Statistical Analysis of Failure Time Data with Missing Information

---

A Dissertation

Presented to

the Faculty of the Graduate School

University of Missouri-Columbia

In Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

---

by

Ping Chen

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

May, 2009

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

STATISTICAL ANALYSIS OF FAILURE TIME DATA WITH MISSING INFORMATION

Presented by Ping Chen,

a candidate for the degree of Doctor of Philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Tony Sun _____

Professor Tim Wright _____

Professor Subharup Guha _____

Professor Min Yang _____

Professor Shuguang Wang _____

ACKNOWLEDGEMENTS

This dissertation was made possible because of the guidance and support of Dr. Tony Sun. He took me under his wing during my second year in the graduate program, and he has been very patient and helpful. I learned a lot from him and he has inspired me a great during my work on this dissertation. I am very fortunate to have found a great advisor and mentor, to whom my deepest gratitude goes.

I would like to express special thanks to the faculty that served on my dissertation committee: Dr. Wright, Dr. Yang, Dr. Guha and Dr. Wang. Thank you for your time and effort. Your input provided me with valuable suggestions which helped me to refine and improve my dissertation research.

I also owe a debt of gratitude to all the faculty who taught me classes: Dr. Athanasios Micheas, Dr. Stas Kolenikov, Dr. Jing Qiu, Dr. Lori Thombs and Dr. Christie Spinka. Your teaching certainly will never be forgotten, and I will apply the knowledge I learned from you. I will also always treasure the friendship and appreciate the help of friends and classmates here at Mizzou. It is all of you that have made the past few years of life fun and enjoyable.

Finally, my husband, Gerhardt, has given steady support throughout my graduate program. He has always been my cheerleader on good days and bad days. And above all, I dedicate all of the effort that went into this study to my parents. Your love and encouragement is where I find my greatest strength to keep going.

Table of Contents

List of Tables

Regression Analysis of Failure Time Data with Missing Information

Ping Chen

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

Failure time data arise in many fields and can involve different types of censoring structures and missing information. We consider three cases: right-censored data with missing censoring indicators, clustered current status data, and clustered interval-censored data. Right-censored data with missing indicators appear when the censoring indicator, the information if the observed time is the survival time of interest or the censoring time, is missing. Clustered current status data arise when the failure times of interest are clustered into small groups and the observed times are either left- or right-censored. Clustered interval-censored data arise when the failure times of interest are clustered into small groups and the observed times are known to fall within certain intervals.

In Chapter 1, three real-life examples are discussed to illustrate right-censored data with missing censoring indicators, clustered current status data and clustered interval-censored data. Also we will review the existing literature on statistical analysis of right-censored failure time data with missing censoring indicators, current status data, interval-censored data and general clustered failure time data.

Chapter 2 discusses regression analysis of right-censored failure time data with missing censoring indicators and presents an efficient estimation procedure based on the EM algorithm. The simulation study performed indicates that the proposed methodology performs well for practical situations. An illustrative example from a breast cancer clinical trial is provided.

Chapter 3 discusses regression analysis of clustered current status data. For inference, a Cox frailty model and a two-step EM algorithm are presented. A simulation study was conducted for the evaluation of the proposed methodology and indicates that the approach performs well for practical situations. An illustrative example from a tumorigenicity experiment is provided.

Chapter 4 generalizes the study of Chapter 3 to clustered interval-censored data. For inference, similar Cox frailty model and two-steps EM algorithm are adopted. Due to the more complex structure of the censoring mechanism, the EM algorithm and the inference procedure are much more complicated for clustered interval-censored data. A simulation study indicates that the approach performs well for practical situations. An illustrative example from a lymphatic filariasis study is provided.

Chapter 5 discusses some directions for future research.

Chapter 1

Introduction

1.1 Introduction

In survival analysis, failure time is commonly defined as the time until the occurrence of some event. Examples of failure times include the death time of a subject and time to occurrences of tumors. Often we can not observe exact failure time due to various reasons including dropouts of study subjects or limited follow-ups. For example, nonlethal tumors in rats usually can not be observed until the rats die or are sacrificed. Death due to breast cancer can not be observed if other reasons caused the death of the individual or the individual left the study. Such incomplete observation of failure time is called censoring. A censoring time can be an examination (observation) time, dropout time, or the time when the study ends.

There are various types of censoring mechanisms: left censoring, right censoring, and interval censoring. Left censoring occurs when the failure has already appeared at the censoring time. Right censoring occurs when the failure has not yet appeared

at the censoring time. Let T_i be the failure time of interest for the ith subject in a study and C_i the subject's censoring time. Under right censoring, T_i is known to be greater than the observed C_i. Under left censoring, T_i is known to be smaller than the observed C_i. Interval censoring occurs when the failure time falls within a time interval. Under interval censoring, T_i is not observed exactly but only known to fall within some interval $(L_i, R_i]$, where L_i and R_i can be regarded as two censoring times for subject i.

If failure times of all subjects in the study are either right-censored or observed exactly, we have right-censored survival data. However, sometimes we do not know if the observed time is the failure time of interest or the censoring time for some subjects. This yields right-censored failure time data with missing censoring indicators. If the failure times of all subjects in the study are either right-censored or left-censored, we have current status survival data. If the failure times of all subjects in the study are interval-censored, we have interval-censored data. Right-censored data and current status data are special cases of interval-censored data. Current status data are also called case I interval-censored data. When subjects under study are from the same family or the same geographic area, there may be non-ignorable correlation among subjects. Current status data with such cluster structure are called clustered current status data. Interval-censored data with such cluster structure are called clustered interval-censored data.

In Section 1.2, we introduce three real life examples exhibiting right-censored data with missing censoring indicators, clustered current status data and clustered interval-

censored data, respectively. In Section 1.3, we discuss existing statistical methods for the analysis of right-censored failure time data with missing censoring indicators. In Section 1.4, we discuss existing statistical methods for the analysis of both independent and clustered interval-censored data with a brief introduction of frailty model. Finally the outline of this dissertation is given in Section 1.5.

1.2 Three Examples

1.2.1 A Stage II Breast Cancer Study

This example comes from a stage II breast cancer clinical trial conducted by the Eastern Cooperative Oncology Group. It was described in Cummings et al. (1986). In this study, 169 elderly women with stage II breast cancer were randomized to receive tamoxifen or placebo for 24 months in a double-blind adjuvant trial. They were stratified prior to randomization on the basis of the number of positive axillary nodes and the estrogen receptor status of their primary tumor. Patients either died from breast cancer, died from other causes or were still alive at the last observed time. The failure time of interest is the death time due to breast cancer. Censoring time is the death time due to other reasons or the last observation time when a subject was still alive. Unfortunately 18 patients had unknown cause of death and thus had missing censoring indicators. This data set is an example of right-censored failure time with missing censoring indicators. In Chapter 2, we will study the effects of the number of positive axillary nodes and the estrogen receptor status of their primary tumor on the

death rate due to breast cancer.

1.2.2 NTP Tumorigenicity Data

This data set comes from a part of an animal tumorigenicity experiment conducted by the National Toxicology Program (NTP). It is a 2-year rodent carcinogenicity study of chloroprene, in which subjects were F344/N rats and B6C3F1 mice of both sexes. The experiment was described in Dunson and Dinse (2002). The experiment contained a control group with no chloroprene and three dose groups with 50 rodents in each group. Rodents in the dose groups were exposed to chloroprene at the concentration of 12.8, 32, and 80 ppm, respectively, 6 hours per day, 5 days per week for up to 2 years. The occurrence of tumors was determined through a pathologic examination when the rodents died. Some rodents died during the study. Those rodents who did not die at the end of the 2-year study were sacrificed regardless of health condition. Since each rat was examined for tumor at the death time, the only information available is whether the rat had suffered adrenal tumor or lung tumor at that time, that is, the onset time of adrenal tumor and lung tumor were either left-censored or right-censored by the death time. The two tumor times of the same rat may be correlated. In Chapter 3, we will study the dose effects on the occurrence rate of both cancers.

1.2.3 Lymphatic Filariasis Data

This example is from a lymphatic filariasis study conducted in Recife, Brazil. Lymphatic filariasis is a debilitating parasitic disease and several worms live together in

several nests. A randomized trial was conducted to compare the effectiveness of co-administration of diethylcarbamazine (DEC)/ albendazole (ALB) versus DEC alone in killing the adult worms. (Dreyer et al., 2006). A total of 47 men participated in the study, 25 in the DEC group and 22 in the DEC/ALB group. The patients were periodically checked up by ultrasound for clearance of worms and thus the data are subject to interval-censoring. Since the nests within one patient are under the influence of the health condition of the same person, they are correlated. In Chapter 4, we will study the dose effects on the clearance rate of the nests of worms.

1.3 Analysis of Right–Censored Data with Missing Censoring Indicators

Consider a survival study that involves n independent subjects and produces right-censored failure time data. For subject i, let T_i denote the failure time of interest and C_i the censoring time with Z_i, a vector of associated covariates, $i = 1, ..., n$. Let $X_i = min(T_i, C_i)$ denote the only observed failure time. The censoring indicator $\delta_i = I(T_i \leq C_i)$ may be missing for some subjects. Define $\epsilon_i = 0$ if δ_i is missing and 1 otherwise. Suppose that the observed information consists of $\{X_i, \epsilon_i, \epsilon_i\delta_i\}$. That is, we have right-censored data with missing censoring indicators. By the terminology of Little and Rubin (1987), δ may be subject to different missing mechanisms. One missing mechanism is missing completely at random (MCAR), meaning that the missing is independent of all variables involved. Another mechanism is missing at random (MAR),

meaning that the missing could depend on observed failure time and/or covariates. We could observe three types of events: (i) $\epsilon = 0$ (ii) $\epsilon = 1$, $\delta = 1$ (iii) $\epsilon = 1$, $\delta = 0$. Define counting processes $N_{i00}(t) = (1 - \epsilon_i)I(X_i \leq t)$, $N_{i11} = \epsilon_i \delta_i I(X_i \leq t)$ and $N_{i10} = \epsilon_i(1 - \delta_i)I(X_i \leq t)$ and the risk process $Y_i(t) = I(X_i \geq t)$.

1.3.1 Analysis of Right–Censored Data with Censoring Indicators Missing Completely At Random

Several authors have investigated the analysis of right-censored failure time data when the censoring indicators are missing completely at random. For example, Lo (1991) considered the nonparametric estimate of the survivor function of $T, S(t) = P(T > t)$, and proposed the following nonparametric estimator:

$$\hat{S}_L(t) = \prod_{X_i \leq t} (1 - \frac{\epsilon_i \delta_i}{\sum_{j=1}^{n} Y_j(X_i)})^{1/\hat{\rho}},$$

where $\hat{\rho} = \sum_{i=1}^{n} \epsilon_i / n$ is the proportion of observed censoring indicators. It is easy to see that $\hat{S}_L$ jumps only at the uncensored failure time with known censoring indicators. Gijbels et al. (1993) observed that $\hat{S}_L$ does not use all of the information from individuals with $\epsilon = 0$, and they proposed a way to make better use of such information. They first estimated the cumulative hazard function of T and then estimated its survival function $S(t)$ by taking a product integral. Their estimator of the cumulative hazard function $\Lambda(t)$ can be expressed as:

$$\hat{\Lambda}_{GLY}(t) = \alpha(t)\hat{\Lambda}_1(t) + (1 - \alpha(t))\hat{\Lambda}_2(t),$$

$$\hat{\Lambda}_1(t) = \hat{\Lambda}_{11}(t)/\hat{\rho}, \quad \hat{\Lambda}_2(t) = \hat{\Lambda}_{00}(t)/(1-\hat{\rho}) - \hat{\Lambda}_{10}(t)/\hat{\rho},$$

$$\hat{\Lambda}_{jk}(t) = \int_0^t \frac{I[Y_{\cdot}(u) > 1]}{Y_{\cdot}(u)} dN_{\cdot jk}(u),$$

$$A_{\cdot}(u) = \sum_{i=1}^{n} A_i(u), \quad , N_{\cdot jk}(u) = \sum_{i=1}^{n} N_{ijk}(u),$$

where $0 \leq \alpha(t) \leq 1$ is a function that minimizes the asymptotic variances of the estimator, $\hat{\Lambda}_{jk}(t)$ is the Nelson-Aalen estimator corresponding to the counting process $N_{jk}(t)$ defined as above.

When some covariate effects are of interest, regression analysis is often performed. The Cox model is often used in regression analysis and assumes that the hazard function of T for subject i given covariates Z_i is

$$\lambda_i(t|Z_i) = \lambda_0(t) \exp(Z_i'\beta),$$

where β is the covariate effect and $\lambda_0(t)$ is an unknown baseline hazard function. To estimate β, Gijbels et al. (1993) proposed an estimating equation based on a weighted average of three partial likelihood score functions under the Cox model:

$$U_{11}(\beta, \infty) + D(\beta, \infty)[U_{00}(\beta, \infty) - \frac{1-\hat{\rho}}{\hat{\rho}} U_{10}(\beta, \infty)] = 0,$$

$$U_{jk}(\beta, \infty) = \sum_{i=1}^{n} \int_0^{\infty} [Z_i - \sum_{i=1}^{n} Z_i Y_i(t) e^{Z_i'\beta} / \sum_{i=1}^{n} Y_i(t) e^{Z_i'\beta}] dN_{ijk}(t),$$

where $D(\beta, \infty)$ is an optimal weight that minimizes the asymptotic variance of $\hat{\beta}$.

