Proportional Hazards Models for Survival Data with Long-term Survivors

by

XIAOBING ZHAO

A thesis submitted in partial fulfillment of
the requirements for the Degree of Doctor of Philosophy
in
Department of Applied Mathematics
The Hong Kong Polytechnic University

Supervisor: Dr. XIAN ZHOU

September 2005

CERTIFICATE OF ORIGINALITY

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institute of higher learning, except where due acknowledgment is made in the text.

XIAOBING ZHAO

ABSTRACT

Survival analysis with long-term survivors has received considerable attention in recent years. It is useful in handling situations in which a proportion of subjects under study may never experience the event of interest. A commonly used approach is to formulate the model as a mixture of two populations, one for "long-term survivors" (subjects that will never experience the event of interest) and the other for "susceptibles" (subjects that will "fail" eventually). This is an attractive approach to analyzing survival data with long-term survivors, in that it contains two parts which can be interpreted separately by adding structure to the standard survival model. Fully parametric approaches have had a long history and recent attention has focused on test for the presence of long-term survivors in the data for mixture model. Recently, Kuk and Chen (1992) proposed a model which is a semiparametric generalization of a parametric model above, which combines a logistic formulation for the probability of occurrence of the event with a proportional hazards specification for the time of occurrence of the event. However the model proposed by Kuk and Chen (1992) *does not* have a proportional hazards structure for the survival function of the entire population, this structure is a desirable property in survival analysis models when doing covariates and is extensively used in survival analysis.

In this dissertation, we will investigate an alternative mixture model with covariates, which *does* have a proportional hazards structure and is proposed by Maller and Zhou (1996) via the different motivation of the model from other cure models and is not further investigated so far. Their idea is to extend Cox model with a parametric or completely unspecified baseline to "improper" Cox model of which the baseline can be modeled as an *improper* and *semiparametric* structure of a combination of the probability of occurrence of the event with a proper survival function for the time of occurrence of the event. Partial and full likelihood methods are used to make statistical inference based on counting processes and martingale technique. In addition, we consider the problem of measurement errors for the covariates, and propose corrected partial and full likelihoods to obtain relevant estimators. We show that the resulting estimators are consistent and asymptotically normal. We study this improper proportional hazards model in both interior and boundary cases by maximum likelihood method, and develop a likelihood ratio test

for the presence of an immune proportion in a population.

Recently, much attention has been attracted to semiparametric transformation model, which provides many interesting statistical models and approaches. In this dissertation, we consider a class of semiparametric transformation models derived from the aforesaid "improper" PH model, which assume a linear relationship between an unknown transformation of the survival time under the proportional hazards model and the covariates. The random errors are modeled by an "improper" extreme value distribution, which is parametrically specified with unknown parameters and covariates. Estimators for the coefficients of covariates are obtained from pseudo Z-estimator approach with censored observations. The consistency and asymptotic normality of the estimators are proved. This transformation model, coupled with proposed inference procedures, provides many alternatives to Cox proportional hazards models for survival analysis with long-term survivors.

Although continuous-time survival analysis is frequently used in many settings, discrete-time survival analysis is often more natural in social and behavioral science applications where the survival data typically posses three features: discrete, ties and contain concomitant information. Inspired by the works of Potts (2004) and Linoff (2004), in this dissertation we will review some existing discrete-time survival models which have already been proposed to analyze survival data from social and behavioral science by these authors, and then generalize these models to accommodate survival data with long-term survivors. As a natural extension of the continuous cases discussed in Chapters 1-4, in Chapter 5 of this dissertation we will predominantly focus on modeling discrete-time survival data which may accommodate proportional hazards structure and propose an alternative discrete-time cure model which does have proportional hazards structure. The maximum likelihood estimation and approximate partial likelihood estimation are proposed to obtain the parameter estimators. The proposed models and approaches can be directly applied to analyze survival data from social and behavioral science such as the economic values for customer retention with long-life customers.

Table of Contents

Acknowledgments

The author wishes to express deep gratitude to his supervisor, Dr. Xian Zhou, for his thoughtful and insightful directions through years. Without his continuous encouragement and support, this dissertation would not be possible.

The author also thanks the staff of the Department of Applied Mathematics of the Hong Kong Polytechnic University for their kind support.

Additionally, the author would like to thank Professor Jinglong Wang of East China Normal University for his tremendous encouragement and support. The author is also grateful to Dr. Xianyi Wu for his helpful suggestions.

Finally, special thanks to my family for their continuing and steady support throughout my studies.

Preface

In this dissertation, we consider a proportional hazards model for survival data with long-term survivors. We begin by introducing the main idea of Cox's proportional hazards model in this preface, which inspired and motivated this research. Next we will briefly explain in words how we extend Cox's proportional hazards model to a new model which may allow a "semiparametric" baseline function. Then the layout of this dissertation is outlined.

A broad family of survival analysis models that has been widely used in the analysis of survival data is best specified via the hazards function. For a constant vector z of explanatory variables (covariates), suppose that the hazards function has the form $\alpha(t) = \alpha_0(t)\Psi(z)$, where $\alpha_0(t)$ is the baseline hazards for an individual with $z = 0$, and $\Psi(0) = 1$. This is the well-known Cox's proportional hazards (PH) model (Cox, 1972, 1975). The function $\Psi(z)$ connects the model with covariates and can be parameterized, commonly by $\Psi(z) = \exp(z^\top \beta)$.

There have been a great deal of literature on the PH model with either time-independent or time-dependent covariates. Two types of methods have been widely used to estimate the PH model: the partial likelihood, which disregards the unknown baseline $\alpha_0(t)$, proposed by Cox (1972); and the full maximum likelihood, which assumes a parameterized baseline hazards function $\alpha_0(t)$, such as the exponential or Weibull hazards functions.

On the other hand, if we consider an equivalent form of the PH model, namely $F(t) = 1 - [1 - F_0(t)]^{\exp(z^\top \beta)}$, then it can be rewritten as $\log\{-\log[1 - F_0(t)]\} = -z^\top \beta + \varepsilon$ (Doksum, 1987), where ε has the extreme value distribution. This model, termed as the *semiparametric transformation model*, has also been extensively investigated by many authors.

This proportional hazards structure is a desirable property in survival analysis models when covariates are involved and is extensively used in survival analysis. To apply it to survival models with long-term survivors, Kuk and Chen (1992) proposed a model which is a semiparametric generalization of a parametric mixture model due to Farewell (1982), which combines a logistic formulation for the probability of occurrence of the event with a proportional hazards specification for the time of occurrence of the event. However, the model proposed by Kuk and Chen (1992) *does not* have a proportional hazards structure for the survival function of the entire population, even if the survival function for the time of occurrence of the event is taken to be proportional hazards structure.

Maller and Zhou (1996) considered both proportional and nonproportional hazards models, such as the exponential mixture model with cdf $p(1-\exp(-\psi t))$ and hazards function $\alpha(t) = p\psi/(1-p\exp(-\psi t)+p)$, where $0 < p < 1$. They also considered the case in which the immune probability p_i and the exponential rate ψ_i of individual i are related to the covariates $z_i = (x_i, y_i)$ in some canonical manner, such as $p_i = \exp(x_i^\top \delta)/(1+\exp(x_i^\top \delta))$ and $\psi_i = \exp(\beta^\top y_i)$. However, just as pointed out by Chen *et al* (1999), this exponential mixture model, when including covariates through parameters p_i, has some drawbacks too. For example, it *does not* lead to a proportional hazards structure, and it yields improper posterior distributions for many types of noninformative improper prior.

An alternative class of survival models, which *do* have proportional hazards and also allow for long-term survivors, was proposed by Maller and Zhou (1996) via a different motivation from other cure models. Their idea is to naturally extend Cox model $S(t) = [S_0(t)]^{\exp(z^\top \beta)}$ with a parametric or completely unspecified baseline $S_0(t)$ to an "improper" (semiparametric baseline) Cox model defined by $S(t) = [1-pF_0(t)]^{\exp(z^\top \beta)}$, where $F_0(t)$ is a proper distribution function which may be parameterized or completely unspecified. We term this model as an *improper proportional hazards model*. This idea can also be depicted through hazards function as follows: $\alpha(t) = \alpha_0(t)\exp(z^\top\beta)$ with $\alpha_0(t) = pf_0(t)/(1 - pF_0(t))$, where $F_0(t)$ is a proper cdf with density $f_0(t)$. This model, however, has not been further investigated by Maller and Zhou or other authors. A main objective of this dissertation is to follow through such an idea to propose an improper mixture proportional hazards model that incorporates the presence of long-term survivors, and develop some efficient methods for statistical inferences in both theoretical and applicable aspects.

The layout of this dissertation is as follows. In Chapter 1, we begin with introducing some background and specification of this improper proportional hazards model, and point out that the proposed model is different from that of Tsodikov (2003) in both motivation and approaches of analyses. For semiparametric model (with nonparametric baseline hazards $F_0(t)$), the partial likelihood is used to estimate the coefficients of covariates, then an estimator of the immune (long-term survival) proportion is derived from the Nelson-Aalen estimator of the cumulative hazards. In Section 1.3, for the full parametric model, we use the classic maximum likelihood to estimate parameters of interest by a parameterized baseline hazards function. In Section 1.4, the asymptotic properties of the estimators are investigated based on the counting processes and martingale technique, and an example is provided to illustrate our statistical inferences. For semiparametric model, an i.i.d. Cox model is assumed (cf. Tsiatis, 1981). The estimation of the asymptotic variances of estimators is discussed in Section 1.5. Some simulation results and an example of application with a set of Leukaemia data are presented in Section 1.7. Section 1.8 concludes this chapter.

In Chapter 2, we consider the improper proportional hazards model, where covariates are subject to linear measurement error. Approximately corrected partial likelihoods are proposed in semiparametric model to reduce the bias of the 'naive' estimators of the covariate coefficients and the long-term survival proportion. Accurately corrected full likelihood is used to obtain our corrected maximum likelihood estimators of parameters in Section 2.2. In Section 2.3, we investigate their asymptotic properties based on the works of Kong and Gu (1999) and Nakamura (1990). Some simulations are performed in Section 2.4 to illustrate our statistical inferences. Concluding remarks are given in Section 2.5.

In Chapter 3, our main objective is to investigate this improper proportional hazards model for the interior and boundary cases. The boundary case is an interesting topic for survival analysis with long-term survivors since we want to know whether an immune proportion is indeed present in the population, which corresponds to testing $H_0 : p = 1$ (see Ghitany, Maller and Zhou (1994), Zhou and Maller (1995), and Vu and Zhou (1997)). Assumptions and model specifications are given in Section 3.2. Section 3.3 presents some preliminary results, and Section 3.4 derives the main results of the chapter.

In Chapter 4, inspired by the works of Hu (1998) and Doksum (1987), we propose a semiparametric transformation model, which is an alternative to our improper proportion hazards model, and is different from other transformation models in that its error variable may depend on the covariates and some unknown parameters. Section 4.2 specifies the model. We then propose a two-step estimation method for this model as follows. A consistent estimator of the unknown transform function is first given for possibly censored observations in Section 4.3. Then in Section 4.4, by inserting the estimated transform function into the likelihood function, termed as the *pseudo likelihood function*, we obtain the maximum likelihood estimators of the coefficients of the covariates by treating the pseudo likelihood function as an ordinary likelihood function. An estimator of the long-term survival proportion is also provided via the estimator of the transform function in Section 4.4. The asymptotic properties of these estimators are derived in Section 4.5. The variance estimators are also proposed in Section 4.5 using the bootstrap method, since the asymptotic variances of the proposed estimators do not have explicit or closed forms. Section 4.6 reports a simulation study and an example of application in criminology is given in Section 4.7. We close up this chapter by some concluding remarks in Section 4.8.

Finally, in Chapter 5, we are interested in discrete-time survival analysis, especially, discrete-time proportional hazards models, for survival data with long-term survivors. Although continuous-time survival analysis is frequently used in many settings, discrete-time survival analysis is often more appropriate in social and behavioral science applications where the survival data typically possess three features: discrete, ties and concomitant information. Inspired by the works of Potts (2004) and Linoff (2004) in which survival analysis is utilized to mine data, in this chapter we will extend this new idea to fitting survival data with long-term survivors. In Section 5.2 we will review some existing discrete-time survival models which have already been proposed to analyze survival data from social and behavioral science in the literature. Then in Section 5.3 we generalize these models to accommodate long-term survivors. The estimations for discrete-time model and discrete-time cure model are investigated in Sections 5.4-5.5. For discrete-time cure model we predominantly focus on left censored or truncated and right censored data which are recorded at some fixed but non-identical points. Then the proposed models and approaches can be directly applied to analyze survival data from social and behavioral science such as the economic values for customer retention with long-life customers. The asymptotic properties are discussed in Section 5.6. Simulation study is reported in Section 5.7 and an example of application is illustrated in Section 5.8. Finally, some concluding remarks are provided in Section 5.9, which include some worthwhile issues for further research.

Chapter 1

Proportional Hazards Model for Survival Data with Long-term Survivors

1.1. Introduction

Conventional event history models typically assume that the entire population is at risk of experiencing the event of interest throughout the observation period. Individuals who will never experience the event are commonly referred to as *long-term survivors*, which is recently attracted a great deal of interests in the analysis of survival data due to its applications in a wide range of areas, including cancer treatment, AIDS study, criminology, marketing, engineering reliability, etc. For a population with long-term survivors, thus the survival data may consist of a mixture of the data from two latent subpopulations: one who has a non-zero risk of experiencing the event, even if they are not observed to do so during the study period, and another who is not subject to the event of interest, and will continue up to the end of the observation period and will therefore always appear as right-censored. This mixture model formulation is an attractive approach to analyzing such data, in that it contains two parts which can be interpreted separately by adding structure to the standard survival model. Yu *et al.* (2004) gave a nice review for this model. The model can be formulated as follows. The survival model is assumed that the failure time can be decomposed

$$T^* = \eta T^s + (1 - \eta)\infty,$$

where $T^s < \infty$ denotes the failure time of susceptible subject and η indicates, by the value 1 or 0, whether the sampled subject is susceptible or not. If we assume that the proportion of the susceptible $Pr(\eta = 1) = p$, where $p \in (0, 1]$, then the distribution function of T^* given by

$$F(x) = Pr(T^s \leq x)P(\eta = 1) + Pr(\infty \leq x)Pr(\eta = 0)$$
$$= pF_0(x) + 0 = pF_0(x),$$

where $F_0(\cdot)$ is the latent distribution function for T^s (no-cure group). Equivalently,

$$S(t) = pS_0(t) + (1 - p),$$

which is also referred as *standard cure* model (SCR).

Common parametric choices for $F_0(t)$ are exponential and Weibull distribution. Nonparametric choices for $F_0(t)$ also have bee considered. The effects of some independent covariates on both the incidence probability p and the survival function $S_0(t)$ for the susceptible group can be modeled. The incidence model is typically given by $p(x) = \exp(x^\top \gamma_1)/(1 + \exp(x^\top \gamma_1))$, where x is a vector for covariates, γ_1 is a parameter to be estimated and $^\top$ denotes the transpose.

For $S_0(t)$ (hence $F_0(t)$), which is assumed to follow a parametric distribution. Farewell (1982) assumed that a Weibull distribution $S_0(t) = \exp[-\psi t^\alpha]$, where α is a parameter to be estimated and ψ also can be modeled as $\alpha = \exp(x^\top \gamma_2)$. Different formulations can also be used in the above setting, especially in the survival function $S_0(t)$ for the susceptible group. Yamaguchi (1992) applied a cure model with a logistic mixture probability model and an accelerated failure time model with generalized gamma distribution. Maller and Zhou (1996) studied the cure model extensively, specifically nonparametric failure time models for one sample and parametric failure time regression models. Recent work has focused on nonparametric failure models. Taylor (1995) assumed a model with a logistic mixture probability and a completely unspecified failure time process, estimated by a Kaplan-Meier type estimator. Most recently, Kuk and Chen (1992), Sy and Taylor (2000) and Peng and Dear (2000) considered a semiparametric Cox *mixture proportional hazards* model for the failure time process, in which $S_0(t)$ is taken to be a proportional hazards structure as $[S_0(t)]^{\exp(z^\top\beta)}$.

In order to keep the proportional hazards structure for survival data with long-term survivors, Yakovlev and Tsodikov (1996), Tsodikov (1998) and Chen, Ibrahim and Sinha (1999) have proposed a model termed as *no-mixture* models, also as the *bounded cumulative hazards* (BCH) model, which can accommodate long-term survivors and also have the proportional hazards structure, and are different from mixture models. In these models, the probability of cure is incorporated into the proportional hazards model by assuming a bounded cumulative hazard $\tilde{G}_0(t)$ as $t \to \infty$ with $\tilde{G}_0(t) \leq \zeta$, $lim_{t\to\infty}\tilde{G}_0(t) = \zeta$. One way to enforce this is to write $\tilde{G}_0(t) = \zeta G_0(t)$, where $G_0(t)$ is the distribution function of a nonnegative random variable. Then the survival distribution $S(t)$ for the population can be written as $S(t) = \exp\{-\zeta G_0(t)\}$ We can see that the cure rate is $lim_{t\to\infty}S(t) = e^{-\zeta}$. Chen *et al.* (1999) showed that if $S(t)$ is taken to have a proportional hazards structure, then the conditional survival function for the susceptible group no longer has no proportional hazards structure. Hence in the non-mixture model, the survival distribution $S(t)$ for the entire population is modeled as a proportional hazard model, whereas in the mixture cure models, the non-cured group is often modeled as a proportional hazards model.

