UNIVERSITY OF CALIFORNIA
Santa Barbara

Middle censoring in the presence of covariates

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Statistics and Applied Probability

by

Elvynna Leong

Committee in Charge:

Professor S. Rao Jammalamadaka, Chair

Professor Yuedong Wang

Professor John Hsu

June 2014

The Dissertation of
Elvynna Leong is approved:

Professor Yuedong Wang

Professor John Hsu

Professor S. Rao Jammalamadaka, Committee Chairperson

June 2014

Middle censoring in the presence of covariates

by

Elvynna Leong

Acknowledgements

First of all, I would like to express my heartfelt gratitude to my advisor, Professor S. Rao Jammalamadaka. He patiently provided the vision, encouragement and advice necessary for me to proceed through the doctoral program and to complete my thesis. He has been a strong and inspirational advisor to me and I could not have imagined having a better advisor and mentor throughout my PhD journey.

I would also like to thank Dr Yuedong Wang and Dr John Hsu for their support, guidance, encouragement and constructive feedback. Their guidance has served me well and I owe them my heartfelt appreciation.

My friends in the PSTAT department, especially Michael Nava and Yitai Chiu, were sources of laughter, joy and support. Their friendship and assistance have meant more to me than I could ever express. It has been a great privilege to spend five years in the Department of Statistics and Applied Probability at University of California, Santa Barbara, and the memories will always remain dear to me.

I would also like to thank my father and my siblings for always being there for me. Their love was my driving force throughout my PhD journey. My husband, Brandon, whose love and continuous encouragement allowed me to finish this journey. There are no words that can express my gratitude and appreciation for all he has done for me. Thank you with all my heart and soul. Finally, I would like to dedicate this work to my mother, who left us too soon. I hope that this work makes you proud.

Curriculum Vitæ

Elvynna Leong

Education

2005	Bachelor of Science in Mathematics, Universiti Brunei Darussalam
2007	Masters of Science in Statistics, University of Manchester
2011	Master of Arts in Statistics, University of California, Santa Barbara
2014	Doctor of Philosophy in Statistics and Applied Probability, University of California, Santa Barbara (expected)

Experience

Dec, 07 - Sep, 09	Lecturer, Faculty of Science, Universiti Brunei Darussalam
Apr, 06 - Sep, 06	Tutor, Faculty of Science, Universiti Brunei Darussalam
Feb, 06 - Mar, 06	Statistics officer, Department of Economic Planning and Development, Brunei Darussalam

Awards

Full scholarship from the Brunei Government for PhD in Statistics at University of California, Santa Barbara, 2009-2014

Full scholarship from the Brunei Government for MSc in Statistics at University of Manchester, 2006-2007

Full scholarship from the Brunei Government for BSc in Mathematics at Universiti Brunei Darussalam, 2001-2005

Best student in BSc Mathematics, 2005

Abstract

Middle censoring in the presence of covariates

Elvynna Leong

Middle censoring refers to data that becomes unobservable if it falls within a random interval (L, R). For some individuals the exact values are available while for others the corresponding intervals of censorship are observed. Left censoring, right censoring and double censoring are special cases of this middle censoring by suitable choices of this censoring interval. Here, we develop new methods for analyzing data subject to middle-censoring when covariates are present. The techniques discussed include parametric models as well as semi-parametric models such as the Cox's Proportional Hazards model and the Accelerated Failure Times model.

In survival studies the values of some covariates may change over time. As such, it is natural to incorporate such time-dependent covariates into the model to be used in survival analysis. The model used in this research integrates both time-independent and time-dependent covariates for middle censored data. Both semiparametric and parametric models are considered when time-dependent covariates are present. Next, discrete lifetime data that follow a geometric distribution, that is subject to middle censoring is considered. Here, we include an extension and generalization to the case where covariates are present and present an alternate approach and proofs which exploit the simple relationship between the geometric and exponential distributions. Also, considered are estimation problems for middle censored data with two independent competing risks, both for parametric and semiparametric context.

Simulation studies are performed to demonstrate the usefulness and accuracy of the methods developed here, and illustrated with a practical example using data from a Stanford Heart Transplant study.

Professor S. Rao Jammalamadaka
Dissertation Committee Chair

Contents

List of Figures

List of Tables

Chapter 1

Survival Analysis

1.1 Introduction

Survival data often consists of a response variable that measures the duration of time until a specified event occurs and a set of independent variables thought to be associated with the event-time variable. This event-time may be death, durations of jobs, conception, the development of some disease, remission after some treatment, survival times in a clinical trial, the appearance of a tumor and equipment breakdown.

The purpose of survival analysis is to model the underlying distribution of event times and to assess the dependence of the event time on other explanatory variables. In many situations, the event time is not observed due to withdrawal or termination of the study; this phenomenon is known as censoring. Survival analysis methods correctly use both the censored and uncensored observations. Let T denote a non-negative random variable representing the failure time of a subject, that is, the survival variable of interest. For inferences about T, the survival function and the hazard function are particularly useful for modeling.

The survival function of T is defined as the probability that T is greater than a certain time, t and is of considerable interest in failure time analysis. Let $S(t)$ denote the survival function of T. Then,

$$S(t) = P(T > t), 0 < t < \infty$$

Assuming that T is continuous and thus its probability density function $f(t)$ exists;

$$S(t) = 1 - F(t) = \textstyle\int_t^\infty f(x)\,\mathrm{d}x.$$

where $F(\cdot)$ is called the Cumulative Distribution Function (CDF). Thus

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

They are monotone, non-increasing functions equal to one at zero and zero as time approaches infinity.

Assume that T is a discrete survival variable taking values $t_1 < t_2 < \ldots$ with probability function$f(t_j) = P(T = t_j); j = 1, 2, \ldots$. Then, the survival function for a discrete random variable T is given by

$$S(t) = P(T > t) = \textstyle\sum_{j:t_j>t} f(t_j)$$

Another quantity of interest is the hazard function of T at time t, and is defined by

$$h(t) = lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t} \tag{1.1.1}$$

It represents the instantaneous probability that a subject fails at time t given that the subject has not failed before t. Assuming that T is a continuous random variable, then Equation (1.1.1) can be rewritten as

$$h(t) = \tfrac{1}{P(T>t)} lim_{\Delta t \to 0} \tfrac{P[t \leq T < t + \Delta t]}{\Delta t} = \tfrac{f(t)}{S(t)} = -\tfrac{dln[S(t)]}{dt}$$

A related quantity is the cumulative hazard function $H(t)$ defined by

$$H(t) = \int_0^t h(u)du = -ln[S(t)]$$

Thus, for continuous lifetimes,

$$S(t) = exp[-H(t)] = exp[-\int_0^t h(u)du]$$

This function is particularly useful in determining the appropriate failure distribution utilizing qualitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time. The restriction on $h(t)$ is that it is non-negative, i.e. $h(t) \geq 0$.

When T is a discrete random variable, the hazard function is given by

$$h(t_j) = P(T = t_j | T \geq t_j) = \frac{p(t_j)}{S(t_{j-1})} \tag{1.1.2}$$

where $j = 1, 2, \cdots$ and $S(t_0) = 1$. Since the relationship between the lifetime model and the survival function is given as

$$p(t_j) = S(t_{j-1}) - S(t_j)$$

then the hazard function in Equation (1.1.2) becomes

$$h(t_j) = P(T = t_j | T \geq t_j) = \frac{p(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S_(t_{j-1})},$$

where $j = 1, 2, \cdots$.

Note that the survival function may be written as the product of conditional survival probabilities

$$S(t) = \prod_{t_j \leq t} \frac{S(t_j)}{S(t_{j-1})} = \prod_{t_j \leq t} [1 - h(t_j)]$$

1.2 Censoring and Truncation

Censoring is one of the unique features of failure time data. By censoring, it means that an observation on a survival time of interest is incomplete, that is, the survival time is observed only to fall into a certain range instead of being known exactly. The various categories of censoring are right censoring, left censoring and interval censoring. Censored data are different from missing data as censored observations still provide some partial information, whereas missing observations provide no information about the variable of interest (Sun, 2006).

Truncation of survival data occurs when only those individuals whose event time lies within a certain observational window (Y_L, Y_R) are observed. An individual whose event time is not in this interval is not observed and no information on this subject is available to the investigator. This is in contrast to censoring where there is at least partial information on each subject. Because we are only aware of individuals with event times in the observational window, the inference for truncated data is restricted to conditional estimation.

The main difference between censoring and truncation is that in truncation, survival data occurs when only those individuals whose event time lies within a certain observational window are observed. Whereas in censoring there is at least partial information about the lifetimes, regardless if the event happens in the observational window or not.

1.2.1 Right Censoring

By right censoring, the failure time of interest is observed either exactly or to be greater than a censoring time. A typical situation that yields right-censored observation is one in which survival study has to end due to, for example, time constraints or resource limitations. In this case, for subjects whose survival events have not occurred at the end of the study, their survival times are not observed exactly but are known to be greater than the study end time i.e. they are right-censored. For subjects who have already failed by the end of the study, their failure times are known exactly. Of course, the study end time could be different for different subjects, and some subjects may withdraw from the study before the end for some reasons. Generally, there are two types of censoring: 1. Type 1 censoring -where censoring occurs at a pre-specified time and 2. Type II censoring - where censoring occurs once a predetermined number of deaths occur (Klein & Moeschberger, 2003).

Type I censoring occurs when the event is observed only if it occurs prior to some pre-specified time. For example, a typical animal study or clinical trial starts with a fixed number of animals or patients to which a treatment is applied. Due to time or cost considerations, the investigator will need to terminate the study or report the results before all subjects realize their events. In this case, all censored observations have times equal to the length of the study period if there are no losses or subject withdrawals.

Assume that there is a lifetime X and a fixed censoring time, C_r where the X's are assumed to be independent and identically distributed. The exact lifetime X of an individual will be known if and only if X is less than or equal to C_r i.e.$X \leq C_r$. However, if X is greater than C_r, the individual is a survivor and so his or her event time is censored at C_r. The data from this experiment can be represented by pairs of random variables (T, δ), where δ indicates whether the lifetime X corresponds to an event $(\delta = 1)$ or is censored $(\delta = 0)$, and T is equal to X, if the lifetime is observed, and to C_r if it is censored, i.e. $T = min(X, C_r)$ (Klein & Moeschberger, 2003). An example would be when animals have different, fixed - sacrifice (censoring) times, this form of Type I censoring is called *progressive Type I censoring*.

Type II censoring is another type of right censoring in which the study continues until the failure of the first r individuals, where r is some predetermined integer $(r < n)$. An example of Type II censoring is when testing equipment life where all items are put on test at the same time, and the test is stopped when r of the n items have failed. An advantage of this type of censoring is that it may save time and money since it could take a long time for all the items to fail. In this case, note that r is the number of failures and $n - r$ the number of censored observations are fixed integers and the censoring time $T_{(r)}$, the rth ordered lifetime is random.

An example is shown in Figure (1.2.1) to illustrate right censoring. The graph describes the experience of several subjects followed over a certain period until death or the end of the study. Some of these subjects may be lost to follow up during the study period, withdraw from the study or still alive at the end of the study . The subjects who died are denoted by X. The figure shows that subjects A and E died before the end of the study so they are uncensored observations. Subject B is lost to follow up while subject D withdraws from the study for some known/unknown reasons. Subjects C and F were alive at the end of the study. Hence, subjects B, C, D and F are considered as right-censored observations.

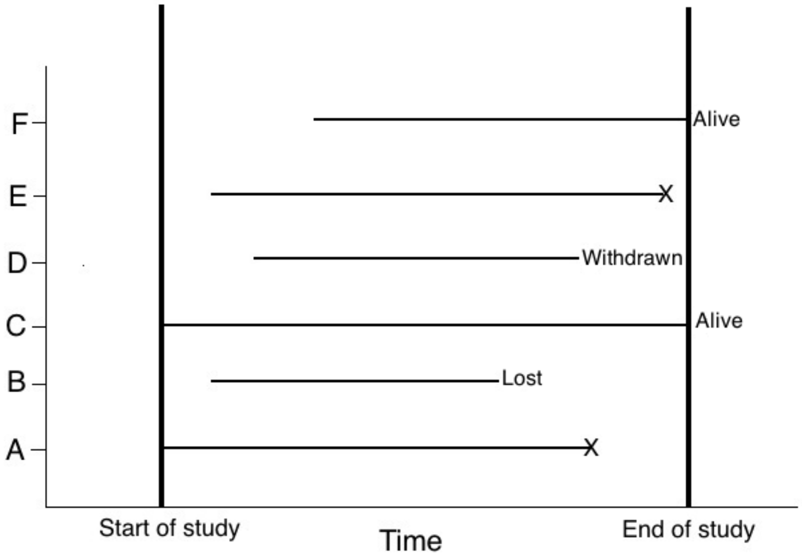


Figure 1.2.1: Study time for six subjects with different status

1.2.2 Left Censoring

Left censoring occurs when a person's true survival time is less than or equal to that person's survival time, that is, the event of interest has already occurred for the individual before that person is observed in the study at censoring time C_l. For such individuals, they have experienced the event before time C_l but their exact event time is unknown. Left censored data can be represented by pairs of random variables (T, η), where T is equal to X if the lifetime is observed and η indicates whether the exact lifetime X is observed $(\eta = 1)$ or not $(\eta = 0)$. Lifetimes are considered **doubly censored** if both left censoring and right censoring occur in a study (Turnbull, 1974).

1.2.3 Interval censoring

Another type of censoring occurs when the lifetime is known only to lie in an interval, instead of being observed exactly. By interval censoring, the study subjects or failure time processes of interest are not under continuous observation. As a consequence, the failure or survival time is not always exactly observed or right-censored. For an interval-censored observation, one only knows of a window, that is, an interval, within which the survival event has occurred. Traditionally the the term "interval censoring" has been used when all the data comes in the form of intervals (cf. Sun, 2006).

Interval-censored failure time data occur in many areas including epidemiological, demographical, financial, sociological and medical studies. An example of interval-censored data occurs in health or medical studies that involve periodic follow-ups. Such interval censoring occurs when patients in a clinical trial or longitudinal study have periodic follow up and the patient's event time is only known to fall in an interval $(L_i, R_i]$.

Let X be a non-negative random variable representing the failure time of an individual in a failure time study. An observation on X is interval-censored if instead of observing X

exactly, only an interval $(L, R]$ is observed such that $X \in (L, R]$ where $P(L \leq R) = 1$ (Sun, 2006).

1.2.4 Likelihood construction for censored data

In constructing a likelihood function for censored data, careful consideration needs to be done on what information each observation gives us. A critical assumption in most studies, is that the lifetimes and censoring times are independent. For an independent and identically distributed sample,

$$T_1, T_2, \cdots, T_n \overset{iid}{\sim} f(t|\theta)$$

with censoring intervals (L_i, R_i). Then the most general likelihood can be written in the form

$$L \propto \prod_{i \in Uncens} f(t_i) \prod_{i \in LC} F(L_i) \prod_{i \in RC} [1 - F(R_i)] \prod_{i \in IC} [F(R_i) - F(L_i)]$$

where $Uncens$ is the set of failure times, LC the set of left-censored observations, RC the set of right-censored observations and IC the set of interval-censored observations. The goal is to try to find the Maximum Likelihood Estimate (MLE) of this equation. However, in most cases there is not an explicit solution hence a numerical approach must be used. Newton-Raphson method is one of the most common approaches used to maximize this.

1.2.5 Left truncation

Truncation of survival data occurs when only those individuals whose event time lies within a certain observational window (Y_L, Y_R) are observed. Left truncation occurs when

Y_R is infinite. Only individuals whose event time T exceeds the truncation time Y_L are observed; that is, T is observed if and only if $Y_L < T$. Left truncation is very common in fields like demography and epidemiology. An example is a study on how long people who have been hospitalized for a heart attack survive taking some treatment at home. The start time is taken to be the time of the heart attack and only those individuals who survive their stay in hospital are able to be included in the study. The truncation time is often called a *delayed entry time* since we only observe subjects from this time until they die or are censored (Klein and Moeschberger, 2005).

1.2.6 Right truncation

Right truncation occurs only when Y_L is equal to zero i.e. the survival time T is observed only when $T \leq Y_R$. For example, right truncation occurs when estimating the distribution of stars from the earth and the stars too far away that are not visible are right truncated.

1.3 Estimating the survival function and the cumulative hazard function

1.3.1 Kaplan-Meier Estimator

The standard non-parametric estimator of the survival function is the Kaplan-Meier estimator introduced by Kaplan and Meier (1958). It is a product-limit estimator and has been extensively studied in the literature. This estimator is defined as follows

$$\hat{S}(t) = \prod_{t_j \leq t}(1 - \frac{d_j}{Y_j})$$

Here t_j denotes the distinct imputed exact failure time, Y_j is the number of individuals who are at risk at time t_j and d_j is the number of failures at time t_j. This estimator is not well defined for values of t beyond the largest observation time. If this corresponds to a failure time, then the estimated survival function is zero beyond this point. If the largest observed time is right censored, the estimator $S(t)$ in undetermined beyond this point. However, Efron (1967) suggests estimating $\hat{S}(t) = 0$ for $t > t_{max}$. This corresponds to assuming that the survivor with the largest time on study would have died immediately after the survivor's censoring time and leads to an estimator which is negatively biased. Other estimates for $\hat{S}(t)$ have been explored as well, as suggested in Gill (1980).

The product-limit estimator, $S(t)$ is a step function with jumps and discontinous at time t_j's. Its pointwise variance estimate is given by the well-known Greenwood's formula (Greenwood, 1926)

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{Y_j(Y_j - d_j)}$$

An alternative estimator of the variance of $\hat{S}(t)$ due to Aalen and Johansen (1978) is given by

$$\tilde{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{Y_j^2}$$

For small to moderate samples, both this estimator and Greenwood's estimator tend to underestimate the true variance of the Kaplan-Meier estimator. On average, Greenwood's estimator tends to come closest to the true variance and has a smaller variance except when Y_j is very small (Klein, 1991).

1.3.2 Nelson-Aalen estimator

The Kaplan-Meier estimator provides an efficient means of estimating the survival function for right censored data. It can also be used to estimated the cumulative hazard function by using the relationship $H(t) = -lnS(t)$. Nelson (1972) suggested an alternate estimator of the cumulative hazard rate, which has better small-sample size performance than the estimator based on the Kaplan-Meier estimator. Aalen (1978b) then rediscovered the estimator who derived the estimator using modern counting estimator, and this estimator is called Nelson-Aalen estimator of the cumulative hazard function i.e.

$$\tilde{H}(t) = \begin{cases} 0 & \text{if } t \leq t_1 \\ \sum_{t_i \leq t} \frac{d_i}{Y_i} & \text{if } t_1 \leq t \end{cases}$$

The estimated variance of the Nelson-Aalen estimator is given by

$$\sigma_H^2(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i^2}$$

Hence, based on the Nelson-Aalen Estimator of the cumulative hazard rate, an alternate estimator of the survival function is given by the equation $\tilde{S}(t) = exp(-\tilde{H}(t))$.

The Nelson-Aalen estimator has two primary uses in analyzing data. The first use is in selecting between parametric models for the time to event. An example is a plot of $\tilde{H}(t)$ against t will be approximately linear if the exponential distribution with hazard rate λ, fits the data well. A second use is in providing the crude estimates of the hazard rate $h(t)$ which is the slope of the Nelson-Aalen estimator.

1.4 Models with covariates

Consider a failure time $X > 0$ and a vector $\mathbf{Z^T} = (Z_1, \cdots, Z_p)$ of explanatory variables associated with failure time X. The explanatory variables $\mathbf{Z^T}$ could either be quantitative variables (such as blood pressure, temperature, age and weight), qualitative variables (such as gender, race, treatment and disease status) and/or time-dependent variables i.e. $\mathbf{Z^T(x)} = [Z_1(x), Z_2(x), \cdots, Z_p(x)]$ which will be discussed in Section 5.

1.4.1 Cox Proportional Hazard function

The Cox Proportional Hazard model (PH) is defined by

$$h(t, \mathbf{Z}) = h_0(t)e^{\sum_{i=1}^{p} \beta_i Z_i} \tag{1.4.1}$$

where $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_p)$ is the explanatory variables and $h_0(t)$ is the baseline hazard function. This model gives an expression for the hazard at time t for an individual with a given specification of a set of explanatory variables $\mathbf{Z}$. An important property of the Cox model is that the baseline hazard, $h_0(t)$ is an unspecified function. It is this property that makes the Cox model a semi-parametric model.

A key reason for the popularity of the Cox model is that, even though the baseline hazard is not specified, reasonably good estimates of regression coefficients, hazard ratio of interest, and adjustable survival curves can be obtained for a wide variety of data situations. Cox PH model is a robust model such that the results from using this model will closely approximate the results for the correct parametric model. Thus, when in doubt, the Cox model will give a reliable enough results so that it is a 'safe' choice of model, and the user does not need to worry about whether the wrong parametric model is chosen.

The goal of this model is not to estimate the baseline hazard function $h_0(t)$ but rather to estimate the effect of the covariates, **Z** on lifetimes. The estimates of the parameters of the Cox model are the maximum likelihood estimates, $\hat{\beta}_i$ and is based on a partial or conditional likelihood rather than a full likelihood approach. Assume that censoring is non-informative in that, given $\mathbf{Z}_j$, the event and censoring time for the jth subject are independent. Also, suppose that there are no ties between the event times. Let $t_1, t_2, \cdots t_D$ denote the ordered event times and $Z_{(i)k}$ be the kth covariate associated with the individual whose failure time is t_i. The risk set at time t_i, $R(t_i)$ is the set of all individuals who are still under study at a time just prior to t_i. The partial likelihood based on the hazard function as specified by (1.4.1) is expressed by

$$L(\beta) = \prod_{i=1}^{D} \frac{exp[\sum_{k=1}^{p} \beta_k Z_{(i)k}]}{\sum_{j \in R(t_i)} exp[\sum_{k=1}^{p} \beta_k Z_{jk}]} \tag{1.4.2}$$

Note that the baseline hazard function does not play a role in the Cox likelihood, thus it plays no role in the estimation of the regression parameters. The term "partial" likelihood is used because the likelihood formula considers probabilities only for those subjects who fail, and does not explicitly consider probabilities for those subjects who are censored. This is treated as a usual likelihood and inference is carried out by usual means. It is of interest to note that the numerator of the likelihood depends only on information from the individual who experiences the event, whereas the denominator utilized information about all individuals who have not yet experienced the event (including the subjects who will be censored later). The (partial) maximum likelihood estimates are found by maximizing (1.4.1) by taking partial derivatives of log of L with respect to each parameter in the model and then solving a system of equations. This solution is carried out using Newton-Raphson iteration.

The PH assumption requires that the Hazard Ratio, HR is constant over time, or equivalently, that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time. Consider HR that compares two different specifications $\mathbf{Z}^*$ and $\mathbf{Z}$ for the explanatory variables used in the Cox model i.e.

$$\hat{HR} = \frac{\hat{h}(t, \mathbf{Z}^*)}{\hat{h}(t, \mathbf{Z})} = e^{\sum_{i=1}^{p} \hat{\beta}_i (Z_i^* - Z_i)}$$

where $\mathbf{Z}^* = (Z_1^*, Z_2^*, \cdots, Z_p^*)$ and $\mathbf{Z} = (Z_1, Z_2, \cdots, Z_p)$ denote the set of $Z's$ for two individuals. Notice that the final expression does not involve the baseline hazard and time t.

1.4.2 Accelerated Failure Times

Many parametric models are Accelerated Failure Time (AFT) models rather than Proportional Hazard (PH) models. The AFT model specifies that

$$\log T = Z^T \beta + W \tag{1.4.3}$$

where β is the vector of regression parameters and W is an error variable with an unknown distribution function. Under model (1.4.3), the underlying assumption for AFT models is that the effect of covariates is multiplicative with respect to survival time. It describes the effect to change the timescale and therefore to accelerate or decelerate the time to failure. This model has been studied by various authors including Miller (1976), Buckley and James

(1979), Koul et al. (1981), Louis (1981), Wei and Gail (1983), James and Smith (1984), Ritov and Wellner (1988), Lai and Ying (1991b), Wei et al. (1990), Tsiatis (1990) and Ritov (1990).

The survival function of the AFT model is given by

$$S(t|\mathbf{Z}) = S_0(exp(\beta^T\mathbf{Z})t)$$

where $exp(\beta^T\mathbf{Z})$ is called the acceleration factor. The acceleration factor is the key measure of association obtained in an AFT model. It allows the investigator to evaluate the effect of predictor variables on survival time just as the hazard ratio allows the evaluation of predictor variables on the hazard.

1.4.3 Time-dependent covariates

In the previous sections, the hazard function of an individual is modeled as a function of fixed time covariates. These are explanatory variables recorded at the start of the study whose values are fixed throughout the course of the study. An example from Klein and Moeschberger (2003, page 295) is from acute leukemia patients who were given a bone marrow transplant where there are three risk groups, donor age, recipient age and several other variables, as fixed time covariates. The basic interest was to evaluate the relationship of the risk groups to the hazard of relapse or death controlling for possible confounding variables which might be related to relapse or death. As is typical in survival studies, individuals are monitored during the study and other explanatory variables are recorded where values may change during the course of the study. Some of these variables may be instrumental in predicting survival and need to be taken into consideration in evaluating the survival distribution. Such variables may change over time are called time-dependent variables. A covariate may be binary with at most one change, depending on the conditions during the study time (Allison, 1995). It is also possible to include time-dependent covariates that are essentially continuous where the value of the covariate is a series of measurements of some explanatory characteristic. Some examples of this type of covariate might be blood pressure, cholesterol, body mass index, size of the tumor, or rate of change in the size of the tumor recorded at different times for a patient.

A common use of time-dependent covariates is for testing the proportional hazards assumption in Section (1.4.1). For time dependent covariates, it is assumed that their value is predictable in the sense that the value of the covariate is known at an instant just prior to time t. The basic model due to Cox (1972) is as in (1.4.1) with $\mathbf{Z}$ replaced by $\mathbf{Z(t)}$,

$$h(t|Z(t)) = h_0(t)e^{\sum_{i=1}^{p}\beta_i Z_i(t)} \tag{1.4.4}$$

where $\mathbf{Z}(t) = [Z_1(t), Z_2(t), \cdots, Z_p(t)]^T$ denote a set of covariates at time t which may effect the survival distributions of X where X denote the time to some event. An assumption of the model 1.4.4 is that the time dependent covariate effect, as measured by its coefficient, does not depend on time.

From a conceptual point of view, the model (1.4.4) becomes more complicated and one should give serious consideration to the nature of any time dependent covariate before including it in the model. Another concern is the potential to over-fit a model when using time dependent covariates. Inclusion of time-dependent covariates should be based on strong clinical evidence. The consequences of using the Cox's proportional hazards model when the hazard ratios are far from constant over time are: (1) The power of corresponding tests decreases because of suboptimal weights for combining the information provided by the risk sets of times where failure occur (Lagakos & Schoenfeld, 1984) and for other covariates with constant hazard ratios, testing power declines as a consequence of an inferior fit of the model. (2) The relative risk for covariates with hazard ratios increasing over time is overestimated while for covariates with converging hazards, perhaps the most frequent violation, the relative risk is underestimated. Fisher and Lin (1999) provide illustrations of the use of time dependent covariates and discuss related conceptual issues and potential problems and biases. Andersen, Borgan, Gill and Keiding (1993), Fleming and Harrington (1991) and Kalbfleisch and Prentice (2002) present the topic from the counting process point of view.

Prediction in survival models with time-dependent covariates would be difficult because of the changing nature of the covariate with time hence its value at different future times will be unknown. Also, the survival curves cannot be estimated in this case. This is because estimating the survival curve would require the value of the time-dependent covariate for the subject. In this case, knowing this value means that this subject has not been observed or may still be alive or is in the risk set. Hence, the survival time for this subject cannot be used to estimate the survivor function (Fisher and Lin, 1999).

Chapter 2

Middle censoring

Jammalamadaka and Mangalam (2003) introduced a general concept of censoring called “Middle Censoring”. Middle censoring refers to data where some of the observations are observed exactly while others become unobservable as and when they fall within a random interval (L, R). Left censoring, right censoring and double censoring are special cases of this middle censoring by suitable choice of this censoring interval.

For some individuals the exact values are available while for others the corresponding intervals of censorship are observed. If a subject is temporarily absent or withdrawn from study and the event of interest occurs during this time interval, the exact time of occurrence can not be observed but instead we only observe an open interval. Other authors have referred to this general scheme as “partly interval censored”, “mixed interval censored” etc. (see Huang, 1999 and Yu et al, 2001)

Suppose there is a random sample of individuals of size n from a specific population whose true life times are $T_1, T_2, \cdots, T_n$. It is assumed that the lifetimes are a random sample from a common distribution, that is

$$T_1, T_2, \cdots, T_n \overset{iid}{\sim} F_0(t)$$

Corresponding to each individual in the sample there is a random censoring interval $(L_1, R_1), (L_2, R_2), \cdots, (L_n, R_n)$ which are independent of the lifetimes. The censoring intervals are taken to be i.i.d bivariate random vectors

$$(L_1, R_1), (L_2, R_2), \cdots, (L_n, R_n) \overset{iid}{\sim} G(l, r)$$

where $P(L_i < R_i) = 1$. This is a standard assumption which is equivalent to saying that the time at which an individual is censored has nothing to do with how long that individual lives. Thus, for the ith individual T_i is observable only if $T_i \notin [L_i, R_i]$. Otherwise, only the censoring interval $[L_i, R_i]$ can be observed corresponding to this individual instead of real observation. Thus the observed data is $X_1, X_2, \cdots, X_n$, where

$$X_i = \begin{cases} T_i & \text{if } T_i \notin (L_i, R_i) \\ (L_i, R_i) & \text{if } T_i \in (L_i, R_i) \end{cases}$$

2.1 Nonparametric Middle-Censoring

In many censoring situations, to estimate the distribution function via the EM algorithm, the resulting algorithm takes the form

$$\hat{F}(t) = E_{\hat{F}}[E_n(t)|\mathbf{X}]$$

as described by Tsai and Crowley (1985), where E_n is the empirical distribution function. This equation was first introduced and referred to as self-consistency equation by Efron (1967).

