since it can take advantage of existing software for variable selection of right-censored data.

Chapter 1

Introduction

1.1 Failure Time Data

Survival analysis is an area of statistics concerned with analyzing failure time data. This means the variable of interest is the time until some event occurs, such as the death of a patient or failure of some mechanical component. This type of data can be found in many fields such as medicine, biology, epidemiology, economics, psychology, and engineering.

One of the key defining properties of failure time data is the presence of censoring. The consequence of censoring is that the event time is only partially known. Censoring occurs frequently in medical studies when, for example, patients drop out of the study. Therefore, their event time is only known to happen over some range of time beyond their last observation time. Some common types of censoring will be discussed in more detail in the next section.

1.1.1 Interval Censoring

There are numerous types of censored data. Here, we will focus on interval-censored data. This is a very general type of censoring that includes several commonly seen data structures as special cases. Interval-censored data is distinguished by the fact that failure times are only observed to belong to some window, or interval, of time. That is, if T is the failure time of interest, we only know T to fall between two values $T \in (L, R]$ where $L \leq R$. So L is the last observation time before the event, and R is the first observation time after the event.

The very well-known right-censored data is a special case of interval censoring. A right-censored observation has L known and $R = \infty$. Right-censored data usually contain a mixture of right-censored observations and exact failure times. One can see that exact observations occur if $L = R$. This type of data is extremely common in medical studies, and a wide range of well-studied techniques for this data exist in the literature.

Another special type of interval-censored data is case I, or current status data. Current status data is unique in that each subject has only one observation time. Consequently, the researcher only knows if the event time occurs before or after the observation time. More specifically, if C is the observation time, one knows $T \in (0, C]$ or $T \in (C, \infty]$. In other words, all of the subjects are either left-censored or right-censored. Current status data is commonly recorded as

$$\{C, \delta = I(T \leq C)\},$$

where I is the indicator function. Therefore, δ indicates whether or not failure has

occurred when the subject is observed. This type of data can be seen frequently in demographical studies and tumorigenicity experiments.

Case II interval-censored data is another type of interval-censored data where at least one interval is above 0 and finite, i.e. $L \in (0, \infty)$ and $R \in (0, \infty)$. A common notation used for case II interval-censored data is

$$\{U, V, \delta_1 = I(T \leq U), \delta_2 = I(U < T \leq V), \delta_3 = 1 - \delta_1 - \delta_2\},$$

where U and V are two observation times with $U \leq V$ and the δ's are censoring indicators. Also, note that case I interval-censored data is a special case of case II data with $U = V = C$. This type of censoring regularly occurs in medical studies where patients are scheduled for regular follow-up visits since patients often miss appointments and return after the event has occurred.

1.1.2 Noninformative and Informative Censoring

Another important aspect of survival analysis is handling the possible relationship between the event time and censoring process. In practice, it is often assumed that the censoring time gives no additional information about the event time. This situation is known as independent or noninformative censoring. For example, a patient drops out of a medical study in order to move to a new state because their spouse found a new job. The time they choose to move would not contain any information about the time they develop a tumor. With current status data, noninformative censoring implies that the failure time and censoring time are independent completely or given covariates. Unfortunately this simple assumption cannot be generalized to case II

interval-censored data since there is an explicit relationship between the event time and the end points of the observation interval, i.e. $L < T \leq R$. Alternatively, as defined in Sun (2006), the noninformative censoring assumption is characterized by

$$P(T \leq t | L = l, R = r, L \leq T < R) = P(T \leq t | l \leq T < r) \tag{1.1}$$

or

$$P(L < T \leq R | L = l, R = r) = P(l < T \leq r). \tag{1.2}$$

The main idea is that $(L, R]$ does not any contain information other than that T is enclosed by the points L and R. So the distribution of T does not involve any parameters in the joint distribution of L and R. The independent censoring assumption has the benefit of not working directly with the distribution of the censoring process.

The noninformative censoring assumption is, of course, not always realistic. In fact, Williams & Lagakos (1977) showed that, under right censoring, likelihood estimation for the event time can ignore the censoring process only when a certain constant sum property is met. Betensky (2000) established a similar condition for case I interval-censored data. It is easy to imagine that the censoring process can contain information about the event time. For example, patients developing tumors might be more likely to make scheduled visits to their doctor. When this is the case, the censoring is referred to as informative censoring. This type of censoring requires more complicated methods.

1.1.3 Examples

As previously mentioned, current status data can often be found in tumorigenicity experiments. One example, the ED_{01} experiment, investigated the time until development of lung and bladder tumors in mice (Lindsey & Ryan, 1993). The study, which lasted 33 months, was conducted on 24,000 female mice at the National Center for Toxicological Research. The mice were randomized into one control group and seven treatment groups. The treatments consisted of different dose levels of 2-acetylaminofluorene (AAF), which is a known carcinogen. The mice were inspected for tumors at the times of their natural death or sacrifice. For this example, the observation time is the mouse's death time, and the failure time is the time of tumor onset. Each mouse has only one observation since they are only inspected at death. Therefore, the only knowledge about tumor onset is that it either did or did not happen prior to death. This of course results in case I interval-censored data.

This study is also an example where one might expect to have informative censoring. If the tumors have any degree of lethality, it is then reasonable to expect that mice who develop tumors earlier would also die earlier. This means that the death time and event time are quite possibly related.

A classic example of case II interval-censored data, discussed in Finkelstein (1986), involves breast cancer patients treated at the Joint Center for Radiation Therapy in Boston from 1976 to 1980. The goal of this particular experiment was to compare the time until cosmetic breast deterioration between patients receiving radiation therapy plus adjuvant chemotherapy, and those receiving only radiation therapy. The retrospective study examined 94 breast cancer patients with 46 treated by radiotherapy alone. The response variable, measuring cosmetic deterioration, was the time until

breast retraction. Patients were scheduled for periodic follow-up, but some missed visits and returned later with changed breast retraction status. Therefore, these patients were interval censored with L equal to the last visit time before breast retraction, and R equal to the first visit time after breast retraction occurred.

1.2 Commonly Used Models

The motivation for performing survival analysis usually falls into one of the following categories: estimation of survival function, treatment comparison, or regression analysis. In this section, we will focus on the situation where a researcher is interested in evaluating covariate effects on survival by performing regression analysis. Several popular regression models will be introduced. First, however, it will be helpful to define a few basic functions frequently used in survival analysis.

As before, let T denote a nonnegative random variable representing the time until some event occurs. For brevity, we will concentrate on the case where T is continuous. The survival function, $S(t)$, is defined as

$$S(t) = P(T \geq t), \quad 0 < t < \infty.$$

This is equal to the probability the event time exceeds the value t. Next, the probability density function, $f(t)$, can be defined as

$$f(t) = -\frac{dS(t)}{dt}.$$

A fundamental quantity in survival analysis is known as the hazard function. This

function, denoted $\lambda(t)$, is defined as

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

We can see that the survival, density, and hazard functions are all related since

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{P(T \geq t)} = \frac{f(t)}{S(t)}.$$

The hazard can be interpreted as the instantaneous failure rate for those subjects still at risk at time t. A related function, the cumulative hazard function, is defined as

$$\Lambda(t) = \int_0^t \lambda(s) ds.$$

1.2.1 Proportional Hazards Model

One of the most popular regression models in survival analysis is the proportional hazards model. This model is also commonly referred to as the Cox model since it was first studied by D.R. Cox in 1972. The proportional hazards model assumes the covariates affect the event time through the following relationship

$$\lambda(t; Z) = \lambda_0(t) \exp(Z'\beta) \tag{1.3}$$

where Z is a vector of covariates, $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function, and β is the vector of regression parameters. The proportional hazards model indicates that the covariates have a multiplicative effect on the hazard. Using (1.3) we can see that the ratio of the hazard functions for two subjects with different

covariates is constant. For example, with a two-sample problem letting Z=1 or Z=0 we have

$$\frac{\lambda(t; Z = 1)}{\lambda(t; Z = 0)} = \exp(\beta).$$

The proportional hazards model derives its name from this property. Also, from (1.3) we can write the density and survival functions of T given the covariates as

$$f(t; Z) = \lambda_0(t) \exp(Z'\beta) \exp[-\Lambda_0(t) \exp(Z'\beta)]$$

and

$$S(t; Z) = \exp[-\Lambda_0(t) \exp(Z'\beta)]$$

where

$$\Lambda_0(t) = \int_0^t \lambda_0(s) ds$$

is the baseline cumulative hazard function. One of the benefits of this model, contributing to its popularity and widespread use, is that the estimation procedure for β is relatively straightforward for right-censored data using the partial likelihood method proposed in Cox (1972).

1.2.2 Additive Hazards Model

Another prevalent regression method is the additive hazards model. For this model, one specifies the relationship between the covariates and hazard using

$$\lambda(t; Z) = \lambda_0(t) + Z'\beta,$$

where again $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard, and β is the vector of regression parameters. The additive hazards model specifies that the covariate effect on the hazard is additive rather than multiplicative, like in the proportional hazards model. To illustrate, again consider the two-sample problem with $Z = 1$ or $Z = 0$,

$$\lambda(t; Z = 1) = \lambda_0(t) + \beta$$

and

$$\lambda(t; Z = 0) = \lambda_0(t)$$

which means β is the difference in hazards between the two groups. Therefore, the additive hazards model provides an alternative interpretation to the proportional hazards model, and does not require the assumption of proportional hazards.

1.2.3 Accelerated Failure Time Model

The accelerated failure time model is another commonly used regression model. This model is a log-linear model, and is defined as

$$\log(T) = Z'\beta + W,$$

where W is an unknown error variable. The accelerated failure time model is appealing because it resembles a standard linear regression model. This means interpretation of the model parameters, in terms of $\log(T)$, is easy and clear.

1.3 Analysis of Interval-Censored Data

In this section we will review a number of methods proposed for analyzing interval-censored data. The focus will be on techniques developed for performing semiparametric regression analysis. So the goal is to assess how covariates influence the failure time. This is currently a very popular topic in the literature. An overview for both case I and case II interval-censored data will be presented.

1.3.1 Case I

Early work on the proportional hazards model with current status data was done by Huang (1996). Here, a maximum likelihood approach was developed and important asymptotic properties were established. Chen *et al.* (2009) studied multivariate current status data, and Sun & Shen (2009) proposed a method for competing risks.

Lin *et al.* (1998) laid the groundwork for analyzing case I interval-censored data using the additive hazards model. Their method utilizes a simple estimating equation procedure based on the partial likelihood by reworking the problem into a proportional hazards setting. This technique conveniently avoids estimating the baseline hazard function. Martinussen & Scheike (2002) proposed a solution that can be more efficient than the one given by Lin *et al.* (1998). However, it is potentially more complicated since it involves estimating the baseline hazard function. Chen & Sun (2009) proposed an imputation approach for the additive hazards model. In addition, Ghosh (2003) developed methods for assessing goodness-of-fit under the additive hazards model.

Several other semiparametric models have been investigated by many authors. Huang (1995) and Rossini & Tsiatis (1996) studied the proportional odds model.

Both Shen (2000) and Xue *et al.* (2004) worked on the accelerated failure time model using sieve methods.

The previously introduced methods were developed under the assumption of non-informative censoring. Significantly less work has been done when this assumption is not valid and one has informative censoring. Much of the early work was done in the context of tumorigenicity experiments using transition functions including Dewanji & Kalbfleisch (1986), Dinse (1991), and Lindsey & Ryan (1993). Lagakos & Louis (1988) proposed a different method that assumes that tumor lethality is known. For general survival data, Zhang *et al.* (2005) introduced a method using frailties to account for the relationship between the event time and censoring time.

1.3.2 Case II

The seminal article analyzing case II interval-censored data under the proportional hazards model was published by Finkelstein (1986). This method proposed a full likelihood approach and found maximum likelihood estimates using the Newton-Raphson method. Score tests were developed in order to test the regression parameters. A downside to this method is that it requires estimation of the baseline cumulative hazard function, which is a nuisance parameter. Satten (1996) and Goggins *et al.* (1998) proposed marginal likelihood methods that do not require estimating the cumulative baseline hazard. However, as discussed in Sun (2006), one still needs to solve complicated score equations. Betensky *et al.* (2002) and Cai & Betensky (2003) proposed methods that are a mixture of full and marginal likelihoods that approximate the infinite-dimensional nuisance parameter with finite-dimensional parameters.

The proportional odds model has also been studied by many authors. Huang

& Wellner (1997) proposed a maximum likelihood approach and proved the useful asymptotic properties of their estimate. Huang & Rossini (1997) and Shen (1998) suggested using sieve maximum likelihood estimation to handle the nuisance function. Also, Rabinowitz *et al.* (2000) investigated a simpler conditional likelihood approach that only involves the regression parameters.