McKeague and Subramanian (1998) proposed a similar estimating equation method that does not depend on the statistics $\hat{\rho}$.

1.3.2 Analysis of Right–Censored Data with Censoring Indicators Missing At Random

Few authors have studied the situation in which the censoring indicators are missing at random for right-censored survival data. Goetghebeur and Ryan (1995) considered proportional hazards models for both failure time and censoring time with a proportional assumption between baseline hazard functions of the two. Let $\lambda_i(t|Z_i)$ and $h_i(t|Z_i)$ be the hazard functions of T_i and C_i for given covariates Z_i. They assumed

$$\lambda_i(t|Z_i) = \lambda(t)e^{Z_i'\beta}, \quad h_i(t|Z_i) = h(t)e^{Z_i'\gamma},$$

where $h(t)$ and $\lambda(t)$ are the baseline hazard functions, γ and β are regression coefficients. They also assumed $h(t) = \exp(\xi)\lambda(t)$, where ξ is an unknown constant.

For estimation, they proposed a method involving two partial likelihood functions. Specifically, let L denote the partial likelihood based on the conditional probability of a specific event given that one event of that type occurs. And let L^* denote the partial likelihood based on the conditional probability of a specific event given one

event occurs without knowing the type. We have

$$L = \prod_{t\geq 0}\prod_{i=1}^{n}\{\frac{w_{i11}}{\sum_{j=1}^{n} Y_j(t)w_{j11}}\}^{dN_{i11}(t)}\{\frac{w_{i10}}{\sum_{j=1}^{n} Y_j(t)w_{j10}}\}^{dN_{i10}(t)}$$
$$\{\frac{w_{i00}}{\sum_{j=1}^{n} Y_j(t)w_{j00}}\}^{dN_{i00}(t)},$$
$$L^* = \prod_{t\geq 0}\prod_{i=1}^{n}\frac{w_{i11}^{dN_{i11}(t)}w_{i10}^{dN_{i10}(t)}w_{i00}^{dN_{i00}(t)}}{\{\sum_{j=1}^{n} Y_j(t)w_{j00}\}^{dN_{i00}(t)+dN_{i11}(t)+dN_{i10}(t)}},$$

where

$$w_{i11} = \exp(Z_i'\beta), w_{i10} = \exp(\xi + Z_i'\gamma), w_{i00} = \exp(Z_i'\beta) + \exp(\xi + Z_i'\gamma).$$

They combined the score functions of L w.r.t β and γ and L^* w.r.t to ξ to give the estimating equation.

1.4 Analysis of Interval–Censored Failure time Data

Consider a survival study that involves n subjects and produces current status data. Let T_i denote the failure time of interest for subject i. Suppose that each subject is observed only once at time C_i and the observed information consists of only C_i and $\delta_i = I(T_i \leq C_i)$. Define the counting process $N_i^c(t) = I(C_i \leq t)$.

1.4.1 Nonparametric Analysis of Independent Interval-censored Data

It is often the case that subjects under the study are independent. The NPMLE of the cumulative distribution function of the survival time $T, F(t) = I(T \leq t)$, is often

important with independent observations.

1.4.1.1 Case I: Current Status Data

Let $C_{(i)}$ be the ith order statistic of $(C_1, \ldots, C_n)$ and let $\delta_{(i)}$ denote the corresponding indicator, i.e., if $C_{(i)} = C_j$, then $\delta_{(i)} = \delta_j$. Then the NPMLE $\hat{F}$ is the maximizer of the following log-likelihood function:

$$L(\tilde{x}) = \sum_{i=1}^{n} \{\delta_{(i)} log(x_i) + (1 - \delta_{(i)}) \log(1 - x_i)\}$$

under the condition $0 \leq x_1 \leq \ldots \leq x_n \leq 1$. The NPMLE $\hat{F}$ can be obtained using either the self-consistency algorithm or the greatest convex minorant algorithm described in Groeneboom and Wellner (1992) and Robertson et al. (1988). Also $\hat{F}$ can be represented by the max-min formula:

$$\hat{F}(C_{(i)}) = \max_{l \leq i} \min_{k \geq i} \frac{\sum_{j=1}^{k} \delta_{(j)}}{k - l + 1}.$$

1.4.1.2 Case II: Interval-censored Data

Suppose $(Li, Ri]$ is the observed interval for subject i to contain Ti. Let $s_j, j = 0, \cdots, m+1$ denote the distinct elements of $\{0; \{(L_i, R_i]\}, \infty\}$ in ascending order. Let α_{ij} be the indicator of the event $s_j \in (L_i; R_i]$ and $p_j = F(s_j) - F(s_{j-1})$. Under this setting, to find the NPMLE of F is equivalent to maximizing the following likelihood $L(p)$ with respect to p with constraints $\sum_{j=1}^{m+1} p_j = 1$ and $p_j > 0, j = 0, \cdots, m+1$.

$$L(p) = \prod_{i=1}^{n} \{F(R_i) - F(L_i)\} = \prod_{i=1}^{n} (\sum_{j=1}^{m+1} \alpha_{ij} p_j)$$

There is no close form for the NPMLE. Turnbull's estimator is commonly used in this situation and can be obtained through a self-consistency algorithm (Turnbull,1976). Turnbull's estimator is easy to implement, but has a slow convergence rate. Another problem to note is that a self-consistent estimator may not be the NPMLE. Gentleman and Geyer (1994) showed that the Kuhn-Tucker conditions are necessary and sufficient conditions for a self-consistent estimator to be the NPMLE by using standard convex optimization techniques. Another method to obtain the NPMLE is to apply the convex minorant algorithm introduced by Groeneboom and Wellner (1992), which converges faster than the self-consistency algorithm. However, the asymptotic distribution of the NPMLE is not established yet for case II interval-censored data although its consistency is already known.

1.4.2 Regression Analysis of Independent Interval–Censored Data

Many authors have discussed regression analysis of interval-censored data when all failure times involved are independent. There are several common semi-parametric models and have been widely studied.

1.4.2.1 Common Semi-parametric Regression Models

Commonly used semi-parametric regression models are: proportional hazards model (Cox model), additive hazards model and proportional odds model are commonly used.

Let $\lambda_i(t|Z_i)$ be the hazard functions of T_i for given covariates Z_i.

Proportional hazards model (Cox model) assumes:

$$\lambda_i(t|Z_i) = \lambda_0(t)e^{Z_i'\beta},$$

where $\lambda_0(t)$ is the unknown baseline hazard function and β is the regression coefficient.

Additive hazards model specifies:

$$\lambda_i(t|Z_i) = \lambda_0(t) + Z_i'\beta.$$

Proportional odds model defines:

$$\frac{1 - S(t|Z_i)}{S(t|Z_i)} = \frac{1 - S_0(t|Z_i)}{S_0(t|Z_i)}e^{Z_i'\beta},$$

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

1.4.2.2 Case I: Current Status Data

Many authors have studied regression analysis of current status data. Huang (1996) discussed the maximum likelihood estimation for the proportional hazards model based on current status data. Under the proportional hazard model as described in Section 1.3.1, the likelihood function is proportional to

$$l(\beta, \Lambda(t)) = \sum_{i=1}^{n} \log \left[\prod_{i=1}^{n} F^{\delta_i}(C_i|Z_i)\{(1 - F(C_i|Z_i))\}^{1-\delta_i} \right],$$

where $F(t|Z_i) = 1 - \exp(-\Lambda(t)e^{Z_i'\beta})$ is the cumulative distribution function. The MLE of $(\beta, \Lambda(t))$ can be calculated through a profile likelihood approach. $\hat{\Lambda}(t)$ is defined to be the NPMLE and can be estimated by iterative convex minorant algorithm (ICM) while fixing β. $\hat{\beta}$ can be estimated by solving score function of l w.r.t. β while fixing Λ. It is shown that the MLE of β is consistent and efficient and it has an asymptotic normal distribution with $n^{1/2}$ convergence rate.

The dimension of NPMLE $\hat{\Lambda}(t)$ will increase as sample size increases. Under some other models, optimization of likelihood w.r.t such NPMLE with high dimension could become difficult. Approximation of the function $\Lambda(t)$ such as sieve approximation with piecewise constant functions can be a solution to reduce the dimension. Rossini and Tsiatis (1998) studied this method under proportional odds model and established the consistency of β and $\hat{\Lambda}(t)$.

Some other authors developed methods for fitting the additive hazards model to the data. Lin et al. (1998) proposed an estimating equation method based on the proportional hazard property of the counting process of $\delta_i N_i^c(t)$ with $N_i^c(t)$ defined above. Ghosh (2001) proposed an efficient estimate procedure similar to that in Huang (1996). Martinussen and Scheike (2002) derived the efficient score function based on both counting processes $\delta_i N_i^c(t)$ and $(1 - \delta_i)N_i^c(t)$ and used the empirical version of the efficient score function as the estimating equation.

1.4.2.3 Case II: Interval–censored Data

There has been a lot of literature on the regression analysis of interval-censored data. Finkelstein, D. (1986) studied estimation under proportional hazards model for interval-censored data. $\hat{\Lambda}(t)$ is defined to be the NPMLE and it is estimated simultaneously with the regression coefficient by solving their score functions jointly. Kooperberg and Stone (1992) and Rosenberg (1995) gave spline-based estimators. More recently, Bebchuk and Betensky (2000) proposed multiple imputation approach, Betensky et al. (2002) explored a local likelihood method, and Cai and Betensky (2003) considered piecewise linear penalized spline. Huang (1997) studied linear sieve estimates under proportional odds model. Zhang, et al. (2005) considered a class of linear transformation models, which includes the proportional odds model as a special case, and proposed a method that does not need to estimate the baseline log odds.

1.4.3 Analysis of Clustered Failure Time Data

For regression analysis of clustered or multivariate failure time data, two approaches are commonly applied: the frailty model approach and the marginal model approach. The former provides a flexible approach for directly modeling the relationship between correlated failure times, while the latter focuses on covariate effects on individual failure times.

The frailty model assumes that for several related survival times T_i, there exists a positive random variable W such that the T_i's are conditionally independent of each

other given W and their conditional survival functions are

$$S_i(t|W) = S_{0i}(t)^W,$$

where $S_{0i}(t)$ is the baseline survival function for T_i. The frailty is usually given a marginal distribution such as gamma distribution. Oakes (1989) considered the estimates of several correlation measurements under the frailty model. The frailty model could be combined with various regression models such as the proportional hazards model and linear transformation models.

There has been a lot of literature on multivariate interval-censored failure time data. Goggins and Finkelstein (2000) and Kim and Xue (2002) considered the use of the marginal proportional hazards model, while Chen et al. (2007) and Tong et al. (2008) investigated the use of the marginal proportional odds model and the marginal additive hazards model, respectively. It is worth noting that multivariate failure time data can be seen as special cases of clustered data with the fixed and same cluster sizes. In general, the cluster size can differ from cluster to cluster.

Estimates under the frailty model have to involve the unknown frailty, which is often handled through means of an integral. The frailty is also often treated as missing data in the EM algorithm.

1.5 Outline

Here is the outline of the rest of the dissertation.

Chapter 2 discusses regression analysis of right-censored failure time data with missing censoring indicators and presents an efficient estimation procedure based on the EM algorithm. The simulation study performed indicates that the proposed methodology performs well for finite sample sizes. An illustrative example from a breast cancer clinical trial is provided.

Chapter 3 discusses regression analysis of clustered current status data. For inference, a Cox frailty model and a two-step EM algorithm are presented. A simulation study is conducted for the evaluation of the proposed methodology and indicates that the approach performs well for practical situations. An illustrative example from a tumorigenicity experiment is provided.

Chapter 4 generalizes the study of Chapter 3 to clustered interval-censored data. For inference, similar Cox frailty model and two-steps EM algorithm are adopted. Due to the more complex structure of the censoring mechanism, the EM algorithm and the inference procedure are much more complicated for clustered interval-censored data. A simulation study indicates that the approach performs well for practical situations. An illustrative example from a lymphatic filariasis study is provided.

Chapter 5 discusses some directions for future research.

Chapter 2

Regression Analysis of Right-censored Failure Time Data with Missing Censoring Indicators

2.1 Introduction

Failure time data arise in many fields and can involve different types of censoring and truncations or structures (Hougaard, 2000; Kalbfleisch and Prentice, 2002; Sun, 2006). One structure that can occur in practice and will be discussed in this chapter is that in the case of right-censored data, the censoring indicator could be missing due to various reasons. For example, in a bioassay experiment, some subjects might not be autopsied to save the expense or the results of an autopsy may be inconclusive. Another example is population mortality studies where relevant death certificate information can be missing due to emigration.

Several authors have investigated the analysis of right-censored failure time data when there exist missing censoring indicators. For a one sample problem with MCAR, for example, Dinse (1982) considered the nonparametric maximum likelihood estimate

(NPMLE) of a survival function and developed an EM algorithm. Lo (1991) discussed the same problem and showed that there exist infinitely many NPMLE and some of them can be inconsistent. In addition, Gijbels et al. (1993), Lo (1991) and McKeague and Subramanian (1998) gave some ad hoc estimators. Assuming the indicators are MAR, van der Laan and McKeague (1998) proposed a sieve NPMLE and showed that it is asymptotically efficient.

