STATISTICAL ANALYSIS OF MULTIVARIATE INTERVAL-CENSORED FAILURE TIME DATA

A Dissertation Presented

to

the Faculty of the Graduate School

University of Missouri-Columbia

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

---

by

MAN-HUA CHEN

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

DECEMBER , 2007

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

STATISTICAL ANALYSIS OF MULTIVARIATE INTERVAL-CENSORED FAILURE TIME DATA

Presented by Man-Hua Chen

A candidate for the degree of Doctor of Philosophy

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

Dr. (Tony) Jianguo Sun _____

Dr. Nancy Flournoy _____

Dr. Lori A. Thombs _____

Dr. Min Yang _____

Dr. J. Wade Davis _____

Dedicated to my family,

Tsuo-Ta Chen, Ching Cheng Chen, (Darren) Chia-Jung Hsu, Benjamin Hsu

ACKNOWLEDGEMENTS

I am indebted to many people for helping me during my doctoral work. While it is not possible to individually thank everyone, I would like to acknowledge several contributions.

I would like to express my sincere gratitude to my advisor, Dr. (Tony) Jianguo Sun for his continual encouragement and endless patience. Without him, I can not go this far and can not make it possible. I extend my gratitude to the members of my advisory committee: Dr. Flournoy, Dr. Thombs, Dr. Yang and Dr. Davis. Special thank is due to Dr. Xingwei Tong for his academic discussion and help.

I truly appreciate all of the faculty members, staff, colleagues in the Department of Statistics, especially Tracy, Judy, Clay, Ray, Do-Hwan, Xin, Liang and Chen-I.

I am deeply grateful to my parents whom I owe everything I am today. Thanks them for endless love and support throughout my life. I would also like to thank my husband and mentor, (Darren) Chia-Jung Hsu. His listening and suggestion have given me a lot of strength in all my work.

Table of Contents

List of Tables

List of Figures

STATISTICAL ANALYSIS OF MULTIVARIATE INTERVAL-CENSORED FAILURE TIME DATA

Man-Hua Chen

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

A voluminous literature on right-censored failure time data has been developed in the past 30 years. Due to advances in biomedical research, interval censoring has become increasingly common in medical follow-up studies. In these cases, each study subject is examined or observed periodically, thus the observed failure time falls into a certain interval.

Additional problems arise in the analysis of multivariate interval-censored failure time data. These include the estimating the correlation among failure times. The first part of this dissertation considers regression analysis of multivariate interval-censored failure time data using the proportional odds model. One situation in which the proportional odds model is preferred is when the covariate effects diminish over time. In contrast, if the proportional hazards model is applied for the situation, one may have to deal with time-dependent covariates. We present an inference approach for fitting the model to multivariate interval-censored failure time data. Simulation studies are conducted and an AIDS clinical trial is analyzed by using this methodology.

The second part of this dissertation is devoted to the additive hazards model for multivariate interval-censored failure time data. In many applications, the proportional hazards model may not be appropriate and the additive hazards model provides an important and useful alternative. The presented estimates of regression parameters are consistent and asymptotically normal and a robust estimate of their covariance matrix is given that takes into account the correlation of the survival variables. Simulation studies are conducted for practical situations.

The third part of this dissertation discusses regression analysis of multivariate interval-censored failure time data using the frailty model approach. Based on the most commonly used regression model, the proportional hazards model, the frailty model approach considers the random effect directly models the correlation between multivariate failure times. For the analysis, we will focus on current status or case I interval-censored data and the maximum likelihood approach is developed for inference. The simulation studies are conducted to asses and compare the finite-sample behaviors of the estimators and we apply the proposed method to an animal tumorigenicity experiment.

Chapter 1

INTRODUCTION

This thesis discusses the analysis of the data in which the response of interest is the time to some event. The occurrences of these events are often referred to as failures. Failure time data occur in numerous fields including the performance of a certain task in a learning experiment in psychology, economic sciences, engineering, public health and epidemiology. In this thesis, we mainly focus on the data from biomedical studies.

One feature of failure time data is incompleteness. Often an observed failure time may fall into a certain range instead of being known exactly and we call this type of incompleteness as censoring. Censored data are different from missing data since the censored data provide partial information, whereas the missing data provide no information. There are usually three types of censored data, right-, left-, and interval-censored data. The right-censored or left-censored data mean that the failure times of interest are observed either exactly, or to be greater or smaller than censoring times, respectively. Kalbfleisch and Prentice (2002) collected statistical models and methods for the analysis of failure time data with right censoring.

The third type of censoring is interval-censoring (Finkelstein, 1986; Sun, 2006). The interval-censored data frequently occur in medical follow-up studies and in these cases, each study subject is examined or observed periodically. In consequence, the survival events of interest are usually observed only to occur between examination times with the exact occurrence times being unknown.

Additional problems arise in the analysis of multivariate failure times. They include estimating the correlation between failure times. A typical example of correlated or multivariate failure time data is given by a twin study and Duffy *et al.* (1990) described such a study comparing monozygotic and dizygotic twins with respect to the strength of dependency of disease risk between pair members. Prentice and Hsu (1997) and Fan *et al.* (2000) gave further details and discussion about the twin study. Among others, Lin (1994) and Wei *et al.* (1989) proposed marginal likelihood estimator for regression parameters under correlated right-censored failure time data. Hougaard (2000) provided more examples of multivariate right-censored failure time data. Goggins and Finkelstein (2000) considered a set of bivariate interval-censored data arising from an AIDS clinical trial on HIV-infected individuals. Bogaerts *et al.* (2002) studied tooth emergence data with multivariate interval-censoring.

The remaining of this Chapter is organized as follows. Section 1.1 introduces a marijuana study of high school boys giving exact, right-censored, or left-censored failure times. Section 1.2 provides four examples of interval-censoring. In Section 1.3, we review some common regression models. Section 1.4 discusses some existing approaches for regression analysis of multivariate interval-censored failure time data. Section 1.5

gives the outline of this thesis.

1.1 Failure Time Data

The failure time data primarily appear in medical and biological sciences and they also widely occur in social and economic sciences as well as in industrial life-testing experiments. In general, we may not always observe exact failure times. Sometimes, we may have smaller or greater failure times than the observation time and are referred to as left- or right-censored failure times, respectively. Censored observations may occur in a number of different areas of research. For example, in the social sciences we may study drop-out times of high school students; in the industry we may test bulb life times, etc. In each case, those students do not drop out, or bulbs still be lighted on the end of the study period and those subjects represent censored observations. In the following, we use an example from a marijuana study to illustrate these details.

1.1.1 Marijuana Study of High School Boys

The marijuana study contains 191 California high school boys. Turnbull and Weiss (1978) discussed this study on the use of marijuana by the high school students and the data are given in Table 1.1. In this study, the interested question is "when did you first use marijuana?". There are three types of answers. Some of the boys remembered the exact age for using marijuana. Some of them only remembered it happen either before or after the current age, and these types of observations are referred to as left- or right-censored observation, respectively.

In Table 1.1, there are one hundred exact failures on the second column, seventy nine left-censored failures on the third column, and twelve right-censored failures on the forth column. References that analyzed this data set include Klein and Moeschberger (2003) and Turnbull and Weiss (1978).

1.2 Failure Time Data with Interval Censoring

Interval-censored data commonly arise in clinical trials and medical studies among others. In such cases, the failure times are not observed exactly but known only within some windows.

A special case of interval-censored failure time data occurs if each study subject is observed only once and in this case the failure time of interest is known only to be smaller or greater than the observation time. In other words, the failure time of interest is either left- or right-censored and such data are often referred to as current status or case I interval-censored data. One example of such data is discussed in Section 1.2.1.

General interval-censored data that are not current status data are usually referred to as case II interval-censored data. We describe such an example in Section 1.2.2. In Section 1.2.3 and 1.2.4, we present two examples of multivariate interval-censored data.

1.2.1 Lung Tumor Data - Case I Interval Censoring

The data are showed in Table 1.2 (Hoel and Walberg, 1972). The RFM mice in a tumorigenicity experiment involved lung tumors. The table includes death times of

144 male RFM mice with lung tumors and the death times are measured in days. The indicators of two treatments, conventional environment (CE) and germ-free environment (GE), are in the first column. The tumor status, with tumor and no tumor, is given in the second column. The death time in days shows in the last columns.

In this study, a question of interest is whether the environment accelerates the time to lung tumor. Lung tumor is often treated as nonlethal, meaning that the occurrence of a tumor does not change the death rate. Then the time to tumor occurrence is only known to be smaller or greater than the observed time of death or sacrifice. That is, we only observe left- or right-censored failure times. As defined above, we call this type of data as current status data or case I interval-censored data.

1.2.2 Breast Cancer Study - Case II Interval Censoring

This set of data concerns the breast cancer patients treated at the Joint Center for Radiation Therapy in Boston between 1976 and 1980. The data were reproduced from Finkelstein and Wolfe (1985). This data contain 94 early breast cancer patients in two treatment groups, radiotherapy (RT) alone and radiation therapy plus adjuvant chemotherapy (RCT). Table 1.3 shows forty six patients in RT group and forty eight patients in RCT group.

In the study, patients were supposed to be seen at clinic visits every 4 to 6 months. Actually, the patients' visit times were varying since some of them missed their visits. For each visit, physicians evaluated the cosmetic appearance of the patient such as breast retraction, a response that had a negative impact on overall cosmetic appearance.

The goal of this study was to detect whether the two treatments are different with respect to their cosmetic effects. By treating the appearance of breast retraction as the failure of interest, we have case II interval-censored data, which are provided in Table 1.3.

1.2.3 NTP Tumor Data - Case I Bivariate Interval Censoring

This data set is a part of an animal tumorigenicity experiment conducted by the National Toxicology Program (NTP). It is a 2-year rodent carcinogenicity study of chloroprene consisting of F344/N rats and B6C3F1 rats with both sexes. The original experiment contained a control group and three dose groups with 50 male and 50 female rats in each group. Rats in the dose groups were exposed to chloroprene at the concentration of 12.8, 32, or 80 ppm, respectively, 6 hours per day, 5 days per week for up to 2 years. The animal either died during the study or was sacrificed at the end of the study. At the death or sacrifice, the presence or occurrence of tumors was determined through a pathologic examination. Thus the tumor occurrence time was not exactly observed but instead known only to be smaller or greater than the death or sacrifice time. As in Section 1.2.1, we call this type of data as current status data or case I interval-censored data.

Duson and Dinse (2002) summarized the data for male rats in the control group and 80 ppm dose group concerning adrenal and lung tumors. This set of bivariate failure times are provided in Table 1.4. In Chapter 4, we study the dose effects associated to the two types of tumors.

1.2.4 AIDS Clinical Trial - Case II Bivariate Interval Censoring

The ACTG 181 study is a part of a comparative clinical trial of three anti-pneumocystis drugs and concerns the opportunistic infection cytomegalovirus (CMV), which is commonly referred to as shedding of the virus. In this study, 204 patients provided urine and blood samples at their clinical visits every 4 weeks and every 12 weeks, respectively, for testing the presence of CMV. Of course, the real sample collection time differed from patient to patient. For instance, some patients missed several visits and returned with changed CMV shedding status, which resulted in interval-censored data. Some patients were already shedding when they began the clinical visits, which provided left-censoring; some patients had not started shedding by the end time of the study, which yielded right-censoring. Goggins and Finkelstein (2000) and Finkelstein *et al.* (2002) gave more detail about this study.

1.3 Regression Models for Failure Time Data

Regression analysis estimates the strength of a modeled relationship between one or more response variables and covariate effects. One major goal of regression analysis is to determine the values of parameters in the model for covariate effects. Covariates could be treatment, age, sex, income, and education.

Let T denote a nonnegative random variable representing the failure time of a subject. The survival function of T is defined as the probability that T exceeds a time t and given by

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

The hazard function is defined differently for continuous and discrete T. Suppose that T is an absolutely continuous failure time and its probability density function $f(t)$ exists. Then $f(t)$ has the form

$$f(t) \,=\, \frac{-\,d\,S(t)}{d\,t}\,.$$

The survival function satisfies

$$S(t) \,=\, \int_t^{\infty} f(s)\,ds\,,\; 0 < t < \infty\,.$$

The hazard function of T at time t is defined as

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

It stands for the instantaneous probability that a subject fails at time t given that the subject has not failed before time t. The relationships between the survival function and the hazard function can be written as

$$\lambda(t) \,=\, \frac{f(t)}{S(t)} \,=\, -\,\frac{d\,log\,S(t)}{d\,t},\; 0 < t < \infty\,.$$

In addition, it can be showed that

$$S(t) \;=\; e^{-\int_0^t \lambda(s)\,ds} \;=\; e^{-\Lambda(t)}$$

and

$$f(t) \;=\; \lambda(t) e^{-\Lambda(t)}\,,$$

where $\Lambda(t) = \int_0^t \lambda(s) ds$, $0 < t < \infty$, which is referred to as the cumulative hazard function of T.

Assume that T is discrete and the probability function of T is defined as

$$f(s_j) \;=\; P(\,T = s_j\,),\;\; j \;=\; 1, 2, \ldots\ldots\,,$$

where $s_1 < s_2 < \cdots$.

The survival function then can be expressed as

$$S(t) \;=\; \sum_{j:t<s_j} f(s_j)\,,\;\; 0 < t < \infty\,.$$

The hazard function at s_j is given by

$$p_j \;=\; P(\,T = s_j \,|\, T \geq s_j\,) \;=\; \frac{f(s_j)}{S(s_{j^-})}\,,$$

where $S(s_{j^-}) = P(T \geq s_j)$, $0 < t < \infty$, and this conditional probability represents the probability that the failure occurs at time s_j given that the failure has not occurred before time s_j, j = 1,2,. . . .

Based on the conditional probability p_j, the survival function can be written as

$$S(t) = \prod_{j:t \geq s_j} (1 - p_j), \ 0 < t < \infty$$

and

$$f(s_j) = p_j \prod_{l=1}^{j-1} (1 - p_l).$$

The following sections describe several continuous semiparametric regression models including the proportional hazards model, the proportional odds model, the additive hazards model and the frailty model. One discrete model, the grouped proportional hazards model, is discussed in the last section of Section 1.3.

1.3.1 The Proportional Hazards Model

In this section we introduce the proportional hazards model, also termed as the Cox model (Cox, 1972; Kalbfleisch and Prentice, 2002). In terms of the hazard function, the model specifies

$$\lambda(t; \mathbf{Z}) = \lambda_0(t) e^{\mathbf{Z}'\beta} \tag{1.1}$$

for given covariates $\mathbf{Z}$, where $\lambda_0(t)$ is an arbitrary unknown baseline hazard function and β denotes a p-dimensional vector of regression parameters.

Under the model (1.1), the conditional density and the survival function of T given $\mathbf{Z}$ have the forms

$$f(t;\mathbf{Z}) \;=\; \lambda_0(t)e^{\mathbf{Z}'\beta}e^{[-\Lambda_0(t)exp(\mathbf{Z}'\beta)]}$$

and

$$S(t;\mathbf{Z}) \;=\; e^{[-\Lambda_0(t)exp(\mathbf{Z}'\beta)]} \;=\; S_0(t)^{exp(\mathbf{Z}'\beta)}\,,\; 0<t<\infty\,,$$

where $\Lambda_0(t)=\int_0^t \lambda_0(s)ds$, the baseline cumulative hazard function and $S_0(t)=e^{-\int_0^t \lambda_0(s)ds}$, the baseline survival function.

The conditional cumulative hazard function of T given $\mathbf{Z}$ has the form

$$\Lambda(t;\mathbf{Z})=\Lambda_0(t)e^{\mathbf{Z}'\beta}\,.$$

In the case of $\mathbf{Z}=0$ or 1, the ratio of the hazard functions under (1.1) has the form

$$\frac{\lambda(t;\mathbf{Z}=1)}{\lambda(t;\mathbf{Z}=0)} \;=\; e^{\beta}\,.$$

This ratio represents that the covariates has multiplicative effects on the hazard function.

The proportional hazards model has been commonly used in survival analysis. The main reason is availability of the partial likelihood approach proposed by Cox (1972, 1975). The approach is efficient since the estimator of β is asymptotically equivalent to the estimator of β given by the full likelihood function. Many authors have studied model (1.1) for regression analysis of right-censored data (Cox and Oakes, 1984; Kalbfleisch and Prentice, 2002).

1.3.2 The Proportional Odds Model

The proportional odds model specifies

$$\frac{F(t;\mathbf{Z})}{1-F(t;\mathbf{Z})} = \frac{F_0(t;\mathbf{Z})}{1-F_0(t;\mathbf{Z})}e^{\mathbf{Z}'\beta} \tag{1.2}$$

or

$$logit\, F(t;\mathbf{Z}) = logit\, F_0(t) + \mathbf{Z}'\beta$$

for given covariates $\mathbf{Z}$, where $F_0(t)$ is an unknown baseline distribution function or the distribution function for the subjects with $\mathbf{Z} = 0$, β denotes a p-dimensional vector of regression parameters and $logit(x) = log(x/(1-x))$.

In the case of $\mathbf{Z} = 0$ or 1, the ratio of the hazard functions under (1.2) is

$$\frac{\lambda(t;\mathbf{Z}=1)}{\lambda(t;\mathbf{Z}=0)} = \frac{1}{1+(e^{-\beta}-1)(1-F_0(t))},$$

which is a monotonic function of t and converges to 1 as $t \to \infty$. In the proportional hazards model, the ratio of the hazards does not change with time t, but, the ratio of the hazards changes with time t under model (1.2).

For fitting a proportional odds regression model to the analysis of right-censored failure time data, Murphy *et al.* (1997) presented a profile likelihood approach. Yang and Prentice (1999) considered the same problem but proposed several classes of regression estimators including the pseudo-maximum likelihood estimators, martingale residual-based estimators, and minimum distance estimators. Chen (2001) used weighted full likelihood for case-cohort studies. For interval-censored failure time data, Rossini and Tsiatis (1996) approximated the full likelihood by treating the baseline log-odds function as a step function for current status data. Huang and Rossini (1997) studied the sieve maximum likelihood estimator that is asymptotically normal with $\sqrt{n}$ convergence rate. Rabinowita *et al.* (2000) provided an conditional logistic regression approximation that involvs the regression parameter only. Zhang *et al.* (2005) investigated a method for linear transform models including the proportional odds model as a special case.

1.3.3 The Additive Hazards Model

Let $\lambda(t; \mathbf{Z})$ represent the hazard function at time t with covariates $\mathbf{Z}$. The additive hazards model assumes that

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

for given covariates $\mathbf{Z}$, where $\lambda_0(t)$ is an arbitrary unspecified baseline hazard function and β denotes a p-dimensional vector of regression parameters.

For fitting an additive hazards regression model to the analysis of right-censored failure time data, Lin and Ying (1994) presented a simple technique by resembling the partial-likelihood-based methods for regression parameters. Kim and Lee (1998) considered goodness-of-fit tests for the two sample problem. Kulich and Lin (2000) studied the case with covariate measurement errors. For current status failure time data, Lin *et al.* (1998) provided estimating equations for regression parameters. Martinussen and Scheike (2002) explored efficient estimation. Zhang *et al.* (2005) investigated informative censoring.

1.3.4 The Frailty Model

Again, let $\lambda(t; \mathbf{Z})$ represent the hazard function at time t with covariates $\mathbf{Z}$. Suppose the failure time t follows a frailty model, that is

$$\lambda(t; \mathbf{Z}) = \lambda_{\mathbf{0}}(\mathbf{t})\mathbf{e}^{\mathbf{Z}'\beta+\mathbf{b}} \tag{1.4}$$

for given covariates $\mathbf{Z}$, where $\lambda_0(t)$ is an unknown baseline hazard function, β denotes a p-dimensional vector of regression parameters and b is the latent variable satisfying $b \sim N(0, \Sigma)$.

Clayton (1978) considered a bivariate family derived by integrating out a gamma-distributed frailty. Shih and Louis (1995) presented a two-stage procedure for estimation in copula models for bivariate data. Glidden (2000) discussed the two-stage

estimator for the Clayton-Oakes model. There is a substantial literature on the use of the frailty for multivariate failure time models. See Oakes (1989), Hougaard (2000), Andersen *et al.* (1993), and Klein and Moeschberger (2003) for more illustration and examples.