The additive hazards model has also been explored by many in the literature. Zeng *et al.* (2006) proposed a maximum likelihood method and Zhu *et al.* (2008) used a transformation approach. An estimating equation procedure that does not require estimation of the baseline hazard was developed by Wang *et al.* (2010). Their approach also has the advantage of being appropriate for both noninformative and informative censoring.

Rabinowitz *et al.* (1995) was one of the first papers to address case II interval-censored data using the accelerated failure time model. They developed an estimation procedure based on a class of score statistics. Also, Betensky *et al.* (2001) studied this model using estimating equations. Both approaches require estimation of the distribution of the error term.

The majority of the methods discussed above assume that the censoring is noninformative. Much less work has been done for case II interval-censored data under the assumption of informative censoring. Zhang *et al.* (2007) proposed a frailty approach for the proportional hazards model and used the EM algorithm for estimation. As mentioned, the work of Wang *et al.* (2010) under the additive hazards model can handle informative censoring. While the model for the failure times is different, both methods modeled the censoring variables using the proportional hazards model.

1.4 Outline

The remainder of this dissertation is organized as follows. In Chapter 2, we introduce work conducted in order to assess model flexibility for regression analysis of case I interval-censored data. Specifically, we are interested in determining the effectiveness of models developed for informative censoring when the censoring process is actually noninformative. This could help provide practical guidelines for determining model choice when one is unsure about the relationship between the failure time and censoring time.

Next, in Chapter 3, we examine model flexibility for regression analysis of case II interval-censored data. Again, the main focus is on investigating whether a technique developed for informative censoring is also reasonable with noninformative censoring. Extensive simulation studies are performed and used to make recommendations for model selection when the dependence between the event time and censoring times is unknown.

Then, in chapter 4, we propose new methodology for variable selection of interval-censored data. This area of research has become extremely popular in recent years as more applied fields are encountering high-dimensional data. Much work has been conducted for variable selection of right-censored data. However, the same cannot be said for interval-censored data. We investigate an approach using imputation that creates an imputed right-censored dataset from the interval-censored data. This enables the researcher to use popular variable selection techniques developed for right-censored data, and the corresponding software, which makes this method easy to use.

Finally, in Chapter 5, we discuss a number of unanswered questions that still exist that could provide motivation for future research in these areas.

Chapter 2

Model Evaluation for Regression Analysis of Current Status Data with Informative Censoring Under Various Conditions

2.1 Introduction

As discussed in the previous chapter, current status data is a type of interval-censored data where each subject is only observed at one time point. Therefore, researchers only know whether the event time of interest is larger or smaller than the observation time. Many methods have been developed in order to analyze current status data when the observation time is independent of the survival time. A brief overview of these methods was presented in Chapter 1. Much recent attention has focused on creating techniques that allow for the observation time and survival time to be dependent, which is frequently referred to as informative censoring. This chapter

aims to evaluate the method developed for regression analysis of current status data with informative censoring proposed by Zhang *et al.* (2005) under an expansive set of potential conditions. The goal of this research is to assess the model's flexibility in handling general situations that might not comply with all of the model assumptions. This could potentially provide further insight towards selecting an appropriate method for a given practical application.

Many different approaches exist for handling problems involving current status data. Several authors, including Peto (1973), Turnbull (1976) and Groeneboom & Wellner (1992), have proposed methods for nonparametric maximum likelihood estimation of the survival time distribution function. Nonparametric options also exist for treatment comparison, such as those described in Andersen & Ronn (1995) and Sun (1999). A wide range of models have also been suggested for regression analysis of current status data. The proportional hazards model was explored in Huang (1996), Lin *et al.* (1998) proposed use of the additive hazards model, and the proportional odds model was investigated in Rossini & Tsiatis (1996). Less work, however, has been done on current status data with informative censoring. Some examples exist in the context of tumorigenicity experiments. A three-state Markov model was discussed in Dewanji & Kalbfleisch (1986), and Lagakos & Louis (1988) proposed a method that utilizes known tumor lethality information. Unfortunately, these methods can be complicated and restrictive. Zhang *et al.* (2005) propose a novel method for regression analysis of current status data with informative censoring that is general and simple to implement. Consequently, this approach will be the focus of this chapter.

2.2 Methodology

2.2.1 Notation and Models

In this section we will outline the methodology proposed in Zhang *et al.* (2005) by first defining the necessary notation and models. Suppose we have a survival study with n independent subjects and define the following variables: the survival time of interest T_i, the observation time C_i, and a p-dimensional vector of possibly time-dependent covariates Z_i, for $i = 1, \ldots, n$.

It is assumed that the relationship between T_i and C_i can be modeled using an arbitrary mean zero random effect $b_i(t)$, which could also depend on time. The dependence between the survival and observation times is then characterized by the specification of their respective hazard functions. Is is assumed that the T_i's follow the additive hazards frailty model, meaning the hazard function at time t is defined by:

$$\lambda_i(t|Z_i(s), b_i(s), s \leq t) = \lambda_1(t) + \beta' Z_i(t) + b_i(t) \tag{2.1}$$

given $\{Z_i(s), b_i(s), s \leq t\}$. Here $\lambda_1(t)$ is an unknown baseline hazard function, and the covariate effect on the survival time is represented by β, a p-dimensional vector of regression parameters.

The C_i's are assumed to follow a proportional hazards frailty model given $\{Z_i(s), b_i(s), s \leq t\}$. That is, the hazard function at time t is given by:

$$\lambda_i^c(t|Z_i(s), b_i(s), s \leq t) = \lambda_2(t) \exp(\gamma' Z_i(t) + b_i(t)) \tag{2.2}$$

where $\lambda_2(t)$ is another unknown baseline hazard function, and γ represents the co-

variate effect on the observation times.

The additive hazards model and proportional hazards model are two of the most well known and most often used survival models. Lin *et al.* (1998) introduced a similar procedure for current status data with noninformative censoring. Also, the use of random effects when dealing with informative censoring has been proposed in many other situations including right-censored data (Huang & Wolfe, 2002) and longitudinal data (Wang & Taylor, 2001).

2.2.2 Parameter Estimation

Now we will consider the proposed method for estimation of the regression parameters. For simplicity, we will first examine the case where there are no censoring times. Assume the survival study gives rise to the following observed data $\{(C_i, \delta_i = I(C_i \leq T_i), Z_i(t), t \leq C_i); i = 1, \ldots, n\}$. Next, define the counting process $N_i(t) = \delta_i I(C_i \leq t) = I(C_i \leq \min(T_i, t))$ and also let $\Lambda_j(t) = \int_0^t \lambda_j(s)\mathrm{d}s$, $j = 1, 2$. Note that this counting process only jumps once if $C_i = t$ and $T_i \geq t$. Also, Zhang *et al.* (2005) show the probability for $\mathrm{d}N_i(t) = 1$ as:

$$\begin{aligned} &\mathrm{d}\ \Pr\{T_i \geq t, C_i = t | Z_i(s), s \leq t\} \\ &\qquad = E\{\mathrm{e}^{-\Lambda_1(t) - B_i(t) - \beta' Z_i^*(t)} \lambda_2(t) \mathrm{e}^{\gamma' Z_i(t) + b_i(t)} \mathrm{d}t\} = \mathrm{e}^{\gamma' Z_i(t) - \beta' Z_i^*(t)} \mathrm{d}\Lambda_0^*(t) \quad (2.3) \end{aligned}$$

where $\mathrm{d}\Lambda_0^*(t) = \mathrm{e}^{-\Lambda_1(t)} E\{\mathrm{e}^{b_i(t) - B_i(t)}\}\mathrm{d}\Lambda_2(t)$, $B_i(t) = \int_0^t b_i(s)\mathrm{d}s$, and $Z_i^*(t) = \int_0^t Z_i(s)\mathrm{d}s$.

This is an interesting result because equation (2.3) is essentially what one would get from a standard proportional hazards model. Therefore, one can define the fol-

lowing martingales:

$$M_i^*(t) = N_i(t) - \int_0^t Y_i(s) \mathrm{e}^{\gamma' Z_i(s) - \beta' Z_i^*(s)} \mathrm{d}\Lambda_0^*(s)$$

where $Y_i(t) = I(C_i \geq t)$ and then use the well-known partial likelihood approach for estimation and inference concerning β and γ.

Along these lines, next define:

$$S^{(0)}(\beta, \gamma, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \mathrm{e}^{\gamma' Z_i(t) - \beta' Z_i^*(t)}$$

$$S^{(1)}(\beta, \gamma, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) Z_i^*(t) \mathrm{e}^{\gamma' Z_i(t) - \beta' Z_i^*(t)}$$

and

$$S^{(2)}(\beta, \gamma, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) Z_i(t) \mathrm{e}^{\gamma' Z_i(t) - \beta' Z_i^*(t)}$$

Now one can estimate β and γ by solving the estimating equations $U_\beta(\beta, \gamma) = 0$ and $U_\gamma(\beta, \gamma) = 0$ where

$$U_\beta(\beta, \gamma) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i^*(t) - \frac{S^{(1)}(\beta, \gamma, t)}{S^{(0)}(\beta, \gamma, t)} \right\} \mathrm{d}N_i(t)$$

and

$$U_\gamma(\beta, \gamma) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i(t) - \frac{S^{(2)}(\beta, \gamma, t)}{S^{(0)}(\beta, \gamma, t)} \right\} \mathrm{d}N_i(t)$$

Zhang *et al.* (2005) point out that $U_\beta(\beta, \gamma, t)$ and $U_\gamma(\beta, \gamma, t)$ are the partial score functions. Therefore, the estimates obtained from this approach are maximum partial likelihood estimates. This method is advantageous for several reasons. First, one

is not required to estimate either of the two baseline hazard functions. Also, let $\hat{\theta} = (\hat{\beta}', \hat{\gamma}')'$ represent the obtained estimates of $\theta = (\beta', \gamma')'$, then $\hat{\theta}$ is a consistent estimator and has an approximately normal asymptotic distribution.

This method is easily extended to the case with censoring times. Suppose now there exists a censoring time C_i^c, which is independent of T_i and C_i, and $C_i^* = \min(C_i, C_i^c)$ is what is observed. Next, let $\xi_i = I(C_i^* = C_i)$ and define a new counting process $N_i^*(t) = \xi_i N_i(t) = I\{C_i \leq \min(T_i, C_i^c, t)\}$. Estimation now proceeds as described above by solving the partial score functions, with the exception that we now define $Y_i(t) = I(C_i^* \geq t)$. The desirable results of consistency and asymptotic normality also hold for this situation.

2.3 Simulation Results

A variety of simulation studies were conducted under differing conditions in order to evaluate the performance of the approach proposed in Zhang *et al.* (2005). The baseline hazard functions, $\lambda_1(t)$ and $\lambda_2(t)$, were set equal to one in all of the simulations. Also, each setup considered the situation with no censoring, 20% censoring, and 40% censoring. This was achieved by setting $C_i^c = \tau$, where τ is a constant used to determine the percentage of censored observations. Each study used a sample size of $n = 200$ with 1000 replications. Results are presented in tables that display the means of $\hat{\beta}$ and $\hat{\gamma}$ for several different combinations of true values for β and γ. Also, each table shows the means of the estimated standard deviations of $\hat{\beta}$ and $\hat{\gamma}$ (SEE) as well as the sample standard deviations of the point estimates (SE). Finally, the 95% empirical coverage probabilities are calculated.

2.3.1 Dependent Censoring

First, we examine the case with informative censoring. This serves as a confirmatory analysis, and gives results that can be used for comparison with the other situations. The setup is the same as in the original paper except here we are only considering the discrete covariate case. Specifically, Z is generated from a Bernoulli distribution with success probability equal to 0.5. Exponential distributions were used for both the survival and observations times with hazards defined in (2.1) and (2.2), respectively. Time-independent random effects were generated from a standard normal distribution.

Tables 2.1, 2.2, and 2.3 show the results for these simulations. It is clear that the method seems to be performing well. Means of the parameter estimates are close to their true values. The variance estimates are close, which suggests that the variance estimation procedure is valid. Also, the empirical 95% coverage probabilities are all fairly close to the desired level. The variance estimates grow as the censoring percentage increases. This can be expected since more information is lost with increased censoring.

2.3.2 Independent Censoring

The second situation considers the case with independent censoring. Using the proposed model, independent censoring is achieved by setting the latent random effects b_i's equal to zero. Then the survival and observation times are both generated from exponential distributions using the hazards defined in (2.1) and (2.2). Here we investigated performance with both a discrete and continuous covariate. For the discrete case, it was assumed Z followed a Bernoulli distribution with success probability 0.5,

and a uniform distribution over $[0, 1]$ was used for Z in the continuous case.