For regression analysis of right-censored failure time with missing censoring indicators, several approaches have been developed under MCAR or MAR (Gijbels et al., 1993; McKeague and Subramanian, 1998; Subramanian, 2000). In particular, Goetghebeur and Ryan (1995) studied situations where the missingness is MAR and both the failure time of interest and the censoring time follow the proportional hazards models marginally. For estimation of regression parameters, they developed an estimating equation approach and established the asymptotic properties of the resulting estimates. However, the approach assumed that the baseline hazard functions for the failure time of interest and the censoring time are proportional to each other, which may not be true in practice. In this chapter, we investigate the same problem and develop an approach that does not require the proportionality assumption.

This chapter is organized as follows. We will begin in Section 2.2 with describing the models and the assumptions that will be used throughout this chapter. In particular, we will assume that both the failure time of interest and the censoring time follow the proportional hazards models marginally. For estimation, we will adopt the sieve approach by approximating the two baseline hazard functions using linear functions.

Section 2.3 presents an EM algorithm for the maximum likelihood estimates of the parameters involved. The estimates of the regression parameters are consistent and have asymptotically normal distributions. Furthermore, they are efficient in that their covariance reaches the information bound. Some numerical results obtained from a simulation study are presented in Section 2.4 and indicate that the proposed approach works well for practical situations. The methodology is applied to a set of right-censored failure time arising from a stage II breast cancer study in Section 2.5 and Section 2.6 contains some concluding remarks.

2.2 Models and Assumptions

Consider a survival study that involves n independent subjects and produces right-censored failure time data. For subject i, let T_i denote the failure time of interest and C_i the censoring time with Z_i a vector of associated covariates, $i = 1, ..., n$. Here we assume that the Z_i's are time-independent for simplicity and the method given below can be easily generalized to situations with time-dependent covariates. Suppose that T_i is independent of C_i given Z_i and the censoring indicator $\delta_i = I(T_i \leq C_i)$ may be missing for some subjects and assume that the missingness is MAR. Define $\epsilon_i = 0$ if δ_i is missing and 1 otherwise, $i = 1, ..., n$. Then the observed data consist of $\{ Y_i = (X_i = \min(T_i, C_i), \epsilon_i, \epsilon_i \times \delta_i) ; i = 1, ..., n \}$.

To specify the covariate effects, following Goetghebeur and Ryan (1995), we will assume that both T_i and C_i follow the proportional hazards models with their marginal

hazard functions given by

$$\lambda(t|Z_i) = \lambda_0(t)\, \exp(Z_i'\beta)$$

and

$$h_c(t|Z_i) = h_0(t)\, \exp(Z_i'\gamma)\,,$$

respectively, where $\lambda_0(t)$ and $h_0(t)$ are unknown baseline hazard functions and β and γ are vectors of regression parameters. Define $\xi = (\beta, \gamma, \lambda_0, h_0)$. Then the log-likelihood function is proportional to

$$l_n(\xi) = \sum_{i=1}^{n} \log p(Y_i|\xi)$$

with

$$p(Y_i|\xi) = \Big[\big\{\lambda_0(X_i)e^{Z_i'\beta}\big\}^{\delta_i} \exp\big\{-\Lambda_0(X_i)e^{Z_i'\beta}\big\}\big\{h_0(X_i)e^{Z_i'\gamma}\big\}^{1-\delta_i} \exp\big\{-H_0(X_i)e^{Z_i'\gamma}\big\}\Big]^{\epsilon_i}$$

$$\times\Big[\big\{\lambda_0(X_i)e^{Z_i'\beta} + h_0(X_i)e^{Z_i'\gamma}\big\} \exp\big\{-\Lambda_0(X_i)e^{Z_i'\beta} - H_0(u)e^{Z_i'\gamma}\big\}\Big]^{1-\epsilon_i},$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ and $H_0(t) = \int_0^t h_0(u)du$, the cumulative baseline hazard functions of T_i and C_i, respectively.

To estimate unknown parameters, it is natural to maximize the log likelihood function $l_n(\xi)$. However, given the dimensions of λ_0 and h_0, this is usually complicated in addition to other problems. To avoid this, we will adopt the sieve approach used by Huang (1997) among others and approximate λ_0 and h_0 using piece wise constant functions. To describe this, suppose $\beta \in A_1 \subset R^p$ and $\gamma \in A_2 \subset R^p$ and for any

variable U with cumulative distribution function F, define

$$a_U = \inf\{x : F(x) > 0\} \text{ and } b_U = \sup\{x : F(x) < 1\}.$$

Suppose that τ is a constant that satisfies $\tau < b_X$ and let $0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = \tau$ be a partition of $[0, \tau]$, where $m = m(n)$ is an integer depending on the sample size n. In the following, we will assume that both λ_0 and h_0 belong to piecewise constant function space

$$\Phi_m = \left\{\phi_m = \sum_{j=1}^{m} b_j I_j(t) : m_0 \le b_j \le M_0, 1 \le j \le m\right\},$$

where $I_j(t) = I(t_{j-1} < t \le t_j)$ and m_0 and M_0 are some known constants. Define $\Theta_n = A_1 \times A_2 \times \Phi_m^2$. Then one can estimate ξ by $\hat{\xi}_n = (\hat{\beta}_n, \hat{\gamma}_n, \hat{\lambda}_n, \hat{h}_n)$ defined as the values of ξ that maximize the log likelihood $l_n(\xi)$ over Θ_n.

2.3 Inference Procedure

To maximize $l_n(\xi)$ over Θ_n, we will apply the EM algorithm. For this, we will assume that the pseudo-complete data include all δ_i's. The log likelihood function of the pseudo-complete data is then given by

$$l_n^c(\xi) = \sum_{i=1}^{n} \log\left[\left\{\lambda_0(X_i)e^{Z_i'\beta}\right\}^{\delta_i} \exp\left\{-\Lambda_0(X_i)e^{Z_i'\beta}\right\}\left\{h_0(X_i)e^{Z_i'\gamma}\right\}^{1-\delta_i}\right.$$

$$\exp\Big\{ - H_0(X_i)e^{Z_i'\gamma}\Big\}\Big] .$$

For the E-step of the EM algorithm, we need to determine the expectation of $l_n^c(\xi)$ conditional on the observed data. For this, note that

$$\tilde{\delta} = E(\delta_i|Y_i) = \delta_i\epsilon_i + (1-\epsilon_i)\frac{\lambda_0(t)e^{Z_i'\beta}}{\lambda_0(t)e^{Z_i'\beta} + h_0(t)e^{Z_i'\gamma}} . \tag{2.1}$$

It follows that

$$E\{l_n^c(\xi)|Y_i's) = \sum_{i=1}^{n}\Big[\tilde{\delta}_i \log\{\lambda_0(X_i)\} + \tilde{\delta}_i Z_i'\beta - \Lambda_0(X_i)e^{Z_i'\beta}\Big]$$

$$+\sum_{i=1}^{n}\Big[(1-\tilde{\delta}_i)\log\{h_0(X_i)\} + (1-\tilde{\delta}_i)Z_i'\gamma - H_0(X_i)e^{Z_i'\gamma}\Big] .$$

For the M-step, assume that

$$\lambda_0(t) = \sum_{j=1}^{m} b_{1j}I_j(t) \;\;,\;\; h_0(t) = \sum_{j=1}^{m} b_{2j}I_j(t) .$$

Bu plugging these into $E\{l_n^c(\xi)\}$ and taking the derivatives with respect to the b_{1j}'s and b_{2j}'s, respectively, we obtain

$$b_{1j}(\beta) = \frac{\sum_{i=1}^{n}\tilde{\delta}_i I_j(X_i)}{S_{0j}(\beta)} \;\;,\;\; b_{2j}(\gamma) = \frac{\sum_{i=1}^{n}(1-\tilde{\delta}_i)I_j(X_i)}{S_{0j}(\gamma)}$$

for given β and γ, $j = 1, ..., m$, where

$$S_{0j}(\beta) = \sum_{i=1}^{n} \exp(Z_i'\beta) \int_0^{X_i} I_j(u)du\,.$$

Define

$$\lambda_0(t,\beta) = \sum_{j=1}^{m} b_{1j}(\beta) I_j(t) \;\;,\;\; h_0(t,\gamma) = \sum_{j=1}^{m} b_{2j}(\gamma) I_j(t)\,. \tag{2.2}$$

By maximizing the profile log likelihood $l_n(\beta, \gamma, \lambda_0(\cdot,\beta), h_0(\cdot,\gamma))$, we have the estimating equations

$$\sum_{i=1}^{n} \tilde{\delta}_i\big(Z_i) - E(\beta, X_i)\big) = 0 \;\;,\;\; \sum_{i=1}^{n} (1-\tilde{\delta}_i)\big(Z_i - E(\gamma, X_i)\big) = 0 \tag{2.3}$$

for β and γ, where

$$E(\beta,t) = \sum_{j=1}^{m} \frac{S_{1j}(\beta)}{S_{0j}(\beta)} I_j(t) \;\;,\;\; S_{1j}(\beta) = \sum_{i=1}^{n} Z_i \exp(Z_i'\beta) \int_0^{X_i} I_j(u)du\,.$$

In summary, the EM algorithm can be described as follows.

Step 0. Choose initial estimates $\beta^{(0)}$, $\gamma^{(0)}$, $\lambda_0^{(0)}$ and $h_0^{(0)}$.

Step 1. At the lth iteration, for given $\beta = \beta^{(l-1)}$, $\gamma = \gamma^{(l-1)}$, $\lambda_0(t) = \lambda_0^{(l-1)}(t, \beta^{(l-1)})$ and $h_0(t) = h_0^{(l-1)}(t, \gamma^{(l-1)})$, calculate $\tilde{\delta}_i^{(l-1)}$ by equation (2.1).

Step 2. Define the updated estimates $\beta^{(l)}$ and $\gamma^{(l)}$ to be the solutions to the equations (2.3) with letting $\tilde{\delta}_i = \tilde{\delta}_i^{(l-1)}$.

Step 3. Define the updated estimates $\lambda_0^{(l)}$ and $h_0^{(l)}$ by using the equation (2.2) with

replacing β and γ by $\beta^{(l)}$ and $\gamma = \gamma^{(l)}$.

Step 4. Go back to step 1 until convergence.

Let $\hat{\beta}_n$, $\hat{\gamma}_n$, $\hat{\lambda}_0$ and $\hat{h}_0$ denote the maximum likelihood estimates derived above. We believe that one can show that they are consistent under some regularity conditions. To describe the asymptotic normality of $\hat{\beta}_n$ and $\hat{\gamma}_n$, let $\alpha_0 = (\beta_0', \gamma_0)'$ denote the true value of $\alpha = (\beta, \gamma)$ and $\hat{\alpha}_n = (\hat{\beta}_n', \hat{\gamma}_n')'$. Define $\rho(t, Z) = P(\varepsilon = 1 | X = t, Z = Z)$, the nonmissing rate of censoring indicators,

$$m(t, Z) = \frac{\lambda_0(t) e^{Z'\beta}}{\lambda_0(t) e^{Z'\beta} + h_0(t) e^{Z'\gamma}} \,,$$

$$p_{11}(t, Z) = \rho(t, Z) \lambda_0(t) e^{Z'\beta} + (1 - \rho(t, Z)) m(t, Z) \lambda_0(t) e^{Z'\beta} \,,$$

$$p_{10}(t, Z) = (1 - \rho(t, Z))(1 - m(t, Z) \lambda_0(t) e^{Z'\beta} \,,$$

$$p_{00}(t, Z) = \rho(t, Z) h_0(t) e^{Z'\gamma} + (1 - \rho(t, Z))(1 - m(t, Z)) h_0(t) e^{Z'\gamma} \,,$$

$$s_0(t) = \begin{pmatrix} E\big[R(t) p_{11}(t, Z)\big] & E\big[R(t) p_{10}(t, Z)\big] \\ E\big[R(t) p_{10}(t, Z)\big] & E\big[R(t) p_{00}(t, Z)\big] \end{pmatrix} ,$$

$$s_1(t) = \begin{pmatrix} E\big[Z(t) R(t) p_{11}(t, Z)\big] & E\big[Z(t) R(t) p_{10}(t, Z)\big] \\ E\big[Z(t) R(t) p_{10}(t, Z)\big] & E\big[Z(t) R(t) p_{00}(t, Z)\big] \end{pmatrix} ,$$

where $R(t) = I(X \geq t)$. Let

$$\big(a^*(u), b^*(u)\big)^T = s_0^{-1}(t) s_1(t) \,.$$

Then one can show that the efficient score function for α has the form $l^*_\alpha(\xi) = (l^{*\prime}_\beta(\xi), l^{*\prime}_\gamma(\xi))'$, where

$$l^*_\beta(\xi) = \int_0^\tau \left[Z dM_1(u|Z) - a^*_1(u) dM_1(u|Z) - b^*_1(u) dM_2(u|Z) \right] \tag{2.4}$$

and

$$l^*_\gamma(\xi) = \int_0^\tau \left[Z dM_2(u|Z) - a^*_2(u) dM_1(u|Z) - b^*_2(u) dM_2(u|Z) \right]. \tag{2.5}$$

Details can be found in Appendix I and II. Furthermore, we believe one can prove that

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0) = I(\alpha_0)^{-1} \sqrt{n} \mathbb{P}_n l^*_\alpha(\xi_0) + o_p(1) \to_d N(0, I(\alpha_0)^{-1}),$$

where $\mathbb{P}_n l^*_\alpha(\xi_0)$ denotes the empirical version of $l^*_\alpha(\xi)$ based on the sample Y_i's and $I(\alpha) = E\{l^*_\alpha l^{*'}_\alpha\}$, the information matrix. That is, $\hat{\alpha}_n$ has an asymptotic normal distribution with the covariance matrix reaching the information bound and thus $\hat{\alpha}_n$ is efficient. The derivation of the efficient score functions (2.4) and (2.5) along with the explicit expression of the information matrix $I(\alpha)$ is given in the appendix.