Covariates can be incorporated into the non-mixture cure model through ζ. One example is to use $\zeta(z) = \exp(z^\top\beta)$. Tsodikov (1998) treated $G_0(t)$ as nuisance and used marginal likelihood to estimate the cure rate $\zeta(x)$. Chen *et al.* (1999) specified a parametric or discrete form for $G_0(t)$ and uses a Bayesian approach. Brown and Ibrahim (2003) have extended this non-mixture cure model to including a longitudinal covariates.

Just as Chen *et al.* (1999) reviewed that the SCR model has a mathematical relationship with the BCH model such that any SCR model can be expressed as the BCH model, but there are some biostatistical differences between them. Especially, when the effects of some independent covariates on both the incidence probability p and the survival function $S_0(t)$ for the susceptible group are modeled, these two cure models have distinct mathematical and statistical differences. The identifiability for the SCR and BCH models are investigated in Appendix B.

For the SCR model, Chen *et al.* (1999) pointed out that relating covariates with p, such as $p = \exp(z^\top\beta)/\left(1 + \exp(z^\top\beta)\right)$, has some drawbacks. First, it does not lead to a proportional hazards structure, which is often preferred for survival models. Second, it yields improper posterior for many types of noninformative improper prior, including the uniform prior for the regression coefficients. For the mixture cure model proposed by Kuk and Chen (1992), on the other hand, it is easy to see that it *does not* have a proportional hazards structure for the entire population even if $S_0(t)$ is taken to be proportional hazards. To remedy these drawbacks for the SCR and Kuk and Chen's (1992) models, in this chapter, we will specify an alternative mixture model with covariates, which *does* have a proportional hazards structure and is proposed by Maller and Zhou (1996) via the different motivation of the model from other cure models, such as BCH and Kuk and Chen's (1992) models, and is not further investigated in Maller and Zhou or by other authors so far. Their idea is to extend Cox model $S(t) = [S_0(t)]^{\exp(z^\top\beta)}$ with a parametric or completely unspecified baseline $S_0(t)$ to "improper" (semiparametric baseline) Cox model defined by

$$S(t) = [1 - pF_0(t)]^{\exp(z^\top\beta)},$$

where $F_0(t)$ is a proper distribution function which may be parameterized or completely unspecified, we term this model as "improper" proportional hazards model.

Maller and Zhou (1996) extensively discussed the estimation of the long-term survivors proportion, also termed as "immune" or "cured" proportion, as well as the survival distribution, via parametric or nonparametric approach. In Maller and Zhou (1996) and other related works, ordinary maximum likelihood approach was successfully applied via mixture models to obtain estimators of the parameter p (the "susceptibles proportion") and the parameters associated with the proper survival function for the susceptible population, and their large-sample properties are obtained via classic approaches. But more advanced theory and techniques, such as the counting process and martingale technique, have not been utilized. The martingale technique, which is based on the statistical theory of counting process initiated by Aalen (1975), and further developed by Andersen and Gill (1982, 1993) and Fleming and Harrington (1991), among others, has since been extensively applied to survival analysis. In this dissertation, we attempt to combine the idea of "long-term survivors" with the Cox proportional hazards model (Cox, 1975) using the martingale technique.

A main objective of this chapter is to follow through such an idea to propose an improper mixture PH model that incorporates the presence of long-term survivors, and apply some martingale techniques to derive relevant large-sample properties. The identifiability for this new proportional hazards cure model is also investigated in Appendix B. It should be noted that our model is fundamentally different from the alternative improper PH model proposed by Tsodikov (1998,2003), in both the motivation of the model and the approach of analyses.

In Section 1.2, we first specify the models and consider the ordinary maximum likelihood function based on the hazard rate; and then give an equivalent reformulation of the model in terms of the intensities of counting processes and the partial likelihood function. Parameter estimation is discussed in Section 1.3 based on the Cox PH model and full likelihood method. In Section 1.4, the martingale theory is applied to derive the asymptotic properties of the statistical procedures related to the model. The estimation of the asymptotic variances of estimators is investigated in Section 1.5. Sections 1.6-1.7 report some simulation results and an application of the model on a set of leukaemia data. Section 1.8 concludes this Chapter.

1.2. Model Specification

Following the usual formulation in survival analysis, we postulate that $T_i^*,\ i = 1, 2, \ldots, n$, are independent and continuously distributed positive random variables representing the failure times of n individuals, each of whom can only be observed within a time interval $[0, c_i]$ subject to censoring times $c_i,\ i = 1, 2, \ldots, n$. The observations for the ith individual consist of $T_i = T_i^* \wedge c_i = \min\{T_i^*, c_i\}$ and $\delta_i = I\{T_i^* \leq c_i\}$, where $I\{\cdot\}$ denotes the indicator of an event. We further assume that T_i^* is independent of c_i for each i, and $\{(T_i^*, c_i), i = 1, \ldots, n\}$ are independent pairs. In addition, $\{c_i\}$ are assumed to follow a common cumulative distribution function G, which is referred to as an independent and identically distribution (i.i.d.) censoring model. The distributions of $T_i^*,\ i = 1, 2, \ldots, n$, on the other hand, are not necessarily identical, and may depend on such covariates as age, gender, treatment method, etc.

Let $F_i(t)$ be the cumulative distribution function (cdf) of T_i^*. The hazard function of T_i^* is the instantaneous rate at which failures occur among remaining survivors. If T_i^* is a continuous random variable with density $f_i(t) = dF_i(t)/dt$, then its hazard function is given by $\alpha_i(t) = f_i(t)/[1 - F_i(t)]$, which is defined at points $0 \leq t < \tau_{F_i}$, where $\tau_{F_i} = \inf\{t \geq 0 : F_i(t) = 1\}$ (so that $F_i(t) < 1$). For a (possibly censored) observation t_i of T_i^* with censoring indicator δ_i, its contribution to the likelihood function can be written in terms of its hazard function $\alpha_i(t)$ by

$$\begin{aligned} L_i(t_i) = f_i(t_i)^{\delta_i}[1 - F_i(t_i)]^{1-\delta_i} &= \left[\frac{f_i(t_i)}{1 - F_i(t_i)}\right]^{\delta_i} [1 - F_i(t_i)] \\ &= \alpha_i(t_i)^{\delta_i} \exp\left\{-\int_0^{t_i} \alpha_i(y) dy\right\}. \end{aligned} \tag{1.1}$$

Let z_i be a $k \times 1$ vector of covariates associated with the survival time T_i^* of an individual under study. The Cox PH model specifies the hazards function of T_i^* with covariate vector z_i by

$$\alpha_i(t) = \alpha_i(t, z) = \alpha_0(t) \exp\{z_i^\top \beta\}, \tag{1.2}$$

where $\alpha_0(t)$ is a baseline hazards function that does not depend on the covariates, and $\beta = (\beta_1, \ldots, \beta_q)^\top$ is an unknown vector of regression parameters (coefficients of covariates) to be estimated.

As discussed in Maller and Zhou (1996), the population contains "long-term survivors" (or "cured" or "immune" individuals) if the cdf $F(t)$ of T^* is *improper*, i.e., $F(\infty) = \mathrm{P}(T^* < \infty) < 1$. To accommodate possible presence of long-term survivors, we allow the baseline cdf of the survival time to be improper of the form $F(t) = pF_0(t)$, where $0 < p \leq 1$ is an additional parameter and $F_0(t)$ is a proper cdf with density $f_0(t)$. The parameter p can be interpreted as the proportion of "susceptible" individuals (those who are not long-term survivors) when covariates have no effects on the survival times. Under such a formulation, the baseline hazard function becomes

$$\alpha_0(t) = \alpha_0(t, p) = \frac{dF(t)/dt}{1 - F(t)} = \frac{d(pF_0(t))/dt}{1 - pF_0(t)} = \frac{pf_0(t)}{1 - pF_0(t))}. \tag{1.3}$$

Therefore we propose to model the hazard rate function of a survival time T_i^* with covariate z_i by

$$\alpha_i(t, z) = \alpha_0(t) \exp\{z_i^\top \beta\} = \frac{pf_0(t)}{1 - pF_0(t)} \exp\{z_i^\top \beta\}. \tag{1.4}$$

By (1.4), if z_i does not depend on t, then the probability for the *ith* individual with covariate z_i to be a long-term survivor is given by

$$\begin{aligned} 1 - \mathrm{P}(T_i^* < \infty) &= \exp\left\{ - \int_0^\infty \alpha_i(t, z) dt \right\} = \exp\left\{ - \int_0^\infty \frac{pf_0(t)dt}{1 - pF_0(t)} \exp(z_i^\top \beta) \right\} \\ &= \exp\left\{ \log(1 - p) \exp(z_i^\top \beta) \right\} = (1 - p)^{\exp(z_i^\top \beta)}, \end{aligned} \tag{1.5}$$

and this reduces to $1 - p$ when there are no covariate effects ($\beta = 0$).

Let $t_1, \ldots, t_n$ be the ordered sample of (possibly censored) survival data with corresponding covariate vectors $z_1, \ldots, z_n$. If the baseline distribution $F_0(t)$ is parameterized with parameter vector ψ, then by (1.1) and (1.3), the likelihood function, which is referred to as the *full likelihood*, can be written as

$$
\begin{aligned}
L_f = L_f(\psi, \beta, p) &= \prod_{i=1}^{n} \alpha_i(t_i, z_i)^{\delta_i} \exp\left\{-\int_0^{t_i} \alpha_i(y, z_i) dy\right\} \\
&= \prod_{i=1}^{n} [\alpha_0(t_i) \exp(z_i^\top \beta)]^{\delta_i} \exp\left\{-\int_0^{t_i} \alpha_0(s) \exp(z_i^\top \beta) ds\right\} \\
&= \prod_{i=1}^{n} \left[\frac{pf_0(t_i)}{1 - pF_0(t_i)} \exp(z_i^\top \beta)\right]^{\delta_i} [1 - pF_0(t_i)]^{\exp(z_i^\top \beta)}. \qquad (1.6)
\end{aligned}
$$

The maximum likelihood estimates of the parameters $(\psi, \beta^\top, p)$ can then be obtained by maximizing the L_f with respect to $(\psi, \beta^\top, p)$.

If $F_0(t)$ is a nonparametric distribution function, then the model in (1.4) is said to be *semi-parametric*. In such a case we can maximize the *partial likelihood* function to estimate β first, then estimate p and $F_0(t)$ nonparametrically. The partial likelihood function, denoted by L_p, is given by

$$
L_p = \prod_i \frac{\exp(z_i^\top \beta)}{\sum_{j \in R(t_i)} \exp(z_j^\top \beta)}, \qquad (1.7)
$$

where $R(t)$ denotes the *risk set* at time t, i.e., the set of individuals who have not failed or been censored prior to time t.

To apply the martingale theory for statistical inference purposes, we now reformulate the Cox PH model in terms of the random intensity of a multivariate counting process. We begin with some introductory discussions on the meanings of the relevant terms. Suppose that individual i has hazard function

$$
\lambda_i(t) = \lim_{h \to 0^+} \frac{1}{h} P[T_i^* \leq t + h \mid T_i^* \geq t], \quad i = 1, 2, \ldots, n.
$$

A multivariate counting process $\tilde{N}(t) = \{N_i(t), i = 1, 2, \ldots, n\}, \ 0 \leq t < \infty$, is a stochastic process with n components that can be thought of as counting the occurences (as time t proceeds) of n different types of events (or the same event for n different individuals). We suppose these events to occur one at a time. The realizations of each component $N_i(\cdot)$, as functions of t, are integer-valued step functions, equal to zero at time zero and with jumps of size 1 only. We assume they are right-continuous, so that $N_i(t)$ is the random number of events of type i in the interval $[0, t]$, and no two components jump at the same time. Define $\mathcal{F}_t$ to be the σ-algebra generated by $\{\tilde{N}(s), s \leq t\}$. Under certain regularity conditions, the process $\tilde{N}$ has intensity process

$$\tilde{\Lambda}(t) = \{\Lambda_i(t) : 0 \leq t < \infty, i = 1, 2, \ldots, n\},$$

where $\Lambda_i(t)dt = \mathrm{P}(N_i \text{ jumps in a small time interval } [t - dt, t] \mid \mathcal{F}_{t-})$.

Now for the full parametric PH model, we define

$$N_i(t) = I\{T_i^* \wedge c_i \leq t, \delta_i = 1\}. \tag{1.8}$$

Then N_i jumps once at time T_i^* if $T_i^* \leq c_i$. Given what happened before the interval $[t - dt, t]$, we know that individual i either failed at the observed time $T_i^* < t \wedge c_i$, or was censored at time $c_i < t$, or is still alive and uncensored at $t - dt$. In the first two cases, we know that N_i either has made its only jump or will never jump, so the probability of a jump in the interval $[t - dt, t]$ is zero. In the last case, either $T_i^* \in [t - dt, t]$ or $T_i^* > t$, so the probability of jump in $[t - dt, t]$ is $\lambda_i(t)dt$. Thus we can write $\Lambda_i(t)dt = Y_i(t)\lambda_i(t)dt$, where

$$\begin{aligned} Y_i(t) &= I\{T_i^* \geq t, c_i \geq t\} = I\{T_i^* \wedge c_i \geq t\} \\ &= \begin{cases} 1, & \text{if individual } i \text{ is still under observation prior to time } t. \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{1.9}$$

Note that given the past prior to (but not including) time t, $Y_i(t)$ and $\Lambda_i(t)$ are fixed (non-random). We say in such a case that $Y_i(t)$ and $\Lambda_i(t)$ are *predictable.*

With $N_i(t)$ defined in (1.8), $\tilde{N}(t) = \{N_i(t), i = 1, \ldots, n\}$ is a multivariate counting process with intensity process $\tilde{\Lambda}(t) = \{\Lambda_i(t), i = 1, \ldots, n\}$ satisfying $\Lambda_i(t)dt = Y_i(t)\lambda_i(t)dt$. We can rewrite the full likelihood in (1.6) as

$$L_f = \prod_{i=1}^{n} \prod_{t>0} \left[\alpha_0(t) \exp(z_i^\top \beta)\right]^{dN_i(t)} \exp\left\{- \int_0^\infty \alpha_0(t) Y_i(t) dt \exp(z_i^\top \beta)\right\}, \tag{1.10}$$

where $\alpha_0(t)$ is given by (1.3), and the partial likelihood in (1.7) is equivalent to

$$L_p = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{Y_i(t) \exp(z_i^\top \beta)}{\sum_{j=1}^{n} Y_j(t) \exp(z_j^\top \beta)} \right\}^{dN_i(t)}. \tag{1.11}$$

Note that $dN_i(t) = 1$ at $t = t_i \leq c_i$ and $dN_i(t) = 0$ otherwise. We take the convention that $a^0 = 1$ for any a. Hence the product over t in (1.10) or (1.11) is actually equal to a single factor with $dN_i(t) = 1$.

In the rest of this section, we show the process $M_i(t) = N_i(t) - \int_0^t \lambda_i(s) Y_i(s) ds$ is a martingale. Let $G(t)$ be the common cdf of the censoring variables c_i's. Since $F_i(t)$ is continuous, $N_i(t) = I\{T_i^* \wedge c_i \leq t, \delta_i = 1\}$ and $Y_i(t) = I\{T_i^* \geq t, c_i \geq t\}$, we have

$$E[N_i(t)] = P(T_i^* \wedge c_i \leq t, \delta_i = 1) = \int_0^t [1 - G(u-)] dF_i(u) \tag{1.12}$$

and

$$E[Y_i(t)] = P(T_i^* \geq t, c_i \geq t) = [1 - F_i(t)][1 - G(t-)]. \tag{1.13}$$

It follows from (1.12)–(1.13) that

$$\begin{aligned} E[dM_i(t)|\mathcal{F}_{t-}] &= E[dN_i(t) - \lambda_i(t) Y_i(t) dt | \mathcal{F}_{t-}] = dE[N_i(t)] - E[Y_i(t)] \lambda_i(t) dt \\ &= [1 - G(t-)] dF_i(t) - [1 - F_i(t)][1 - G(t-)] \frac{dF_i(t)}{1 - F_i(t)} = 0. \end{aligned}$$

Thus $M_i(t)$ is a martingale. Note that $F_i(t)$ can be an improper distribution function, such as $F_i(t) = 1 - (1 - pF_0(t))^{\exp(z_i^\top \beta)}$ under the preceding formulation.