Table 2.4 and Table 2.5 show the simulation results for a discrete and continuous covariate, respectively, with no censoring. These results display a number of important characteristics. Overall the point estimates for β and γ seem to be unbiased for both types of covariates. In general, the SE and SEE are reasonably close, which indicates that the variance estimates are sensible. Moreover, the coverage probabilities are largely accurate. Results with 20% and 40% censoring for both the discrete and continuous case can be found in Tables 2.9-2.12. These additional results mainly tell the same story. Estimates for both the discrete and continuous case tend to be unbiased, and the coverage probabilities are all quite close to the desired size. The important difference that can been seen is in the variance estimates. Specifically, the variance tends to increase for both types of covariates as the censoring percentage increases.

2.3.3 Multivariate Random Effects

The next simulation study investigated performance when the structure of the random effects is misspecified. The proposed method assumes that the hazard functions for the survival and observation times share a random effect b_i. We examined a more general case where the relationship between the survival and observation times are characterized by an arbitrary random vector, $\mathbf{b}_i(t) = (b_i^1(t), b_i^2(t))$, with mean zero. The hazard functions for the survival and observation times, respectively, are now defined as:

$$\lambda_i(t|Z_i(s), b_i^1(s), s \leq t) = \lambda_1(t) + \beta' Z_i(t) + b_i^1(t) \tag{2.4}$$

$$\lambda_i^c(t|Z_i(s), b_i^2(s), s \leq t) = \lambda_2(t) \exp(\gamma' Z_i(t) + b_i^2(t)) \tag{2.5}$$

where $\lambda_1(t)$, $\lambda_2(t)$, β, and γ are the same as in (1) and (2).

For this case, we created current status data by first generating time-independent random effects assuming $\mathbf{b}_i \sim \text{MVN}(0, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

where ρ, representing different levels of correlation between the random effects, was set to 0.3 and 0.5. For this situation only the discrete covariate was considered with Z coming from a Bernoulli distribution with success probability equal to 0.5. Finally, survival and observation times were generated from exponential distributions using the hazards defined in equations (2.4) and (2.5).

Results for these simulation studies with no censoring are presented in tables 2.6 and 2.7. Table 2.6 shows outcomes with $\rho = 0.3$ and table 2.7 has the results when $\rho = 0.5$. Once again, additional simulations with 20% and 40% censoring can be found in Tables 2.13-2.16. It is clear from the results that the existence of a multivariate random effect causes serious problems for parameter estimation at both levels of correlation. The estimates for β and γ are biased in all cases. The coverage probabilities vary widely and are not close to the desired size. Similar conclusions can be seen when censoring is present, and once again the variance estimates increase as the censoring percentage increases.

2.3.4 Model Misspecification

The final simulation study examined the case where the hazard function for the observation times is misspecified. Unlike Zhang *et al.* (2005), where a proportional hazards frailty model is assumed, we investigated the situation where the observation times follow an additive hazards frailty model, i.e. the hazard function for the observation times is given by:

$$\lambda_i^c(t|Z_i(s), b_i(s), s \leq t) = \lambda_3(t) + \gamma' Z_i(t) + b_i(t) \tag{2.6}$$

where $\lambda_3(t)$ is an unspecified baseline hazard function, and γ once again denotes the covariate effect on the observation times.

As with the last case, we focused only on the situation where Z was generated from a Bernoulli distribution with success probability 0.5. Here, current status data was created by first generating the b_i's from a standard normal distribution. Then, survival and observation times were produced from exponential distributions with hazards given by (2.1) and (2.6), with $\lambda_3(t) = 1$.

Table 2.8 shows the simulation results for this setup with no censoring. Upon inspection, it seems that the results in this case are mixed. The method performs adequately for certain parameter combinations and poorly for others. When γ is equal to zero, the bias for $\hat{\beta}$ and $\hat{\gamma}$ is small, and the coverage probabilities are fairly accurate. However, the results deteriorate as γ increases. Bias increases for both parameters and coverage probability drops. This could possibly be explained by the fact that the additive hazards model and proportional hazards model are similar for certain parameter values. Analysis of the results with censoring, which can be found

in Tables 2.17-2.18, shows a similar outcome. The results are reasonable when γ is equal to zero and get worse as γ increases. Also, it can be seen that the variance estimates increase as the censoring percentage increases.

2.3.5 Efficiency Comparison

It is also interesting to compare the efficiency of the method of Zhang *et al.* (2005) with one designed for independent censoring when the censoring is in fact noninformative. This could help a researcher determine which approach to use when it is either known, or assumed, that the censoring is noninformative. In order to accomplish this, another simulation study was conducted under the setup described in Lin *et al.* (1998). This method was chosen for comparison because the model specification is the same as that in Zhang *et al.* (2005) except for the frailty term. Specifically, the failure times and observation times were generated according to exponential distributions with the following hazards

$$\lambda_i(t|Z_i(s), b_i(s), s \leq t) = \lambda_1(t) + \beta' Z_i(t) \tag{2.7}$$

and

$$\lambda_i^c(t|Z_i(s), b_i(s), s \leq t) = \lambda_2(t) \exp(\gamma' Z_i(t)). \tag{2.8}$$

This represents the case where the censoring is independent. The baseline hazard function for the failure times is set to be $\lambda_1(t) = 1$, and the baseline hazard for the censoring times takes the values $\lambda_2(t) = 0.5, 1.0,$ and 1.5. The true value of β is taken to be 0.5 and the covariate is generated from a uniform distribution over $(0, \sqrt{12})$. Samples sizes of $n = 100$ and $n = 200$ were investigated. 10,000 iterations were used for each combination of parameters.

The results for this simulation experiment can be found in table 2.19. The top half of the table shows the original results from Lin *et al.* (1998). The bottom portion shows the outcomes using the method of Zhang *et al.* (2005). Bias and coverage probabilities are very close for both methods. However, the method of Lin *et al.* (1998) does have smaller SE and SEE. This indicates that it might be preferred to use their method when one believes the censoring is truly noninformative. This result makes sense since the approach of Lin *et al.* (1998) was developed for this type of censoring.

2.4 Discussion

In this chapter we investigated a number of situations in order to explore the flexibility of the model proposed by Zhang *et al.* (2005). We found through extensive simulation studies that this procedure can handle certain cases beyond its intended use. This model, which was designed for use with informative censoring, satisfactorily managed data generated under noninformative censoring. This suggests that this approach would be a reasonable choice when the true nature of the censoring is unknown, provided the other assumptions hold. This technique, however, was not malleable enough to adequately deal with other situations that violate the assumptions of the model, including a bivariate random effect, and a misspecified hazard function for the observation times. In these cases, the model performed poorly in terms of both bias and empirical coverage probability. Therefore, a researcher would be wise to check the proportional hazards assumption on the observation times. This should be relatively simple since the observation times give either exact or right-censored data,

and there exist many well-known model-checking methods in this context.

This work could easily be expanded in a number of other interesting directions. For example, one could potentially investigate the effectiveness of other models intended for current status data with informative censoring when analyzing data with noninformative censoring. Also, similar questions could be asked for methods developed for the more general case II interval censored data. This particular topic will be explored in Chapter 3.

Table 2.1: Dependent censoring with discrete covariate and no censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	0.004	0.284	0.282	0.951	0.006	0.385	0.387	0.964
	0.5	0.032	0.316	0.306	0.946	0.567	0.536	0.500	0.957
	1	0.024	0.324	0.330	0.963	1.098	0.659	0.641	0.960
0.2	0	0.209	0.295	0.278	0.934	0.009	0.428	0.403	0.957
	0.5	0.215	0.302	0.298	0.946	0.545	0.547	0.516	0.948
	1	0.243	0.324	0.321	0.954	1.133	0.684	0.655	0.950
0.5	0	0.490	0.278	0.272	0.945	-0.027	0.448	0.442	0.958
	0.5	0.517	0.294	0.292	0.949	0.550	0.551	0.554	0.958
	1	0.516	0.309	0.312	0.954	1.086	0.695	0.683	0.961

Table 2.2: Dependent censoring with discrete covariate and 20% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	0.003	0.328	0.317	0.943	-0.005	0.591	0.558	0.949
	0.5	-0.001	0.349	0.330	0.942	0.528	0.670	0.637	0.944
	1	0.022	0.351	0.346	0.947	1.082	0.741	0.738	0.964
0.2	0	0.208	0.328	0.313	0.946	0.015	0.624	0.594	0.947
	0.5	0.192	0.325	0.324	0.952	0.502	0.681	0.669	0.951
	1	0.231	0.324	0.338	0.967	1.087	0.772	0.768	0.965
0.5	0	0.501	0.315	0.309	0.951	-0.001	0.659	0.656	0.961
	0.5	0.500	0.333	0.319	0.949	0.528	0.760	0.727	0.947
	1	0.515	0.334	0.331	1.077	0.822	0.821	0.961	

Table 2.3: Dependent censoring with discrete covariate and 40% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.017	0.373	0.369	0.963	-0.018	0.915	0.923	0.956
	0.5	0.020	0.392	0.379	0.953	0.591	1.003	0.997	0.958
	1	0.013	0.403	0.390	0.950	1.061	1.118	1.082	0.963
0.2	0	0.209	0.385	0.369	0.945	0.033	1.064	1.008	0.952
	0.5	0.186	0.369	0.376	0.959	0.470	1.058	1.064	0.968
	1	0.213	0.387	0.387	0.956	1.063	1.171	1.152	0.958
0.5	0	0.505	0.390	0.368	0.946	0.058	1.199	1.146	0.949
	0.5	0.518	0.399	0.375	0.944	0.521	1.307	1.200	0.949
	1	0.550	0.382	0.382	0.959	1.176	1.277	1.270	0.958

Table 2.4: Independent censoring with discrete covariate and no censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.006	0.29	0.29	0.952	-0.0009	0.457	0.435	0.952
	0.5	0.019	0.332	0.313	0.946	0.55	0.585	0.537	0.947
	1	0.026	0.329	0.336	0.949	1.11	0.679	0.669	0.971
0.2	0	0.226	0.294	0.286	0.94	0.007	0.471	0.448	0.96
	0.5	0.217	0.304	0.306	0.953	0.547	0.556	0.553	0.968
	1	0.218	0.329	0.325	0.954	1.10	0.709	0.675	0.959
0.5	0	0.505	0.29	0.282	0.945	0.06	0.508	0.491	0.956
	0.5	0.512	0.305	0.299	0.954	0.553	0.624	0.593	0.955
	1	0.534	0.32	0.319	0.955	1.12	0.733	0.722	0.959

Table 2.5: Independent censoring with continuous covariate and no censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean $\hat{\gamma}$	SE	SEE	CP	Mean $\hat{\beta}$	SE	SEE	CP
0	0	-0.017	0.510	0.507	0.955	-0.012	0.784	0.754	0.947
	0.5	-0.002	0.504	0.536	0.967	0.521	0.863	0.888	0.960
	1	0.071	0.599	0.566	0.944	1.147	1.092	1.048	0.943
0.2	0	0.189	0.482	0.492	0.968	-0.013	0.756	0.763	0.968
	0.5	0.200	0.523	0.522	0.945	0.529	0.936	0.909	0.955
	1	0.216	0.554	0.550	0.952	1.050	1.080	1.057	0.953
0.5	0	0.502	0.489	0.480	0.953	0.004	0.835	0.823	0.956
	0.5	0.480	0.499	0.503	0.957	0.491	0.952	0.947	0.960
	1	0.480	0.540	0.526	0.940	0.970	1.134	1.082	0.946

Table 2.6: Multivariate random effect with $\rho = 0.3$ and no censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.001	0.249	0.247	0.964	0.005	0.348	0.315	0.979
	0.5	-0.074	0.260	0.257	0.942	0.169	0.411	0.359	0.740
	1	-0.129	0.282	0.272	0.915	0.450	0.512	0.448	0.602
0.2	0	0.170	0.261	0.244	0.932	0.047	0.392	0.340	0.968
	0.5	0.095	0.259	0.252	0.927	0.191	0.438	0.371	0.754
	1	0.036	0.257	0.267	0.913	0.483	0.534	0.462	0.630
0.5	0	0.401	0.245	0.239	0.914	0.076	0.413	0.364	0.973
	0.5	0.328	0.248	0.248	0.896	0.254	0.475	0.408	0.806
	1	0.302	0.260	0.261	0.867	0.571	0.561	0.501	0.731