In the middle censored cases, the self-consistent estimator (SCE) (see Jammalamadaka and Mangalam, 2003), $\hat{F}_n$ satisfies the following equation:

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i I_{[T_i \le t]} + (1-\delta_i) I_{[R_i \le t]} + (1-\delta_i) I_{[t \in (L_i, R_i)]} \frac{\hat{F}_n(t) - \hat{F}_n(L_i)}{\hat{F}_n(R_i-) - \hat{F}_n(L_i)}\right) \tag{2.1.1}$$

where $\delta_i = I\{T_i \notin (L_i, R_i)\}$

An exact solution to equation (2.1.1) does not exist but can be solved via the EM algorithm. It can be computed by the iterative formula

$$\hat{F}^{(m+1)}(t) = E_{\hat{F}^{(m)}}[E_n(t)|\mathbf{X}]$$

where $E_n(t)$ is the empirical distribution function.

Jammalamadaka and Mangalam (2003) showed that the NPLME satisfies the self-consistency equation, SCE (2.1.1) and is listed below.

Theorem 2.1. *The NPMLE satisfies equation (2.1.1).*

A question of interest is whether or not the NPMLE will have all its mass on the uncensored observation. The authors prove the following proposition in answer to this question.

Proposition 2.2. *If each observed censored interval (L_i, R_i) contains at least one uncensored observation $X_j, j \neq i$, then any distribution function that satisfies equation (2.1.1) attaches all its mass on the uncensored observations.*

Jammalamadaka and Mangalam (2003) also consider the case when the censoring interval, (L_i, R_i), does not contain an uncensored observation. They suggest assigning the mass corresponding to that interval to its midpoint when this happens. Hence, the initial estimator may give equal mass to all uncensored observations and to the midpoints of those finite censored intervals that contain no uncensored observations. If an infinite censoring interval happens to be empty of uncensored observations, one can then assign the mass to any arbitrary point inside this interval for the estimator to have a maximum.

The following theorem by Jammalamadaka and Mangalam (2003) answers questions about consistency of the SCE under some mild conditions.

Theorem 2.3. *The self consistency equation, SCE is uniformly strongly consistent.*

See Jammalamadaka and Mangalam (2003) for a thorough proof of the above statements. Jammalamadaka and Iyer (2004) considered a slight variation of this estimator and proved that it is consistent and converges weakly to a Gaussian process.

2.2 Parametric Middle-Censoring

Middle censoring in the parametric context was first discussed by Iyer, Jammalamadaka and Kundu (2008) (IJK from now on). Lifetimes, in the parametric context, are assumed to be an i.i.d sample from a known distribution and the censoring intervals are also i.i.d bivariate random vectors from another known distribution and are taken to be independent of the lifetimes; a common assumption in survival analysis. See Kaplan and Meier (1958), Turnbull (1974), Jammalamadaka and Mangalam (2003) and IJK (2008).

2.2.1 Exponential distribution

In IJK (2008), the lifetimes T_i are exponentially distributed with mean $\frac{1}{\theta_0}$. The left point of the censored interval, L_i is an Exponential random variable with mean $\frac{1}{\alpha}$ and the length of the censored interval, U_i is Exponentially distributed with mean $\frac{1}{\beta}$ i.e.

$$\begin{aligned} T_1, T_2, \cdots, T_n &\overset{iid}{\sim} f(t|\theta_0) = \theta_0 e^{-\theta_0 t} \\ L_1, L_2, \cdots, L_n &\overset{iid}{\sim} f(l|\alpha) = \alpha e^{-\alpha l} \\ U_1, U_2, \cdots, U_n &\overset{iid}{\sim} f(u|\beta) = \beta e^{-\beta u} \end{aligned}$$

for $t > 0, l > 0, u > 0$. Moreover, T_i's, L_i's and U_i's are all independent of each other. The observed data X_i's, just as in the nonparametric set-up in Section (2.1), are given as

$$X_i = \begin{cases} T_i & \text{if } T_i \notin (L_i, R_i) \\ \\ (L_i, R_i) & \text{if } T_i \in (L_i, R_i) \end{cases}$$

Since the model is parametrized, the MLE of θ_0 can then be solved. Assuming that there are $n_1 > 0$ uncensored observations and $n_2 > 0$ censored observations, without loss of generality, re-order the data into uncensored and censored observations. The observed data is

$$T_1, \cdots, T_{n_1}, (L_{n_1+1, R_{n_1+1}}), \cdots, (L_{n_1+n_2, R_{n_1+n_2}})$$

where $n_1 + n_2 = n$. The likelihood can then be written as

$$L(\theta) = c\theta^{n_1} e^{-\theta \sum_{i=1}^{n_1} t_i} \prod_{i=n_1+1}^{n_1+n_2} (e^{-\theta l_i} - e^{-\theta r_i}) \tag{2.2.1}$$

where c is a normalizing constant depending on α and β. However, the estimation of α and β are not of interest, thus they are left as a constant. The log-likelihood can be written as

$$l(\theta) = logc + n_1 ln\theta - \theta \sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} ln(e^{-\theta l_i} - e^{-\theta r_i}) \tag{2.2.2}$$

The EM algorithm is used to find the MLE of θ_0 since this is an example of incomplete data. The algorithm is an iterative procedure that finds MLE's in parametric estimation for incomplete data by repeating the following steps:

1. E-step: Calculates the conditional expectation of the complete data log-likelihood given the observed data and the parameter estimates

2. M-step: Given a complete data log-likelihood, the M step finds the parameter estimates to maximize the complete data log-likelihood from the E-step.

The two steps are iterated until the iteration coverage.

Applying integration by parts gives the following equation

$$E[T|L < T < R] = \frac{e^{-\theta L}(L + \frac{1}{\theta}) - e^{-\theta R}(R + \frac{1}{\theta})}{e^{-\theta L} - e^{-\theta R}} \tag{2.2.3}$$

Equation (2.2.3) is used as the E-Step in the EM algorithm and then the required log-likelihood is given by

$$l^*(\theta) \propto nln\theta - \theta \left[\sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} t_i^* \right]$$

where

$$t_i^* = E[T_i|L_i < T_i < R_i] = \frac{e^{-\theta L_i}(L_i + \frac{1}{\theta}) - e^{-\theta R_i}(R_i + \frac{1}{\theta})}{e^{-\theta L_i} - e^{-\theta R_i}} \tag{2.2.4}$$

Hence, the EM algorithm can be set up as follows. Choose $\theta_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $\theta_{(j)}$ is the jth estimate
- Step 2: Compute T_i^* using Equation (2.2.4) with $\theta = \theta_{(j)}$
- Step 3: Set $\theta_{(j+1)} = \frac{n}{\sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} t_i^*}$
- Step 4: Repeat until convergence criteria is met

IJK (2008) give sufficient conditions for this algorithm to converge, which is quoted below.

Theorem 2.4. *The iterative process will converge if*

$$\sum_{i=n_1+1}^{n} r_i \leq 2\sum_{i=1}^{n_1} t_i + 3\sum_{i=n_1+1}^{n} l_i$$

To prove that it is a global maximum, IJK (2008) show this in Lemmas (2.5) and (2.6).

Lemma 2.5. $\frac{1}{n} l(\theta) \to g(\theta)$ *a.s.*

Lemma 2.6. $g(\theta)$ *is a unimodal function with unique maximum.*

It is then proved that the MLE of θ converges to the point at which $g(\theta)$ attains its maximum. IJK (2008) also gave the asymptotic distribution of the MLE of θ which is given in Theorem (2.7).

Theorem 2.7. *The MLE of θ i.e. $\hat{\theta}$ has the following asymptotic distribution*

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{dist}{\to} N\left(0, \frac{\sigma^2}{c^2}\right)$$

where c and σ have explicit, closed forms. For proofs of the statements see IJK (2008).

To conclude their work, IJK (2008) conducted some numerical simulations to illustrate these methods and found that the numerics were consistent with the theory in all cases. It even converges with a large proportions of censored observations.

IJK (2008) also considered a Bayesian approach to this problem. They gave a priori distribution on θ and since the gamma distribution is a conjugate prior to the exponential distribution, they use a $Gamma(a, b)$ prior which is given by

$$\pi(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}$$

When there is at least one censored observation, the posterior distribution is somewhat cumbersome. However, if there is no censored observations, then the posterior is a $Gamma(a + n, b + \sum_{i=1}^{n} t_i)$. The restricted distribution of T is given by

$$f_{T|T\in(L,R)}(t|\theta) = \frac{\theta e^{-\theta t}}{e^{-\theta L} - e^{-\theta R}} \tag{2.2.5}$$

for $t \in (L < R)$.

IJK (2008) propose using a Gibbs sampling technique to obtain a Bayes estimate of θ and the steps are given below.

- Step 1: Generate $\theta_{(1)}$ from a $Gamma(a + n_1, b + \sum_{i=1}^{n_1} t_i)$

- Step 2: From the restricted distribution (2.2.5), generate the incomplete data t_i^*,

- Step 3: Generate $\theta_{(2)}$ from $Gamma(a+n, b+\sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n} t_i^*)$

- Step 4: Go to Step 2 and replace $\theta_{(1)}$ by $\theta_{(2)}$. Repeat Steps 2 and 3, N times.

Under squared error loss, the Bayes estimate is given by

$$\hat{\theta}_{Bayes} = \frac{1}{N-M}\sum_{1=M+1}^{N}\theta_{(i)}$$

where M is the burn-in sample size. The Bayesian results were found to be consistent with the EM algorithm results.

Bennett (2011) considered lemma (2.8) and then gave theorem (2.9) to show that a similar EM algorithm will converge for a much richer class of distributions under the same censoring mechanism.

Lemma 2.8. *Let $x \in \mathbb{R}, \theta_i \in (\alpha_i, \beta_i)$ for $i \in \{1, 2, \cdots, k\}$, where the interval can be infinite. Let $f(x, \theta) : \mathbb{R}^{k+1} \to \mathbb{R}$ be a continuous function. Define $F(\theta) = \int_a^b f(x, \theta)dx$ where a, b are finite constants. Then $F(\theta)$ is a continuous function.*

Theorem 2.9. *Let $x_1, \cdots, x_{n_1}, (l_{n_1+1, r_{n_1+1}}), \cdots, (l_{n_1+n_2}, r_{n_1+n_2})$ be the observed middle - censored data from a continuous exponential family distribution*

$$f(x|\phi) = h(x)c(\phi)exp\left[\sum_{j=1}^{k} w_j(\phi)t_j(x)\right]$$

such that $h(x), t_j(x), c(\phi)$ and $w_j(\phi)$ are all continuous functions. Then the EM algorithm will converge for this data.

See Bennett (2011) for proofs of lemma (2.8) and theorem (2.9).

2.2.2 Weibull distribution

Bennett (2011) considered lifetime that are Weibull distributed i.e.

$$T_1, \cdots, T_n \overset{iid}{\sim} f(t|a,b) = abt^{a-1}exp(-bt^a)$$

for $t > 0$. The censoring mechanism is the same as before i.e. the left censoring point for each individual L_i is assumed to be an exponential random variable with mean $\frac{1}{\alpha}$ and the length $Z_i = R_i - L_i$ is assumed to be another independent exponential random variable with mean $\frac{1}{\beta}$. Moreover, the T_i's and L_i's and Z_i's are all independent of each other.

Reordering the data into the uncensored and censored observations, the observed data is written as $T_1, \cdots, T_{n_1}, (L_{n_1+1}, R_{n_1+1}), \cdots, (L_{n_1+n_2}, R_{n_1+n_2})$ where $n_1 + n_2 = n$. The log-likelihood can be written as

$$l(a,b) \propto n_1 ln(a) + n_1 ln(b) + (a-1)\sum_{i=1}^{n_1} ln(t_i) - b\sum_{i=1}^{n_1} t_i^a + \sum_{i=n_1+1}^{n_1+n_2} ln(e^{-bl_i^a} - e^{-br_i^a})$$

The conditional expectations needed in order to use the EM-algorithm are given as

$$E[T^a|L < T < R] = \frac{\int_L^R t^a abt^{a-1}e^{-bt^a}dt}{exp[-bl_i^a] - exp[-br_i^a]} \tag{2.2.6}$$

$$E[ln(T)|L < T < R] = \frac{\int_L^R ln(t)abt^{a-1}e^{-bt^a}dt}{exp[-bl_i^a] - exp[-br_i^a]} \tag{2.2.7}$$

The conditional expectations (2.2.6) and (2.2.7) do not have a closed form like Equation (2.2.3) but they can be found numerically. The required log-likelihood for the M-step is given as

$$\begin{aligned} l^*(a,b) \propto & nln(a) + nln(b) + (a-1)\left[\sum_{i=1}^{n_1} ln(t_i) + \sum_{i=n_1+1}^{n_1+n_2} ln(t_i)^*\right] \\ & - b\left[\sum_{i=1}^{n_1} t_i^a + \sum_{i=n_1+1}^{n_1+n_2} (t_i^a)^*\right] \end{aligned} \tag{2.2.8}$$

where the t_a^*'s and $ln(t)^*$ are found using Equations (2.2.6) and (2.2.7) respectively.

Hence, the EM-algorithm is set up as follows. Choose $(a,b)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(a,b)_{(j)}$ is the jth estimate
- Step 2: Compute the incomplete data, T_i^*, using Equations (2.2.6) and (2.2.7) with $(a,b) = (a,b)_{(j)}$
- Step 3: Solve equation (2.2.8) for its maximum and set $(a,b)_{(j+1)}$ as that maximum
- Step 4: Repeat until convergence criteria is met

Bennett (2011) showed that the EM algorithm converges by using Theorem (2.9). The author also conducted some numerical simulations to illustrate these methods and found that the numerics were consistent with the theory in all cases. It even converges with a large proportions of censored observations.

2.3 Parametric models in the presence of time-independent covariates

2.3.1 AFT model and theoretical results

Bennett (2011) considered a p-parameter AFT model in middle censoring for parametric models in the presence of covariates. A general case of these models is where $t_1, \cdots, t_{n_1}, (l_{n_1+1}, r_{n_1+1}), \cdots, (l_{n_1+n_2}, r_{n_1+n_2})$ are the observed middle-censored data from a p-parameter AFT model with a baseline density being a continuous k-parameter exponential family distribution

$$f(t|Z) = f_0(e^{\theta^T z}t) = h(e^{\theta^T z}t)c(\phi)exp\left[\sum_{j=1}^{k} w_j(\phi)\nu_j(e^{\theta^T z}t)\right]$$

where $h(\cdot), \nu_j(\cdot), c(\phi)$ and $w_j(\phi)$ are all continuous functions.

It is assumed that there is at least one uncensored observation and at least one uncensored observation, hence $n_1 > 0$ and $n_2 > 0$. This gives the following complete likelihood:

$$\begin{aligned} l(\phi, \theta) =& nln[c(\phi)] + \sum_{i=1}^{n_1}\left[ln[h(e^{\theta^T z_i}t_i)] + \sum_{j=1}^{k} w_j(\phi)\nu_j(e^{\theta^T z_i}t_i)\right] \\ &+ \sum_{i=n_1+1}^{n_1+n_2}\left[ln[h(e^{\theta^T z_i}l_i)] + \sum_{j=1}^{k} w_j(\phi)\nu_j(e^{\theta^T z_i}l_i)\right] \end{aligned} \tag{2.3.1}$$

To find the MLE of the parameters in this model, the EM algorithm is implemented. The estimates are updated with the following steps:

- Step 1: Suppose that $(\phi, \theta)_{(j)}$ is the jth estimate
- Step 2: Compute T_i^* by calculating $E[T_i|l_i < T_i < r_i, (\phi, \theta) = (\phi, \theta)_{(j)}]$

- Step 3: Solve Equation (2.3.1) with the T_i^*'s imputed for the censored observations for its maximum and set $(\phi, \theta)_{(j+1)}$ as the values that maximize that equation

- Step 4: repeat until convergence criteria is met

Bennett (2011) showed that the algorithm converges and is stated in Theorem (2.10).

Theorem 2.10. *Let $t_1, \cdots, t_{n_1}, (l_{n_1,1}, r_{n_1+1}), \cdots, (l_{n_1+n_2}, r_{n_1+n_2})$ be the observed middle -censored data from an AFT model with p regression parameters and a baseline survival function coming from a continuous k-parameter exponential family distribution. Then the EM-algorithm will converge for this data.*

The author considered the Exponential AFT and Weibull AFT models and shown in Sections (2.3.2) and (2.3.3).

2.3.2 Exponential AFT model

An exponential AFT model with middle censoring is considered where each person has a survival time, T_i, and covariates specific to that individual, Z_i. The lifetimes are $Exp(aexp[\theta^T \mathbf{Z}])$ where θ is the effect of each covariate i.e.

$$T_i \sim aexp\left(\theta^T \mathbf{Z}_i\right) exp\left[-aexp[\theta^T Z_i]t\right]$$

for $t > 0, i = 1, 2, \cdots, n$.

Again, consider the censoring mechanism given in Section (2.2) and assume that there is at least one censored observation, hence $n_2 > 0$. With this set-up, the log-likelihood is given by

$$
\begin{aligned}
l(a,\theta) \propto & n_1 ln(a) + \sum_{i=1}^{n_1} \theta^T \mathbf{Z}_i - a \sum_{i=1}^{n_1} exp\left[\theta^T \mathbf{Z}_i\right] t_i \\
& + \sum_{i=n_1+1}^{n} ln\left[exp[-aexp\left[\theta^T \mathbf{Z}_i\right] l_i] - exp[-aexp(\theta^T \mathbf{Z}_i) r_i]\right]
\end{aligned}
$$

Applying the EM-algorithm in the same fashion as Section (2.2). The conditional expectation is

$$
E[T_i | L_i < T_i < R_i] = \frac{e^{-aexp[a^T \mathbf{Z}_i] L_i}\left(L_i + \frac{1}{ae^{\theta^T \mathbf{Z}_i}}\right) - e^{-aexp[a^T \mathbf{Z}_i] R_i}\left(R_i + \frac{1}{ae^{\theta^T \mathbf{Z}_i}}\right)}{e^{-aexp[\theta^T \mathbf{Z}_i] L_i} - e^{-aexp[\theta^T \mathbf{Z}_i] R_i}} \tag{2.3.2}
$$

Then the required log-likelihood is given by

$$
l^*(a,\theta) \propto nln(a) + \sum_{i=1}^{n} \theta^T \mathbf{Z}_i - a\left(\sum_{i=1}^{n_1} exp[\theta^T \mathbf{Z}_i] t_i + \sum_{i=n_1+1}^{n} exp[\theta^T \mathbf{Z}_i] t_i^*\right) \tag{2.3.3}
$$

where the t_i^*'s are found using Equation (2.3.2).

To run the EM-algorithm, choose $(a,\theta)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(a,\theta)_{(0)}$ is the jth estimate
- Step 2: Compute T_i^* using Equation (2.3.2) with $(a,\theta) = (a,\theta)_{(j)}$
- Step 3: Solve equation (2.3.3) for its maximum and set $(a,\theta)_{(j+1)}$ as that maximum
- Step 4: Repeat until convergence criteria is met

By Theorem (2.10), it is known that the above algorithm will converge, but further research needs to be done to ensure that convergence point of this algorithm is a global maximum and hence the MLE of (a, θ) (Bennett, 2011). The asymptotic distribution of (a, θ) is also of interest. One should perform the usual check and run this algorithm many times with different initial values for (a, θ) to ensure that the algorithm is not trapped at local extrema.

2.3.3 Weibull AFT model

In this section, each person has a survival time T_i and covariates specific to that individual $\mathbf{Z}_i$. The baseline lifetimes will be $Weibull(a, b)$ thus the distribution of T_i is

$$f(t|Z) = a(bexp[a\theta^T\mathbf{Z}])t^{a-1}exp[-(bexp[a\theta^T\mathbf{Z}])t^a] \tag{2.3.4}$$

for $t > 0$.

If the density is as above, each individual lifetime T_i has a $Weibull(a, bexp[a\theta^T\mathbf{Z}_i])$ distribution. Again, consider the censoring mechanism given in Section (2.2) and also assume that there is at least one censored observation, hence $n_2 > 0$. With this setup, the log-likelihood is given by

$$\begin{aligned} l(a, b, \theta) \propto & n_1 ln(a) + n_1 ln(b) + a\sum_{i=1}^{n_1}\theta^T\mathbf{Z}_i + (a-1)\sum_{i=1}^{n_1} ln(t_i) - b\sum_{i=1}^{n_1} t_i^a exp(a\theta^T\mathbf{Z}_i) \\ & + \sum_{i=n_1+1}^{n} ln\left(exp[-be^{a\theta^T\mathbf{Z}_i}l_i^a] - exp[-be^{a\theta^T\mathbf{Z}_i}r_i^a]\right) \end{aligned} \tag{2.3.5}$$

Next, apply the EM-algorithm in the same fashion as Section (2.2). Then the conditional expectation is needed to do this is given by

$$g(T_i)^* = E[g(T_i)|L_i < T_i < R_i] = \int_{L_i}^{R_i} g(t)f(t|Z_i)dt \tag{2.3.6}$$

where $f(t|Z_i)$ is given by Equation (2.3.4). Then, the log-likelihood required for the EM-algorithm is

$$\begin{aligned} l^*(a,b,\theta) \propto & nln(a) + nln(b) + a\sum_{i=1}^{n}\theta^T Z_i + (a-1)\sum_{i=1}^{n_1} ln(t_i) - b\sum_{i=1}^{n_1} t_i^a exp(a\theta^T Z_i) \\ & + (a-1)\sum_{i=n_1+1}^{n} ln(t_i^*) - b\sum_{i=n_1+1}^{n} (t_i^*)^a exp(a\theta^T \mathbf{Z}_i) \end{aligned} \tag{2.3.7}$$

where the t_i^*'s are found using Equation (2.3.6).

Now to run the EM-algorithm, choose $(a,b,\theta)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(a,b,\theta)_{(j)}$ is the jth estimate
- Step 2: Compute $g(T_i^*)$ using Equation (2.3.6) with $(a,b,\theta) = (a,b,\theta)_{(j)}$
- Step 3: Solve Equation (2.3.7) for its maximum and set $(a,b,\theta)_{(j+1)}$ as that maximum
- Step 4: Repeat until convergence criteria is met

The previous theorems on Exponential family members cannot be used here, but it has been proved in Bennett (2011) that the EM-algorithm converges in this case as well which is given in Theorem (2.11).

Theorem 2.11. *Let $t_1, \cdots, t_{n_1}, (l_{n_1+1}, r_{n_1+1}), \cdots, (l_{n_1+n_2}, r_{n_1+n_2})$ be the observed middle - censored data from a Weibull AFT model. Then the EM-algorithm will converge for this data.*

2.4 Semiparametric models in the presence of time - independent covariates

2.4.1 Proportional Hazards Model

Bennett (2011) discussed the Cox proportional hazard model for the semiparametric models in the presence of covariates in middle censoring. The model has been studied extensively in the case of right censoring (see Cox, 1972 and Efron, 1977). The Cox model is given by

$$S(t|Z) = S_0(t)^{exp(\theta Z)} \tag{2.4.1}$$

where $S(t)$ is the survival function for a non-negative random variable. With this semiparametric setup, the density of lifetimes is given by

$$f(t|Z) = -\frac{\partial}{\partial t}S(t|Z) = f_0(t)exp(\theta Z)S_0(t)^{exp(\theta Z)-1} \tag{2.4.2}$$

where the survival function, $S(t|Z)$ is given in Equation (2.4.1). The baseline survival function, $S_0(t)$ is treated as a nuisance parameter and is not estimated.

Without loss of generality let $t_1, \cdots, t_{n_1}, (l_{n_1+1}, r_{n_1+1}), \cdots, (l_{n_1+n_2}, r_{n_1+n_2})$ be the middle-censored data from equation (2.4.2) under the general censoring scheme. Then the full likelihood is given by

$$L(\theta) = \prod_{uncens} f(t|Z) \prod_{cens} (S(l|Z) - S(r|Z))$$

The corresponding log-likelihood is

$$l_{full}(\theta) = l_{uncens}(\theta) + l_{cens}(\theta)$$

where

$$l_{uncens}(\theta) = \sum_{i=1}^{n_1} ln[f_0(t_i)] + \theta \sum_{i=1}^{n_1} Z_i + \sum_{i=1}^{n_1} [exp(\theta Z_i) - 1] ln[S_0(t_i)] \tag{2.4.3}$$

$$l_{cens}(\theta) = \sum_{i=1}^{n_2} ln[S(l_i|Z_i) - S(r_i|Z_i)] \tag{2.4.4}$$

From equations (2.4.3) and (2.4.4), the estimation of the baseline survival function $S_0(t)$ and the baseline density $f_0(t)$ are required in order to estimate the covariate effect, θ. The survival function is estimated nonparametrically using the self consistent estimator in Equation (2.1.1) given in Jammalamadaka and Mangalam (2003). However, the problem arises when trying to estimate the baseline density function, $f_0(t)$. A possible approach, as suggested by the author, is to fit a smoothing spline to the estimate of the baseline survival function $S_0(t)$ and differentiate this to approximate the desired density. However, while this is possible, there is a nicer and simpler way to avoid doing this in order to obtain the MLE of θ.

The key is to write out the derivative of the log-likelihood as follows

$$l'(\theta) = \frac{\partial}{\partial\theta} l_{uncens}(\theta) + \frac{\partial}{\partial\theta} l_{cens}(\theta) \tag{2.4.5}$$

where the derivatives of the uncensored and censored data are given in Equations (2.4.6) and (2.4.7)

$$\frac{\partial}{\partial\theta} l_{uncens}(\theta) = \sum_{i=1}^{n_1} Z_i + \sum_{i=1}^{n_1} Z_i exp(\theta Z_i) ln(S_0(t_i)) \tag{2.4.6}$$

$$\frac{\partial}{\partial\theta} l_{cens}(\theta) = \sum_{cens} \left[\frac{Z_i e^{\theta Z_i} ln(S_0(l_i)) S_0(l_i)^{e^{\theta Z_i}} - Z_i e^{\theta Z_i} ln(S_0(r_i)) S_0(r_i)^{e^{\theta Z_i}}}{S(l_i, Z_i) - S(r_i, Z_i)} \right] \tag{2.4.7}$$

From Equations (2.4.6) and (2.4.7), the baseline density $f_0(t)$ is not present. Hence, it is not needed in order to solve for the roots of this equation. Again, while there is no general closed form solution to Equation (2.4.5), it can be solved numerically. The algorithm to find the Maximum Likelihood estimate of the regression parameter θ and also to estimate the baseline survival function $S_0(t)$ in a Cox proportional hazard model where the distribution of covariate values Z_i follow a $Binomial(1, p)$ distribution is constructed by Bennett (2011) as follows:

- Step 1: Estimate $S_0^{(1)}(t)$ via NPMLE or SCE only using the data with no covariate effect.
- Step 2: Estimate $\theta_{(1)}$ by solving the root of equation (2.4.5) and using $S_0^{(1)}(t)$ to solve for the necessary probability in it.
- Step 3: Find $\tilde{t}_i = (S_0^{(1)})^{-1}\left[S_0^{(1)}(t_i)^{exp(\theta^{(1)}Z_i)}\right]$. One can find $\tilde{l}_i$ and $\tilde{r}_i$ similarly. Note: If $z_i = 0$, then $\tilde{t}_i = t_i$ by definition of the Cox model.
- Step 4: Estimate $S_0^{(2)}(t)$ via SCE (or NPMLE) using all $\tilde{t}_i$, $\tilde{l}_i$ and $\tilde{r}_i$ as your data.
- Step 5: Estimate $\theta_{(2)}$ by solving equation (2.4.5) and using $S_0^{(2)}(t)$ to solve necessary probability.
- Step 6: Repeat steps (3) - (5) until convergence.

The author conducted some numerical simulations to illustrate these methods and found that even under high amount of censoring, the fitted CDF and MLE of the parameter θ perform remarkably well.

2.4.2 Accelerated Failure Times model

Bennett (2011) then discussed the use of Accelerated Failure Times, AFT model in this context. The survival function in this case is given as

$$S(t|Z) = S_0(te^{\theta Z})$$

Hence, the density function is given by

$$f(t|Z) = -\frac{d}{dt}S(t,Z) = e^{\theta Z}f_0(te^{\theta Z})$$

The author stated that a very large underlying assumption needs to be considered which is, the baseline survival function $S_0(t)$ is known. This is a common practice in engineering and in reliability studies where the exponential distribution is one of the more commonly assumed lifetime distributions.

Recall that the Cox PH assumption is equivalent to the AFT assumption when the baseline distribution is an exponential distribution i.e. if $T \sim Exp(a)$ then for $t > 0$, $S(t) = e^{-at}$. Hence,

$$S_0(te^{\theta\mathbf{Z}}) = exp(-ate^{\theta\mathbf{Z}}) = e^{-ate^{\theta\mathbf{Z}}} = S_0(t)^{e^{\theta\mathbf{Z}}}$$

which shows that these two models are equivalent.

Hence, if it is assumed that the true distribution of lifetimes is an exponential distribution, then the same methodology described in Section (2.4.1) for the Cox PH model can be used. The algorithm is given as follows:

1. Estimate $S_0^{(1)}(t)$ via NPMLE or SCE using all of the data

2. Estimate $\theta^{(1)}$ by solving the root of Equation (2.4.5) and using $S_0^{(1)}(t)$ to solve the necessary probabilities in it.

3. Find $\tilde{t_i} = t_i exp(\theta^{(1)} Z_i)$. One can find $\tilde{l_i}$ and $\tilde{r_i}$ in the same fashion.

4. Estimate $S_0^{(2)}(t)$ via SCE or NPMLE using all of the $\tilde{t_i}$, $\tilde{l_i}$ and $\tilde{r_i}$ as your data

5. Estimate $\theta^{(2)}$ by solving the root of Equation (2.4.5) and using $S_0^{(2)}(t)$ to solve for the necessary probabilities in it

6. Repeat steps (3)-(5) until convergence

The author conducted some numerical simulations to illustrate these methods and found that even under high amount of censoring, the fitted CDF and MLE of the parameter θ perform remarkably well and the procedure gives extremely accurate results..

2.5 Discrete middle-censored data

An example of discrete lifetimes is a study in which a women who stop using oral contraception are followed until pregnancy. The number of cycles rather than the time to pregnancy is used because the cycle length varies among women and a woman ovulates only once per menstrual cycle. The number of cycles is a discrete outcome. Davarzani and Parsian (2011) first discussed middle censoring in a discrete set-up.

Following the spirit of Iyer, Jammalamadaka and Kundu (2008), it is assumed that there is a random sample of individuals of size n from a specific population with lifetimes

$T_1, T_2, \cdots, T_n$. Corresponding to every individual in the sample, there is a random censoring interval $[L_1, R_1]$, $[L_2, R_2]$, $\cdots$, $[L_n, R_n]$ which are independent of the lifetimes. Just as in the previous set-up the observed data X_i's are given by

$$X_i = \begin{cases} T_i & \text{if } T_i \notin (L_i, R_i) \\ \\ (L_i, R_i) & \text{if } T_i \in (L_i, R_i) \end{cases}$$

for $i = 1, 2, \cdots, n$.