1.3. Estimation of Parameters

We first consider the semiparametric Cox PH model. Cox (1975) derived

$$L_p = L_p(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{Y_i(t) \exp(z_i^\top \beta)}{\sum_{j=1}^{n} Y_j(t) \exp(z_j^\top \beta)} \right\}^{dN_i(t)} \tag{1.14}$$

as a partial likelihood function. Let $\bar{N}(t) = \sum_{i=1}^{n} N_i(t)$. Then by (1.14),

$$\log L_p = \sum_{i=1}^{n} \int_0^\infty z_i^\top \beta dN_i(t) - \int_0^\infty \log \left\{ \sum_{j=1}^{n} Y_j(t) \exp(z_j^\top \beta) \right\} d\bar{N}(t). \tag{1.15}$$

An estimator $\hat{\beta}$ of β is given by maximizing the L_p with respect to β, or equivalently, solving the following equation:

$$U(\beta) = \frac{\partial}{\partial \beta} \log L_p = \sum_{i=1}^{n} \int_0^\infty z_i dN_i(s) - \int_0^\infty \frac{\sum_{i=1}^n Y_i(s) z_i \exp(z_i^\top \beta)}{\sum_{i=1}^n Y_i(s) \exp(z_i^\top \beta)} d\bar{N}(s) = 0. \tag{1.16}$$

Breslow (1974) suggested that the baseline cumulative hazard $\Lambda_0(t) = \int_0^t \alpha_0(s)ds$ can be estimated by

$$\hat{\Lambda}_0(t) = \int_0^t \frac{dN(s)}{\sum_{i=1}^n Y_i(s) \exp(z_i^\top \hat{\beta})}. \tag{1.17}$$

Then an estimator of the baseline survival function $S_0(t)$ is $\hat{S}_0(t) = \exp\left(-\hat{\Lambda}_0(t)\right)$. Furthermore, since

$$\Lambda_0(\infty) = \int_0^\infty \alpha_0(s)ds = \int_0^\infty \frac{pf_0(s)}{1 - pF_0(s)} ds = -\log(1-p),$$

using $\hat{\Lambda}_0(\cdot)$ given by (1.17), p can be estimated by

$$\hat{p} = 1 - \exp -(\hat{\Lambda}_0(+\infty)). \tag{1.18}$$

For the full parametric Cox PH model, let $F(t, \psi)$ be the parameterized baseline cumulative distribution function with parameter vector ψ, and write $\theta^\top = (\psi, \beta^\top, p)$. Then the log-likelihood function is given by

$$\log L_f = \sum_{i=1}^{n} \left\{ \int_0^\infty \log\{\alpha_i(t, \theta, z_i)\} dN_i(t) - \int_0^\infty \alpha_i(t, \theta, z_i) Y_i(t) dt \right\}, \tag{1.19}$$

where $\alpha_i(t, \theta, z_i) = \alpha_0(t, p, \psi) \exp(z_i^\top \beta)$ and $\alpha_0(t, p, \psi) = pf_0(t, \psi)/[1 - pF_0(t, \psi)]$.

Assuming the order of differentiation and integration to be interchangeable, the first derivative vector of the log-likelihood with respect to θ is given by

$$U(\theta) = \frac{\partial \log L_f}{\partial \theta} = \sum_{i=1}^{n} \int_0^\infty \frac{\partial}{\partial \theta} \log \alpha_i(t, \theta, z_i) dM_i(t), \tag{1.20}$$

which is known as the *score process*, where

$$M_i(t) = N_i(t) - \int_0^t \alpha_i(s, \theta, z_i) Y_i(s) ds, \quad i = 1, \ldots, n.$$

As discussed in Section 1.2, $\{M_i, i = 1, \ldots, n\}$ is a sequence of martingale. Thus an estimator $\hat{\theta}$ can be found by solving the score function (1.20).

1.4. Large-sample Properties of Estimators

1.4.1. Semi-parametric PH Model

In this subsection, we derive the large sample properties of the estimators for the covariate coefficients β, the baseline cumulative hazards function $\Lambda(t)$, and the baseline susceptible proportion p. Our results are mainly based on the Lenglart's inequality and the martingale central limit theory. We consider in this section the case of i.i.d. random covariates $\{z_i\}$ (cf. Andersen and Gill, 1982).

The Cox PH model has been studied extensively in the statistical literature, which includes Andersen and Gill (1982), Gill (1984), Andersen, Borgan, Gill, Keiding (1993), Kalbfleish and Prentice (2002), Fleming and Harington (1991), among others. In particular, the results in Andersen and Gill (1982) can be employed to well fit our model with long-term survivors. Let "$\xrightarrow{p}$" and "$\xrightarrow{d}$" denote convergence in probability and in distribution respectively.

Define $\tau_L = \inf\{t \geq 0 : L(t) = 1\}$ for any distribution function $L(t)$, with $\tau_L = \infty$ if $L(t) < 1$ for all $t \geq 0$. We now state our main results of the asymptotic properties for aforementioned estimators.

Theorem 1.1. *Suppose that (N_i, Y_i, z_i), $i = 1, 2, \ldots, n$, are i.i.d. replicates of (N, Y, z), where $N(t)$ and $Y(t)$ are counting process and predictable process corresponding to an improper survival function $(1 - pF_0(t))^{\exp(z^\top \beta)}$ with $0 < p < 1$, z is bounded, and the matrix Σ defined by (1.22) in lemma 1.1 below is positive definite. If either $\tau_G = \infty$ or $\tau_{F_0} \leq \tau_G < \infty$, then with probability approaching 1, equation (1.16) has a solution $\hat{\beta}$ such that $\hat{\beta} \xrightarrow{p} \beta_0$ and $n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma^{-1})$ as $n \to \infty$.*

The next theorem gives the asymptotic normality of the Nelson-Aalen estimator for the improper baseline function.

Theorem 1.2. *Under the conditions of Theorem 1.1, for* $t \in [0, \infty]$,

$$n^{1/2}\{\hat{\Lambda}_0(t) - \Lambda_0(t)\} \xrightarrow{d} N(0, b^2(t) + a(t)^\top \Sigma^{-1} a(t)), \quad \textit{as } n \to \infty,$$

where $b(t)$ *and* $a(t)$ *are defined in Lemma 1.2 below,* $\hat{\Lambda}_0(t)$ *is defined in (1.17), and* $N_i(t)$ *and* $Y_i(t)$ *are defined in Theorem 1.1.*

The asymptotic properties of the estimator $\hat{p}$ of p in (1.18) is given by the following theorem.

Theorem 1.3. *Under the conditions of Theorem 1.1,* $\hat{p} \xrightarrow{p} p$ *as* $n \to \infty$, *and*

$$n^{1/2}(\hat{p} - p) \xrightarrow{d} N(0, \exp(-2\Lambda_0(+\infty))\Sigma^{-1}_{\hat{\Lambda}_0(+\infty)}),$$

where $\Sigma_{\hat{\Lambda}_0(+\infty)}$ *is the asymptotic variance of* $\hat{\Lambda}_0(+\infty)$ *(whose estimation was discussed by Tsiatis (1981), see (1.27) below).*

1.4.2. Full Maximum Likelihood Model

Now let $F_0(t, \psi)$ denote $F_0(t)$ and $1 - F_h(t, \psi) = \{1 - pF_0(t, \psi)\}^{\exp(z_h^\top \beta)}$ in the full maximum likelihood model. Then $\theta^\top = (\psi, \beta^\top, p)$ is the parameter vector to be estimated. To avoid confusion we will denote by $\theta_0^\top = (\psi_0, \beta_0^\top, p_0)$ the true value of the parameter and reserve θ for the free parameter in the log-likelihood function, the score function, etc. Let $\tilde{N} = \{N_h, h = 1, \ldots, n\}$ be a sequence of counting processes with intensity processes $\lambda(t) = \{\lambda_h(t), h = 1, \ldots, n\}$, which have a parametric form $\lambda_h(t) = \lambda_h(t, \theta_0) = \alpha_h(t, \theta_0) Y_h(t)$ with

$$\alpha_h(t, \theta_0) = \frac{p f_0(t, \psi)}{1 - pF_0(t, \psi)} \exp(z_h^\top \beta), \quad h = 1, \ldots, n.$$

To simplify the notation we write $\frac{\partial}{\partial \theta_j} g(\theta_0)$ for $\frac{\partial}{\partial \theta_j} g(\theta)|_{\theta=\theta_0}$ and assume covariates $z_1, \ldots, z_n$ to be one-dimensional and non-random.

The following theorem provides the asymptotic theory for full likelihood model. We say that a sequence of events $\{A_n\}$ occurs with probability approaching 1 (WPA1) if $\Pr\{A_n\} \to 1$ as $n \to \infty$.

Theorem 1.4. *Suppose that there exists a neighborhood Θ_0 of θ_0 such that $F_0(t,\psi)$ and $f_0(t,\psi)$ are three-times continuously differentiable with respect to ψ in Θ_0; and that the derivatives of $\int_0^t \alpha_h(s,\theta)ds$ can be taken inside the integral. Let*

$$g_h^{(j)}(t,\theta_0) = \left\{\frac{\partial}{\partial\theta_j}\log\alpha_h(t,\theta_0)\right\}^2 \quad \text{and} \quad g_h^{(jl)}(t,\theta_0) = \left\{\frac{\partial}{\partial\theta_j\partial\theta_l}\log\alpha_h(t,\theta_0)\right\}^2$$

for $j,l = 1,2,3$. If the following regularity conditions (referred to as Condition G) hold:

(G1) there exist functions $\gamma(t)$ and $\rho(t)$ such that for all j,l,m and $t \in [0,\infty]$,

$$\sup_{\theta\in\Theta_0}\left|\frac{\partial^3\alpha_h(t,\theta)}{\partial\theta_j\partial\theta_l\partial\theta_m}\right| \le \gamma(t), \quad \sup_{\theta\in\Theta_0}\left|\frac{\partial^3\log\alpha_h(t,\theta)}{\partial\theta_j\partial\theta_l\partial\theta_m}\right| \le \rho(t);$$

(G2) the limits of the sequences of functions

$$n^{-1}\sum_{h=1}^n g_h(t,\theta_0)\alpha_h(t,\theta_0)(1-F_h(t,\theta_0)) \quad \text{and} \quad n^{-1}\sum_{h=1}^n \gamma(t)(1-F_h(t,\theta_0))$$

exist, where $g_h(t,\theta_0)$ represents one of $g_h^{(j)}(t,\theta_0)$, $g_h^{(jl)}(t,\theta_0)$ and $\rho(t)$;

(G3) there exist integrable functions $\pi(t,\theta_0)$ and $q(t,\theta_0)$ such that

$$n^{-1}\sum_{h=1}^n g_h(t,\theta_0)\alpha_h(t,\theta_0)(1-F_h(t,\theta_0)) < \pi(t,\theta_0),$$

and $n^{-1}\sum_{h=1}^n \gamma(t)(1-F_h(t,\theta_0)) < q(t,\theta_0)$,

then with probability approaching 1, the equation $U(\theta) = 0$, where $U(\theta)$ is given in (1.20), has a solution $\hat\theta$ such that $\hat\theta \xrightarrow{p} \theta_0$ and $n^{1/2}(\hat\theta - \theta_0) \xrightarrow{d} N(0,\Sigma^{-1})$ as $n \to \infty$, where $\Sigma = \{\sigma_{jl}(\theta_0)\}$ is defined by part D of Condition 6.1.1 in Andersen et al. (1993, p. 421) (referred to as A-K Condition), e.i., as $n \to \infty$,

$$n^{-2}\int_0^\infty \sum_{h=1}^n \{\frac{\partial}{\partial\theta_j}\log\alpha_h(t,\theta_0)\}\{\frac{\partial}{\partial\theta_l}\log\alpha_h(t,\theta_0)\}\alpha_h(t,\theta_0)dt \xrightarrow{p} \sigma_{jl}(\theta_0).$$

1.4.3. Exponential Mixture Model with Covariates.

Let us consider an example with an exponential cdf $F_0(t, \psi_0) = 1 - e^{-\psi_0 t}$. Then the survival function for individual h is

$$1 - F_h(t, \theta_0) = (1 - p_0 F_0(t, \psi_0))^{\exp(z_h \beta_0)} = (1 - p_0 + p_0 e^{-\psi_0 t})^{\exp(z_h \beta_0)}$$

and so the hazard rate is

$$\alpha_h(t, \theta_0) = \frac{p_0 \psi_0 \exp(-\psi_0 t)}{1 - p_0 F_0(t, \psi_0)} \exp(z_h \beta_0).$$

Assume that $z \in [-c_1, c_2]$, $\beta \in [-b_1, b_2]$, $p_0 \in (0, 1)$, $\psi_0 \in (0, \infty)$, where c_i, b_i $(i = 1, 2)$ are positive constants, and let $\Theta = [-c_1, c_2] \otimes [-b_1, b_2] \otimes (0, 1) \otimes (0, \infty)$. We can verify easily (thought a bit lengthy) that Condition G of Theorem 4 holds. For example, by differentiating $\alpha_h(t, \theta)$ in the above (related derivatives are listed in Appendix A), we can see that, for $(\theta_1, \theta_2, \theta_3) = (\psi, \beta, p)$,

$$\sup_{\theta \in \Theta_0} \left| \frac{\partial^3 \log \alpha_h(t, \theta)}{\partial \theta_j \partial \theta_l \partial \theta_m} \right| \leq \rho(t), \quad \text{for all } t > 0 \text{ and } i, j, m \in \{1, 2, 3\},$$

where

$$\rho(t) = \frac{t^3 (4p_2^3 - 5p_2^2 + 4p_2)}{(1 - p_2)^3} + \frac{t^2 (2p_2 - 1)}{(1 - p_2)^3} + \frac{2t}{(1 - p_2)^3} + \frac{4}{(1 - p_2)^3} + \frac{2}{p_1^2} + \frac{1}{\psi_1^3}$$

with $p_2 = p_0 + \delta$, $p_1 = p_0 - \delta$, $\psi_1 = \psi_0 - \delta$ for sufficient small $\delta > 0$ such that $\Theta_0 \in \Theta$. Further, $\pi^{(3)}(t, \theta_0)$, $\pi^{(33)}(t, \theta_0)$ and $\pi_\rho(t, \theta_0)$ can be taken as

$$\pi^{(3)}(t, \theta_0) = \left[\frac{1}{p_0} + \frac{1 - \exp(-\psi_0 t)}{1 - p_0} \right]^2 \frac{c p_0 \psi_0 \exp(-\psi_0 t)}{1 - p_0},$$

$$\pi^{(33)}(t, \theta_0) = \left\{ \frac{1}{p_0^2} + \left[\frac{1 - \exp(-\psi_0 t)}{1 - p_0} \right]^2 \right\}^2 \frac{c p_0 \psi_0 \exp(-\psi_0 t)}{1 - p_0}$$

and

$$\pi_\rho(t, \theta_0) = \rho(t) \frac{c p_0 \psi_0 \exp(-\psi_0 t)}{1 - p_0},$$

where c is an upper bound of $\exp(z_h\beta)$. Similarly, we can find $\gamma(t)$ and consequently $q(t,\theta_0) = \gamma(t)cp_0\psi_0\exp(-\psi_0 t)/(1-p_0)$. The other functions required in Condition G can be found easily as well.

Note that the function $t^{n_1}\exp(-\psi_0 t^{n_2})$ is integrable for positive integers n_1 and n_2. Hence it is easy to see that the above functions satisfy Condition G. Thus the conclusions of Theorem 1.4 hold for exponential $F_0(t,\psi)$.

1.4.4. Proofs

The proofs of Theorems 1.1-1.3 draw on the results of Anderson and Gill (1982), although some conditions need to be checked in fine details as shown in Lemma 1.1 below. We begin with Lemma 1.1 from Andersen and Gill (1982).

Lemma 1.1. *Let (N_i, Y_i, z_i), $i = 1, 2, \dots, n$, be i.i.d. replicates of (N, Y, z), where $N(t)$ and $Y(t)$ are counting process and predictable process corresponding to a survival function $S(t)$ with $E[N(\infty)] < \infty$, and z is bounded. Assume that, for each $\tau < \infty$,*

$$\mathrm{P}(Y(t) = 1, \forall t \leq \tau) > 0 \tag{1.21}$$

and the matrix

$$\Sigma = \int_0^\infty v(\beta_0, t)s^{(0)}(\beta_0, t)\lambda_0(t)dt \tag{1.22}$$

is positive definite, where

$$v(\beta,t) = \frac{s^{(2)}(\beta,t)}{s^{(0)}(\beta,t)} - \left[\frac{s^{(1)}(\beta,t)}{s^{(0)}(\beta,t)}\right]^2, \quad s^{(i)}(\beta,t) = E[Y(t)z^i\exp(z^\top\beta)],\ i = 0,1,2.$$

Then with probability approaching 1, the equation in (1.16) has a solution $\hat\beta$ such that $\hat\beta \xrightarrow{p} \beta_0$, and $n^{1/2}(\hat\beta - \beta_0) \xrightarrow{d} N(0, \Sigma^{-1})$ as $n \to \infty$.