Table 2.7: Multivariate random effect with $\rho = 0.5$ and no censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.006	0.263	0.238	0.940	0.052	0.400	0.298	0.942
	0.5	-0.176	0.275	0.255	0.900	0.061	0.383	0.343	0.643
	1	-0.224	0.267	0.265	0.840	0.366	0.417	0.397	0.541
0.2	0	0.142	0.274	0.236	0.920	-0.038	0.381	0.315	0.940
	0.5	-0.031	0.247	0.247	0.920	0.124	0.492	0.375	0.620
	1	0.035	0.244	0.260	0.900	0.562	0.419	0.434	0.664
0.5	0	0.492	0.230	0.232	0.940	0.150	0.334	0.352	0.981
	0.5	0.259	0.258	0.242	0.840	0.167	0.258	0.385	0.847
	1	0.301	0.271	0.256	0.860	0.654	0.673	0.490	0.643

Table 2.8: Additive hazards model for observation times with no censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	0.07	0.241	0.232	0.940	0.001	0.202	0.161	0.988
	0.5	0.031	0.280	0.281	0.952	0.624	0.438	0.415	0.953
	1	0.05	0.318	0.314	0.957	1.197	0.684	0.637	0.953
0.2	0	0.130	0.236	0.239	0.947	-0.206	0.234	0.223	0.899
	0.5	0.113	0.278	0.267	0.933	0.446	0.406	0.359	0.877
	1	0.130	0.293	0.302	0.946	1.032	0.603	0.580	0.919
0.5	0	0.314	0.260	0.248	0.876	-0.363	0.342	0.305	0.763
	0.5	0.177	0.238	0.239	0.694	0.072	0.280	0.230	0.406
	1	0.231	0.293	0.284	0.808	0.738	0.504	0.498	0.822

Table 2.9: Independent censoring with discrete covariate and 20% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.003	0.331	0.321	0.944	0.021	0.605	0.576	0.958
	0.5	-0.006	0.346	0.335	0.955	0.523	0.672	0.652	0.966
	1	0.036	0.362	0.352	0.95	1.11	0.772	0.759	0,962
0.2	0	0.211	0.308	0.316	0.961	0.027	0.608	0.609	0.961
	0.5	0.207	0.342	0.329	0.938	0.548	0.702	0.684	0.956
	1	0.238	0.346	0.344	0.949	1.14	0.835	0.79	0.953
0.5	0	0.513	0.321	0.213	0.945	0.037	0.694	0.668	0.949
	0.5	0.507	0.323	0.322	0.954	0.529	0.757	0.741	0.962
	1	0.525	0.344	0.336	0.952	1.09	0.871	0.841	0.956

Table 2.10: Independent censoring with discrete covariate and 40% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	0.012	0.372	0.371	0.949	0.032	0.915	0.919	0.962
	0.5	0.0009	0.396	0.382	0.949	0.479	1.06	0.994	0.945
	1	0.011	0.398	0.394	0.953	1.12	1.10	1.08	0.956
0.2	0	0.185	0.366	0.368	0.954	-0.037	1.01	0.996	0.959
	0.5	0.191	0.383	0.377	0.947	0.491	1.11	1.05	0.949
	1	0.21	0.391	0.386	0.963	1.05	1.18	1.14	0.955
0.5	0	0.509	0.388	0.369	0.947	0.005	1.19	1.14	0.942
	0.5	0.522	0.384	0.377	0.956	0.558	1.28	1.20	0.948
	1	0.517	0.379	0.384	0.962	1.06	1.28	1.26	0.95

Table 2.11: Independent censoring with continuous covariate and 20% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.021	0.558	0.554	0.954	-0.026	1.024	0.986	0.944
	0.5	0.024	0.584	0.578	0.954	0.578	1.153	1.102	0.951
	1	-0.001	0.618	0.598	0.944	1.024	1.262	1.230	0.955
0.2	0	0.170	0.571	0.547	0.940	-0.039	1.11	1.051	0.950
	0.5	0.208	0.583	0.567	0.944	0.519	1.202	1.163	0.951
	1	0.239	0.602	0.586	0.954	1.118	1.361	1.288	0.951
0.5	0	0.496	0.541	0.537	0.959	-0.063	1.187	1.161	0.954
	0.5	0.541	0.558	0.552	0.956	0.600	1.289	1.260	0.956
	1	0.515	0.572	0.571	0.958	0.987	1.369	1.379	0.956

Table 2.12: Independent censoring with continuous covariate and 40% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.019	0.607	0.639	0.960	-0.077	1.609	1.590	0.956
	0.5	0.008	0.681	0.659	0.946	0.518	1.783	1.699	0.939
	1	-0.02	0.701	0.678	0.943	0.945	1.847	1.815	0.955
0.2	0	0.163	0.625	0.636	0.960	-0.108	1.769	1.719	0.943
	0.5	0.202	0.639	0.650	0.963	0.496	1.761	1.815	0.966
	1	0.214	0.659	0.667	0.957	1.043	1.953	1.940	0.940
0.5	0	0.522	0.643	0.633	0.948	0.003	1.976	1.95	0.946
	0.5	0.534	0.653	0.644	0.954	0.573	2.063	2.053	0.956
	1	0.540	0.714	0.658	0.936	1.078	2.296	2.152	0.944

Table 2.13: Multivariate random effect with $\rho = 0.3$ and 20% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	0.002	0.295	0.284	0.941	0.034	0.544	0.566	0.973
	0.5	-0.074	0.289	0.292	0.963	0.213	0.667	0.601	0.907
	1	-0.059	0.311	0.302	0.925	0.733	0.704	0676	0.932
0.2	0	0.196	0.297	0.281	0.952	0.119	0.638	0.618	0.991
	0.5	0.103	0.294	0.289	0.927	0.244	0.682	0.635	0.928
	1	0.073	0.294	0.297	0.933	0.610	0.750	0.681	0.874
0.5	0	0.427	0.251	0.278	0.965	0.106	0.670	0.679	0.956
	0.5	0.317	0.270	0.280	0.883	0.350	0.688	0.691	0.940
	1	0.339	0.310	0.289	0.938	0.681	0.722	0.743	0.925

Table 2.14: Multivariate random effect with $\rho = 0.3$ and 40% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	0.033	0.373	0.338	0.930	0.049	1.233	1.101	0.940
	0.5	-0.065	0.344	0.344	0.970	0.180	1.09	1.12	0.942
	1	-0.180	0.3877	0.349	0.900	0.243	1.218	1.147	0.883
0.2	0	0.169	0.396	0.341	0.910	0.162	1.279	1.232	0.971
	0.5	0.150	0.307	0.342	0.990	0.370	1.170	1.241	0.986
	1	0.119	0.363	0.347	0.940	0.742	1.325	1.250	0.940
0.5	0	0.409	0.311	0.333	0.980	-0.015	1.348	1.351	0.960
	0.5	0.389	0.305	0.339	0.950	0.500	1.271	1.358	0.991
	1	0.326	0.386	0.342	0.850	0.763	1.547	1.383	0.954

Table 2.15: Multivariate random effect with $\rho = 0.5$ and 20% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.046	0.284	0.285	0.980	-0.029	0.666	0.587	0.926
	0.5	-0.029	0.327	0.295	0.940	0.247	0.614	0.624	0.921
	1	-0.013	0.326	0.302	0.920	0.588	0.711	0.657	0.867
0.2	0	0.142	0.289	0.282	0.960	-0.07	0.502	0.605	0.983
	0.5	0.090	0.260	0.283	0.940	0.352	0.534	0.630	0.982
	1	0.019	0.279	0.294	0.940	0.548	0.733	0.686	0.922
0.5	0	0.374	0.238	0.278	0.940	0.103	0.607	0.674	0.960
	0.5	0.356	0.338	0.282	0.880	0.218	0.747	0.699	0.901
	1	0.321	0.342	0.291	0.860	0.666	0.861	0.740	0.922

Table 2.16: Multivariate random effect with $\rho = 0.5$ and 40% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.070	0.330	0.343	0.980	-0.185	1.057	1.110	0.961
	0.5	-0.058	0.349	0.348	0.960	0.200	1.131	1.152	0.942
	1	-0.043	0.372	0.353	0.940	0.781	1.239	1.154	0.943
0.2	0	0.139	0.364	0.336	0.940	-0.071	1.267	1.164	0.967
	0.5	0.137	0.435	0.341	0.860	0.238	1.503	1.219	0.925
	1	0.073	0.385	0.350	0.880	0.790	1.406	1.262	0.908
0.5	0	0.510	0.341	0.337	0.980	0.581	1.879	1.375	0.927
	0.5	0.415	0.263	0.337	0.960	0.494	1.165	1.331	0.981
	1	0.306	0.335	0.341	0.940	0.639	1.335	1.395	0.920

Table 2.17: Additive hazards model for observation times with 20% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	0.012	0.301	0.303	0.956	0.010	0.552	0.536	0.966
	0.5	0.009	0.321	0.317	0.959	0.543	0.641	0.629	0.968
	1	0.020	0.356	0.335	0.944	1.131	0.839	0.765	0.953
0.2	0	0.162	0.318	0.303	0.950	-0.177	0.600	0.585	0.952
	0.5	0.099	0.336	0.313	0.922	0.420	0.684	0.638	0.940
	1	0.106	0.325	0.327	0.949	0.977	0.785	0.751	0.939
0.5	0	0.320	0.303	0.301	0.917	-0.352	0.659	0.641	0.918
	0.5	0.222	0.312	0.307	0.852	0.197	0.674	0.643	0.925
	1	0.259	0.332	0.321	0.881	0.822	0.784	0.749	0.933

Table 2.18: Additive hazards model for observation times with 40% censoring

		$\hat{\gamma}$				$\hat{\beta}$			
γ	β	Mean	SE	SEE	CP	Mean	SE	SEE	CP
0	0	-0.010	0.372	0.363	0.945	-0.023	1.065	1.042	0.952
	0.5	0.019	0.378	0.375	0.952	0.556	1.173	1.127	0.951
	1	0.003	0.406	0.385	0.945	1.015	1.306	1.213	0.943
0.2	0	0.149	0.396	0.364	0.929	-0.119	1.195	1.120	0.945
	0.5	0.089	0.393	0.370	0.928	0.360	1.228	1.150	0.940
	1	0.127	0.393	0.381	0.944	1.022	1.314	1.238	0.938
0.5	0	0.315	0.356	0.364	0.929	-0.344	1.269	1.230	0.939
	0.5	0.233	0.377	0.368	0.881	0.200	1.261	1.207	0.947
	1	0.268	0.378	0.376	0.906	0.851	1.350	1.290	0.951

Table 2.19: Efficiency comparison with method of Lin *et al.* (1998)

	n=100			n=200		
	$\lambda_{c,0} = 0.5$	1.0	1.5	$\lambda_{c,0} = 0.5$	1.0	1.5
Bias	0.04	0.03	0.02	0.02	0.02	0.01
SE	0.38	0.42	0.50	0.25	0.29	0.33
SEE	0.38	0.41	0.49	0.25	0.28	0.33
95% CP	0.95	0.96	0.95	0.95	0.95	0.95
Bias	0.07	0.03	0.02	0.03	0.03	0.01
SE	0.66	0.75	0.92	0.42	0.49	0.60
SEE	0.61	0.72	0.87	0.40	0.48	0.58
95% CP	0.95	0.96	0.95	0.95	0.95	0.95

Chapter 3

Model Evaluation for Regression Analysis of Case II Interval-Censored Data with Informative Censoring Under Various Conditions

3.1 Introduction

As we have seen in earlier chapters, interval-censored data is a common occurrence in survival analysis. Again, by interval censored data, we mean that the failure time under study is only known to belong to some window of observation times. In this chapter we will explore case II interval-censored data, which is a more general type of interval censoring than was discussed in Chapter 2. Recently, there has been an increased interest in developing techniques to handle what is known as informative

interval censoring. This type of interval censoring arises when the event time is somehow related to, or dependent on, the observation process. The purpose of this chapter is to evaluate one method designed for regression analysis of informative interval-censored data that was proposed by Zhang *et al.* (2007). We hope to establish a range of conditions under which this method performs well, and also strive to identify situations where this approach is not appropriate.

There are many methods available for analyzing noninformative interval censored data. Several authors considered the well-known proportional hazards model. Leading the way was the influential work of Finkelstein (1986), which used a maximum likelihood approach. Other models have also been explored. For example, the proportional odds model was examined by Huang & Wellner (1997) and Betensky *et al.* (2001) utilized the accelerate failure time model. Less work, however, has been completed for informative interval censoring. A popular approach with dependent interval censoring is to make use of frailty terms. Zhang *et al.* (2005) proposed such a method for current status data. Finkelstein *et al.* (2002) and Betensky & Finkelstein (2002) considered general interval censored data with informative censoring. One limitation of the method of Betensky & Finkelstein (2002) is the requirement of follow-up after the event time. The method proposed by Zhang *et al.* (2007) for regression analysis of data with informative interval censoring avoids this issue. Moreover, it uses the popular proportional hazards model with frailty terms. These benefits motivated our evaluation of this approach.