1.3.5 The Grouped Proportional Hazards Model

All regression models discussed in the previous sections are for continuous failure times. In this section we introduce the grouped PH model (Pierce *et al.*, 1979; Prentice and Gloeckler, 1978). Let $\mathbf{Z}$ be covariates as before and let $S_0(s_j)$ represent the survivor function for subjects with $\mathbf{Z} = 0$. The survivor function at time s_j for a subject with $\mathbf{Z}$ is

$$S(s_j; \mathbf{Z}) = S_0(s_j)^{exp(\mathbf{Z}'\beta)}$$

$$= \prod_{k=1}^{j} q_k(s_j; \mathbf{Z})^{exp(\mathbf{Z}'\beta)},$$

where $q_j(s_j; \mathbf{Z}) = P(T > s_j | T \geq s_j, \mathbf{Z}) = S_0(s_j)/S_0(s_{j-1})$.
In practice, it is convenient to eliminate the parameter range restriction by taking $\alpha_j = log(-log\, q_j)$. Using the new parameters, we have

$$S(s_j; \mathbf{Z}) = \prod_{k=1}^{j} e^{-exp(\alpha_k + \mathbf{Z}'\beta)}. \tag{1.5}$$

Prentice and Gloeckler (1978) considered the grouped proportional hazards model

and Lawless (2003) gave more discuession of the grouped proportional hazards model.

1.4 Regression Analysis of Multivariate Interval-censored Data

For regression analysis of multivariate failure time data, two approaches are commonly used. They are the marginal model approach and the random effect model approach.

1.4.1 Marginal Model Approach

The marginal model approach assumes that each of the correlated failure times follows some marginal models. It deals with the marginal likelihood function directly and ignores the dependence structure between the failure times. Under this working independence assumption, one needs to use the robust covariance estimation for regression coefficients.

There exits extensive literature about the marginal model approach. For instance, Wei *et al.*, (1989) and Guo and Lin (1994) modeled the marginal distributions using the proportional hazards model. The former considered continuous failure times, whereas the latter discussed discrete times. Cai and Prentice (1995) investigated the same problem and provided weighted partial likelihood score equations for inference about regression parameters. Goggins and Finkelstein (2000) and Kim and Xue (2002) applied the same approach to the analysis for interval-censored data under the marginal proportional hazards model.

1.4.2 Random Effect Model Approach

Random effect model approach assumes that there exists a common and an unobserved latent random variable, a positive random variable, and the correlated failure times are independent given the latent variable. The latent random variable is also called frailty, which reflects the dependence of the correlated failure times. Compared with the marginal model approach, one advantage of this frailty model approach is that it directly models the correlation of failure times.

Clayton and Cuzick (1985) extended the proportional hazards model and included a random effect representing heterogeneity of frailty to failure times. Oakes (1989) considered the frailty model for bivariate failure time data. Huang and Wolf (2002) treated censoring as informative for the frailty model. For further extended discussion and illustration for multivariate failure time data, see Hougaard (2000) and Klein and Moeschberger (2003).

1.5 Outline of the Dissertation

This dissertation contains three main parts concerning statistical analysis of multivariate interval-censored failure time data. In Chapter 2, we consider the proportional odds model for multivariate interval-censored failure time data. We begin with describing the proportional odds model and then present an inference approach for fitting the model to multivariate discrete interval-censored failure time data. Simulation studies show that the proposed method works well for practical censoring percentages. The

method is applied to a set of bivariate interval-censored data arising from ACTG 181 data described in Section 1.2.4.

Chapter 3 discusses the multivariate interval-censored data analysis using the additive hazards model. For the analysis, we develop a marginal model approach. The resulting estimates of regression parameters are consistent and asymptotically normal and a robust estimate of their covariance matrix is given that takes into account the correlation of failure times. Simulation results indicate that the presented inference approach works reasonably well.

In Chapter 4, we consider the fitting of the marginal frailty model to multivariate current status data and for inference, the maximum likelihood approach is developed. In particular,an EM algorithm is presented for estimation of parameters. The simulation studies are performed to asses and compare the finite-sample behaviors of the estimators and we apply the proposed methods to an animal tumorigenicity experiment discussed in Section 1.2.3.

Some future research directions are addressed in Chapter 5.

Chapter 2

THE PROPORTIONAL ODDS MODEL FOR MULTIVARIATE INTERVAL-CENSORED FAILURE TIME DATA

2.1 Introduction

This chapter discusses the fitting of the proportional odds model to multivariate interval-censored failure time data. The model is one of the most commonly used regression models in failure time data analysis. As the proportional hazards model, it specifies that covariates have multiplicative effects but on the odds function rather than on the hazard function as the former. One situation in which the proportional odds model is preferred is when the covariate effects diminish over time. In contrast, if the proportional hazards model is applied for the situation, one may have to deal with time-dependent covariates.

Many authors have considered inference for the proportional odds model. For exam-

ple, Bennett (1983) and Murphy *et al.* (1997) discussed the use of the profile likelihood approach for fitting the proportional odds model to right-censored failure time data. Yang and Prentice (1999) considered the same problem but proposed several classes of regression estimators including the pseudo-maximum likelihood estimators. Huang and Rossini (1997) and Rabinowitz *et al.* (2000) investigated regression analysis of current status data and interval-censored data, respectively, using the proportional odds model. All of the approaches described above are for univariate failure time data.

As mentioned before, multivariate failure time data occur when one is interested in several related failure times (Lin, 1994; Hougaard, 2000). In particular, interval-censored data include right-censored data as a special case. One field in which interval-censored failure time data frequently occur is medical follow-up studies and in these cases, each study subject is commonly examined or observed periodically. Goggins and Finkelstein (2000) presented a set of bivariate interval-censored data arising from an AIDS clinical trial on HIV-infected individuals. Kim and Xue (2002) discussed an ongoing clinical trial involving subjects with systemic lupus erythematosus.

A number of authors have discussed regression analysis of multivariate right-censored failure time data and the readers are referred to references described in Chapter 1. In contrast, there exists limited research on the analysis of multivariate interval-censored data. Among others, Betensky and Finkelstein (1999) and Gentleman and Vandal (2002) discussed nonparametric estimation of bivariate survival function and Wang and Ding (2000), Ding and Wang (2004) and Sun *et al.* (2006) studied estimation of the association parameter between two correlated survival variables. Also Dunson

and Dinse (2002) considered the analysis of multivariate case I interval-censored data using Bayesian approaches. For regression analysis, Goggins and Finkelstein (2000) and Kim and Xue (2002) proposed to fit the marginal proportional hazards model to the data and to base the inference on the full likelihood function obtained by treating the correlated survival variables independent. It is well-known that the proportional hazards model may not fit failure time data well sometimes and one simple situation is when treatment effects change with time. Corresponding to this, in the following, we investigate the proportional odds model.

The remainder of this chapter is organized as follows. We begin with describing the model and assumptions in Section 2.2. Section 2.3 presents an inference approach for fitting the proportional odds model to multivariate discrete interval-censored failure time data. For the construction of the likelihood function, following Goggins and Finkelstein (2000) and Wei *et al.* (1989), we use the working independence assumption that assumes that the survival variables of interest are independent. The resulting estimates of regression parameters are consistent and asymptotically normal. For the covariance matrix of the estimated parameters, a robust estimate is given that takes into account the correlation of the survival variables. In Section 2.4, some simulation results are presented and indicate that the presented inference approach works well for practical situations. In Section 2.5, we apply the approach to an AIDS clinical trial and Section 2.6 provides some discussion.

2.2 Models and Assumptions

Consider a survival study that involves K possibly correlated failure times $(T_1, \cdots, T_K)$. Suppose that the T_k's can be observed only to belong to one of M different intervals given by or each study subject is observed only at M time points $0 = t_0 < t_1 < t_2 < \cdots < t_M < t_{M+1} = \infty$. This is usually the case in medical follow-up studies or clinical trials. Note that for simplicity, we use the same set of values or intervals for all T_k here and if they have different sets of intervals, the inference procedure can be developed similarly as below. For T_k, we assume that its marginal distribution function $F_k(t)$ satisfies the following proportional odds model

$$\frac{F_k(t)}{1 - F_k(t)} = \frac{F_{0k}(t)}{1 - F_{0k}(t)} e^{\mathbf{Z}'\beta} \tag{2.1}$$

given $\mathbf{Z}$. In model (2.1), $\mathbf{Z}$ is a p-dimensional vector of covariates, $F_{0k}(t)$ denotes a completely unknown baseline distribution function and β is the p-dimensional vector of regression parameters.

Note that model (2.1) assumes that the baseline distribution functions may be different for different failure times, but the covariate effects for all K failure variables are the same. A situation may occur where the covariate effects may also be different and in this case, one can define a common, big covariate vector. Let $A_{0k}(t) = F_{0k}(t)/(1 - F_{0k}(t))$, the baseline odds function of T_k, $k = 1, \ldots, K$. Then under model (2.1) and given $\mathbf{Z}$, the probability that T_k is observed to belong to the mth interval $(t_{m-1}, t_m]$ is

given by

$$p_{km}(\mathbf{Z}) = \frac{1}{1 + A_{0k}(t_{m-1}) \exp(\mathbf{Z}'\beta)} - \frac{1}{1 + A_{0k}(t_m) \exp(\mathbf{Z}'\beta)}$$

for $m = 1, \ldots, M$, $k = 1, \ldots, K$.

For inference about β, we assume that only interval-censored data about the T_k's are available and they have the form

$$\{ (L_{ik}, R_{ik}] , Z_i ; i = 1, \ldots, n, k = 1, \ldots, K \} .$$

In the above, $(L_{ik}, R_{ik}]$ denotes the interval within which the kth failure of the ith subject is observed to occur and n the number of subjects under study. Here we use the convention that $L_{ik} = R_{ik}$ means that we have an exact observation on the kth failure time of the ith subject and $R_{ik} = t_{M+1} = \infty$ means that the observation on T_{ik} is right-censored. In the following, we assume that $\{L_{ik}, R_{ik}\} \subseteq \{t_m\}$ and define $\alpha_{ikm} = 1$ if $(L_{ik}, R_{ik}]$ contains t_m and $\alpha_{ikm} = 0$ otherwise, $m = 1, \ldots, M + 1$, $k = 1, \ldots, K$, $i = 1, \ldots, n$. Then the likelihood contribution from the kth type of failure of the ith subject is given by

$$L_{ik}(\beta, \underline{A}_k) = \sum_{m=1}^{M+1} \alpha_{ikm} \, p_{km}(\mathbf{Z}_i) ,$$

where $\underline{A}_k = (A_{0k}(t_1), \ldots, A_{0k}(t_M))'$ and $p_{k\,M+1}(\mathbf{Z}_i) = (1 + A_{0k}(t_M) \exp(\mathbf{Z}_i'\beta))^{-1}$. In the next section, we discuss estimation of regression parameters β along with other parameters.

2.3 Parameter Estimation

For estimation of regression parameters β as well as the $\underline{A}_k$'s, following Goggins and Finkelstein (2000) and Wei *et al.* (1989), first we assumes that the K failure types are independent. Under this working independence assumption, the full log-likelihood has the form

$$l(\beta, \underline{A}_k) = \sum_{k=1}^{K} \sum_{i=1}^{n} \log \{L_{ik}(\beta, \underline{A}_k)\}$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} \log \left\{ \sum_{m=1}^{M+1} \alpha_{ikm} \left(\frac{1}{1 + A_{0k}(t_{m-1}) e^{\mathbf{Z}_i'\beta}} - \frac{1}{1 + A_{0k}(t_m) e^{\mathbf{Z}_i'\beta}} \right) \right\}$$

$$= \sum_{k=1}^{K} \sum_{i=1}^{n} \log \left\{ \alpha_{ik1} + \sum_{m=1}^{M} (\alpha_{ikm+1} - \alpha_{ikm}) \left(1 + A_{0k}(t_m)\, e^{\mathbf{Z}_i'\beta} \right)^{-1} \right\}$$

and one can maximize $l(\beta, \underline{A}_k)$ over β and the $\underline{A}_k$'s subject to

$$0 \leq A_{0k}(t_1) \leq \cdots \leq A_{0k}(t_M)\,,$$

since the distributions F_{0k} are all nondecreasing.

In practice, it is convenient to eliminate the parameter range restriction. To this end, define $\Delta_{km} = A_{0k}(t_m) - A_{0k}(t_{m-1})$ and $\gamma_{km} = \log \Delta_{km}$, $m = 1, \ldots, M$, $k = 1, \ldots, K$. Let $\gamma_k = (\gamma_{k1}, \cdots, \gamma_{kM})'$ and $\gamma = (\gamma_1', \cdots, \gamma_K')'$. Then the full log-likelihood can be rewritten as

$$l(\beta, \gamma) = \sum_{k=1}^{K} \sum_{i=1}^{n} \log \Big\{ \alpha_{ik1} + \sum_{m=1}^{M} \Big(\alpha_{ikm+1} - \alpha_{ikm} \Big) \Big(1 + \sum_{j=1}^{m} e^{\gamma_{kj} + \mathbf{Z}_i'\beta} \Big)^{-1} \Big\}\,. \qquad (2.2)$$

To maximize $l(\beta,\gamma)$, one can use, for example, the Newton-Raphson algorithm. For this, we need the first derivatives of $l(\beta,\gamma)$ and they have the forms

$$U_\beta(\beta,\gamma) \;=\; \frac{\partial l(\beta,\gamma)}{\partial \beta} \;=\; -\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ik}^{-1}\, V_{\beta,ik}\,,$$

and

$$U_{\gamma_{km}}(\beta,\gamma) \;=\; \frac{\partial l(\beta,\gamma)}{\partial \gamma_{km}} \;=\; -\sum_{i=1}^{n} W_{ik}^{-1}\, V_{\gamma,ikm}\,,$$

where

$$W_{ik} \;=\alpha_{ik1} \;+\; \sum_{m=1}^{M}(\alpha_{ikm+1}-\alpha_{ikm})(1+\sum_{j=1}^{m} e^{\gamma_{kj}+\mathbf{Z}_i'\beta})^{-1}\,,$$

$$V_{\beta,ik} \;=\; -\frac{\partial W_{ik}}{\partial \beta} \;=\; \sum_{m=1}^{M}\mathbf{Z}_i(\alpha_{ik\,m+1} \;-\; \alpha_{ikm})\sum_{j=1}^{m} e^{\gamma_{kj}+\mathbf{Z}_i'\beta}\Big/(1+\sum_{j=1}^{m} e^{\gamma_{kj}+\mathbf{Z}_i'\beta})^2\,,$$

and

$$V_{\gamma,ikm} \;=\; -\frac{\partial W_{ik}}{\partial \gamma_{km}} \;=\; \sum_{s=m}^{M}(\alpha_{ik\,s+1} \;-\; \alpha_{iks})\, e^{\gamma_{km}+\mathbf{Z}_i'\beta}\Big/(1+\sum_{j=1}^{s} e^{\gamma_{kj}+\mathbf{Z}_i'\beta})^2\,.$$

The second derivatives of $l(\beta,\gamma)$ are given in the Appendix A.

For $k \;=\; 1,\ldots,K$, let $U_{k,\gamma}(\beta,\gamma) \;=\; (U_{\gamma_{k1}}(\beta,\gamma),\cdots,U_{\gamma_{kM}}(\beta,\gamma))'$. Then one can estimate β and γ by $\hat\beta$ and $\hat\gamma$ defined as the solution to the equations

$$U_\beta(\beta,\gamma) \;=\; 0 \;\;,\;\; U_{1,\gamma}(\beta,\gamma) \;=\; 0 \;\;,\;\; \cdots,\;\; U_{M,\gamma}(\beta,\gamma) \;=\; 0\,. \tag{2.3}$$

For large n, the distribution of $(\hat\beta',\hat\gamma')'$ can be approximated by the $(p+KM)$-variate

normal distribution with mean $(\beta_0', \gamma_0')'$ (Goggins and Finkelstein, 2000; Kim and Xue, 2002; Guo and Lin, 1994), where β_0 and γ_0 denote the true values of β and γ, respectively.

For estimation of the covariance matrix of $(\hat{\beta}', \hat{\gamma}')'$, let

$$U_{\beta,i}(\beta, \gamma) \;=\; -\sum_{k=1}^{K} W_{ik}^{-1}\, V_{\beta,ik}\,,$$

and

$$U_{\gamma,ik}(\beta, \gamma) \;=\; -\,W_{ik}^{-1}\,(V_{\gamma,ik1}, \cdots, V_{\gamma,ikM})'$$

for $i = 1, \ldots, n$, $k = 1, ldots, K$. Also let $D_i(\beta, \gamma) = (U'_{\beta,i}(\beta, \gamma), U'_{\gamma,i1}(\beta, \gamma), \cdots, U'_{\gamma,iK}(\beta, \gamma))'$, a $(p + K \times M)$-dimensional vector. Then a consistent estimate of the covariance matrix is given by

$$I^{-1}(\hat{\beta}, \hat{\gamma})\, D(\hat{\beta}, \hat{\gamma})\, I^{-1}(\hat{\beta}, \hat{\gamma})$$

(Kim and Xue, 2002; Guo and Lin, 1994; White, 1982). In the above,

$$D(\beta, \gamma) \;=\; \sum_{i=1}^{n} D_i(\beta, \gamma)\, D_i'(\beta, \gamma)\,,$$

and $I(\beta,\gamma)$ denotes the observed Fisher information matrix and has the form

$$I(\beta,\gamma) = \begin{pmatrix} I_{\beta\beta}(\beta,\gamma) & I_{\beta\gamma_1}(\beta,\gamma) & \cdots & I_{\beta\gamma_K}(\beta,\gamma) \\ I'_{\beta\gamma_1}(\beta,\gamma) & I_{\gamma_1\gamma_1}(\beta,\gamma) & \cdots & I_{\gamma_1\gamma_K}(\beta,\gamma) \\ \vdots & \vdots & \vdots & \vdots \\ I'_{\beta\gamma_K}(\beta,\gamma) & I_{\gamma_K\gamma_1}(\beta,\gamma) & \cdots & I_{\gamma_K\gamma_K}(\beta,\gamma) \end{pmatrix},$$

where $I_{\beta\beta}(\beta,\gamma) = -\partial^2 l(\beta,\gamma)/\partial\beta\partial\beta'$, $I_{\beta\gamma_k}(\beta,\gamma) = -\partial^2 l(\beta,\gamma)/\partial\beta\partial\gamma_k$ and $I_{\gamma_k\gamma_j}(\beta,\gamma) = -\partial^2 l(\beta,\gamma)/\partial\gamma_k\partial\gamma_j$, which are given in the Appendix A as mentioned before, $k,j = 1,\ldots,K$.

Sometimes it is reasonable to assume that the K failure times have the same baseline distribution functions. That is,

$$F_{01}(t) = \cdots = F_{0K}(t) \tag{2.4}$$

in model (2.1). This means that $\gamma_1 = \cdots = \gamma_K$. Let $\gamma^* = (\gamma_1^*,\cdots,\gamma_M^*)'$ denote these common γ_k's. Then the full log-likelihood function $l(\beta,\gamma)$ reduces to $l^*(\beta,\gamma^*) = \sum_{i=1}^n l_i^*(\beta,\gamma^*)$, where

$$l_i^*(\beta,\gamma^*) = \sum_{k=1}^{K} \log\left\{\alpha_{ik1} + \sum_{m=1}^{M}(\alpha_{ikm+1} - \alpha_{ikm})\,(\sum_{j=1}^{m} e^{\gamma_j^* + \mathbf{Z}_i'\beta} + 1)^{-1}\right\}. \tag{2.5}$$

Let $\tilde{\beta}$ and $\tilde{\gamma}^*$ denote the estimates of β and γ^* defined as the solution to the score

equations

$$U^*_\beta(\beta,\gamma^*) \;=\; \frac{\partial l^*(\beta,\gamma^*)}{\partial\beta} \;=\; 0 \;\;,\;\; U^*_{\gamma^*}(\beta,\gamma^*) \;=\; \frac{\partial l^*(\beta,\gamma^*)}{\partial\gamma^*} \;=\; 0\,.$$

Then as with $\hat{\beta}$ and $\hat{\gamma}$, for large n, one can approximate the joint distribution of $\tilde{\beta}$ and $\tilde{\gamma}^*$ by the $(p+M)$-variate normal distribution with mean $(\beta_0', \gamma_0^{*\prime})'$ and covariance matrix $I^{*-1}(\tilde{\beta},\tilde{\gamma}^*)\, D^*(\tilde{\beta},\tilde{\gamma}^*)\, I^{*-1}(\tilde{\beta},\tilde{\gamma}^*)$. In the above, γ_0^* denotes the true value of γ^*,

$$I^*(\beta,\gamma^*) \;=\; -\left(\begin{array}{cc} \partial^2 l^*(\beta,\gamma^*)/\partial\beta\partial\beta' & \partial^2 l^*(\beta,\gamma^*)/\partial\beta\partial\gamma^{*\prime} \\ \partial^2 l^*(\beta,\gamma^*)/\partial\gamma^*\partial\beta & \partial^2 l^*(\beta,\gamma^*)/\partial\gamma^*\partial\gamma^{*\prime} \end{array}\right),$$

and

$$D^*(\beta,\gamma^*) \;=\; \sum_{i=1}^{n} D_i^*\, D_i^{*\prime} \;=\; \sum_{i=1}^{n} \left\{ \frac{\partial l_i^*(\beta,\gamma^*)}{\partial(\beta,\gamma^*)} \right\} \left\{ \frac{\partial l_i^*(\beta,\gamma^*)}{\partial(\beta,\gamma^*)} \right\}'.$$

The expressions of all first and second derivatives used above are given in the Appendix A.