Similar to the proof of Lemma 1.1, we can proceed to prove Theorem 1.1.

Proof of Theorem 1.1. Note that $N(t)$ and $Y(t)$ correspond to an improper cumulative distribution function $F(t) = 1 - (1 - pF_0(t))^{\exp(z^\top\beta)}$, and that $p < 1$, we have

$E[N(\infty)] < \infty$. Similar to the proof of Lemma 1.1 (see Andersen and Gill (1982)), it suffices to show that (1.21) holds either for $\tau < \tau_{F_0}$ or for $\tau < \infty$.

Since $F(t)$ is improper, $\mathrm{P}(T^* > \tau) = 1 - F(\tau) > 0$ for any $\tau < \infty$. Furthermore, if $\tau_G = \infty$, then $\mathrm{P}(c \geq \tau) = 1 - G(\tau-) > 0$ for any $\tau < \infty$. Hence

$$\mathrm{P}(Y(t) = 1, \forall t \leq \tau) = \mathrm{P}(T^* \wedge c \geq \tau) = \mathrm{P}(T^* \geq \tau)\mathrm{P}(c \geq \tau) > 0 \tag{1.23}$$

for any $\tau < \infty$. If $\tau_{F_0} \leq \tau_G < \infty$, on the other hand, then $\mathrm{P}(c \geq \tau) = 1 - G(\tau-) > 0$ for any $\tau < \tau_{F_0} \leq \tau_G$, which shows that (1.23) holds for $\tau < \tau_{F_0}$, and so does (1.21). This completes the proof. ∎

The next lemma on the cumulative hazards estimation is due to Andersen and Gill (1982) as well (cf. also, Kalbfleisch and Prentice (2002), p.177).

Lemma 1.2. *Under the same conditions as for Lemma 1.1, as $n \to \infty$,*

$$n^{1/2}\{\hat{\Lambda}_0(t) - \Lambda_0(t)\} \xrightarrow{d} N(0, b^2(t) + a(t)^\top \Sigma^{-1} a(t)), \quad t \in [0, \infty],$$

where

$$b(t) = \int_0^t \frac{\alpha_0(s)ds}{s^{(0)}(\beta, s)} \quad \text{and} \quad a(t) = \int_0^t \frac{s^{(1)}(\beta, s)}{s^{(0)}(\beta, s)} \alpha_0(s)ds,$$

$\hat{\Lambda}_0(t)$ is defined in (1.17), and $N_i(t)$ and $Y_i(t)$ are as in Lemma 1.1.

Proof of Theorem 1.2. This follows directly from Lemma 1.2. ∎

Proof of Theorem 1.3. This follows from the well known Slutsky theorem and delta-method together with Theorem 1.2. ∎

The following lemma is due to Gill (1983a) (see also Andersen, Borgan, Gill and Keiding (1993), p. 85) and plays a key role in proving the Theorem 1.4.

Lemma 1.3 *Let $\{X^{(n)}(s)\}$ be a sequence of stochastic processes with a deterministic limit $f(s)$ as $n \to \infty$ for almost all $t \in [0, \infty)$, where $\int_0^\infty |f(s)|ds < \infty$. Furthermore, for all $\delta > 0$, there exists $k_\delta(s)$ with $\int_0^\infty k_\delta(s)ds < \infty$ such that*

$$\liminf_{n\to\infty} \mathrm{P}(|X^{(n)}(s)| \leq k_\delta(s) \text{ for all } s\ ) \geq 1 - \delta.$$

Then

$$\sup_t \left| \int_0^\infty X^{(n)}(s)ds - \int_0^\infty f(s)ds \right| \xrightarrow{p} 0.$$

Proof of Theorem 1.4. To prove Theorem 4, we need the following "in probability linear bound" by Daniels (1945). Let H_n be the empirical distribution based on a random sample of size n from a continuous distribution H. Then $\mathrm{P}\{1 - H_n(s) \leq \delta^{-1}(1 - H(s))\} \geq 1 - \delta$ for any $\delta \in (0, 1)$ and all $s \in [0, \infty]$. We first use this result on the empirical distribution of the uncensored t_i^* to obtain $\mathrm{P}\{I(T_i^* \geq t) \leq \delta^{-1}(1 - F_i(t, \theta_0))\} \geq 1 - \delta$. Since $Y_i(t) = I(T_i \geq t) \leq I(T_i^* \geq t)$, we get $\mathrm{P}\{(Y_i(t) \leq \delta^{-1}(1 - F_i(t, \theta_0))\} \geq 1 - \delta$ for all $t \in [0, \infty)$. Then for a function $g_i(t, \theta_0)$, we have

$$\mathrm{P}\left\{ \frac{1}{n} \sum_{i=1}^n A_i \leq \frac{1}{\delta n} \sum_{i=1}^n B_i \right\} \geq 1 - \delta \quad \text{for all } t \in [0, \infty), \tag{1.24}$$

where $A_i = g_i(t, \theta_0)\alpha_i(t, \theta_0)Y_i(t)$ and $B_i = g_i(t, \theta_0)\alpha_i(t, \theta_0)[1 - F_i(t, \theta_0)]$.

On the other hand, as $E[Y_i(t)] = \mathrm{P}(T_i^* \geq t, c_i \geq t) = (1 - F_i(t, \theta_0))(1 - G(t-))$ and $Var(Y_i(t)) \leq E[Y_i(t)]$, by Condition $(G2)$ and the Markov weak law of large numbers (Sen and Singer (1993), p. 63, let $\delta = 2$), we can easily see that

$$\frac{1}{n} \sum_{i=1}^n A_i \xrightarrow{p} y(t), \tag{1.25}$$

where $y(t) = \lim\limits_{n \to \infty} n^{-1} \sum_{i=1}^n B_i[1 - G(t-)]$.

Using Condition G together with Lebesgue Dominated Convergence Theorem, we have $\int_0^\infty y(t)dt < \infty$. Hence by Lemma 1.3 and (1.24)–(1.25),

$$\frac{1}{n} \int_0^\infty \sum_{i=1}^n g_i(t, \theta_0)\alpha_i(t, \theta_0)Y_i(t)dt \xrightarrow{p} \int_0^\infty y(t)dt. \tag{1.26}$$

Now let $g_h(t, \theta_0) = g_h^{(j)}(t, \theta_0) = \{\frac{\partial}{\partial \theta_j} \log \alpha_h(t, \theta_0)\}^2$. Then part (C) of A-K Condition follows from (1.26). By the Cauchy-Schwarz inequality, part (B) of A-K Condition also holds. Similarly, by letting

$$g_h(t, \theta_0) = g_h^{(jl)}(t, \theta_0) = \left\{ \frac{\partial}{\partial \theta_j \partial \theta_l} \log \alpha_h(t, \theta_0) \right\}^2$$

or $\rho(t)$, we can see that (1.26) together with (G2) imply (E) of A-K Condition. In particular, if

$$g_{il}(t,\theta_0) = \left\{ \frac{\partial}{\partial \theta_j} \log \alpha_h(t,\theta_0) \right\} \left\{ \frac{\partial}{\partial \theta_l} \log \alpha_h(t,\theta_0) \right\},$$

then (B) of A-K Condition leads to $\sigma_{jl}(\theta_0) = \int_0^\infty y(t)dt$, provided that $\Sigma = \{\sigma_{jl}(\theta_0)\}$ is positive define (in part (D) of A-K Condition). Thus A-K Condition holds and Theorem 1.4 then follows from Lemma 1.3 and Theorem 6.1.1 in Andersen et al. (1993). ∎

1.5. Estimation of the Asymptotic Variance of Estimators

In this section, we will discuss the estimators of the asymptotic variances. For simplicity we consider the case of 1-dimensional i.i.d. covariates $z_1, \ldots, z_n$, which are realizations of a random variable z. Censoring is assumed independent of the failure time and z. Let $t_1, t_2, \ldots, t_n$ be the ordered observations. We can estimate β by maximizing the logarithm of the Cox partial likelihood of β:

$$\log L_p = \sum_{i=1}^{n} \int_0^\infty \left\{ z_i\beta - \log \left(\sum_{j=1}^{n} Y_j(t) \exp(z_j\beta) \right) \right\} dN_i(t).$$

Given the estimator $\hat{\beta}$, the baseline cumulative hazard $\Lambda_0(t)$ is estimated by

$$\hat{\Lambda}_0(t) = \int_0^t \frac{dN(s)}{\sum_{i=1}^n Y_i(s) \exp\,[z_i\hat{\beta}]}.$$

The asymptotic variance $V = Var(\hat{\Lambda}_0(t))$ of the above estimator, discussed by Tsiatis (1981), can be estimated by

$$\hat{V} = \int_0^t \frac{dN(s)}{\{\sum_{i=1}^n Y_i(s) \exp[z_i\hat{\beta}]\}^2} + I_n^{-1}(\hat{\beta}, t) \left\{ \int_0^t \frac{\epsilon(\hat{\beta}, s)dN(s)}{\sum_{i=1}^n Y_i(s) \exp[z_i\hat{\beta}]} \right\}^2, \qquad (1.27)$$

where

$$\epsilon(\hat{\beta}, s) = \sum_{i=i}^{n} z_i Y_i(s) \exp(z_i\hat{\beta}) / \sum_{i=1}^{n} Y_i(s) \exp(z_i\hat{\beta})$$

and $I_n(\beta, t) = -\frac{\partial^2}{\partial \beta^2} \log L_p$ is the observed information on β (Kalbfleish and Prentice 2002 p. 174):

$$I_n(\beta, t) = \int_0^t \frac{\sum_{i=1}^n [z_i - \epsilon(\beta, s)]^2 Y_i(t) \exp(z_i\beta)}{\sum_{i=1}^n Y_i(s) \exp(z_i\beta)} dN(s),$$
$$= \int_0^t \left\{ \frac{\sum_{i=i}^n z_i^2 Y_i(s) \exp(z_i\beta)}{\sum_{i=1}^n Y_i(s) \exp(z_i\beta)} - \left[\frac{\sum_{i=i}^n z_i Y_i(s) \exp(z_i\beta)}{\sum_{i=1}^n Y_i(s) \exp(z_i\beta)} \right]^2 \right\} dN(s).$$

Using the above, $Var(\hat{\beta})$ can be estimated by

$$\hat{V}ar(\hat{\beta}) = I_n^{-1}(\hat{\beta}, +\infty). \tag{1.28}$$

By the delta method, we have $Var(\hat{p}) \approx Var(\hat{\Lambda}_0(+\infty)) \exp(-2\hat{\Lambda}_0(+\infty))$, hence an estimator of $Var(\hat{p})$ is given by

$$\hat{V}ar(\hat{p}) = \hat{V}ar(\hat{\Lambda}_0(+\infty)) \exp(-2\hat{\Lambda}_0(+\infty)), \tag{1.29}$$

where $\hat{V}ar(\hat{\Lambda}_0)$ is given in (1.27). Following the results of Section 7.2.2 in Andersen *et al.* (1993, pp.496-501), it is easy to show that under the conditions of Theorem 1, $\hat{V}ar(\hat{\beta})$ and $\hat{V}ar(\hat{p})$ are consistent estimators of $Var(\hat{\beta})$ and $Var(\hat{p})$, respectively.

For the full likelihood model, the log-likelihood function (apart from a constant) takes the form

$$\log L_f = \int_0^\infty \sum_{i=1}^n \log \alpha_i(t, \theta, z_i) dN_i(t) - \int_0^\infty \sum_{i=1}^n \alpha_i(t, \theta, z_i) Y_i(t) dt,$$

and is maximized by $\hat{\theta}$. The observed information of θ can be easily found by the results of Kalbfleish and Prentice (2002, p. 63, p. 180). The observed information of p is $I_n(p, t) = -\int_0^t \sum_{i=1}^n \frac{\partial^2}{\partial p^2} \log \alpha_i(t, \theta, z_i) dM_i(t)$. Clearly,

$$I_n(p, +\infty) = \sum_{i=1}^n \left\{ \frac{\delta_i(1 - 2pF_0(\psi, t_i))}{p^2[1 - pF_0(\psi, t_i)]^2} + \frac{\exp(z_i\beta) F_0^2(\psi, t_i)}{[1 - pF_0(\psi, t_i)]^2} \right\}. \tag{1.30}$$

Proceeding along the lines of the above we can find the observed information of β as

$$I_n(\beta, +\infty) = -\sum_{i=1}^n z_i^2 \exp(z_i\beta) \log(1 - pF_0(\psi, t_i)), \tag{1.31}$$

and the observed information of ψ is

$$I_n(\psi, +\infty) = \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\psi^2} + \frac{\psi^2 t_i^2 (p - p^2) \exp(-\psi t_i)[\delta_i - \exp(z_i\beta)]}{\psi^2 [1 - pF_0(\psi, t_i)]^2} \right\}. \tag{1.32}$$

The variances of the estimators of p, β, ψ are estimated by replacing p, β, ψ in (1.30)–(1.32) with their maximum likelihood estimators. Then we have

$$\hat{V}(\hat{p}) = I_n^{-1}(\hat{p}, +\infty), \hat{V}(\hat{\beta}) = I_n^{-1}(\hat{\beta}, +\infty), \hat{V}(\hat{\psi}) = I_n^{-1}(\hat{\psi}, +\infty).$$

These estimators are consistent by standard maximum likelihood theory.

1.6. Simulation Results

In the simulation study, we compare the performance of the parametric maximum likelihood estimators (MLEs) with the maximum partial likelihood estimators (MPLEs). The calculations can be seen more clearly in special cases. We consider the two sample problem with exponentially distributed lifetimes. The two samples are of sizes, say, n_0 and n_1, respectively, $n_0 + n_1 = n$, with sample membership being indicated by the dummy variable

$$z_i = \begin{cases} 0, & \text{if individual } i \text{ is in sample 0} \\ 1, & \text{if individual } i \text{ is in sample 1} \end{cases}, \quad i = 1, 2, \ldots, n.$$

Data are generated from the survival functions $S_0(t) = 1 - pF_0(t)$ (with respect to sample 0) and $S_1(t) = (1 - pF_0(t))^{\exp(\beta)}$ (with respect to sample 1). Let $F_0(t)$ be an exponential distribution with parameter $\psi = 0.058$. The proportion of the susceptibles is $p = 0.90$ and the coefficient of covariates is $\beta = -0.3851$. Censoring times c are generated from an uniform distribution between 0 and 100. For this simulation, samples of sizes $n_1 = n_2 = 100$ and $n_1 = n_2 = 400$ were replicated 10000 times. In the first simulation study, the result is on the mean and standard deviation of the estimators. In the second simulation study, three (survival) distribution functions are compared in each of the two samples.

Let p_0 and p_1 denote the susceptible proportions in samples 0 and 1 respectively. Then $p_0 = p = 0.9$ and $p_1 = 1 - (1-p)^{\exp(\beta)} = 1 - 0.1^{\exp(-0.3851)} = 0.79125$. The means and standard deviations of the simulated estimates are displayed in Table 1.1 below, where $p_0^{(f)}$ and $p_1^{(f)}$ denote the maximum likelihood estimators of p_0 and p_1 respectively under the full likelihood model; $p_0^{(s)}$ and $p_1^{(s)}$ the estimators of p_0 and p_1 respectively under the semiparametric PH model; and $\beta^{(s)}$ and $\beta^{(f)}$ the estimators of the coefficient β under semiparametric Cox model and the full likelihood model respectively.

Table 1.1. Summary of the simulation studies on the estimators of β and p

$n_1 = n_2$		$\beta^{(s)}$	$\beta^{(f)}$	$p_0^{(s)}$	$p_1^{(s)}$	$p_0^{(f)}$	$p_1^{(f)}$	$\hat{\psi}$
100	Mean	-0.3945	-0.3903	0.8931	0.7818	0.9206	0.8022	0.0588
	STD	0.0782	0.0858	0.0368	0.0493	0.0823	0.0481	0.0082
400	Mean	-0.3870	-0.3863	0.8959	0.7857	0.9000	0.7918	0.0581
	STD	0.0392	0.0407	0.0191	0.0255	0.0184	0.0236	0.0036

From Table 1.1 we can see that the estimates are close to the true values of the parameters and the accuracy improves as the sample sizes increase from 100 to 400.

The true and estimated distribution function curves are shown in Figures 1.1 and 1.2 below (with $n_1 = n_2 = 100$), where $F(t)$ denotes the true distribution for the simulation, FP denotes the estimates under the full parametric model and SP the estimates under the semiparametric model. These two figures show that the two estimators are close to the true distribution function curve.