Davarzani and Parsian (2011) constructed the following censoring mechanism. Assume that $T_1, T_2, \cdots, T_n$ are an $i.i.d$ sample from a geometric distribution with mean $\frac{1-\theta_0}{\theta_0}$ i.e. with probability function

$$P(T_i = t_i) = \theta_0(1 - \theta_0)^{t_i}$$

for $t_i = 0, 1, 2, \cdots$.

The left point of the censored interval, L_i is a geometric random variable with mean $(1 - p_l)/p_l$ and the length of the censored interval, S_i is a geometric random variables with mean $1/p_s$ i.e.

$$L_i \overset{iid}{\sim} P(L_i = l_i) = p_l(1 - p_l)^{l_i}$$

$$S_i \overset{iid}{\sim} P(S_i = s_i) = p_s(1 - p_s)^{s_i - 1}$$

for $l_i = 0, 1, 2, \cdots$, $s_i = 1, 2, 3, \cdots$ and $S_i = R_i - L_i$.

The lifetimes, T_i, L_i and S_i are independent for all i. Davarzani and Parsian (2011) proposed solving for the MLE in the case where the lifetimes T_i's have a geometric distribution i.e. $T_i \sim Geometric(\theta)$ by using the EM-algorithm.

The model is completely parametrized, hence the likelihood can be easily written down and the MLE of θ can be solved. Assume that there are $n_1 > 0$ uncensored observations and $n_2 > 0$ censored observations where $n = n_1 + n_2$. Without loss of generality, after re-ordering the data, it is assumed that the first n_1 are the uncensored observations while the remaining n_2 are the uncensored observations. Hence the observed data is

$$T_1, T_2, \cdots, T_{n_1}, [L_{n_1+1}, R_{n_1+1}], [L_{n_1+2}, R_{n_1+2}] \cdots, [L_{n_1+n_2}, R_{n_1+n_2}]$$

The likelihood function of the observed data is written as

$$L(\theta) = c\theta^{n_1}(1-\theta)^{(\sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} l_i)} \prod_{i=n_1+1}^{n_1+n_2} (1-(1-\theta)^{s_i+1}) \tag{2.5.1}$$

where $c = c_1^{n_2} c_2^{n_1}$ is the normalizing constant which does not depend on θ. From Equation (2.5.1), the log-likelihood function of θ is

$$l(\theta) = ln(c) + n_1 ln(\theta) + \sum_{i=1}^{n_1} t_i ln(1-\theta) + \sum_{i=n_1+1}^{n_1+n_2} l_i ln(1-\theta) + \sum_{i=n_1+1}^{n_1+n_2} ln(1-(1-\theta)^{s_i+1}) \tag{2.5.2}$$

Taking derivative of equation (2.5.2) with respect to θ gives the following equation

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{n_1}{\theta} - \frac{1}{1-\theta}\left(\sum_{i=1}^{n_1} t_i\right) - \frac{1}{1-\theta}\left(\sum_{i=n_1+1}^{n_1+n_2} l_i\right) + \sum_{i=n_1+1}^{n_1+n_2} \frac{(s_i+1)(1-\theta)^{s_i}}{1-(1-\theta)^{s_i+1}} \tag{2.5.3}$$

Equation (2.5.3) shows that it is not possible the obtain the MLE of θ in an explicit form. Also, this is an example of incomplete data. Hence the authors suggested using EM algorithm to find the MLE of θ.

$$\begin{aligned} E_\theta(T|L \le T \le R) &= \sum_{t=L}^{R} t P_\theta(T = t | L \le T \le R) \\ &= \left[L + \frac{(1-\theta) - (S+1)(1-\theta)^{S+1} + S(1-\theta)^{S+2}}{(1-(1-\theta)^{S+1})(1-(1-\theta))}\right] \end{aligned} \tag{2.5.4}$$

Equation (2.5.4) is used as the E-step in the EM algorithm and the required log-likelihood is given as

$$l^*(\theta) \propto nln(\theta) + \sum_{i=1}^{n_1} t_i ln(1-\theta) + \sum_{i=n_1+1}^{n_1+n_2} t_i^* ln(1-\theta) \qquad (2.5.5)$$

where

$$t_i^* = E_\theta(T_i|L_i \leq T_i \leq R_i) = \left[L_i + \frac{(1-\theta)-(S_i+1)(1-\theta)^{S_i+1}+S_i(1-\theta)^{S_i+2}}{(1-(1-\theta)^{S_i+1})(1-(1-\theta))}\right]$$

Thus, the EM algorithm can be set up as follows. Choose $\theta_{(0)}$ to be the MLE of the uncensored data i.e. $\theta_{(0)} = \frac{n_1}{n_1+\sum_{i=1}^{n_1} t_i}$. Update the estimates with the following steps.

- Step 1: Suppose that $\theta_{(j)}$ is the jth estimate
- Step 2: Compute T_i^* by using Equation (2.5.4) with $\theta = \theta_{(j)}$
- Step 3: Set $\theta_{(j+1)} = \frac{n}{n+\sum_{i=1}^{n_1} t_i+\sum_{i=n_1+1}^{n_1+n_2} t_i^*}$
- Step 4: Repeat until convergence is met.

Davarzani and Parsian (2011) gave a sufficient condition for the algorithm to converge, proved that the MLE of θ is a consistent estimator and gave an asymptotic distribution of the MLE.

Chapter 3

Discrete middle-censored data

Middle censoring in a discrete set-up was first discussed by Davarzani and Parsian (2011) (DP from now on) where the lifetimes as well as the lower limit and length of censoring interval are assumed to have geometric distribution. In this chapter, lifetimes that follow a geometric distribution as in DP (2011) are considered, but we generalize their set-up to the important case where covariates are present, as well as provide alternate results and proofs by exploiting the simple relationship between the exponential and geometric distributions.

In Section 3.1.1 this connection is discussed, and used in Section 3.2 to find the maximum likelihood estimates (MLEs) under middle-censoring in the presence of covariates using Accelerated Failure Time model for the geometric case, and discuss the EM algorithm for obtaining them. The novelty of this approach, contrasted with that in DP (2011), is to adapt the methods of Iyer, Jammalamadaka and Kundu (2008) to the geometric case. Simulation studies are carried out, in support of the theory to indicate how well the proposed estimation methods work. The asymptotic distribution of the MLE in terms of Fisher information is also considered here. Section 3.3 illustrates the application of the proposed model to the Stanford Heart transplant study from Crowley and Hu (1977).

3.1 An Alternate Approach to Discrete Lifetimes and with Covariates

In this section, it can be shown that one can utilize the connection between the geometric and exponential distributions, so that the results in DP (2011) can be subsumed by what has been done in Iyer, Jammalamadaka and Kundu (2008) for exponential data. This connection will also allow us to more readily extend the results to the case of covariates, as we do in Section 3.2.

3.1.1 An important link

As is known, the geometric distribution can be thought of as the discrete analogue of the exponential distribution, and the following well-known lemma provides the elegant connection between the two. We add the short proof for completeness as well as to introduce the notations:

Lemma 3.1. *If X is an exponentially distributed random variable with parameter λ, then $Y = \lfloor X \rfloor$ where $\lfloor\ \rfloor$ is the floor function (the integer part of x), is a geometrically distributed random variable with parameter $p = 1 - e^{-\lambda}$.*

Proof. Let $X \sim exp(\lambda)$ with p.d.f. $f(x) = \lambda e^{-\lambda x}$. Suppose we have $Y = \lfloor X \rfloor$. Then,

$$P(\lfloor X \rfloor = a) = P(a \leq X < a+1) = e^{-a\lambda}(1 - e^{-\lambda}) = (1-p)^a p$$

Therefore, $Y = \lfloor X \rfloor \sim geometric(p)$ where $p = 1 - e^{-\lambda}$, $a = 0, 1, 2, \cdots$, $0 \leq p \leq 1$ and $\lambda > 0$. □

The geometric distribution also inherits the interesting property known as the memoryless property which the exponential distribution has. For integers, $s > t$, it is the case that

$$P(X > s | X > t) = P(X > s - t)$$

that is, the geometric distribution 'forgets' what has occurred. The probability of getting an additional $s - t$ failures, having already observed t failures, is the same as the probability of observing $s - t$ failures at the start of the sequence.

Applying the property of memorylessness and using the relationship of exponential and geometric distributions from Lemma (3.1) to middle censoring, the geometric lifetimes can be generated from the exponentially distributed lifetime, $T_i \sim Exp(\lambda)$.The geometric distributed lifetimes is $Y_i = \lfloor T_i \rfloor \sim geometric(p)$ with probability function

$$P(Y_i = y_i) = p(1-p)^{y_i}$$

for $y_i = 0, 1, 2, \cdots$ and $p = 1 - e^{-\lambda}$.

The left point of the censored interval, $U_i = \lfloor L_i \rfloor \sim geometric(p_u)$ with probability function

$$P(U_i = u_i) = p_u(1-p_u)^{u_i}$$

where $L_i \sim Exp(\alpha)$, $p_u = 1 - e^{-\alpha}$ and $u_i = 0, 1, 2, \cdots$ while the length of the censored interval is $W_i = \lfloor S_i \rfloor \sim geometric(p_w)$ with probability function

$$P(W_i = w_i) = p_w(1 - p_w)^{w_i - 1}$$

where $S_i \sim Exp(\beta)$,$S_i = R_i - L_i$, $p_w = 1 - e^{-\beta}$, $w_i = 1, 2, \cdots$ and $W_i = V_i - U_i$ where V_i is the right point of the censored interval. The lifetimes Y_i, U_i and W_i are independent for all i.

3.2 Geometric model in the presence of covariates

In this section, a geometric lifetime with middle censoring in the presence of covariates is considered. The covariates that are considered here are fixed, that is, known at baseline or entry to the study. The relationship between an exponential distribution and the geometric distribution discussed in Section (3.1.1) can also be applied here for a geometric lifetime in the presence of covariates. Here, each person has a survival time, T_i and covariates specific to that individual $\boldsymbol{Z}_i$.

It may be recalled that when the baseline distribution is an exponential, the Cox proportional hazard assumption is equivalent to the accelerated failure time assumption. See e.g. Cox and Oakes (1984, pages 70-72) who show that the exponential regression model is an example of an accelerated failure time model with proportional hazards. Hence, the lifetimes, T_i are first generated from an exponential accelerated failure model or a Cox PH model when the baseline distribution is the exponential distribution i.e. $T_i \sim$

$Exponential(\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i})$ with p.d.f.

$$f(t|\boldsymbol{Z}_i) = \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} exp(-\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} t)$$

where $\boldsymbol{\theta}$ is the effect of each covariate $\boldsymbol{Z}$, and the superscript T stands for transpose operation. Hence one can generate geometric lifetimes from the generated exponential lifetime that is, $Y_i = \lfloor T_i \rfloor \sim geometric(p_i)$ where $p_i = 1 - e^{-\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}$. We take the left end-point of the censored interval $\boldsymbol{U}_i \sim geometric(p_u)$ while the width of the censored interval is taken to be $\boldsymbol{W}_i \sim geometric(p_w)$ where $\boldsymbol{W}_i = \boldsymbol{V}_i - \boldsymbol{U}_i$ and $\boldsymbol{V}_i$ is the right censored point of the censored interval.

Since the model is completely parametric, the likelihood can be written down and the MLE of p can be solved. Suppose that there are $n_1 > 0$ uncensored observations and $n_2 > 0$ censored observations where $n = n_1 + n_2$. After re-ordering the data, without loss of generality, it is assumed that the first n_1 are the uncensored observations while the remaining n_2 are the censored observations. Hence the observed data is

$$\{Y_1, Y_2, \cdots, Y_{n_1}, [U_{n_1+1}, V_{n_1+1}], [U_{n_1+2}, V_{n_1+2}] \cdots, [U_{n_1+n_2}, V_{n_1+n_2}]\}.$$

Similar to the methods used in DP (2011) in Section 2, the likelihood function of the observed data is written as

$$L(p_i) = c p_i^{n_1} (1-p_i)^{(\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n_1+n_2} u_i)} \prod_{i=n_1+1}^{n_1+n_2} (1 - (1-p_i)^{w_i+1}) \tag{3.2.1}$$

where $c = c_1^{n_2} c_2^{n_1}$ is the normalizing constant which does not depend on p_i where $p_i = 1 - e^{-\lambda e^{\boldsymbol{\theta}^T \boldsymbol{z}_i}}$. From Equation (3.2.1), the log-likelihood function of p_i is

$$\begin{aligned} l(p_i) =& ln(c) + n_1 ln(p_i) + \sum_{i=1}^{n_1} y_i ln(1-p_i) + \sum_{i=n_1+1}^{n_1+n_2} u_i ln(1-p_i) \\ &+ \sum_{i=n_1+1}^{n_1+n_2} ln(1-(1-p_i)^{w_i+1}) \end{aligned} \tag{3.2.2}$$

Applying the EM algorithm to find the MLE of p, the following conditional expectation is required:

$$\begin{aligned} E_p(Y|U \le Y \le V) &= \sum_{y=U}^{V} y P_p(Y=y|U \le Y \le V) \\ &= \left[U + \frac{(1-p) - (W+1)(1-p)^{W+1} + W(1-p)^{W+2}}{(1-(1-p)^{W+1})(1-(1-p))} \right] \end{aligned} \tag{3.2.3}$$

Equation (3.2.3) is used as the E-step in the EM algorithm and the required log-likelihood is given as

$$l^*(p_i) \propto n ln(p_i) + \sum_{i=1}^{n_1} y_i ln(1-p_i) + \sum_{i=n_1+1}^{n_1+n_2} y_i^* ln(1-p_i) \tag{3.2.4}$$

where

$$y_i^* = E_{p_i}(Y_i|U_i \le Y_i \le V_i) = \left[U_i + \frac{(1-p_i) - (W_i+1)(1-p_i)^{W_i+1} + W_i(1-p_i)^{W_i+2}}{(1-(1-p_i)^{W_i+1})(1-(1-p_i))} \right]$$

Thus, the EM algorithm can be set up as follows. Choose p_0 to be the MLE of the uncensored data. Update the estimates with the following steps.

- Step 1: Suppose that $p_{(j)}$ is the jth estimate

- Step 2: Compute Y_i^* by using Equation (3.2.3) with $p = p_{(j)}$

- Step 3: Solve Equation (3.2.4) for its maximum and set $p_{(j+1)}$ as that maximum.

- Step 4: Repeat until a convergence criterion is met.

A simulation study is performed to illustrate the usefulness of this approach. Simulations are carried out in R using $N = 100$ replications with a common sample size $n = 250$. Each sample is then censored and the EM algorithm described above is applied to the censored data. The censoring mechanism is as follows; the left endpoint of the censored interval is geometric distributed with mean 0.5 and the length of the censored interval is also geometric distributed with mean 0.1. Three covariates are used in this simulation. The covariates Z_1 and Z_2 are generated from a binomial distribution with one trial and probability of success equal to 0.5. The third covariate, Z_3 is generated from a standard normal distribution. Three cases for the true covariate effects are considered here, similar to Pan (1999). They are $\theta = (1, 1, 1), \theta = (1, 0, 0)$ and $\theta = (0, 0, 1)$ and are chosen since they represent the case where the covariates have an equal effect, where only one Bernoulli covariate has one effect, and where only the normally distributed covariate had an effect. The true values of λ are chosen to be 0.5 and 0.3 as shown in Table 3.1. A number of different starting points were used in the EM-algorithm in order to capture the global maximum.

The 'Initial est' is the average value of the $N = 100$ estimates using only the uncensored observations, 'MLE' reported is the average value of the $N = 100$ estimates obtained using the model and the estimated mean-squared error, $EMSE$ is calculated using the equation

$$EMSE(\hat{p}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{p} - p)^2$$

where $\hat{p}$ is the estimate of p and a total of N simulation were computed. The standard deviation (SD) of the estimates are evaluated in the simulation and their confidence intervals could be evaluated. The ‘Censored proportion’ line in the table gives the mean proportion of censoring in the $N = 100$ simulated samples.

In the $N = 100$ simulations, the samples were found to be between 14% and 29% censored. The MLE’s of $\lambda, \theta_1, \theta_2, \theta_3$ were computed using the EM algorithm described above. See Table 3.1 for the results from these simulations. The maximum likelihood estimates, (MLE) are fairly close to the actual value compared to the initial estimates (Initial Est) and the estimated mean-squared errors (EMSE) are small. The standard deviation (SD) of the estimates are evaluated in the simulation and their confidence intervals could be evaluated. This approach yields very useful, accurate and reliable results. Note that we initialized the EM-algorithm from a number of different starting points and it shows that the likelihood does have a unique maximum.

As an illustration, box plots shown in Figures 3.2.1-3.2.4 are constructed to compare the difference between the estimates found using just the uncensored observations and the MLE using our model when the true values are $\lambda = 0.5, \theta_1 = 1, \theta_2 = 1$ and $\theta_3 = 1$. For each of the figures, the box plot on the left represents the estimates found using only the uncensored observations while the box plot on the right shows the MLEs found using our model. All four figures show that the MLEs using our model, evident from the right box plots, are closer to the true values compared to the estimates using the uncensored observations only. This shows how well the information lost due to middle censoring is recovered using our estimation methods.

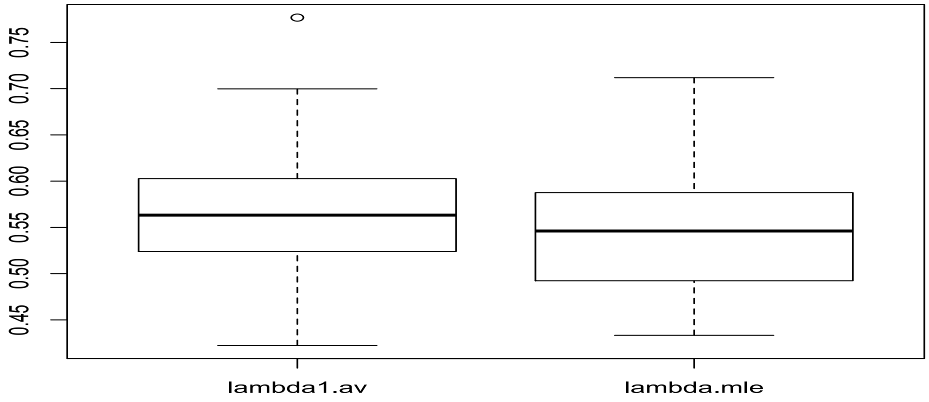


Figure 3.2.1: True value, $\lambda = 0.5$

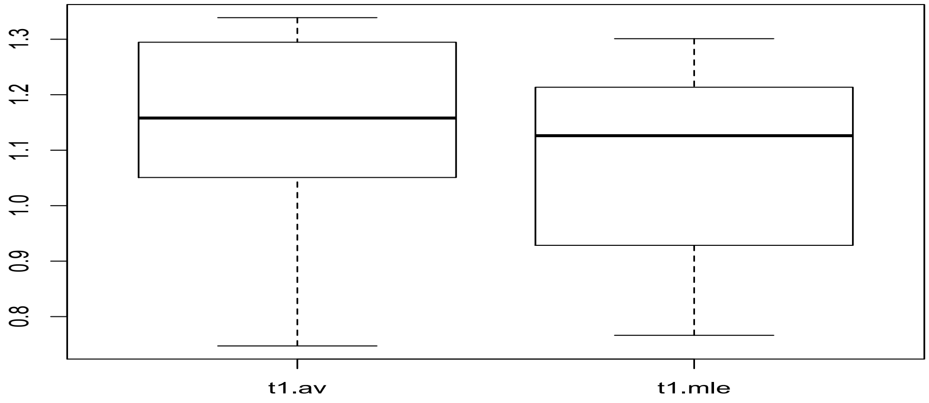


Figure 3.2.2: True value, $\theta_1 = 1$

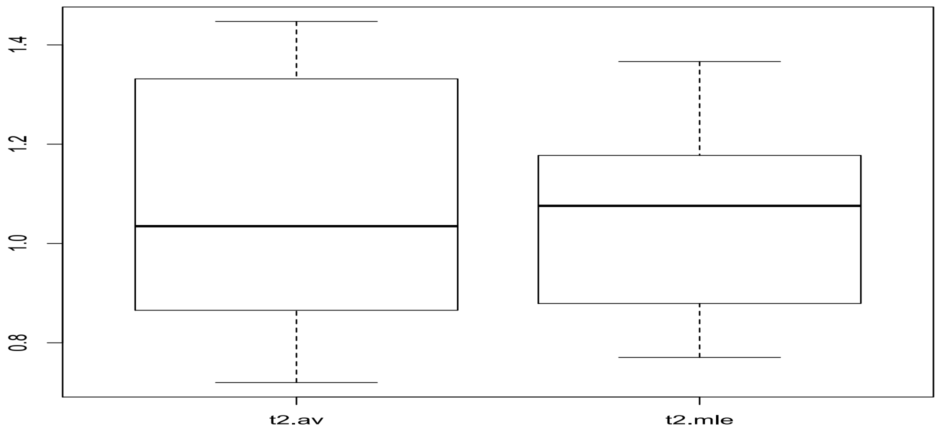


Figure 3.2.3: True value, $\theta_2 = 1$

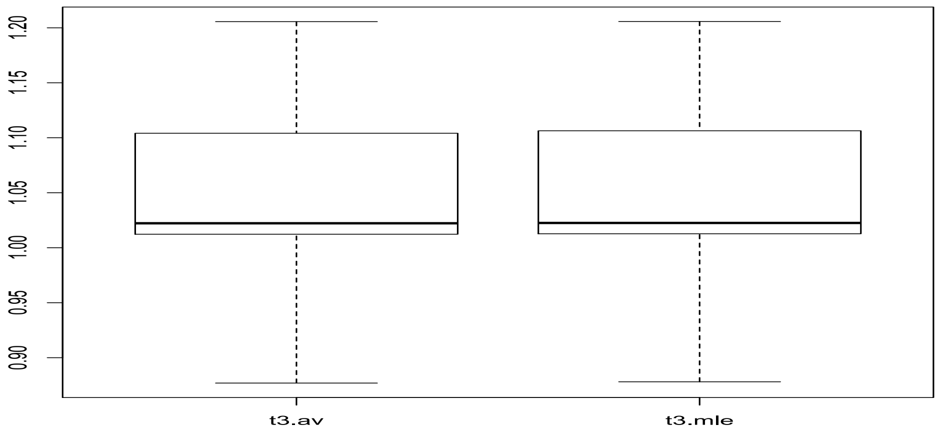


Figure 3.2.4: True value, $\theta_3 = 1$

Parameter	True Value	Initial Est	MLE	SD	EMSE	Censored Prop
λ	0.5	0.5636	0.5312	0.1161	0.0150	0.1452
θ_1	1.0	1.0838	1.0308	0.2869	0.0855	
θ_2	1.0	1.0432	1.0387	0.2788	0.0790	
θ_3	1.0	1.0673	1.0597	0.1888	0.0365	
λ	0.5	0.5553	0.5428	0.1074	0.0119	0.1444
θ_1	1.0	1.0745	1.0678	0.2229	0.0501	
θ_2	0.0	0.1038	0.0627	0.2256	0.0589	
θ_3	0.0	0.0088	0.0067	0.1120	0.01307	
λ	0.5	0.5698	0.5394	0.0699	0.0051	0.1688
θ_1	0.0	0.0992	0.0774	0.2929	0.0869	
θ_2	0.0	-0.0199	-0.0101	0.2927	0.0876	
θ_3	1.0	1.1005	1.0878	0.1745	0.0315	
λ	0.3	0.3510	0.3401	0.0516	0.0028	0.2006
θ_1	1.0	1.1099	1.0343	0.2741	0.0787	
θ_2	1.0	1.1212	1.0456	0.2750	0.0766	
θ_3	1.0	1.1129	1.0376	0.1457	0.0222	
λ	0.3	0.0366	0.3365	0.0513	0.0027	0.2910
θ_1	1.0	1.2802	1.1600	0.2126	0.0476	
θ_2	0.0	0.0187	0.0174	0.2592	0.0688	
θ_3	0.0	0.0211	0.0128	0.0876	0.0078	
λ	0.3	0.3450	0.3365	0.0388	0.0016	0.2990
θ_1	0.0	-0.0099	-0.0086	0.3144	0.0989	
θ_2	0.0	0.0651	0.0541	0.3122	0.0978	
θ_3	1.0	1.1089	1.0811	0.1585	0.0287	

Table 3.1: Simulation results for Geometric model in the presence of 3 covariates

3.2.1 The case of no covariates

The case where there are no covariates which is considered in DP (2011) comes out as a special case of what we already have, by taking $Z = 0$. In this case, the likelihood function of the observed data is written as in equation (3.2.1) but with $p = 1 - e^{-\lambda}$. The log-likelihood function of p, the conditional expectation for the E-step of the EM algorithm and the log-likelihood required for the M-step are shown in equations (3.2.2), (3.2.3) and (3.2.4) respectively. Hence the EM algorithm is set up as follows. Choose $p_{(0)}$ to be the MLE of the uncensored data i.e. $p_{(0)} = \frac{n_1}{n_1+\sum_{i=1}^{n_1} y_i}$. Update the estimates with the following steps.

- Step 1: Suppose that $p_{(j)}$ is the jth estimate
- Step 2: Compute Y_i^* by using Equation (3.2.3) with $p = p_{(j)}$
- Step 3: Set $p_{(j+1)} = \frac{n}{n+\sum_{i=1}^{n_1} y_i+\sum_{i=n_1+1}^{n_1+n_2} y_i^*}$
- Step 4: Repeat until convergence is met.

Simulations are run in order to test the validity of the program. We considered different sample sizes namely $n = 50, 100, 250$ and 500. For each sample size n, $N = 100$ samples were simulated. Each sample was then censored and the EM algorithm described above was applied to the censored data. See Table 3.2 for the results from these simulations. The 'p est' reported is the average value of the $N = 100$ estimates obtained.

The geometric model converges numerically to the true value in all cases, which is consistent with the result found in DP (2011) for the geometric lifetimes. The estimates are converging to the true values as the sample size, n increases but it appears to converge rather slowly.

n	(p_l, p_z)	(0.5,0.9)	(0.2,0.9)	(0.3,0.8)
50	p est	0.3159	0.3109	0.3097
	EMSE of p	0.0017	0.0019	0.0018
	SD	0.0288	0.0276	0.0309
	Censored Proportion	0.1526	0.1538	0.0912
100	p est	0.3092	0.3095	0.3086
	EMSE of p	0.0009	0.0008	0.0008
	SD	0.0236	0.0229	0.0227
	Censored Proportion	0.1438	0.1628	0.1003
250	p est	0.3056	0.3077	0.3050
	EMSE of p	0.0003	0.0004	0.0003
	SD	0.0178	0.0176	0.0170
	Censored Proportion	0.1471	0.1592	0.0968
500	p est	0.3054	0.3065	0.3035
	EMSE of p	0.0002	0.0002	0.0002
	SD	0.0136	0.0135	0.0131
	Censored Proportion	0.1468	0.1605	0.0946

Table 3.2: Simulation results for Geometric (0.3) lifetimes

3.2.2 Asymptotic distribution of the MLE

It can be checked that the conditions for the validity of the properties of the MLEs, hold. For completeness, we give below the derivatives of the log-likelihood function from equation (3.2.2) where $p = 1 - e^{-\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}$:

$$\frac{\partial l}{\partial \lambda} = \sum_{i=1}^{n_1} \frac{e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}{e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}} - 1} - \sum_{i=1}^{n_1} y_i e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} - \sum_{i=n_1+1}^{n_1+n_2} u_i e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} + \sum_{i=n_1+1}^{n_1+n_2} \frac{e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}(w_i+1)}{e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}(w_i+1)} - 1}$$

and for $j = 1, 2, 3$,

$$\frac{\partial l}{\partial \boldsymbol{\theta}_j} = \sum_{i=1}^{n_1} \frac{\boldsymbol{Z}_j \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}{e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}} - 1} - \sum_{i=1}^{n_1} y_i \boldsymbol{Z}_j \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} - \sum_{i=n_1+1}^{n_1+n_2} u_i \boldsymbol{Z}_j \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} + \sum_{i=n_1+1}^{n_1+n_2} \frac{\boldsymbol{Z}_j \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}(w_i+1)}{e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}(w_i+1)} - 1}$$

The second derivatives are given by

$$\frac{\partial^2 l}{\partial \lambda^2} \;=\; -\sum_{i=1}^{n_1} \frac{e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i+2(\boldsymbol{\theta}^T \boldsymbol{Z}_i)}}}{(e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}-1)^2} - \sum_{i=n_1+1}^{n_1+n_2} \left(\frac{(w_i+1)^2 e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i (w_i+1)+2(\boldsymbol{\theta}^T \boldsymbol{Z}_i)}}}{(e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i (w_i+1)}}-1)^2} \right)$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \lambda \partial \theta_j} =& -\sum_{i=1}^{n_1} \frac{Z_j e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} (\lambda e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}+\boldsymbol{\theta}^T \boldsymbol{Z}_i} - e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}+1)}{(e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}-1)^2} - \sum_{i=1}^{n_1} y_i Z_j e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} - \sum_{i=n_1+1}^{n_1+n_2} u_i Z_j e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} \\ & - \sum_{i=n_1+1}^{n_1+n_2} \frac{(w_i+1) Z_j e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} (\lambda (w_i+1) e^{\lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}+\boldsymbol{\theta}^T \boldsymbol{Z}_i} - e^{\lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}+1)}{(e^{\lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}-1)^2} \end{aligned} \tag{3.2.5}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \theta_j^2} =& -\sum_{i=1}^{n_1} \frac{Z_j^2 \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} (\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i+\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}} - e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}+1)}{(e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}-1)^2} - \sum_{i=1}^{n_1} y_i Z_j^2 \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} - \sum_{i+n_1+1}^{n_1+n_2} u_i Z_j^2 \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} \\ & - \sum_{i=n_1+1}^{n_1+n_2} \frac{Z_j^2 \lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} (\lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i+\lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}} - e^{\lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}+1)}{(e^{\lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}-1)^2} \end{aligned} \tag{3.2.6}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \theta_j \partial \theta_k} =& -\sum_{i=1}^{n_1} \frac{Z_j Z_k \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} (\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i+\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}} - e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}+1)}{(e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}-1)^2} - \sum_{i=i}^{n_1} y_i Z_j Z_k \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} \\ & - \sum_{i=n_1+1}^{n_1+n_2} u_i Z_j Z_k \lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} \\ & - \sum_{i=n_1+1}^{n_1+n_2} \frac{Z_j Z_k \lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i} (\lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i+\lambda (w_i+1) e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}} - e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}+1)}{(e^{\lambda e^{\boldsymbol{\theta}^T \boldsymbol{Z}_i}}-1)^2} \end{aligned} \tag{3.2.7}$$

By substituting the MLE found by using the algorithm above into the information matrix, we obtain the "observed information" matrix, namely the Hessian matrix of the log-likelihood function (see Efron and Hinkley, 1978) as follows

$$\hat{I}_{4\times 4} = \begin{bmatrix} \frac{\partial^2 l}{\partial \lambda^2} & \frac{\partial^2 l}{\partial \lambda \partial \theta_1} & \frac{\partial^2 l}{\partial \lambda \partial \theta_2} & \frac{\partial^2 l}{\partial \lambda \partial \theta_3} \\ \frac{\partial^2 l}{\partial \theta_1 \partial \lambda} & \frac{\partial^2 l}{\partial \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_3} \\ \frac{\partial^2 l}{\partial \theta_2 \partial \lambda} & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_2^2} & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_3} \\ \frac{\partial^2 l}{\partial \theta_3 \partial \lambda} & \frac{\partial^2 l}{\partial \theta_3 \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_3 \partial \theta_2} & \frac{\partial^2 l}{\partial \theta_3^2} \end{bmatrix}$$

where $\lambda = \hat{\lambda}$ and $\theta_i = \hat{\theta}_i$. Hence $\hat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ is asymptotically Normal with mean zero and covariance $I(\boldsymbol{\theta})^{-1}$. This large sample approximation can be used to construct the required confidence intervals, as we do in Section 3.2 and the ensuing illustration.