3.2 Methodology

3.2.1 Notation and Models

First we will describe the methodology proposed in Zhang *et al.* (2007) by defining the necessary notation and introducing the models. Suppose we have a survival study with n independent subjects, and define T_i to be the failure time for subject $i = 1, \ldots, n$. We also need two additional random variables, U_i and V_i, such that $U_i \leq V_i$. Since we have interval censored data, T_i is not observed directly. It is only known whether T_i is less than or equal to U_i, between U_i and V_i, or greater than V_i. Next, we consider observing a $p \times 1$ vector of covariates Z_i for each subject. Finally, define $W_i = U_i - V_i$ to be the gap time between our two observation times.

It is assumed that the failure time and observation times are related through an unobserved random vector $b_i = (b_{1i}, b_{2i}, b_{3i})'$. The relationship among the variables is then modeled using the following hazard functions for T_i, U_i and W_i:

$$\lambda_i^{(T)}(t|Z_i, b_i) = \lambda_{t0}(t) \exp(\beta_t' Z_i + b_{1i}) \tag{3.1}$$

$$\lambda_i^{(U)}(t|Z_i, b_i) = \lambda_{u0}(t) \exp(\beta_u' Z_i + b_{2i}) \tag{3.2}$$

$$\lambda_i^{(W)}(t|Z_i, b_i) = \lambda_{w0}(t) \exp(\beta_w' Z_i + b_{3i}) \tag{3.3}$$

where β_t, β_u and β_w are $p \times 1$ vectors of regression parameters, and $\lambda_{t0}(t)$, $\lambda_{u0}(t)$ and $\lambda_{w0}(t)$ are unknown baseline hazard functions.

In addition, it is assumed that the latent random vector b_i follows a multivariate

normal distribution. Specifically, $b_i \sim$ i.i.d. $\mathrm{N}(0, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}$$

The values of the various σ's in the covariance matrix describe the relationship between the failure time and observation times. For example, if $\sigma_{12} = 0$ then the failure time T_i is independent of the first observation time U_i, given Z_i.

Next, we will describe construction of the likelihood function. Note that the exact value of T_i in not known since we have interval censored data. Therefore, the observed data can be expressed by $\{U_i, V_i, \delta_{1i}, \delta_{2i}, Z_i\}$ for $i = 1, \ldots, n$, where $\delta_{1i} = I(T_i \leq U_i)$ and $\delta_{2i} = I(U_i < T_i \leq V_i)$ are indicators identifying the interval containing T_i. Now define

$$(L_i, R_i] = \begin{cases} (0, U_i] & \text{if } \delta_{1i} = 1 \\ (U_i, V_i] & \text{if } \delta_{2i} = 1 \\ (V_i, \infty] & \text{otherwise} \end{cases}$$

and let $0 = s_0 < s_1 < \ldots < s_m = \infty$ be the set of times containing each L_i and R_i. Also, consider $\Lambda_{t0}(t) = \int_0^t \lambda_{t0}(u)\mathrm{d}u$ and $\gamma_j = \log \Lambda_{t0}(s_j)$. Finally, now also let $\gamma = (\gamma_1, \ldots, \gamma_{m-1})'$, $\theta = (\beta_t', \beta_u', \beta_w', \gamma', \sigma_{kl}, 1 \leq k \leq l \leq 3)'$, and $\Delta_i = (\delta_{1i}, \delta_{2i})$. One can now build the likelihood function of the observed data. Note that conditional on (U_i, W_i, Z_i, b_i), the likelihood for subject i is given by:

$$L_{\Delta_i|U_i,W_i,b_i}(\theta) = \sum_{j=1}^{m} \alpha_{ij}[\exp\{-\exp(\beta_t' Z_i + b_{1i} + \gamma_{j-1})\} - \exp\{-\exp(\beta_t' Z_i + b_{1i} + \gamma_j)\}]$$

where $\alpha_{ij} = 1$ if $(s_{j-1}, s_j]$ is a subset of $(L_i, R_i]$ and 0 otherwise. The likelihood functions for U_i and W_i, conditional on (Z_i, b_i), have the forms:

$$L_{U_i|b_i}(\theta) = \lambda_{u0}(U_i)\ \exp\{\beta_u' Z_i + b_{2i}\}\ \exp\{-\exp(\beta_u' Z_i + b_{2i})\Lambda_{u0}(U_i)\}$$

and

$$L_{W_i|b_i}(\theta) = \{\lambda_{w0}(W_i)\ \exp\{\beta_w' Z_i + b_{3i}\}\ \exp\{-\exp(\beta_w' Z_i + b_{3i})\Lambda_{w0}(W_i)\}\}^{\Psi_i}$$

where $\Lambda_{u0}(t) = \int_0^t \lambda_{u0}(u)\mathrm{d}u$, $\Lambda_{w0}(t) = \int_0^t \lambda_{w0}(u)\mathrm{d}u$, and $\Psi_i = I(W_i < \infty)$. Now define $O_i = \{\Delta_i, \Psi_i, U_i, W_i, Z_i\}$ to be the full observed data from subject i, and let $O = \{O_1, \ldots, O_n\}$ denote the combined data from all subjects. We can now represent the full likelihood by:

$$L_O(\theta) = \prod_{i=1}^{n} L_i(\theta; O_i) = \prod_{i=1}^{n} \int L_{\Delta_i|U_i,W_i,b_i}(\theta) L_{U_i|b_i}(\theta) L_{W_i|b_i}(\theta) f(b_i; \Sigma)\mathrm{d}b_i$$

where $f(b_i; \Sigma)$ is the density function of b_i.

3.2.2 Parameter Estimation

Maximization of the likelihood function is not straightforward since the b_i's are unknown. Therefore, the authors propose using the EM algorithm in order to estimate the unknown parameters. The complete data is defined to be $\{(O_i, b_i), i = 1, \ldots, n\}$, and in typical fashion, one alternates between calculating the expectation of the log-likelihood, and then updating the estimate by maximizing the complete data likelihood. Variance estimation is achieved using Louis' formula. The procedure is quite

complex, and a detailed discussion including recommendations for implementing the algorithm can be found in Zhang *et al.* (2007).

3.3 Simulation Results

A number of different simulation studies were conducted in order to assess the performance of the methodology proposed in Zhang *et al.* (2007). For simplicity, the baseline hazard functions $\lambda_{t0}(t)$, $\lambda_{u0}(t)$ and $\lambda_{w0}(t)$ were set equal to one in each setup. Also, the off-diagonal elements of the covariance matrix were all set equal to 0.03. This produces correlation coefficients of 0.75 among b_{1i}, b_{2i} and b_{3i}. Covariates were generated for both the continuous and discrete case. The Z_i were generated either from a uniform distribution over $[-1, 1]$ or a Bernoulli distribution with success probability of 0.5. Each simulation used a sample size of $n = 200$ with 1000 iterations unless otherwise noted.

Results are summarized in tables using several different statistics. The bias in estimating β_t, β_u and β_w is calculated by taking the mean of the parameter estimates minus the true value. Each table also shows the sample standard deviation of the point estimates (SE) as well as the mean of the estimated standard errors (SEE). Finally, 95% empirical coverage probabilities are calculated.

3.3.1 Dependent Censoring

The first setup was a confirmatory simulation with dependent censoring. These simulations verify the original results and serve as a point of comparison for all the new cases. Here, the data is generated according to the assumptions of the paper. Survival

times were generated from an exponential distribution with hazard function given by (3.1). The first observation time and gap time were also generated from exponential distributions with hazards (3.2) and (3.3), respectively. In accordance with the original paper, we set $\beta_t = \beta_u = \beta_w = 1$.

The dependent censoring results for both a continuous covariate and a discrete covariate are in table 3.1. Both results look very good. The bias is small for all three parameters. The variance estimates are all similar, and each coverage probability is around the specified 95%. Also, all of these outcomes closely match the corresponding results found in Zhang et al. (2007).

3.3.2 Independent Censoring

The next simulations investigated performance when the censoring is independent. This setup can be obtained from the proposed model by setting the latent random effects equal to zero for each subject, i.e. $b_{1i} = b_{2i} = b_{3i} = 0$ for all i. Next, the survival time, first observation time, and gap time are generated from exponential distributions with hazards defined according to equations (3.1), (3.2) and (3.3), respectively. Also, the true parameter values used here are $\beta_t = \beta_u = \beta_w = 1$.

Table 3.2 shows the outcomes for these simulations. The results seem to indicate that the proposed method does well in the presence of independent censoring. Bias for all three parameters is reasonably small for both the continuous and discrete covariate. The SE and SEE are always quite close implying that the variance estimation is doing a good job, and the 95% coverage probabilities are also fairly close to the desired values. Moreover, these results are comparable to the dependent censoring setup in table 3.1.

3.3.3 Model Misspecification

The next simulation studies examined cases where the hazard functions were misspecified. The proposed method assumes that all three hazard functions follow a proportional hazards model, and we were interested in evaluating the effectiveness of this approach when these assumptions are violated. To achieve this, the proportional hazards model was replaced with the additive hazards model in various combinations.

The first setup was a worst-case scenario where all three hazard functions were misspecified. That is, the hazard functions for T_i, U_i and W_i were defined as:

$$\lambda_i^{(T)}(t|Z_i, b_i) = \lambda_{t0}(t) + \beta_t^{'} Z_i + b_{1i} \tag{3.4}$$

$$\lambda_i^{(U)}(t|Z_i, b_i) = \lambda_{u0}(t) + \beta_u^{'} Z_i + b_{2i} \tag{3.5}$$

$$\lambda_i^{(W)}(t|Z_i, b_i) = \lambda_{w0}(t) + \beta_w^{'} Z_i + b_{3i} \tag{3.6}$$

where once again β_t, β_u and β_w are $p \times 1$ vectors of regression parameters, and $\lambda_{t0}(t)$, $\lambda_{u0}(t)$ and $\lambda_{w0}(t)$ are unknown baseline hazard functions.

The values for T_i, U_i and W_i were again generated from exponential distributions, but the hazard functions were specified using equations (3.4), (3.5) and (3.6). The true values for β_u and β_w were always equal to 1, and β_t took the values 0, 0.5, and 1.

The results for these simulations with $\beta_t = 1$ can be found in table 3.3. It is clear from these simulations that the method performed poorly under these conditions. The bias is very large for all three parameters with both the continuous and discrete covariates, and the coverage probabilities are terrible. The results are similar when

β_t is 0 and 0.5. The only exception is that the results for β_t improve as β_t decreases for the uniform covariate. This could possibly be explained by the fact that the proportional hazards model and additive hazards model behave alike under these conditions. The full tables with results for all parameter values are in the appendix in table 3.4.

Next, we considered the situation where only the hazards for U and W were misspecified. The survival times were generated from an exponential distribution using the correct proportional hazards model given by equation (3.1). The values for U and W are also from an exponential distribution, but the hazards are defined using the additive hazards model and equations (3.5) and (3.6), respectively. Once again the true values for β_u and β_w are set equal to 1, and β_t takes the values 0, 0.5, and 1.

Table 3.5 summarizes the outcomes of these simulations with $\beta_t = 1$. There are still serious problems with the estimates for β_u and β_w in terms of bias and coverage probability. This can be expected since U and W were generated with incorrect hazard functions. However, the results are interesting because the estimation for β_t is quite good for both cases. The bias is small, the variance estimates are close, and the coverage probabilities are right around 95%. Results are nearly identical when β_t is equal to 0 and 0.5, and once again these additional results can be found in table 3.6. This is promising since the primary goal is to estimate the covariate effect on the survival times. However, more research should go into explaining this curious outcome. It is possible that the variances and covariances of the random effects are small enough that they behave as if they are independent.

The last setup for these simulations tested the case where only the survival time hazard function is misspecified. Therefore, T_i was generated from an exponential

distribution with hazard function given by equation (3.4) while U_i and W_i were generated from exponential distributions using hazards defined in (3.2) and (3.3). We investigated an extensive range of possible values for β_t, β_u, and β_w for both covariate types.

The results for the continuous covariate are displayed in table 3.7. Table 3.9 has the results when Z_i follows a Bernoulli distribution. Since the influence of the covariates on T_i is the main focus of the analysis, we only present the information for β_t in order to reduce potential confusion, and make the table easier to digest. Here we can see that these simulations had a mixed outcome. It seems that the estimation procedure is performing well only for certain parameter combinations. This could possibly be explained by the fact that the proportional hazards model and additive hazards model are similar for some values. Nevertheless, the results become significantly worse as β_t gets larger. Also, the omitted data for β_u and β_w indicate that the model performs well in estimating these parameters. This conclusion is not too surprising since U and W were properly specified with the proportional hazards model.