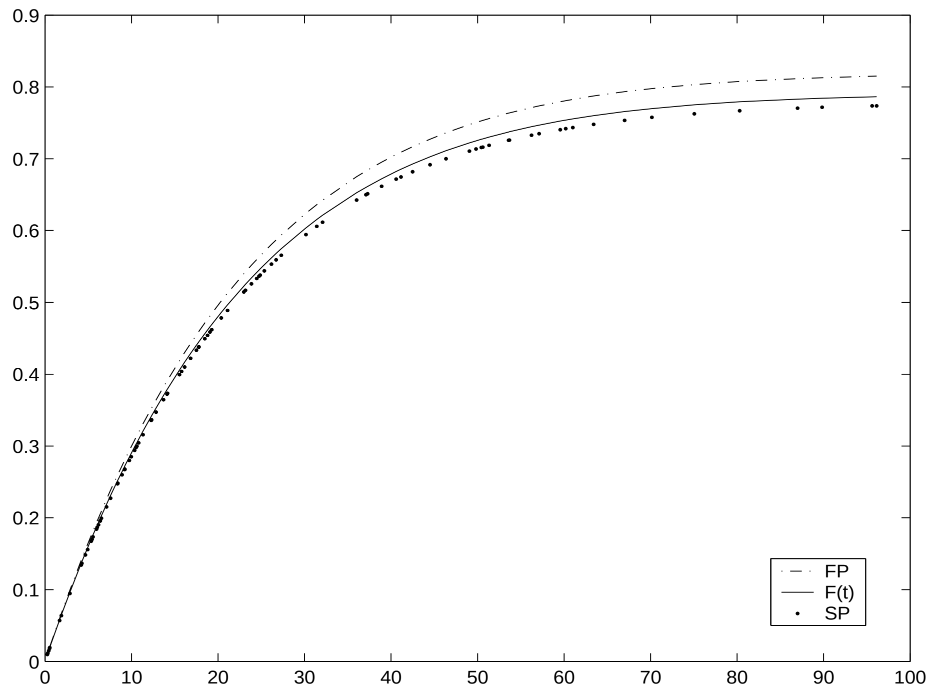


Figure 1.1. The Estimated Distribution Curve for Two Models in Sample 0

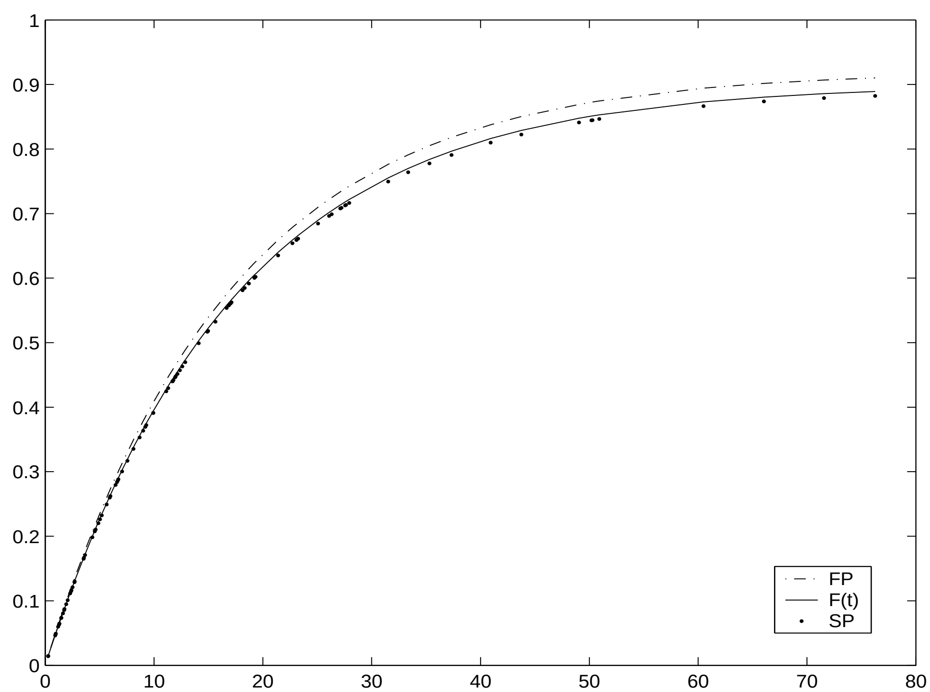


Figure 1.2. The Estimated Distribution Curve for Two Models in Sample 1

1.7. An Example of Application

We consider the leukaemia data analyzed by Goldman *et al.* (1984) and further discussed by Maller and Zhou (1996). Our main interest is to compare the estimates between the semiparametric (partial likelihood) model and the full likelihood model under the PH structure. We take $F_0(t, \psi)$ to be an exponential distribution with parameter ψ. The estimates of the parameters p, β and ψ are listed in Table 1.2 below:

Table 1.2. Estimates for the Leukaemia Data ($n_0 = 46,\ n_1 = 44$)

$\beta^{(s)} = 0.3950$	$p_0^{(s)} = 0.6811$	$p_1^{(s)} = 0.8167$	
$\beta^{(f)} = 0.3690$	$p_0^{(f)} = 0.6892$	$p_1^{(f)} = 0.8155$	$\hat{\psi} = 1.9414$

where $\beta^{(s)}$ and $\beta^{(f)}$ denote the estimates of β based on the semiparametric model and the full likelihood model, respectively; $p_i^{(s)}$ and $p_i^{(f)}$ denote the estimates of p based on the two models with $i = 0$ for Group one (allogeneic transplants) and $i = 1$ for Group two (autologous transplants) data. The distribution curves are displayed in Figures 1.3 and 1.4 below. We can see that the estimators based on the two models are relatively close to each other, and the discrepancies between the two models may be explained by small sample sizes and the choice for the baseline survival function $F_0(t)$.

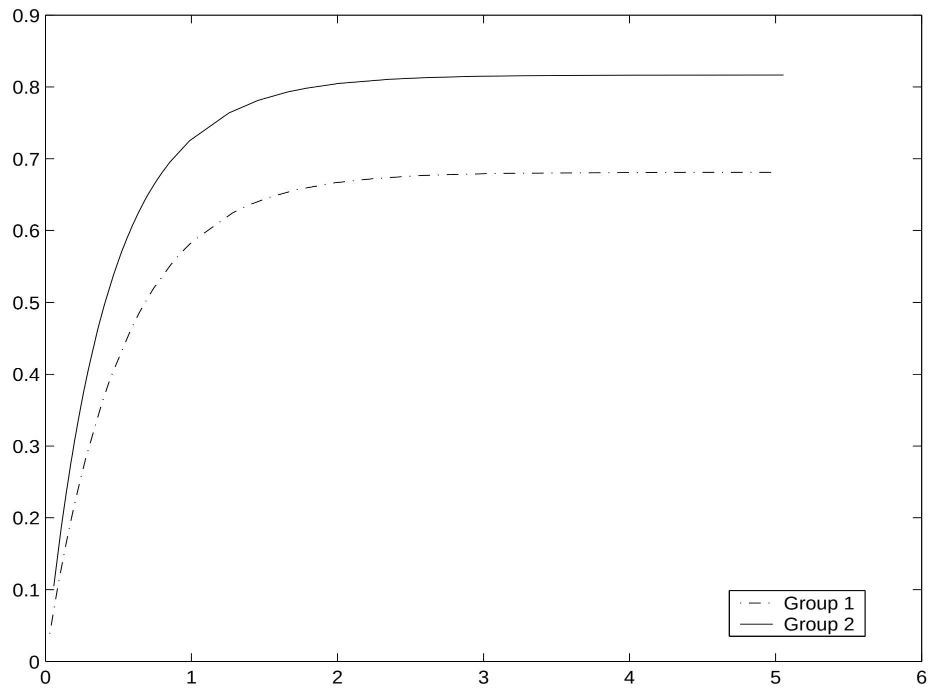


Figure 1.3. cdf's for leukaemia data under the partial likelihood model

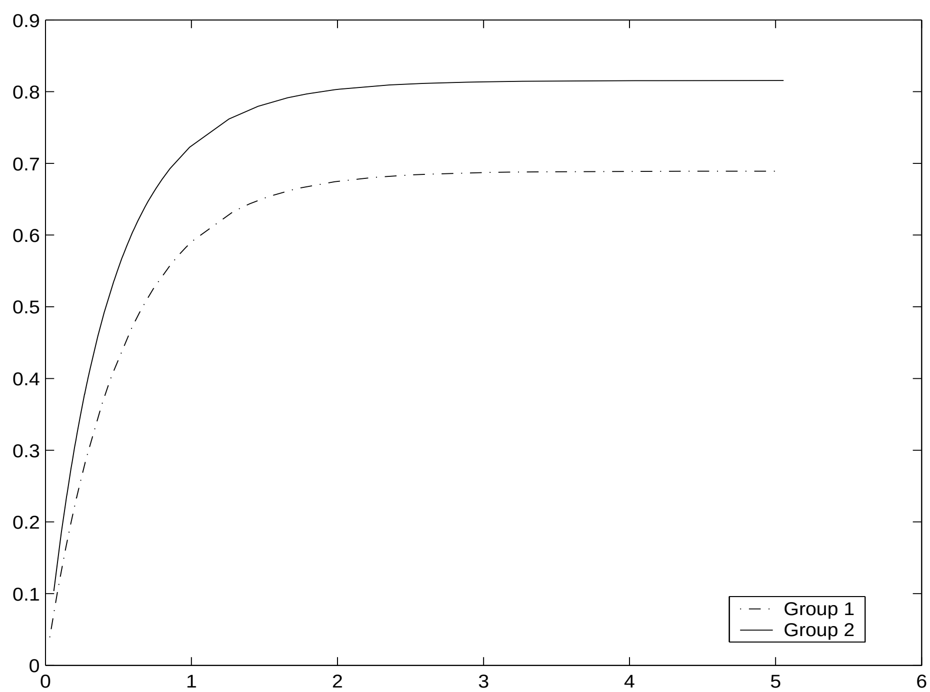


Figure 1.4. cdf's for leukaemia data under the full likelihood model

1.8. Concluding Remarks

In this Chapter we proposed a new improper PH model to analyze survival data with long-term survivors and covariates. Our model, which is based on the mixture model, can overcome the drawbacks mentioned by Chen *et al.* (1999). Our approach is based on the counting processes and martingale techniques, which has rarely been used to investigate the survival data with long-term survivors. The partial maximum likelihood and full maximum likelihood estimators of the parameters are obtained and their asymptotic properties are investigated.

The simulation study indicates that the proposed model and estimation procedures produce efficient estimators for both semiparametric and parametric settings. An application is also demonstrated with a set of Leukaemia data.

Our motivation to model this improper PH model is based on the SCR model, which leads to a natural extension of the usual PH model. Hence methods for the standard PH model may be extended to our improper model, such as Bayesian analysis with an improper noninformative prior, measurement error for the covariates, which may be longitudinal data, and so on, but further investigations are needed. Another issue worth for further consideration is hypothesis testing based on our model, which may involve tests under nonstandard conditions (e.g., p on the boundary of its parameter space). Final issue for further study is the extension of continuous-time improper PH model to discrete-time PH models for the survival data with long-term survivors duo to their extensively applications to economic, , business, marketing, sociological and behavioral statistics.

Chapter 2

Measurement Error in Proportional Hazards Models for Survival Data with Long-term Survivors

2.1. Introduction

In Chapter 1, we reviewed some cure models and Cox mixture proportional hazards models, and thus proposed an alternative "improper" proportional hazards model, which may allow a semiparametric baseline. We also investigated its parameter estimation by using partial likelihood and full likelihood approaches, and then derived their asymptotic properties via the martingale techniques.

In that chapter, our statistical inferences are based on time-independent covariates. We also must know the values of the covariates for all subjects at risk at any failure time points, with the risk set being defined as the set of all patients who are still under study prior to that time point. In many clinical studies, however, it is quite often that some or even all of the covariates are measured with error or are misspecified. A common consequence of such measurement error or misspecification is that the parameter estimates are *attenuated*, in the sense that the estimators shrink towards zero. A more serious consequence is that such error could lead to violation of the correct model relationship (Kong and Gu, 1999).

Considering conventional Cox proportional hazards models with covariates subject to measurement error, Kong and Gu (1999) mentioned two approaches of parameter estimation. One approach, proposed by Prentice (1982), is believed to give consistent *induced maximum partial likelihood estimator* and further explored by Pepe, Self and Prentice (1989), Tsiatis, DeGruttola and Wulfsohn (1995), Wulfson and Tsiatis (1997) and Zhou and Pepe (1995). However, so far there has been no effective way to derive this estimator. Another approach, the *corrected likelihood*, developed by Stefanski (1989) and Nakamura (1990), is to construct an unbiased score function. However, as Stefanski (1989) argued, such a corrected score function does not exist for the partial likelihood score. Therefore, Nakamura (1992) introduced an approximately corrected partial likelihood score and derived an estimator that was less biased than the 'naive' maximum partial likelihood estimator (MPLE), as shown through simulations. Kong and Gu (1999) showed that Nakamura's approach can produce a consistent estimator.

In this chapter, we will apply some martingale techniques and combine the mixture models with PH models to tackle the problem of measurement error in time-independent covariates. An accurately corrected maximum likelihood score and an approximately corrected partial likelihood score are used to obtain the corrected maximum likelihood estimators (CMLEs) and the corrected maximum partial likelihood estimators (CMPLEs), respectively. The asymptotic properties are derived based on the works of Kong and Gu (1999) and Nakamura (1990).

We would like to point out that our proposed method is different from the recent work of Thomas (2004), in which an exact corrected log-likelihood is proposed under the assumption of piecewise constant hazard, that is, the hazard is assumed to be a constant in the interval of two adjacent failure points.

The structure of this chapter is briefly described as follows. Parameter estimations are discussed in Section 2.2 based on Cox proportional hazards model and full likelihood model. In Section 2.3, the martingale theory is applied to derive the asymptotic properties of the statistical procedures related to the model. Section 2.4 reports some simulation results. Section 2.5 concludes this chapter.

2.2. Estimation of Parameters

2.2.1. Proportional Hazards Model

In proportional hazards model for survival data with long-term survivors, recall that the hazard function of an individual with a covariate vector z is given by (cf.(1.4))

$$\alpha(t, \theta, z) = \frac{pf_0(t)}{1 - pF_0(t)} \exp(z^\top \beta).$$

Now we shall denote by Z and $\tilde{Z}$ the set of z's and $\tilde{z}$'s, respectively. Let $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa^\top$ for a column vector a. We introduce the following notations:

$$S^{(r)}(\beta, t, Z) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) z_i^{\otimes r} \exp(z_i^\top \beta), \quad r = 0, 1, 2.$$

$$E(\beta, t, Z) = \frac{S^{(1)}(\beta, t, Z)}{S^{(0)}(\beta, t, Z)}, \quad V(\beta, t, Z) = \frac{S^{(2)}(\beta, t, Z)}{S^{(0)}(\beta, t, Z)} - E^{\otimes 2}(\beta, t, Z),$$

and

$$s^{(r)}(\beta, t) = E[S^{(r)}(\beta, t, Z)], \quad r = 0, 1, 2,$$

where the expectation is taken with respect to the true distribution of (T^*, c, z). Under the above assumptions, as $n \to \infty$, $S^{(r)}(\beta, t, Z) \xrightarrow{p} s^{(r)}(\beta, t)$ for $r = 0, 1, 2$, so that $E(\beta, t, Z) \xrightarrow{p} e(\beta, t)$ and $V(\beta, t, Z) \xrightarrow{p} v(\beta, t)$, where

$$e(\beta, t) = \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} \quad \text{and} \quad v(\beta, t) = \frac{s^{(2)}(\beta, t)}{s^{(0)}(\beta, t)} - e^{\otimes 2}(\beta, t).$$

Define

$$\Sigma = \int_0^t v(\beta, s) s^{(0)}(\beta, s) \alpha_0(s) ds.$$

Let $\bar{N}(t) = \sum_{i=1}^n N_i(t)$ and D be the observed values of $(N_i(\cdot), Y_i(\cdot), t_i), i = 1, \ldots, n$. Then by (1.11) we have

$$\log L_p = \sum_{i=1}^{n} \int_0^\infty z_i^\top \beta dN_i(t) - \int_0^\infty \log \left\{ \sum_{j=1}^{n} Y_j(t) \exp(z_j^\top \beta) \right\} d\bar{N}(t), \qquad (2.1)$$

and the score function corresponding to (2.1) is given by

$$U(\beta, Z, D) = \frac{\partial}{\partial \beta} \log L_p = \sum_{i=1}^{n} \int_0^\infty z_i dN_i(t) - \int_0^\infty E(\beta, t, Z) d\bar{N}(t). \tag{2.2}$$

From Chapter 1, the solution of $U(\beta, Z, D) = 0$ has some attractive properties when the covariates are observed accurately (without measurement error). But in reality it is quite common that some or even all of the covariates are measured with error or are misspecified. Thus we assume an additive measurement error of the form:

$$\tilde{z}_i = z_i + \varepsilon_i, \quad \text{for} \quad i = 1, \ldots, n, \tag{2.3}$$

with normal random variables ε_i independent of z_i. The main assumption on ε_i is that it has mean zero and a variance-covariance matrix $\Lambda_{q\times q}$ that is not dependent on time t. In this chapter we focus on model (2.2) under conditions (2.3) and i.i.d. random measurement errors $\varepsilon_i,\ i = 1, \ldots, n$.