3.3 A practical example

In this section, the proposed techniques are applied to a Stanford Heart Transplant study in Crowley and Hu (1977) as in Section (5.4). We take the variables namely *age* at acceptance in years (Z_1) and prior *surgery* (Z_2) as the covariates with respective regression coefficients θ_1 and θ_2.

For the complete data set it is observed that the maximum likelihood estimates of λ, θ_1 and θ_2 are 0.00824, 0.02231 and -0.40420 respectively. In order to create a set of middle-censored data, we randomly choose several actual failure data and replace them by random censoring intervals. The data were censored by a random interval whose left end was a geometric random variable with mean 300 and the width was geometric with mean 1000. It is found that 21.51$\%$ of data were censored resulting in 135 uncensored observations and 37 censored observations. Applying the model given in Section 3.2, it is found that the estimates of the regression coefficients are $\hat{\lambda} = 0.00737, \hat{\theta}_1 = 0.01989$ and $\hat{\theta}_2 = -0.62468$. The 95$\%$ confidence intervals based on the asymptotic distribution of λ, θ_1 and θ_2 are (0.00605,0.00870), (0.00246,0.03731) and (-1.07574,-0.17361) respectively. These results are consistent with what the other studies discovered.

In order to assess how much of a change it makes in the estimates or confidence intervals when one uses the discretized geometric distribution in lieu of the original exponential distribution, we fit this model with the exponential distribution instead of the geometric distribution. The estimates of the regression coefficients are $\hat{\lambda} = 0.00736, \hat{\theta}_1 = 0.02714$ and $\hat{\theta}_2 = -0.77792$. The 95% confidence intervals based on the asymptotic distribution of λ, θ_1 and θ_2 are (0.00726,0.00747), (0.00840, 0.04587) and (-1.15766,-0.39818) respectively. The data were censored exactly like the geometric case resulting in 19.77% censored observations, specifically 138 uncensored observations and 34 censored observations. These comparisons show that the estimates are very close as are the confidence intervals.

Chapter 4

Competing risks models for middle censored data

4.1 Introduction

Standard survival analysis focuses on failure-time data that have a single type of failure. Competing risks arise when a failure can result from one of several causes and one cause precludes the others (Andersen et al., 1995; Klein, 2010; Klein and Moeschberger, 2003; Marubini and Valsecchi, 1995; Pintilie, 2006). An investigator is often interested in the assessment of a specific risk in the presence of other risk factors. Examples of competing risks can occur in many fields, including public health, reliability analysis, and demography. For example, a person can die from lung cancer or from a stroke, but not from both (although he could have both lung cancer and atherosclerosis before he dies). A logical objective for competing risks data is to assess the relationship of relevant predictors to the failure rate or corresponding survival probability of any one of the possible events allowing for the competing risks of the other ways to fail.

In analyzing the competing risks model, it is assumed that data consists of a failure time and an indicator denoting the cause of failure. Several studies have carried out under this assumption for both the parametric and the nonparametric set up. For the parametric set up, it is assumed that different lifetime distributions follow some special parametric distribution such as exponential, Weibull and gamma. Berkson and Elveback (1960), Cox (1959), David and Moeschberger (1978) considered this problem from the parametric point of view. Kaplan and Meier (1958), Efron (1967) and Peterson (1977) analyzed the nonparametric version of this model. Miyawaka (1984) considered a model where the failure time of that item/individual is observed but the corresponding cause of failure is not observed.

In classical competing risks, the observed outcome comprises of the time to failure, T and the cause or type of failure, C. The failure time T is taken to be continuous and the cause C takes one of a number of values labeled $1, 2, \cdots, k$. The probability framework is a bivariate distribution in which the component C is discrete and T is continuous. It is assumed that to every failure, only one cause is assigned from the given set of k causes. For data with competing risks, the cause-specific hazard function denoted by $h_j(t)$, describe the rate of failure at time t for event type j given that the follow up time is at least t is

$$h_j(t) = lim_{\triangle t \to 0} \frac{P(t < T < t + \triangle t, C = j | T \geq t)}{\triangle t}$$

If the failure at time t for event type j is assumed to be mutually exclusive and that follow up terminates from one and only one event, then the overall hazard function is

$$h(t) = \sum_{j=1}^{k} h_j(t)$$

It follows that the overall, event free, survival function is

$$S(t) = exp(-H(t)) = exp\left(-\sum_{j=1}^{k} H_j(t)\right) = exp\left(-\sum_{j=1}^{k} \int_0^t h_j(u)du\right)$$

The cumulative incidence function denoted by $F_j(t)$ is the probability of failure due to cause j prior to time t

$$F_j(t) = P(T \leq t, C = j) = \int_0^t h_j(u)S(u)du$$

where $k = 1, 2, \cdots, K$.

Consider a life test with n items with two competing risks causes which are independently distributed with cumulative distribution functions $F_1(t), F_2(t)$ and the related probability density functions $f_1(t), f_2(t)$ respectively.

Suppose the total competing risks middle censored data is expressed as

$$\{(T_1, \delta_1), (T_2, \delta_2), \cdots, (T_{n_1}, \delta_{n_1}), [(L_{n_1+1}, R_{n_1+1}), \delta_{n_1+1}], \cdots, [(L_{n_1+n_2}, R_{n_1+n_2}), \delta_{n_1+n_2}]\}$$

where $n_1 + n_2 = n$ and δ_i is the indicator function defined as

$$\delta_i = \begin{cases} 1 & \text{if failure is due to cause 1} \\ 2 & \text{if failure is due to cause 2} \end{cases}$$

The joint likelihood function of failure is given by

$$L(t) = \prod_{i=1}^{n_1} [f_1(t_i)\bar{F}_2(t_i)]^{I(\delta_i=1)} [f_2(t_i)\bar{F}_1(t_i)]^{I(\delta_i=2)} \prod_{j=1}^{2} \prod_{i=n_1+1}^{n} [F_j(R_i) - F_j(L_i)]^{I(\delta_i=j)}$$

where $I(\cdot)$ denotes the indicator function. Note that $\sum_{j=1}^{2} I(\delta_i = j) = 1$ for $i = 1, 2, \cdots, n$, $\sum_{i=1}^{n} I(\delta_i = j)$ denotes the number of failure by cause j, $j = 1, 2$ and $\sum_{j=1}^{2} \sum_{i=1}^{n} I(\delta_i = j) = n$.

4.2 Parametric models

Lifetimes that follow an exponential distribution and Weibull distribution are considered in this section. Under the same censoring mechanism, theorem (4.1) shows that the EM algorithm will converge for these distributions.

Theorem 4.1. *Let*

$$\{(x_1,\delta_1),(x_2,\delta_2),\cdots,(x_{n_1},\delta_{n_1}),[(L_{n_1+1},R_{n_1+1}),\delta_{n_1+1}],\cdots,[(L_{n_1+n_2},R_{n_1+n_2}),\delta_{n_1+n_2}]\}$$

be the the observed middle-censored data where δ_i is the indicator function defined as

$$\delta_i = \begin{cases} 1 & \textit{if failure is due to cause 1} \\ \\ 2 & \textit{if failure is due to cause 2} \end{cases}$$

from a continuous exponential family distribution

$$f_s(x|\phi) = h_s(x)c_s(\phi)exp\left[\sum_{j=1}^{k} w_{sj}(\phi)t_{sj}(x)\right]$$

such that $h(x), t_j(x), c(\phi)$ and $w_j(\phi)$ are all continuous functions and $s = 1, 2$. Then the EM algorithm will converge for this data.

Proof. The complete log-likelihood is given as

$$\begin{aligned} l(\phi) \propto & \sum_{i=1}^{n_1} I(\delta_i = 1)ln[f_1(x_i)\bar{F_2}(x_i)] + \sum_{i=1}^{n_1} I(\delta_i = 2)ln[f_2(x_i)\bar{F_1}(x_i)] \\ & + \sum_{i=n_1+1}^{n} I(\delta_i = 1)ln[f_1(x_i)\bar{F_2}(x_i)] + \sum_{i=n_1+1}^{n} I(\delta_i = 2)ln[f_2(x_i)\bar{F_1}(x_i)] \end{aligned}$$

where $n_1 + n_2 = n$.

The conditional expectation $E[t_{sj}(x_i)|\phi^*, a_i < x_i < b_i]$ are continuous functions by Lemma (6) of Bennett (2011) since each function is continuous in each argument. Note that $F_s(x|\phi)$ is a continuous function by the same lemma.

Hence $E[l(\phi|\text{complete data})|\phi^*, \text{censored data}]$ is a continuous function in both ϕ and ϕ^*. Thus, the EM algorithm will converge by Theorem 2 of Wu (1983).

□

4.2.1 Exponential distributed lifetimes

Consider that the lifetime distributions T_1 and T_2 due to cause 1 and cause 2 are both exponential distributed with CDFs and PDFs

$$F_j(t;\theta_j) = 1 - e^{-\theta_j t} \text{ and } f_j(t;\theta_j) = \theta_j e^{-\theta_j t}$$

respectively where $\theta_j > 0, j = 1, 2$ are unknown parameters and $t > 0$.

Again, we can reorder the data into the uncensored and censored observations. Hence, our observed data is

$$\{(T_1,\delta_1),(T_2,\delta_2),\cdots,(T_{n_1},\delta_{n_1}),[(L_{n_1+1},R_{n_1+1}),\delta_{n_1+1}],\cdots,[(L_{n_1+n_2},R_{n_1+n_2}),\delta_{n_1+n_2}]\}$$

where $n_1 + n_2 = n$. Now, the likelihood function can be written as

$$L(t;\theta) \propto \prod_{i=1}^{n_1} \theta_1^{I(\delta_i=1)}\theta_2^{I(\delta_i=2)} \prod_{i=1}^{n_1}(e^{-t_i\theta_1 - t_i\theta_2}) \prod_{j=1}^{2}\prod_{i=n_1+1}^{n} (e^{-l_i\theta_j} - e^{-r_i\theta_j})^{I(\delta_i=j)} \tag{4.2.1}$$

Based on equation (4.2.1), the corresponding log-likelihood is

$$l(\theta_1, \theta_2) \propto \sum_{i=1}^{n_1} I(\delta_i = 1) ln\theta_1 + \sum_{i=1}^{n_1} I(\delta_i = 2) ln\theta_2 - \sum_{i=1}^{n_1} (t_i\theta_1 + t_i\theta_2) + \sum_{j=1}^{2} \sum_{i=n_1+1}^{n} ln[e^{-L_i\theta_j} - e^{-R_i\theta_j}]^{I(\delta_i=j)} \tag{4.2.2}$$

The EM algorithm is used to find the MLEs of the parameters involved. In the E-step, the conditional expectation needed for the incomplete data is given as

$$E[T|L < T < R] = \frac{\int_L^R t\theta_1^{I(\delta_i=1)}\theta_2^{I(\delta_i=2)}(e^{-t\theta_1 - t\theta_2})dt}{F(R|\theta_1, \theta_2) - F(L|\theta_1, \theta_2)} \tag{4.2.3}$$

Thus the required log-likelihood is

$$l(\theta_1, \theta_2) = \sum_{i=1}^{n} I(\delta_i = 1) ln\theta_1 + \sum_{i=1}^{n} I(\delta_i = 1) ln\theta_2 - \sum_{i=1}^{n_1} (\theta_1 t_i + \theta_2 t_i) - \sum_{i=n_1+1}^{n} (\theta_1 t_i^* + \theta_2 t_i^*) \tag{4.2.4}$$

where the t^* is found using Equation (4.2.3).

The EM Algorithm is used to solve for the MLE's of θ_1 and θ_2. The steps involved in algorithm are as follows

- Step 1: Suppose that $\theta_j^{(0)}, j = 1, 2$ are the initial guess values of the maximum likelihood estimate of θ_j.

- Step 2: Compute T_i^* using equation (4.2.3) with $(\theta_1, \theta_2) = (\theta_1, \theta_2)_{(j)}$

- Step 3: Solve equation (4.2.4) for its maximum and set $(\theta_1, \theta_2)_{(j+1)}$ as the maximum

- Step 4: Repeat until a convergence criterion is met

$(c_1,d_1),(c_2,d_2)$	n	$\hat{\theta}_1$	$\hat{\theta}_2$	$Cens_1$	$Cens_2$
(1,1.2) , (1,1.2)	50	1.1379	1.1032	0.12	0.11
	$EMSE$	0.1017	0.1517		
	100	1.0859	1.0782	0.11	0.11
	$EMSE$	0.0377	0.0652		
	250	1.0467	1.0336	0.11	0.13
	$EMSE$	0.0160	0.0219		
	500	1.0214	1.0162	0.11	0.11
	$EMSE$	0.0119	0.0114		
$(0.5,1.5),(0.5,1.5)$	50	1.1312	1.1414	0.06	0.07
	$EMSE$	0.1987	0.2589		
	100	1.0913	1.0831	0.07	0.08
	$EMSE$	0.1890	0.2178		
	250	1.0626	1.0388	0.06	0.07
	$EMSE$	0.1094	0.0251		
	500	1.0313	1.0027	0.07	0.07
	$EMSE$	0.0099	0.0098		
$(1,2),(1,2)$	50	1.1211	1.0705	0.07	0.09
	$EMSE$	0.1835	0.1258		
	100	1.0634	1.0378	0.08	0.09
	$EMSE$	0.0813	0.0545		
	250	1.0504	1.0244	0.09	0.09
	$EMSE$	0.0284	0.0444		
	500	1.0151	1.0093	0.09	0.07
	$EMSE$	0.0119	0.0065		

Table 4.1: Numerical results for Exponential model

A simulation study was performed to illustrate the usefulness of the approach. For each sample size $n, N = 50$ replications were simulated. Each sample was censored and the above EM algorithm was applied to the censored data. The censoring mechanism is as follows; the left endpoint and right endpoint of the censored interval for cause i are exponentially distributed with mean c_i and d_i respectively for $i = 1, 2$. The true values for the simulation are $\theta_1 = 1, \theta_2 = 1$. The MLE's of θ_1 and θ_2 were computed using the EM algorithm and the results are shown in Table 4.1. The estimates of all the parameters in all cases are close to the actual value and they are converging to the true values as the sample size, n increases. The estimated mean squared errors *(EMSE)* for the estimates are small. Hence, this approach yields very useful, accurate and reliable results.

4.2.2 Weibull distributed lifetimes

In this section, consider that the lifetime distributions T_1 and T_2 due to cause 1 and cause 2 are both Weibull distributed with CDFs and PDFs

$$F_j(t; a_j, b_j) = 1 - e^{-b_j t^{a_j}}$$

$$f_j(t; a_j, b_j) = a_j b_j t^{a_j - 1} e^{-b_j t^{a_j}}$$

respectively where $a_j > 0, \frac{1}{b_j} > 0, j = 1, 2$ are unknown parameters and $t > 0$.

The likelihood function is written as

$$\begin{aligned} L(t; a, b) \propto & \prod_{i=1}^{n_1} [a_1 b_1 t_i^{a_1 - 1} e^{-b_1 t^{a_1}} e^{-b_2 t^{a_2}}]^{I(\delta_i = 1)} [a_2 b_2 t_i^{a_2 - 1} e^{-b_2 t^{a_2}} e^{-b_1 t^{a_1}}]^{I(\delta_i = 2)} \\ & \prod_{j=1}^{2} \prod_{i=n_1+1}^{n} (e^{-b_j l_i^{a_j}} - e^{-b_j r_i^{a_j}})^{I(\delta_i = j)} \end{aligned}$$

where the corresponding log-likelihood is

$$\begin{aligned} l(a_1, b_1, a_2, b_2) \propto & \sum_{i=1}^{n_1} I(\delta_i = 1)[ln(a_1 b_1) + (a_1 - 1) ln(t_i)] - \sum_{i=1}^{n_1} (b_1 t_i^{a_1} - b_2 t_i^{a_2}) \\ & + \sum_{i=1}^{n_1} I(\delta_i = 2)[ln(a_2 b_2) + (a_2 - 1) ln(t_i)] \\ & + \sum_{j=1}^{2} \sum_{i=n_1+1}^{n} I(\delta_i = j) ln[exp(-b_j l_i^{a_j}) - exp(-b_j r_i^{a_j})] \end{aligned} \tag{4.2.5}$$

We need to use the EM algorithm to find the MLEs of the parameters involved. In the E-step, the conditional expectations for the incomplete data are given as

$$E[T^{a_j}|L<T<R] = \frac{\int_L^R t^{a_j}(a_1b_1t^{a_1-1}e^{-b_1t^{a_1}-b_2t^{a_2}})^{I(\delta_i=1)}(a_2b_2t^{a_2-1}e^{-b_2t^{a_2}-b_1t^{a_1}})^{I(\delta_i=2)}dt}{F(R|a_1,b_1,a_2,b_2)-F(L|a_1,b_1,a_2,b_2)} \tag{4.2.6}$$

$$E[ln(T)|L<T<R] = \frac{\int_L^R ln(t)(a_1b_1t^{a_1-1}e^{-b_1t^{a_1}-b_2t^{a_2}})^{I(\delta_i=1)}(a_2b_2t^{a_2-1}e^{-b_2t^{a_2}-b_1t^{a_1}})^{I(\delta_i=2)}dt}{F(R|a_1,b_1,a_2,b_2)-F(L|a_1,b_1,a_2,b_2)} \tag{4.2.7}$$

Then the required log-likelihood is

$$\begin{aligned} l(a_1,b_1,a_2,b_2) = &\sum_{i=1}^{n} I(\delta_i=1)ln(a_1b_1) + \sum_{i=1}^{n} I(\delta_i=2)ln(a_2b_2) + (a_1-1)\sum_{i=1}^{n_1} I(\delta_i=1)ln(t_i) \\ &+ (a_1-1)\sum_{i=n_1+1}^{n} I(\delta_i=1)ln(t_i)^* + (a_2-1)\sum_{i=1}^{n_1} I(\delta_i=2)ln(t_i) \\ &+ (a_2-1)\sum_{i=n_1+1}^{n} I(\delta_i=2)ln(t_i)^* - \sum_{i=1}^{n_1}(b_1t_i^{a_1}+b_2t_i^{a_2}) - \sum_{i=n_1+1}^{n}(b_1t_i^{a_1*}+b_2t_i^{a_2*}) \end{aligned} \tag{4.2.8}$$

where the $t_i^{a_j}$ and $ln(t_i)^*$ are given in Equations (4.2.6) and (4.2.7) respectively.

The EM algorithm can now be set up as follows. To make this more explicit, choose $(a_1,b_1,a_2,b_2)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(a_1,a_2,b_1,b_2)_{(j)}$ is the jth estimate
- Step 2: Compute equations (4.2.6) and (4.2.7) with $(a_1,b_1,a_2,b_2) = (a_1,b_1,a_2,b_2)_{(j)}$
- Step 3: Solve equation (4.2.8) for its maximum and set $(a_1,b_1,a_2,b_2)_{(j+1)}$ as that maximum
- Step 4: Repeat until a convergence criterion is met

A simulation study was performed to illustrate the usefulness of the approach. For each sample size $n, N = 50$ replications were simulated. Each sample was censored and the above EM algorithm was applied to the censored data. The censoring mechanism is as follows; the left endpoint and right endpoint of the censored interval for cause i are exponentially distributed with mean c_i and d_i respectively for $i = 1, 2$. The true values for the simulation are $a_1 = 2, b_1 = 1, a_2 = 2$ and $b_2 = 1$. The MLE's of a_1, b_1, a_2 and b_2 were computed using the EM algorithm and the results are shown in Table 4.2.

$(c_1, d_1), (c_2, d_2)$	n	$\hat{a}_1$	$\hat{b}_1$	$\hat{a}_2$	$\hat{b}_2$	$Cens_1$	$Cens_2$
$(1, 1), (1, 1)$	50	2.2541	0.8505	2.2417	0.9088	0.12	0.11
	$EMSE$	*0.0020*	*0.0014*	*0.0018*	*0.0006*		
	100	2.2015	0.8819	2.1822	0.9119	0.19	0.12
	$EMSE$	*0.0012*	*0.0009*	*0.0010*	*0.0005*		
	250	2.1992	0.9124	2.1461	0.9322	0.13	0.11
	$EMSE$	*0.0012*	*0.0005*	*0.0006*	0.0003		
	500	2.1106	0.9671	2.0835	0.9621	0.12	0.13
	$EMSE$	*0.0004*	*0.0001*	*0.0002*	*0.0001*		
$(1, 0.5), (1, 0.5)$	50	2.2299	0.9119	2.2956	0.9110	0.24	0.18
	$EMSE$	*0.0017*	*0.0005*	*0.0027*	*0.0004*		
	100	2.2173	0.9280	2.2502	0.9378	0.20	0.18
	$EMSE$	*0.0016*	*0.0003*	*0.0021*	*0.0002*		
	250	2.1926	0.9948	2.1698	0.9711	0.20	0.21
	$EMSE$	*0.0014*	*0.0000*	*0.0009*	*0.0000*		
	500	2.1653	0.9998	2.1622	0.9964	0.19	0.22
	$EMSE$	*0.0008*	*0.0000*	*0.0008*	*0.0000*		
$(2, 1), (2, 1)$	50	2.2450	0.9012	1.9552	0.9003	0.22	0.24
	$EMSE$	*0.0013*	*0.0009*	*0.0009*	*0.0006*		
	100	2.1279	0.9340	2.1236	0.9191	0.26	0.24
	$EMSE$	*0.0005*	*0.0003*	*0.0005*	*0.0004*		
	250	2.1218	0.9434	2.1236	0.9400	0.20	0.24
	$EMSE$	*0.0004*	*0.0002*	*0.0005*	*0.0003*		
	500	2.0498	0.9564	2.0992	0.9667	0.21	0.22
	$EMSE$	*0.0001*	*0.0001*	*0.0003*	*0.0001*		

Table 4.2: Numerical results for Weibull model

The estimates of all the parameters in all cases are close to the actual value and they are converging to the true values as the sample size, n increases, but it appears to be converging slowly. It also performs reasonably well even with a large proportion of censored observations and the estimated mean squared errors *(EMSE)* for the estimates are small. Hence, this approach yields very useful, accurate and reliable results.

4.3 Parametric models in the presence of covariates

Lifetimes that follow an exponential AFT model and Weibull AFT model are considered in this section. It is shown in Theorem (4.2) that under the same censoring mechanism, the EM-algorithm will converge for these distributions. Note that the exponential AFT model is a special case of the Weibull AFT model. At this stage of development, we only deal with one single covariate Z.

Theorem 4.2. *Let*

$$\{(x_1,\delta_1),(x_2,\delta_2),\cdots,(x_{n_1},\delta_{n_1}),[(L_{n_1+1},R_{n_1+1}),\delta_{n_1+1}],\cdots,[(L_{n_1+n_2},R_{n_1+n_2}),\delta_{n_1+n_2}]\}$$

be the the observed middle-censored data from a Weibull Accelerated Failure Time model where δ_i is the indicator function defined as

$$\delta_i = \begin{cases} 1 & \textit{if failure is due to cause 1} \\ 2 & \textit{if failure is due to cause 2} \end{cases}$$

Then the EM algorithm will converge for this data.

Proof. The complete log-likelihood is given as

$$\begin{aligned}
l(a_j, b_j, \theta_j) = & \sum_{i=1}^{n_1} I(\delta_i = 1) log(a_1 b_1) + a_1 \sum_{i=1}^{n_1} I(\delta_i = 1)\theta_1 Z_i \\
& + (a_1 - 1) \sum_{i=1}^{n_1} I(\delta_i = 1) ln(t_i) - b_1 \sum_{i=1}^{n_1} I(\delta_i = 1) t_i^{a_1} exp(a_1 \theta_1 Z_i) \\
& - b_2 \sum_{i=1}^{n_1} I(\delta_i = 1) t_i^{a_2} exp(a_2 \theta_2 Z_i) + \sum_{i=1}^{n_1} I(\delta_i = 2) log(a_2 b_2) \\
& + a_2 \sum_{i=1}^{n_1} I(\delta_i = 2)\theta_2 Z_i + (a_2 - 1) \sum_{i=1}^{n_1} I(\delta_i = 2) ln(t_i) \\
& - b_2 \sum_{i=1}^{n_1} I(\delta_i = 2) t_i^{a_2} exp(a_2 \theta_2 Z_i) - b_1 \sum_{i=1}^{n_1} I(\delta_i = 2) t_i^{a_1} exp(a_1 \theta_1 Z_i) \\
& + (a_1 - 1) \sum_{i=n_1+1}^{n} I(\delta_i = 1) ln(t_i) - b_1 \sum_{i=n_1}^{n} I(\delta_i = 1) t_i^{a_1} exp(a_1 \theta_1 Z_i) \\
& - b_2 \sum_{i=n_1+1}^{n} I(\delta_i = 1) t_i^{a_2} exp(a_2 \theta_2 Z_i) + (a_2 - 1) \sum_{i=n_1+1}^{n} I(\delta_i = 2) ln(t_i) \\
& - b_2 \sum_{i=n_1+1}^{n} I(\delta_i = 2) t_i^{a_2} exp(a_2 \theta_2 Z_i) - b_1 \sum_{i=n_1+1}^{n} I(\delta_i = 2) t_i^{a_1} exp(a_1 \theta_1 Z_i)
\end{aligned}$$

The conditional expectations

$$\begin{aligned}
& E[ln(t) | a_j^*, b_j^*, \theta_j^*, l < t < r] = \\
& \quad \int_l^r ln(t) \left[a_1^* b_1^* exp(a_1^* \theta_1^* Z_i) t^{a_1^* - 1} exp\left(-b_1^* e^{a_1^* \theta_1^* Z_i} t^{a_1^*}\right) exp\left(-b_2^* e^{a_2^* \theta_2^* Z_i} t^{a_2^*}\right)\right]^{I(\delta_i = 1)} \\
& \quad \times \left[a_2^* b_2^* exp(a_2^* \theta_2^* Z_i) t^{a_2^* - 1} exp\left(-b_2^* e^{a_2^* \theta_2^* Z_i} t^{a_2^*}\right) exp\left(-b_1^* e^{a_1^* \theta_1^* Z_i} t^{a_1^*}\right)\right]^{I(\delta_i = 2)} dt
\end{aligned}$$

$$\begin{aligned}
& E[t^{a_j} exp(a_j \theta_j Z_i) | a_j^*, b_j^*, \theta_j^*, l < t < r] = \\
& \int_l^r t^{a_j^*} exp(a_j^* \theta_j^* Z_i) \left[a_1^* b_1^* exp(a_1^* \theta_1^* Z_i) t^{a_1^* - 1} exp\left(-b_1^* e^{a_1^* \theta_1^* Z_i} t^{a_1^*}\right) exp\left(-b_2^* e^{a_2^* \theta_2^* Z_i} t^{a_2^*}\right)\right]^{I(\delta_i = 1)} \\
& \times \left[a_2^* b_2^* exp(a_2^* \theta_2^* Z_i) t^{a_2^* - 1} exp\left(-b_2^* e^{a_2^* \theta_2^* Z_i} t^{a_2^*}\right) exp\left(-b_1^* e^{a_1^* \theta_1^* Z_i} t^{a_1^*}\right)\right]^{I(\delta_i = 2)} dt
\end{aligned}$$

are both continuous functions by Lemma 6 of Bennett (2011) since each function is continuous is its argument(s). Hence, $E[l(\phi, \theta)|\text{complete data}|\phi^*, \theta^*, \text{censored data}]$ is continuous in $a_1, a_1^*, b_1, b_1^*, a_2, a_2^*, b_2, b_2^*, \theta_1, \theta_1^*, \theta_2$ and θ_2^*. Thus, the EM algorithm will converge by Theorem 2 in Wu (1983).

□

4.3.1 Exponential Accelerated Failure Time model

Consider that the lifetime distributions T_1 and T_2 due to cause 1 and cause 2 are both exponential Accelerated Failure Time distributed with CDFs and PDFs,

$$F_j(t; a_j, \theta_j) = 1 - exp(-a_j e^{\theta_j^T Z} t)$$

$$f_j(t; a_j, \theta_j) = a_j e^{\theta_j^T Z} exp(-a_j e^{\theta_j Z} t)$$

respectively where $\theta_j, j = 1, 2$ are unknown parameters and $t > 0$. The censoring mechanism is as follows. The left point of the censored interval is an exponential random variable with mean $1/c$ and the length of the censored interval is an exponentially distributed with mean $1/d$.