For inference about α, one needs to estimate the information matrix $I(\alpha)$. For this, one can first estimate p_{11}, p_{10} and p_{00} with their empirical versions with replacing β, γ, λ_0 and h_0 with their estimates and also do the same with $s_0(t)$ and $s_1(t)$. In addition, we need to estimate the non-missing rate $\rho(t, Z)$ for which a natural estimate is given

by the kernel estimate

$$\hat{\rho}_n(t, Z) = \frac{\sum_{i=1}^{n} \epsilon_i V_a(X_i - t) W_b(Z_i - Z)}{\sum_{i=1}^{n} V_a(X_i - t) W_b(Z_i - Z)}, \tag{2.6}$$

where $V_a(t) = a^{-1}V(t/a)$ and $W_b(t) = b^{-1}W(t/b)$ with $V(t)$ and $W(Z)$ being some kernel functions and a and b being bandwidths. Let $\hat{s}_0(t)$ and $\hat{s}_1(t)$ denote the empirical estimates of $s_0(t)$ and $s_1(t)$ and $(\hat{a}^*, \hat{b}^*)^T = \hat{s}_0^{-1}\hat{s}_1$. Then the information matrix $I(\alpha)$ can be estimated by the empirical version of the expression given in the Appendix II with replacing all unknown functions by their estimates. For the integration involved, one could use recursive adaptive Simpson quadrature or recursive adaptive Lobatto quadrature.

For estimation of λ_0 and h_0 in (2.2), one needs to choose the number of partition points m and the partition points t_j's. From the asymptotic point view, m is usually taken to be $O(n^{1/3})$ (Huang, 1997). For the selection of the partition points t_j's, a natural choice is to evenly divide $[0, \tau]$ into m different parts or to divide $[0, \tau]$ into m parts such that each interval contains roughly the same numbers of observed times X_i's.

2.4 A Simulation Study

We conducted an extensive simulation study to evaluate the finite sample performance of the inference procedure proposed in the previous sections. In the study, we con-

sidered both MCAR and MAR mechanisms and assumed that Z_i is a scalar variable. We generated the failure time of interest T_i from the exponential distribution with the hazard function $\lambda_i(t|Z_i) = e^{Z_i\beta}$ and the censoring time C_i also from the exponential distribution with the hazard function $h(t|Z_i) = e^{Z_i\gamma}$ and truncated at 1.5. That is, $\tau = 1.5$. For the missingness of the censoring indicator under MCAR, we assumed that ϵ_i followed the Bernoulli distribution with success probability p. That is, δ_i is known with probability p. For situations with MAR, the missing indicator ϵ_i was generated from the Bernoulli distribution with the success probability $\rho(X) = \{1 + \exp(X - e^{0.3})\}^{-1}$ for given the observed time X. The results given below are based on 1000 replications.

Table 2.1 and 2.2 present the results for estimation of β and γ under the MCAR mechanism with $p = 0.5$ or 0.7, the true values of (β, γ) being $(0.5, 0.5)$, $(0.5, 0)$ or $(0, 0.5)$, and $n = 200$. Here we assumed that Z_i is a binary variable taking value 0 or 1 with probability 0.5, and took $m = 20$ for λ_0 and h_0 with $[0, \tau]$ divided such that all partition intervals contain roughly the same numbers of the observed failure times. The results include the averages of the point estimates $\hat{\beta}$ and $\hat{\gamma}$ (Mean), the sample standard deviations of the point estimates (SSD), the averages of the estimated standard errors (ESE), and the 95% empirical coverage probabilities (CP). Table 2.1 is for $p = 0.5$, while in Table 2.2, $p = 0.7$. These results suggest that the inference procedure proposed in the previous sections works reasonably well, especially when the missing rate of censoring indicators is not too high.

For situations with the MAR mechanism, in addition to the binary covariate, we also considered the covariate from the uniform distribution over $(0, 1)$ and the effect

of the sample size on the performance of the proposed inference procedure. Table 2.3 and Table 2.4 give the results obtained under the set-ups similar to those in Table 2.1 except the different missing mechanism and with $n = 200$ and 400, respectively. We chose $m = 10$ for $n = 200$ and $m = 20$ for $n = 400$. Here for estimation of the missing censoring indicator rate, we took V in (2.6) to be the Gaussian kernel function with the bandwidths a to be the one chosen by leave-one-out cross validation which gives the smallest mean square error and W to be constant 1. The results for estimation of β and γ with the Z_i's from $U(0, 1)$ are displayed in Tables 2.5 and 2.6. It can be seen that these tables yield conclusions similar to those given by Table 2.1 and Table 2.2. In particular, the procedure seems to perform better when the sample size increases and the bias seems to decrease. It appears that the regression parameters can be more accurately estimated with $Z_i \sim B(0, 0.5)$ than $Z_i \sim U(0, 1)$.

2.5 An Illustrative Example

We illustrate the proposed method using data from a clinical trial in stage II breast cancer used in Goetghebeur and Ryan (1995). In the study, 169 elderly women either died from breast cancer, died from other causes or were still alive at the last observed time. Death time due to breast cancer is the interested failure time and others are censoring times. It the end, 18 women had unknown cause of death and thus had missing censoring indicators. Two covariates were suggested be important and used in the regression: $Z_1 = 1$ if there are more than 4 nodes and 0 otherwise and $Z_2 = 1$ if ER is negative and 0 otherwise. First 5 intervals containing roughly same numbers

of observed times were chosen for the estimate of baseline hazard functions. The estimate for regression coefficients are $\hat{\beta}_1 = 0.57455\,(0.2667)$and $\hat{\beta}_2 = 1.6245\,(0.4234)$ for $Z1$ and Z_2, respectively. Comparing to the results given by Goetghebeur and Ryan (1995): $\hat{\beta}_1 = 0.57\,(0.2803)$ and $\hat{\beta}_2 = 1.59\,(0.4822)$, the point estimates are close but the proposed estimates have smaller standard deviations. For the regression of the censoring time, the estimates of the regression coefficients are $\hat{\gamma}_1 = -0.2632\,(0.1929)$ and $\hat{\gamma}_2 = -0.84724\,(1.0068)$ and they are all insignificant. We also tried 10 time intervals for the sieve estimation of the baseline hazard functions. The results are close to the result for 5 time intervals: $\hat{\beta}_1 = 0.5843\,(0.2562)$, $\hat{\beta}_2 = 1.6761\,(0.3967)$ $\hat{\gamma}_1 = -0.2672\,(0.1928)$ and $\hat{\gamma}_2 = -0.8558\,(1.0096)$.

2.6 Discussion and Concluding Remarks

This chapter considered regression analysis of right censored data with censoring indicators missing. For the analysis, the proportional hazards model was employed and an estimation procedure was developed. A major advantage of the proposed approach over the existing methods is that it allows missingness of censoring indicators to depend on the observed time or covariates and it does not impose the proportionality assumption between the baseline hazard functions of the survival time of interest and the censoring time. The simulation study suggests that the approach works well in practical situations.

We assumed conditional independence between the survival time of interest and the censoring time. Further study under the situation in which the censoring process is

informative could be considered. One way is to employ a frailty model which imposes a common frailty shared by the hazard functions of the survival time of interest and the censoring time. Competing risks models with missing failure type indicators, under which different types of failure times are usually correlated, have similar data structures and could be considered the same way.

Chapter 3

Statistical Analysis of Clustered Current Status Data

3.1 Introduction

Clustered current status data arise when the failure times of interest are clustered into small groups and the observed times are either left- or right-censored. Furthermore, the failure times within a group are correlated and the group sizes may differ from one another. In other words, study subjects are sampled in clusters and observed only once and thus no exact failure times are available. Examples of clustered current status data can be found in many areas such as cross-sectional studies and tumorigenicity experiments. In the former, one may be interested in studying times to some disease for family members based on some survey data. And for the latter, the times to tumor occurrences of litter-mates may be involved. The presence or absence of the tumor is known only at the time of animal's death as is often the case.

Current status data are also often referred to as case I interval-censored data and

their analysis has been discussed by many authors when all failure times involved are from independent subjects (Keiding, 1990; Sun, 2006). Among others, Huang (1995) investigated the fitting of the proportional odds model to current status data and Huang (1996) discussed the maximum likelihood estimation for the proportional hazards model based on current status data. Ghosh (2001), Lin et al. (1998) and Martinussen and Scheike (2002) developed methods for fitting the additive hazards model to the data. More references on this can be found in Sun (2006).

A few methods have been proposed for regression analysis of clustered right-censored failure time data (Cai and Prentice, 1997; Cai et al., 2000; Hougaard, 2000; Lu and Wang, 2005; Zeng et al., 2008). For example, Cai et al. (2002) and Zeng et al. (2008) investigated the fitting of semiparametric linear transformation models to clustered right-censored data. However, it does not seem to exist any method for regression analysis of clustered current status data except some methods for regression analysis of multivariate interval-censored failure time data (Chen et al., 2007; Goggins and Finkelstein, 2000; Kim and Xue, 2002; Tong et al, 2008). Here the interval-censored data mean that the failure time of interest is observed only to belong to a window instead of being observed exactly or right-censored. For regression analysis of multivariate interval-censored data, Goggins and Finkelstein (2000) and Kim and Xue (2002) considered the use of the marginal proportional hazards model, while Chen et al. (2007) and Tong et al. (2008) investigated the use of the marginal proportional odds model and the marginal additive hazards model, respectively. It is worth noting that multivariate failure time data can be seen as special cases of clustered data with

the fixed and same cluster sizes. In general, the cluster size can differ from cluster to cluster.

For regression analysis of clustered or multivariate failure time data, two commonly used approaches are the frailty model approach (Cai et al, 2002; Clayton and Cuzick, 1985; Oakes, 1989; Zeng et al., 2008) and the marginal model approach (Chen et al., 2007; Kim and Xue, 2002). The former provides a flexible way for directly modeling the relationship between correlated failure times (Hougaard, 2000), while the latter focuses on covariate effects on individual failure times. In this chapter, we will adopt the frailty model approach for the analysis using the Cox frailty model. The model and the likelihood function will be described in Section 3.2. For parameter estimation, we will apply the maximum likelihood estimation approach and a twp-step EM algorithm is presented in Section 3.3 (Dempster et al., 1977). Section 3.4 presents some results from a simulation study evaluating the performance of the presented methodology and an illustrative example from a tumorigenicity experiment is provided in Section 3.5. Some concluding remarks are given in Section 3.6.

3.2 Model and the Likelihood Function

Consider a survival study that involves n small clusters of subjects. Let T_{ij} denote the failure time of interest from subject j in cluster i with a vector of associated covariates Z_{ij}, $j = 1, ..., n_i$, $i = 1, ..., n$. Suppose that the T_{ij}'s may be dependent within each cluster but independent between the clusters. Also suppose that each subject is observed only once at time Y_{ij} and the observed information consists of only

Y_{ij} and $\delta_{ij} = I(T_{ij} \leq Y_{ij})$. That is, we have current status data. For regression analysis, we will assume that for each cluster, there exists a latent variable b_i and given b_i, the hazard function of T_{ij} is given by

$$\lambda(t|Z_{ij}, b_i) = \lambda_0(t) \exp\{Z'_{ij}\beta + b_i\}, \tag{3.1}$$

where $\lambda_0(t)$ is an unknown baseline hazard function and β denotes the regression parameter. Note that here for simplicity, we assume that the T_{ij}'s have the same baseline hazard function and the covariate effects on them are the same. Some comments on this will be given below. Also we will assume that given the b_i's, all failure times T_{ij}'s are independent.

The model (3.1) is often referred to the proportional hazards frailty model and has been extensively used for regression analysis of clustered right-censored failure time data (Cai and Prentice, 1997; Clayton and Cuzick, 1985; Hougaard, 2000; Lee et al., 1992). It is easy to see that if the variance of the b_i's is equal to zero, model (3.1) reduces to the proportional hazards model, the most commonly used regression model for failure time data (Cox, 1972, Kalbfleisch and Prentice, 2002). In the following, we assume that the b_i's follow a parametric model with mean zero and the density function $f(b, \theta)$, where θ denotes unknown parameters. Also we assumes that T_{ij} is independent of Y_{ij} given Z_{ij}. Then the log likelihood function is proportional to

$$l(\beta, \Lambda_0(t), \theta) = \sum_{i=1}^{n} \log \left[\int \prod_{j=1}^{n_i} F^{\delta_{ij}}(Y_{ij}|Z_{ij}, b_i)\{1 - F(Y_{ij}|Z_{ij}, b_i)\}^{1-\delta_{ij}} f(b_i, \theta)\, db_i \right],$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$, the baseline cumulative hazard function of T_{ij}, and

$$F(Y_{ij}|Z_{ij}, b_i)) = 1 - \exp\{-\Lambda_0(t)e^{Z'_{ij}\beta+b_i}\},$$

the cumulative distribution function of T_{ij}.