For observable covariates $\tilde{z}_i$, $i = 1, \ldots, n$, if $S^{(r)}(\beta, t, \tilde{Z}),\ r = 0, 1, 2$, $E(\beta, t, \tilde{Z})$, $V(\beta, t, \tilde{Z})$ and $U(\beta, \tilde{Z}, D)$ are defined simply by replacing Z with $\tilde{Z}$ in the above definitions, under observed values of $(\tilde{z}_i, t_i, \delta_i),\ i = 1, \ldots, n$, then the observed score $U(\beta, \tilde{Z}, D)$ is biased. So some modified methods were presented to deal with this problem. In this chapter, Nakamura's corrected score function method will be used to find a corrected estimator for the "improper" proportional hazards model.

Let $\tilde{Z}$ denote a set of observable $\tilde{z}$'s as in (2.3) and replace Z with $\tilde{Z}$ in (2.2). Then the score function becomes

$$U(\beta, \tilde{Z}, D) = \sum_{i=1}^{n} \int_0^\infty \tilde{z}_i dN_i(t) - \int_0^\infty E(\beta, t, \tilde{Z}) d\bar{N}(t), \tag{2.4}$$

which is called a *naive* score function. Let $\beta_{\tilde{z}}$ satisfy the equation $U(\beta_{\tilde{z}}, \tilde{Z}, D) = 0$. Then $\beta_{\tilde{z}}$ is biased (Prentice 1982). Nakamura (1990) proposed a correction of the bias by using a *corrected score function* $U^*(\beta, \tilde{Z}, D)$ whose expectation $E^*[U^*(\beta, \tilde{Z}, D)]$ with respect to the $\varepsilon' s$ given D and Z coincides with $U(\beta, Z, D)$.

Let β^* be a solution of the equation $U^*(\beta^*, \tilde{Z}, D) = 0$ with a positive-definite I_n^*, where I_n^* is the corrected observed information matrix. Such a β^* is called a *corrected estimator*. Stefanski (1989) proved that no exact U^* exists for proportional hazards model, thus we shall find an approximately corrected score function. Applying a first-order approximation, we have (see Nakamura (1992))

$$E^*[E(\tilde{Z}|\beta, D)] \approx E(Z|\beta, D) + \Lambda\beta. \tag{2.5}$$

With Z being replaced by $\tilde{Z}$ and $E(\beta, t, Z)$ by $E(\beta, t, \tilde{Z}) - \Lambda\beta$ in (2.2), the corrected score function is obtained as

$$\begin{aligned} U^*(\beta, \tilde{Z}, D) &= \sum_{i=1}^{n} \int_0^\infty \tilde{z}_i dN_i(t) - \int_0^\infty \left\{ E(\beta, t, \tilde{Z}) - \Lambda\beta \right\} d\bar{N}(t) \\ &= \sum_{i=1}^{n} \int_0^\infty \left\{ \tilde{z}_i - E(\beta, t, \tilde{Z}) + \Lambda\beta \right\} dN_i(t) \\ &= U(\beta, \tilde{Z}, D) + D^*\Lambda\beta, \end{aligned} \tag{2.6}$$

where D^* is the total number of failed subjects. If there exists a solution β^* to $U^*(\beta^*, \tilde{Z}, D) = 0$, then β^* is called the *corrected maximum partial likelihood estimator* (CMPLE).

Breslow (1974) suggested to estimate the underlying cumulative hazard function $\Lambda_0(t) = \int_0^t \alpha_0(s, p) ds$ by

$$\hat{\Lambda}_0(t, Z) = \int_0^t \frac{d\bar{N}(s)}{\sum_{i=1}^n Y_i(s) \exp(z_i^\top \beta^*)} = \int_0^t \frac{d\bar{N}(s)/n}{S^{(0)}(\beta^*, s, Z)}.$$

If we simply replace Z by $\tilde{Z}$ in $\hat{\Lambda}_0(t, Z)$, $\hat{\Lambda}_0(t, \tilde{Z})$ will be a biased estimator of $\Lambda_0(t)$. The corrected score function method can be used to find asymptotically unbiased estimators. If there exists an estimator $\Lambda_0^*(t, \tilde{Z})$ such that

$$E^*[\Lambda_0^*(t, \tilde{Z})] = \hat{\Lambda}_0(t, Z),$$

then $\Lambda_0^*(t, \tilde{Z})$ is called a *corrected estimator* of $\Lambda_0(t)$. The following formula:

$$E^*[\exp(\tilde{z}_i^\top \beta)] = \exp(z_i^\top \beta + \xi) \tag{2.7}$$

will be used to find $\Lambda_0^*(t, \tilde{Z})$, where $\xi = \frac{1}{2}\beta^\top \Lambda \beta$.

Let $A = \sum_{i=1}^n Y_i(t) \exp(\tilde{z}_i^\top \beta)$. Then $E^*(A) = \sum_{i=1}^n Y_i(t) \exp(z_i^\top \beta) \exp(\xi)$. Now define $f(A) = 1/A$. Since $A \xrightarrow{p} E^*(A)$ as $\xi \to 0$, expanding f at $A = E^*(A)$ to the first order and taking the expectation yield an approximate equality with small ξ (Kendall and Stuart, 1977, p. 260),

$$E^*(1/A) \approx 1/E^*(A). \tag{2.8}$$

Then the corrected estimator of $\Lambda_0(t)$ is

$$\Lambda_0^*(t, \tilde{Z}) = \int_0^t \frac{\exp(\xi^*) d\bar{N}(s)}{\sum_{i=1}^n Y_i(s) \exp(\tilde{z}_i^\top \beta^*)} = \int_0^t \frac{\exp(\xi^*) d\bar{N}(s)/n}{S^{(0)}(\beta^*, s, \tilde{Z})}, \tag{2.9}$$

where $\xi^* = \frac{1}{2}\beta^{*\top} \Lambda \beta^*$, which was also given by Kong, Huang and Li (1998) from a different motivation. Hence an estimator of the baseline survival function with "long-term" survivors can be given by

$$\hat{S}_0(t) = \exp\left\{-\Lambda_0^*(t, \tilde{Z})\right\},$$

and the parameter p, the proportion of susceptible individuals, can be estimated by

$$p^* = 1 - \exp\left\{-\Lambda_0^*(+\infty, \tilde{Z})\right\}. \tag{2.10}$$

Remark 2.1: We can obtain more accurate approximate results if the terms on the right-hand sides of equations (2.5) and (2.8) are replaced with their respective second-order approximations.

2.2.2. Maximum likelihood Model

Now we denote $F_0(t)$ by $F_0(t, \psi)$ in maximum likelihood model so that parameters can be easily estimated. Then $\theta^\top = (\psi, \beta^\top, p)$ is the parameter to be estimated. Recall the following likelihood function (cf. (1.6)):

$$\begin{aligned} L_f(\theta, Z, \tilde{D}) &= \prod_{i=1}^n \left[\frac{p f_0(t_i)}{1 - pF_0(t_i)} \exp(z_i^\top \beta)\right]^{\delta_i} \exp\left\{-\int_0^{t_i} \frac{p f_0(s)}{1 - pF_0(s)} \exp(z_i^\top \beta) ds\right\}, \\ &= \prod_{i=1}^n \left[\frac{p f_0(t_i)}{1 - pF_0(t_i)} \exp(z_i^\top \beta)\right]^{\delta_i} [1 - pF_0(t_i)]^{\exp(z_i^\top \beta)}, \end{aligned}$$

where $\tilde{D} = \left\{ \tilde{D}_i = (t_i, \delta_i), i = 1, \ldots, n \right\}$.

Denote by $l(\theta, Z, \tilde{D})$, $U(\theta, Z, \tilde{D})$ and $I_n(\theta, Z, \tilde{D})$ the log-likelihood, score function and observed information $-\partial U(\theta, Z, \tilde{D})/\partial\theta$, respectively, of θ given Z and $\tilde{D}$.

The point θ_z that satisfies $U(\theta, Z, \tilde{D}) = 0$, if attainable, is a maximum likelihood estimate of θ. When z is subject to measurement error and $\tilde{z}$ is the observable value of z, $U(\theta, \tilde{Z}, \tilde{D})$ obtained from $U(\theta, Z, \tilde{D})$ by simply replacing z with $\tilde{z}$ is termed a *naive score function*. The *naive maximum likelihood estimator* $\theta_{\tilde{z}}$, however, is often inconsistent. The corrected score function method to obtain a consistent estimator starts with finding function $l^*(\theta, \tilde{Z}, \tilde{D})$ such that its conditional expectation (also denote by $E^*(.)$) with respective to $\tilde{Z}$ given Z and $\tilde{D}$ is $l(\theta, Z, \tilde{D})$. If $E^*(.)$ and $\partial\theta$ are interchangeable, we have

$$E^*[U^*(\theta, \tilde{Z}, \tilde{D})] = U(\theta, Z, \tilde{D}),$$

where $U^*(\theta, \tilde{Z}, \tilde{D}) = \partial l^*(\theta, \tilde{Z}, \tilde{D})/\partial\theta$. The point $\theta^* = (\beta^*, p^*, \psi^*)$ that solves the equation $U^*(\theta^*, \tilde{Z}, \tilde{D}) = 0$ with a positive definite $I_n^*(\theta, \tilde{Z}, \tilde{D})$ is called a *corrected maximum likelihood estimate*, where

$$I_n^*(\theta, \tilde{Z}, \tilde{D}) = -\partial U^*(\theta, \tilde{Z}, \tilde{D})/\partial\theta.$$

Given observed values of $\tilde{D}_i = (t_i, \delta_i)$ and z_i, the log-likelihood function $l_i(\theta, z_i, \tilde{D}_i)$ is given by

$$\delta_i \left[\ln\left(\frac{p f_0(t_i, \psi)}{1 - pF_0(t_i, \psi)} \right) + z_i^\top \beta \right] + \exp\left(z_i^\top \beta \right) \ln(1 - pF_0(t_i, \psi)). \tag{2.11}$$

Following (2.7), we can define a score function $l_i^*(\theta, \tilde{z}_i, \tilde{D}_i)$ by

$$l_i^*(\theta, \tilde{z}_i, \tilde{D}_i) = \delta_i \left[\ln\left(\frac{p f_0(t_i, \psi)}{1 - pF_0(t_i, \psi)} \right) + \tilde{z}_i^\top \beta \right] + \exp\left(\tilde{z}_i^\top \beta - \xi_i \right) \ln(1 - pF_0(t_i, \psi)), \tag{2.12}$$

where $\xi_i = \frac{1}{2}\beta^\top \Lambda \beta$. Then

$$E^*[l^*(\theta, \tilde{Z}, \tilde{D})] = l(\theta, Z, \tilde{D}),$$

and therefore $l^*(\theta, \tilde{Z}, \tilde{D})$ is a corrected log-likelihood function. Differentiating with respect to p, β, ψ, we get

$$U_p^*(\theta, \tilde{Z}, \tilde{D}) = \sum_{i=1}^{n} \left\{ \delta_i \left[\frac{1}{p} + \frac{F_0(t_i)}{1 - pF_0(t_i)} \right] - \exp(\tilde{z}_i^\top \beta - \xi_i) \frac{F_0(t_i)}{1 - pF_0(t_i)} \right\}, \quad (2.13)$$

$$U_\beta^*(\theta, \tilde{Z}, \tilde{D}) = \sum_{i=1}^{n} \left\{ \delta_i \tilde{z}_i + (\tilde{z}_i - \sigma_i) \exp(\tilde{z}_i^\top \beta - \xi_i) \ln(1 - pF_0(t_i)) \right\}, \quad (2.14)$$

and

$$\begin{aligned} U_\psi^*(\theta, \tilde{Z}, \tilde{D}) &= \sum_{i=1}^{n} \left\{ \delta_i \left[\frac{1}{\psi} - t_i + \frac{p\psi \exp(-\psi t_i)}{1 - pF_0(t_i)} \right] \right\} \\ &\quad - \sum_{i=1}^{n} \left\{ \exp(\tilde{z}_i^\top \beta - \xi_i) \frac{p\psi \exp(-\psi t_i)}{1 - pF_0(t_i)} \right\}. \end{aligned} \quad (2.15)$$

The corrected maximum likelihood estimator ${\theta^*}^\top = (\psi^*, {\beta^\top}^*, p^*)$ is easily obtained by solving equations (2.13)–(2.15).

2.3. Asymptotic Properties of Parameter Estimators

2.3.1. Proportional hazards Model

In this section, we derive the large sample properties of the parameter estimators β^* and p^* for model (1.11) with additive measurement error (2.3). For Cox PH model (1.11), the consistency of the estimators of β and the cumulative baseline hazard function $\Lambda_0(t, Z)$ has been presented by Andersen and Gill (1982) for the observed time $t \in [0, \tau]$ for any $\tau < \infty$. Moreover, for the i.i.d. case, the asymptotic properties of the above estimators have also been established by Andersen and Gill (1982) for observed time $t \in [0, \infty]$. On the other hand, under model (1.11) with error (2.3) (for the i.i.d. case), Kong and Gu (1999) established the asymptotic properties of the corrected maximum partial likelihood estimator β^* and the corresponding $\Lambda^*(t, \tilde{Z})$ for the observed time $t \in [0, \tau]$ for $\tau < \infty$ under the regularity conditions in Andersen and Gill (1982). The results of Kong and Gu (1999) also hold under the assumptions of Theorem 4.2 (i.i.d. case) in Andersen and Gill (1982). Thus our asymptotic properties can be established by incorporating the results of Andersen and Gill (1982) with those of Kong and Gu (1999).

Theorem 2.1. *Suppose that $(N_i, Y_i, \tilde{Z}_i)$, $i = 1, \ldots, n$ are i.i.d. replicates of $(N, Y, \tilde{Z})$, where $N(t)$ and $Y(t)$ are counting process and predictable process corresponding to an improper survival function $(1 - pF_0(t))^{\exp(z^\top \beta)}$ with $0 < p < 1$, $\tilde{Z}$ is bounded and Σ is positive definite. If either $\tau_G = \infty$ or $\tau_{F_0} \le \tau_G < \infty$, then with probability approaching 1, equation (2.5) has a solution β^* such that $\beta^* \xrightarrow{p} \beta_0$ and $n^{1/2}(\beta^* - \beta_0)$ converges in distribution to a normal distribution with mean zero and a variance-covariance matrix that can be consistently estimated by (2.17) below.*

The next theorem gives the asymptotic properties for the estimator of p.

Theorem 2.2. *Under the conditions of Theorem 2.1, $p^* \xrightarrow{p} p$ as $n \to \infty$, and*

$$n^{1/2}(p^* - p) \xrightarrow{d} N(0, \exp(-2\Lambda_0^*(+\infty, \tilde{Z}))\Sigma^{-1}_{\Lambda_0^*(+\infty,\tilde{Z})}),$$

where $\Sigma_{\Lambda_0^*(+\infty,\tilde{Z})}$ is the asymptotic variance of $\Lambda_0^*(+\infty, \tilde{Z})$ (whose estimation is given by (2.18) below).

2.3.2. Maximum Likelihood Model

For simplicity we consider the case where the covariate z is an univariate random variable.

Theorem 2.3. *Let $(t_k^*, c_k, z_k, \varepsilon_k), k = 1, \ldots, n$, be independent random variables, ε_i are normal distribution random variables with variance σ_i^2. Then under the* *conditions in Lemma 2.3 below and*

$$\sum_{k=1}^{\infty} k^{-2}\left\{2\xi\delta_k^2 + W_k\left[4\xi\delta_k \exp(z_k\beta) + M_k\right]\right\} < \infty,$$

the equations (2.13)–(2.15) each has a root θ^* *which is consistent as* $n \to \infty$ *and* $\theta^{*\top} = (\psi^*, \beta^{*\top}, p^*)$ *is normally distributed with finite variance, where* $\xi = \sigma^2\beta^2/2$, $W_k = \left[\log(1 - pF_0(t_k, \psi))\right]^2$ *and* $M_k = \exp(2z_k\beta + 2\xi) - \exp(2z_k\beta)$.

2.3.3. Proofs

To begin, we state some lemmas.