As in Section (4.2.1), reorder the data into the uncensored and censored observations as follows.

$$\{(T_1, \delta_1), (T_2, \delta_2), \cdots, (T_{n_1}, \delta_{n_1}), [(L_{n_1+1}, R_{n_1+1}), \delta_{n_1+1}], \cdots, [(L_{n_1+n_2}, R_{n_1+n_2}), \delta_{n_1+n_2}]\}$$

where $n_1 + n_2 = n$. Now, the likelihood function is

$$L(a_1, a_2, \theta_1, \theta_2) = \prod_{i=1}^{n_1} \{a_1^{I(\delta_i=1)} a_2^{I(\delta_i=2)} \left[e^{\theta_1 Z_i} exp(-a_1 e^{\theta_1 Z_i} t_i) exp(-a_2 e^{\theta_2 Z_i} t_i)\right]^{I(\delta_i=1)}$$
$$\times \left[e^{\theta_2 Z_i} exp(-a_2 e^{\theta_2 Z_i} t_i) exp(-a_1 e^{\theta_1 Z_i} t_i)\right]^{I(\delta_i=2)}\}$$
$$\times \prod_{j=1}^{2} \prod_{i=n_1+1}^{n} (S_j(L_i) - S_j(R_i))^{I(\delta_i=j)} \tag{4.3.1}$$

The corresponding log-likelihood is

$$l(a_1, a_2, \theta_1, \theta_2) = \sum_{i=1}^{n_1} I(\delta_i = 1) ln(a_1) + \sum_{i=1}^{n_1} I(\delta_i = 2) ln(a_2)$$
$$+ \sum_{i=1}^{n_1} I(\delta_i = 1)(\theta_1 Z_i - a_1 e^{\theta_1 Z_i} t_i - a_2 e^{\theta_2 Z_i} t_i)$$
$$+ \sum_{i=1}^{n_1} I(\delta_i = 2)(\theta_2 Z_i - a_2 e^{\theta_2 Z_i} t_i - a_1 e^{\theta_1 Z_i} t_i)$$
$$+ \sum_{j=1}^{2} \sum_{i=n_1+1}^{n} I(\delta_i = j) ln \left[exp(-a_j e^{\theta_j^T Z_i} l_i) - exp(-a_j e^{\theta_J^T Z_i} r_i)\right] \tag{4.3.2}$$

The EM algorithm is used to find the MLEs of the parameters involved. The conditional expectation needed for the incomplete data is given as

$$E[T|L < T < R] = \frac{\int_L^R t f(t|a_1, a_2, \theta_1, \theta_2) dt}{F(R|a_1, a_2, \theta_1, \theta_2) - F(L|a_1, a_2, \theta_1, \theta_2)} \tag{4.3.3}$$

The required log-likelihood is

$$l(a_1, a_2, \theta_1, \theta_2) = \sum_{i=1}^{n} I(\delta_i = 1) ln(a_1) + \sum_{i=1}^{n} I(\delta_i = 2) ln(a_2)$$
$$+ \sum_{i=1}^{n_1} I(\delta_i = 1)(\theta_1 Z_i - a_1 e^{\theta_1 Z_i} t_i - a_2 e^{\theta_2 Z_i} t_i)$$
$$+ \sum_{i=n_1+1}^{n} I(\delta_i = 1)(\theta_1 Z_i - a_1 e^{\theta_1 Z_i} t_i^* - a_2 e^{\theta_2 Z_i} t_i^*)$$
$$+ \sum_{i=1}^{n_1} I(\delta_i = 2)(\theta_2 Z_i - a_2 e^{\theta_2 Z_i} t_i - a_1 e^{\theta_1 Z_i} t_i)$$
$$+ \sum_{i=n_1+1}^{n} I(\delta_i = 2)(\theta_2 Z_i - a_2 e^{\theta_2 Z_i} t_i^* - a_1 e^{\theta_1 Z_i} t_i^*) \tag{4.3.4}$$

where the t_i^* is found using equation (4.3.3).

The EM Algorithm is used to solve for the MLE's of a_1, a_2, θ_1 and θ_2 and it is set up as follows

- Step 1: Suppose that $a_j^{(0)}$ and $\theta_j^{(0)}, j = 1, 2$ are the initial guess values of the maximum likelihood estimate of a_j and θ_j respectively.
- Step 2: Compute T_i^* using equation (4.3.3) with $(a_1, a_2, \theta_1, \theta_2) = (a_1, a_2, \theta_1, \theta_2)_{(j)}$
- Step 3: Solve equation (4.3.4) for its maximum and set $(a_1, a_2, \theta_1, \theta_2)_{(j+1)}$ as the maximum
- Step 4: Repeat until convergence criteria is met

A simulation study was performed to illustrate the usefulness of the approach. For each sample size $n, N = 100$ replications were simulated. Each sample was censored and the above EM algorithm was applied to the censored data. The censoring mechanism is as follows; the left end point and right end point of the censored interval for cause i are exponentially distributed with mean c_i and d_i respectively for $i = 1, 2$. The true values for the simulation are $a_1 = 1, a_2 = 1, \theta_1 = 1, \theta_2 = 1$ and the covariate Z is generated from a standard normal distribution. The MLE's of $a_1, a_2, \theta_1, \theta_2$ were computed using the EM algorithm and the results are shown in Table 4.3.

The estimates of all the parameters in all cases are close to the actual value and they are converging to the true values as the sample size, n increases. It performs quite well even with a large proportion of censored observations and the estimated mean squared errors

(*EMSE*) for the estimates are very small. Hence, this approach yields very useful, accurate and reliable results.

$(c_1, d_1), (c_2, d_2)$	n	$\hat{a_1}$	$\hat{a_2}$	$\hat{\theta_1}$	$\hat{\theta_2}$	$Cens_1$	$Cens_2$
$(1, 2), (1, 2)$	50	1.1595	1.4054	1.0193	1.0182	0.29	0.30
	$EMSE$	*0.0004*	*0.0082*	*0.0001*	*0.0002*		
	100	1.0900	1.2984	1.0144	1.0192	0.27	0.29
	$EMSE$	*0.0001*	*0.0045*	*0.0000*	*0.0000*		
	500	1.0885	1.0654	1.0119	1.0076	0.27	0.29
	$EMSE$	*0.0000*	*0.0000*	*0.0000*	*0.0000*		
$(1, 3), (1, 3)$	50	1.2153	1.3204	1.0542	1.0444	0.30	0.30
	$EMSE$	*0.0015*	*0.0051*	*0.0004*	*0.0001*		
	100	1.1543	1.2609	1.0534	1.0367	0.32	0.32
	$EMSE$	*0.0008*	*0.0034*	*0.0001*	*0.0000*		
	500	1.1076	1.1420	1.0326	1.0055	0.33	0.33
	$EMSE$	*0.0000*	*0.0000*	*0.0001*	*0.0000*		
$(3, 5), (3, 5)$	50	1.1080	1.4274	1.0342	1.0221	0.19	0.20
	$EMSE$	*0.0004*	*0.0091*	*0.0001*	*0.0000*		
	100	1.0734	1.1581	1.0115	1.0073	0.20	0.20
	$EMSE$	*0.0002*	*0.0013*	*0.0000*	*0.0000*		
	500	1.0611	1.0940	1.0028	1.0032	0.20	0.20
	$EMSE$	*0.0000*	*0.0000*	*0.0000*	*0.0000*		

Table 4.3: Numerical results for Exponential AFT model

4.3.2 Weibull Accelerated Failure Time model

Consider that the lifetime distributions T_1 and T_2 due to cause 1 and cause 2 are both Weibull Accelerated Failure Time distributed with CDFs and PDFs,

$$F_j(t; a_j, b_j, \theta_j) = 1 - exp(-b_j e^{a_j \theta_j^T Z} t^{a_j})$$

$$f_j(t; a_j, b_j, \theta_j) = a_j b_j exp(a_j \theta_j^T Z) t^{a_j - 1} exp(-b_j e^{a_j \theta_j^T Z} t^{a_j})$$

respectively where $a_j, b_j, \theta_j, j = 1, 2$ are unknown parameters and $t > 0$. Again, we will consider the censoring mechanism as follows; the left point of the censored interval is an exponential random variable with mean $1/c$ and the length of the censored interval is an exponentially distributed with mean $1/d$. It is also assumed that there is at least one censored observation i.e. $n_2 > 0$. The likelihood function is then given as

$$\begin{aligned} L(a_1, a_2, b_1, b_2, \theta_1, \theta_2) = & \prod_{i=1}^{n_1} (a_1 b_1)^{I(\delta_i=1)} (a_2 b_2)^{I(\delta_i=2)} \\ & \times \left[e^{a_1\theta_1 Z_i} t_i^{a_1-1} exp(-b_1 e^{a_1\theta_1 Z_i} t_i^{a_1}) exp(-b_2 e^{a_2\theta_2 Z_i} t_i^{a_2}) \right]^{I(\delta_i=1)} \\ & \times \left[e^{a_2\theta_2 Z_i} t_i^{a_2-1} exp(-b_2 e^{a_2\theta_2 Z_i} t_i^{a_2}) exp(-b_1 e^{a_1\theta_1 Z_i} t_i^{a_1}) \right]^{I(\delta_i=2)} \\ & \times \prod_{j=1}^{2} \prod_{i=n_1+1}^{n} (S_j(L_i) - S_j(R_i))^{I(\delta_i=j)} \end{aligned} \tag{4.3.5}$$

Based on equation (4.3.5), the corresponding log-likelihood is given as

$$\begin{aligned} l(a_1, a_2, b_1, b_2\theta_1, \theta_2) = & \sum_{i=1}^{n_1} I(\delta_i = 1) ln(a_1 b_1) + \sum_{i=1}^{n_1} I(\delta_i = 2) ln(a_2 b_2) \\ & + a_1 \sum_{i=1}^{n_1} I(\delta_i = 1)\theta_1 Z_i + a_2 \sum_{i=1}^{n_1} I(\delta_i = 2)\theta_2 Z_i \\ & + (a_1 - 1) \sum_{i=1}^{n_1} I(\delta_i = 1) ln(t_i) + (a_2 - 1) \sum_{i=1}^{n_1} I(\delta_i = 2) ln(t_i) \\ & - b_1 \sum_{i=1}^{n_1} e^{a_1\theta_1 Z_i} t_i^{a_1} - b_2 \sum_{i=1}^{n_1} e^{a_2\theta_2 Z_i} t_i^{a_2} \\ & + \sum_{j=1}^{2} \sum_{i=n_1+1}^{n} I(\delta_i = j) ln \left[exp(-b_j e^{a_j\theta_j^T Z_i} l_i^{a_j}) - exp(-b_j e^{a_j\theta_j^T Z_i} r_i^{a_j}) \right] \end{aligned} \tag{4.3.6}$$

Applying the EM algorithm in the same fashion as Section (4.3.1), the conditional expectation needed is given as

$$
\begin{aligned}
&E[ln(T)|L < T < R] = \\
&\quad \frac{\int_L^R ln(t)(a_1b_1e^{a_1\theta_1Z_i}t_i^{a_1-1})^{I(\delta_i=1)}(a_2b_2e^{a_2\theta_2Z_i}t_i^{a_2-1})^{I(\delta_i=2)}e^{-b_1e^{a_1\theta_1Z_i}t_i^{a_1}}e^{-b_2e^{a_2\theta_2Z_i}t_i^{a_2}}dt}{F(R|a_1, a_2b_1, b_2, \theta_1, \theta_2) - F(L|a_1, a_2, b_1, b_2\theta_1, \theta_2)}
\end{aligned}
\tag{4.3.7}
$$

$$
\begin{aligned}
&E[T^{a_j}|a_1^*, a_2^*, b_1^*, b_2^*, \theta_1^*, \theta_2^*, L < T < R] = \\
&\quad \frac{\int_L^R t^{a_j}(a_1b_1e^{a_1\theta_1Z_i}t_i^{a_1-1})^{I(\delta_i=1)}(a_2b_2e^{a_2\theta_2Z_i}t_i^{a_2-1})^{I(\delta_i=2)}e^{-b_1e^{a_1\theta_1Z_i}t_i^{a_1}}e^{-b_2e^{a_2\theta_2Z_i}t_i^{a_2}}dt}{F(R|a_1, a_2b_1, b_2, \theta_1, \theta_2) - F(L|a_1, a_2, b_1, b_2\theta_1, \theta_2)}
\end{aligned}
\tag{4.3.8}
$$

Then the required log-likelihood is

$$
\begin{aligned}
l(a_1, a_2,&b_1, b_2, \theta_1, \theta_2) = \\
&\sum_{i=1}^{n} I(\delta_i = 1)ln(a_1b_1) + \sum_{i=1}^{n} I(\delta_i = 2)ln(a_2b_2) + a_1\sum_{i=1}^{n} I(\delta_i = 1)\theta_1 Z_i \\
&+ a_2\sum_{i=1}^{n} I(\delta_i = 2)\theta_2 Z_i + (a_1 - 1)\left[\sum_{i=1}^{n_1} I(\delta_i = 1)ln(t_i) + \sum_{i=n_1+1}^{n} I(\delta_i = 1)ln(t_i)^*\right] \\
&+ (a_2 - 1)\left[\sum_{i=1}^{n_1} I(\delta_i = 2)ln(t_i) + \sum_{i=n_1+1}^{n} I(\delta_i = 2)ln(t_i)^*\right] \\
&- b_1\left[\sum_{i=1}^{n_1} e^{a_1\theta_1Z_i}t_i^{a_1} + \sum_{n_1+1}^{n} e^{a_1\theta_1Z_i}t_i^{a_1*}\right] - b_2\left[\sum_{i=1}^{n_1} e^{a_2\theta_2Z_i}t_i^{a_2} + \sum_{n_1+1}^{n} e^{a_2\theta_2Z_i}t_i^{a_2*}\right]
\end{aligned}
\tag{4.3.9}
$$

where $ln(t)^*$ and t^{a*} are found using equations (4.3.7) and (4.3.8) respectively.

The EM algorithm used to solve for the MLE's of $a_1, b_1a_2, b_2, \theta_1$ and θ_2 can now be set up as follows.

- Step 1: Suppose that $a_j^{(0)}, b_j^{(0)}$ and $\theta_j^{(0)}, j = 1, 2$ are the initial guess values of the maximum likelihood estimate of a_j, b_j and θ_j respectively.

- Step 2: Compute equations (4.3.7) and (4.3.8) with

  $(a_1, b_1, a_2, b_2, \theta_1, \theta_2) = (a_1, b_1, a_2, b_2, \theta_1, \theta_2)_{(j)}$

- Step 3: Solve equation (4.3.9) for its maximum and set $(a_1, b_1, a_2, b_2, \theta_1, \theta_2)_{(j+1)}$ as the maximum

- Step 4: Repeat until convergence criteria is met

A simulation study was performed to illustrate the usefulness of the approach. For each sample size $n, N = 100$ replications were simulated. Each sample was censored and the above EM algorithm was applied to the censored data. The censoring mechanism is as follows; the left end point and right end point of the censored interval for cause i are exponentially distributed with mean c_i and d_i respectively for $i = 1, 2$. The true values for the simulation are $a_1 = 2, b_1 = 1, a_2 = 2, b_2 = 1, \theta_1 = 1, \theta_2 = 1$ and the covariate Z is generated from a standard normal distribution. The MLE's of $a_1, b_1, a_2, b_2, \theta_1, \theta_2$ were computed using the EM algorithm and the results are reported in Table 4.4. The estimates of all the parameters in all cases are close to the actual value and they are converging to the true values as the sample size, n increases. Again the estimated mean squared errors (*EMSE*) for the estimates are found to be very small. Hence, this method yields reliable results with reasonable accuracy for the Weibull AFT model.

Although Theorem (4.2) shows that the EM algorithm will converge, one needs to confirm that the convergence is the global maximum. To that end, one should run this algorithm many times with different initial values for $(a_1, b_1, a_2, b_2, \theta_1, \theta_2)$ in our case, and check that the algorithm is not trapped at local extrema, as suggested by Bennett (2011).

(c_1, d_1) & (c_2, d_2)	n	$\hat{a_1}$	$\hat{b_1}$	$\hat{a_2}$	$\hat{b_2}$	$\hat{\theta_1}$	$\hat{\theta_2}$	$Cens_1$	$Cens_2$
$(1, 2)$	50	1.9320	0.8911	1.9662	0.9011	1.0529	1.0361	16.06	16.07
&	$EMSE$	*0.0011*	*0.0012*	*0.0008*	*0.0010*	*0.0010*	*0.0011*		
$(1, 2)$	100	1.9554	0.9219	1.9700	0.9219	1.0361	1.0265	17.40	16.00
	$EMSE$	*0.0009*	*0.0007*	*0.0006*	*0.0008*	*0.0008*	*0.0009*		
	250	1.9622	0.9422	1.9798	0.9518	1.0222	1.0197	13.90	16.90
	$EMSE$	*0.0006*	*0.0004*	*0.0003*	*0.0005*	*0.0006*	*0.0008*		
	500	1.9929	0.9699	1.9910	0.9789	1.0110	1.0187	19.16	13.64
	$EMSE$	*0.0004*	*0.0001*	*0.0001*	*0.0002*	*0.0004*	*0.0005*		
$(2, 3)$	50	1.9494	0.9012	1.9529	0.9281	0.9644	0.9440	14.40	7.40
&	$EMSE$	*0.0012*	*0.0016*	*0.0009*	*0.0013*	*0.0010*	*0.0010*		
$(3, 2)$	100	1.9544	0.9213	1.9699	0.9310	0.9661	0.9559	13.60	7.90
	$EMSE$	*0.0010*	*0.0011*	*0.0007*	*0.0010*	*0.0008*	*0.0008*		
	250	1.9801	0.9666	1.9771	0.0.9577	0.9819	0.9799	12.00	8.40
	$EMSE$	*0.0004*	*0.0004*	*0.0005*	*0.0008*	*0.0007*	*0.0005*		
	500	2.0129	0.9987	1.9839	0.9831	1.0144	1.0039	14.84	7.60
	$EMSE$	*0.0001*	*0.0003*	*0.0003*	*0.0005*	*0.0005*	*0.0002*		
$(3, 5)$	50	1.9356	0.9110	1.9217	0.9210	0.9693	0.9551	11.96	11.04
&	$EMSE$	*0.0016*	*0.0012*	*0.0009*	*0.0015*	*0.0008*	*0.0009*		
$(3, 5)$	100	1.9440	0.9299	1.9501	0.9567	0.9801	0.9610	11.00	10.40
	$EMSE$	*0.0011*	*0.0009*	*0.0007*	*0.0012*	*0.0006*	*0.0005*		
	250	1.9501	0.9587	1.9605	0.9771	0.9889	0.9899	11.80	9.60
	$EMSE$	*0.0007*	*0.0005*	*0.0004*	*0.0002*	*0.0003*	*0.0004*		
	500	1.9737	0.9777	1.9843	0.9810	1.0031	1.0045	11.88	11.11
	$EMSE$	*0.0004*	*0.0003*	*0.0002*	*0.0000*	*0.0001*	*0.0002*		

Table 4.4: Numerical results for Weibull AFT model

4.4 Semiparametric models in the presence of covariates

The most general proportional hazards model has cause-specific hazard function

$$h_{0j}(t)exp(\theta_j^T Z)$$

which allows for a cause-specific baseline hazard and cause-specific estimates of effect. With this semiparametric setup, the density of lifetimes is given by

$$f_j(t|Z) = f_{0j}(t)e^{\theta_j^T Z}S_{0j}(t)^{exp(\theta_j^T Z)-1}$$

Again, at this stage of development we will only deal with a single covariate, Z. Without loss of generality, let our observed data be

$$\{(T_1, \delta_1), (T_2, \delta_2), \cdots, (T_{n_1}, \delta_{n_1}), [(L_{n_1+1}, R_{n_1+1}), \delta_{n_1+1}], \cdots, [(L_{n_1+n_2}, R_{n_1+n_2}), \delta_{n_1+n_2}]\}$$

where $n_1 + n_2 = n$. Assuming that we have at least one censored observations i.e. $n_2 > 0$, the likelihood function of failure is given by

$$L(\theta_1, \theta_2) \;=\; \prod_{i=1}^{n_1}[f_1(t_i)\bar{F}_2(t_i)]^{I(\delta_i=1)}[f_2(t_i)\bar{F}_1(t_i)]^{I(\delta_i=2)}\prod_{j=1}^{2}\prod_{i=n_1+1}^{n}[F_j(R_i) - F_j(L_i)]^{I(\delta_i=j)}$$

The corresponding log-likelihood is

$$l_{full}(\theta_1, \theta_2) = l_{uncens}(\theta_1, \theta_2) + l_{cens}(\theta_1, \theta_2)$$

where

$$\begin{aligned} l_{uncens}(\theta_1, \theta_2) = \sum_{i=1}^{n_1}(I(\delta_i = 1)[lnf_{01}(t_i) + \theta_1^T Z_i + (e^{\theta_1^T Z_i} - 1)lnS_{01}(t_i) + e^{\theta_2^T Z_i}lnS_{02}(t_i)] \\ + I(\delta_i = 2)[lnf_{02}(t_i) + \theta_2^T Z_i + (e^{\theta_2^T Z_i} - 1)lnS_{02}(t_i) + e^{\theta_1^T Z_i}lnS_{01}(t_i)]) \qquad (4.4.1) \end{aligned}$$

$$l_{cens}(\theta_1, \theta_2) = \sum_{j=1}^{2} \sum_{i=n_1+1}^{n} ln[S_j(l_i) - S_j(r_i)]^{I(\delta_i=j)} \tag{4.4.2}$$

From equations (4.4.1) and (4.4.2), in order to estimate the parameters θ_1 and θ_2, we require the estimation of the baseline survival function $S_{0j}(t)$ for $j = 1, 2$ and the baseline density $f_{0j}(t)$. The survival function can be estimated nonparametrically by using the self-consistent estimator given in Jammalamadaka and Mangalam (2003) but difficulty arises when wanting to estimate the baseline density function. This can be avoided by taking the derivative of the log-likelihood.

$$l'(\theta) = \frac{\partial l_{uncens}(\theta)}{\partial \theta_j} + \frac{\partial l_{cens}(\theta)}{\partial \theta_j} \tag{4.4.3}$$

where the derivatives of the uncensored and censored data are given below

$$\frac{\partial l_{uncens}(\theta_1, \theta_2)}{\partial \theta_1} = \sum_{i=1}^{n_1} I(\delta_i = 1)\left(Z_i + lnS_{01}(t_i)Z_i e^{\theta_1^T Z_i}\right) + I(\delta_i = 2)\left(lnS_{01}(t_i)Z_i e^{\theta_1^T Z_i}\right) \tag{4.4.4}$$

$$\frac{\partial l_{uncens}(\theta_1, \theta_2)}{\partial \theta_2} = \sum_{i=1}^{n_1} I(\delta_i = 1)\left(lnS_{02}(t_i)Z_i e^{\theta_2^T Z_i}\right) + I(\delta_i = 2)\left(Z_i + lnS_{02}(t_i)Z_i (e^{\theta_2^T Z_i})\right) \tag{4.4.5}$$

$$\frac{\partial l_{cens}(\theta_1, \theta_2)}{\partial \theta_1} = \sum_{i=n_1+1}^{n} I(\delta_i = 1)\left(\frac{Z_i e^{\theta_1^T Z_i} ln(S_{01}(l_i))(S_{01}(l_i))^{e^{\theta_1^T Z_i}} - Z_i e^{\theta_1^T Z_i} ln(S_{01}(r_i))(S_{01}(r_i))^{e^{\theta_1^T Z_i}}}{S_{01}(l_i)^{e^{\theta_1^T Z_i}} - S_{01}(r_i)^{e^{\theta_1^T Z_i}}}\right) \tag{4.4.6}$$

$$\frac{\partial l_{cens}(\theta_1, \theta_2)}{\partial \theta_2} = \sum_{i=n_1+1}^{n} I(\delta_i = 2)\left(\frac{Z_i e^{\theta_2^T Z_i} ln(S_{02}(l_i))(S_{02}(l_i))^{e^{\theta_2^T Z_i}} - Z_i e^{\theta_2^T Z_i} ln(S_{02}(r_i))(S_{02}(r_i))^{e^{\theta_2^T Z_i}}}{S_{02}(l_i)^{e^{\theta_2^T Z_i}} - S_{02}(r_i)^{e^{\theta_2^T Z_i}}}\right) \tag{4.4.7}$$

From equations (4.4.4), (4.4.5), (4.4.6) and (4.4.7), the baseline density $f_{0j}(t)$ is not present, hence we do not need to estimate it in order to solve the roots of this equation. Since there is no general and closed form solution to equation (4.4.3), it can be solved numerically. Here, it is assumed that Z_i follows a $Binomial(1, p)$ distribution. The algorithm to find the MLE of the regression parameters θ_1 and θ_2 is given below. For $j = 1, 2$,

- Step 1: Estimate $S_{0j}^{(1)}(t)$ by using SCE (or NPMLE) using the data with no covariate effect i.e. use all of the t_i such that $Z_i = 0$
- Step 2: Estimate $\theta_j^{(1)}$ by solving the root of (4.4.3) and using $S_{0j}^{(1)}(t)$ to solve for the necessary probabilities in it
- Step 3: Obtain $\tilde{t}_i = {S_{0j}^{(1)}}^{-1}\left[S_{0j}^{(1)}(t_i)^{exp(\theta_j^{(1)})Z_i}\right]$. $\tilde{l}_i$ and $\tilde{r}_i$ can be found the same way.
- Step 4: Estimate $S_{0j}^{(2)}(t)$ by using SCE (or NPMLE) using all of the $\tilde{t}_i, \tilde{l}_i$ and $\tilde{r}_i$ as your data
- Step 5: Find $\theta_j^{(2)}$ by solving for the root of (4.4.3) and using $S_{0j}^{(2)}(t)$ to solve for the necessary probabilities in it
- Step 6: Repeat steps (3)-(5) until convergence criteria is met

To ensure that the algorithm works, a simulation is done. The baseline density for cause 1, $f_{01}(t)$ and cause 2, $f_{02}(t)$ are assumed to be exponential with mean 10 and 8 respectively. The censoring mechanism is as follows: the left censoring point for each individual is assumed to be an exponential random variable with mean 5 and the length of the interval is assumed to be another independent exponential random variable with mean 6. The covariates, Z_i were generated from a $Binomial(1, 0.5)$ distribution and the true covariate effects are $\theta_1 = 1$ and $\theta_2 = 1$. Graphs of the final estimate of $F_0(t)$ with sample sizes $n = 50, 100, 250$ and 1000 are shown in Figures (4.4.1) - (4.4.4). In each of these figures, the empirical CDF of the uncensored data is given as a starting point, the true CDF and the fitted CDF are shown. From the figures, it is clear that as the sample size n increases, the estimated CDF does a significantly better job at estimating the true distribution.