2.4 Simulation Study

To assess the performance of the inference procedure presented in the previous sections, we conducted a simulation study with the focus on estimation of regression parameters. In the study, we considered the situation where there exist $K = 2$ correlated failure times T_1 and T_2 and a scale covariate $\mathbf{Z}$, taking value 0 or 1 with probability 0.5. The joint distribution of T_1 and T_2 was assumed to be given by the bivariate

exponential distribution (Gumbel, 1960)

$$F(t_1, t_2 \,|\, \mathbf{Z}) \;=\; F_1(t_1 \,|\, \mathbf{Z})\, F_2(t_2 \,|\, \mathbf{Z})\, [1 \,+ \alpha\, (1 - F_1(t_1 \,|\, \mathbf{Z}))\, (1 - F_2(t_2 \,|\, \mathbf{Z}))]\,.$$

Here α represents the association parameter, giving the correlation $\rho = \alpha/4$, and F_1 and F_2 denote the exponential distributions, the marginal distributions of T_1 and T_2, specified by the proportional odds model (2.1) with

$$\frac{F_{01}(t)}{1 \,-\, F_{01}(t)} \;=\; 0.2t \;\;,\;\; \frac{F_{02}(t)}{1 \,-\, F_{02}(t)} \;=\; 0.25t\,.$$

For the generation of interval-censored data, motivated by medical follow-up studies, we assumed that each study subject was supposed to be observed at $M = 8$ different time points. At each time point, it was assumed that a subject was observed with probability $1 - q$ independent of observations at the other time points. For subject i and the kth failure time, the observed interval is thus given by the observation times that are right before and after the generated failure time T_{ik}. The results given below are based on 1000 replications and for sample sizes $n = 100$ and 200.

Tables 2.1 and 2.2 present the simulation results obtained for situations where $n = 100$, $\beta_0 = 0$ or 0.5, $\rho = 0.1$, 0.3, 0.5, or 0.7, and $q = 0.1$, 0.3, 0.5, or 0.7. In these tables, we calculated based on simulated data the averages of the regression parameter estimates $\hat{\beta}$ (AVE), the averages of the estimated standard errors of $\hat{\beta}$ (SEE), the sample standard deviations of $\hat{\beta}$ (SSE), and the 95% empirical coverage probabilities (CP). The results suggest that the proposed estimate of the regression parameter seems

to be unbiased and the estimate of the standard error, which is close to the sample standard deviation for most situations considered, is reasonably reliable, especially when the censoring percentage is not too high. The results obtained for $n = 200$ are given in Tables 2.3 and 2.4 and similar to those presented in Tables 2.1 and 2.2. As expected, compared to the case of $n = 100$, both the bias and the standard error become smaller and the empirical coverage probabilities are better. The results also indicate that for situations with smaller sample sizes or large q, the method could give low empirical coverage probabilities. This is because the large q, meaning high censoring percentage, corresponds to wide observed intervals for the survival times of interest and thus means less information about the survival times. More comments on this are given below.

To assess the normal approximation to the distribution of $\hat{\beta}$, we studied the quantile plots of the standardized $\hat{\beta}$ against the standard normal variable for various situations considered in Tables 2.1 - 2.4. Figures 2.1 and 2.2 display the situation where $\beta_0 = 0$, $\rho = 0.5$, $q = 0.3$ and $n = 100$ and $n = 200$, respectively. Among others, not presented here, indicate that the normal approximation seems reasonable. In the simulation study, we also investigated other set-ups including those with different values of β_0 and obtained similar results.

2.5 Analysis of an AIDS Clinical Trial

Now we apply the inference procedure proposed in the previous sections to a set of bivariate interval-censored data arising from an AIDS clinical trial, ACTG 181, on

HIV-infected individuals (Sun, 2006; Goggins and Finkelstein, 2000; Finkelstein *et al.*, 2002). We have described this data set in Section 1.2.4.

Let T_1 and T_2 denote the CMV shedding times in blood and urine, respectively. Following Goggins and Finkelstein (2000), define the covariate Z to be 1 if the baseline CD4 count was less than 75 (in late stage of the HIV disease) or 0 otherwise. To study the effect of the baseline CD4 count on CMV shedding in either blood or urine, we first check if T_1 and T_2 can be described by the proportional odds model (2.1). For this, we obtained the separate nonparametric estimates of the marginal survival functions of T_1 and T_2 for the patients with $Z = 0$ and 1, respectively, using the self-consistency algorithm (Peto, 1973; Turnbull, 1976). Figure 2.3 displays these estimated survival functions of T_1, the time to CMV shedding in blood, along with the estimates of the same functions assuming that model (2.1) is correct for the patients with their baseline CD4 counts both less than 75 and greater than or equal to 75. The same estimates but for T_2, the time to CMV shedding in urine, are given in Figure 2.4. The two figures suggest that model (2.1) provides a reasonable fit to the data set considered here.

For the effect of the baseline CD4 count on blood and urine shedding times, the inference procedure given in the previous sections gave $\hat{\beta} = 1.2637$ with the estimated standard error of 0.1686. This suggests that the patients with baseline CD4 counts lower than 75 were at a significant higher risk of CMV shedding in both the blood and urine than those with baseline CD4 counts over 75. Goggins and Finkelstein (2000) analyzed the same data set assuming that both shedding times follow the marginal proportional hazards model. Based on the maximum likelihood approach, they obtained

0.97 for the estimated effect of the covariate Z defined above with the estimated standard error of 0.197. This result is similar to that given above using the proportional odds model with the latter giving more significant covariate effect. However, it should be noted that the estimated covariate effect given here is on the odds function, while the one given in Goggins and Finkelstein (2000) is on the hazard or survival function.

2.6 Discussion

In this chapter, we considered the fitting of the marginal proportional odds model to multivariate interval-censored failure time data and for inference, the maximum likelihood approach was developed. The proportional odds model is appealing in many situations, especially preferred when covariate or treatment effects change with time and approach to one. In the development of the estimation procedure, following other authors, we made use of the working independence assumption and gave a robust estimate of the covariance matrix of the estimated regression parameters. The method reduces to the usual maximum likelihood approach if there exists only one survival variable of interest ($K = 1$). Numerical studies were performed and they suggest that the presented approach works well for most of practical situations. However, care needs to be exercised if the sample size is small and the censoring percentage is high since in these situations, the coverage probability could be low as shown in the simulation studies. For the situation where the censoring percentage is very high, say, 70% or more, one method that can be used to low the percentage is grouping to reduce the number of probability points or M.

As discussed above, an alternative to the approach developed in this chapter is to fit multivariate interval-censored failure time data to the marginal proportional hazards model using the approach given in Goggins and Finkelstein (2000). A natural question is when one should apply the proportional odds model and the inference procedure given in the previous sections. In practice, two types of situations can occur. One is that the observed failure time data are generated from or could be better approximated by the proportional odds model than the proportional hazards model. The other is that there is no clear choice between the two. For the former, it is apparent that one should choose the proportional odds model since one would get invalid estimated covariate effects otherwise. To demonstrate this, consider the set-up in Table 2.2 with $\rho = 0.5$ and $q = 0.3$. For the simulated data there, instead of fitting the proportional odds model as did in the table, the fitting of the proportional hazards model gave AVE = 0.2810, the average of the estimated covariate effect, which is clearly biased. For the latter type, the decision may be difficult and should be decided based on the subject matter and the interest of the investigator. For example, as pointed out before, the proportional odds model may be a better choice if covariate effects are known to change with time such as diminishing.

In addition to the proportional odds model and the proportional hazards model for regression analysis of multivariate interval-censored data, other alternatives include the additive hazards model, frailty model, and the linear transformation model. All these models have been extensively discussed for univariate failure time data, but there exists limited research on them for multivariate failure time data, especially for multivariate

interval-censored data. We will discussed the analysis of these models in the following chapters.

Chapter 3

THE ADDITIVE HAZARDS MODEL FOR MULTIVARIATE INTERVAL-CENSORED FAILURE TIME DATA

3.1 Introduction

This chapter discusses regression analysis of multivariate interval-censored failure time data using the additive hazards model. As discussed before, an example of such data arises from tumorgenicity experiments. In these studies, one often concerns different types of tumors, tumors in different body locations of animals, or different types of tumors in different locations. Although the time to tumor occurrence is usually the variable of interest, the tumor presence or absence can only be observed when animals die or are sacrificed in most situations. Thus the survival time of interest is either left- or right-censored.

In the following, we first introduce the model and assumptions in Section 3.2. Sec-

tion 3.3 presents an inference approach for regression analysis of multivariate discrete interval-censored failure time data using the additive hazards model. For the construction of the likelihood function, following Goggins and Finkelstein (2000) and Wei *et al.* (1989), we use the working independence assumption that assumes that the survival variables of interest are independent. The resulting estimates of regression parameters are consistent and asymptotically normal and a robust estimate of their covariance matrix is given that takes into account the correlation of the survival variables. In Section 3.4, we present some simulation results that indicate that the presented inference approach works reasonably well for practical situations. Section 3.5 contains some concluding remarks.

3.2 Models and Assumptions

As in Chapter 2, consider a survival study that involves K possibly correlated failure times $(T_1, \cdots, T_K)$. Suppose that the T_k's can be observed only to belong to one of M different intervals given by or each study subject is observed only at M time points $0 = t_0 < t_1 < t_2 < \cdots < t_M < t_{M+1} = \infty$. This is usually the case in medical follow-up studies or clinical trials. Note that for simplicity, we use the same set of values or intervals for all T_k here and if they have different sets of intervals, the inference procedure can be developed similarly as below. Given a p-dimensional vector of covariates $\mathbf{Z}$, it will be assumed that the hazard function of T_k is given by

the following additive hazards model

$$\lambda_k(t) \;=\; \lambda_{0k}(t) \;+\; \mathbf{Z}'\beta\,, \tag{3.1}$$

where $\lambda_{0k}(t)$ denotes a completely unknown baseline hazard function and β is the p-dimensional vector of regression parameters.

Note that model (3.1) assumes that the baseline hazard functions may be different for different failure times, but the covariate effects for all K failure variables are the same. A situation may occur where the covariate effects may also be different and in this case, one can define a common, big covariate vector (Goggins and Finkelstein, 2000). The same is true for the covariates $\mathbf{Z}$. Let $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(s)ds$, the cumulative baseline hazard function of T_k, $k = 1, \ldots, K$. Then under model (3.1) and given $\mathbf{Z}$, the probability that T_k is observed to belong to the mth interval $(t_{m-1}, t_m]$ is given by

$$p_{km}(\mathbf{Z}) \;=\; e^{-\Lambda_{0k}(t_{m-1})-x_{m-1}(\mathbf{Z})'\beta} \;-\; e^{-\Lambda_{0k}(t_m)-x_m(\mathbf{Z})'\beta}\,,$$

where $x_m(\mathbf{Z}) = \int_0^{t_m} \mathbf{Z}(s)ds$, $m = 1, \ldots, M$, $k = 1, \ldots, K$.

For inference about β, we assume that only interval-censored data about the T_k's are available and they have the form

$$\{\, (L_{ik}, R_{ik}]\,,\; Z_i(t)\,;\; i \;=\; 1, \ldots, n,\; k \;=\; 1, \ldots, K\,\}\,.$$

In the above, $(L_{ik}, R_{ik}]$ denotes the interval within which the kth failure of the ith

subject is observed to occur and n the number of subjects under study. Here we use the convention that $L_{ik} = R_{ik}$ means that we have an exact observation on the kth failure time of the ith subject and $R_{ik} = t_{M+1} = \infty$ means that the observation on T_{ik} is right-censored. In the following, we assume that $\{L_{ik}, R_{ik}\} \subseteq \{t_m\}$ and define $\alpha_{ikm} = 1$ if $(L_{ik}, R_{ik}]$ contains t_m and $\alpha_{ikm} = 0$ otherwise, $m = 1, \ldots, M+1$, $k = 1, \ldots, K$, $i = 1, \ldots, n$. Then the likelihood contribution from the kth type of failure of the ith subject is given by

$$L_{ik}(\beta, \underline{\Lambda}_k) = \sum_{m=1}^{M+1} \alpha_{ikm}\, p_{km}(\mathbf{Z}_i)\,,$$

where $\underline{\Lambda}_k = (\Lambda_{0k}(t_1), \ldots, \Lambda_{0k}(t_M))'$ and $p_{kM+1}(\mathbf{Z}_i) = e^{-\Lambda_{0k}(t_M) - x'_M(\mathbf{Z}_i)\beta}$. In the next section, we discuss estimation of regression parameters β along with other parameters.

3.3 Parameter Estimation

To estimate β as well as the $\underline{\Lambda}_k$'s, following Goggins and Finkelstein (2000), and Wei *et al.* (1989), we use the working independence assumption that assumes that the K failure types are independent. Under this working independence assumption, the full log-likelihood function has the form

$$l(\beta, \underline{\Lambda}_k) = \sum_{k=1}^{K} \sum_{i=1}^{n} \log\{L_{ik}(\beta, \underline{\Lambda}_k)\}\,,$$

and one can maximize $l(\beta, \underline{\Lambda}_k)$ over β and the $\underline{\Lambda}_k$'s subject to

$$0 \leq \Lambda_{0k}(t_1) \leq \cdots \leq \Lambda_{0k}(t_M)\,,$$

where $k = 1, \ldots, K$.

To eliminate the parameter range restriction, define $\Delta_{km} = \Lambda_{0k}(t_m) - \Lambda_{0k}(t_{m-1})$ and $\gamma_{km} = \log \Delta_{km}$, $m = 1, \ldots, M$, $k = 1, \ldots, K$. Let $\gamma_k = (\gamma_{k1}, \cdots, \gamma_{kM})'$ and $\gamma = (\gamma_1', \cdots, \gamma_K')'$. Then the full log-likelihood function can be rewritten as

$$l(\beta, \gamma) = \sum_{k=1}^{K} \sum_{i=1}^{n} \log \left\{ \alpha_{ik1} + \sum_{m=1}^{M} (\alpha_{ikm+1} - \alpha_{ikm})\, e^{-\sum_{j=1}^{m} \exp(\gamma_{kj}) - x_m'(\mathbf{Z}_i)\beta} \right\}. \quad (3.2)$$

To maximize $l(\beta, \gamma)$, one can use, for example, the Newton-Raphson algorithm. For this, we need the first derivatives of $l(\beta, \gamma)$ and they have the forms

$$U_\beta(\beta, \gamma) = \frac{\partial l(\beta, \gamma)}{\partial \beta} = -\sum_{i=1}^{n} \sum_{k=1}^{K} W_{ik}^{-1} V_{\beta, ik}\,,$$

and

$$U_{\gamma_{km}}(\beta, \gamma) = \frac{\partial l(\beta, \gamma)}{\partial \gamma_{km}} = -\sum_{i=1}^{n} W_{ik}^{-1} V_{\gamma, ikm}\,,$$

where

$$W_{ik} = \alpha_{ik1} + \sum_{m=1}^{M} (\alpha_{ik\,m+1} - \alpha_{ikm})\, e^{-\sum_{j=1}^{m} \exp(\gamma_{kj}) - x_m'(\mathbf{Z}_i)\beta}\,,$$

$$V_{\beta, ik} = -\frac{\partial W_{ik}}{\partial \beta} = \sum_{m=1}^{M} x_{im} (\alpha_{ik\,m+1} - \alpha_{ikm})\, e^{-\sum_{j=1}^{m} \exp(\gamma_{kj}) - x_m'(\mathbf{Z}_i)\beta}\,,$$

and

$$V_{\gamma,ikm} \;=\; -\,\frac{\partial W_{ik}}{\partial \gamma_{km}} \;=\; \sum_{s=m}^{M} \left(\alpha_{ik\,s+1} \;-\; \alpha_{iks}\right) e^{\gamma_{km} - \sum_{j=1}^{s} \exp(\gamma_{kj}) - x'_m(\mathbf{Z}_i)\beta}\,.$$

The second derivatives of $l(\beta, \gamma)$ are given in the Appendix B.

For $k = 1, \ldots, K$, let $U_{k,\gamma}(\beta, \gamma) = (U_{\gamma_{k1}}(\beta, \gamma), \ldots, U_{\gamma_{kM}}(\beta, \gamma))'$. Then one can estimate β and γ by $\hat{\beta}$ and $\hat{\gamma}$ defined as the solution to the equations

$$U_\beta(\beta, \gamma) \;=\; 0 \;\;,\;\; U_{1,\gamma}(\beta, \gamma) \;=\; 0 \;\;,\;\; \cdots,\;\; U_{K,\gamma}(\beta, \gamma) \;=\; 0\,. \tag{3.3}$$

For large n, the distribution of $(\hat{\beta}', \hat{\gamma}')'$ can be approximated by the $(p + KM)$-variate normal distribution with mean $(\beta_0', \gamma_0')'$ (Goggins and Finkelstein, 2000; Guo and Lin, 1994; Kim and Xue, 2002), where β_0 and γ_0 denote the true values of β and γ, respectively.

To estimate the covariance matrix of $(\hat{\beta}, \hat{\gamma})$, let

$$U_{\beta,i}(\beta, \gamma) \;=\; -\sum_{k=1}^{K} W_{ik}^{-1}\, V_{\beta,ik}\,,$$

and

$$U_{\gamma,ik}(\beta, \gamma) \;=\; -\,W_{ik}^{-1}\,(V_{\gamma,ik1}, \cdots, V_{\gamma,ikM})'$$

for $i = 1, \ldots, n$, $k = 1, \ldots, K$. Also let $D_i(\beta, \gamma) = (U'_{\beta,i}(\beta, \gamma), U'_{\gamma,i1}(\beta, \gamma), \cdots, U'_{\gamma,iK}(\beta, \gamma))'$, a $(p + K \times M)$-dimensional vector. Then a consistent estimate of covariance matrix is given by

$$I^{-1}(\hat{\beta}, \hat{\gamma})\, D(\hat{\beta}, \hat{\gamma})\, I^{-1}(\hat{\beta}, \hat{\gamma})$$

(Kim and Xue, 2002; Guo and Lin, 1994; White, 1982). In the above,

$$D(\beta,\gamma) \;=\; \sum_{i=1}^{n} D_i(\beta,\gamma)\, D_i'(\beta,\gamma)$$

and $I(\beta,\gamma)$ denotes the observed Fisher information matrix and has the form

$$I(\beta,\gamma) \;=\; \begin{pmatrix} I_{\beta\beta}(\beta,\gamma) & I_{\beta\gamma_1}(\beta,\gamma) & \cdots & I_{\beta\gamma_K}(\beta,\gamma) \\ I'_{\beta\gamma_1}(\beta,\gamma) & I_{\gamma_1\gamma_1}(\beta,\gamma) & \cdots & I_{\gamma_1\gamma_K}(\beta,\gamma) \\ \vdots & \vdots & \vdots & \vdots \\ I'_{\beta\gamma_K}(\beta,\gamma) & I_{\gamma_K\gamma_1}(\beta,\gamma) & \cdots, & I_{\gamma_K\gamma_K}(\beta,\gamma) \end{pmatrix},$$

where $I_{\beta\beta}(\beta,\gamma) = -\partial^2 l(\beta,\gamma)/\partial\beta\partial\beta'$, $I_{\beta\gamma_k}(\beta,\gamma) = -\partial^2 l(\beta,\gamma)/\partial\beta\partial\gamma_k$ and $I_{\gamma_k\gamma_j}(\beta,\gamma) = -\partial^2 l(\beta,\gamma)/\partial\gamma_k\partial\gamma_j$, which are given in the Appendix B as mentioned before, $k, j = 1,\ldots,K$.