For estimation of all parameters, it is natural to maximize the log likelihood function $l(\beta, \Lambda_0(t), \theta)$. In the next section, we will present an EM type algorithm for the maximization.

3.3 Parameter Estimation

For estimation of all unknown parameters β, $\Lambda_0(t)$ and θ, one way is to directly maximize the log likelihood function $l(\beta, \Lambda_0(t), \theta)$. However, it is easy to see that this will be quite complicated (Huang, 1996). An alternative is to apply the sieve approach or to approximate $\Lambda_0(t)$ using a piecewise constant function, which makes the problem more tractable and has been used by Rossini et al. (1996) among others. In the following, we will adopt this approach and suppose that there exist J different predetermined time points $0 = t_0 < t_1 < ... < t_J = \tau$ such that the cumulative baseline hazard function $\Lambda_0(t)$ can be approximated by

$$\Lambda_0(t) = \sum_{j=1}^{J} I_j(t) \sum_{k=1}^{j} e^{\gamma_k}, \tag{3.2}$$

where τ denotes the largest follow-up time, $I_j(t) = I(t_{j-1} < t \leq t_j)$, and $\gamma = (\gamma_1, ..., \gamma_J)'$ are unknown parameters. Then the log likelihood function l becomes the function of β, γ and θ.

To maximize l with respect to β, γ and θ, we present a two-step algorithm below that iterates between the estimation of β and γ and the estimation of θ while fixing others. For the estimation of β and γ, we will employ the EM algorithm, while for the estimation of θ, we will directly maximize the log likelihood function l.

For the E-step in estimation of β and γ, we will assume that the b_i's are observed, which gives the pseudo-complete data $\{Y_{ij}, \delta_{ij}, Z_{ij}, b_i\}$ and the pseudo-complete data log likelihood function $l^c(\beta, \gamma, \theta) = \sum_{i=1}^n l_i^c(\beta, \gamma, \theta)$, where

$$l_i^c(\beta, \gamma, \theta) = \log f(b_i, \theta) + \sum_{j=1}^{n_i} \log \left\{ F^{\delta_{ij}}(Y_{ij}|Z_{ij}, b_i)(1 - F(Y_{ij}|Z_{ij}, b_i))^{1-\delta_{ij}} \right\} .$$

It follows that the conditional expectation of the pseudo-complete data log likelihood function has the form

$$E\{l^c(\beta, \gamma, \theta)\} = \sum_{i=1}^n E\{l_i^c(\beta, \gamma, \theta)\} = \sum_{i=1}^n \int l_i^c(\beta, \gamma, \theta) f(b_i|O_i, \beta^{(m)}, \gamma^{(m)}, \theta^{(m)}) db_i$$

given the observed data and the current estimates of the parameters, where $O_i = \{Y_{ij}, \delta_{ij}, Z_{ij} \, ; \, j = 1, ..., n_i\}$ and

$$f(b_i|O_i, \beta, \gamma, \theta) = \frac{\prod_{j=1}^{n_i} F^{\delta_{ij}}(Y_{ij}|Z_{ij}, b_i)\{1 - F(Y_{ij}|Z_{ij}, b_i)\}^{1-\delta_{ij}} f(b_i, \theta)}{\int \prod_{j=1}^{n_i} F^{\delta_{ij}}(Y_{ij}|Z_{ij}, b_i)\{1 - F(Y_{ij}|Z_{ij}, b_i)\}^{1-\delta_{ij}} f(b_i, \theta) \, db_i} ,$$

the conditional density function of b_i given the observed data O_i.

It is apparent that the computation of the expectation above has no closed form. Therefore, we need some numerical approximation. In general, we need to evaluate the integral of the following form

$$E(h(b_i)|O_i, \beta^{(m)}, \gamma^{(m)}, \theta^{(m)}) = \int h(b_i) f(b_i|O_i, \beta^{(m)}, \gamma^{(m)}, \theta^{(m)}) db_i$$

for any function $h(b_i)$ of b_i. The integral could be evaluated through various methods and below we will use the recursive Lobatto quadrature provided by *Quadl* function in Matlab.

For the M-Step in estimation of β and γ, one can easily derive the score functions with respect to β and γ as

$$U_\beta(\beta, \gamma, \theta) = \frac{\partial E\{l^c(\beta, \gamma, \theta)\}}{\partial \beta} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} E\left[\left\{ \frac{\delta_{ij}}{1 - F(Y_{ij}|Z_{ij}, b_i)} - 1 \right\} e^{Z'_{ij}\beta + b_i} Z_{ij} \right],$$

$$U_{\gamma_k}(\beta, \gamma, \theta) = \frac{\partial E\{l^c(\beta, \gamma, \theta)\}}{\partial \gamma_k} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} I(Y_{ij} \geq t_{k-1}) e^{\gamma_k} E\left[\left\{ \frac{\delta_{ij}}{1 - F(Y_{ij}|Z_{ij}, b_i)} - 1 \right\} e^{Z'_{ij}\beta + b_i} \right],$$

$k = 1, ..., J$. Again E is the conditional expectation.

In summary, the two-step algorithm can be summarized as follows.

Step 0. Choose the initial estimates of β, γ and θ.

Step 1. Let $\beta^{(m)}$, $\gamma^{(m)}$ and $\theta^{(m)}$ denote the estimates obtained after the mth iteration. At the $(m+1)$th iteration, define the updated estimate $\beta^{(m+1)}$ as the solution to the equation $U_\beta(\beta, \gamma^{(m)}, \theta^{(m)}) = 0$.

Step 2. Define the updated estimate $\gamma^{(m+1)}$ as the solution to the equations

$$U_{\gamma_k}(\beta^{(m+1)}, \gamma, \theta^{(m)}) = 0\,,\ k = 1, ..., J\,.$$

Step 3. Define the updated estimate $\theta^{(m+1)}$ as the solution to the equation

$$U_\theta(\beta^{(m+1)}, \gamma^{(m+1)}, \theta) = \frac{\partial l(\beta, \gamma, \theta)}{\partial \theta}\,\big|_{\beta=\beta^{(m+1)}, \gamma=\gamma^{(m+1)}} = 0\,.$$

Step 4. Go back to Step 1 until the convergence.

In the above algorithm, all equations can be solved by, for example, the Newton-Rapson algorithm. In the numerical evaluation and the example below, we used the Matlab function *fminunc*. Also we assumed that the b_i's follow the normal distribution with mean zero and the standard deviation σ and took $\theta = -2\log(\sigma)$ to get rid of the positive constraint of σ. Then we have

$$U_\theta(\beta, \gamma, \theta) = \sum_{i=1}^{n} \frac{\int \{\prod_{j=1}^{n_i} F^{\delta_{ij}}(Y_{ij}|Z_{ij}, b_i)(1 - F(Y_{ij}|Z_{ij}, b_i))^{1-\delta_{ij}}\} f'_\theta(b_i, \theta) db_i}{\int \{\prod_{j=1}^{n_i} F^{\delta_{ij}}(Y_{ij}|Z_{ij}, b_i)(1 - F(Y_{ij}|Z_{ij}, b_i))^{1-\delta_{ij}}\} f(b_i, \theta) db_i}\,,$$

where

$$f'_\theta(b, \theta) = \frac{1}{2\sqrt{2\pi}} e^{\frac{1}{2}(\theta - b^2 e^\theta)}(1 - b^2 e^\theta)\,.$$

Note that as an alternative to the above two-step algorithm, one can directly develop

an EM algorithm for estimation of β, γ and θ together. However, we found that the direct EM is much slower than the two two-step EM algorithm given above.

Let $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\theta}$ denote the estimates of β, γ and θ obtained above. Then one can approximate their distributions by the normal distributions. For their variance estimates, we suggest to apply the profile likelihood method (Murphy et al., 1997). Specifically, assume that $b_i \sim N(0, \sigma^2)$ and one is only interested in parameters β and θ. Define the profile likelihood functions for β and θ as

$$pl(\beta) = \frac{1}{n} \sup_{\gamma,\theta} l(\beta, \gamma, \theta) \ , \ pl(\theta) = \frac{1}{n} \sup_{\beta,\gamma} l(\beta, \gamma, \theta) \, ,$$

which give the information matrices for β and θ as

$$I_\beta = -E \left\{ \frac{\partial^2 pl(\beta)}{\partial \beta^2} \right\} \ , \ I_\theta = -E \left\{ \frac{\partial^2 pl(\theta)}{\partial \theta^2} \right\} \, ,$$

respectively. In practice, one can estimate these information matrices by

$$\hat{I}_\beta = h_n^{-2} \{ 2pl(\hat{\beta}) - pl(\hat{\beta} - h_n) - pl(\hat{\beta} + h_n) \}$$

and

$$\hat{I}_\theta = h_n^{-2} \{ 2pl(\hat{\theta}) - pl(\hat{\theta} - h_n) - pl(\hat{\theta} + h_n) \}$$

and thus the variances of $\hat{\beta}$ and $\hat{\theta}$ by the inverse of $\hat{I}_\beta$ and $\hat{I}_\theta$, respectively. In the estimates above, h_n is a constant and a common choice for it is $n^{-1/2}$.

3.4 A Simulation Study

A simulation study was conducted to evaluate the finite sample properties of the method proposed in the previous sections with the focus on estimation of β. In the study, the failure time of interest was generated from the proportional hazards model

$$\lambda_{ij}(t) = \lambda_0(t) \exp(Z'_{ij}\beta + b_i)\,.$$

The latent variables b_i's were assumed to follow the normal distribution with mean zero and the standard deviation σ, and the cluster size n_i was generated from the uniform distribution $U\{2, 3, 4\}$. Furthermore, the observation time Y_{ij} was assumed to follow the uniform distribution over $(0, 1.5)$. The results given below are based on $n = 100$ with 100 replications.

Table 3.1 presents the results obtained for estimation of β with Z_{ij} being a binary variable from $B(1, 0.5)$ and $\lambda_0(t) = 1$. Here we considered the true values of β being -0.5, 0 or 0.5 and the true values of σ being 1 or 0.2. The results include the averages of the point estimates $\hat{\beta}$ (Mean), the sample standard deviations of the point estimate (SSD), the averages of the estimated standard errors (ESE), and the 95% empirical coverage probabilities (CP). For the variance estimation, we took $h_n = n^{-1/2}$ and the cumulative baseline function was approximated by a five steps piecewise constant function (3.2) (J=5) with the intervals divided to have roughly same numbers of subjects in each. The results suggest that the proposed estimate of regression parameter β seems to be unbiased and the suggested variance estimate also seems reasonable.

The case considered above corresponds to the two-sample situation with the constant hazard function. We also studied other situations where Z_{ij} is continuous and/or the hazard function is nonconstant. For example, Table 3.2 gives the simulation results with Z_{ij} generated from the standard normal distribution with all other set-up being the same as in Table 3.1. They gave similar conclusions to those in Table 3.1. Table 3.3 displays the results obtained with Z_{ij} arising from the uniform distribution over $(0, 1)$ and $\lambda_0(t) = kt$, where $k = 3, 3.5, 4$ for $\beta = 0.5, 0, -0.5$, respectively. The true value of σ is 0.2. It seems that biases still exist for $n = 100$. Bigger sample size may be needed to produce more accurate estimates.

Limited results for the estimation of θ are also presented in Table 3.4. The estimates were obtained from the same simulated date sets as ones from Table 3.1 and Table 3.2 when $\sigma = 1$ or equivalently, $\theta = 0$. Again it seems that biases still exist for $n = 100$. Bigger sample size may be needed to produce more accurate estimates.

3.5 An Illustrative Example

In this section, we apply the proposed method to an animal tumorigenicity experiment conducted by the National Toxicology Program(NTP). It is a 2 year rodent carcinogenicity study of chloroprene consisting of 50 rats with both sexes in each of the three dose groups and one control group. The animal either died during the study or was sacrificed at the end of the study. At the death or sacrifice, the presence of tumors was determined through a pathologic examination. Thus the tumors occurrence time were not directly observed and we only have current status data about these tumor

times. In the analysis below, we will focus on two types of tumors, adrenal and lung tumors, from the male rats in the control and 8 ppm dose groups. The goal is to compare the tumor growth rates between the control and dose groups.

For the analysis, let T_{i1} and T_{i2} denote the occurrence times of adrenal and lung tumors, respectively, for the ith animal and define $Z_{ij} = 1$ if the ith animal was in the dose group and 0 otherwise. Assume that the tumor occurrence times can be described by model (3.1). The application of the method proposed in the previous sections gave $\hat{\beta} = 0.5424$ with the estimated standard error of 0.3174. This corresponds to the p-value of 0.0875 for testing no does effect on the tumor growth. Here for the result, we divided the positive half line into 4 time intervals with roughly equal numbers of observations within each interval for the sieve estimation of the cumulative baseline hazard function in (3.2) ($J = 4$). By using 9 time intervals with also roughly equal numbers of observations, we obtained $\hat{\beta} = 0.5213$ with the estimated standard error of 0.3302, giving the p-value of 0.1144 for testing the dose effect.