In this simulation, $N = 30$ samples were run with sample sizes $n = 50, 100, 250, 500$. The results of this study are given in Table 4.5. As the sample size increases, the accuracy of the MLEs of θ_1 and θ_2 increases slowly. Their EMSE decreases as shown in the table which means that their variability decreases. The variables $Censored_i$ for $i = 1, 2$ are the average amount of censoring in these $N = 30$ samples. On average, 12% of these observations are censored.

n	50	100	250	500
$\hat{\theta}_1$	0.8022	0.8515	0.9368	0.9783
$\hat{\theta}_2$	0.8420	0.8887	0.9062	0.9501
$EMSE(\hat{\theta}_1)$	0.0009	0.0005	0.0001	0.00001
$EMSE(\hat{\theta}_2)$	0.0012	0.0003	0.0002	0.00006
$Censored_1$	0.11	0.11	0.13	0.13
$Censored_2$	0.12	0.14	0.11	0.11

Table 4.5: Simulations for Cox model with Binomial covariates

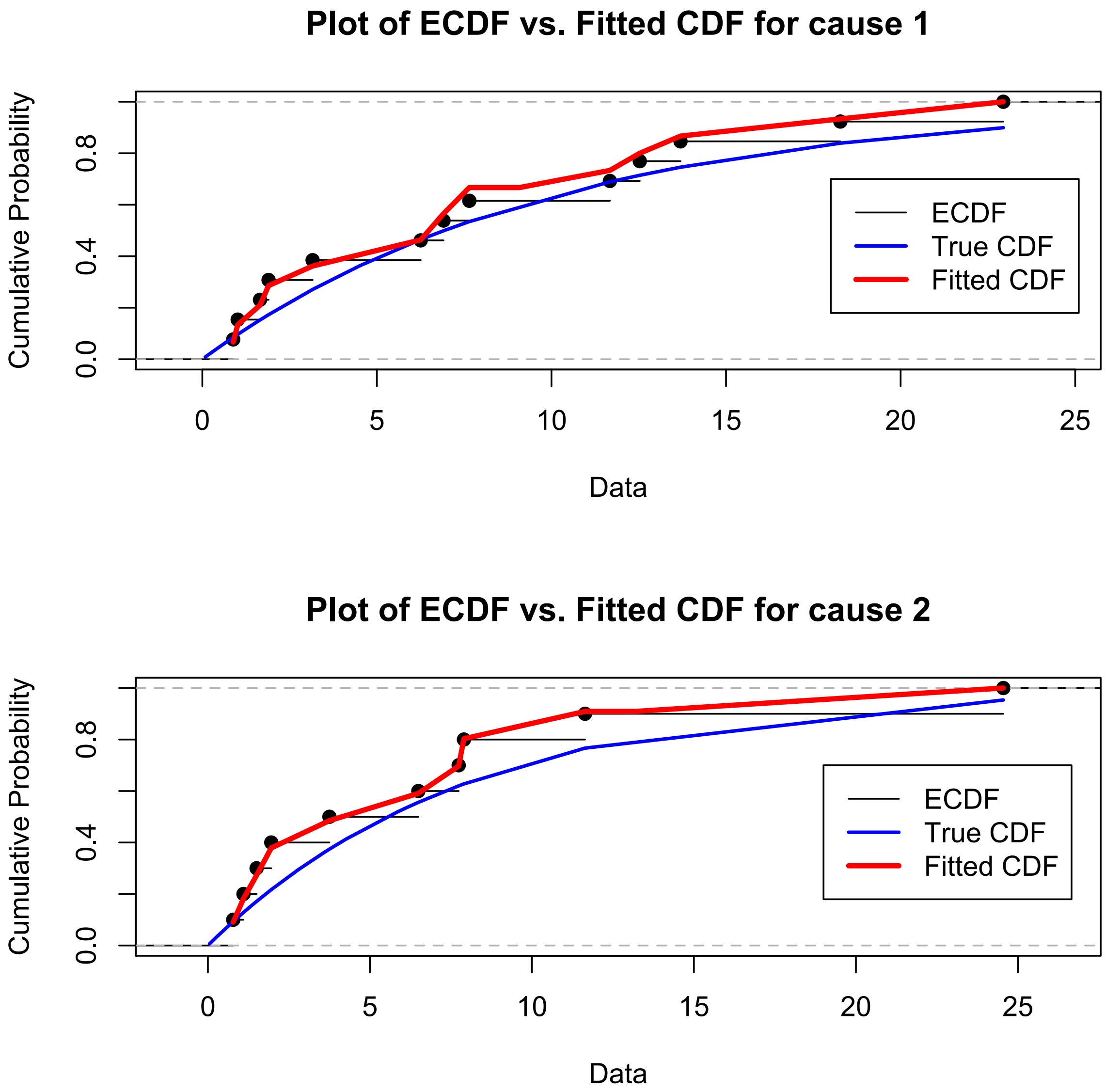
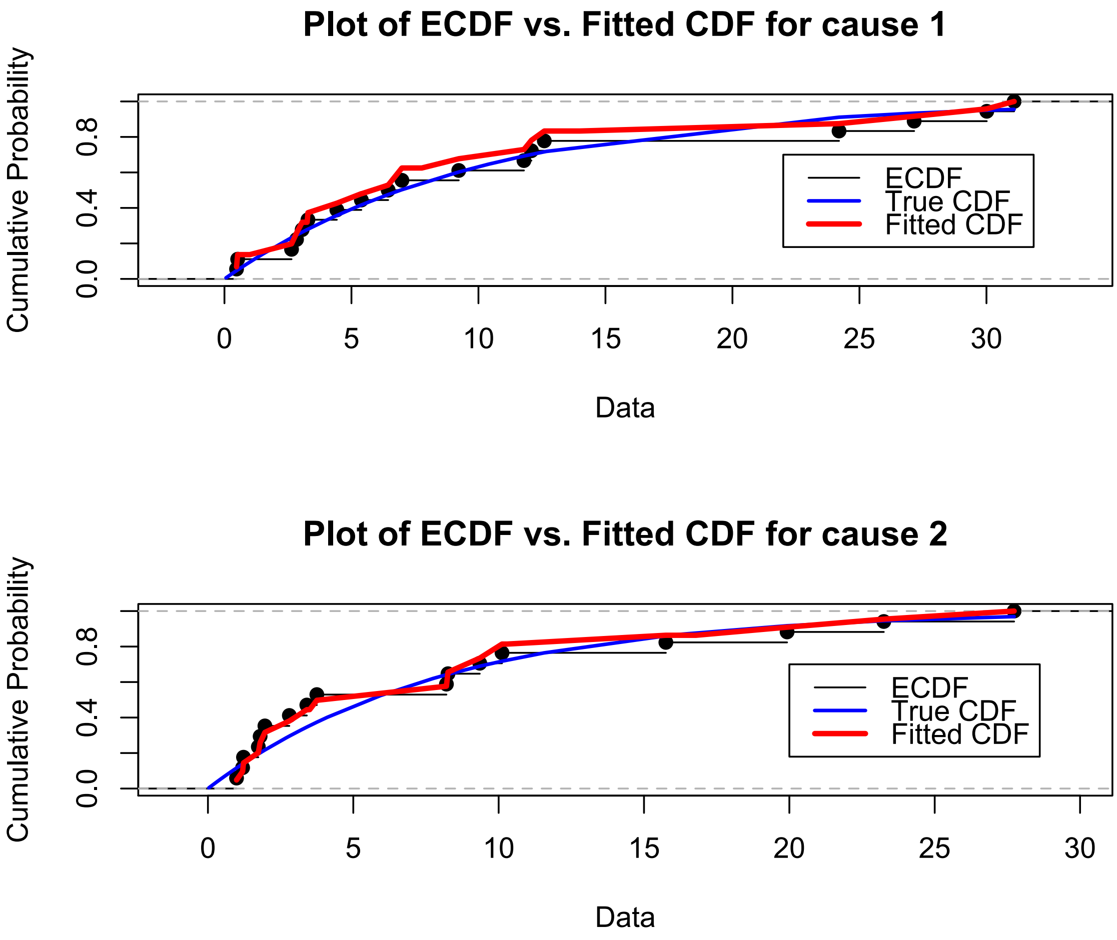


Figure 4.4.1: Semi-parametric model, $n = 50$


Plot of ECDF vs. Fitted CDF for cause 1

Plot of ECDF vs. Fitted CDF for cause 2

Figure 4.4.2: Semi-parametric model, $n = 100$

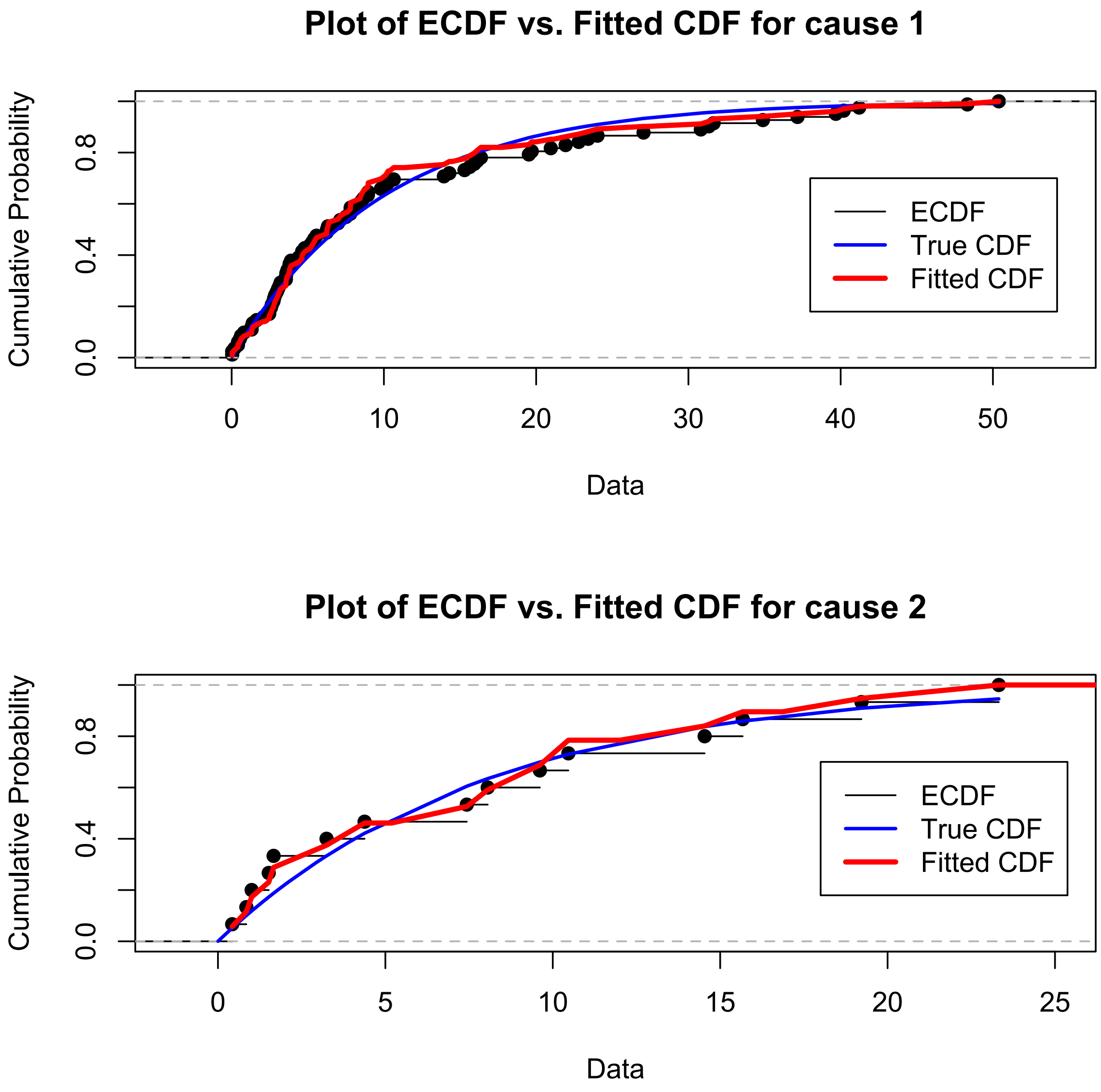


Figure 4.4.3: Semi-parametric model, $n = 250$

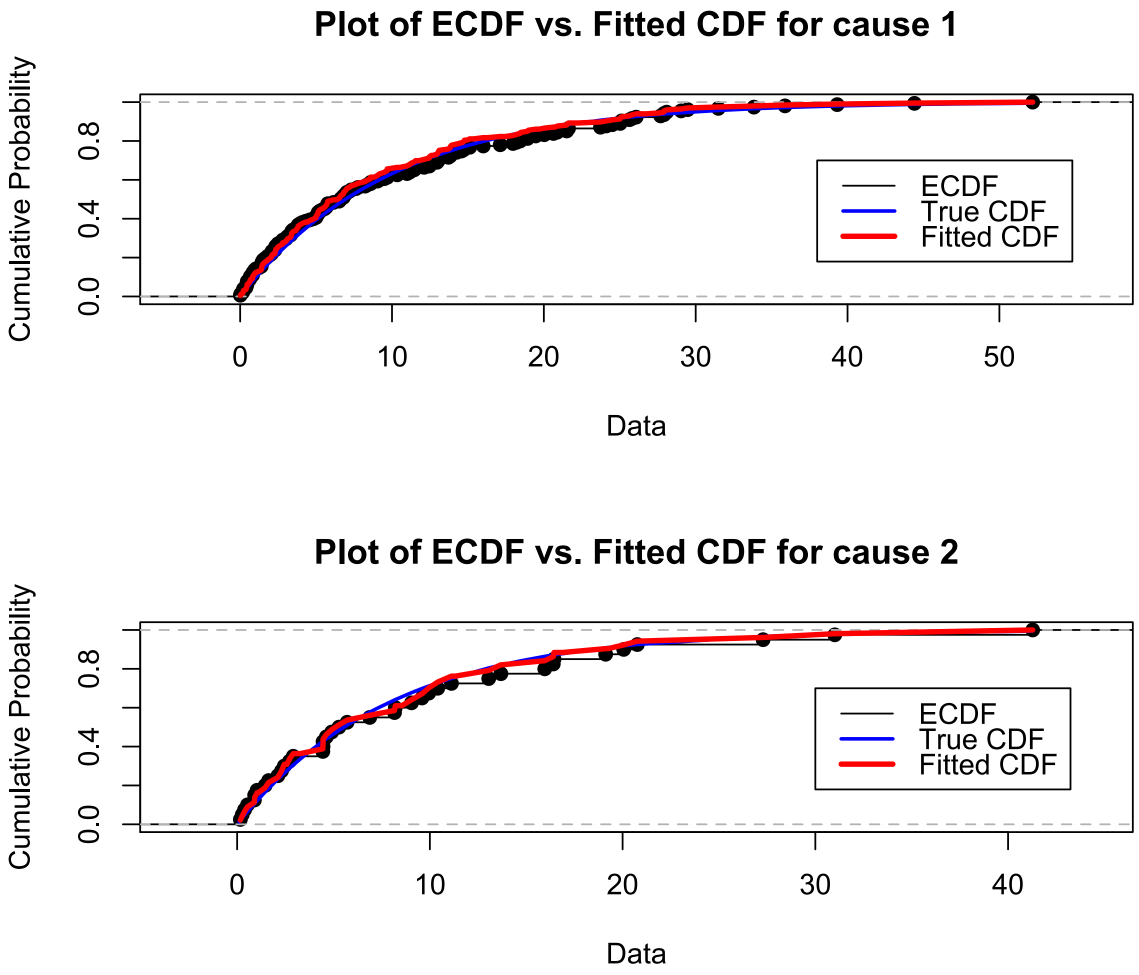


Figure 4.4.4: Semi-parametric model, $n = 500$

4.5 A practical example

In this section, again we apply the proposed techniques to a Stanford Heart Transplant study in Crowley and Hu (1977) as in Section (5.4) and (3.3). Previous open-heart surgery for each patient is indicated by the variable *surgery* taken as the time-independent covariate Z with respective regression coefficient θ_i for $i = 1, 2$. We create a variable called *risk* and randomly chose two causes of risk generated from a binomial distribution with probability 0.50.

In order to create a set of middle-censored data, we randomly choose several actual failure data and replace them by random censoring intervals. For cause 1, the data were censored by a random interval whose left end was an exponential random variable with mean 1 and the width was exponential with mean 10 while for cause 2, the data were censored by a random interval whose left end was an exponential random variable with mean 1 and the width was exponential with mean 12. It is found that 7% of data were censored for cause 1 resulting in 7 uncensored observations and 30 censored observations while for cause 2, 10% of data were censored resulting in 10 uncensored observations and 56 censored observations. It is observed that the maximum likelihood estimates of θ_1 and θ_2 of the complete data are -0.6240 and -0.6838 respectively. Applying the model given in Section 4.4, it is found that the estimates of the regression coefficients are $\hat{\theta}_1 = -0.6141$ and $\hat{\theta}_2 = -0.7253$. These estimates are close to the maximum likelihood estimates found using the complete data which shows that this approach yields reliable, useful and accurate results.

Chapter 5

Middle censoring in the presence of time-dependent covariates

5.1 Time-dependent covariates

One advantage of using Cox proportional hazard model is its ability to incorporate time-dependent covariates (see e.g. Cox, 1975 and Therneau and Grambsch, 2000). The time dependent covariates refers to a variable whose value itself changes over the duration of follow up. Examples of time-dependent covariate in biomedical research are 1. cumulative dosage of radiation or of a pharmaceutical agent, 2. receipt of an organ transplant and 3. compliance with a medication intended for chronic use (Austin, 2011). In the first example, cumulative dosage of radiation is a continuous time-dependent covariate, whose value is non-decreasing over time. In the second example, receipt of an organ transplant is a dichotomous exposure of treatment. For example, subjects may change their exposure status from unexposed to exposed at most once during the follow-up interval. It is assumed that once exposed, the subject remains exposed for the duration of follow up. One common approach to model this type of covariate is to use a step function which takes the value 1 for smoking and 0 otherwise within each follow-up interval (Fisher and Lin, 1999). In the third example, current medication use also represents a dichotomous exposure. However in this case, subjects may move from unexposed to exposed and from exposed to unexposed during the course of follow up.

Cox (1975) proposed the use of time-dependent covariate in the Cox Proportional Hazard models and gave the partial likelihood analysis and also generated the partial likelihood function for censored data. An algorithm was introduced by Peterson (1986) for estimating the parameters of parametric models in the presence of time-dependent covariates. Seaman and Bird (2001) discuss the proportional hazard model with interval censored data when time-dependent covariates are present. Parametric survival models for interval-censored data with time-dependent covariates were discussed in Sparling et al (2006).

If time-dependent variables are considered in a study, the Cox model from Equation (2.4.1) may still be used but such a model no longer satisfies the proportional hazard assumption. The model that integrates both time-independent and time-dependent covariates is called the extended Cox model (Kleinbaum and Klein, 2012, pages 244-245).

The hazard function for the extended Cox model is given by

$$h(t|\mathbf{Z}, \mathbf{Z}(t)) = h_0(t)e^{\sum_{i=1}^{p1} \theta_i Z_i + \sum_{j=1}^{p2} \delta_j Z_j(t)} \tag{5.1.1}$$

The assumption is that the effect of a time-dependent variable $Z_j(t)$ on the survival probability at time t depends on the value of this variable at that same time t, and not on the value at an earlier or later time. i.e.

$$h(t|\{Z(\mu), \mu \in [0,t]\}) = h(t|Z(t))$$

The extended Cox model in (5.1.1) can be written in another way that simultaneously considers all time-independent variables of interest. This model is written as

$$h(t|\mathbf{Z}, \mathbf{Z}(t)) = h_0(t)e^{\sum_{i=1}^{p_1}\theta_i Z_i + \sum_{j=1}^{p_2}\delta_j Z_j g_j(t)}$$

where $g_j(t)$ is some function of time for the jth variable. Some common examples of $g(t)$ are t, $log(t)$ and the heaviside functions (Kleinbaum and Klein, 2012, pages 254-255). The choice of time-dependent covariates may be based on theoretical considerations and strong clinical evidence.

The hazard ratio, HR at time t for the two individuals with different covariates Z and Z^* is given by

$$\hat{HR}(t) = exp\left(\sum_{i=1}^{p_1}\hat{\theta}_i[Z_i^* - Z_i] + \sum_{j=1}^{p_2}\hat{\delta}_j[Z_j^*(t) - Z_j(t)]\right)$$

Note that in this hazard ratio formula, δ_j represents overall effect of $Z_j(t)$ considering all times at which this variable has been measured in this study. However, the hazard ratio depends on time, t which means that the hazards of event at time t is no longer proportional, and the model is no longer a PH model.

To check the PH assumption using a statistical test, consider $H_0 : \delta_1 = \delta_2 = \cdots = \delta_p = 0$. Under H_0, the model reduces to the PH model. The test can be carried out using a likelihood ratio, LR test which computes the difference between the log likelihood statistic, $-2lnL$ for the proportional hazard, PH model and the log likelihood statistic for the extended Cox model. The test statistic has approximately a chi-square distribution with p degrees of freedom under the null hypothesis, where p denotes the number of parameters being set equal to zero under H_0. i.e. Under H_0, the likelihood ratio test,

$$LR = -2lnL_{PH} - (-2lnL_{ExtCox}) \sim \chi^2_p$$

In addition to considering time-dependent variable for analyzing a time - independent variable not satisfying the PH assumption, there are variables that are inherently defined as time-dependent variables. One of the earliest applications of the use of time-dependent covariates is in the report by Crowley and Hu (1977) on the Stanford Heart Transplant study. Time-dependent variables are usually classified to be internal or external. An internal time-dependent variable is one that the change of covariate over time is related to the characteristics or behavior of the individual. For example, blood pressure, disease complications, etc. The external time-dependent variable is one whose value at a particular time does not require subjects to be under direct observation, i.e., values change because of characteristics external to the individuals such as the level of air-pollution.

When no particular distribution is known, the semiparametric model is used with no specific form for the hazard function. If a particular distribution can be specified for the survival data, parametric model is the appropriate approach. The estimates obtained from the parametric model will be more accurate and the relative efficiency against the semiparametric model increases. Both semiparametric and parametric models may be extended to accommodate time-dependent covariates as shown in Section (5.3) and (5.2) respectively.

5.2 Parametric model

Lifetimes that follow an exponential and Weibull distribution model are considered in this section. Two types of time-dependent covariates which are a dichotomous time-dependent covariate that changes at most once from one status to another and a continuous time-dependent covariate are explored.

5.2.1 A dichotomous time-dependent covariate

Suppose that the time-dependent covariate is a dichotomous time-dependent covariate that can change at most once from untreated to treated. Let t_0 denote the time at which the time-varying covariate changes from untreated $(Z = 0)$ to treated $(Z = 1)$. Hence, we have

$$Z(t) = \begin{cases} 0 & \text{if } t < t_0 \\ \\ 1 & \text{if } t \geq t_0 \end{cases}$$

Lifetimes that follow an exponential distribution and a Weibull distribution are considered in this section.

Exponential distributed model

Consider an exponential distributed model with middle censoring in the presence of time-dependent covariate. Here, each person has a survival time, T and covariates specific to that individual, Z and $Z(t)$. The lifetimes will be $Exponential(\lambda exp(\theta^T \mathbf{Z} + \delta^T \mathbf{Z}(t)))$ where θ and δ are the effect of each covariate.

Suppose that there is one time-independent covariate, Z and one time-dependent covariate denoted by $Z(t)$. Let θ be the vector of regression coefficients associated with the vector of fixed covariate, Z while δ is the regression coefficient associated with $Z(t)$. Hence, the hazard function, $h(t)$ is given as

$$h(t|Z, Z(t)) = \lambda exp(\theta Z + \delta Z(t))$$

and the cumulative hazard function, $H(t)$ is written as

$$H(t|Z, Z(t)) = \int_0^t \lambda exp(\theta Z + \delta Z(u))du$$

The cumulative hazard function, $H(t)$ for $t < t_0$ and $t \geq t_0$ is given as the following. The derivation of this expression is presented in Appendix A.

$$H(t|Z, Z(t)) = \begin{cases} \lambda exp(\theta Z)t & \text{if } t < t_0 \\ \\ \lambda exp(\theta Z)(t_0 + exp(\delta)t - exp(\delta)t_0) & \text{if } t \geq t_0 \end{cases}$$

The survival function of the above model is then given as

$$S(t|Z, Z(t)) = \begin{cases} exp(-\lambda exp(\theta Z)t) & \text{if } t < t_0 \\ \\ exp(-\lambda exp(\theta Z)(t_0 + exp(\delta)t - exp(\delta)t_0) & \text{if } t \geq t_0 \end{cases}$$

where $S(t)$ is defined as $S(t) = exp(-H(t))$. The above survival function can be rewritten as

$$S(t|Z, Z(t)) = exp\left[-\lambda e^{\theta Z}[t(1 - Z(t)) + e^{\delta}tZ(t) + (1 - e^{\delta})t_0 Z(t)]\right]$$

Hence, the density of lifetimes is given by

$$\begin{aligned} f(t|Z, Z(t)) =& \lambda e^{\theta Z}[1 - Z(t) + e^{\delta}Z(t)] \\ & \times exp\left[-\lambda e^{\theta Z}[t(1 - Z(t)) + e^{\delta}tZ(t) + (1 - e^{\delta})t_0 Z(t)]\right] \end{aligned} \tag{5.2.1}$$

Assume that there is at least one censored observation i.e. $n_2 > 0$ and consider the censoring mechanism given in Section (2.2). The log-likelihood of the exponential lifetimes with one time-independent covariate, Z and one time-dependent covariate, $Z(t)$ is

$$\begin{aligned} l(t|Z, Z(t)) = n_1 ln(\lambda) + \sum_{i=1}^{n_1} (\theta Z_i + ln[1 - Z_i(t_i) + exp(\delta) Z_i(t_i)]) \\ - \sum_{i=1}^{n_1} \lambda e^{\theta Z_i} \left[t_i(1 - Z_i(t_i)) + exp(\delta) t_i Z_i(t_i) + (1 - exp(\delta)) t_{0i} Z_i(t_i)\right] \\ + \sum_{i=n_1+1}^{n} ln(S(l_i|Z_i, Z_i(l_i)) - S(r_i|Z_i, Z_i(r_i)) \end{aligned} \tag{5.2.2}$$

Now, applying the EM algorithm in the same fashion as Section (2.2), the following conditional expectation required is

$$t^* = \frac{\int_l^r tf(t|Z, Z(t))dt}{S(l|Z, Z(l)) - S(r|Z, Z(r))} \tag{5.2.3}$$

Then the log-likelihood required is written as

$$\begin{aligned} l^*(\lambda, \theta, \delta) =& nln(\lambda) + \sum_{i=1}^{n} \left[\theta Z_i + ln(1 - Z_i(t_i) + exp(\delta) Z_i(t_i))\right] \\ & - \sum_{i=1}^{n_1} \lambda exp(\theta Z_i)[t_i(1 - Z_i(t_i)) + exp(\delta) t_i Z_i(t_i)] \\ & - \sum_{i=n_1+1}^{n} \lambda exp(\theta Z_i)[t_i^*(1 - Z_i(t_i)) + exp(\delta) t_i^* Z_i(t_i)] \\ & - \sum_{i=1}^{n} \lambda exp(\theta Z_i) \left[1 - exp(\delta)\right] t_{i0} Z_i(t_i) \end{aligned} \tag{5.2.4}$$

where the t_i^*'s are found using Equation (5.2.3).

The EM algorithm can now be set up as follows. To make this more explicit, choose $(\lambda, \theta, \delta)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(\lambda, \theta, \delta)_{(j)}$ is the jth estimate
- Step 2: Compute Equation (5.2.3) with $(\lambda, \theta, \delta) = (\lambda, \theta, \delta)_{(j)}$
- Step 3: Solve Equation (5.2.4) for its maximum and set $(\lambda, \theta, \delta)_{(j+1)}$ as that maximum
- Step 4: Repeat until convergence criteria is met

A simulation study is performed to illustrate the usefulness of this approach. Simulations are carried out in R using $N = 100$ replications with a common sample size of $n = 250$. The censoring mechanism is that the left endpoint of the censored interval is exponentially distributed with mean 1 while the length of the censored interval is also exponentially distributed with mean 1. The covariate, Z is generated from a binomial distribution. The covariate, $Z(t)$ is a dichotomous time-dependent covariate that can change at most once from untreated to treated. Let t_0 denote the time at which the time-varying covariate changes from unexposed $(Z = 0)$ to exposed $(Z = 1)$. Hence, we have

$$Z(t) = \begin{cases} 0 & \text{if } t < t_0 \\ 1 & \text{if } t \geq t_0 \end{cases}$$

The survival time can be simulated from the following equations.

$$T = \begin{cases} \frac{-ln(u)}{\lambda exp(\theta Z)} & \text{if } -ln(u) < \lambda exp(\theta Z) t_0 \\ \frac{-ln(u) - \lambda exp(\theta Z) t_0 + \lambda exp(\theta Z + \delta) t_0}{\lambda exp(\theta Z + \delta)} & \text{if } -ln(u) \geq \lambda exp(\theta Z) t_0 \end{cases}$$

where $u \sim Uniform(0, 1)$. The derivation of the expression is shown in Appendix A.

Three cases for the true covariate effects were considered. They are $(\theta, \delta) = (1, 1), (1, 0)$ and $(0, 1)$. The true value for λ is taken to be 0.5 in all cases and t_0 is taken to be exponential distributed with rate 0.5. In these $N = 100$ simulations, the samples were in average between 23% to 25% censored. The true values λ, θ and δ were compared with their MLE's using the EM algorithm described above. Table 5.1 reports the results from these simulations. The MLE's of all the parameters are quite close to the actual value with small estimated mean squared error *(EMSE)*. The estimated mean squared error *EMSE* is calculated using the equation

$$EMSE(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta} - \theta)^2$$

Hence, this approach yields accurate, useful and reliable results.

	λ	θ	δ
True Value	0.5	1	1
MLE	0.5048	0.9468	1.0632
EMSE	0.0021	0.0358	0.0832
Censored Proportion	0.2468		
True Value	0.5	1	0
MLE	0.4996	1.0463	0.1069
EMSE	0.0015	0.01813	0.0184
Censored Proportion	0.2342		
True Value	0.5	0	1
MLE	0.5003	0.0209	1.0216
EMSE	0.0040	0.0064	0.0491
Censored Proportion	0.2526		

Table 5.1: Numerical results for exponential AFT model in the presence of a dichotomous time-dependent covariate

Weibull distributed model

Consider a Weibull distributed model with shape α and scale β with middle censoring in the presence of one time-independent covariate, Z and one time-dependent covariate, $Z(t)$. Each person has a survival time T and covariates specific to that individual, Z and $Z(t)$. As in the exponential model, let θ be the vector of regression coefficients associated with the vector of fixed covariates, Z while δ is the regression coefficient associated with $Z(t)$. The hazard function, $h(t)$ is given as

$$h(t|Z, Z(t)) = \alpha\beta t^{\alpha-1} exp(\theta Z + \delta Z(t))$$

and the cumulative hazard function, $H(t)$ is given as

$$H(t|Z, Z(t)) = \int_0^t \alpha\beta u^{\alpha-1} exp(\theta Z + \delta Z(u)) du$$

The cumulative hazard function, $H(t)$ for $t < t_0$ and $t \geq t_0$ is given as the following. The derivation of this expression is presented in Appendix C.

$$H(t|Z, Z(t)) = \begin{cases} \beta exp(\theta Z) t^{\alpha} & \text{if } t < t_0 \\ \\ \beta exp(\theta Z)(t_0^{\alpha} + exp(\delta) t^{\alpha} - exp(\delta) t_0^{\alpha}) & \text{if } t \geq t_0 \end{cases}$$

The survival function of the above model is then given as

$$S(t|Z, Z(t)) = \begin{cases} exp(-\beta exp(\theta Z) t^{\alpha}) & \text{if } t < t_0 \\ \\ exp(-\beta exp(\theta Z)(t_0^{\alpha} + exp(\delta) t^{\alpha} - exp(\delta) t_0^{\alpha})) & \text{if } t \geq t_0 \end{cases}$$

where $S(t)$ is defined as $S(t) = exp(-H(t))$. The above survival function can be rewritten as

$$S(t|Z, Z(t)) = exp[-\beta exp(\theta Z)(t^{\alpha}(1 - Z(t)) + exp(\delta) t^{\alpha} Z(t) + (1 - exp(\delta)) t_0^{\alpha} Z(t))]$$

Hence, the density of lifetimes is given by

$$\begin{aligned} f(t|Z,Z(t)) =&\beta exp(\theta Z)[(1-Z(t))\alpha t^{\alpha-1}+exp(\delta)Z(t)\alpha t^{\alpha-1}]\times \\ & exp[-\beta e^{\theta Z}(t^{\alpha}(1-Z(t))+e^{\delta}t^{\alpha}Z(t)+(1-e^{\delta})t_0^{\alpha}Z(t))] \end{aligned} \tag{5.2.5}$$

Again, assume that there is at least one censored observation i.e. $n_2 > 0$ and consider the censoring mechanism given in Section (2.2). The log-likelihood is

$$\begin{aligned} l(t|Z,Z(t)) = n_1 ln(\beta) &+ \sum_{i=1}^{n_1}(\theta Z_i + ln[(1-Z_i(t_i))\alpha t_i^{\alpha-1}+exp(\delta)Z_i(t_i)\alpha t_i^{\alpha-1}]) \\ &- \sum_{i=1}^{n_1}\beta e^{\theta Z_i}\,[t_i^{\alpha}(1-Z_i(t_i))+exp(\delta)t_i^{\alpha}Z_i(t_i)+(1-exp(\delta))t_{0i}^{\alpha}Z_i(t_i)] \\ &+ \sum_{i=n_1+1}^{n} ln(S(l_i|Z_i,Z_i(l_i))-S(r_i|Z_i,Z_i(r_i)) \end{aligned} \tag{5.2.6}$$

Applying the EM algorithm in the same fashion as Section (2.2), the following conditional expectations required are

$$t^{*\alpha} = \frac{\int_l^r t^{\alpha}\times f(t|Z,Z(t))dt}{S(l|Z,Z(l))-S(r|Z,Z(r))} \tag{5.2.7}$$

$$t^{*\alpha-1} = \frac{\int_l^r t^{\alpha-1}\times f(t|Z,Z(t))dt}{S(l|Z,Z(l))-S(r|Z,Z(r))} \tag{5.2.8}$$

Then the log-likelihood required is

$$\begin{aligned} l^*(\alpha,\beta,\theta,\delta) =& nln\beta + \sum_{i=1}^{n}(\theta Z_i) - \sum_{i=1}^{n}\beta exp(\theta Z_i)[(1-exp(\delta))t_{i0}^{\alpha}Z(t_i)] \\ &+ \sum_{i=1}^{n_1} ln[(1-Z_i(t_i))\alpha t_i^{\alpha-1} + exp(\delta)Z(t_i)\alpha t_i^{\alpha-1}] \\ &+ \sum_{i=n_1+1}^{n} ln[1-Z_i(t_i))\alpha t_i^{*\alpha-1} + exp(\delta)Z(t_i)\alpha t_i^{*\alpha-1}] \\ &- \sum_{i=1}^{n_1}\beta exp(\theta Z_i)[t_i^{\alpha}(1-Z(t_i)) + exp(\delta)t_i^{\alpha}Z(t_i)] \\ &- \sum_{i=n_1+1}^{n}\beta exp(\theta Z_i)[t_i^{*\alpha}(1-Z(t_i)) + exp(\delta)t_i^{*\alpha}Z(t_i)] \end{aligned} \tag{5.2.9}$$

where the $t_i^{*\alpha}$'s and the $t_i^{*\alpha-1}$ are found using Equation (5.2.7) and (5.2.8) respectively.