Sometimes it is reasonable to assume that the K failure times have the same marginal baseline hazard function. That is,

$$\lambda_{01}(t) \;=\; \cdots \;=\; \lambda_{0K}(t) \tag{3.4}$$

in model (3.1). This means that $\gamma_1 = \cdots = \gamma_K$. Let $\gamma^* = (\gamma_1^*, \cdots, \gamma_M^*)'$ denote these common γ_k's. Then the full log-likelihood function $l(\beta,\gamma)$ reduces to $l^*(\beta,\gamma^*) =$

$\sum_{i=1}^{n} l_i^*(\beta, \gamma^*)$, where

$$l_i^*(\beta, \gamma^*) \;=\; \sum_{k=1}^{K} \log \left\{ \alpha_{ik1} \;+\; \sum_{m=1}^{M} \left(\alpha_{ikm+1} \;-\; \alpha_{ikm}\right) e^{-\sum_{j=1}^{m} \exp(\gamma_j^*) - x'_m(\mathbf{Z}_i)\beta} \right\} . \qquad (3.5)$$

Let $\tilde{\beta}$ and $\tilde{\gamma}^*$ denote the estimates of β and γ^* defined as the solution to the score equations

$$U_\beta^*(\beta, \gamma^*) \;=\; \frac{\partial l^*(\beta, \gamma^*)}{\partial \beta} \;=\; 0 \;\;,\;\; U_{\gamma^*}^*(\beta, \gamma^*) \;=\; \frac{\partial l^*(\beta, \gamma^*)}{\partial \gamma^*} \;=\; 0\,.$$

Then as with $\hat{\beta}$ and $\hat{\gamma}$, for large n, one can approximate the joint distribution of $\tilde{\beta}$ and $\tilde{\gamma}^*$ by the $(p+M)$-variate normal distribution with mean $(\beta_0', \gamma_0^{*'})'$ and covariance matrix $I^{*-1}(\tilde{\beta}, \tilde{\gamma}^*)\, D^*(\tilde{\beta}, \tilde{\gamma}^*)\, I^{*-1}(\tilde{\beta}, \tilde{\gamma}^*)$. In the above, γ_0^* denotes the true values of γ^*,

$$I^*(\beta, \gamma^*) \;=\; -\begin{pmatrix} \partial^2 l^*(\beta, \gamma^*)/\partial\beta\partial\beta' & \partial^2 l^*(\beta, \gamma^*)/\partial\beta\partial\gamma^{*'} \\ \partial^2 l^*(\beta, \gamma^*)/\partial\gamma^*\partial\beta & \partial^2 l^*(\beta, \gamma^*)/\partial\gamma^*\partial\gamma^{*'} \end{pmatrix},$$

and

$$D^*(\beta, \gamma^*) \;=\; \sum_{i=1}^{n} D_i^*\, D_i^{*'} \;=\; \sum_{i=1}^{n} \left\{ \frac{\partial l_i^*(\beta, \gamma^*)}{\partial(\beta, \gamma^*)} \right\} \left\{ \frac{\partial l_i^*(\beta, \gamma^*)}{\partial(\beta, \gamma^*)} \right\}' .$$

The expressions of all first and second derivatives used above are given in the Appendix B.

3.4 Simulation Study

We conducted a simulation study for evaluating the performance of the methodology presented in the previous sections with the focus on estimation of regression parameters. In the study, we considered the situation where there exist $K = 2$ correlated failure times T_1 and T_2 and they were generated from the bivariate exponential distribution function (Gumbel, 1960)

$$F(t_1, t_2 \,|\, \mathbf{Z}) \;=\; F_1(t_1 \,|\, \mathbf{Z})\, F_2(t_2 \,|\, \mathbf{Z})\, [1 \,+ \alpha\, (1 - F_1(t_1 \,|\, \mathbf{Z}))\, (1 - F_2(t_2 \,|\, \mathbf{Z}))]\,.$$

In the above, α is the association parameter, giving the correlation $\rho = \alpha/4$, F_1 and F_2 are the marginal exponential distributions with the hazard functions $\lambda_1 + \mathbf{Z}\beta$ and $\lambda_2 + \mathbf{Z}\beta$, respectively, where λ_1 and λ_2 are constants and the covariate $\mathbf{Z}$ is a binary variable taking values 0 and 1 with probability 0.5.

For the generation of interval-censored data, we assumed that each study subject was supposed to be observed at $M = 8$ time points, which is similar to situations in many medical follow-up studies. At each time point, a subject is observed with probability $1 - q$ and independent of observations at the other time points. Thus L_{ik} and R_{ik} are the actual observation times that immediately before and after the true failure time T_{ik}. The results given below are based on $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, and 1000 replications.

Table 3.1 and 3.2 present the simulation results obtained for situations where $n = 100$, $\beta_0 = 0$ or 1, $\rho = 0.1$, 0.3, 0.5, 0.7, and $q = 0.1$, 0.3, 0.5. The tables includes the

averages of the regression parameter estimates $\hat{\beta}$ (AVE) based on simulated data, the averages of the estimated standard errors of $\hat{\beta}$ (SEE), the sample standard deviations of $\hat{\beta}$ (SSE), and the 95% empirical coverage probabilities (cp). The results suggest that the proposed estimate of the regression parameter seems to be unbiased and the estimate of the standard error is close to the sample, standard deviation for most situations considered. To see the effect of sample sizes, Table 3.3 and 3.4 gives the simulation results obtained under the same set-ups as in Table 3.3 and 3.4 except $n = 200$. It gives similar conclusions and as expected, the estimated standard error becomes smaller when the sample size increases.

In some cases, the proposed estimate shows the regression parameter has a bias. If there are no covariate effects, $\beta_0 = 0$, the high censored rates, 0.5, cause biases for small sample sizes, $n = 100$, and it can be improved by larger sample sizes, $n = 200$. If there are strong covariate effects, $\beta_0 = 1$, and high correlation, $\rho = 0.7$, cause biases. Compaired to the high correlation $\rho = 0.7$, we considered covariate effects $\beta_0 = 0.5$, $q = 0.1,\ 0.3,\ 0.5$ and $n = 100,\ 200$ in Table 3.5. All the results suggest that the proposed estimate of the regression parameter seems to be unbiased.

To investigate the normal approximation to the distribution of $\hat{\beta}$, we studied the quantile plots of the standardized $\hat{\beta}$ against the standard normal variable for various practical situations considered in Tables 3.1-3.4. These plots indicate that the approximation seems to be reasonably good, especially for the case of $n = 200$. For example, Figure 3.1 and displays the quantile plot for the case where $\beta_0 = 0$, $n = 100$, $(q, \rho) = (0.3, 0.3)$ and Figure 3.2 displays $\beta_0 = 0$, $n = 200$, $(q, \rho) = (0.3, 0.3)$ respec-

tively. Other plots are similar and omitted.

3.5 Discussion

This chapter discussed regression analysis of multivariate interval-censored failure time data using the additive hazards model and a marginal inference approach is presented. The simulation study was conducted and indicates that the presented approach works well for practical situations. However, in some cases of the high censored rates, the coverage probability could be low as shown in the simulation studies ($\beta_0 = 0$). For the situation where the censoring percentage is very high, say 50% or more, one method that can be used to low the percentage is grouping to reduce the number of probability points or M.

Chapter 4

A FRAILTY MODEL APPROACH FOR MULTIVARIATE CURRENT STATUS DATA

4.1 Introduction

This chapter discusses the fitting of the frailty model to multivariate current status data. The frailty model is commonly used in the analysis of multivariate failure time data and provides a flexible approach for directly modeling the relationship among correlated failure times. In the preceding chapters, we have developed the marginal model approach (Goggins and Finkelstein, 2000; Kim and Xue, 2002) for multivariate interval-censored failure time data using the proportional odds model and the additive hazards model. In this chapter, we will discuss the random effect model approach for regression analysis of multivariate current status data.

Random effect model approach assumes that there exists a common and an unobserved latent random variable, also called frailty, and correlated failure times are independent given the frailty variable. The frailty variable represents the dependence

of the correlated failure times. Compared with the marginal model approach, one advantage of the frailty model approach is that it directly models the correlation of failure times.

Among these, Clayton and Cuzick (1985) extended the proportional hazards model and included a random effect representing heterogeneity of subjects. Oakes (1989) considered the frailty model for bivariate failure time data. Huang and Wolf (2002) treated censoring as informative by using the frailty model approach.

In the following sections, we first present models and assumptions in Section 4.2. Section 4.3 discusses estimation of unknown parameters by maximizing the likelihood function. For this procedure, we develop an EM algorithm. The key behind using the EM algorithm is to treat the unobservable random variable as the missing value. The resulting estimates of regression parameters are consistent and asymptotically normal. For the covariance matrix of the estimated parameters, a robust estimate is given that takes into account the correlation of the survival variables. In Section 4.4, some simulation results are presented and indicate that the presented inference approach works well for practical situations. We apply the approach to the NTP tumor data in Section 4.5. Section 4.6 contains some discussion.

4.2 Models and Assumptions

Consider a survival study consisting of K failure times $(T_1, \cdots, T_K)$ and suppose that for T_k, its hazard function is given by

$$\lambda_k(t) = \lambda_{0k}(t)\, e^{x_k'\beta + b_k}, \tag{4.1}$$

where $b_1, \cdots, b_K$ are the latent variables satisfying

$$(b_1, \cdots, b_K) \sim N(0, \Sigma).$$

In the following, we assume that given $(b_1, \cdots, b_K)$, $T_1, \cdots, T_K$ are independent. Let C_k be the monitoring censoring time for kth type failure time and τ be a preselected constant. Note from Wang and Ding (2000) and Ding and Wang (2004) that if $T_1, \cdots, T_K$ are measured from the same subjects, then the C_k's can be assumed to be equal, that is $C_1 = \cdots = C_K = C$. In the following, we will focus on this univariate censoring situation and the approach proposed below can be easily generalized to the unequal monitoring times cases. Then given b_k and current status data $\delta_k = I(T_k \leq C \wedge \tau)$ and $C = t$, we can write the probability of observing δ_k at time t as

$$P(\delta_k | x_k, b_k, C = t) = [1 - e^{-\Lambda_{0k}(t)\exp(x_k'\beta + b_k)}]^{\delta_k} e^{-(1-\delta_k)\Lambda_{0k}(t)\exp(x_k'\beta + b_k)}.$$

In the above expression for the likelihood, the functions $\Lambda_{0k}(t)$ are all unknown, so

they also need to be estimated. To this end, the time interval $(0, \tau]$ is divided by J intervals $I_j = (s_{j-1}, s_j]$ for $j = 1, \cdots, J$ with $0 = s_0 < s_1 < \cdots < s_J = \tau$ and $\Lambda_{0k}(t)$ is assumed be a step function being a constant within each interval. To be specific, we assume that $\Lambda_{0k}(t) = \sum_{j=1}^{J} e^{\gamma_{kj}} I_j(t)$, where $I_j(t)$ takes the value one if t is in the interval I_j and zero otherwise and $\gamma_k = (\gamma_{k1}, \cdots, \gamma_{kJ})'$ are $J-$ dimensional parameters to be estimated. Since the functions Λ_{0k} are increasing, the parameters $\gamma_{k1}, \cdots, \gamma_{kJ}$ have the restriction $\gamma_{k1} \leq \gamma_{k2} \leq \cdots \leq \gamma_{kJ}$. In order to remove this range restriction, we re-parameterize the γ_k's as

$$\Lambda_{0k}(t) = \sum_{j=1}^{J} I_j(t) \sum_{a=1}^{j} e^{\gamma_{ka}}.$$

After being parameterized by the above expressions, the range restriction for the parameters γ_k's disappears. Write $\gamma = (\gamma_1, \cdots, \gamma_K)$.

For each subject, since the multivariate failure times $(T_1, \cdots, T_K)$ are independent conditional on the frailty vector $b = (b_1, \cdots, b_K)'$ and covariate $x = (x_1, \cdots, x_K)'$, the likelihood can be expressed as the product of the above K probabilities, that is

$$\begin{aligned} L^*(\theta; O) = \int \prod_{k=1}^{K} \sum_{j=1}^{J} \Big(I_j(t) [1 - e^{-\sum_{a=1}^{j} \exp(\gamma_{ka}) \exp(x_k' \beta + b_k)}]^{\delta_k} \\ \times e^{-(1-\delta_k) \sum_{a=1}^{j} \exp(\gamma_{ka}) \exp(x_k' \beta + b_k)} \Big) f(b; \Sigma) db, \end{aligned} \tag{4.2}$$

where $\theta = (\beta, \Sigma, \gamma)$ is the parameter to be estimated, $O = (\delta, x, t)$ with $\delta = (\delta_1, \cdots, \delta_K)'$ is the observed dataset, and $f(b; \Sigma)$ is the density function of the normal distribution with mean zero and covariance Σ.

Note that the joint density function of O and b is the integral part in (4.2). Then the conditional density function of b given O is given by

$$f(b|O,\theta) = \frac{1}{L^*(\theta;O)} \prod_{k=1}^{K} \sum_{j=1}^{J} \Big(I_j(t)[1 - e^{-\sum_{a=1}^{j} \exp(\gamma_{ka}) \exp(x_k'\beta + b_k)}]^{\delta_k} \times e^{-(1-\delta_k)\sum_{a=1}^{j} \exp(\gamma_{ka}) \exp(x_k'\beta + b_k)} \Big) f(b;\Sigma) \tag{4.3}$$

Then given the observation data set $O_i = (\delta_i, x_i, t_i), (i = 1, \cdots, n)$ for n independent subjects, the likelihood can be written as the product of the n likelihoods, that is $L(\theta;O) = \prod_{i=1}^{n} L^*(\theta;O_i)$. The conditional density function $b_i = (b_{i1}, \cdots, b_{iK})'$ given the observation set O is only related to the ith observation O_i, that is, $f(b_i|O,\theta) = f(b_i|O_i,\theta)$, which is defined in (4.3) just replacing (O, x_k, δ_k, t_k) with the ith observational value $(O_i, x_{ik}, \delta_{ik}, t_i)$, that is

$$f(b_i|O_i,\theta) = \frac{1}{L^*(\theta;O_i)} \prod_{k=1}^{K} \sum_{j=1}^{J} I_j(t) \Big[1 - e^{-\sum_{a=1}^{j} \exp(\gamma_{ka}) \exp(x_{ik}'\beta + b_{ik})} \Big]^{\delta_{ik}}$$

$$\times e^{-(1-\delta_{ik})\sum_{a=1}^{j} \exp(\gamma_{ka}) \exp(x_{ik}'\beta + b_{ik})} f(b_i;\Sigma).$$

4.3 Estimation Procedure

In this section, we discuss estimation of unknown parameters θ by maximizing the likelihood function $L(\theta;O)$. For this, we develop an EM algorithm.

E-step

The key behind using the EM estimation procedure is to treat the unobservable

random variable b_i as the missing value. To be specific, the complete data consist of two parts, one is the observable data $O = (O_1, \cdots, O_n)$ and the other is the missing data $b = (b_1, \cdots, b_n)'$. First, we can write the likelihood of the complete data and then compute the expectation of the log-likelihood with respect to the conditional density function of b given O if the parameter θ is the "true" $\theta^{(m)}$ at the mth iteration. Given the complete data, the log-likelihood can be written as $l(\theta; O, b) = \sum_{i=1}^{n} l_i(\theta; O_i, b_i)$, where

$$l_i = \log f(b_i; \Sigma) + \sum_{k=1}^{K} \sum_{j=1}^{J} I_j(t_i) \Big[\delta_{ik} \log \Big(1 - e^{-\sum_{a=1}^{j} \exp(\gamma_{ka}) \ \exp(x'_{ik}\beta + b_{ik})} \Big)$$

$$- (1 - \delta_{ik}) \sum_{a=1}^{j} e^{\gamma_{ka}} e^{x'_{ik}\beta + b_{ik}} \Big].$$

Then the expectation of the above log-likelihood can be written as

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

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

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

Let $v_i = [\Sigma^{(m)}]^{-1/2} b_i$, then $b_i = [\Sigma^{(m)}]^{1/2} v_i$. Then the above equality can be rewrit-

ten as the function of v_i. It is apparent that the denominator and numerator of the above equality can be interpreted as the expectation of one function for v_i with respect to a standard normal distribution. To be specific, $E(h(b_i)|O_i, \theta^{(m)})$ equals to

$$E(h(b_i)|O_i, \theta^{(m)}) = \frac{E[\psi(v_i; \theta^{(m)}, O_i)h([\Sigma^{(m)}]^{1/2}v_i)]}{E\psi(v_i; \theta^{(m)}, O_i)},$$

where

$$\psi(v_i; \theta^{(m)}, O_i) = \prod_{k=1}^{K}\sum_{j=1}^{J}\Big\{ I_j(t_i)\Big[1 - e^{-\sum_{a=1}^{j}\exp(\gamma_{ka}^{(m)})\exp(x_{ik}'\beta^{(m)} + b_{ik})}\Big]^{\delta_{ik}}$$

$$\times e^{-(1-\delta_{ik})\sum_{a=1}^{j}\exp(\gamma_{ka}^{(m)})\exp(x_{ik}'\beta^{(m)} + b_{ik})}\Big\}$$

and the expectation on the numerator and denominator is taken with respect to the standard $K-$ dimensional normal distribution. Therefore, for sufficiently large L, the expectation $E(h(b_i)|O_i, \theta^{(m)})$ can be approximated by

$$E(h(b_i)|O_i, \theta^{(m)}) \simeq \widehat{E}(h(b_i)) = \frac{\sum_{j=1}^{L}\psi(v_j; \theta^{(m)}, O_i)h([\Sigma^{(m)}]^{1/2}v_j)}{\sum_{j=1}^{L}\psi(v_j; \theta^{(m)}, O_i)}, \tag{4.4}$$

where $v_j = (v_{j1}, \cdots, v_{jK})' \overset{\text{i.i.d}}{\sim} N_K(0, I_K)$.

M-Step

In this maximization step, we need to maximize the conditional expectation $El_i(\theta; O_i, b_i)$ and then obtain the $m+1$th iteration estimator $\theta^{(m+1)}$. Using the ideas in the E-step,

all the values $h(b_i)$ involved in the expression $l(\theta; O, b)$ are all replaced by their approximation $\widehat{E}(h(b_i))$. First, taking derivatives of $l(\theta; O, b)$ with respect to Σ, we obtain the $m+1$th maximum likelihood estimator of Σ as

$$\Sigma^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{E}(b_i b_i'), \tag{4.5}$$

where $\widehat{E}(b_i b_i')$ can be approximated by (4.4).

Secondly, we turn to solve the maximum likelihood estimator of parameters β and γ. Taking derivatives of $El(\theta; O, b)$ with respect to β and γ gives to the score equations as

$$\begin{aligned} U_\beta(\beta,\gamma) = &\quad \frac{\partial l(\theta; O, b)}{\partial \beta} = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{a=1}^{j} \exp(\gamma_{ka}) I_j(t_i) x_{ik} E\psi_{ikj}^{(1)}(b_i; \beta, \gamma_k), \\ U_{\gamma_{kj}}(\beta,\gamma) = &\quad \frac{\partial l(\theta; O, b)}{\partial \gamma_{kj}} = \sum_{i=1}^{n}\sum_{h=j}^{J} I_h(t_i)\exp(\gamma_{kj}) E\psi_{ikh}^{(1)}(b_i; \beta, \gamma_k), \end{aligned}$$

where

$$\psi_{ikj}^{(1)}(b_i; \beta, \gamma_k) = \Big[\frac{\delta_{ik}}{1 - \exp[-\sum_{a=1}^{j}\exp(\gamma_{ka})\exp(x_{ik}'\beta + b_{ik})]} - 1\Big]\exp(x_{ik}'\beta + b_{ik}).$$

Similar to the above arguments, the expectations expressed in the score equations have no closed form, therefore we also need approximations. Using the approximations $\widehat{E}$ in (4.4) to replace the E's terms and noticing that the denominators in (4.4) are all constants, we can obtain the working score equations as follows

$$\widehat{U}_{\beta}(\beta,\gamma)=\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{J}\sum_{a=1}^{j}\exp(\gamma_{ka})I_j(t_i)x_{ik}\frac{\sum_{l=1}^{L}\psi_{ikj}^{(1)}(b_l;\beta,\gamma_k)\psi(v_l;\theta^{(m)},O_i)}{\sum_{l=1}^{L}\psi(v_l;\theta^{(m)},O_i)},$$

$$\widehat{U}_{\gamma_{kj}}(\beta,\gamma)=\sum_{i=1}^{n}\sum_{h=j}^{J}I_h(t_i)\exp(\gamma_{kj})\frac{\sum_{l=1}^{L}\psi_{ikh}^{(1)}(b_l;\beta,\gamma_k)\psi(v_l;\theta^{(m)},O_i)}{\sum_{l=1}^{L}\psi(v_l;\theta^{(m)},O_i)},$$

where $b_l=[\Sigma^{(m)}]^{1/2}v_l$ and $v_1,\cdots,v_L$ are i.i.d random samples from the normal distribution $N_K(0,I_K)$.