The previous results motivated one final model misspecification simulation. The setup is nearly identical except the sample size is increased to $n = 400$, and the only parameter combinations examined were those that exhibited large bias in the preceding simulation. A summary of this simulation study can be found in table 3.8. Increasing the sample size unfortunately did not lead to a reduction in bias. The variance estimates did get smaller but this resulted in the coverage probabilities becoming far worse.

3.4 Discussion

This chapter examined a variety of situations in order to assess the flexibility of the method proposed by Zhang *et al.* (2007) for regression analysis of survival data with informative interval censoring. The simulation studies were designed to evaluate performance under circumstances beyond the intended scope of the original work. Overall the results were mixed, but we now have a more complete understanding of when this method produces accurate parameter estimates.

This method worked quite well when the censoring was noninformative even though it was developed to deal specifically with informative censoring. This suggests that this approach could be an acceptable choice when a researcher is uncertain whether or not they informative censoring. Also, our results indicate that this method is somewhat sensitive to the proper specification of the hazard functions. Generating data using the additive hazards model caused bias in the estimation of the corresponding parameters. Interestingly, bias only existed for the parameters associated with the variables that had improper hazard functions. In order to explain these results, it would be beneficial to explore these setups in more detail by examining a wider range of variances and correlations for the random effects.

Table 3.1: Dependent censoring

	Continuous covariate				Discrete covariate			
Parameter	Bias	SE	SEE	CP	Bias	SE	SEE	CP
β_t	-0.0224	0.1846	0.1856	0.955	0.0210	0.1485	0.1473	0.948
β_u	-0.0266	0.1454	0.1401	0.939	-0.0185	0.1620	0.1570	0.941
β_w	-0.0251	0.1463	0.1400	0.937	-0.0133	0.1586	0.1570	0.948

Table 3.2: Independent censoring

	Continuous covariate				Discrete covariate			
Parameter	Bias	SE	SEE	CP	Bias	SE	SEE	CP
β_t	-0.0044	0.1836	0.1858	0.947	0.0238	0.1330	0.1463	0.965
β_u	0.0059	0.1430	0.1410	0.950	0.0105	0.1648	0.1583	0.940
β_w	0.0075	0.1442	0.1411	0.947	0.0216	0.1522	0.1585	0.957

Table 3.3: Additive hazard model for T, U, and W with $\beta_t = 1$

	Continuous covariate				Discrete covariate			
Parameter	Bias	SE	SEE	CP	Bias	SE	SEE	CP
β_t	-0.2424	0.1766	0.1866	0.738	-0.2796	0.1441	0.1463	0.496
β_u	-0.2916	0.2563	0.2514	0.795	-0.2817	0.1492	0.1510	0.510
β_w	-0.2652	0.2636	0.2518	0.795	-0.2651	0.1512	0.1515	0.549

Table 3.4: Full results for additive hazard model for T, U, and W

	Continuous covariate											
	$\hat{\beta}_t$				$\hat{\beta}_u$				$\hat{\beta}_w$			
β_t	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
0	-0.0118	0.1860	0.1906	0.957	-0.2746	0.2624	0.2518	0.792	-0.2723	0.2627	0.2523	0.793
0.5	-0.0784	0.1685	0.1855	0.953	-0.2967	0.2648	0.2521	0.751	-0.2894	0.2585	0.2523	0.783
1	-0.2424	0.1766	0.1866	0.738	-0.2916	0.2563	0.2514	0.795	-0.2652	0.2636	0.2518	0.795
	Discrete covariate											
	$\hat{\beta}_t$				$\hat{\beta}_u$				$\hat{\beta}_w$			
β_t	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
0	0.7122	0.1473	0.1469	0.001	-0.2712	0.1568	0.1514	0.540	-0.2767	0.1563	0.1511	0.555
0.5	0.2105	0.1448	0.1459	0.708	-0.2619	0.1522	0.1513	0.582	-0.2578	0.1538	0.1516	0.590
1	-0.2796	0.1441	0.1463	0.496	-0.2817	0.1492	0.1510	0.510	-0.2651	0.1512	0.1515	0.549

Table 3.5: Additive hazard model for U, and W with $\beta_t = 1$

	Continuous covariate				Discrete covariate			
Parameter	Bias	SE	SEE	CP	Bias	SE	SEE	CP
β_t	0.0118	0.1801	0.1901	0.962	0.0115	0.1485	0.1511	0.953
β_u	-0.2786	0.2667	0.2519	0.777	-0.2704	0.1528	0.1513	0.563
β_w	-0.2701	0.2468	0.2520	0.814	-0.2599	0.1524	0.1515	0.590

Table 3.6: Full results for additive hazard model for U, and W

	Continuous covariate											
	$\hat{\beta}_t$				$\hat{\beta}_u$				$\hat{\beta}_w$			
β_t	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
0	-0.0095	0.1829	0.1897	0.957	-0.2953	0.2592	0.2520	0.771	-0.2857	0.2736	0.2517	0.764
0.5	0.0248	0.1766	0.1863	0.965	-0.3123	0.2606	0.2517	0.741	-0.2864	0.2638	0.2521	0.785
1	0.0118	0.1801	0.1901	0.962	-0.2786	0.2667	0.2519	0.777	-0.2701	0.2468	0.2520	0.814
	Discrete covariate											
	$\hat{\beta}_t$				$\hat{\beta}_u$				$\hat{\beta}_w$			
β_t	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
0	-0.0026	0.1507	0.1469	0.950	-0.2731	0.1589	0.1514	0.544	-0.2578	0.1542	0.1511	0.608
0.5	0.0073	0.1414	0.1445	0.961	-0.2682	0.1561	0.1515	0.565	-0.2625	0.1594	0.1514	0.573
1	0.0115	0.1485	0.1511	0.953	-0.2704	0.1528	0.1513	0.563	-0.2599	0.1524	0.1515	0.590

Table 3.7: Additive hazard model for T with continuous covariate

β_t	β_u	β_w	Bias $\hat{\beta}_t$	SE	SEE	CP
0	0	0	-0.0115	0.1847	0.1934	0.962
		0.5	-0.0254	0.1857	0.1911	0.959
		1	-0.0143	0.1844	0.1914	0.957
	0.5	0	-0.0065	0.1851	0.1922	0.958
		0.5	-0.0264	0.1877	0.1899	0.953
		1	-0.0230	0.1837	0.1905	0.956
	1	0	-0.0249	0.1877	0.1951	0.964
		0.5	-0.0288	0.1848	0.1946	0.961
		1	-0.0139	0.1946	0.1954	0.952
0.5	0	0	-0.0728	0.1816	0.1974	0.960
		0.5	-0.0715	0.1845	0.1944	0.944
		1	-0.0769	0.1811	0.1923	0.943
	0.5	0	-0.0730	0.1843	0.1920	0.955
		0.5	-0.0746	0.1866	0.1879	0.934
		1	-0.0764	0.1789	0.1871	0.942
	1	0	-0.0798	0.1772	0.1929	0.949
		0.5	-0.0699	0.1794	0.1880	0.946
		1	-0.0731	0.1785	0.1876	0.944
1	0	0	-0.2412	0.1895	0.2061	0.793
		0.5	-0.2415	0.2078	0.2004	0.733
		1	-0.2367	0.1960	0.1985	0.763
	0.5	0	-0.2464	0.1921	0.1958	0.757
		0.5	-0.2436	0.1823	0.1911	0.754
		1	-0.2365	0.1887	0.1887	0.758
	1	0	-0.2404	0.1833	0.1933	0.767
		0.5	-0.2385	0.1763	0.1881	0.766
		1	-0.2461	0.1868	0.1848	0.734

Table 3.8: Additive hazard model for T with continuous covariate, $n = 400$

β_t	β_u	β_w	Bias $\hat{\beta}_t$	SE	SEE	CP
1	0	0	-0.2444	0.1321	0.1402	0.590
		0.5	-0.2333	0.1285	0.1363	0.595
		1	-0.2278	0.1300	0.1353	0.613
	0.5	0	-0.2416	0.1226	0.1337	0.548
		0.5	-0.2421	0.1229	0.1306	0.528
		1	-02384	0.1172	0.1289	0.540
	1	0	-0.2502	0.1202	0.1315	0.513
		0.5	-0.2529	0.1200	0.1284	0.474
		1	-0.2426	0.1249	0.1268	0.508

Table 3.9: Additive hazard model for T with discrete covariate

β_t	β_u	β_w	Bias $\hat{\beta}_t$	SE	SEE	CP
0	0	0	-0.0095	0.1442	0.1478	0.957
		0.5	-0.0129	0.1526	0.1474	0.948
		1	-0.0105	0.1473	0.1494	0.965
	0.5	0	-0.0186	0.1479	0.1471	0.948
		0.5	-0.0192	0.1476	0.1475	0.952
		1	-0.0182	0.1497	0.1510	0.955
	1	0	-0.0219	0.1514	0.1516	0.957
		0.5	-0.0212	0.1546	0.1533	0.951
		1	-0.0340	0.1615	0.1601	0.946
0.5	0	0	-0.0789	0.1508	0.1522	0.915
		0.5	-0.0872	0.1502	0.1498	0.915
		1	-0.0840	0.1512	0.1507	0.913
	0.5	0	-0.0798	0.1474	0.1471	0.920
		0.5	-0.0934	0.1447	0.1453	0.900
		1	-0.0828	0.1456	0.1464	0.928
	1	0	-0.0871	0.1463	0.1481	0.924
		0.5	-0.0833	0.1468	0.1467	0.913
		1	-0.0962	0.1437	0.1495	0.919
1	0	0	-0.2665	0.1604	0.1614	0.599
		0.5	-0.2587	0.1577	0.1586	0.597
		1	-0.2757	0.1613	0.1572	0.578
	0.5	0	-0.2750	0.1538	0.1531	0.537
		0.5	-0.2751	0.1480	0.1489	0.525
		1	-0.2707	0.1609	0.1486	0.532
	1	0	-0.2808	0.1535	0.1508	0.516
		0.5	-0.2849	0.1465	0.1467	0.492
		1	-0.2938	0.1463	0.1461	0.460

Chapter 4

An Imputation Approach for Variable Selection of Interval-Censored Survival Data

4.1 Introduction

When conducting a survival analysis, it is often of interest to determine which of a set of variables are significantly related with the event time under study. For example, a physician might be interested in identifying a number of diagnostic tests or measurements that predict a patient's disease progression. This problem of variable selection has received a considerable amount of recent attention. Much of the motivation for research in this area is due to the increased availability of high-dimensional genetic data where researchers attempt to find significant genes that are related with patient survival. Currently, the vast majority of the variable selection work in the survival setting has focused on right-censored data. This chapter will take a first step

in extending some of the popular variable selection techniques to interval-censored survival data.

Variable selection in survival analysis draws heavily from the extensive research conducted for classic linear models. One can choose a model based on information criteria such as AIC (Akaike, 1974) and BIC (Schwarz *et al.*, 1978), or perform some kind of stepwise procedure. An outline of these approaches are described in Collett (2003). These methods, however, lack stability in terms of the variables selected (Breiman *et al.*, 1996). Also, special procedures need to be considered for increasingly high-dimension data where the number of covariates exceed the number of observations. A number of more advanced approaches exist to address these issues. Faraggi & Simon (1998) proposed a Bayesian variable selection technique for survival data. A random forest based approach for high-dimensional data was investigated by Ishwaran *et al.* (2010). Also, penalized likelihood procedures have recently grown in popularity. Methods such as LASSO (Tibshirani, 1996) and SCAD (Fan & Li, 2001) estimate regression coefficients and choose important variables simultaneously. This is accomplished by forcing the parameter estimates for insignificant covariates to be equal to zero. The selected variables are then those predictors that have nonzero coefficients. Many of these penalized likelihood procedures have been extended to handle right-censored data, such as the predominant LASSO (Tibshirani, 1997) and SCAD (Fan & Li, 2002), amongst others. Here, we will propose an imputation approach for variable selection of interval-censored survival data that enables the use of these popular penalized likelihood procedures.

Imputing interval-censored data has been successful in many different situations. It enables one to reduce the complex problem to the simpler and well-studied right-

censoring case where numerous tools are available. Bebchuk & Betensky (2000) use imputation to estimate the hazard function for interval-censored data. Pan (2000b) developed a two-sample test for comparing survival distributions using multiple imputation. Moreover, Pan (2000a) created an imputation method for regression analysis of interval-censored data. Our proposed approach for variable selection of interval-censored data involves imputing failure times for observations that are not right-censored, and then using one of the many available penalized likelihood procedures in order to identify the important covariates. This method is very straight-forward, and takes advantage of existing software, which makes it easy to implement.