Lemma 2.1. *Let* $(N_i, Y_i, \tilde{Z}_i)$, $i = 1, \ldots, n$, *be i.i.d. replicates of* $(N, Y, \tilde{Z})$, *where* $N(t)$ *and* $Y(t)$ *are counting process and predictable process corresponding to a survival function* $S(t)$ *with* $E[N(\infty)] < \infty$, $\tilde{Z}$ *is bounded and* Σ *is positive definite. Assume that for each* $\tau < \infty$

$$P(Y(t) = 1, \forall t \leq \tau) > 0. \tag{2.16}$$

Then with probability approaching 1, the score equation for (2.4) has a solution β^* *such that* $\beta^* \xrightarrow{p} \beta_0$ *and* $n^{1/2}(\beta^* - \beta_0)$ *has asymptotically a normal distribution with mean zero and a variance-covariance matrix consistently estimated by*

$$I_n^*(\beta^*, \tilde{Z})^{-1} J_n^*(\beta^*, \tilde{Z}) I_n^*(\beta^*, \tilde{Z})^{-1}, \tag{2.17}$$

where $I_n^*(\beta, \tilde{Z})$ *is the corrected observed information,*

$$J_n^*(\beta^*, \tilde{Z}) = \frac{1}{n}\sum_{i=1}^{n} C_{1i}(\beta^*, \tilde{Z})^{\otimes 2},$$

$$C_{1i}(\beta, \tilde{Z}) = \int_0^\infty \left\{\tilde{z}_i - \frac{S^{(1)}(\beta, s, \tilde{Z})}{S^{(0)}(\beta, s, \tilde{Z})} + \Lambda\beta\right\} dN_i(s)$$

$$- \int_0^\infty \left\{\tilde{z}_i - \frac{S^{(1)}(\beta, s, \tilde{Z})}{S^{(0)}(\beta, s, \tilde{Z})}\right\} \frac{Y_i(s)\exp(\tilde{z}_i^\top\beta)}{S^{(0)}(\beta, s, \tilde{Z})} \frac{d\bar{N}_i(s)}{n},$$

$$S^{(0)}(\beta, t, \tilde{Z}) = \exp(\xi)\left\{S^{(0)}(\beta, t, Z) + \frac{1}{\sqrt{n}}Q_0(\beta, t)\right\},$$

$$S^{(1)}(\beta,t,\tilde{Z}) = \exp(\xi)\left\{S^{(1)}(\beta,t,Z) + \Lambda\beta S^{(0)}(\beta,t,Z)\frac{1}{\sqrt{n}}Q_1(\beta,t)\right\},$$

$$Q_0(\beta,t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i(t)\exp(z_i^\top\beta)\left\{\exp(\varepsilon_i^\top\beta - \xi) - 1\right\},$$

and

$$Q_1(\beta,t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i(t)\exp(z_i^\top\beta)[z_i\exp(\varepsilon_i^\top\beta - \xi) - \{z_i + \Lambda\beta\}].$$

Proof: This is a direct consequence of Theorem 4.2 in Andersen and Gill (1982) together with Theorem 3.1 in Kong and Gu (1999). ∎

Lemma 2.2. *Under the assumptions of Lemma 2.1, $\Lambda_0^*(t,\tilde{Z})$, as defined in (2.9), is a consistent estimator of $\Lambda_0(t,\tilde{Z})$. Also, on $[0,\infty]$, $n^{1/2}\{\Lambda_0^*(t,\tilde{Z}) - \Lambda_0(t,\tilde{Z})\}$ can be expressed as a sum of independent random processes and so it converges to a Gaussian process with mean zero and a variance function that can be consistently estimated by*

$$\frac{1}{n}\sum_{i=1}^{n}\left\{nH^*(\beta^*,t)I^*(\beta^*,\tilde{Z})^{-1}C_{1i}(\beta^*,\tilde{Z},t) + C_{2i}(\beta^*,\tilde{Z},t)\right\}^{\otimes 2}, \tag{2.18}$$

where

$$H^*(\beta^*,t) = -\int_0^t \frac{\exp(\xi)}{S^{(0)}(\beta^*,s,\tilde{Z})}\left\{\frac{S^{(1)}(\beta^*,s,\tilde{Z})}{S^{(0)}(\beta^*,s,\tilde{Z})} + \Lambda\beta^*\right\}\frac{d\bar{N}_i(s)}{n},$$

$C_{1i}(\beta^*,X,t)$ is defined in lemma 2.1 with integration interval $[0,\infty]$ being replaced with $[0,t]$, and

$$C_{2i}(\beta^*,\tilde{Z},t) = \int_0^t \frac{\exp(-\xi)}{S^{(0)}(\beta^*,s,\tilde{Z})}\left\{dN_i(s) - \frac{Y_i(s)\exp(\tilde{z}_i^\top\beta^*)}{S^{(0)}(\beta^*,s,\tilde{Z})}\frac{d\bar{N}_i(s)}{n}\right\}.$$

Proof: This is straightforward by combining Theorem 4.2 in Andersen and Gill (1982) with Theorem 4.1 in Kong and Gu (1999). ∎

Lemma 2.3. *Let $l^*(\theta,\tilde{Z},\tilde{D})$ and $l(\theta,Z,\tilde{D})$ be, respectively, corrected log-likelihood function and log-likelihood function. Suppose that $l^*(\theta,\tilde{Z},\tilde{D})$ and $l(\theta,Z,\tilde{D})$ are* *differentiable in an open convex subset of a parameter space that contains the "true" parameter* θ_0,

$$\sum_{k=1}^{\infty} k^{-2} Var^* \left\{ l^* \left(\theta, \tilde{z}_k, \tilde{D}_k \right) \right\} < \infty, \tag{2.19}$$

θ is identifiable and the $\tilde{D}_i$, $i = 1, 2, \ldots,$ are mutually independent. Then the equation $U^*(\theta, \tilde{Z}, \tilde{D}) = 0$ has a root θ^* which is consistent as $n \to \infty$. Furthermore, under certain regularity conditions on $U^*(\theta, \tilde{Z}, \tilde{D})$, the unconditional distribution of θ^* is asymptotically normal with mean θ and variance

$$I^+(\theta, Z) E^+ \left[Var^* \left\{ U^*(\theta, \tilde{Z}, \tilde{D}) \right\} \right] I^+(\theta, Z)^{-1} + I^+(\theta, Z)^{-1},$$

where $I^+(\theta, Z) = E^+[I(\theta, Z, \tilde{D})]$, the expectations are with respective to the random variable $\tilde{D}$, and $Var^*(\cdot)$ denotes the conditional variance given Z and $\tilde{D}$.

Proof: See Nakamura (1990). ∎

Proof of Theorem 2.1. Since $p < 1$, we have $E[N(\infty)] < \infty$. Note that $N(t)$ and $Y(t)$ correspond to an improper distribution $F(t) = 1 - (1 - pF_0(t))^{\exp(z^\top \beta)}$. Similar to the proof of Lemma 2.1 (see also Andersen and Gill, 1982), it suffices to show that (2.16) holds either for $\tau < \tau_{F_0}$ or for $\tau < \infty$.

Since $F(t)$ is improper, $P(T^* > \tau) = 1 - F(\tau) > 0$ for any $\tau < \infty$. Furthermore, if $\tau_G = \infty$, then $P(c \geq \tau) = 1 - G(\tau-) > 0$ for any $\tau < \infty$. Hence

$$P(Y(t) = 1, \forall t \leq \tau) = P(T^* \wedge c \geq \tau) = P(T^* \geq \tau) P(c \geq \tau) > 0. \tag{2.20}$$

for any $\tau < \infty$. On the other hand, if $\tau_{F_0} \leq \tau_G < \infty$, then $P(c \geq \tau) = 1 - G(\tau-) > 0$ for any $\tau < \tau_{F_0} \leq \tau_G$, which shows that (2.20) holds for $\tau < \tau_{F_0}$, and so does (2.16). This completes the proof. ∎

Proof of Theorem 2.2. This follows from the well known Slutsky theorem and delta-method together with Lemma 2.2. ∎

Proof of Theorem 2.3. It is sufficient to prove condition (2.19) in Lemma 2.3. Note that

$$E^*\{\exp(\tilde{z}\beta)\} = \exp(z\beta + \xi), \tag{2.21}$$

$$E^*\{\tilde{z}\exp(\tilde{z}\beta)\} = (z + \sigma^2\beta)\exp(z\beta + \xi), \tag{2.22}$$

and

$$E^*\{\tilde{z}^2\exp(\tilde{z}\beta)\} = \{\sigma^2 + (z + \sigma^2\beta)^2\}\exp(z\beta + \xi). \tag{2.23}$$

Then Theorem 2.3 follows. ∎

Before closing this section, we consider variance estimation for the estimators based on the full likelihood approach in Section 2.2.2. Nakamura (1990) gave two asymptotically equivalent estimators of the variance of θ^* (cf. Nakamura, 1990, p.130). For simplicity, the variance of θ^* can be estimated by

$$\hat{V}ar(\theta^*) = I^*(\theta^*, \tilde{Z}, \tilde{D})^{-1} S(\theta^*, \tilde{Z}, \tilde{D}) I^*(\theta^*, \tilde{Z}, \tilde{D})^{-1},$$

where $I^*(\theta^*, \tilde{Z}, \tilde{D})$ is the corrected observed information and

$$S(\theta^*, \tilde{Z}, \tilde{D}) = \sum_{i=1}^{n} U^*(\theta, \tilde{z}_k, \tilde{D}_k) U^*(\theta, \tilde{z}_k, \tilde{D}_k)^\top.$$

2.4. Simulation Results

In this section, we report some simulations results to assess the accuracy of p^*, β^* and ψ^* when the error magnitude of covariates varies for the case of one-dimensional covariate. We chose the sample size n to be $n = 200, 400, 1000$. Let z be a set of n uniform random numbers in $(0, 12^{1/2})$ so that the standard deviation of covariate z is 1. For a given β and each covariate value z, a failure time t is generated from the survival function $(1 - pF_0(t))^{\exp(z^\top\beta)}$ with the susceptible proportion $p =$ 0.9, 0.8, 0.6, 0.4 and an exponential $F_0(t)$ with hazard rate $\lambda_0(t) = 0.058$. To save space, Tables 2.1–2.8 present only part of the simulation results for sample size

$n = 1000$, 400 and 200, and $p = 0.9$. The results for other values of n and p are similar and available from the authors. For a fixed σ and a measurement error ε generated from $N(0,1)$, an observed covariates $\tilde{z}$ is the sum of z and $\sigma\varepsilon$. Given n independent pairs of $(\tilde{z}, t)$, we derive the corrected maximum partial likelihood estimator (CMPLE) and the corrected maximum likelihood estimator (CMLE). This procedure was repeated 500 times, and the averages of the CMPLEs and CMLEs were calculated.

We preformed a simulation to fit the two models with censoring time independently generated from uniform distribution on $[0, d]$, where d is a positive number. For a fixed β, d can be chosen such that the percentage of censoring is a desired value. We choose $d = 100$ for simplicity. In Tables 2.1–2.8 below, $M(\cdot)$ represents the mean of the naive estimator (denoted by β_x, p_x and ψ_x) or the corrected estimator (β^*, p^* and ψ^*). The standard deviation of the estimators is denoted by $SD(\cdot)$ and the average censoring proportion of the failure data for each combination β and σ is denoted by r.

First a simulation was performed to compare the corrected estimators with corresponding naive estimators of p and β. The results of simulation are listed in Tables 2.1 and 2.2, from which we can find that the naive estimations are biased, and the bias are larger when σ are increasing.

Next, we performed simulations to choose a reasonable sample size. The sample size n was chosen to be $200, 400, 1000$, and a recommended sample size was given.

We must note that the CMPLE and CMLE may not always exist because the derivative of the corrected score function is not always negative in a neighborhood of the true θ. This usually happens when the measurement error is large. In our simulation study, we find that the CMPLE often fails to converge when $\beta = 2.0$ and $\sigma \geq 0.3$, and the CMLE often fails to converge when $\beta = 2.0$ and $\sigma \geq 1$.

The simulation results indicate that the CMPLE and CMLE perform quite well for the given sample sizes. We see from these tables that the CMPLE and CMLE are more accurate when the measurement error is small and the sample size is increasing (in Tables 2.3–2.8). We also see that for the given sample sizes, the estimators of β are biased upwards and the estimations of p are biased downwards for semiparametric model, whereas in maximum likelihood estimation, the estimators of β and p are biased upwards. The results also indicate that the CMPLE is more robust than the CMLE for the selected sample sizes and $|\sigma\beta|$. For example, when sample size $n = 200$, the CMLE cannot be accepted for their large biases (cf. Table 2.6), while the CMPLE as listed in Table 2.3 are reasonably accurate. When the sample size is large, the procedures produce accurate estimators. Overall, the CMPLE preforms better than the CMLE and sample size $n = 400$ is recommended for the CMPLE. Finally it should be mentioned that the second order Taylor expansion can be considered to obtain more accurate approximate results (see Nakamura 1992).

Table 2.1. CMPLE and corresponding naive estimator with $n = 400$

β	σ	$M(\beta)$	$M(p)$	$M(\beta_x)$	$M(p_x)$
0.1	0.1	0.0998	0.8931	0.0995	0.8903
0.2	0.2	0.2006	0.8902	0.1916	0.8944
0.3	0.3	0.3016	0.8914	0.2752	0.9005
0.4	0.4	0.4048	0.8879	0.3435	0.9006
0.5	0.5	0.4957	0.8884	0.3842	0.9182
0.5	0.6	0.5068	0.8889	0.3559	0.9268
0.6	0.5	0.6053	0.8876	0.4599	0.9210
0.7	0.5	0.7059	0.8828	0.5117	0.9243
0.8	0.4	0.8039	0.8824	0.6435	0.9170
0.9	0.3	0.9044	0.8825	0.7777	0.9054
1.0	0.3	1.0035	0.8813	0.8579	0.9064
1.0	0.5	1.0158	0.8835	0.7155	0.9290

Table 2.2. CMLE and corresponding naive estimator with $n = 400$

β	σ	$M(\beta)$	$M(p)$	$M(\psi)$	$M(\beta_x)$	$M(p_x)$	$M(\psi_x)$
0.1	0.1	0.0975	0.9012	0.0584	0.0111	0.9022	0.0581
0.2	0.2	0.2001	0.9026	0.0582	0.1949	0.9029	0.0582
0.3	0.3	0.3022	0.9025	0.0583	0.2752	0.9005	0.0590
0.4	0.4	0.3999	0.9036	0.0586	0.3435	0.9006	0.0612
0.5	0.5	0.5104	0.9044	0.0572	0.3842	0.9182	0.0623
0.5	0.6	0.5072	0.9016	0.0578	0.3559	0.9268	0.0645
0.6	0.5	0.6058	0.9085	0.0579	0.4599	0.9210	0.0658
0.7	0.5	0.7091	0.9048	0.0579	0.5412	1.0398	0.0737
0.8	0.4	0.8084	0.9049	0.0582	0.6815	1.0040	0.0693
0.9	0.3	0.9058	0.9081	0.0577	0.8157	0.9496	0.0658
1.0	0.3	1.0043	0.9082	0.0570	0.9123	0.9120	0.0663
1.0	0.5	1.0075	0.8835	0.0557	0.7155	0.9290	0.0698

Table 2.3. CMPLEs of β and p and their standard deviations with $n = 200$

r	β	σ	$M(\beta^*)$	$M(p^*)$	$SD(\beta^*)$	$SD(p^*)$
0.79	0.1	0.1	0.1036	0.8854	0.0830	0.0480
0.83	0.2	0.2	0.1935	0.8856	0.0803	0.0491
0.85	0.3	0.3	0.3034	0.8845	0.0860	0.0488
0.82	0.4	0.4	0.4032	0.8818	0.0996	0.0594
0.89	0.5	0.5	0.5066	0.8763	0.1045	0.0563
0.89	0.5	0.6	0.5098	0.8765	0.1180	0.0572
0.91	0.6	0.5	0.6194	0.8745	0.1142	0.0599
0.92	0.7	0.5	0.7147	0.8708	0.1255	0.0602
0.93	0.8	0.4	0.8249	0.8678	0.1187	0.0660
0.93	0.9	0.3	0.9137	0.8624	0.1114	0.0646
0.94	1.0	0.3	1.0123	0.8582	0.1235	0.0669
0.94	1.0	0.5	1.0158	0.8687	0.1568	0.0720

Table 2.4. CMPLEs of β and p and their standard deviations with $n = 400$

r	β	σ	$M(\beta^*)$	$M(p^*)$	$SD(\beta^*)$	$SD(p^*)$
0.79	0.1	0.1	0.0998	0.8931	0.0561	0.0336
0.83	0.2	0.2	0.2066	0.8902	0.0595	0.0357
0.85	0.3	0.3	0.3016	0.8914	0.0589	0.0335
0.82	0.4	0.4	0.4048	0.8879	0.0618	0.0389
0.89	0.5	0.5	0.4957	0.8884	0.0756	0.0397
0.89	0.5	0.6	0.5086	0.8889	0.0830	0.0430
0.91	0.6	0.5	0.6053	0.8876	0.0785	0.0439
0.92	0.7	0.5	0.7059	0.8828	0.0876	0.0464
0.93	0.8	0.4	0.8039	0.8824	0.0883	0.0497
0.93	0.9	0.3	0.9044	0.8825	0.0759	0.0474
0.94	1.0	0.3	1.0035	0.8813	0.0836	0.0480
0.94	1.0	0.5	1.0158	0.8835	0.1089	0.0552