Computational Issues

Let $\widehat{U}_{\gamma_k}(\beta,\gamma)=(\widehat{U}_{\gamma_{k1}}(\beta,\gamma),\cdots,\widehat{U}_{\gamma_{kJ}}(\beta,\gamma))'=\widehat{U}_{\gamma_k}(\beta,\gamma_k)$ be the $J-$ vector. Then the working score estimation equations can be written as

$$\widehat{U}(\beta,\gamma)=(\widehat{U}_{\beta}(\beta,\gamma)',\widehat{U}_{\gamma_1}(\beta,\gamma_1)',\cdots,\widehat{U}_{\gamma_K}(\beta,\gamma_K)')'=0.$$

From the expressions above, one can see that it is not easy to solve the above $p+K\times J$ equations simultaneously. To overcome this difficulty, by noting that the equations $\widehat{U}_{\gamma_k}(\beta,\gamma_k)$ only involve the parameters γ_k and β but not any other parameters γ_j for any $j\neq k$, we can solve these equations separately as below. First obtain reasonable initial values $\widehat{\theta}^{(0)}$ for all parameters and at the mth stage, carry out the following steps.

Step 1. Estimate $\widehat{\Sigma}^{(m+1)}$ by (4.5).

Step 2. Estimate $\widehat{\beta}^{(m+1)}$ by solving the working score equation $\widehat{U}_{\beta}(\beta,\gamma^{(m)})=0$.

Step 3. After obtaining $\widehat{\beta}^{(m+1)}$, estimate the $\widehat{\gamma}_1^{(m+1)}, \cdots, \widehat{\gamma}_K^{(m+1)}$ by solving the equations $\widehat{U}_{\gamma_1}(\widehat{\beta}^{(m+1)}, \gamma_1) = 0, \cdots, \widehat{U}_{\gamma_K}(\widehat{\beta}^{(m+1)}, \gamma_K) = 0$ one by one.

Repeat Steps 1-3 until convergence.

The procedure described above has several advantages: (1). it only involves solving several low-dimensional equations instead of high-dimensional equations; (2). the computation is more stable and efficient; (3). the solution to the $m+1$th iterative equations can be easily implemented via the usual Newton-Raphson algorithm. These tasks can be also done through the software packages such as Matlab.

As for the variance estimation of the estimator $\widehat{\theta}$, one can use the inverse of the observed information matrix $I(\widehat{\theta})$, which is given in the Appendix C.

4.4 Simulation Studies

Simulation studies were carried out to assess the finite sample performance of the inference procedure presented in the previous sections. In the study, we considered the situation where there exist $K = 2$ correlated failure times T_1 and T_2 and a scale covariate $\mathbf{X}$ taking value 0 or 1 with probability 0.5. The joint cumulative distribution of T_1 and T_2 was assumed to be

$$F(t_1, t_2 \,|\, b, \mathbf{X}) \;=\; F_1(t_1 \,|\, b, \mathbf{X})\, F_2(t_2 \,|\, b, \mathbf{X})\,.$$

In the above, F_1 and F_2 denote the exponential distributions with the hazard functions $\lambda_{01}e^{X\beta_0+b_1}$ and $\lambda_{02}e^{X\beta_0+b_2}$, respectively.

In the study, we assumed that the monitoring times are the same and there exist J different time points for the monitoring time. We generated an indicator interval of monitoring time for each subject. Based on the indicator interval and failure times T_1, T_2, we have current status data δ_1 and δ_2. We used $L = 30$ for the approximation of the expectation $E(h(b_i)|O_i, \theta^{(m)})$. For the results presented below, we took the baseline hazard functions λ_{01} and λ_{02} to be $0.04t_1$ and $0.02t_2$, respectively, with 1000 replications and sample sizes $n = 100$ and 200.

In the first simulation study, we considered three different monitoring intervals, J=10, 15, 20 with the standard deviation of the latent variables b_1 and b_2 being 0.1. Table 4.1 presents the simulation results obtained for situations where n=100, and $\beta_0 = 0$, 1, -1. The table includes the averages of the regression parameter estimates $\hat{\beta}$ (AVE), the averages of the estimated standard errors of $\hat{\beta}$ (SEE), the sample standard deviations of $\hat{\beta}$ (SSE), and the 95% empirical coverage probabilities (CP). The results suggest that the proposed estimate of the regression parameter seems to be unbiased and the estimate of the standard error, which is close to the sample standard deviation for most situations considered, is reasonably reliable. To see the effect of sample sizes, Table 4.2 gives the simulation results obtained under the same set-ups as in Table 4.1 except n=200. It gives similar conclusions and as expected the estimated standard error becomes smaller when the sample size increases.

In the second simulation study, we considered smaller monitoring intervals, J=5,

10 with the standard deviation of the latent variables b_1 and b_2 being 0.5. Table 4.3 presents the simulation results obtained for situations where n=100 and n=200, and $\beta_0 = 0.5$, -0.5. Again, the estimates of the regression parameter seem to be unbiased, SEEs and SSEs are quite close, and the 95% coverage probabilities are close to 0.95. When the sample size increases, it also gives similar conclusions and as expected the estimated standard error becomes smaller.

To assess the normal approximation to the distribution of $\hat{\beta}$, we studied the quantile plots of the standardized $\hat{\beta}$ against the standard normal variable for various situations considered in Tables 4.1 and 4.2. Figures 4.1 and 4.2 display the situation where $\beta_0 = 0.5$, J=10, $\sigma = 0.5$ (the standard deviation of the latent variables b_1 and b_2) and $n = 100$ and $n = 200$, respectively. They indicate that the normal approximation seems reasonable. Other plots are similar and omitted.

4.5 Analysis of a National Toxicology Program Study

In this section, we apply the methodology presented in the previous sections to the animal tumorigenicity experiment described in Section 1.2.3. As described before, the animals either died during the study or were sacrificed at the end of the study. At the death or sacrifice, the presence or absence of tumors was determined through a pathologic examination. Thus the tumor occurrence times were not exactly observed but instead known only to be smaller or greater than the death or sacrifice time. In other words, we only have interval-censored observations on tumor occurrence times. Following Dunson and Dinse (2002), we will focus on two types of tumors, adrenal and

lung tumors, on the male rats from the control and 80 ppm dose groups. The goal here is to compare the tumor growth rates between the control and dose groups based on bivariate case I interval-censored data and we assume that the death or sacrifice times follow the same distribution for all animals.

To apply the presented methodology, let T_{i1} and T_{i2} denote the occurrence times of adrenal and lung tumors for the ith animal and define $X_i = 1$ if the ith animal was in the dose group and 0 otherwise. To give a graphical view about the dose effect on the tumor growth, Figures 4.3 and 4.4 present the estimated marginal survival functions for the time to the adrenal and lung tumor for animals in the control and dose groups given by the nonparametric maximum likelihood (NPML) function approach (Barlow *et al.*, 1972; Robertson *et al.*, 1988). They suggest that the two types of tumors seem to have different baseline hazard functions and survival functions and indicate that there seems to exist significant dose effect on both adrenal and lung tumor.

By assuming that the baseline hazard functions are different, the application of the method gave $\hat{\beta} = 0.7958$ with the estimated standard error being 0.3665. This suggests that the animals in the dose group had significantly higher occurrence rates of both adrenal and lung tumors. For comparison, we also fitted the data to the proportional hazards model without frailty assuming different baseline hazard functions for adrenal and lung tumors but the same dose effect, we obtained the estimated dose effect of 0.6270 and the p-value for testing no dose effect being 0.1114.

Figures 4.5-4.6 present the estimated marginal survival functions for time to adrenal tumor under model (4.1) and the proportional hazards model, respectively. Figures

4.7-4.8 present the estimated marginal survival functions for time to lung tumor under model (4.1) and the proportional hazards model, respectively. These estimated survival functions of adrenal and lung tumors along with the the estimates of the same functions assuming that model (4.1) is correct for the existing significant dose effect on the adrenal and lung tumors. Again it indicate that our proposed method seems to be more reasonable and appropriate than the marginal proportional hazards model.

4.6 Discussion

In this chapter, we considered the fitting of the marginal frailty model to multivariate current status data and for inference, the maximum likelihood approach was developed. For estimation of parameters, an EM algorithm was developed that only involves several low-dimensional equations and is more stable and efficient. The simulation study indicated that our method works well for both of shorter and longer monitoring times. The method was applied to a set of bivariate case I interval-censored data.

In both the preceding chapter and this chapter, we mentioned the same problem of choosing an appropriate model among all available models. There does not exist an approach in the literature that can be used to distinguish the semiparameter regression models for both univariate and multivariate interval-censored failure time data. For the future work, a relatively simple question is to develop statistical methods to assess the goodness-of-fit for these semiparameter regression models. Also, we only considered the situations where study subjects are observed or monitored at finite time points for

all our works in Chapters 2 and 3. Although this covers many studies yielding interval-censored data, it would be useful to develop statistical methods for situations in which both survival times of interest and observation or monitoring times are continuous variables.

Chapter 5

FUTURE RESEARCH

In the preceding chapters, we mentioned the problem of choosing an appropriate model among all available semiparameter regression models. To the best of our knowledge, there does not exist an approach in the literature that can be used to distinguish the semiparameter regression models for both univariate and multivariate interval-censored failure time data. We will briefly discuss this question in Section 5.1.

We have developed the random effect model approach for multivariate current status data. Huang and Wolf (2002) treated censoring as informative for the frailty model. We will briefly discuss informative censoring in Section 5.2

5.1 A Goodness-of-fit Test for Multivariate Interval-censored Failure Times

One question for regression analysis of multivariate interval-censored data is how one can choose an appropriate model among all available models. Except the graphical

approach used in Sections 2.5 and 4.5, there does not seem to exist an approach in the literature that can be used to choose or distinguish these different models and it is apparent that the development of such approach would be very useful. A topic for future research is to develop statistical methods to assess the goodness-of-fit for the each of these common semiparametric regression models with multivariate interval-censored data.

5.2 A Frailty Model Approach for Case II Multivariate Informative Interval Censoring

In Chapter 4, we developed the random effect model approach for multivariate current status data. However, we didn't consider for informative censoring. It is useful to study how to generalize the proposed methods to censoring by some causes to be analyzed as informative while treating censoring by other causes as noninformative. Moreover, we focus on current status or case I interval-censored data in Chapter 4. For future research, developing a frailty model approach for case II multivariate interval-censored data is needed.

APPENDIX

Appendix A: Expression of the Observed Fisher Information Matrix in Chapter 2

Let the W_{ik}'s, $V_{\beta,ik}$'s, $V_{\gamma,ikm}$'s, $I(\beta,\gamma)$, $I^*(\beta,\gamma^*)$ and their components be defined as in the previous sections. Define $b_{ikm} = \sum_{j=1}^{m} \exp(\gamma_{kj} + Z_i'\beta)$ and

$$V_{\beta\beta,ik} = \sum_{m=1}^{M} Z_i Z_i' \left(\alpha_{ikm+1} - \alpha_{ikm}\right) b_{ikm}(b_{ikm}-1)/(1+b_{ikm})^3,$$

$$V_{\gamma\gamma,ikm} = \sum_{s=m}^{M} \left(\alpha_{iks+1} - \alpha_{iks}\right)(1-b_{iks})/[(1+b_{iks})^3 \exp(\gamma_{km} + Z_i'\beta)]\,,$$

$$V_{\gamma\beta,ikm} = -\frac{\partial V_{\gamma,ikm}}{\partial \beta} = \sum_{s=m}^{M} Z_i(\alpha_{iks+1} - \alpha_{iks}) \exp(\gamma_{km} + Z_i'\beta)(b_{iks}-1)/(1+b_{iks})^3\,,$$

and for $j < s$

$$V_{\gamma\gamma,ikmj} = -\frac{\partial V_{\gamma,ikm}}{\partial \gamma_{kj}} = 2\sum_{s=m}^{M}(\alpha_{iks+1} - \alpha_{iks}) \exp(\gamma_{km} + \gamma_{kj} + 2Z_i'\beta)/(1+b_{iks})^3\,.$$

Then we have

$$I_{\beta\beta}(\beta,\gamma) \;=\; \sum_{i=1}^{n}\sum_{k=1}^{K}\left(\, W_{ik}^{-2}V_{\beta,ik}V_{\beta,ik}' \;-\; W_{ik}^{-1}V_{\beta\beta,ik}\,\right),$$

$$-\,\frac{\partial^2 l(\beta,\gamma)}{\partial\gamma_{km}\partial\beta} \;=\; \sum_{i=1}^{n}\left(\, W_{ik}^{-2}V_{\gamma,ikm}V_{\beta,ik} \;-\; W_{ik}^{-1}V_{\gamma\beta,ikm}\,\right),$$

the mth column of $I_{\beta,\gamma_k}(\beta,\gamma)$, and $I_{\gamma_k,\gamma_j}(\beta,\gamma) \;=\; 0_{M\times M}$ for any $1 \le k \ne j \le K$. Furthermore, for $I_{\gamma_k,\gamma_k}(\beta,\gamma)$, its (m,m)th diagonal element is given by

$$-\,\frac{\partial^2 l(\beta,\gamma)}{\partial\gamma_{km}\partial\gamma_{km}} \;=\; \sum_{i=1}^{n}\left(\, W_{ik}^{-2}V_{\gamma,ikm}^{2} \;-\; W_{ik}^{-1}V_{\gamma\gamma,ikm}\,\right)$$

and for any $j \;<\; m$, its (m,j)th element has the form

$$-\,\frac{\partial^2 l(\beta,\gamma)}{\partial\gamma_{km}\partial\gamma_{kj}} \;=\; \sum_{i=1}^{n}\left(\, W_{ik}^{-2}V_{\gamma,ikm}V_{\gamma,ikj} \;-\; W_{ik}^{-1}V_{\gamma\gamma,ikmj}\,\right).$$

Now we consider the special situation where (2.4) is true. For this, for $k = 1,\cdots,K$ and $m = 1,\cdots,M$, define $b_{im} = \sum_{j=1}^{m}\exp(\gamma_j^* + Z_i'\beta)$ and

$$W_{ik} \;= \alpha_{ik1} \;+\; \sum_{m=1}^{M}(\alpha_{ikm+1} \;-\; \alpha_{ikm})\,(\sum_{j=1}^{m}\exp(\gamma_j^* + Z_i'\beta) + 1)^{-1},$$

$$V_{\gamma^*,ikm} \;=\; -\,\frac{\partial W_{ik}}{\partial\gamma_m^*} \;=\; \sum_{s=m}^{M}(\alpha_{iks+1} - \alpha_{iks})\,\exp(\gamma_m^* + Z_i'\beta)/(1+b_{im})^2,$$

$$V_{\beta,ik} \;=\; -\,\frac{\partial W_{ik}}{\partial\beta} \;=\; \sum_{m=1}^{M} Z_i(\alpha_{ikm+1} - \alpha_{ikm})b_{im}/(1+b_{im})^2\,,$$

$$V_{\beta\beta,ik} \;=\; -\,\frac{\partial V_{\beta,ik}}{\partial \beta'} \;=\; \sum_{m=1}^{M} Z_i Z_i'(\alpha_{ikm+1}-\alpha_{ikm})\, b_{im}(b_{im}-1)/(1+b_{im})^3\,,$$

$$V_{\gamma^*\beta,ikm} \;=\; -\,\frac{\partial V_{\gamma^*,ikm}}{\partial \beta} \;=\; \sum_{s=m}^{M} Z_i(\alpha_{iks+1}-\alpha_{iks})\exp(\gamma_m^*+Z_i'\beta)(b_{im}-1)/(1+b_{im})^3,$$

$$V_{\gamma^*\gamma^*,ikm} \;=\; -\,\frac{\partial V_{\gamma^*,ikm}}{\partial \gamma_m^*} \;=\; \sum_{s=m}^{M}(\alpha_{iks+1}-\alpha_{iks})\exp(\gamma_m^*+Z_i'\beta)\Big[\frac{2\exp(\gamma_m^*+Z_i'\beta)}{1+b_{im}}-1\Big]/(1+b_{im})^2,$$

and for $j \neq m$,

$$V_{\gamma^*\gamma^*,ikmj} \;=\; -\,\frac{\partial V_{\gamma^*,ikm}}{\partial \gamma_j^*} \;=\; \sum_{s=m}^{M} 2(\alpha_{is+1}-\alpha_{is})\exp(\gamma_m^*+\gamma_j^*+2Z_i'\beta)/(1+b_{im})^3\,.$$

Using the notation defined above, we have

$$U_\beta^*(\beta,\gamma^*) \;=\; -\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ik}^{-1}\, V_{\beta,ik}\;\;,$$

$$U_{\gamma^*}^*(\beta,\gamma^*) \;=\; -\sum_{i=1}^{n}\sum_{k=1}^{K} W_{ik}^{-1}(V_{\gamma^*,ik1},\cdots,V_{\gamma^*,ikM})\,,$$

$$D_i^* \;=\; -\sum_{k=1}^{K} W_{ik}^{-1}(V_{\beta,ik}',V_{\gamma^*,ik1},\cdots,V_{\gamma^*,ikM})'\,,$$

$$-\frac{\partial^2 l^*(\beta,\gamma^*)}{\partial\beta\partial\beta'} \;=\; \sum_{i=1}^{n}\sum_{k=1}^{K}\left(W_{ik}^{-2}V_{\beta,ik}V_{\beta,ik}' \;-\; W_{ik}^{-1}V_{\beta\beta,ik}\right)\,,$$

$$-\frac{\partial^2 l^*(\beta,\gamma^*)}{\partial\gamma_m^*\partial\beta} \;=\; \sum_{i=1}^{n}\sum_{k=1}^{K}\left(W_{ik}^{-2}V_{\gamma^*,ikm}V_{\beta,ik} \;-\; W_{ik}^{-1}V_{\gamma^*\beta,ikm}\right)\,,$$

$$-\frac{\partial^2 l^*(\beta,\gamma^*)}{\partial\gamma_m^*\partial\gamma_m^*} \;=\; \sum_{i=1}^{n}\sum_{k=1}^{K}\left(W_{ik}^{-2}V_{\gamma^*,ikm}^2 \;-\; W_{ik}^{-1}V_{\gamma^*\gamma^*,ikm}\right)\,,$$

and for $j \neq m$,

$$-\frac{\partial^2 l^*(\beta,\gamma^*)}{\partial\gamma_m^*\partial\gamma_j^*} = \sum_{i=1}^{n}\sum_{k=1}^{K}\left(W_{ik}^{-2}V_{\gamma^*,ikm}V_{\gamma^*,ikj} - W_{ik}^{-1}V_{\gamma^*\gamma^*,ikmj}\right).$$

Appendix B: Expression of the Observed Fisher Information Matrix in Chapter 3

Let the W_{ik}'s, $V_{\beta,ik}$'s, $V_{\gamma,ikm}$'s, $I(\beta,\gamma)$, $I^*(\beta,\gamma^*)$ and their components be defined as in the previous sections. Define

$$V_{\beta\beta,ik} = \sum_{m=1}^{M} x_{im}x'_{im}\left(\alpha_{ikm+1} - \alpha_{ikm}\right)\exp(-\sum_{j=1}^{m} e^{\gamma_{kj}} - x'_{im}\beta)\,,$$

$$V_{\gamma\gamma,ikm} = \sum_{s=m}^{M}\left(\alpha_{iks+1} - \alpha_{iks}\right)\exp(2\gamma_{km} - \sum_{j=1}^{s} e^{\gamma_{kj}} - x'_{im}\beta)\,,$$

$$V_{\gamma\beta,ikm} = -\frac{\partial V_{\gamma,ikm}}{\partial\beta} = \sum_{s=m}^{M} x_{im}(\alpha_{iks+1} - \alpha_{iks})\exp(\gamma_{km} - \sum_{j=1}^{s} e^{\gamma_{kj}} - x'_{im}\beta)\,,$$

and

$$V_{\gamma\gamma,ikmj} = -\frac{\partial V_{\gamma,ikm}}{\partial\gamma_{kj}} = \sum_{s=m}^{M}(\alpha_{iks+1} - \alpha_{iks})\exp(\gamma_{km} + \gamma_{kj} - \sum_{j=1}^{s} e^{\gamma_{kj}} - x'_{ikm}\beta)\,.$$

Then we have

$$I_{\beta\beta}(\beta,\gamma) = \sum_{i=1}^{n}\sum_{k=1}^{K}\left(W_{ik}^{-2}V_{\beta,ik}V'_{\beta,ik} - W_{ik}^{-1}V_{\beta\beta,ik}\right)\,,$$

$$-\frac{\partial^2 l(\beta,\gamma)}{\partial\gamma_{km}\partial\beta} = \sum_{i=1}^{n}\left(W_{ik}^{-2}V_{\gamma,ikm}V_{\beta,ik} - W_{ik}^{-1}V_{\gamma\beta,ikm}\right)\,,$$

the mth column of $I_{\beta,\gamma_k}(\beta,\gamma)$, and $I_{\gamma_k,\gamma_j}(\beta,\gamma) = 0_{M\times M}$ for any $1 \le k \ne j \le K$.