4.2 Penalized Likelihood Review

In this section, we present a brief review of variable selection using penalized likelihood. As mentioned in the previous section, one of the reasons these procedures are attractive is because they estimate the coefficients and select variables simultaneously. This is accomplished by optimizing the likelihood subject to some penalty function. For example, minimize

$$-l_n(\beta) + \sum_{j=1}^{d} P_\lambda(\beta_j) \tag{4.1}$$

where l_n is the logarithm of the likelihood function, λ is an unknown regularization parameter, and $P(.)$ is the penalty function. The penalty function places a higher cost on more complicated models. Optimizing this penalized likelihood constrains some of the coefficients forcing them to zero, which enables variables selection.

A number of different penalized likelihood procedures have been studied in the literature. The various approaches differ in their choice of penalty function, which

results in methods with distinctive properties. Next, we will give a short overview of two of the leading penalized likelihood procedures most commonly seen in the literature.

4.2.1 Least Absolute Shrinkage and Selection Operator

One of the pioneers in this field is the least absolute shrinkage and selection operator, or LASSO, proposed by Tibshirani (1996). The LASSO uses an L_1 penalty, which constrains the sum of the absolute values of the regression parameters to be less than some value. Therefore, the LASSO uses the penalty:

$$P_\lambda(\beta) = \lambda|\beta_j| \tag{4.2}$$

This method was first extended to the survival setting for right-censored data under the proportional hazards model (Tibshirani, 1997). In this situation, the log-likelihood in equation 4.1 is replaced by the logarithm of the partial likelihood from the proportional hazards model. One benefit of using this penalty function is that it results in a concave optimization problem, which eases the computational burden. Also, the LASSO is available in the 'glmnet' R package.

The usefulness and popularity of the LASSO has generated a large number of derivative procedures for both classic linear models and survival analysis. Zou & Hastie (2005) proposed the elastic net, which is a combination of the LASSO and ridge regression. This enables a grouping effect for the covariates. Also, Zou (2006) introduced the adaptive LASSO, which allows the penalty parameter to change for each covariate.

4.2.2 Smoothly Clipped Absolute Deviation

One of the main alternatives to the LASSO in the smoothly clipped absolute deviation penalty, which is also known as SCAD (Fan & Li, 2001). This approach uses a penalty function of the following form:

$$P'_{\lambda}(\beta) = I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \tag{4.3}$$

for some $a > 2$.

SCAD was extended to right-censored survival data under the Cox model in Fan & Li (2002), and the SCAD procedure is available in R using the 'SIS' package.

The SCAD developers argue for their approach as one of the methods satisfying the following three properties of a good variable selection procedure:

1. Unbiasedness
2. Sparsity
3. Continuity

Moreover, the proponents of SCAD established that this procedure, unlike the original LASSO, satisfies an oracle property. Meaning, it performs as well as knowing the set of important variables ahead of time. This is a potential benefit over the LASSO. However, the SCAD procedure requires optimizing a non-convex penalty function. This creates additional challenges in practical situations, and can be a potential downside to this method.

4.3 Methodology

This section will describe our proposed methodology for variable selection of interval-censored survival data in detail. We will adopt the same notation used in previous chapters. Mainly, we have interval-censored data where the failure time T falls between two values $T \in (L, R]$ where $L \leq R$, and a vector of covariates Z. The main idea is quite simple: impute failure times for the finitely interval-censored observations, i.e. observations with $R < \infty$, and then use existing variable selection methods on the resulting imputed data that is a mix of exact imputed failure times and right-censored observations. Therefore, a simplified version of the algorithm can be broken down into three steps:

1. Impute failure times.

2. Use desired penalized likelihood method.

3. Select important variables.

At the first stage, one begins with a dataset containing the interval-censored failure times. For all the subjects who are not right-censored, an event time needs to be imputed. Here, we investigate two general ways to perform the imputation: single point imputation, and multiple imputation.

4.3.1 Single Point Imputation

The simplest method for imputing the failure times is using single point imputation, where one value from the interval $(L, R]$ is selected. Common choices for the imputed point include: selecting the midpoint of the interval, the left end point, the right end

point, or a random point within the interval. This decision could potentially be based on some prior knowledge that the true event times are close to one particular end of the interval. Here, we will consider three possible single point imputation approaches. Denote T^* as the imputed failure time. First, we examine left point imputation, where $T^* = L$. Second, we look at right point imputation, where $T^* = R$. Finally, we consider midpoint imputation, where $T^* = \frac{L+R}{2}$.

After imputation, the desired penalized likelihood procedure is used on the new data. We investigate the LASSO, SCAD, and elastic net. Variable selection is then performed as discussed in previous sections, where the significant variables are those with nonzero coefficient estimates.

The main advantage of single point imputation is how easy the approach is to implement. Also, if the observed intervals are narrow, then this approach can perform reasonably well. However, this method might not be reliable in all situations. This motivates an extension to multiple imputation.

4.3.2 Multiple Imputation

With multiple imputation, as the name suggests, several different values from $(L, R]$ are imputed. This creates m different imputed datasets. Each dataset is then analyzed separately, and the results are combined. The process is then repeated until the parameter estimates converge.

The multiple imputation procedure adopted here uses the following steps:

- Step 0 (Initialization): Set $\hat{\beta}^{(0)} = 0$. Produce m datasets by first generating m initial failure times for each individual from the uniform distribution $U(L_i, R_i)$. Then calculate Breslow's estimate for the baseline cumulative hazard function,

$\hat{\Lambda}_{0,k}^{(0)}$, for each of the datasets. Next, the baseline survival functions can be estimated by taking $\hat{S}_{0,k}^{(0)} = \exp(-\hat{\Lambda}_{0,k}^{(0)})$. Finally, combine the m estimates to get $\hat{S}_0^{(0)} = \sum_{k=1}^{m} \hat{S}_{0,k}^{(0)}/m$.

- Step 1: Assume the current estimates of the regression parameters and baseline survival functions are $\hat{\beta}^{(i)}$ and $\hat{S}_0^{(i)}$ respectively. Generate m sets of exact failure times for the interval-censored observations by sampling $T_{i,k}$ from the distribution $[\hat{S}_0^{(i)}]^{\exp(Z_i\hat{\beta}^{(i)})}$ conditional on $L_i < T_i \leq R_i$ for $k = 1, \ldots, m$.

- Step 2: For each imputed dataset estimate the regression parameters $\hat{\beta}_k^{(i)}$ using the desired method (LASSO, Elastic Net, and SCAD).

- Step 3: Estimate baseline survival $\hat{S}_{0,k}^{(i)}$ using Breslow's method and the values for T_i, k and $\hat{\beta}_k^{(i)}$ for $k = 1, \ldots, m$.

- Step 4: Update the estimates by combining the $\hat{\beta}_k^{(i)}$ and $\hat{S}_{0,k}^{(i)}$ to get $\hat{\beta}^{(i+1)}$ and $\hat{S}_0^{(i)}$.

- Step 5: Return to step 1 and repeat steps 1 through 5 until $\hat{\beta}^{(i)}$ converges.

Following the recommendation of Pan (2000a), we set the number of imputed datasets to create equal to 10, i.e. $m = 10$. Another important consideration is how the updated estimates are calculated in Step 4. Pan (2000a) uses the sample mean of the $\hat{\beta}_k^{(i)}$. However, one nonzero $\hat{\beta}_k^{(i)}$ outlier will cause the mean to be nonzero. This is a possible disadvantage of this approach, as it might inflate the model size and result in more false positives. Therefore, to update the estimates, we used both the mean and median of the $\hat{\beta}_k^{(i)}$.

4.4 Simulation Studies

A number of simulation studies were performed in order to study the effectiveness of the proposed imputation method for variable selection of interval-censored data. Survival times are generated using the proportional hazards model. That is, the hazard function is defined as:

$$\lambda(t|x) = \lambda_0(t)\exp(Z'\beta) \tag{4.4}$$

where $\lambda_0(t)$ is the baseline hazard function, Z is the covariate vector, and β is the vector of regression parameters. An exponential distribution with baseline hazard $\lambda_0(t) = 0.1$, and sample size of $n = 100$ is used. Similar to Tibshirani (1997), each Z_i marginally follows a standard normal distribution with the correlation between Z_i and Z_j given by $\rho^{|i-j|}$ with $\rho = 0.5$. Here, the true value for β is taken to be $\beta = (-0.8, 0, 0, -0.8, -0.8, 0, 0, 0, 0)'$.

Two different approaches for generating the censoring intervals are explored. For the first approach, the censoring intervals are created by first generating two independent uniform random variables $c_1 \sim U(0, a)$ and $c_2 \sim U(0, a)$, and then setting $L = T - c_1$ and $R = T + c_2$. Each dataset is analyzed using all three single point imputation methods, as well as the multiple imputation approach, in order to compare the various imputation procedures. Three values for a are used: $a = 0.4, 0.6$ and 0.8.

The second setup was designed to mimic what commonly occurs in medical studies with periodic follow-up. For each subject, there are $k+1$ examination times. The first examination time Y_i is generated from a $U(0, 1)$ random variable. Then, subsequent examination intervals are created by taking $(0, Y_i], (Y_i, Y_i + len], \ldots, (Y_i + k * len, \infty)$,

and $(L_i, R_i]$ are chosen to satisfy $L_i < T_i \leq R_i$. The values for k and len are selected as 3 and 5, respectively. Again, each dataset is analyzed using both single and multiple imputation for comparison purposes. Similar setups are used in Pan (2000a) and Sun *et al.* (2013), amongst others.

For each data setup, we utilized three competing penalized likelihood procedures: LASSO, SCAD, and elastic net. The values of λ for the LASSO and elastic net are chosen using 10-fold cross validation. The regularization parameter for SCAD is chosen by minimizing the generalized cross-validation statistic given in Fan & Li (2002). Also, as recommend in Fan & Li (2002), we fixed the value $a = 3.7$ when using SCAD.

For each simulation, performance is evaluated by calculating the mean model size of the 100 iterations. It is desirable for this value to be close to three since the only significant predictors are Z_1, Z_4, and Z_5. It is also important to ensure that the method is selecting the correct variables. Therefore, we also calculated the mean number of both the correctly and incorrectly identified nonzero coefficients as a measure of the true positives and false positives, respectively.

4.4.1 Simulation Study 1

We will first examine the results for the simulations using single imputation under the first data setup. Table 4.1 has these results for all three single point imputation procedures. The mean model sizes are all relatively close the the desired value of three. SCAD seems to have the mean model sizes closest to the true value overall. LASSO and elastic net are generally quite close in performance with mean models sizes around 3.5. Interestingly, LASSO and elastic net always correctly identify all three

significant variables, while SCAD occasionally misses one of the important covariates. Also, the mean number of false positives is low for all three methods. The results are consistent for left, midpoint, and right imputation. In some cases, there is a slight decrease in performance as the size of the interval increases, which is not a surprise. This effect is most notable in the results for the elastic net. Overall the results are quite good.

Next, table 4.2 has the results for multiple imputation under the first data setting. The first thing that stands out about these results are the differences between using the mean and median to combine the multiple datasets. The median universally outperforms the mean. When focusing on the results using the median, there does not appear to be a large difference between LASSO, SCAD, and elastic net. All three methods using the median are performing well with mean model sizes close to the optimal value of three. Also, the number of correctly identified significant variables is always equal to the actual value of three, and the number of false positives is low. As was the case with single imputation, there is a slight drop-off in performance as the length of the intervals gets larger.

4.4.2 Simulation Study 2

Results for the second data setup using single point imputation can found found in table 4.3. In general, the results are quite similar to those under setup 1. The mean model sizes are close to the true value for LASSO, SCAD, and elastic net using all three imputation strategies. Also, the number of false positives is reasonably low in all cases. There is one slight difference, as the true positive rate has slightly decreased when compared with the first setup. The number of correctly identified important

variables is still very good overall since the value never drops below 2.96.

A similar picture can be seen in the results for multiple imputation using this second data generation process, which can be found in table 4.4. Again, combining the parameter estimates using the median seems to outperform use of the mean. The mean model sizes are all very good, and there are a small number of false positives for all three penalized likelihood procedures when using the median. Also, like with single imputation, there is a very slight decrease in the number of correctly identified significant predictors. However, it does not appear to be serious as the values are still extremely close to the true value.

4.5 Discussion

In this chapter we proposed a general imputation approach for variable selection of interval-censored survival data that can take advantage of popular penalized likelihood procedures. Several different imputation techniques were investigated under a number of different simulation settings. Overall, the approach performed quite well in terms of average model size, true positives, and false positives. This is promising as there currently is a lack of available methods for variable selection of interval-censored data. Moreover, this research could easily be extending in a number of different directions. Some possibilities for future work in this area will be discussed in the next chapter.

There are several important points that should be considered when conducting an analysis using this method. First, it is essential that the covariates are all on the same scale, or standardized, prior to analysis. This ensures that the regularization

treats each variable fairly. Also, the choice of imputation procedure should be carefully determined. In general, if no prior information indicates the true failure times are close to one end of the interval, then a multiple imputation procedure is recommended. Finally, one should hesitate before using this method if the data is ultra high dimension because some type of screening procedure is recommended to reduce the size of the set of covariates, and such a method for interval-censored data does not yet exist.