The EM algorithm can now be set up as follows. To make this more explicit, choose $(\alpha,\beta,\theta,\delta)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(\alpha,\beta,\theta,\delta)_{(j)}$ is the jth estimate
- Step 2: Compute Equations (5.2.7) and (5.2.8) with $(\alpha,\beta,\theta,\delta) = (\alpha,\beta,\theta,\delta)_{(j)}$
- Step 3: Solve Equation (5.2.9) for its maximum and set $(\alpha,\beta,\theta,\delta)_{(j+1)}$ as that maximum
- Step 4: Repeat until convergence criteria is met

A simulation study is performed to illustrate the usefulness of this approach. Simulations are carried out in R using $N = 100$ replications with a common sample size of $n = 250$. The censoring mechanism is that the left endpoint of the censored interval is exponentially distributed with mean 1 while the length of the censored interval is also exponentially distributed with mean 1. The covariate, Z is generated from a binomial distribution. The covariate, $Z(t)$ is a dichotomous time-dependent covariate that can change at most once from untreated to treated. Let t_0 denote the time at which the time-varying covariate changes from unexposed $(Z = 0)$ to exposed $(Z = 1)$. Hence, we have

$$Z(t) = \begin{cases} 0 & \text{if } t < t_0 \\ \\ 1 & \text{if } t \geq t_0 \end{cases}$$

The survival time can be simulated from the following equations.

$$T = \begin{cases} \left(\frac{-ln(u)}{\beta exp(\theta Z)}\right)^{1/\alpha} & \text{if } -ln(u) < \beta exp(\theta Z)t_0^\alpha \\ \\ \left(\frac{-ln(u)-\beta exp(\theta Z)t_0^\alpha+\beta exp(\theta Z+\delta)t_0^\alpha}{\beta exp(\theta Z+\delta)}\right)^{1/\alpha} & \text{if } -ln(u) \geq \beta exp(\theta Z)t_0^\alpha \end{cases}$$

where $u \sim Uniform(0, 1)$. The derivation of the expression is shown in Appendix C.

Three cases for the true covariate effects were considered. They are $(\theta, \delta) = (0.5, 0.5)$, $(1.0, 0.5)$ and $(0.5, 1.0)$. The true value for α and β are taken to be 0.6 and 0.5 respectively in all cases and t_0 is taken to be exponential distributed with rate 0.5. In these $N = 100$ simulations, the samples were in average between 17% to 22% censored. The true values α, β, θ and δ were compared with their MLE's using the EM algorithm described above. Table 5.2 reports the results from these simulations. The MLE's of all the parameters are close to the actual value with small estimated mean squared error *(EMSE)*. As expected, this approach yields reliable, useful and accurate results.

	α	β	θ	δ
True Value	0.6	0.5	0.5	0.5
MLE	0.6327	0.4743	0.5243	0.4821
EMSE	0.0080	0.0039	0.0065	0.0153
Censored Proportion	0.1760			
True Value	0.6	0.5	1.0	0.5
MLE	0.6049	0.4777	1.1845	0.4779
EMSE	0.0003	0.0039	0.0052	0.0378
Censored Proportion	0.1840			
True Value	0.6	0.5	0.5	1.0
MLE	0.5915	0.4512	0.5254	1.1491
EMSE	0.0003	0.0055	0.0059	0.0072
Censored Proportion	0.2240			

Table 5.2: Numerical results for Weibull AFT model in the presence of a dichotomous time-dependent covariate

5.2.2 Continuous time-dependent covariate

Suppose that we have a continuous time-dependent covariate whose value is non - decreasing over time. Just as a matter of illustration, the time-dependent covariate $Z(t)$ is assumed to be a linear function $Z(t) = s + kt$, where $k > 0$ and $t > 0$. This would be the case when a subject is exposed to a uniform dose during each unit of time during follow up. An example of this case would be when subjects are exposed to a fixed dose of medication each day or when patients take a certain dose of medication each day (Austin, 2011). Lifetimes that follow an exponential distribution and a Weibull distribution are considered in this section.

Exponential distributed model

Consider an Exponential distributed model with rate λ in the presence of one time-independent covariate, Z and one time-dependent covariate, $Z(t)$ where $Z(t) = s + kt$. Let θ be the vector of fixed covariates, Z and δ is the regression coefficient associated with $Z(t)$. The cumulative hazard function, $H(t)$ is given as the following. The derivation of this expression is presented in Appendix B.

$$H(t|Z, Z(t)) = \tfrac{\lambda exp(\theta Z+\delta s)}{\delta k}[exp(\delta kt) - 1]$$

The survival function of the above model is given as

$$S(t|Z, Z(t)) = exp(-\tfrac{\lambda exp(\theta Z+\delta s)}{\delta k}[exp(\delta kt) - 1])$$

where $S(t)$ is defined as $S(t) = exp(-H(t))$.

Therefore, the density of lifetimes is given by

$$f(t|Z, Z(t)) = \lambda exp\left(\frac{\lambda}{\delta k}e^{\theta Z+\delta Z(t)}(e^{-\delta kt} - 1) + \theta Z + \delta Z(t)\right)$$

Now assume that there is at least one censored observation i.e. $n_2 > 0$ and consider the censoring mechanism given in Section (2.2). The log-likelihood is written as

$$\begin{aligned} l(t|Z, Z(t)) =& n_1 ln(\lambda) + \sum_{i=1}^{n_1}\left(\frac{\lambda}{\delta k}e^{\theta Z_i+\delta Z_i(t_i)}(e^{-\delta kt_i} - 1) + \theta Z_i + \delta Z_i(t_i)\right) \\ &+ \sum_{i=n_1+1}^{n} ln(S(l_i|Z_i, Z_i(l_i)) - S(r_i|Z_i, Z_i(r_i)) \end{aligned} \quad (5.2.10)$$

The conditional expectation required for applying the EM algorithm is given as

$$t^* = \frac{\int_l^r t\lambda exp\left(\frac{\lambda}{\delta k}e^{\theta Z+\delta Z(t)}(e^{-\delta kt} - 1) + \theta Z + \delta Z(t)\right)dt}{S(l|Z, Z(l)) - S(r|Z, Z(r))} \quad (5.2.11)$$

Then the log-likelihood required is

$$
\begin{aligned}
l^*(\lambda,\theta,\delta) =& nln\lambda + \sum_{i=1}^{n}\theta Z_i + \delta\left(\sum_{i=1}^{n_1} Z_i(t_i) + \sum_{i=n_1+1}^{n} Z_i(t_i^*)\right) \\
&+ \frac{\lambda}{\delta k}\left(\sum_{i=1}^{n_1}\left[e^{\theta Z_i+\delta Z_i(t_i)}(e^{-\delta k t_i - 1})\right] + \sum_{i=n_1+1}^{n}\left[e^{\theta Z_i+\delta Z_i(t_i^*)}(e^{-\delta k t_i^* - 1})\right]\right)
\end{aligned}
\tag{5.2.12}
$$

where the t_i^*'s are found using Equation (5.2.11).

The EM algorithm can now be set up as follows. To make this more explicit, choose $(\lambda,\theta,\delta)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(\lambda,\theta,\delta)_{(j)}$ is the jth estimate
- Step 2: Compute Equation (5.2.11) with $(\lambda,\theta,\delta) = (\lambda,\theta,\delta)_{(j)}$
- Step 3: Solve Equation (5.2.12) for its maximum and set $(\lambda,\theta,\delta)_{(j+1)}$ as that maximum
- Step 4: Repeat until convergence criteria is met

Now, a simulation study is performed to illustrate the usefulness of this approach. Simulations are carried out in R using $N = 100$ replications with a common sample size of $n = 250$. The censoring mechanism is that the left endpoint of the censored interval is Exponentially distributed with mean 1 while the length of the censored interval is also Exponentially distributed with mean 1. The covariate, Z is generated from a Binomial distribution. The covariate, $Z(t)$ is a continuous time-dependent where $Z(t) = s + kt$ and

$s = 1, k = 5$. The survival time can be simulated from Equation (5.2.13)

$$T = \frac{1}{\delta k} ln \left[1 + \frac{\delta k(-ln(u))}{\lambda exp(\theta Z + \delta s)} \right] \tag{5.2.13}$$

where $u \sim Uniform(0, 1)$. The derivation of the expression is shown in Appendix B.

Three cases for the true covariate effects were considered. They are $(\theta, \delta) = (1.0, 1.0)$, $(1.0, 0.5)$ and $(0.0, 1.0)$. The true value for λ is taken to be 0.5 in all cases. In these $N = 100$ simulations, the samples were in average between 14% to 25% censored. The true values λ, θ and δ were compared with their MLE's using the EM algorithm described above. Table 5.3 reports the results from these simulations. The MLE's of all the parameters are good as they are close to the actual value. Again, the estimated mean squared errors *(EMSE)* are small for all parameters. Again, this approach yields useful, reliable and accurate results.

	λ	θ	δ
True Value	0.5	1.0	1.0
MLE	0.5327	1.0373	1.0078
EMSE	0.0038	0.0042	0.0007
Censored Proportion	0.1488		
True Value	0.5	1.0	0.5
MLE	0.5213	1.2232	0.5038
EMSE	0.0106	0.0577	0.0062
Censored Proportion	0.2127		
True Value	0.5	0.0	1.0
MLE	0.5089	-0.0955	1.0324
EMSE	0.0033	0.0205	0.0046
Censored Proportion	0.2539		

Table 5.3: Numerical results for Exponential AFT model in the presence of continuous time-dependent covariate

Weibull distributed model

When the event times follow a Weibull distribution with shape α and scale β, the cumulative hazard function, $H(t)$ is given as equation (5.2.14). The derivation of this expression is presented in Appendix D. Again, suppose there is one time-independent covariate, Z and one time-dependent covariate, $Z(t)$ where $Z(t) = s + kt$ with $k > 0$.

$$H(t|Z, Z(t)) = \frac{\alpha\beta\delta}{1+\alpha} exp(\theta Z + \delta s) \left[exp(kt^{1+\alpha}) - 1\right] \tag{5.2.14}$$

The survival function of the above model is given as

$$S(t|Z, Z(t)) = exp\left(-\tfrac{\alpha\beta\delta}{1+\alpha} exp(\theta Z + \delta s) \left[exp(kt^{1+\alpha}) - 1\right]\right)$$

where $S(t)$ is defined as $S(t) = exp(-H(t))$.

Hence, the density of lifetimes is given by

$$f(t|Z, Z(t)) = \alpha\beta\delta k t^{\alpha} exp\left(\frac{-\alpha\beta\delta(e^{kt^{\alpha+1}} - 1)e^{\theta Z + \delta s}}{\alpha+1} + \theta Z + kt^{\alpha+1} + \delta s\right)$$

Now assume that there is at least one censored observation i.e. $n_2 > 0$ and consider the censoring mechanism given in Section (2.2). The log-likelihood is written as

$$\begin{aligned} l(t|Z, Z(t)) =& n_1 ln(\alpha\beta\delta k) + \sum_{i=1}^{n_1}\left(-\frac{\alpha\beta\delta}{\alpha+1}(e^{kt_i^{\alpha+1}} - 1)e^{\theta Z_i + \delta s} + \theta Z_i + kt_i^{\alpha+1} + \delta s\right) \\ &+ \alpha \sum_{i=1}^{n_1} ln(t_i) + \sum_{i=n_1+1}^{n} ln(S(l_i|Z_i, Z_i(l_i)) - S(r_i|Z_i, Z_i(r_i))) \end{aligned} \tag{5.2.15}$$

Applying the EM algorithm in the same fashion as Section (2.2), the following conditional expectations required are

$$ln(t)^* = \frac{\int_l^r ln(t)\alpha\beta\delta k t^{\alpha} exp\left(\frac{-\alpha\beta\delta(e^{kt^{\alpha+1}}-1)e^{\theta Z+\delta s}}{\alpha+1} + \theta Z + kt^{\alpha+1} + \delta s\right)dt}{S(l|Z,Z(l)) - S(r|Z,Z(r))} \tag{5.2.16}$$

$$t^{*a+1} = \frac{\int_l^r t^{a+1}\alpha\beta\delta k t^{\alpha} exp\left(\frac{-\alpha\beta\delta(e^{kt^{\alpha+1}}-1)e^{\theta Z+\delta s}}{\alpha+1} + \theta Z + kt^{\alpha+1} + \delta s\right)dt}{S(l|Z,Z(l)) - S(r|Z,Z(r))} \tag{5.2.17}$$

Then the log-likelihood required is

$$\begin{aligned} l^*(\alpha,\beta,\theta,\delta) =& nln(\alpha\beta\delta k) + \alpha\left(\sum_{i=1}^{n_1} ln(t_i) + \sum_{i=n_1+1}^{n} ln(t_i)^*\right) + \sum_{i=1}^{n}(\theta Z_i + \delta s) \\ & - \frac{\alpha\beta\delta}{1+\alpha}\left[\sum_{i=1}^{n_1}(e^{kt_i^{\alpha+1}} - 1)e^{\theta Z_i+\delta s} + \sum_{i=n_1+1}^{n}(e^{kt_i^{*\alpha+1}} - 1)e^{\theta Z_i+\delta s}\right] \\ & + k\left[\sum_{i=1}^{n_1} t_i^{a+1} + \sum_{i=n_1+1}^{n} t_i^{*a+1}\right] \end{aligned} \tag{5.2.18}$$

where the $ln(t_i)^*$'s and t_i^{*a+1}'s are found using Equation (5.2.16) and (5.2.17) respectively.

The EM algorithm can now be set up as follows. To make this more explicit, choose $(\alpha,\beta,\theta,\delta)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(\alpha,\beta,\theta,\delta)_{(j)}$ is the jth estimate
- Step 2: Compute Equation (5.2.16) and (5.2.17) with $(\alpha,\beta,\theta,\delta) = (\alpha,\beta,\theta,\delta)_{(j)}$
- Step 3: Solve Equation (5.2.18) for its maximum and set $(\alpha,\beta,\theta,\delta)_{(j+1)}$ as that maximum
- Step 4: Repeat until convergence criteria is met

A simulation study is performed to illustrate the usefulness of this approach. It is carried out in R using $N = 100$ replications with a common sample size of $n = 250$. The censoring mechanism is that the left endpoint of the censored interval is Exponentially distributed with mean 1 while the length of the censored interval is also Exponentially distributed with mean 1. The covariate, Z is generated from a Binomial distribution. The covariate, $Z(t)$ is a continuous time-dependent where $Z(t) = s + kt$ and $s = 1, k = 5$. The survival time can be simulated from Equation (5.2.19)

$$T = \left[\frac{1}{k}ln\left(1 + \frac{(1+a)(-ln(u))}{\alpha\beta\delta exp(\theta Z + \delta s)}\right)\right]^{1/1+\alpha} \tag{5.2.19}$$

where $u \sim Uniform(0, 1)$. The derivation of the expression is shown in Appendix D.

Three cases for the true covariate effects were considered. They are $(\theta, \delta) = (1.0, 1.0)$, $(1.0, 0.5)$ and $(0.5, 1.0)$. The true value for α and β are taken to be 2 and 1 respectively in all cases. In these $N = 100$ simulations, the samples were in average between 22% to 29% censored. The true values α, β, θ and δ were compared with their MLE's using the EM-algorithm described above. Table 5.4 reports the results from these simulations. The MLE's of all the parameters are close to the actual value with small estimated mean squared error *(EMSE)*. This approach yields reliable, useful and accurate results.

	α	β	θ	δ
True Value	2.0	1.0	1.0	1.0
MLE	1.9330	0.9223	0.9754	1.0643
EMSE	0.0360	0.0221	0.0189	0.0100
Censored Proportion	0.2288			
True Value	2.0	1.0	1.0	0.5
MLE	1.9895	1.0919	1.0038	0.5430
EMSE	0.0234	0.0673	0.0518	0.0064
Censored Proportion	0.2996			
True Value	2.0	1.0	0.5	1.0
MLE	1.9674	0.9905	0.5631	0.9872
EMSE	0.0252	0.0133	0.0117	0.0064
Censored Proportion	0.2376			

Table 5.4: Numerical results for Weibull AFT model in the presence of continuous time-dependent covariate

5.3 Semiparametric model

The extended Cox model is introduced in Section (5.1) which incorporates both time-independent and time-dependent covariates. The hazard function for the extended Cox model is given by

$$h(t|\mathbf{Z},\mathbf{Z}(t)) = h_0(t)exp\left(\theta^T\mathbf{Z} + \delta^T\mathbf{Z}(t)\right)$$

where $h_0(t)$ is the baseline hazard function. In this model, the baseline hazard function, $h_0(t)$ is interpreted as the hazard function for whom all the variables are zero at the time origin and remain at this same value through time.

The survival function for the model can be defined as

$$S(t|\mathbf{Z},\mathbf{Z}(t)) = exp\left(-\int_0^t h_0(u)exp(\theta^T\mathbf{Z} + \delta^T\mathbf{Z}(u))du\right)$$

This function depends on the baseline hazard function, $h_0(t)$, the fixed covariates $\mathbf{Z}$ and the time-dependent covariates $\mathbf{Z}(t)$ from time 0 to t. Hence, $S(t)$ depends on the future values for the time-dependent covariates which are generally unknown (Collett, 2003).

With this semiparametric set-up, the density of lifetimes is given by

$$f(t|\mathbf{Z},\mathbf{Z}(t)) \qquad = \qquad h_0(t)exp\left(\theta^T\mathbf{Z}+\delta^T\mathbf{Z}(t)-\int_0^t h_0(u)e^{\theta^T\mathbf{Z}+\delta^T\mathbf{Z}(u)}du\right) \quad (5.3.1)$$

Without loss of generality, let $t_1,\cdots,t_{n1},(l_{n_1+1},r_{n_1+1}),\cdots,(l_{n_1+n_2},r_{n_1+n_2})$ be the middle censored data from equation (5.3.1). Then the full likelihood is given by

$$L(\theta,\delta)=\prod_{uncens} f(t|\mathbf{Z},\mathbf{Z}(t))\prod_{cens}(S(l|\mathbf{Z},\mathbf{Z}(l))-S(r|\mathbf{Z},\mathbf{Z}(r)))$$

The corresponding log-likelihood is

$$l_{full}(\theta,\delta)=l_{uncens}(\theta,\delta)+l_{cens}(\theta,\delta)$$

where

$$l_{uncens}(\theta,\delta)=\sum_{i=1}^{n_1} ln(h_0(t_i))+\sum_{i=1}^{n_1}\left[\theta^T Z_i+\delta^T Z_i(t_i)-\int_0^t h_0(u_i)e^{\theta^T Z_i+\delta^T Z_i(u_i)}du\right] \quad (5.3.2)$$

$$l_{cens}(\theta,\delta)=\sum_{j=1}^{n_2} ln\left(e^{-\int_0^l h_0(u_i)e^{\theta^T Z_i+\delta^T Z_i(u_i)}du}-e^{-\int_0^r h_0(u_i)e^{\theta^T Z_i+\delta^T Z_i(u_i))}du}\right) \quad (5.3.3)$$

From Equations (5.3.2) and (5.3.3), the estimation of the baseline hazard function, $h_0(t)$ is required to estimate the covariate effect θ and δ. One approach to estimate the baseline hazard function is to fit a smoothing spline. See Wang (2011) for details on smoothing spline.

To obtain the MLE of θ, find the derivative of the log-likelihood

$$\frac{\partial}{\partial\theta}l(\theta,\delta)=\frac{\partial}{\partial\theta}l_{uncens}(\theta,\delta)+\frac{\partial}{\partial\theta}l_{cens}(\theta,\delta) \quad (5.3.4)$$

where

$$\frac{\partial}{\partial\theta} l_{uncens}(\theta,\delta) = \sum_{i=1}^{n_1} Z_i - \left[\sum_{i=1}^{n_1} \int_0^{t_i} h_0(u_i) Z_i exp(\theta^T Z_i + \delta^T Z_i(u_i)) du \right] \tag{5.3.5}$$

$$\frac{\partial}{\partial\theta} l_{cens}(\theta,\delta) = \sum_{j=1}^{n_2} \left(\frac{\frac{\partial}{\partial\theta}(e^{-\int_0^{l_j} h_0(u_j) e^{\theta^T Z_j + \delta^T Z_j(u_j)} du} - e^{-\int_0^{r_j} h_0(u_j) e^{\theta^T Z_j + \delta^T Z_j(u_j))} du})}{e^{-\int_0^{l_j} h_0(u_j) e^{\theta^T Z_j + \delta^T Z_j(u_j)} du} - e^{-\int_0^{r_j} h_0(u_j) e^{\theta^T Z_j + \delta^T Z_j(u_j))} du}} \right) \tag{5.3.6}$$

Similarly, to find the MLE of δ, find the derivative of the log-likelihood

$$\frac{\partial}{\partial\delta} l(\theta,\delta) = \frac{\partial}{\partial\delta} l_{uncens}(\theta,\delta) + \frac{\partial}{\partial\delta} l_{cens}(\theta,\delta) \tag{5.3.7}$$

where

$$\frac{\partial}{\partial\delta} l_{uncens}(\theta,\delta) = \sum_{i=1}^{n_1} Z_i(t_i) - \left[\sum_{i=1}^{n_1} \int_0^{t_i} h_0(u_i) Z_i(u_i) exp(\theta^T Z_i + \delta^T Z_i(u_i)) du \right] \tag{5.3.8}$$

$$\frac{\partial}{\partial\delta} l_{cens}(\theta,\delta) = \sum_{j=1}^{n_2} \left(\frac{\frac{\partial}{\partial\delta}(e^{-\int_0^{l_j} h_0(u_j) e^{\theta^T Z_j + \delta^T Z_j(u_j)} du} - e^{-\int_0^{r_i} h_0(u_j) e^{\theta^T Z_j + \delta^T Z_j(u_j))} du})}{e^{-\int_0^{l_i} h_0(u_i) e^{\theta^T Z_j + \delta^T Z_j(u_j)} du} - e^{-\int_0^{r_i} h_0(u_j) e^{\theta^T Z_j + \delta^T Z_j(u_j))} du}} \right) \tag{5.3.9}$$

While there is no general closed form solution to Equations (5.3.4) and (5.3.7), the maximum likelihood estimate of θ and δ can be solved numerically. To do this, first estimate $h_0(t)$ by fitting a smoothing spline.

Then, the EM algorithm can now be set up as follows. To make this more explicit, choose $(\theta,\delta)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(\theta,\delta)_{(j)}$ is the jth estimate
- Step 2: Compute $g(T_i^*) = E[g(T_i|a_i < g(T_i) < b_i), \theta = \theta_{(j)}, \delta = \delta_{(j)}]$
- Step 3: Solve the log-likelihood in the M-step with the $g(T_i)^*$'s imputed for the censored observations for its maximum and set $(\theta,\delta)_{(j+1)}$ as the values that maximize that equation
- Step 4: Repeat until convergence criteria is met

5.3.1 A simulation study

A simulation study is performed to illustrate the usefulness of this approach. Simulations are carried out in R using $N = 100$ replications with sample sizes $n = 50, 100, 250$ and 500. The censoring mechanism is that the left endpoint of the censored interval is exponentially distributed with mean 1 while the length of the censored interval is also exponentially distributed with mean 1. The true covariate effects are $\theta = 1$ and $\delta = 1$ in all cases and t_0 is taken to be exponential distributed with rate 0.5. The covariate, Z is generated from a binomial distribution. The covariate, $Z(t)$ is a dichotomous time-dependent covariate that can change at most once from untreated to treated. Let t_0 denote the time at which the time-varying covariate changes from unexposed $(Z = 0)$ to exposed $(Z = 1)$. Hence, we have

$$Z(t) = \begin{cases} 0 & \text{if } t < t_0 \\ \\ 1 & \text{if } t \geq t_0 \end{cases}$$

Again, suppose that there is one time-dependent covariate, Z and one time-dependent covariate denoted by $Z(t)$. Let θ be the vector of regression coefficients associated with the vector of fixed covariates, $Z(t)$. Suppose that the baseline hazard function is a cubic spline as in equation (5.3.10).

$$h_0(t) = 1.125 + 2.75(t - 0.5) + 1.5(t - 0.5)^2 + (t - 0.5)^3 \tag{5.3.10}$$

Hence, the hazard function, $h(t)$ is given as

$$h(t|Z, Z(t)) = (1.125 + 2.75(t - 0.5) + 1.5(t - 0.5)^2 + (t - 0.5)^3)exp(\theta Z + \delta Z(t))$$

and the cumulative hazard function, $H(t)$ is given as

$$H(t|Z, Z(t)) = \begin{cases} exp(\theta Z)\left(t^2 + \frac{t^4}{4}\right) & \text{if } t < t_0 \\ \\ \left(t_0^2 + \frac{t_0^4}{4}\right)[exp(\theta Z) - exp(\theta Z + \delta)] + \left(t^2 + \frac{t^4}{4}\right)exp(\theta Z + \delta) & \text{if } t \geq t_0 \end{cases}$$

The survival function of the above model is then given as

$$S(t|Z,Z(t)) = \begin{cases} exp\left(-e^{\theta Z}\left(t^2+\frac{t^4}{4}\right)\right) & \text{if } t < t_0 \\ exp\left(-\left(t_0^2+\frac{t_0^4}{4}\right)\left[e^{\theta Z}-e^{\theta Z+\delta}\right]-\left(t^2+\frac{t^4}{4}\right)e^{\theta Z+\delta}\right) & \text{if } t \geq t_0 \end{cases}$$

where $S(t)$ is defined as $S(t) = exp(-H(t))$. The survival function can be rewritten as

$$S(t|Z,Z(t)) = exp\left[-e^{\theta Z}\left[\left(t^2+\frac{t^4}{4}\right)\left[1-Z(t)+e^{\delta}Z(t)\right]+\left(t_0^2+\frac{t_0^4}{4}\right)(1-e^{\delta})Z(t)\right]\right]$$

Hence, the density of lifetimes is given by

$$\begin{aligned} f(t|Z,&Z(t)) = (t^3+2t)(e^{\delta}Z(t)-Z(t)+1)\times \\ &exp\left[-e^{\theta Z}\left(t_0^2+\frac{t_0^4}{4}\right)\left(1-e^{\delta}\right)Z(t)+\theta Z-\left(t^2+\frac{t^4}{4}\right)\left(e^{\delta}Z(t)-Z(t)+1\right)e^{\theta Z}\right] \end{aligned}$$

Now, assume that there is at least one censored observation i.e. $n_2 > 0$ and consider the censoring mechanism given in Section (2.2). The log-likelihood is

$$\begin{aligned} l(t|Z,Z(t)) &= \sum_{i=1}^{n_1} ln(t_i^3+2t_i)+\sum_{i=1}^{n_1} ln[e^{\delta}Z_i(t_i)-Z_i(t_i)+1]+\sum_{i=1}^{n_1}\theta Z_i \\ &+\sum_{i=1}^{n_1}\left[-e^{\theta Z_i}\left(t_{i0}^2+\frac{t_{i0}^4}{4}\right)\left(1-e^{\delta}\right)Z_i(t_i)-\left(t_i^2+\frac{t_i^4}{4}\right)\left(e^{\delta}Z_i(t_i)-Z_i(t_i)+1\right)e^{\theta Z_i}\right] \\ &+\sum_{i=n_1+1}^{n} ln[S(l|Z_i,Z_i(l_i))-S(r_i|Z_i,Z_i(r_i))] \end{aligned}$$

Applying the EM algorithm in the same fashion as Section (2.2), the following conditional expectations required are

$$t^* = \frac{\int_l^r tf(t|Z,Z(t))dt}{S(l|Z,Z(l))-S(r|Z,Z(r))} \tag{5.3.11}$$

$$t^{2*} = \frac{\int_l^r t^2f(t|Z,Z(t))dt}{S(l|Z,Z(l))-S(r|Z,Z(r))} \tag{5.3.12}$$

$$t^{3*} = \frac{\int_l^r t^3f(t|Z,Z(t))dt}{S(l|Z,Z(l))-S(r|Z,Z(r))} \tag{5.3.13}$$

$$t^{4*} = \frac{\int_l^r t^4f(t|Z,Z(t))dt}{S(l|Z,Z(l))-S(r|Z,Z(r))} \tag{5.3.14}$$

Then the log-likelihood required is

$$\begin{aligned} l^*(\alpha,\beta,\theta,\delta) = & \sum_{i=1}^{n_1} log(t_i^3+2t_i) + \sum_{i=n_1+1}^{n} log(t_i^{3*}+2t_i^*) + \sum_{i=1}^{n} log[e^{\delta}Z_i(t_i) - Z_i(t_i) + 1] \\ & + \sum_{i=1}^{n}\theta Z_i + \sum_{i=1}^{n}\left[-e^{\theta Z_i}\left(t_{i0}^2+\frac{t_{i0}^4}{4}\right)(1-e^{\delta})Z_i(t_i)\right] \\ & - \sum_{i=1}^{n_1}\left(t_i^2+\frac{t_i^4}{4}\right)\left(e^{\delta}Z_i(t_i)-Z_i(t_i)+1\right)e^{\theta Z_i} \\ & - \sum_{i=n_1+1}^{n}\left(t_i^{2*}+\frac{t_i^{4*}}{4}\right)\left(e^{\delta}Z_i(t_i)-Z_i(t_i)+1\right)e^{\theta Z_i} \end{aligned} \tag{5.3.15}$$

where the t_i^*'s, t_i^{2*}'s,t_i^{3*}'s and t_i^{4*}'s are found using Equations (5.3.11), (5.3.12), (5.3.13) and (5.3.14) respectively.