Furthermore, for $I_{\gamma_k,\gamma_k}(\beta,\gamma)$, its (m,m)th diagonal element is given by

$$-\frac{\partial^2 l(\beta,\gamma)}{\partial\gamma_{km}\partial\gamma_{km}} = \sum_{i=1}^{n}\left(W_{ik}^{-2}V_{\gamma,ikm}^2 - W_{ik}^{-1}V_{\gamma\gamma,ikm} + W_{ik}^{-1}V_{\gamma,ikm}\right)$$

and for any $j < m$, its (m,j)th element has the form

$$-\frac{\partial^2 l(\beta,\gamma)}{\partial\gamma_{km}\partial\gamma_{kj}} = \sum_{i=1}^{n}\left(W_{ik}^{-2}V_{\gamma,ikm}V_{\gamma,ikj} + W_{ik}^{-1}V_{\gamma\gamma,ikmj}\right).$$

Now we consider the special situation where (3.4) is true. For this, define

$$W_{ik} = \alpha_{ik1} + \sum_{m=1}^{M}(\alpha_{ikm+1}-\alpha_{ikm})\exp(-\sum_{j=1}^{m}e^{\gamma_j^*} - x_{im}'\beta),$$

$$V_{\gamma^*,ikm} = -\frac{\partial W_{ik}}{\partial\gamma_m^*} = \sum_{s=m}^{M}(\alpha_{iks+1}-\alpha_{iks})\exp(\gamma_m^* - \sum_{j=1}^{s}e^{\gamma_j^*} - x_{is}'\beta),$$

$$V_{\beta,ik} = -\frac{\partial W_{ik}}{\partial\beta} = \sum_{m=1}^{M}x_{im}(\alpha_{ikm+1}-\alpha_{ikm})\exp(-\sum_{j=1}^{m}e^{\gamma_j^*} - x_{im}'\beta),$$

$$V_{\beta\beta,ik} = -\frac{\partial V_{\beta,ik}}{\partial\beta} = \sum_{m=1}^{M}x_{im}x_{im}'(\alpha_{ikm+1}-\alpha_{ikm})\exp(-\sum_{j=1}^{m}e^{\gamma_j^*} - x_{im}'\beta),$$

$$V_{\gamma^*\beta,ikm} = -\frac{\partial V_{\gamma^*,ikm}}{\partial\beta} = \sum_{s=m}^{M}x_{im}(\alpha_{iks+1}-\alpha_{iks})\exp(\gamma_m^* - \sum_{j=1}^{s}e^{\gamma_j^*} - x_{im}'\beta),$$

$$V_{\gamma^*\gamma^*,ikm} = \sum_{s=m}^{M}(\alpha_{iks+1}-\alpha_{iks})\exp(2\gamma_m^* - \sum_{j=1}^{s}e^{\gamma_j^*} - x_{im}'\beta),$$

and for $j < m$,

$$V_{\gamma^*\gamma^*,ikmj} = -\frac{\partial V_{\gamma^*,ikm}}{\partial \gamma_j^*} = \sum_{s=m}^{M}(\alpha_{iks+1} - \alpha_{iks}) \exp(\gamma_m^* + \gamma_j^* - \sum_{j=1}^{s} e^{\gamma_j^*} - x_{im}'\beta)\,.$$

Using the notation defined above, we have

$$U_\beta^*(\beta,\gamma^*) = -\sum_{k=1}^{K}\sum_{i=1}^{n} W_{ik}^{-1} V_{\beta,ik} \;\;,\;\; U_{\gamma^*}^*(\beta,\gamma^*) = -\sum_{k=1}^{K}\sum_{i=1}^{n} W_{ik}^{-1}(V_{\gamma^*,ik1},...,V_{\gamma^*,ikM})'\,,$$

$$D_i^* = -\sum_{k=1}^{K} W_{ik}^{-1}(V_{\beta,ik}', V_{\gamma^*,ik1},...,V_{\gamma^*,ikM})'\,,$$

$$\frac{\partial^2 l^*(\beta,\gamma^*)}{\partial\beta\partial\beta'} = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(W_{ik}^{-2}V_{\beta,ik}V_{\beta,ik}' - W_{ik}^{-1}V_{\beta\beta,ik}\right),$$

$$-\frac{\partial^2 l^*(\beta,\gamma^*)}{\partial\gamma_m^*\partial\beta} = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(W_{ik}^{-2}V_{\gamma^*,ikm}V_{\beta,ik} - W_{ik}^{-1}V_{\gamma^*\beta,ikm}\right),$$

$$-\frac{\partial^2 l^*(\beta,\gamma^*)}{\partial\gamma_m^*\partial\gamma_m^*} = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(W_{ik}^{-2}V_{\gamma^*,ikm}^2 - W_{ik}^{-1}V_{\gamma^*\gamma^*,ikm} + W_{ik}^{-1}V_{\gamma^*,ikm}\right),$$

and

$$-\frac{\partial^2 l^*(\beta,\gamma^*)}{\partial\gamma_m^*\partial\gamma_j^*} = \sum_{k=1}^{K}\sum_{i=1}^{n}\left(W_{ik}^{-2}V_{\gamma^*,ikm}V_{\gamma^*,ikj} + W_{ik}^{-1}V_{\gamma^*\gamma^*,ikmj}\right)$$

for $j < m$.

Appendix C: Variance Estimation in Chapter 4

As for the variance estimation of the estimator $\widehat{\theta}$, we adopt the inverse of the observed information matrix $I(\widehat{\theta})$, which is calculated by

$$I(\widehat{\theta}) = -E\Big(\frac{\partial^2 l(\theta; O, b)}{\partial\theta\partial\theta'}|O, \widehat{\theta}\Big).$$

Recall that $l(\theta; O, b)$ is the logarithm of the likelihood for the complete data and the score equations $U(\beta, \gamma) = \partial l(\theta; O, b)/\partial\theta$. Define

$$\psi^{(2)}_{ikj}(b_i; \beta, \gamma_k) = \frac{\delta_{ik}\exp[-\sum_{a=1}^{j}\exp(\gamma_{ka})\exp(x'_{ik}\beta + b_{ik})]}{\Big[1 - \exp[-\sum_{a=1}^{j}\exp(\gamma_{ka})\exp(x'_{ik}\beta + b_{ik})]\Big]^2} e^{2x'_{ik}\beta + 2b_{ik}}$$

Then the second derivatives of $l(\theta; O, b)$ are

$$\begin{aligned}
-\frac{\partial^2 l(\theta; O, b)}{\partial\beta\partial\beta'} &= \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{J}[\sum_{a=1}^{j}\exp(\gamma_{ka})]I_j(t_i)[\psi^{(2)}_{ikj}(b_i; \beta, \gamma_k)\sum_{a=1}^{j}\exp(\gamma_{ka}) - \psi^{(1)}_{ikj}(b_i; \beta, \gamma_k)]x_{ik}x'_{ik}, \\
-\frac{\partial^2 l(\theta; O, b)}{\partial\gamma_{kj}\partial\beta} &= \sum_{i=1}^{n}\sum_{h=j}^{J} I_h(t_i)\exp(\gamma_{kj})[\psi^{(2)}_{ikh}(b_i; \beta, \gamma_k)\sum_{a=1}^{h}\exp(\gamma_{ka}) - \psi^{(1)}_{ikh}(b_i; \beta, \gamma_k)]x_{ik}, \\
-\frac{\partial^2 l(\theta; O, b)}{\partial\gamma_{kj}\partial\gamma_{k'j'}} &= \begin{cases} 0 & \text{if } k \neq k', \\ \sum_{i=1}^{n}\sum_{h=j}^{J} I_h(t_i)\exp(\gamma_{kj} + \gamma_{kj'})\psi^{(2)}_{ikh}(b_i; \beta, \gamma_k) & \text{if } k = k',\ j' < j, \\ \sum_{i=1}^{n}\sum_{h=j}^{J} I_h(t_i)e^{\gamma_{kj}}[\psi^{(2)}_{ikh}(b_i; \beta, \gamma_k)e^{\gamma_{kj}} - \psi^{(1)}_{ikh}(b_i; \beta, \gamma_k))] & \text{if } k = k',\ j' = j. \end{cases}
\end{aligned}$$

Then the variance of $\widehat{\theta}$ is the inverse of the expectation $I(\widehat{\theta})$, which can be approximated by (5).

BIBLIOGRAPHY

Andersen, P. K. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. New York: John Wiley.

Bennett S. (1983). Log-logistic regression models for survival data. *Applied Statistics*, **32**, 165-171.

Betensky R. A. and Finkelstein D. M. (1999). A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, **55**, 940-943.

Bogaerts, K., Leroy, R. Lesaffre, E. and Declerck, D. (2002). Modelling tooth emergence data based on multivariate interval-censored data. *Statistics in Medicine*, **21**, 3775-3787.

Cai, J. (1999). Hypothesis testing of hazard ration parameters in marginal models for multivariate failure time data. *Lifetime Data Analysis*, **5**, 39-53.

Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ration parameters based on correlated failure time data. *Biometrika*, **82**, 151-164.

Chen, H. Y. (2001). Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design. *Journal of the American Statistical Association*, **96**, 1446-1457.

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

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.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, **34**, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Cox, D. R. and Oakes (1984).*Analysis of Survival Data*. New York: Chapman and Hall.

Ding, A. A. and Wang, W. (2004). Testing independence for bivariate current status data. *Journal of the American Statistical Association*, **99**, 145-155.

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

Duffy, D. L., Martin, N. G. and Matthews, J. D. (1990). Appendectomy in Australian

twins. *The American Journal Human Genetics*, **47**, 590-592.

Fan, J., Hsu, L., and Prentice, R. L. (2000). Dependence estimation over a finite bivariate failure time region. *Lifetime Data Analysis*, **6**, 343-355.

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

Finkelstein, D. M., Goggins, W. B. and Schoenfeld, D. A. (2002). Analysis of failure time data with dependent interval censoring. *Biometrics*, **58**, 298-304.

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, **41**, 933-945.

Gentleman, R. and Vandal, A. C. (2002). Nonparametric estimation of the bivariate CDF for arbitrarily censored data. *The Canadian Journal of Statistics*, **30**, 557-572.

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

Glidden, D. V. (2000). A two-stage estimator of the dependence parameter for the Clayton-Oakes model. *Lifetime Data Analysis*, **6**, 141-156.

Gumbel, E. J. Bivariate exponential distribution. (1960). *Journal of the American Statistical Association*, **55**, 698-707.

Guo, S.W. and Lin, D. Y. (1994). Regression analysis of multivariate grouped survival data. *Biometrics*, **50**, 632-639.

He, W. and Lawless, J. F. (2003). Flexible maximum likelihood methods for bivariate proportional hazards models. *Biometrics*, **59**, 837-848.

Hoel, D. G. and Walberg, H. E. (1972). Statistical analysis of survival experiments. *Journal of National Cancer Institute*, **49**, 361-372.

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

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

Huang, X. and Wolf, R. A. (2002). A frailty model for informative censoring. *Biometrics*, **58**, 510-520.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267-278.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Second edition, New York: John Wiley.

Kim, J. and Lee, S. (1998). Two-sample goodness-of-fit tests for additive risk models with censored observations. *Biometrika*, **85**, 593-603.

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

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis*, New York: Springer-

Verlag.

Kulich, M. and Lin, D. Y. (2000). Additive hazards regression with covariate measurement error. *Journal of the American Statistical Association*, **95**, 238-248.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. New York: John Wiley.

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

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

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61-71.

Li, Q. H. and Lagakos, S. W. (2004). Comparisons of test statistics arising from marginal analyses of multivariate survival data. *Lifetime Data Analysis*, **10**, 389-405.

Martinussen, T. and Scheike, T. H. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika*, **89**, 649-658.

Murphy, S. A., Rossini, A. J. and Van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, **92**, 968-976.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American*

Statistical Association, **84**, 487-493.

Peto, R. (1973). Experimental Survival Curves for Interval-Censored Data. *Applied Statistics*, **22**, 86 -91.

Pierce, D. A., Stewart, W. H. and Kopecky, K. J. (1979). Distribution free regression analysis of grouped survival data. *Biometrics*, **35**, 785-793.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, **34**, 57-67.

Prentice, R. L. and Hsu, L. (1997). Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika*, **84**, 349-363.

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

Robertson, T., Wright, F. T., and Dykstra, R. (1988). *Order Restricted Statistical Inference*. New York: John Wiley.

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

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, **51**, 1384-1399.

Spiekerman, C. F. and Lin, D. Y. (1996). Checking the marginal Cox model for correlated failure time data. *Biometrika*, **83**, 143-156.

Sun, J. (2005). Interval Censoring. *Encyclopedia of Biostatistics*, Second edition, New York: John Wiley, 2603-2609.

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

Sun, L., Wang, L. and Sun, J. (2006). Estimation of the association for bivariate interval-censored failure time data. *Scandinavian Journal of Statistics*, **33**, 637-649.

Turnbull, B. W. (1976). The Empirical Distribution Function from Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society, Series B*, **38**, 290 -295.

Turnbull, B. W. and Weiss, L. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, **34**, 367-375.

Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data. *Biometrika*, **87**, 879-893.

Wei, L. J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distribution. *Journal of the American Statistical Association*, **84**, 1065-1073.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*,

50, 1-25.

Yang, S. and Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association*, **94**, 125-136.

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

Zhang, Z. G., Sun, J. and Sun, J. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine*, **24**, 1399-1407.

Table 1.1: Ages in years to the first use of marijuana

Age	No. of exact obs.	No. of left-censored obs.	No. of right-censored obs.
10	4	0	0
11	12	0	0
12	19	2	0
13	24	15	1
14	20	24	2
15	13	18	3
16	3	14	2
17	1	6	3
18	0	0	1
> 18	4	0	0

Table 1.2: Death times in days for 144 male RFM mice with lung tumors

Group	Tumor status	Death times
CE	With tumor	381, 477, 485, 515, 539, 563, 565, 582, 603, 616
		624, 650, 651, 656, 659, 672, 679, 698, 702, 709
		723, 731, 775, 779, 795, 811, 839
	No tumor	45, 198, 215, 217, 257, 262, 266, 371, 431, 447
		454, 459, 475, 479, 484, 500, 502, 503, 505, 508
		516, 531, 541, 553, 556, 570, 572, 575, 577, 585
		588, 594, 600, 601, 608, 614, 616, 632, 632, 638
		642, 642, 642, 644, 644, 647, 647, 653, 659, 660
		662, 663, 667, 667, 673, 673, 677, 689, 693, 718
		720, 721, 728, 760, 762, 773, 777, 815, 886
GE	With tumor	546, 609, 692, 692, 710, 752, 773, 781, 782, 789
		808, 810, 814, 842, 846, 851, 871, 873, 876, 888
		888, 890, 894, 896, 911, 913, 914, 914, 916, 921
		921, 926, 936, 945, 1008
	No tumor	412, 524, 647, 648, 695, 785, 814, 817, 851, 880
		913, 942, 986

Table 1.3: Observed intervals in months for times to breast retraction of early breast cancer patients

Group	Observed intervals in months
RT	(45,], (25,37], (37,], (4,11], (17,25], (6,10], (46,], (0,5], (33,], (15,]
	(0,7], (26,40], (18,], (46,], (19,26], (46,], (46,], (24,], (11,15], (11,18]
	(46,], (27,34], (36,], (37,], (22,], (7,16], (36,44], (5,12], (38,], (34,]
	(17,], (46,], (19,35], (46,], (5,12], (9,14], (36,48], (17,25], (36,], (46,]
	(37,44], (37,], (24,], (0,8], (40,], (33,]
RCT	(8,12], (0,5], (30,34], (16,20], (13,], (0,22], (5,8], (13,], (30,36], (18,25]
	(24,31], (12,20], (10,17], (17,24], (18,24], (17,27], (11,], (8,21], (17,26]
	(35,], (17,23], (33,40], (4,9], (16,60], (33,], (24,30], (31,], (11,], (15,22]
	(35,39], (16,24], (13,39], (15,19], (23,], (11,17], (13,], (19,32], (4,8], (22,]
	(44,48], (11,13], (34,], (34,], (22,32], (11,20], (14,17], (10,35], (48,]

Table 1.4: NTP study: the occurance of adrenal and lung tumors by the time of death for 100 male rats

Age at death[a] (months)	Control group 0 ppm				High dose group 80 ppm			
11	0	0	0	0[b]	2	0	0	0
16	1	0	0	0	2	0	0	0
17	1	0	0	0	1	0	0	0
18	4	0	0	0	5	0	0	1
19	3	0	0	0	3	0	0	0
20	4	2	0	0	4	0	0	0
21	2	2	1	0	7	3	0	1
22	5	3	1	0	5	0	0	1
23	0	0	0	0	2	2	1	0
24	3	4	0	0	0	5	0	0
25	1	0	0	0	0	1	0	0
25[c]	5	8	0	0	0	2	0	2
Total	29	19	2	0	31	13	1	5

[a] Day of death data are grouped into month.

[b] Number of rats with no tumors, only adrenal, only lung, and both tumors, respectively.

[c] Animals sacrificed at the end of study.