Table 4.1: Single Imputation for Simulation 1

	Method	$a = 0.4$			$a = 0.6$			$a = 0.8$		
		Size	Correct	Incorrect	Size	Correct	Incorrect	Size	Correct	Incorrect
Left	LASSO	3.52	3.00	0.52	3.44	3.00	0.44	3.41	3.00	0.41
	SCAD	3.05	2.95	0.10	3.05	2.95	0.10	3.06	2.96	0.10
	Elastic Net	3.50	3.00	0.50	3.69	3.00	0.60	3.64	3.00	0.64
Mid	LASSO	3.46	3.00	0.46	3.44	3.00	0.44	3.46	3.00	0.46
	SCAD	3.04	2.94	0.10	3.05	2.95	0.10	3.04	2.95	0.09
	Elastic Net	3.50	3.00	0.50	3.66	3.00	0.66	3.68	3.00	0.68
Right	LASSO	3.38	3.00	0.38	3.45	3.00	0.45	3.45	3.00	0.45
	SCAD	3.05	2.94	0.11	3.05	2.94	0.11	3.05	2.94	0.11
	Elastic Net	3.44	3.00	0.44	3.63	3.00	0.63	3.66	3.00	0.66

Table 4.2: Multiple Imputation for Simulation 1

Method		$a = 0.4$			$a = 0.6$			$a = 0.8$		
		Size	Correct	Incorrect	Size	Correct	Incorrect	Size	Correct	Incorrect
LASSO	Mean	3.94	3.00	0.94	3.80	3.00	0.80	3.81	3.00	0.81
	Median	3.39	3.00	0.39	3.40	3.00	0.40	3.42	3.00	0.42
SCAD	Mean	3.39	3.00	0.39	3.42	3.00	0.42	3.45	3.00	0.45
	Median	3.30	3.00	0.30	3.33	3.00	0.33	3.33	3.00	0.33
Elastic Net	Mean	4.08	3.00	1.08	3.95	3.00	0.95	3.96	3.00	0.96
	Median	3.51	3.00	0.51	3.52	3.00	0.52	3.53	3.00	0.53

Table 4.3: Single Imputation for Simulation 2

	Method	Size	Correct	Incorrect
Left	LASSO	3.32	2.99	0.33
	SCAD	3.23	2.96	0.27
	Elastic Net	3.40	3.00	0.40
Mid	LASSO	3.31	2.99	0.32
	SCAD	3.24	2.96	0.28
	Elastic Net	3.38	3.00	0.38
Right	LASSO	3.29	2.97	0.32
	SCAD	3.27	2.97	0.30
	Elastic Net	3.36	2.99	0.37

Table 4.4: Multiple Imputation for Simulation 2

Method		Size	Correct	Incorrect
LASSO	Mean	3.76	3.00	0.76
	Median	3.34	2.99	0.35
SCAD	Mean	3.50	2.99	0.51
	Median	3.26	2.97	0.29
Elastic Net	Mean	3.89	3.00	0.89
	Median	3.37	3.00	0.37

Chapter 5

Future Research

A plethora of open questions still exist in regression analysis of interval-censored data. In this chapter, we will discuss a number of possible directions for future research based on the questions explored in the previous chapters.

5.1 Model Evaluation of Case I Data

For current status data, we examined an approach assuming that the failure times follow an additive hazards model and that the observation times follow a proportional hazards model. These are, of course, not the only options for model choices. It could potentially be interesting to investigate the performance of other models for informative censoring under the various conditions discussed in Chapter 2.

As we saw in Chapter 2, the method of Zhang *et al.* (2005) produced unbiased results under informative and noninformative censoring. However, this approach was not as efficient as the method proposed by Lin *et al.* (1998) when the censoring

is independent. Therefore, it is difficult to suggest an optimal, or default choice, when analyzing current status data. Developing a procedure to test if the censoring is informative or noninformative would be extremely helpful for determining model choice when faced with a real-world problem.

5.2 Model Evaluation of Case II Data

Similar to the case with current status data, we only investigated one particular model for flexibility in different censoring setups for case II interval-censored data. Consequently, more could be gained by performing comparable simulations using methods that assume a different model, such as the additive hazards model. Also, as mentioned in the previous section, it would be quite useful to develop a general testing procedure to determine if the censoring in case II interval-censored data is informative or noninformative.

5.3 Variable Selection of Interval-Censored Data

There are vast opportunities for future work with variable selection of interval-censored data since there is currently very little research in this area. The work in chapter 5 could be complimented with simulations evaluating performance under other models where penalized likelihood procedures exist for right-censored data, such as the additive hazards model. Also, an important future direction is to develop some sort of screening procedure when one is faced with ultra high dimensional data. Another interesting possibility is to directly extend penalized likelihood procedures, such as

LASSO and SCAD, to interval-censored data since this would no longer require using imputation.

Bibliography

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 716–723.

Andersen, Per Kragh, & Ronn, Birgitte B. 1995. A nonparametric test for comparing two samples where all observations are either left-or right-censored. *Biometrics*, 323–329.

Bebchuk, Judith D, & Betensky, Rebecca A. 2000. Multiple imputation for simple estimation of the hazard function based on interval censored data. *Statistics in medicine*, **19**(3), 405–419.

Betensky, Rebecca A. 2000. On nonidentifiability and noninformative censoring for current status data. *Biometrika*, **87**(1), 218–221.

Betensky, Rebecca A, & Finkelstein, Dianne M. 2002. Testing for dependence between failure time and visit compliance with interval-censored data. *Biometrics*, **58**(1), 58–63.

Betensky, Rebecca A, Rabinowitz, Daniel, & Tsiatis, Anastasios A. 2001. Com-

putationally simple accelerated failure time regression for interval censored data. *Biometrika*, **88**(3), 703–711.

Betensky, Rebecca A, Lindsey, Jane C, Ryan, Louise M, & Wand, MP. 2002. A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, **21**(2), 263–275.

Breiman, Leo, *et al.* 1996. Heuristics of instability and stabilization in model selection. *The annals of statistics*, **24**(6), 2350–2383.

Cai, Tianxi, & Betensky, Rebecca A. 2003. Hazard regression for interval-censored data with penalized spline. *Biometrics*, **59**(3), 570–579.

Chen, Ling, & Sun, Jianguo. 2009. A multiple imputation approach to the analysis of current status data with the additive hazards model. *Communications in StatisticsTheory and Methods*, **38**(7), 1009–1018.

Chen, Man-Hua, Tong, Xingwei, & Sun, Jianguo. 2009. A frailty model approach for regression analysis of multivariate current status data. *Statistics in medicine*, **28**(27), 3424–3436.

Collett, David. 2003. *Modelling survival data in medical research.* CRC press.

Cox, David R. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, **34**, 187–220.

Dewanji, Anup, & Kalbfleisch, JD. 1986. Nonparametric methods for survival/sacrifice experiments. *Biometrics*, 325–341.

Dinse, Gregg E. 1991. Constant risk differences in the analysis of animal tumorigenicity data. *Biometrics*, 681–700.

Fan, Jianqing, & Li, Runze. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**(456), 1348–1360.

Fan, Jianqing, & Li, Runze. 2002. Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, 74–99.

Faraggi, David, & Simon, Richard. 1998. Bayesian variable selection method for censored survival data. *Biometrics*, 1475–1485.

Finkelstein, Dianne M. 1986. A proportional hazards model for interval-censored failure time data. *Biometrics*, 845–854.

Finkelstein, Dianne M, Goggins, William B, & Schoenfeld, David A. 2002. Analysis of failure time data with dependent interval censoring. *Biometrics*, **58**(2), 298–304.

Ghosh, Debashis. 2003. Goodness-of-fit methods for additive-risk models in tumorigenicity experiments. *Biometrics*, **59**(3), 721–726.

Goggins, William B, Finkelstein, Dianne M, Schoenfeld, David A, & Zaslavsky, Alan M. 1998. A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics*, 1498–1507.

Groeneboom, Piet, & Wellner, Jon A. 1992. *Information bounds and nonparametric maximum likelihood estimation*. Vol. 19. Springer.

Huang, Jian. 1995. Maximum likelihood estimation for proportional odds regression model with current status data. *Analysis of Censored Data, IMS Lecture Notes-Monograph Series*, 129–145.

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

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

Huang, Jian, & Wellner, Jon A. 1997. Interval censored survival data: a review of recent progress. *Pages 123–169 of: Proceedings of the First Seattle Symposium in Biostatistics*. Springer.

Huang, Xuelin, & Wolfe, Robert A. 2002. A frailty model for informative censoring. *Biometrics*, **58**(3), 510–520.

Ishwaran, Hemant, Kogalur, Udaya B, Gorodeski, Eiran Z, Minn, Andy J, & Lauer, Michael S. 2010. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, **105**(489), 205–217.

Lagakos, Stephen W, & Louis, Thomas A. 1988. Use of tumour lethality to interpret tumorigenicity experiments lacking cause-of-death data. *Applied Statistics*, 169–179.

Lin, DY, Oakes, David, & Ying, Zhiliang. 1998. Additive hazards regression with current status data. *Biometrika*, **85**(2), 289–298.

Lindsey, Jane C, & Ryan, Louise M. 1993. A three-state multiplicative model for rodent tumorigenicity experiments. *Applied Statistics*, 283–300.

Martinussen, Torben, & Scheike, Thomas H. 2002. Efficient estimation in additive hazards regression with current status data. *Biometrika*, **89**(3), 649–658.

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

Pan, Wei. 2000b. A two-sample test with interval censored data via multiple imputation. *Statistics in Medicine*, **19**(1), 1–11.

Peto, Richard. 1973. Experimental survival curves for interval-censored data. *Applied Statistics*, 86–91.

Rabinowitz, Daniel, Tsiatis, Anastasios, & Aragon, Jorge. 1995. Regression with interval-censored data. *Biometrika*, **82**(3), 501–513.

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

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

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

Schwarz, Gideon, *et al.* 1978. Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

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

Shen, Xiaotong. 2000. Linear regression with current status data. *Journal of the American Statistical Association*, **95**(451), 842–852.

Sun, Jianguo. 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, Jianguo. 2006. *The statistical analysis of interval-censored failure time data*. Vol. 2. Springer.

Sun, Jianguo, & Shen, Junshan. 2009. Efficient estimation for the proportional hazards model with competing risks and current status data. *Canadian Journal of Statistics*, **37**(4), 592–606.

Sun, Jianguo, Feng, Yanqin, & Zhao, Hui. 2013. Simple estimation procedures for regression analysis of interval-censored failure time data under the proportional hazards model. *Lifetime data analysis*, 1–18.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, Robert. 1997. The lasso method for variable selection in the Cox model. *Statistics in medicine*, **16**(4), 385–395.

Turnbull, Bruce W. 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 290–295.

Wang, Lianming, Sun, Jianguo, & Tong, Xingwei. 2010. Regression analysis of case II interval-censored failure time data with the additive hazards model. *Statistica Sinica*, **20**(4), 1709.

Wang, Yan, & Taylor, Jeremy M G. 2001. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association*, **96**(455), 895–905.

Williams, JS, & Lagakos, SW. 1977. Models for censored survival analysis: constant-sum and variable-sum model. *Biometrika*, **64**(2), 215–224.

Xue, Hongqi, Lam, KF, & Li, Guoying. 2004. Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association*, **99**(466), 346–356.

Zeng, Donglin, Cai, Jianwen, & Shen, Yu. 2006. Semiparametric additive risks model for interval-censored data. *Statistica Sinica*, **16**(1), 287.

Zhang, Zhigang, Sun, Jianguo, & Sun, Liuquan. 2005. Statistical analysis of current status data with informative observation times. *Statistics in Medicine*, **24**(9), 1399–1407.

Zhang, Zhigang, Sun, Liuquan, Sun, Jianguo, & Finkelstein, Dianne M. 2007. Regression analysis of failure time data with informative interval censoring. *Statistics in Medicine*, **26**(12), 2533–2546.

Zhu, Liang, Tong, Xingwei, & Sun, Jianguo. 2008. A transformation approach for the analysis of interval-censored failure time data. *Lifetime data analysis*, **14**(2), 167–178.

Zou, Hui. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**(476), 1418–1429.

Zou, Hui, & Hastie, Trevor. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.

VITA

Tyler Cook was born in 1985 in Durham, North Carolina. He received his B.S. in mathematics and statistics from the University of Missouri in December of 2008. Then, in the fall of 2009, he began graduate studies in the statistics department at the University of Missouri. He will graduate with his PhD in statistics in the May of 2015.