Note that model (3.1) assumes that the two different types of tumors have the identical baseline hazard functions, which may not be realistic. In other words, it may be more reasonable to assume different baseline hazards for the two different types of tumors and it is actually straightforward to generalize the method given above to this more general situation. By assuming different baseline hazard functions for the two tumors, we got $\hat{\beta} = 0.9772$ with the estimated standard error being 0.5394 and the p-value being 0.07 for testing no dose effect by using the same 4 time intervals as above. In this case, we also obtained $\hat{\sigma} = 1.2453$ with the estimated standard error

of 0.6028. By using the same 9 time intervals used above, the estimated dose effect is $\hat{\beta} = 0.9336$ with the estimated standard error being 0.5164, which again yielded the p-value of 0.07. For the latent variable, we got $\hat{\sigma} = 1.1297$ with the estimated standard error of 0.6137. These results indicate that there was some mild dose effect and the animals in the dose group seem to have higher tumor occurrence rates than these in the control group. Also it seems that the occurrence rates between the two types of tumors were significantly correlated. Similar results were obtained for other partitions with respect to the estimation of the cumulative hazard function.

3.6 Discussion and Concluding Remarks

This chapter considered regression analysis of clustered current status data which often arise in, for example, cross-sectional and animal studies. For the analysis, the proportional hazards frailty model was employed and an estimation procedure was developed. A major advantage of the proposed approach over the existing methods for multivariate current status data is that it allows the group sizes to be different. The simulation study suggests that the approach works well in practical situations. As mentioned before, the approach can be easily generalized to situations where the baseline hazard function in model (3.1) is different for subjects in different groups. The same is true if different covariate effects are assumed in model (3.1).

There exist several directions for future research. One is that one may be interested

in generalizing model (3.1) to

$$\lambda(t|Z_{ij}, X_{ij}, b_i) = \lambda_0(t) \exp\{Z'_{ij}\beta + X'_{ij}b_i\}$$

assuming that there exists another vector of covariates X_{ij}. Here X_{ij} may be completely different from, overlap with, or the same as Z_{ij} and represent covariates that may affect the correlation among different types of events. Of course, one can replace $\lambda_0(t)$ and β by different baseline hazard functions and different β for subjects in different groups as discussed above. Another direction is to generalize the proposed method to clustered case II or general interval censored data (Sun, 2006). For this, the idea described above should directly apply. However, estimation or the algorithm needed for the determination of the estimates of unknown parameters would be much more difficult. Of course, it would also be useful to develop some procedures for checking model (3.1).

Chapter 4

Statistical Analysis of Clustered Interval–Censored Data

4.1 Introduction

Clustered interval-censored data occur when failure times of interest are clustered into small groups and the observed times are known to lie in certain intervals. Furthermore, the failure times within a group are correlated and the group sizes may differ from one another. For example, the response times arise from a clinical trial or a longitudinal study in which there is a periodic follow up. An individual who is monitored weekly for a response may miss visits for a few weeks, and return in a changed response state, thus contributing an interval-censored observation.

As discussed before, the analysis of interval-censored data has been discussed by many authors when all failure times involved are from independent subjects. Among others, Huang (1997) investigated the fitting of the proportional odds model to interval-censored data. More references on this can be found in Sun (2006).

A few methods have been proposed for regression analysis of clustered right-censored failure time data (Cai and Prentice, 1997; Cai et al., 2000; Hougaard, 2000; Lu and Wang, 2005; Zeng et al., 2008). For example, Cai et al. (2002) and Zeng et al. (2008) investigated the fitting of semi-parametric linear transformation models to clustered right-censored data. However, there does not seem to exist any method for regression analysis of clustered interval-censored data except some methods for regression analysis of multivariate interval-censored failure time data (Chen et al., 2007; Goggins and Finkelstein, 2000; Kim and Xue, 2002; Tong et al, 2008).

As noted before, for regression analysis of clustered or multivariate failure time data, two commonly used approaches are the frailty model approach (Cai et al, 2002; Clayton and Cuzick, 1985; Oakes, 1989; Zeng et al., 2008) and the marginal model approach (Chen et al., 2007; Kim and Xue, 2002). The former provides a flexible way for directly modeling the relationship between correlated failure times (Hougaard, 2000), while the latter focuses on covariate effects on individual failure times. In this chapter, we will adopt the frailty model approach for the analysis using the Cox frailty model. The model and the likelihood function will be described in Section 4.2. For parameter estimation, we will apply the maximum likelihood estimation approach and a two-step EM algorithm is presented in Section 4.3 (Dempster et al., 1977). Section 4.4 presents some results from a simulation study evaluating the performance of this methodology, which is applied to the lymphatic filariasis study discussed before in Section 4.5. Some concluding remarks are given in Section 4.6.

4.2 Model and the Likelihood Function

Consider a survival study that involves n small clusters of subjects. Let T_{ij} denote the failure time of interest from subject j in cluster i with a vector of associated covariates Z_{ij}, $j = 1, ..., n_i$, $i = 1, ..., n$. Suppose that the T_{ij}'s may be dependent within each cluster but independent between the clusters. Also suppose that each subject is observed to lie between inspecting time L_{ij} and R_{ij}, and the observed information consists of only L_{ij} and R_{ij}. That is, we have interval-censored data. For regression analysis, we will assume that for each cluster, there exists a latent variable b_i and given b_i, the hazard function of T_{ij} is given by

$$\lambda(t|Z_{ij}, b_i) = \lambda_0(t) \exp\{\beta' Z_{ij} + b_i\}, \tag{4.1}$$

where $\lambda_0(t)$ is an unknown baseline hazard function and β denotes the regression parameter. Note that here for simplicity, we assume that the T_{ij}'s have the same baseline hazard function and the covariate effects are the same. Some comments on this will be given below. Also we will assume that given the b_i's, all failure times T_{ij}'s are independent.

The model (4.1) is often referred to the proportional hazards frailty model and has been extensively used for regression analysis of clustered right-censored failure time data (Cai and Prentice, 1997; Clayton and Cuzick, 1985; Hougaard, 2000; Lee et al., 1992). It is easy to see that if the variance of the b_i's is equal to zero, model (4.1) reduces to the proportional hazards model, the most commonly used regression model

for failure time data (Cox, 1972, Kalbfleisch and Prentice, 2002). In the following, we assume that the b_i's follow a parametric model with mean zero and the density function $f(b, \theta)$, where θ denotes unknown parameters. Also we assume that L_{ij} and R_{ij} contain no more information about T_{ij} except $L_{ij} \leq T_{ij} \leq R_{ij}$. Then the log likelihood function is proportional to:

$$l(\beta, \Lambda_0(t), \theta) = \sum_{i=1}^{n} \log \int \{\prod_{j=1}^{n_i} (S(L_{ij}|Z_{ij}, b_i) - S(R_{ij}|Z_{ij}, b_{ij})\} f(b_i, \theta) db_i.$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$, the baseline cumulative hazard function of T_{ij}, and

$$S(t|Z_{ij}, b_i) = \exp\{-\Lambda_0(t) e^{\beta' Z_{ij} + b_i}\},$$

the survival function of T_{ij}. For estimation of all parameters, it is natural to maximize the log likelihood function $l(\beta, \Lambda_0(t), \theta)$. In the next section, we will present an EM type algorithm for the maximization.

4.3 Parameter Estimation

For estimation of all unknown parameters β, $\Lambda_0(t)$ and θ, one way is to directly maximize the log likelihood function $l(\beta, \Lambda_0(t), \theta)$. However, it is easy to see that this will be quite complicated (Huang, 1996). An alternative is to apply the sieve approach or to approximate $\Lambda_0(t)$ using a piecewise linear function, which has been used by

Huang (1997) among others. In the following, we will adopt this approach and suppose that there exist J different predetermined time points $0 = t_0 < t_1 < ... < t_J = \tau$ such that the cumulative baseline hazard function $\Lambda_0(t)$ can be approximated by

$$\Lambda_0(t) = \sum_{j=1}^{J} I_j(t)(\phi_{j-1} + \frac{\phi_j - \phi_{j-1}}{t_j - t_{j-1}}(t - t_{j-1})) = \sum_{j=1}^{J} I_j(t)(\sum_{k=0}^{j-1} e^{r_k} + e^{r_j}\frac{t - t_{j-1}}{t_j - t_{j-1}}), \quad (4.2)$$

where τ denotes the largest follow-up time, $I_j(t) = I(t_{j-1} < t \leq t_j)$, $\phi = (\phi_0, ..., \phi_j)$ are the values of $\Lambda_0(t)$ at time points $(t_0, t_1, ..., t_j)$ and $\gamma = (\gamma_0, \gamma_1, ..., \gamma_J)$ are some unknown parameters.

To maximize l with respect to β, γ and θ, we present a two-step algorithm below that iterates between the estimation of β and γ and the estimation of θ while fixing others. For the estimation of β and γ, we will employ the EM algorithm, while for the estimation of θ, we will directly maximize the log likelihood function l.

For the E-step in estimation of β and γ, we will assume that the b_i's are observed, which gives the log likelihood of the pseudo-complete data $\{L_{ij}, R_{ij}, Z_{ij}, b_i\}$ and its log likelihood $l^c(\beta, \gamma, \theta) = \sum_{i=1}^{n} l_i^c(\beta, \gamma, \theta)$, where

$$l_i^c(\beta, \gamma, \theta) = \log f(b_i, \theta) + \sum_{j=1}^{ni} \log(S(L_{ij}|Z_{ij}, b_i) - S(R_{ij}|Z_{ij}, b_i)).$$

It follows that the expectation of the above log-likelihood can be written as

$$El^c(\beta, \gamma, \theta) = \sum_{i=1}^{n} El_i^c(\beta, \gamma, \theta) = \sum_{i=1}^{n} \int l_i^c(\theta, O_i, b_i) f(b_i|O_i, \beta^{(m)}, \gamma^{(m)}, \theta^{(m)}) db_i \quad (4.1)$$

given the observed data and the current estimates of the parameters, where $O_i = \{L_{ij}, R_{ij}, Z_{ij}\,;\, j = 1, ..., n_i\}$ and

$$f(b_i|O_i, \beta, \gamma, \theta) = \frac{\prod_{j=1}^{n_i}[S(L_{ij}|Z_{ij}, b_i) - S(R_{ij}|Z_{ij}, b_i)]f(b_i, \theta)}{\int \prod_{j=1}^{n_i}[S(L_{ij}|Z_{ij}, b_i) - S(R_{ij}|Z_{ij}, b_i)]f(b_i, \theta)db_i}$$

the conditional density function of b_i given the observed data O_i.

It is apparent that the computation of the expectation has no closed form. Therefore, we need numerical computation. Generally, we need to evaluate integrals of the following forms: for any function $h(b_i)$ of b_i,

$$E(h(b_i)|O_i, \beta^{(m)}, \gamma^{(m)}, \theta^{(m)}) = \int h(b_i) f(b_i|O_i, \beta^{(m)}, \gamma^{(m)}, \theta^{(m)})db_i,$$

For the M-Step, one can easily derive the score functions with respect to β and γ as

$$U_\beta(\beta, \gamma, \theta) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} E\frac{e^{\beta' Z_{ij}+b_i} Z_{ij}[S(L_{ij})\Lambda_0(L_{ij}) - S(R_{ij})\Lambda_0(R_{ij})]}{S(L_{ij}) - S(R_{ij})};$$
$$U_{\gamma_k}(\beta, \gamma, \theta) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} E\frac{e^{\beta' Z_{ij}+b_i}[S(L_{ij})\Lambda'_{\gamma_k}(L_{ij}) - S(R_{ij})\Lambda'_{\gamma_k}(R_{ij})]}{S(L_{ij}) - S(R_{ij})};$$
$$\Lambda'_{\gamma_k}(t) = \frac{\partial\Lambda_0(t)}{\partial\gamma_k} = [I(t > t_k) + \frac{t - t_{j-1}}{t_j - t_{j-1}} I_k(t)]e^{r_k}$$

where $k = 1, ..., J$. Again E is the conditional expectation.

In summary, the two-step algorithm can be summarized as follows.

Step 0. Choose the initial estimates of β, γ and θ.

Step 1. Let $\beta^{(m)}$, $\gamma^{(m)}$ and $\theta^{(m)}$ denote the estimates obtained after the mth iteration. At the $(m+1)$th iteration, define the updated estimate $\beta^{(m+1)}$ as the solution to the equation $U_\beta(\beta, \gamma^{(m)}, \theta^{(m)}) = 0$.

Step 2. Define the updated estimate $\gamma^{(m+1)}$ as the solution to the equations

$$U_{\gamma_k}(\beta^{(m+1)}, \gamma, \theta^{(m)}) = 0\,,\ k = 1, ..., J\,.$$

Step 3. Define the updated estimate $\theta^{(m+1)}$ as the solution to the equation

$$U_\theta(\beta^{(m+1)}, \gamma^{(m+1)}, \theta) = \frac{\partial l(\beta, \gamma, \theta)}{\partial \theta}\,\Big|_{\beta=\beta^{(m+1)}, \gamma=\gamma^{(m+1)}} = 0\,.$$

Step 4. Go back to Step 1 until the convergence $\hat{U}_\theta(\beta^{(m+1)}, \gamma^{(m+1)}) = 0$;

In the above algorithm, all equations can be solved by, for example, the Newton-Rapson algorithm. In the numerical evaluation and the example below, we used the Matlab function *fminunc*. Also we assumed that the b_i's follow the normal distribution with mean zero and the standard deviation σ and took $\theta = -2\log(\sigma)$ to get rid of the positive constraint of σ. Then we have

$$U_\theta = \sum_{i=1}^{n} \frac{\int \{\prod_{j=1}^{n_i} (S(L_{ij}|Z_{ij}, b_i) - S(R_{ij}|Z_{ij}, b_i))\} f'_\theta(b_i, \theta) db_i}{\int \{\prod_{j=1}^{n_i} (S(L_{ij}|Z_{ij}, b_i) - S(R_{ij}|Z_{ij}, b_i))\} f(b_i, \theta) db_i}$$

$$f'_\theta(b, \theta) = \frac{1}{2\sqrt{2\pi}} e^{\frac{1}{2}(\theta - b^2 e^\theta)}(1 - b^2 e^\theta);$$

Note that as an alternative to the above two-step algorithm, one can directly develop an EM algorithm for estimation of β, γ and θ together. However, we found that the direct EM is much slower than the two two-step EM algorithm given above.