Table 2.1: Estimates of the regression parameter for $n = 100$ and $\beta_0 = 0$

Correlation ρ	Censoring percentage q	AVE	SEE	SSE	CP
0.1	0.1	-0.00061	0.25839	0.26176	0.950
	0.3	-0.01941	0.25622	0.25588	0.952
	0.5	-0.00622	0.25313	0.28590	0.927
	0.7	0.00819	0.27218	0.29224	0.924
0.3	0.1	-0.00692	0.26562	0.26395	0.952
	0.3	0.00661	0.26465	0.26505	0.955
	0.5	0.00011	0.25949	0.27917	0.928
	0.7	-0.00454	0.27766	0.30548	0.912
0.5	0.1	0.00540	0.27340	0.27970	0.953
	0.3	0.00709	0.27224	0.27359	0.954
	0.5	0.00497	0.26723	0.28643	0.937
	0.7	-0.01248	0.28110	0.30336	0.917
0.7	0.1	0.00622	0.28113	0.28892	0.946
	0.3	-0.00377	0.28114	0.28321	0.954
	0.5	0.01855	0.27440	0.29159	0.930
	0.7	0.02063	0.28986	0.32677	0.891

Table 2.2: Estimates of the regression parameter for $n = 100$ and $\beta_0 = 0.5$

Correlation	Censoring percentage				
ρ	q	AVE	SEE	SSE	CP
0.1	0.1	0.50887	0.29209	0.25555	0.962
	0.3	0.50839	0.28698	0.26993	0.945
	0.5	0.51280	0.27511	0.27417	0.929
	0.7	0.48494	0.28561	0.29988	0.915
0.3	0.1	0.49474	0.29947	0.26108	0.966
	0.3	0.50779	0.29799	0.28069	0.939
	0.5	0.49452	0.28104	0.28214	0.949
	0.7	0.53961	0.28592	0.30643	0.913
0.5	0.1	0.49784	0.31554	0.26959	0.956
	0.3	0.52168	0.30851	0.28305	0.948
	0.5	0.50341	0.28973	0.28727	0.939
	0.7	0.51636	0.28839	0.30730	0.918
0.7	0.1	0.51103	0.31693	0.27232	0.964
	0.3	0.50565	0.31482	0.28621	0.954
	0.5	0.51253	0.29603	0.28363	0.942
	0.7	0.51052	0.29255	0.31811	0.918

Table 2.3: Estimates of the regression parameter for $n = 200$ and $\beta_0 = 0$

Correlation	Censoring percentage				
ρ	q	AVE	SEE	SSE	CP
0.1	0.1	0.00054	0.17969	0.17887	0.949
	0.3	0.00078	0.18115	0.17960	0.950
	0.5	0.00763	0.18445	0.18433	0.953
	0.7	0.00135	0.19462	0.20553	0.929
0.3	0.1	-0.00383	0.18537	0.19145	0.949
	0.3	0.00044	0.18637	0.17781	0.962
	0.5	0.00065	0.18977	0.19782	0.940
	0.7	0.00561	0.20148	0.20825	0.932
0.5	0.1	0.00430	0.19102	0.18412	0.959
	0.3	-0.00001	0.19261	0.19692	0.951
	0.5	0.00672	0.19535	0.19167	0.949
	0.7	-0.00281	0.20347	0.21499	0.932
0.7	0.1	-0.00005	0.19622	0.19846	0.947
	0.3	0.01670	0.19800	0.20389	0.945
	0.5	-0.00392	0.19899	0.20133	0.955
	0.7	0.00002	0.20755	0.21900	0.932

Table 2.4: Estimates of the regression parameter for $n = 200$ and $\beta_0 = 0.5$

Correlation	Censoring percentage				
ρ	q	AVE	SEE	SSE	CP
0.1	0.1	0.52061	0.19058	0.17605	0.961
	0.3	0.50090	0.19129	0.18483	0.946
	0.5	0.50619	0.19603	0.19368	0.944
	0.7	0.50626	0.20869	0.20383	0.934
0.3	0.1	0.50395	0.19692	0.19563	0.940
	0.3	0.50639	0.19787	0.19245	0.947
	0.5	0.51286	0.19991	0.19740	0.951
	0.7	0.51199	0.20751	0.21169	0.927
0.5	0.1	0.50847	0.20284	0.18907	0.960
	0.3	0.50934	0.20361	0.19358	0.958
	0.5	0.52033	0.20156	0.19869	0.937
	0.7	0.49255	0.20785	0.21637	0.938
0.7	0.1	0.50473	0.20917	0.19175	0.969
	0.3	0.50336	0.20861	0.19526	0.954
	0.5	0.49752	0.20836	0.20178	0.945
	0.7	0.50923	0.21322	0.21603	0.937

Table 3.1: Estimates of the regression parameter for $n = 100$ and $\beta_0 = 0$

Correlation	Censoring percentage				
ρ	q	AVE	SEE	SSE	CP
0.1	0.1	0.0013	0.0216	0.0224	0.944
	0.3	-0.0005	0.0244	0.0238	0.937
	0.5	0.0026	0.0254	0.0246	0.888
0.3	0.1	0.0001	0.0222	0.0239	0.952
	0.3	0.0011	0.0226	0.0251	0.961
	0.5	0.0035	0.0257	0.0258	0.889
0.5	0.1	0.0007	0.0235	0.0257	0.949
	0.3	-0.0013	0.0265	0.0265	0.943
	0.5	0.0016	0.0257	0.0270	0.895
0.7	0.1	-0.0002	0.0251	0.0255	0.937
	0.3	-0.0008	0.0253	0.0266	0.937
	0.5	0.0046	0.0270	0.0267	0.885

Table 3.2: Estimates of the regression parameter for $n = 100$ and $\beta_0 = 1$

Correlation	Censoring percentage				
ρ	q	AVE	SEE	SSE	CP
0.1	0.1	0.9930	0.1392	0.1417	0.948
	0.3	1.0347	0.1402	0.1432	0.950
	0.5	1.0804	0.1403	0.1488	0.943
0.3	0.1	1.0115	0.1373	0.1571	0.954
	0.3	0.9945	0.1392	0.1558	0.953
	0.5	1.0359	0.1602	0.1612	0.949
0.5	0.1	0.9866	0.1497	0.1675	0.947
	0.3	1.0041	0.1487	0.1738	0.953
	0.5	0.9930	0.1510	0.1733	0.950
0.7	0.1	0.9419	0.1412	0.1733	0.905
	0.3	0.9419	0.1404	0.1726	0.915
	0.5	0.9528	0.1527	0.1821	0.934

Table 3.3: Estimates of the regression parameter for $n = 200$ and $\beta_0 = 0$

Correlation	Censoring percentage				
ρ	q	AVE	SEE	SSE	CP
0.1	0.1	0.0005	0.0144	0.0148	0.957
	0.3	0.0002	0.0149	0.0154	0.954
	0.5	0.0009	0.0167	0.0163	0.932
0.3	0.1	0.0005	0.0155	0.0159	0.955
	0.3	0.0009	0.0168	0.0164	0.943
	0.5	-0.0001	0.0178	0.0172	0.924
0.5	0.1	-0.0002	0.0158	0.0166	0.956
	0.3	-0.0005	0.0168	0.0171	0.947
	0.5	0.0010	0.0199	0.0186	0.942
0.7	0.1	0.0003	0.0163	0.0167	0.948
	0.3	0.0009	0.0169	0.0173	0.951
	0.5	0.0006	0.0178	0.0178	0.932

Table 3.4: Estimates of the regression parameter for $n = 200$ and $\beta_0 = 1$

Correlation	Censoring percentage				
ρ	q	AVE	SEE	SSE	CP
0.1	0.1	1.0113	0.0937	0.0998	0.963
	0.3	0.9980	0.0944	0.1002	0.955
	0.5	0.9629	0.0938	0.1022	0.938
0.3	0.1	1.0089	0.1004	0.1089	0.962
	0.3	1.0063	0.0994	0.1092	0.957
	0.5	1.0236	0.1095	0.1117	0.949
0.5	0.1	0.9847	0.1065	0.1156	0.948
	0.3	1.0076	0.1017	0.1179	0.948
	0.5	1.0060	0.1124	0.1203	0.952
0.7	0.1	0.9416	0.1008	0.1138	0.896
	0.3	0.9440	0.1016	0.1183	0.907
	0.5	0.9475	0.1166	0.1241	0.903

Table 3.5: Estimates of the regression parameter for $\rho = 0.7$ and $\beta_0 = 0.5$

Sample size	Censoring percentage				
n	q	AVE	SEE	SSE	CP
100	0.1	0.4797	0.0817	0.1028	0.954
	0.3	0.4774	0.0898	0.1059	0.930
	0.5	0.5009	0.0901	0.1053	0.944
200	0.1	0.5080	0.0571	0.0699	0.948
	0.3	0.5082	0.0615	0.0701	0.941
	0.5	0.4889	0.0610	0.0717	0.963

Table 4.1: Estimates of the regression parameter for $n = 100$

β_0	J	AVE	SEE	SSE	CP
0	10	0.0472	0.2550	0.2514	0.954
	15	0.0403	0.2673	0.2583	0.944
	20	0.0466	0.2710	0.2623	0.943
1	10	1.1255	0.3126	0.2982	0.946
	15	1.0468	0.3393	0.3059	0.933
	20	1.1423	0.3051	0.3069	0.948
-1	10	-1.0022	0.3118	0.3094	0.960
	15	-0.9728	0.3162	0.3141	0.953
	20	-0.9671	0.3149	0.3216	0.959

Table 4.2: Estimates of the regression parameter for $n = 200$

β_0	J	AVE	SEE	SSE	CP
0	10	0.0233	0.1723	0.1742	0.957
	15	0.0167	0.1807	0.1759	0.952
	20	0.0139	0.1801	0.1804	0.952
1	10	1.0779	0.1932	0.1978	0.945
	15	1.0697	0.2005	0.1985	0.944
	20	1.0845	0.2050	0.2017	0.943
-1	10	-0.9727	0.2074	0.2098	0.941
	15	-0.9996	0.2271	0.2206	0.955
	20	-0.9820	0.2246	0.2225	0.940

Table 4.3: Estimates of the regression parameter for $\beta_0 = 0.5$ or -0.5

n	β_0	J	AVE	SEE	SSE	CP
100	0.5	5	0.5337	0.2801	0.2701	0.954
		10	0.5517	0.2793	0.2396	0.950
	-0.5	5	-0.5076	0.2761	0.2565	0.942
		10	-0.5443	0.2897	0.2816	0.960
200	0.5	5	0.5192	0.1957	0.1888	0.945
		10	0.5152	0.1919	0.1898	0.948
	-0.5	5	-0.4833	0.1938	0.1789	0.934
		10	-0.5021	0.1921	0.1906	0.951

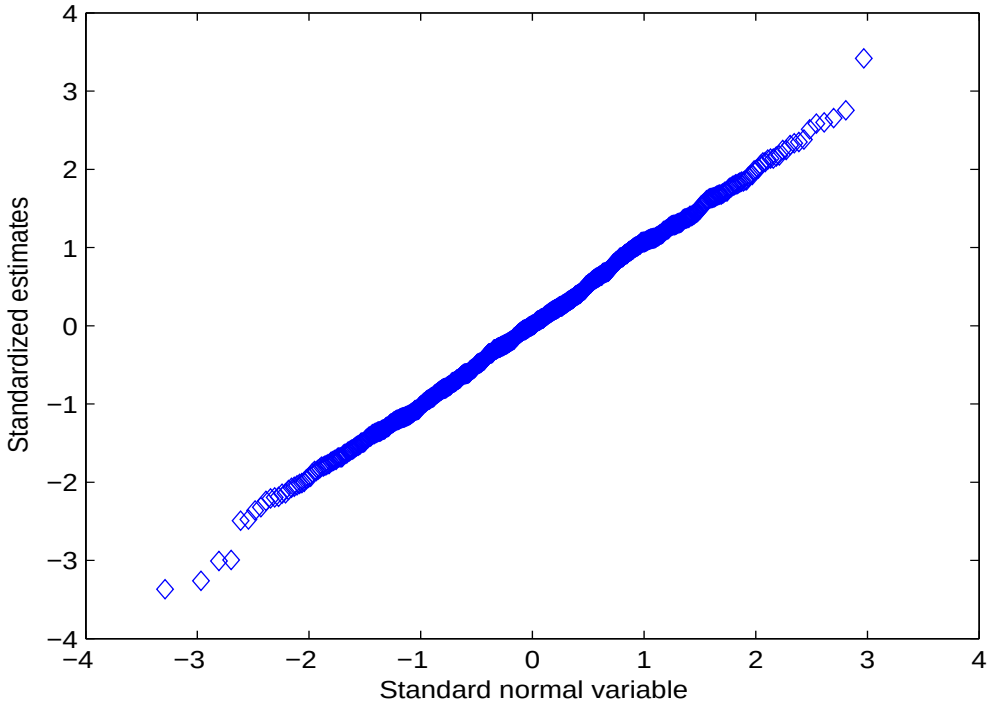

Figure 2.1: Quantile plot of the standardized parameter estimates for n=100

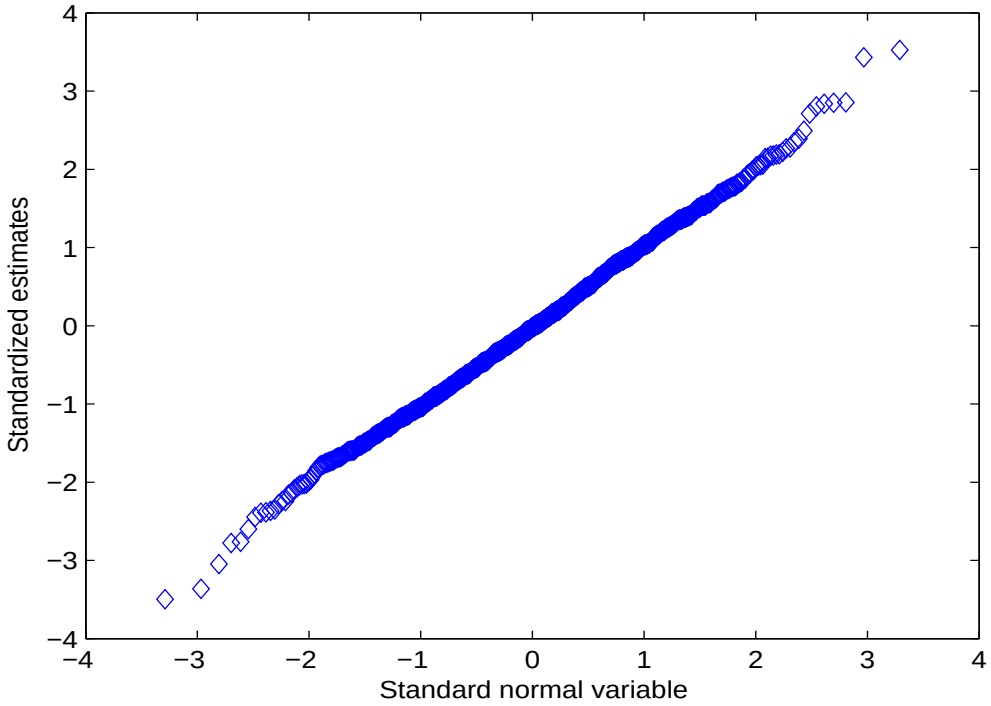

Figure 2.2: Quantile plot of the standardized parameter estimates for n=200

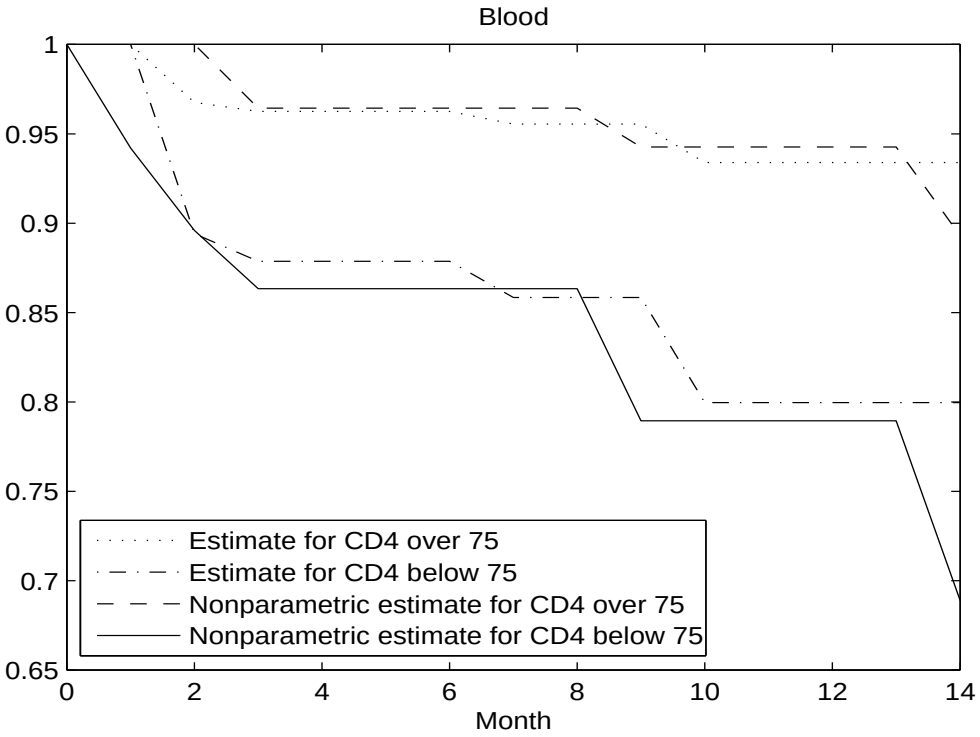

Figure 2.3: Estimated of the marginal survival function for time to CMV shedding in blood

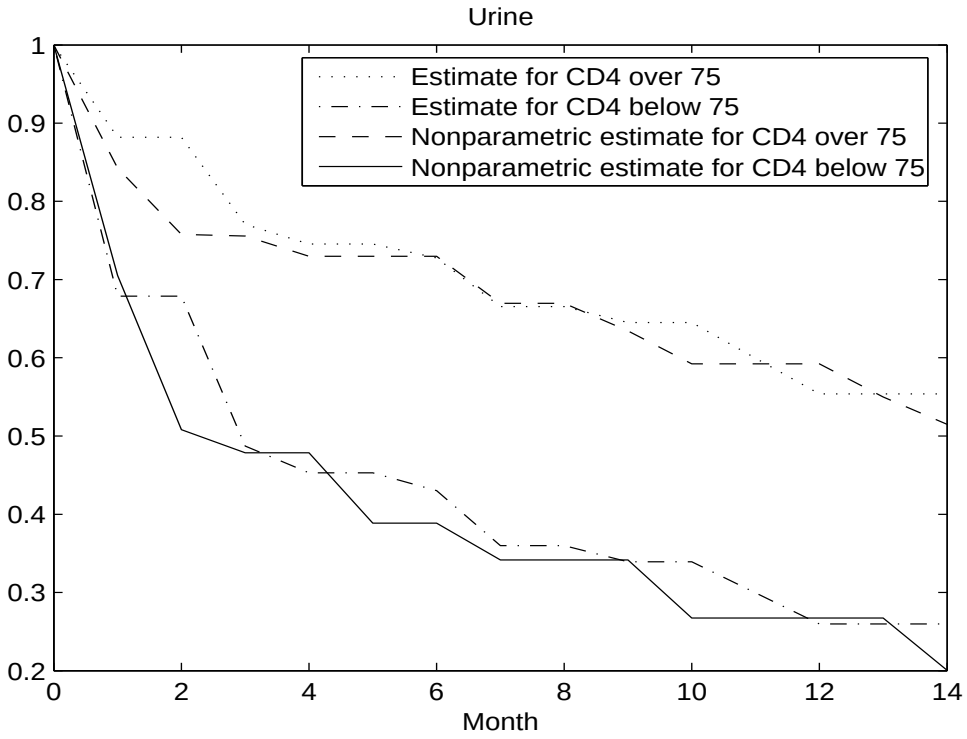

Figure 2.4: Estimated of the marginal survival function for time to CMV shedding in urine