The survival time can be simulated from the following equations.

$$T = \begin{cases} \left[2\left(1-\frac{ln(u)}{e^{\theta Z}}\right)^{1/2}-2\right]^{1/2} & \text{if } -ln(u) < e^{\theta Z}\left(t_0^2+\frac{t_0^4}{4}\right) \\ \left[2\left(1+\frac{-ln(u)-\left(t_0^2+\frac{t_0^4}{4}\right)(e^{\theta Z}-e^{\theta Z+\delta})}{e^{\theta Z+\delta}}\right)^{1/2}-2\right]^{1/2} & \text{if } -ln(u) \geq e^{\theta Z}\left(t_0^2+\frac{t_0^4}{4}\right) \end{cases}$$

where $u \sim Uniform(0,1)$. The derivation of the expression is shown in Appendix E.

In these $N = 100$ simulations, the samples were in average between 29% to 30% censored. The true values θ and δ were compared with their MLE's using the EM algorithm described above. Table 5.5 reports the results from these simulations. The MLE's of all the parameters are close to the actual value with small estimated mean squared error *(EMSE)*.

n	50	100	250	500
$\hat{\theta}_1$	0.9047	0.9394	0.9589	0.9960
$\hat{\delta}_1$	0.9489	0.9623	1.0128	1.0106
$EMSE(\hat{\theta}_1)$	0.0007	0.0002	0.0000	0.0000
$EMSE(\hat{\delta}_1)$	0.0009	0.0005	0.0000	0.0000
$Censored$	0.2938	0.2969	0.3000	0.3042

Table 5.5: Simulations for semiparametric model with a dichotomous time-dependent covariate

5.4 A practical example

In a 1977 report on the Stanford Heart Transplant study, patients identified as being eligible for a heart transplant were followed until death or censorship (Crowley and Hu, 1977). The Stanford Heart Transplant program began in October 1967 where patients are admitted to the program after review by a committee and then they wait for donor hearts to become available. The data contains survival time in days of 103 patients, 69 of whom received transplants. For each patient in the program, there is a birth date (*birth.dt*), date of acceptance into the program (*accept.dt*), the date of transplant (*tx.date*), date of the end of follow-up (*fu.date*). The survival time is defined as *fu.date* - *accept.dt*, denoted by *futime* in days. The survival time is said to be censored (*fustat=0*) or uncensored (*fustat=1*) depending on whether *fu.date* is the date of death or the closing date of the study. The age of acceptance into the program is denoted by *age* in years. Previous open-heart surgery for each patient is denoted by the variable *Surgery*. The waiting period in days for a transplant recipient is denoted by *wait.time*. For each transplant recipient, there are three other variables (*mismatch, hla.a2, mscore*) of tissue-type mismatching. Patients are accepted into the study if physicians judge them suitable for heart transplant. When a donor becomes available, physicians choose the transplant recipients according to various medical criteria. The variable of interest here is heart transplant status (*transplant*) at time t defined as

$$\textit{transplant} = \begin{cases} 0 & \text{if the patient has not received the transplant by time } t \text{ i.e. if } t < \textit{wait.time} \\ \\ 1 & \text{if the patient has received the transplant at time } t \text{ i.e. if } t \geq \textit{wait.time} \end{cases} \tag{5.4.1}$$

Hence, for a patient who did not receive a transplant during the study, the value of *transplant* will remain 0 at all times. For a patient receiving a transplant, the value of *transplant* is 0 from the start of eligibility until the time at which the patient receives the transplant which then changes to 1 and remains 1 throughout the study.

We take the variables namely age at acceptance in years, *age* and heart transplant status, *transplant* as the covariates with respective regression coefficients θ and δ. For the complete data set it is observed that the maximum likelihood estimates of θ and δ are 0.6301 and 691.9513 respectively. In order to create a set of middle-censored data, we randomly choose several actual failure data and replace them by random censoring intervals. The data were censored by a random interval whose left end was an exponential random variable with mean 1 and the width was exponential with mean 15. It is found that 18.44% of data were censored resulting in 84 uncensored observations and 19 censored observations. We fit a cubic spline to estimate the baseline hazard function using just the uncensored observations. This is shown in figure (5.4.1). Applying the model given in Section 5.3, it is found that the estimates of the regression coefficients are $\hat{\theta} = 0.6511$ and $\hat{\delta} = 692.1225$. These estimates are close to the maximum likelihood estimates found using the complete data which shows that this approach yields reliable, useful and accurate results.

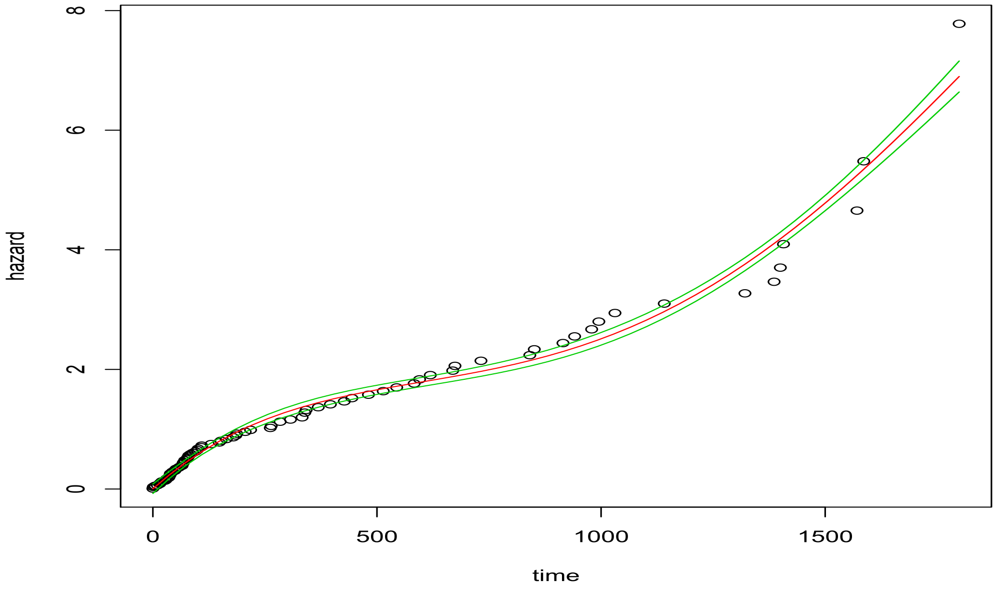


Figure 5.4.1: Baseline hazard function using cubic spline

It needs to be noted that interpretation of the parameter estimates needs to be done carefully depending on the type of time-dependent covariates, that is, if they are an internal covariate or an external covariate.

Chapter 6

Conclusions and Future Work

New methods for analyzing data subject to middle censoring when covariates are present are developed in this dissertation. In all cases, the EM algorithms are used which are theoretically shown to converge, and further demonstrated through simulations that the convergence is to estimates that are close to the true parameter values.

In Chapter 3, we considered inference for discrete lifetimes, when the data is middle-censored and extend it to the case when covariates are present. We also validate and confirm the estimation and inference procedures discussed, from extensive simulation studies, which show that the MLE of the regression coefficients are very close to the true values in all cases.

The competing risks model in Chapter 4 focuses only on one covariate. Again, it would be of interest to extend this methodology to include more than one covariate. Also, for the semiparametric model, only covariates from a Binomial distribution is explored in this dissertation. Thus, the next step would be to extend the existing methodology when the covariate comes from a continuous distribution. However, Bennett (2011) suggested to bin such continuous responses into discrete groups and transforming the continuous covariate into a categorical variable. With this approach, the methods that was discussed in this thesis could be used.

In Chapter 5, we study middle censoring in the presence of only one time-dependent covariate. The next logical step is to extend this methodology to include more than one covariate. Through an iterative method, these methods are assumed to work in high dimensions, but more work needs to be done. Two types of time-dependent covariates are considered here which are 1. A dichotomous time-dependent covariate that can change at most once from untreated to treated and 2. A continuous time-dependent covariate whose value is non-decreasing over time. However, another time-dependent covariate that may be interesting to explore is a dichotomous time-dependent covariate with multiple changes between treated and untreated which is common in survival analysis. This is the case where a subject may repeatedly move between untreated and treated conditions and assuming that all subjects are untreated at time, $t = 0$.

In this dissertation, we assumed that for each individual in the sample, there is a random censoring interval which are independent of the lifetimes. This is a pretty standard assumption equivalent to saying that the time at which an individual is censored has nothing to do with how long that individual lives. However, the assumption of independence between the survival time and the censoring time is open to debate. For example, in some clinical trials, potentially aggressive therapies may have side effects on patients depending on their tolerance, and the patients may need to be withdrawn from the clinic for a period of recovery. The time of withdrawal in such case depends on the risk of time-to-event and if the event happens within the censoring interval, dependent middle censored data is encountered. Davarzani and Parsian (2010) considered dependent right-censoring where the lifetime and censoring variables have a Marshall-Olkin Bivariate Exponential distribution. Following this context, research could be done for dependent middle-censored data with the presence of covariates.

Bayesian approach to middle censoring in both the parametric models and semi-parametric models could be a very interesting topic to consider. For the exponential and geometric models respectively, Iyer, Jammalamadaka and Kundu (2008) and Davarzani and Parsian (2011) considered Bayes approaches using prior distributions on all the parameters and evaluating the performance of the Bayes estimates obtained by doing this. That would be an interesting extension which is worth pursuing.

Appendix A

Exponential distribution lifetimes with a dichotomous time-dependent covariate

If the event times follow an exponential distribution, then when $t < t_0$,

$$\begin{aligned} H(t|Z, Z(t)) &= \int_0^t \lambda(exp(\theta Z + \delta Z(u)))du \\ &= \int_0^t \lambda exp(\theta Z)du \\ &= \lambda exp(\theta Z)t \end{aligned}$$

When $t \geq t_0$

$$\begin{aligned} H(t|Z, Z(t)) &= \int_0^t \lambda(exp(\theta Z + \delta Z(u)))du \\ &= \lambda exp(\theta Z) \int_0^t exp(\delta Z(u))du \\ &= \lambda exp(\theta Z) \left[\int_0^{t_0} exp(\delta Z(u))du + \int_{t_0}^t exp(\delta Z(u))du \right] \\ &= \lambda exp(\theta Z) \left[\int_0^{t_0} du + \int_{t_0}^t exp(\delta)du \right] \\ &= \lambda exp(\theta Z)(t_0 + exp(\delta)t - exp(\delta)t_0) \end{aligned}$$

Appendix A. Exponential distribution lifetimes with a dichotomous time-dependent covariate

The domain of $H(t|Z, Z(t))$ can be partitioned into two mutually exclusive intervals: $A_1 = (0, t_0)$ and $A_2 = [t_0, \infty)$. The range of the cumulative hazard function over each of A_1 and A_2 are $R_1 = (0, \lambda exp(\theta Z)t_0)$ and $R_2 = [\lambda exp(\theta Z)t_0, \infty)$ respectively.

When $H(t|Z, Z(t)) < \lambda exp(\theta Z)t_0$, we have

$$t = \frac{H(t|Z, Z(t))}{\lambda exp(\theta Z)}$$

Hence, the inverse cumulative hazard function

$$H^{-1}(t|Z, Z(t)) = \frac{t}{\lambda exp(\theta Z)}$$

if $t < \lambda exp(\theta Z)t_0$

When $H(t|Z, Z(t)) \geq \lambda exp(\theta Z)t_0$,

$$t = \frac{H(t|Z, Z(t)) - \lambda exp(\theta Z)t_0 + \lambda exp(\theta Z + \delta)t_0}{\lambda exp(\theta Z + \delta)}$$

The inverse cumulative hazard function is given as

$$H^{-1}(t|Z, Z(t)) = \frac{t - \lambda exp(\theta Z)t_0 + \lambda exp(\theta Z + \delta)t_0}{\lambda exp(\theta Z + \delta)}$$

if $t \geq \lambda exp(\theta Z)t_0$

Hence, the survival time can be simulated from

$$T = \begin{cases} \frac{-log(u)}{\lambda exp(\theta Z)} & \text{if } -log(u) < \lambda exp(\theta Z)t_0 \\ \\ \frac{-log(u) - \lambda exp(\theta Z)t_0 + \lambda exp(\theta Z + \delta)t_0}{\lambda exp(\theta Z + \delta)} & \text{if } -log(u) \geq \lambda exp(\theta Z)t_0 \end{cases}$$

where $u \sim Uniform(0, 1)$.

Appendix B

Exponential distribution lifetimes with continuous time-dependent covariate

If the event times follow an exponential distribution, then

$$\begin{aligned}
H(t|Z, Z(t)) &= \int_0^t \lambda(exp(\theta Z + \delta Z(u)))du \\
&= \lambda exp(\theta Z) \int_0^t exp(\delta Z(u))du \\
&= \lambda exp(\theta Z) \int_0^t exp(\delta(s + ku))du \\
&= \frac{\lambda exp(\theta Z + \delta s)}{\delta k} \left[exp(\delta kt) - 1\right]
\end{aligned}$$

We then have

$$t = \frac{1}{\delta k} ln \left(1 + \frac{\delta k(H(t|Z, Z(t)))}{\lambda exp(\theta Z + \delta s)}\right)$$

The corresponding inverse cumulative hazard function is

$$H^{-1}(t|Z, Z(t)) = \frac{1}{\delta k} ln \left(1 + \frac{\delta k(t)}{\lambda exp(\theta Z + \delta s)}\right)$$

where $u \sim Uniform(0, 1)$ and $k > 0$.

Hence, the survival time can be simulated from

$$T = \frac{1}{\delta k} ln \left(1 + \frac{\delta k(-ln(u))}{\lambda exp(\theta Z + \delta s)}\right)$$

Appendix C

Weibull distribution lifetimes with a dichotomous time-dependent covariate

If the event times follow a Weibull distribution, then when $t < t_0$,

$$\begin{aligned} H(t|Z, Z(t)) &= \int_0^t \alpha\beta t^{\alpha-1} exp(\theta Z + \delta Z(u))du \\ &= \beta exp(\theta Z) \int_0^t \alpha u^{\alpha-1} du \\ &= \beta exp(\theta Z) t^{\alpha} \end{aligned}$$

When $t \geq 0$

$$\begin{aligned} H(t|Z, Z(t)) &= \int_0^t \alpha\beta u^{\alpha-1}(exp(\theta Z + \delta Z(u)))du \\ &= \alpha\beta exp(\theta Z) \int_0^t u^{\alpha-1} exp(\delta Z(u))du \\ &= \alpha\beta exp(\theta Z) \left[\int_0^{t_0} u^{\alpha-1} exp(\delta Z(u))du + \int_{t_0}^t u^{\alpha-1} exp(\delta Z(u))du \right] \\ &= \alpha\beta exp(\theta Z) \left[\frac{t_0^{\alpha}}{\alpha} + exp(\delta) \left(\frac{t^{\alpha}}{\alpha} - \frac{t_0^{\alpha}}{\alpha} \right) \right] \\ &= \beta exp(\theta Z)(t_0^{\alpha} + exp(\delta)t^{\alpha} - exp(\delta)t_0^{\alpha}) \end{aligned}$$

The domain of $H(t|Z, Z(t))$ can be partitioned into two mutually exclusive intervals: $A_1 = (0, t_0)$ and $A_2 = [t_0, \infty)$. The range of the cumulative hazard function over each of A_1 and A_2 are $R_1 = (0, \beta exp(\theta Z)t_0^{\alpha})$ and $R_2 = [\beta exp(\theta Z)t_0^{\alpha}, \infty)$ respectively.

When $H(t|Z, Z(t)) < \beta exp(\theta Z)t_0^{\alpha}$, we have

$$t = \left(\frac{H(t|Z, Z(t))}{\beta exp(\theta Z)}\right)^{1/\alpha}$$

Hence, the inverse cumulative hazard function

$$H^{-1}(t|Z, Z(t)) = \left(\frac{t}{\beta exp(\theta Z)}\right)^{1/\alpha}$$

if $t < \beta exp(\theta Z)t_0^{\alpha}$

When $H(t|Z, Z(t)) \geq \beta exp(\theta Z)t_0^{\alpha}$,

$$t = \left(\frac{H(t|Z, Z(t)) - \beta exp(\theta Z)t_0^{\alpha} + \beta exp(\theta Z + \delta)t_0^{\alpha}}{\beta exp(\theta Z + \delta)}\right)^{1/\alpha}$$

The inverse cumulative hazard function is given as

$$H^{-1}(t|Z, Z(t)) = \left(\frac{t - \beta exp(\theta Z)t_0^{\alpha} + \beta exp(\theta Z + \delta)t_0^{\alpha}}{\beta exp(\theta Z + \delta)}\right)^{1/\alpha}$$

if $t \geq \beta exp(\theta Z)t_0^{\alpha}$

Hence, the survival time can be simulated from

$$T = \begin{cases} \left(\frac{-log(u)}{\beta exp(\theta Z)}\right)^{1/\alpha} & \text{if } -log(u) < \beta exp(\theta^T Z)t_0^{\alpha} \\ \left(\frac{-log(u) - \beta exp(\theta Z)t_0^{\alpha} + \beta exp(\theta Z + \delta)t_0^{\alpha}}{\beta exp(\theta Z + \delta)}\right)^{1/\alpha} & \text{if } -log(u) \geq \beta exp(\theta Z)t_0^{\alpha} \end{cases}$$

where $u \sim Uniform(0, 1)$.

Appendix D

Weibull distribution lifetimes with continuous time-dependent covariate

If the event times follow a Weibull distribution, then

$$\begin{aligned} H(t|Z, Z(t)) &= \int_0^t \alpha\beta u^{\alpha-1} exp(\theta Z + \delta Z(u))du \\ &= \alpha\beta exp(\theta Z) \int_0^t u^{\alpha-1} exp(\delta(s + ku))du \\ &= \frac{\alpha\beta\delta exp(\theta Z + \delta s)}{1+\alpha} \left[exp(kt^{1+\alpha}) - 1\right] \end{aligned}$$

We then have

$$t = \left[\frac{1}{k} ln\left(1 + \frac{(1+\alpha)H(t|Z, Z(t))}{\alpha\beta\delta exp(\theta Z + \delta s)}\right)\right]^{1/1+\alpha}$$

The corresponding inverse cumulative hazard function is

$$H^{-1}(t|Z, Z(t)) = \left[\frac{1}{k} ln\left(1 + \frac{(1+\alpha)t}{\alpha\beta\delta exp(\theta Z + \delta s)}\right)\right]^{1/1+\alpha}$$

where $u \sim Uniform(0, 1)$ and $k > 0$.

Hence, the survival time can be simulated from

$$T = \left[\frac{1}{k} ln\left(1 + \frac{(1+\alpha)(-ln(u))}{\alpha\beta\delta exp(\theta Z + \delta s)}\right)\right]^{1/1+\alpha}$$

Appendix E

Semi-parametric model with cubic spline as the baseline hazard function

When the baseline hazard function is assumed to be a cubic spline, specifically

$$h_0(t) = 1.125 + 2.75(t - 0.5) + 1.5(t - 0.5)^2 + (t - 0.5)^3$$

then when $t < t_0$,

$$\begin{aligned} H(t|Z, Z(t)) &= \int_0^t \left(1.125 + 2.75(u - 0.5) + 1.5(u - 0.5)^2 + (u - 0.5)^3\right) exp(\theta Z + \delta Z(u))du \\ &= \int_0^t \left(1.125 + 2.75(u - 0.5) + 1.5(u - 0.5)^2 + (u - 0.5)^3\right) exp(\theta Z)du \\ &= exp(\theta Z) \left(t^2 + \frac{t^4}{4}\right) \end{aligned}$$

When $t \geq t_0$

$$\begin{aligned} H(t|Z, Z(t)) &= \int_0^t \left(1.125 + 2.75(u - 0.5) + 1.5(u - 0.5)^2 + (u - 0.5)^3\right) exp(\theta Z + \delta Z(u))du \\ &= exp(\theta Z) \int_0^t \left(1.125 + 2.75(u - 0.5) + 1.5(u - 0.5)^2 + (u - 0.5)^3\right) exp(\delta Z(u))du \\ &= exp(\theta Z)[\int_0^{t_0} \left(1.125 + 2.75(u - 0.5) + 1.5(u - 0.5)^2 + (u - 0.5)^3\right) du \\ &+ \int_{t_0}^t \left(1.125 + 2.75(u - 0.5) + 1.5(u - 0.5)^2 + (u - 0.5)^3\right) exp(\delta)du] \\ &= \left(t_0^2 + \frac{t_0^4}{4}\right) [exp(\theta Z) - exp(\theta Z + \delta)] + \left(t^2 + \frac{t^4}{4}\right) exp(\theta Z + \delta) \end{aligned}$$

The domain of $H(t|Z, Z(t))$ can be partitioned into two mutually exclusive intervals: $A_1 = (0, t_0)$ and $A_2 = [t_0, \infty)$. The range of the cumulative hazard function over each of A_1 and A_2 are $R_1 = \left(0, exp(\theta Z)\left(t_0^2 + \frac{t_0^4}{4}\right)\right)$ and $R_2 = \left[exp(\theta Z)\left(t_0^2 + \frac{t_0^4}{4}\right), \infty\right)$ respectively.

When $H(t|Z, Z(t)) < exp(\theta Z)\left(t_0^2 + \frac{t_0^4}{4}\right)$, we have

$$t = \left[2\left(1 + \frac{H(t)}{exp(\theta Z)}\right)^{1/2} - 2\right]^{1/2}$$

Hence, the inverse cumulative hazard function

$$H^{-1}(t|Z, Z(t)) = \left[2\left(1 + \frac{t}{exp(\theta Z)}\right)^{1/2} - 2\right]^{1/2}$$

if $t < exp(\theta Z)\left(t_0^2 + \frac{t_0^4}{4}\right)$

When $H(t|Z, Z(t)) \geq exp(\theta Z)\left(t_0^2 + \frac{t_0^4}{4}\right)$,

$$t = \left[2\left(1 + \frac{H(t) - \left(t_0^2 + \frac{t_0^4}{4}\right)[exp(\theta Z) - exp(\theta Z + \delta)]}{exp(\theta Z + \delta)}\right)^{1/2} - 2\right]^{1/2}$$

The inverse cumulative hazard function is given as

$$H^{-1}(t|Z, Z(t)) = \left[2\left(1 + \frac{t - \left(t_0^2 + \frac{t_0^4}{4}\right)[exp(\theta Z) - exp(\theta Z + \delta)]}{exp(\theta Z + \delta)}\right)^{1/2} - 2\right]^{1/2}$$

if $t \geq exp(\theta Z)\left(t_0^2 + \frac{t_0^4}{4}\right)$

Hence, the survival time can be simulated from

$$T = \begin{cases} \left[2\left(1 + \frac{-ln(u)}{e^{\theta Z}}\right)^{1/2} - 2\right]^{1/2} & \text{if } -ln(u) < e^{\theta Z}\left(t_0 + \frac{t_0^2}{2}\right) \\ \left[2\left(1 + \frac{-ln(u) - \left(t_0^2 + \frac{t_0^4}{4}\right)[e^{\theta Z} - e^{\theta Z + \delta}]}{e^{\theta Z + \delta}}\right)^{1/2} - 2\right]^{1/2} & \text{if } -ln(u) \geq e^{\theta Z}\left(t_0 + \frac{t_0^2}{2}\right) \end{cases}$$

where $u \sim Uniform(0, 1)$.

Bibliography

O.O. Aalen. Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6(4):701–726, 1978a.

O.O. Aalen and S. Johansen. An empirical transition matrix for nonhomogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):141–150, 1978.

P.D. Allison. *Survival Analysis using SAS: A practical guide*. SAS Publishing, NC, 1995.

P.K Andersen, O Borgan, Gill R.D, and Keiding N. *Statistical Models Based on Counting Processes*. Springer Publishers, New York, 1993.

P.K Andersen, O Borgan, Gill R.D, and Keiding N. *Statistical Models Based on Counting Processes*. Springer Publishers, New York, 1995.

P.C. Austin. Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31:3946–3958, 2012.

N.A. Bennett. *Some Contributions to Middle-Censoring*. PhD thesis, University of California Santa Barbara, 2011.

J. Berkson and L. Elveback. Competing exponential risks with particular inference to the study of smoking lung cancer. *Journal of the American Statistical Association*, 55(291):415–428, 1960.

J.D. Buckley and I.R. James. Linear regression with censored data. *Biometrika*, 66(3):429–436, 1979.

D.R. Cox. The analysis of exponentially distributed life-times with two types of failures. *Journal of the Royal Statistical Society B*, 21(2):411–421, 1959.

D.R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Societ, Series B*, 34:187–220, 1972.

D.R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

D.R. Cox and D. Oakes. *Analysis of Survival Data*. Chapman Hall, London, U.K., 1984.

J. Crowley and M. Hu. Covariance analysis of heart transplant data. *Journal of the American Statistical Association*, 72(357):27–36, 1977.

N. Davarzani and A. Parsian. Bayesian inference in dependent right censoring. *Communications in Statistics - Theory and Methods*, 39:1270–1288, 2010.

N. Davarzani and A. Parsian. Statistical inference for discrete middle-censored data. *Journal of Statistical Planning and Inference*, 141(4):1455–1462, 2011.

H.A. David and M.L. Moeschberger. *The Theory of Competing Risks*. Macmillan Publishing, London, U.K., 1978.

B. Efron. The two-sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 4:831–853, 1967.

B. Efron and D.V. Hinkley. Assessing the accuracy of the maxmimum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–482, 1978.

L.D. Fisher and D.Y. Lin. Time dependent covariates in the cox proportional-hazards regression model. *Annual Review of Public Health*, 20:145–157, 1999.

T.R. Fleming and D.P. Harrington. *Counting Process and Survival Analysis*. John Wiley & Sons, Inc., New York, 1991.

T.R. Fleming and D.P. Harrington. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser, Basel, 1992.

R.D. Gill. Censoring and stochastic integrals. *Statistica Neerlandica*, 34(2):124–124, 1980.

B. Gompertz. On the nature of the function expressive of the law of human mortality and on the new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London A*, 115:513–583, 1825.

M. Greenwood. The natural duration of cancer, reports on public health and medical subjects. *Reports on Public Health and Medical Subjects*, 33:1–26, 1926.

J. Huang. Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, 9:501–519, 1999.

S.K. Iyer, S.Rao Jammalamadaka, and D. Kundu. Analysis of middle-censored data with exponential lifetime distributions. *Journal of Statistical Planning and Inference*, 138(11):3550–3560, 2008.

I.R. James and P.J. Smith. Consistency results for linear regression with censored data. *The Annals of Statistics*, 12(2):590–600, 1984.

S.Rao Jammalamadaka and S.K. Iyer. Approximate self consistency for middle-censored data. *Journal of Statistical Planning and Inference*, 124(1):75–86, 2004.

S.Rao Jammalamadaka and V. Mangalam. Nonparametric estimation for middle-censored data. *Journal of Nonparametric Statistics*, 15(2):253–265, 2003.

S.Rao Jammalamadaka and V. Mangalam. A general censoring scheme for circular data. *Statistical Methodology*, 6(3):280–289, 2009.

J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Inc., New York, second edition, 2002.

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

J.P. Klein. Small-sample moments of some estimators of the variance of the kaplan-meier and nelson-aalen estimators. *Scandinavian Journal of Statistics*, 18(4):333–340, 1991.

J.P. Klein and M.L. Moeschberger. *Survival analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2003.

D.G Kleinbaum and M. Klein. *Logistic Regression - A self learning Text-Third Edition*. Springer, New York, 2010.

D.G. Kleinbaum and M. Klein. *Survival analysis*. Springer, New York, 2012.

H. Koul, V. Susarla, and J. Van Ryzin. Regression analysis with randomly right censored data. *The Annals of Statistics*, 9(6):1276–1288, 1981.

T.L. Lai and Z. Ying. Large sample theory of a modified buckley-james estimator for regression analysis with censored data. *The Annals of Statistics*, 19(3):1370–1402, 1991.

T.A. Louis. Nonparametric analysis of an accelerated failure time model. *Biometrika*, 68(2):381–390, 1981.

E. Marubini and M.G. Valsecchi. *Analyzing Survival Data from Clinical Trials and Observational Studies*. John Wiley & Sons, Ltd., Chichester, U.K., 1995.

R.G. Miller. Least squares regression with censored data. *Biometrika*, 63(3):449–464, 1976.

M. Miyawaka. Analysis of incomplete data in competing risks model. *IEEE Transactions on Reliability*, 33(4):293–296, 1984.

W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–965, 1972.

W. Pan. Extending the iterative convex minorant algorithm to the cox model for interval-censored data. *Journal of Computational and Graphical Statistics*, 8(1):109–120, 1999.

A.P. Peterson. Expressing the kaplan-meier estimator as a function of empirical survival functions. *Journal of the American Statistical Association*, 72(360):854–858, 1977.

T. Peterson. Fitting parametric survival models with time-dependent covariates. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 35(3):281–288, 1986.

M. Pintilie. *Competing Risks, A Practical Perspective*. John Wiley & Sons, Ltd., Chichester, West Sussex, England, 2006.

Y. Ritov. Estimation in a linear regression model with censored data. *The Annals of Statistics*, 18(1):303–328, 1990.

Y. Ritov and J.A. Wellner. Censoring, martingales and the cox model. *Contemporary Mathematics*, 80:191–219, 1988.

S.R. Seaman and S.M. Bird. Proportional hazards model for interval-censored failure times and time-dependent covariates: application to hazard of hiv infection of injecting drug users in prison. *Statistics in medicine*, 20:1855–1870, 2001.

Y.H. Sparling, N. Younes, and J.M. Lachin. Parametric survival models for interval-censored data with time-dependent covariates. *Biostatistics*, 7(4):599–614, 2006.

J. Sun. *The Statistical Analysis of Interval-censored Failure Time Data*. Springer, New York, 2006.

T.M. Therneau and P.M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York, 2000.

W.Y. Tsai and J. Crowley. A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. *The Annals of Statistics*, 13(4):1317–1334, 1985.

A.A. Tsiatis. Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, 18(1):354–372, 1990.

B.W. Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69(345):169–173, 1974.

Y. Wang. *Smoothing Splines Methods and Applications*. Taylor and Francis Group, Florida, 2011.

L.J. Wei and M.H. Gail. Nonparametric estimation for a scale-change with censored observations. *Journal of the American Statistical Association*, 79(383):649–652, 1983.

L.J. Wei, Z. Ying, and D.Y. Lin. Linear regression analysis of censored survival data based on rank tests. *Biometrika*, 77(4):845–852, 1990.

C.F.J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Q. Yu, G.Y.C. Wong, and L. Li. Asymptotic properties of self-consistent estimators with mixed-interval censored data. *Annals of the Institute of Statistical Mathematics*, 53:469–486, 2001.

S.L. Zeger and C.Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, 1986.