Let $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\theta}$ denote the estimates of β, γ and θ obtained above. Then one can approximate their distributions by the normal distributions. For their variance estimates, we suggest to apply the profile likelihood method (Murphy et al., 1997). Specifically, assume that $b_i \sim N(0, \sigma^2)$ and one is only interested in parameters β and θ. Define the profile likelihood functions for β and θ as

$$pl(\beta) = \frac{1}{n} \sup_{\gamma,\theta} l(\beta, \gamma, \theta) \ , \ pl(\theta) = \frac{1}{n} \sup_{\beta,\gamma} l(\beta, \gamma, \theta) \, ,$$

which give the information matrices for β and θ as

$$I_\beta = -E \left\{ \frac{\partial^2 pl(\beta)}{\partial \beta^2} \right\} \ , \ I_\theta = -E \left\{ \frac{\partial^2 pl(\theta)}{\partial \theta^2} \right\} ,$$

respectively. In practice, one can estimate these information matrices by

$$\hat{I}_\beta = h_n^{-2} \{ 2pl(\hat{\beta}) - pl(\hat{\beta} - h_n) - pl(\hat{\beta} + h_n) \}$$

and

$$\hat{I}_\theta = h_n^{-2} \{ 2pl(\hat{\theta}) - pl(\hat{\theta} - h_n) - pl(\hat{\theta} + h_n) \}$$

and thus the variances of $\hat{\beta}$ and $\hat{\theta}$ by the inverse of $\hat{I}_\beta$ and $\hat{I}_\theta$, respectively. In the

estimates above, h_n is a constant and a common choice for it is $n^{-1/2}$.

4.4 A Simulation Study

A simulation study was conducted to evaluate the finite sample properties of the method proposed in the previous sections with the focus on estimation of β. In the study, the failure time of interest was generated from the proportional hazards model

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\beta' Z_{ij} + b_i) .$$

The latent variables b_i's were assumed to follow the normal distribution with mean zero and the standard deviation σ, and the cluster size n_i was generated from the uniform distribution $U\{2, 3, 4\}$. Furthermore, the follow up times are from 0.1 to 1.3 with a gap time 0.3. A subject is observed at each follow up time with probability 0.5. The results given below are based on $n = 100$ with 100 replications.

Table 4.1 presents the results obtained for estimation of β with Z_{ij} being a binary variable from $B(1, 0.5)$ and $\lambda_0(t) = 1$. Here we considered the true values of β being -0.5, 0 or 0.5 and the true values of σ being 1 or 0.2. The results include the averages of the point estimates $\hat{\beta}$ (Mean), the sample standard deviations of the point estimate (SSD), the averages of the estimated standard errors (ESE), and the 95% empirical coverage probabilities (CP). For the variance estimation, we took $h_n = n^{-1/2}$ and the cumulative baseline function was approximated by a five steps piecewise linear function (4.2) ($J = 5$) with the partition time points to be 0.1 to 1.3 with a gap time

0.3. The results suggest that the proposed estimate of regression parameter β seems to be unbiased and the suggested variance estimate also seems reasonable.

The situations considered above correspond to the two-sample problem with the constant hazard function. We also studied other set-ups such that Z_{ij} is continuous. For example, Table 4.2 gives the simulation results with Z_{ij} generated from the standard normal distribution with all other set-up being the same as in Table 4.1. They gave similar conclusions to those in Table 4.1.

4.5 An Illustrative Example

In this section, we apply the proposed method to the study of lymphatic filariasis described in Chapter 1. Lymphatic filariasis is a debilitating parasitic disease and several worms live together in several nests. A randomized trial was conducted in Recife, Brazil to compare the effectiveness of co-administration of DEC/ALB versus DEC alone in killing the adult worms. A total of 47 men participated in the study, 25 in the DEC group and 22 in the DEC/ALB group. The patients were periodically checked up by ultrasound for clearance of worms and thus the data are subject to interval-censoring. Here the cluster is the subject and a subunit within the cluster is a nest of adult worms.

For the analysis, let T_{ij} denote the clearance times of the jth nest of worm of the ith subject and define $Z_{ij} = 1$ if the ith subject was in the DEC/ALB group and 0 otherwise. Assume that the clearance times can be described by model (4.1). The application of the method proposed in the previous sections gave $\hat{\beta} = 1.4049$ with the

estimated standard error of 0.7495. This corresponds to the p-value of 0.08 for testing no co-administration effect. Here for the result, we used the ultrasound check up time as the partition of the time intervals. For the latent variable, we got $\hat{\sigma} = 2.8157$ with the estimated standard error of 0.5368. By using another 4 intervals partition $\hat{\beta} = 1.3205$ with the estimated standard error of 0.7957, giving the p-value of 0.0970 for testing the effect. For the latent variable, we got $\hat{\sigma} = 2.4587$ with the estimated standard error of 0.5196. These results indicate that there were no strong difference between the two treatments. Also the occurrence rates between nests of worms within same subject seem to be significantly correlated.

4.6 Discussion and Concluding Remarks

This chapter considered regression analysis of clustered interval-censored data. For the analysis, the proportional hazards frailty model was employed and an estimation procedure was developed. A major advantage of the proposed approach over the existing methods for multivariate interval-censored data is that it allows the group sizes to be different. The simulation study suggests that the approach works well in practical situations. As mentioned before, the approach can be easily generalized to situations where the baseline hazard function in model (4.1) is different for subjects in different groups. The same is true if different covariate effects are assumed in model (4.1).

Chapter 5

Future Research

In this chapter, we briefly discuss two future research directions related to the research presented in Chapters 2 and 3.

Through the whole dissertation, we assumed the conditional independence between the survival time of interest and the censoring time given covariates. That is, the censoring is noninformative. Sometimes there may exist situations when the censoring process is informative. For this, one could use a frailty model which imposes a common frailty shared by the hazard functions of the survival time of interest and the censoring time. That is, we can assume the hazard functions of T_i and C_i are:

$$\lambda_i(t|Z_i, b_i) = \lambda_0(t)\exp(Z_i'\beta + b_i), \quad h_i(t|Z_i, b_i) = h_0(t)\exp(Z_i'\gamma + \alpha b_i),$$

respectively, and T_i and C_i are independent given the covariates Z_i and the frailty b_i. Different constant α reflects different dependence relationship between the two. Given covariates, T_i and C_i are positively correlated with $\alpha > 0$, negatively correlated

with $\alpha < 0$ and independent with $\alpha = 0$. Similar inference procedure based on EM algorithm could be derived.

Another direction is that one may be interested in generalizing model (4.1) to

$$\lambda(t|Z_{ij}, X_{ij}, b_i) = \lambda_0(t) \exp\{\beta' Z_{ij} + b_i' X_{ij}\}$$

assuming that there exists another vector of covariates X_{ij}. Here X_{ij} may be completely different from, overlap with, or the same as Z_{ij} and represent covariates that may affect the correlation among different types of events. It would also be useful to develop some procedures for checking model (4.1).

Appendix

Appendix I: Derivation of the Efficient Score Functions (2.4) and (2.5)

Define two processes

$$\begin{aligned} M_1(t|z) &= \int_0^t dN_{11}(u) + m(u,z)dN_{00}(u) - R(u)d\Lambda_T(u|z), \\ M_2(t|z) &= \int_0^t dN_{10}(u) + (1-m(u,z))dN_{00}(u) - R(u)d\Lambda_C(u|z), \end{aligned} \tag{5.1}$$

where

$$m(t,z) = \frac{\lambda(t|Z)}{\lambda(t|Z) + h(t|z)}.$$

Notice that

$$M_1(t|z) = M_{11}(t,z) + \int_0^t m(u,z)dM_{00}(u|z), M_2(t|z) = M_{10}(t,z) + \int_0^t (1-m(u,z))dM_{00}(u|z),$$

where

$$M_{11}(t|z) = N_{11}(t) - \int_0^t R(u)\rho(u,z)\lambda(u|z)du,$$
$$M_{10}(t|z) = N_{10}(t) - \int_0^t R(u)\rho(u,z)h(u|z)du,$$
$$M_{00}(t|z) = N_{00}(t) - \int_0^t R(u)(1-\rho(u))(\lambda(u|z)+h(u|z))du,$$

are martingales with respect to the filtration generated by

$$\big\{I(X_i \le s, \xi = j, \delta\xi = k), Z_i, i = 1, \ldots, n, (j,k) \in \{(1,1),(1,0),(0,0)\}, s \le t\big\},$$

so $M_1(t|z)$ and $M_2(t|z)$ are two local martingales.

$$\dot{l}_\beta = Z\big[\epsilon\delta + (1-\epsilon)m(t,Z) - \Lambda(t)e^{Z\beta}\big] = \int_0^\tau Z dM_1(u|Z);$$
$$\dot{l}_\gamma = Z\big[\epsilon(1-\delta) + (1-\epsilon)[1-m(t,Z)] - H(t)e^{Z\gamma}\big] = \int_0^\tau Z dM_2(u|Z);$$
$$\text{Let}\frac{\partial[\log\lambda(t)]}{\partial\eta_1} = a(t), \quad \frac{\partial[\log h(t)]}{\partial\eta_2} = b(t);$$
$$\dot{l}\eta_1 = \epsilon\delta a(t) + (1-\epsilon)m(t)a(t) - e^{Z\beta}\int_0^\tau \lambda(u)a(u)R(u)du = \int_0^\tau a(u)dM_1(u|Z);$$
$$\dot{l}\eta_2 = \epsilon(1-\delta)b(t) + (1-\epsilon)(1-m(t))b(\tau) - e^{Z\gamma}\int_0^\tau h(u)b(u)R(u)du = \int_0^\tau b(u)dM_2(u|Z)$$

The variance processes of the martingales are

$$\begin{aligned}
d\langle M_1\rangle(u|Z) &= d\langle M_{11}\rangle(u|Z) + m(u,Z)^2 d\langle M_{00}\rangle(u|Z) \\
&= R(u)\rho(u,Z)\lambda(u|Z)du + R(u)(1-\rho(u,Z))m(u,Z)\lambda(u|Z)du; \\
d\langle M_2\rangle(u|Z) &= d\langle M_{10}\rangle(u|Z) + (1-m(u,Z))^2 d\langle M_{00}\rangle(u|Z) \\
&= R(u)\rho(u,Z)h(u|Z)du + R(u)(1-\rho(u,Z))(1-m(u,Z))h(u|Z)du; \\
d\langle M_1, M_2\rangle(u|Z) &= m(u,Z)(1-m(u,Z))\langle dM_{00}\rangle(u|Z) = R(u)(1-\rho(u,Z))(1-m(u,Z))\lambda(u|Z)du
\end{aligned}$$

The equations to solve the efficient score are:

$$\begin{aligned}
E(\dot{l}_\beta - \dot{l}_{\eta^*_{1\beta}} - \dot{l}_{\eta^*_{2\beta}})\dot{l}_{\eta_1} = 0; \quad E(\dot{l}_\beta - \dot{l}_{\eta^*_{1\beta}} - \dot{l}_{\eta^*_{2\beta}})\dot{l}_{\eta_2} = 0; \\
E(\dot{l}_\gamma - \dot{l}_{\eta^*_{1\gamma}} - \dot{l}_{\eta^*_{2\gamma}})\dot{l}_{\eta_1} = 0; \quad E(\dot{l}_\gamma - \dot{l}_{\eta^*_{1\gamma}} - \dot{l}_{\eta^*_{2\gamma}})\dot{l}_{\eta_2} = 0;
\end{aligned}$$

Now we use the first two equations as an example:

$$E\{\int_0^\tau [(Z-a_1^*(u))dM_1(u|Z)-b_1^*(u)dM_2(u|Z)]\int_0^\tau a(u)dM_1(u|Z)\}=0;$$
$$E\{\int_0^\tau [(Z-a_1^*(u))dM_1(u|Z)-b_1^*(u)dM_2(u|Z)]\int_0^\tau b(u)dM_2(u|Z)\}=0;$$
$$E\{\int_0^\tau a(u)[Zd\langle M_1\rangle(u|Z)-a_1^*(u)d\langle M_1\rangle(u|Z)-b_1^*(u)d\langle M_1,M_2\rangle(u|Z)]\}=0;$$
$$E\{\int_0^\tau b(u)[Zd\langle M_1,M_2\rangle(u|Z)-a_1^*(u)d\langle M_1,M_2\rangle(u|Z)-b_1^*(u)d\langle M_2\rangle(u|Z)]\}=0;$$
$$E[Zp_{11}(u)R(u)-a_1^*(u)p_{11}(u|Z)R(u)-b_1^*(u)p_{10}(u|Z)R(u)]=0;$$
$$E[Zp_{10}(u)R(u)-a_1^*(u)p_{10}(u|Z)R(u)-b_1^*(u)p_{00}(u|Z)R(u)]=0;$$
$$(a_1^*(u),b_1^*(u))'=s_0(u)^{-1}s_1(u);$$
$$s_0(u)=\begin{pmatrix}E[R(u)p_{11}(u|Z)] & E[R(u)p_{10}(u|Z)]\\ E[R(u)p_{10}(u|Z)] & E[R(u)p_{00}(u|Z)]\end{pmatrix}$$
$$s_1(u)=\Big(E[ZR(u)p_{11}(u|Z)],E[ZR(u)p_{10}(u|Z)]\Big)'$$