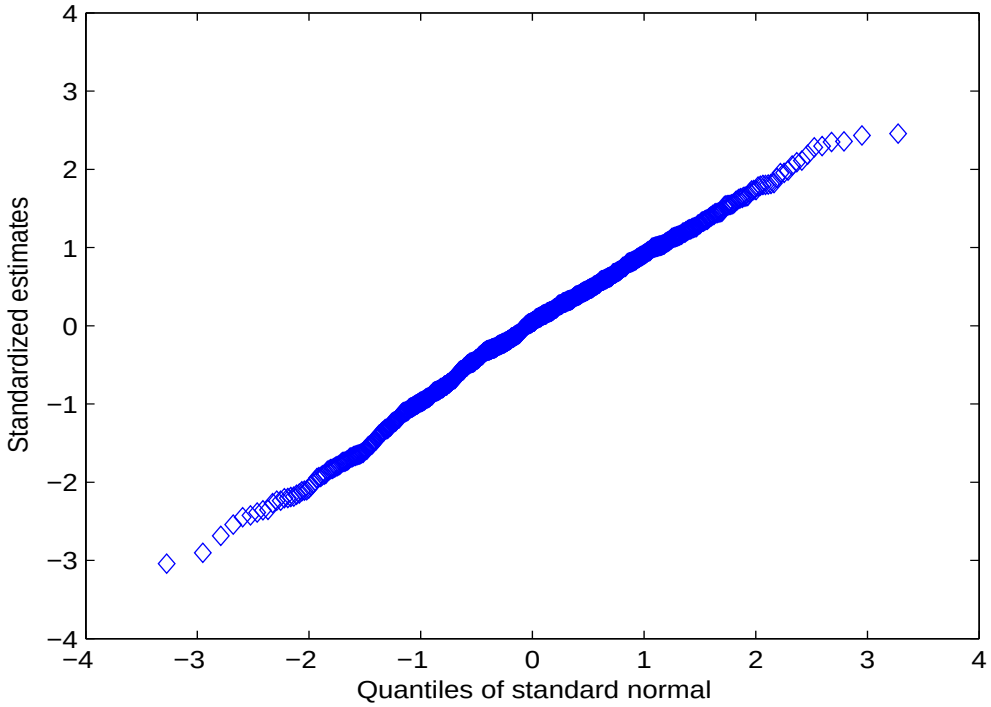

Figure 3.1: Quantile plot of the standardized parameter estimates for n=100

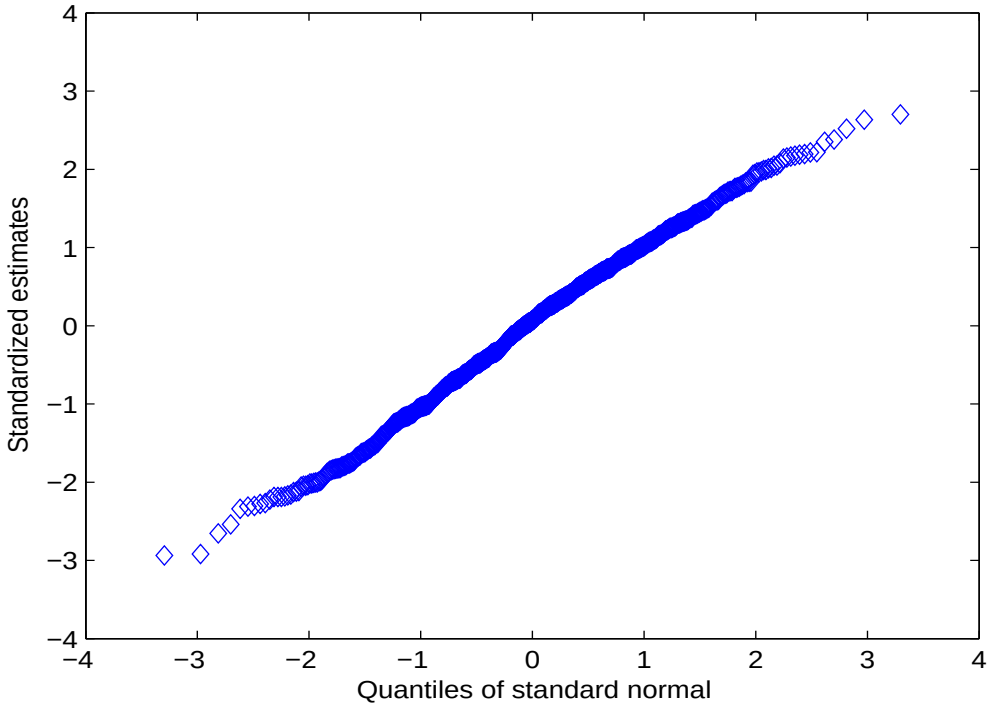

Figure 3.2: Quantile plot of the standardized parameter estimates for n=200

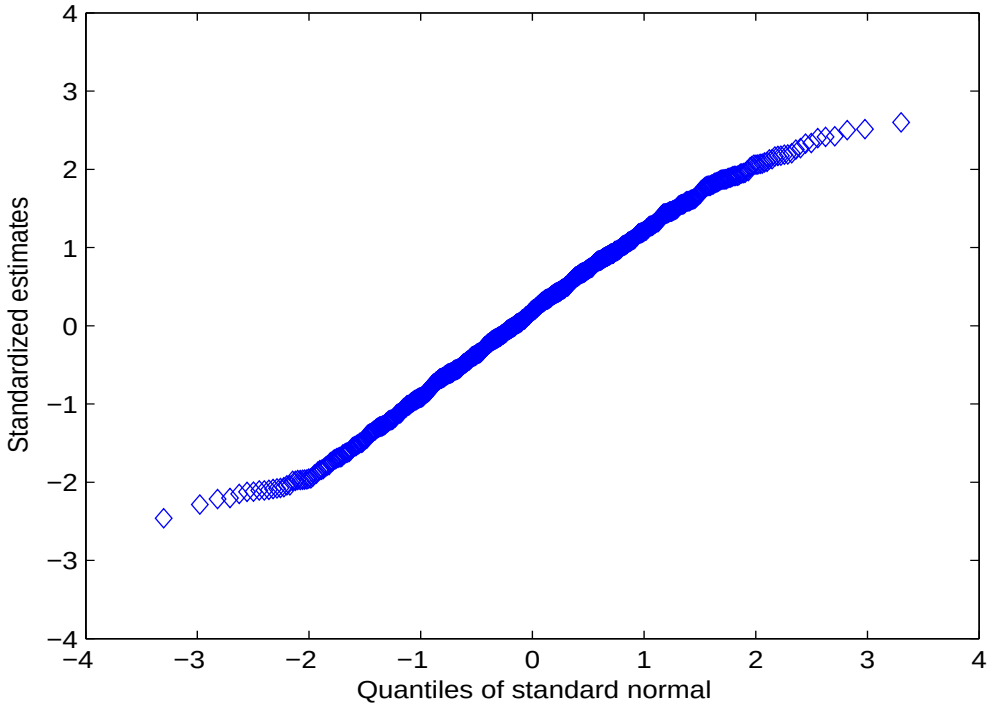

Figure 4.1: Quantile plot of the standardized parameter estimates for n=100

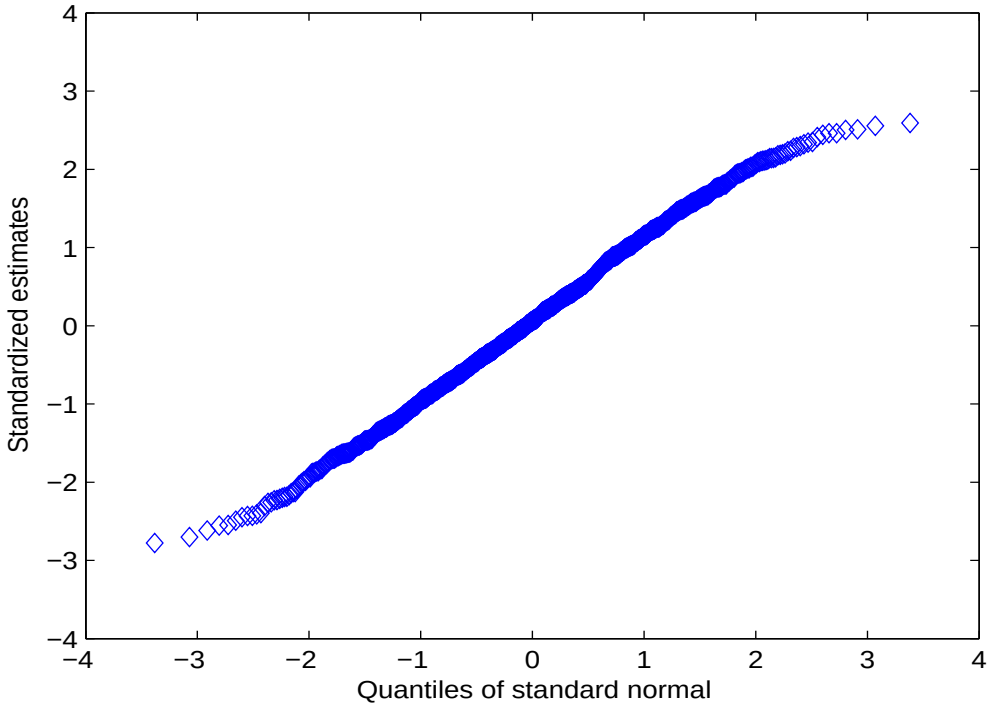

Figure 4.2: Quantile plot of the standardized parameter estimates for n=200

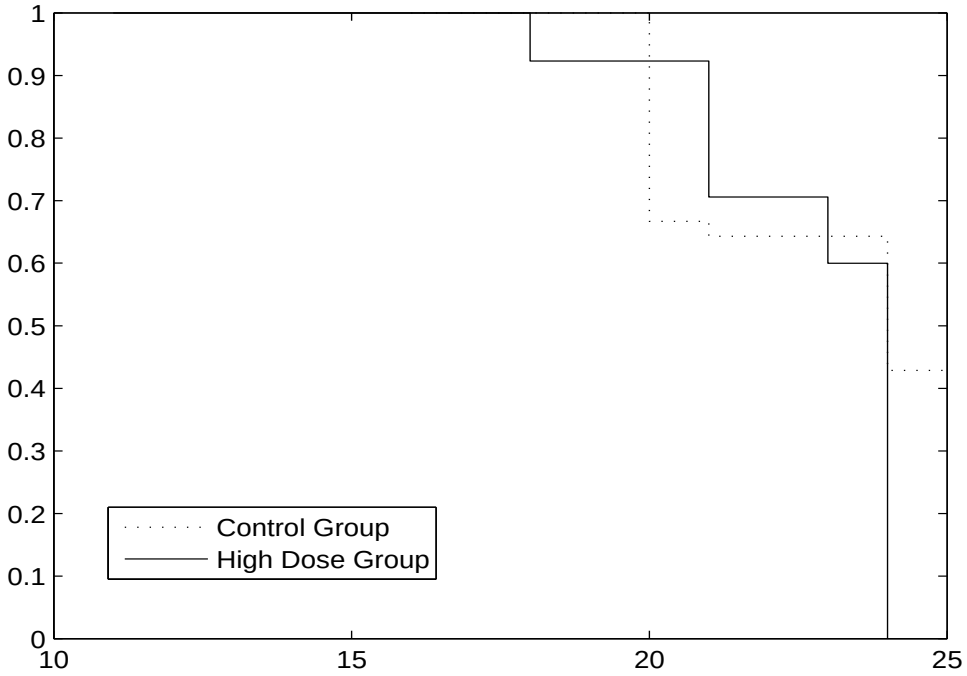

Figure 4.3: Estimated marginal survival functions for adrenal tumor

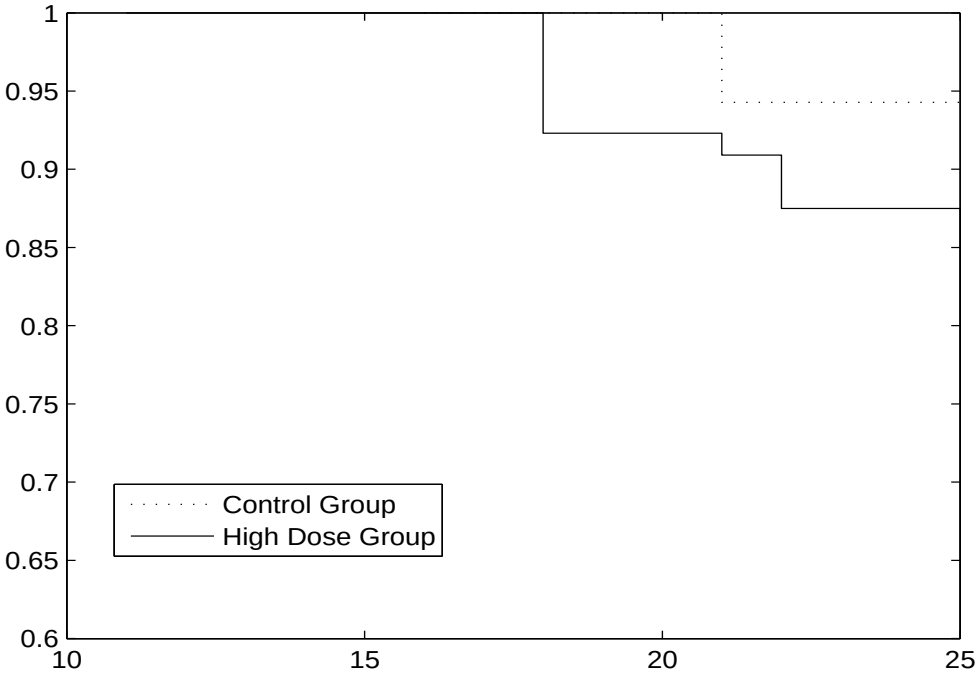

Figure 4.4: Estimated marginal survival functions for lung tumor

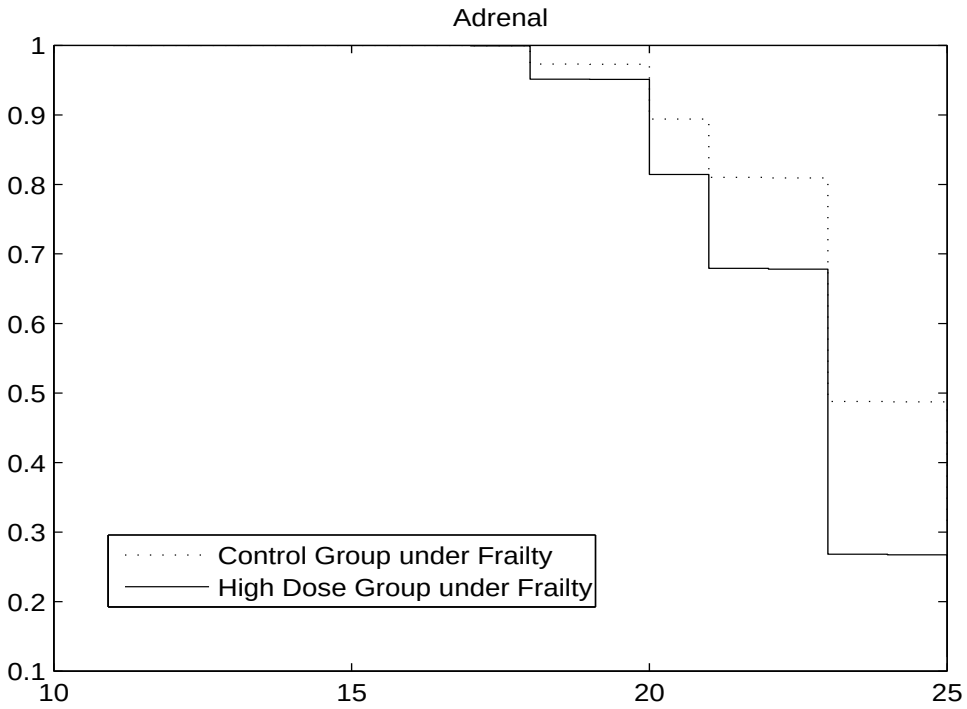

Figure 4.5: Estimated marginal survival functions for time to adrenal tumor under frailty model

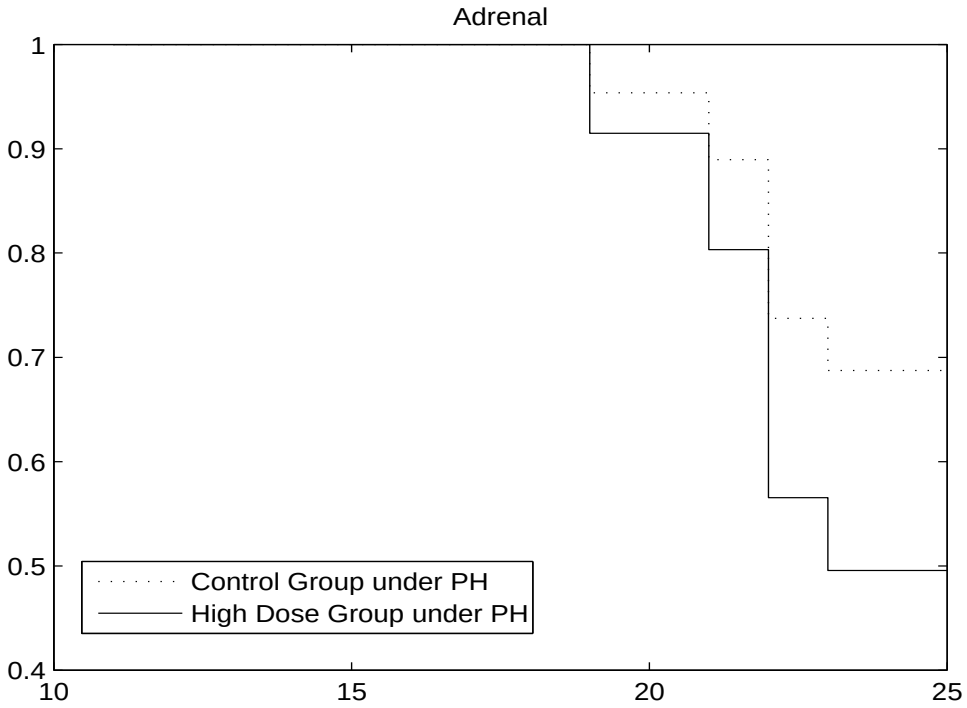

Figure 4.6: Estimated marginal survival functions for time to adrenal tumor under PH model

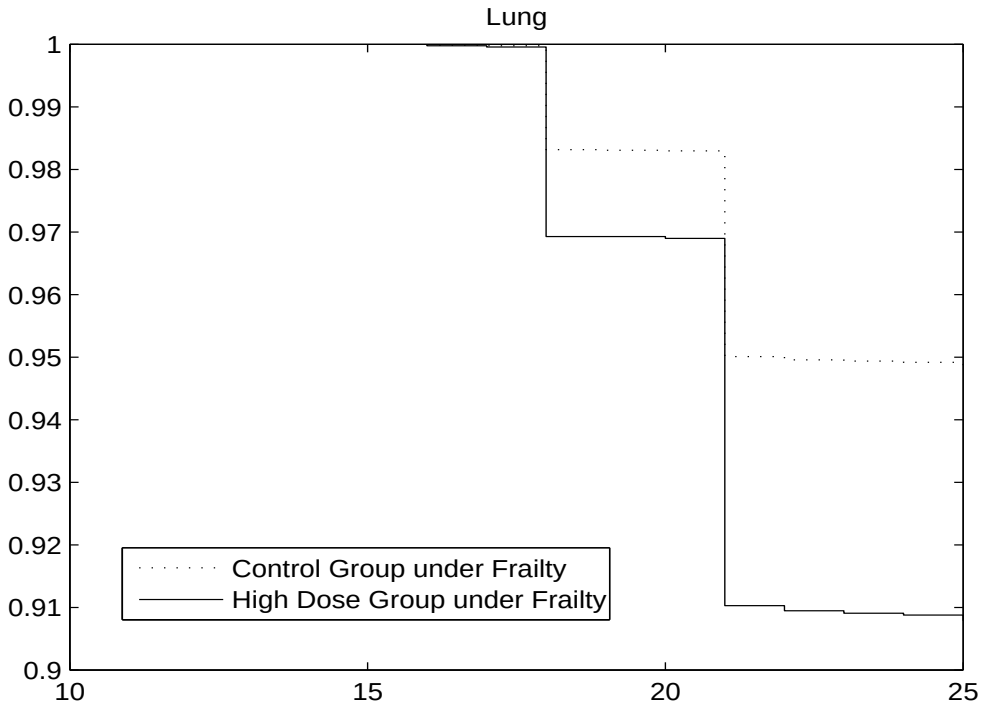

Figure 4.7: Estimated marginal survival functions for time to lung tumor under frailty model

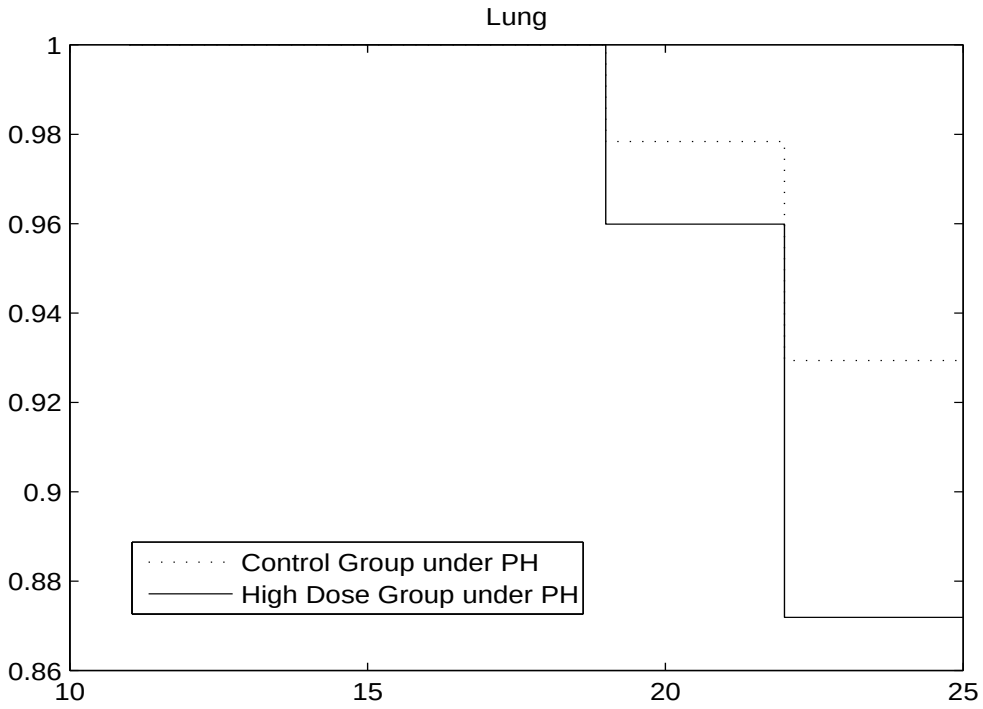

Figure 4.8: Estimated marginal survival functions for time to lung tumor under PH model

VITA

Man-Hua Chen was born on December 15, 1973, in Taipei, Taiwan. She received her B.A. in Statistics from Feng Chia University in Taichung in 1997 and M.B.A. in Statistics from Tamkang University in Tamsui in 2000. In the same year, she served as a teaching assistant in the Department of Statistics at the Feng Chia Universtiy. She joined the graduate program in the Department of Statistics at the University of Missouri-Columbia in January 2003. She will receive her Ph.D. in Statistics in December 2007. As of January 2008, she will be serving as a Postdoctoral Fellow at the Institute of Statistical Science of Academia Sinica in Taipei, Taiwan.