With the same process, we have

$$(a_2^*(u),b_2^*(u))^u=s_0(u)^{-1}s_1^*(u);\quad s_1^*(u)=\Big(E[ZR(u)p_{10}(u|Z)],E[ZR(u)p_{00}(u|Z)]\Big)'$$

Appendix II: Explicit Expression of the Information Matrix of Chapter 2

This appendix will present the explicit expression of the information matrix $I(\alpha)$. For this, let

$$I(\alpha) = \begin{pmatrix} I_{11}(\alpha) & I_{12}(\alpha) \\ I_{12}(\alpha) & I_{22}(\alpha) \end{pmatrix},$$

where I_{11} and I_{22} have the same dimension as β and γ. Then we have

$$I_{11}(\alpha) = E\int_0^\tau \big(Z(u) - a_1^*(u))^{\otimes 2}\big) R(u) p_{11}(u, Z) du + E\int_0^\tau b_1^*(u)^{\otimes 2} R(u) p_{00}(u, Z) du$$

$$-2E\int_0^\tau \big(Z(u) - a_1^*(u)\big) b_1^{*T}(u) R(u) p_{10}(u, Z) du\,,$$

$$I_{12}(\alpha_0) = E\int_0^\tau \Big[\big(Z(u) - a_1^*(u)\big)\big(Z(u) - b_2^*(u)\big) + a_2^* b_1^*\Big] R(u) p_{10}(u, Z) du$$

$$-E\int_0^\tau \big(Z(u) - b_2^*(u)\big) b_1^{*T}(u) R(u) p_{00}(u, Z) du$$

$$-E\int_0^\tau \big(Z(u) - a_1^*(u)\big) a_2^{*T}(u) R(u) p_{11}(u, Z) du$$

and

$$I_{22}(\alpha_0) = E\int_0^\tau \big(Z(u) - b_2^*(u))^{\otimes 2}\big) R(u) p_{00}(u, Z) du + E\int_0^\tau a_2^*(u)^{\otimes 2} R(u) p_{11}(u, Z) du$$

$$-2E\int_0^\tau \big(Z(u) - b_2^*(u)\big) a_2^{*T}(u) R(u) p_{10}(u, Z) du\,.$$

BIBLIOGRAPHY

1. Cai, J. W. and Prentice, R. L. (1997). Regression estimation using multivariate failure time data and a common baseline hazard function model. *Lifetime Data Anal.* 3, 197-213

2. Cai, T., Cheng, S. C. and Wei, L. J. (2002). Semiparametric mixed-effects models for clustered failure time data. *J. Amer. Statist. Assoc.*, 97, 514-522.

3. Cai, T., Wei, L. J. and Wilcox, M. (2000). Semiparametric regression analysis for clustered failure time data. *Biometrika*, 87, 867-878.

4. Cox, D. R. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B 34, 187-220.

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

6. Clayton, D. G. and Cuzick, J. (1985). Multivariate generations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A*, 148, 82-117.

7. Cummings FJ, Gray R, Davis TE, Tormey DC, Harris JE, Falkson GG and Arseneau J. (1986). Tamoxifen versus placebo: double-blind adjuvant trial in elderly women with stage II breast cancer. *NCI Monogr.* , 1, 119-23.

8. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

9. Dinse, G. E. (1982). Nonparametric estimation for partially-complete time and type of failure data. *Biometrics*, 38, 417-431.

10. Dreyer, G., Addiss, D., Williamson, J.M. and Noroes, J. (2006). Efficacy of co-administered diethylcarbamazine and albendazole against adult Wuchereria bancrofti. *Transactions of the Royal Society for Tropical Medicine and Hygiene*, 100, 1118-1125.

11. Dunson, D. B. and Dinse, G. E. (2002). Bayesian models for multivariate current status data with informative censoring. *Biometrics*, 58, 79-88.

12. Finkelstein, DM. (1986). A proportional hazard model for interval-censored failure time data. *Biometrics*, 42, 845-854.

13. Gijbels, I., Lin, D. and Ying, Z. (1993). Non- and semi-parametric analysis of failure time data with missing failure indicators. *Tech. Report 039-93, Mathematical Sciences Research Institute, Berkeley*.

14. Goetghebeur, E. J. and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika*, 82, 821-833.

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

16. Ghosh, D. (2001). Efficiency considerations in the additive hazards model with current status data. *Statist. Neerlandica* 55, 367–376.

17. Groeneboom and Wellner (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser, Boston.

18. Hougaard, P. (2000). *Analysis of multivariate survival data.* Springer-Verlag: New York.

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

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

21. Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*, Wiley: New York.

22. Kim, M. Y. and Xue, X. (2002). The analysis of multivariate interval-censored survival data. *Statistics in Medicine*, 21, 3715-3726.

23. Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley: New York.

24. Lee, E. W., Wei, L. J. and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art*, Ed.J. P. Klein and P. K. Goel, 237–47. Dordrecht: Kluwer Academic.

25. Lin, D. Y. , Oakes, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika*, 85(2), 289-298.

26. Lo, S. H. (1991). Estimating a survival function with incomplete cause-of-death data. *J. Multivariate Anal*, 39, 217-235.

27. Lu, S., and Wang, M. (2005). Marginal analysis for clustered failure time data. *Lifetime Data Analysis*, 11, 61-79.

28. Martinussen, T. and Scheike, H. T. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika*, 89, 3, 649C658.

29. McKeague, I. W. and Subramanian, S. (1998). Product-limit estimators and Cox regression with missing censoring information. *Scand. J. Statist*, 25, 589-601.

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

31. Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*,84, 487-493.

32. Robertson, T., Wright, F.T. and Dykstra, R.L. (1998). *Order restricted statistical inference*, Wiley.

33. Subramanian, S. (2000). Efficient estimation of regression coefficients and baseline hazard under proportionality of conditional hazards. *J. Statist. Plann. Inference*, 84, 81-94.

34. Sun, J. (2006). *The Statistical Analysis of Interval-censoring Failure Time Data*, New York: Springer-Verlag.

35. Zeng, D., Lin, D. Y. and Lin, X. (2008). Semiparametric transformation models with random effects for clustered failure time data. *Statist. Sinica*, 18, 355-377.

36. Van der Laan, M. J. and McKeague, I. W. (1998). Efficient estimation from right-censored data when failure indicators are missing at random. *Ann. Statist.*, 26, 164-182.

Table 2.1: Estimation of β and γ under MCAR and $p = 0.5$

Parameter	True value	Mean	SSD	SSE	CP
β	0.5	0.5245	0.2689	0.2509	0.931
γ	0.5	0.4661	0.2489	0.2418	0.950
β	0	-0.0094	0.3066	0.2736	0.935
γ	0.5	0.4604	0.2410	0.2215	0.931
β	0.5	0.5290	0.2569	0.2315	0.925
γ	0	-0.0184	0.2728	0.2626	0.947

Table 2.2: Estimation of β and γ under MCAR and $p = 0.7$

Parameter	True value	Mean	SSD	SSE	CP
β	0.5	0.5037	0.2404	0.2273	0.942
γ	0.5	0.4826	0.2196	0.2214	0.955
β	0	-0.0015	0.2646	0.2455	0.937
γ	0.5	0.4822	0.2045	0.2032	0.949
β	0.5	0.5024	0.2092	0.2103	0.954
γ	0	-0.0011	0.2407	0.2381	0.945

Table 2.3: Estimation of β and γ under MAR, $Z \sim B(1, 0.5)$ and $n = 200$

Parameter	True value	Mean	SSD	SSE	CP
β	0.5	0.5088	0.2291	0.2284	0.946
γ	0.5	0.4899	0.2318	0.2232	0.945
β	0	-0.0013	0.2493	0.2469	0.951
γ	0.5	0.4749	0.2077	0.2054	0.949
β	0.5	0.5127	0.2281	0.2117	0.932
γ	0	-0.0111	0.2375	0.2402	0.963

Table 2.4: Estimation of β and γ under MAR, $Z \sim B(1, 0.5)$ and $n = 400$

Parameter	True value	Mean	SSD	SSE	CP
β	0.5	0.5036	0.1675	0.1598	0.941
γ	0.5	0.4930	0.1573	0.1568	0.951
β	0	0.0004	0.1737	0.1732	0.950
γ	0.5	0.4755	0.1477	0.1439	0.945
β	0.5	0.5037	0.1514	0.1488	0.953
γ	0	-0.0000	0.1684	0.1674	0.950

Table 2.5: Estimation of β and γ under MAR, $Z \sim U(0, 1)$ and $n = 200$

Parameter	Value	Mean	SSD	SSE	CP
β	0.5	0.5146	0.3893	0.3762	0.946
γ	0.5	0.4858	0.3686	0.3679	0.950
β	0	-0.0074	0.4554	0.4215	0.925
γ	0.5	0.4719	0.3516	0.3504	0.955
β	0.5	0.5162	0.3797	0.3614	0.936
γ	0	-0.0198	0.3908	0.4064	0.953

Table 2.6: Estimation of β and γ under MAR, $Z \sim U(0, 1)$ and $n = 400$

Parameter	True value	Mean	SSD	SSE	CP
β	0.5	0.5089	0.2795	0.2757	0.943
γ	0.5	0.4929	0.2692	0.2688	0.946
β	0	0.0025	0.3019	0.2955	0.944
γ	0.5	0.4779	0.2624	0.2476	0.944
β	0.5	0.5144	0.2665	0.2552	0.944
γ	0	-0.0071	0.2786	0.2846	0.947

Table 3.1: Estimation of β with binary Z_{ij} and $\lambda_0(t) = 1$

σ	β	Mean	SSD	ESE	CP
1	0.5	0.5018	0.2289	0.2144	0.94
	0	-0.0026	0.2165	0.2139	0.96
	-0.5	-0.5330	0.2215	0.2225	0.95
0.2	0.5	0.5432	0.1844	0.1788	0.96
	0	-0.0068	0.1697	0.1787	0.95
	-0.5	-0.5225	0.1994	0.1938	0.95

Table 3.2: Estimation of β with normal Z_{ij} and $\lambda_0(t) = 1$

σ	β	Mean	SSD	ESE	CP
1	0.5	0.4925	0.1368	0.1348	0.94
	0	0.0429	0.1187	0.1201	0.96
	-0.5	-0.5176	0.1268	0.1353	0.96
0.2	0.5	0.5261	0.1202	0.1095	0.94
	0	0.0094	0.0917	0.0944	0.96
	-0.5	-0.5104	0.1073	0.1081	0.96

Table 3.3: Estimates of β with uniform Z_{ij} and $\lambda_0(t) = kt$

β	Mean	SSD	ESE	CP
0.5	0.4616	0.3938	0.3781	0.95
0	0.0335	0.4275	0.3955	0.96
-0.5	0.5420	0.3900	0.3984	0.95

Table 3.4: Estimation of θ with $\theta = 0$

Z	β	Mean	SSD	ESE	CP
Bernoulli	0.5	0.1022	0.5087	0.4554	0.94
	0	-0.0028	0.4731	0.4383	0.97
	-0.5	0.0798	0.4635	0.4482	0.95
Normal	0.5	0.0585	0.5049	0.4776	0.97
	0	0.0231	0.4074	0.4515	0.98
	-0.5	0.0415	0.5422	0.4818	0.98

Table 4.1: Estimation of β with binary Z_{ij}

σ	β	Mean	SSD	ESE	CP
1	0.5	0.5109	0.0892	0.0868	0.94
	0	0.0209	0.1600	0.1552	0.94
	-0.5	-0.4895	0.1527	0.1611	0.95
0.2	0.5	0.5103	0.1346	0.1293	0.95
	-0.5	-0.5024	0.1429	0.1421	0.96
	0	-0.0165	0.1337	0.1323	0.94

Table 4.2: Estimation of β with normal Z_{ij}

σ	β	Mean	SSD	ESE	CP
1	0.5	0.5101	0.0900	0.0868	0.94
	-0.5	-0.4906	0.0866	0.0863	0.94
	0	0.0044	0.0784	0.0777	0.96
0.2	0.5	0.5013	0.0734	0.0736	0.94
	0	-0.0011	0.0685	0.0652	0.95
	-0.5	-0.5024	0.0782	0.0734	0.94

VITA

Ping Chen was born on July 1st, 1984, in Xinyang, Henan Province, People's Republic of China. After attending public schools in Xinyang, she received her B.A. in Finance from The University of Science and Technology of China in 2005. She joined the graduate program in the Department of Statistics at the University of Missouri-Columbia in August of 2005. She will receive her Ph.D. in Statistics in May of 2009. As of June 2009, she will be serving as a health care performance analyst for Barnes-Jewish Christian Health Systems in St. Louis, MO.