Table 2.5. CMPLEs of β and p and their standard deviations with $n = 1000$

r	β	σ	$M(\beta^*)$	$M(p^*)$	$SD(\beta^*)$	$SD(p^*)$
0.79	0.1	0.1	0.0988	0.8964	0.0352	0.0208
0.83	0.2	0.2	0.1989	0.8967	0.0384	0.0216
0.85	0.3	0.3	0.3022	0.8935	0.0370	0.0232
0.82	0.4	0.4	0.4021	0.8934	0.0430	0.0264
0.89	0.5	0.5	0.5034	0.8943	0.0449	0.0272
0.89	0.5	0.6	0.5037	0.8922	0.0481	0.0298
0.91	0.6	0.5	0.6012	0.8936	0.0235	0.0321
0.92	0.7	0.5	0.7031	0.8934	0.0521	0.0329
0.93	0.8	0.4	0.8041	0.8925	0.0533	0.0334
0.93	0.9	0.3	0.9034	0.8899	0.0493	0.0353
0.94	1.0	0.3	1.0039	0.8916	0.0529	0.0346
0.94	1.0	0.5	1.0049	0.8912	0.0759	0.0374

Table 2.6. CMLEs of β and p and their standard deviations with $n = 200$

r	β	σ	$M(\beta^*)$	$M(p^*)$	$M(\psi^*)$	$SD(\beta^*)$	$SD(p^*)$	$SD(\psi^*)$
0.79	0.1	0.1	0.1016	0.9507	0.0585	0.0847	1.1541	0.0093
0.83	0.2	0.2	0.1934	0.9512	0.0593	0.0797	1.0899	0.0090
0.85	0.3	0.3	0.3018	0.9156	0.0579	0.0950	0.1058	0.0103
0.82	0.4	0.4	0.4024	0.9721	0.0584	0.1056	0.5518	0.0147
0.89	0.5	0.5	0.5035	1.1523	0.0583	0.1151	2.6054	0.0168
0.89	0.5	0.6	0.5102	1.0486	0.0576	0.1196	1.2716	0.0169
0.91	0.6	0.5	0.6190	1.0615	0.0559	0.1234	1.0380	0.0189
0.92	0.7	0.5	0.7143	1.0705	0.0561	0.1121	2.3077	0.0163
0.93	0.8	0.4	0.8215	1.0519	0.0558	0.1209	1.1718	0.0243
0.93	0.9	0.3	0.9138	1.1844	0.0556	0.0959	2.7637	0.0163
0.94	1.0	0.3	1.0087	1.1342	0.0563	0.1167	3.4083	0.0226
0.94	1.0	0.5	1.0135	0.9882	0.0559	0.1359	0.3046	0.0208

Table 2.7. CMLEs of β and p and their standard deviations with $n = 400$

β	σ	$M(\beta^*)$	$M(p^*)$	$M(\psi^*)$	$SD(\beta^*)$	$SD(p^*)$	$SD(\psi^*)$
0.1	0.1	0.0975	0.9012	0.0584	0.0552	0.0552	0.0059
0.2	0.2	0.2001	0.9026	0.0582	0.0576	0.0576	0.0065
0.3	0.3	0.3022	0.9025	0.0583	0.0581	0.0334	0.0068
0.4	0.4	0.3999	0.9036	0.0586	0.0596	0.0371	0.0069
0.5	0.5	0.5104	0.9044	0.0572	0.0570	0.2632	0.0113
0.5	0.6	0.5072	0.9016	0.0578	0.0760	0.0760	0.0090
0.6	0.5	0.6058	0.9085	0.0579	0.0776	0.0735	0.0125
0.7	0.5	0.7091	0.9048	0.0579	0.0830	0.5547	0.0113
0.8	0.4	0.8084	0.9049	0.0582	0.0827	0.1824	0.0152
0.9	0.3	0.9058	0.9081	0.0577	0.0741	0.7566	0.0185
1.0	0.3	1.0043	0.9082	0.0570	0.0667	0.5927	0.0100
1.0	0.5	1.0075	0.8835	0.0557	0.0821	0.4197	0.0115

Table 2.8. CMLEs of β and p and their standard deviations with $n = 1000$

β	σ	$M(\beta^*)$	$M(p^*)$	$M(\psi^*)$	$SD(\beta^*)$	$SD(p^*)$	$SD(\psi^*)$
0.1	0.1	0.0986	0.9015	0.0581	0.0348	0.0188	0.0036
0.2	0.2	0.1994	0.9016	0.0580	0.0349	0.0198	0.0038
0.3	0.3	0.3020	0.8998	0.0582	0.0360	0.0201	0.0040
0.4	0.4	0.4018	0.9022	0.0579	0.0380	0.0224	0.0045
0.5	0.5	0.5043	0.9029	0.0576	0.0419	0.0252	0.0490
0.5	0.6	0.5047	0.8999	0.0577	0.0403	0.0255	0.0049
0.6	0.5	0.6001	0.9012	0.0578	0.0402	0.0256	0.0050
0.7	0.5	0.7046	0.9038	0.0575	0.0454	0.0421	0.0056
0.8	0.4	0.8034	0.9044	0.0577	0.0431	0.0311	0.0053
0.9	0.3	0.9029	0.9044	0.0575	0.0400	0.0379	0.0053
1.0	0.3	1.0036	0.9051	0.0575	0.0403	0.0339	0.0053
1.0	0.5	1.0021	0.9066	0.0580	0.0544	0.0443	0.0068

2.5. Concluding Remarks

In this Chapter we also investigate an improper PH model to analyze survival data with long-term survivors. We applied some martingale techniques to tackle the problem of measurement error in time-independent covariates. An accurately corrected maximum likelihood score and an approximately corrected partial likelihood score are used to obtain the corrected maximum likelihood estimators (CMLEs) and the corrected maximum partial likelihood estimators (CMPLEs), respectively. The asymptotic properties are derived based on the works of Kong and Gu (1999) and Nakamura (1990).

The simulation study indicates that the proposed model and estimation procedures produce efficient estimators.

In survival analysis, a frequent objective is to characterize the relationship between survival times and covariates. Recently, modeling the event time process using longitudinal data as time-dependent covariates in a proportional hazards model is a standard framework (such as, Wulfsohn and Tsiatis, (1997) and Tsiatis and Davidian (2001)). Other than the above study in chapters 1-2 where the covariates are time-independent, longitudinal data study where repeated measurements on a continuous response, an observation on a possible censored time-to-event, and additional covariates information are collected on each participant may allow "intermittent" and "error" measurements and also differs from the time-dependent covariates study where the values of all subjects in the risk at any failure points are known (cf. Sec.6.3, Kalbfeisch and Prentice (2002)). This issue is worth for further studies.

Finally we can consider some more complex structure of measurement error by extending or modifying our model in two aspects. One is to extend our model $S(t) = [1 - pF_0(t)]^{\exp(z^\top \beta)}$ to more general cases, such as a nonparametric model $\psi(x)$, or a seimparametric partially linear model $\psi(x) + z^\top \beta$, for the covariates; and measurement error may be added to $\psi(x)$. We expect our statistical inference would still work for the above models if $\psi(x)$ is modeled by local polynomials. This method for nonlinear models was considered by many authors such as, Liang *et al.* (1999), Carroll *et al.* (1999), Iturria *et al.* (1999), Wang *et al.* (1998), among others. Another direction is to consider multiplicative measurement error instead of additive ones, such as $x\varepsilon$ with some covariates x.We have proposed a transform $\log\{-\log(\cdot)\}$ to the model in Chapter 4. Based on such transform some results for multiplicative measurement error, such as those in Iturria *et al.* (1999) and Eckert *et al.* (1997), may be considered together with our statistical inference in Chapter 4.

Chapter 3

Proportional Hazards Model for Survival Data with Long-term Survivors: Interior and Boundary Cases

3.1. Introduction

Suppose that $F_0(t, \psi)$ is a proper cumulative distribution function (cdf) with parameter (vector) ψ and $p \in (0, 1]$. Then the survival function of the standard cure rate (SCR) model can be written as

$$S(t) = 1 - pF_0(t, \psi). \tag{3.1}$$

Recently, an alternative cure rate model $S(t) = \exp\{-\zeta G_0(t, \psi)\}$ with $\zeta \in (0, \infty)$ and $G_0(t, \psi)$ to be a proper cdf, termed as the bounded cumulative hazard (BCH) model, was introduced by Yakovlev *et al.* (1993) and subsequently investigated by many authors including Yakovlev (1994), Chen *et al.* (1999) and Tsodikov *et al.* (2003), among others. The SCR model has a mathematical relationship with the BCH model such that any SCR model can be expressed as the BCH model (Chen *et al.*, 1999), and vice versa. If allowing for covariates, the SCR model

$$S(t) = 1 - pF_0(t, \psi), \quad \text{with} \quad p = \frac{\exp(z^\top \beta)}{1 + \exp(z^\top \beta)}, \tag{3.2}$$

was proposed by Maller and Zhou (1996), and the BCH model with $\theta = \exp(\beta^\top z)$ has been discussed by Tsodikov (2003), where β is a vector of regression coefficients, z is a vector of covariates and $^\top$ denotes the transpose. However, Chen *et al.* (1999) found that the SCR model in (3.2) has some drawbacks. First, it does not have a proportional hazards structure, which is a desirable property for survival models.

Second, it yields improper posterior for many types of noninformative improper prior, including the uniform prior for the regression coefficients. Maller and Zhou (1996) raised an idea to incorporate Cox proportional hazards (PH) models with mixture models and proposed an intensity function of the form

$$\alpha(t) \exp\left(\beta^\top z\right) \quad \text{with} \quad \alpha(t) = \frac{p f_0(t, \psi)}{1 - p F_0(t, \psi)}, \tag{3.3}$$

where $\alpha(t)$ is a baseline intensity function with respect to an improper distribution function $pF_0(t, \psi)$ and $f_0(t, \psi)$ is the density function associated with $F_0(t, \psi)$. This improper PH model, having been investigated in Chapters 1-2 with $p \in (0, 1)$, can overcome the drawbacks mentioned above. It should be noted that many aspects for this improper PH model, such as the motivation and the research approach to be utilized in this chapter, are different from those of the improper PH model proposed by Tsodikov (2003).

For the above models (3.1)-(3.3), our main interests lie in two aspects: one is to estimate the parameters $\theta^\top = (\psi, \beta^\top, p)$, where ψ is a parameter associated with $F_0(t, \psi)$ to be estimated, β is a coefficient vector of the covariates and p is related to the proportion of "susceptibles" in the population. The other is, when the true parameters are allowed to be on the boundary of the parameter space, i.e., $p_0 = 1$ (which is referred to as the *boundary case*), how to obtain the asymptotic distributions of the MLEs and the likelihood ratio statistic used to test $H_0 : p_0 = 1$. Following the procedure of Self and Liang (1987), Fahrmeir and Kaufmann (1985), Zhou and Maller (1995) investigated model (3.1) with common p and ψ for all individuals (i.e., with no covariate information for p and ψ). Model (3.2) were also discussed by Ghitany, Maller and Zhou (1994), when F_0 is exponential with hazard rate $\psi = \exp(\gamma^\top x)$ and $p = \exp(\beta^\top z)/(1+\exp(\beta^\top z))$. Vu, Maller and Zhou (1998) further extended the work to the case where the failure time distribution belongs to an exponential family.

In this chapter, following the procedure of Vu, Maller and Zhou (1998), we study model (3.1). We will establish the existence and consistency of the MLEs, say $\hat{\theta}_n$, which may lie on the boundary of the parameter space, and find the asymptotic distributions of $\hat{\theta}_n$ and the related likelihood ratio statistic. However, our model is not limited to an exponential family anymore. Moreover, Vu, Maller and Zhou (1998) assumed a common covariate effect on p for all individuals (in other words, there is no covariate effect on the proportion of "immunes"), this reduces to the case of Zhou and Maller (1995). In this chapter, we allow different effects of covariates on p between individuals and extend the works of Vu, Maller and Zhou (1998) beyond the exponential family of failure distributions.

In Section 3.2, we first specify the model and then make some assumptions. Preliminary and main results for parameter estimation based on the likelihood method are discussed in Section 3.3, while Section 3.4 deals with the likelihood ratio statistics. The proofs for the main results are given in Sections 3.5–3.6.

3.2. Model Specification and Assumptions

Following the notations in Section 1.2 of Chapter 1, we let $t_1, \ldots, t_n$ be a sample of (possibly censored) survival data with corresponding covariate vectors $z_1, \ldots, z_n$. If the baseline distribution $F_0(t, \psi)$ is parameterized with parameter vector ψ, then by (1.6), the likelihood function can be written as

$$\begin{aligned} L_f = L_f(\psi, \beta^\top, p) &= \prod_{i=1}^{n} \alpha_i(t_i, z_i)^{\delta_i} \exp\left\{-\int_0^{t_i} \alpha_i(y, z_i) dy\right\} \\ &= \prod_{i=1}^{n} [\alpha_0(t_i) \exp(\beta^\top z_i)]^{\delta_i} \exp\left\{-\int_0^{t_i} \alpha_0(s) \exp(\beta^\top z_i) ds\right\} \\ &= \prod_{i=1}^{n} \left[\frac{p f_0(t_i, \psi)}{1 - pF_0(t_i, \psi)} \exp(\beta^\top z_i)\right]^{\delta_i} [1 - pF_0(t_i, \psi)]^{\exp(\beta^\top z_i)}. \qquad (3.4) \end{aligned}$$

For example, if $F_0(t, \psi)$ is an exponential distribution, then $F_0(t, \psi) = 1 - \exp(\psi t)$. Let $l_i(\theta)$ be the contribution to the log-likelihood $l(\theta) = \log L_f$ by individual i.

Then $l(\theta) = \sum_{i=1}^{n} l_i(\theta)$ and

$$l_i(\theta) = \delta_i\{\log(p) + \log(\psi) - \psi t_i + \beta^\top z_i + [\exp(\beta^\top z_i) - 1]\log(1 - pF_0(t_i, \psi))\}$$
$$+ (1 - \delta_i)\exp(\beta^\top z_i)\log(1 - pF_0(t_i, \psi)). \tag{3.5}$$

We assume that the parameter spaces for ψ and β are open sets $\Theta_\psi = (0, \infty)$ and $\Theta_\beta = (-b, b)$, $b > 0$, respectively. Denote the "true" parameter value of θ by $\theta_0^\top = (\psi_0, \beta_0^\top, p_0)$. We consider two cases in this chapter:

(1) **The interior case:** $\theta_0 \in \Theta = \Theta_\psi \otimes \Theta_\beta \otimes \Theta_p$, where $\Theta_p = (0, 1)$, then θ_0 is an interior point of Θ.

(2) **The boundary case:** $\theta_0 \in \Theta = \Theta_\psi \otimes \Theta_\beta \otimes \Theta_p$, where $\Theta_p = (0, 1]$, then θ_0 may be an boundary point of Θ.

For the interior and boundary cases, the maximum likelihood estimators of the parameters $(\psi, \beta^\top, p)$ can then be obtained by maximizing the $l(\theta)$ with respected to $\theta = (\psi, \beta^\top, p)$.

Define the "link" functions

$$\eta_i = \exp(k_i(\beta)), \quad k_i(\beta) = \beta^\top z_i, \quad k_{i0} = k_i(\beta_0) = \beta_0^\top z_i. \tag{3.6}$$

From (3.5)-(3.6) we can calculate the vector derivatives $\partial l(\theta)/\partial\psi$, $\partial l(\theta)/\partial\beta$ and $\partial l(\theta)/\partial p$, and then write

$$S_n(\theta) = \frac{\partial l(\theta)}{\partial\theta} = \begin{bmatrix} \frac{\partial l(\theta)}{\partial\psi} & \frac{\partial l(\theta)}{\partial\beta} & \frac{\partial l(\theta)}{\partial p} \end{bmatrix}^\top = \sum_{i=1}^{n} X_i S_i(\theta), \tag{3.7}$$

where

$$S_i(\theta) = \begin{bmatrix} s_{i1} \\ s_{i2} \\ s_{i3} \end{bmatrix} = \begin{bmatrix} s_{i1}(\theta) \\ s_{i2}(\theta) \\ s_{i3}(\theta) \end{bmatrix} = \begin{bmatrix} \frac{\partial l_i(\theta)}{\partial\psi} & \frac{\partial l_i(\theta)}{\partial\beta} & \frac{\partial l_i(\theta)}{\partial p} \end{bmatrix}^\top$$

and

$$X_i = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & & & 0 \\ \vdots & & z_i & & \vdots \\ & & & \ddots & \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{(q+2)\times 3} \tag{3.8}$$

are non-stochastic matrices. Similarly, we obtain the minus second derivative matrix of $l(\theta)$ as

$$F_n(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2} = -\begin{bmatrix} \frac{\partial^2 l(\theta)}{\partial \psi^2} & \frac{\partial^2 l(\theta)}{\partial \psi \partial \beta} & \frac{\partial^2 l(\theta)}{\partial \psi \partial p} \\ \frac{\partial^2 l(\theta)}{\partial \beta \partial \psi} & \frac{\partial^2 l(\theta)}{\partial \beta^2} & \frac{\partial^2 l(\theta)}{\partial \beta \partial p} \\ \frac{\partial^2 l(\theta)}{\partial p \partial \psi} & \frac{\partial^2 l(\theta)}{\partial p \partial \beta} & \frac{\partial^2 l(\theta)}{\partial p^2} \end{bmatrix} = \sum_{i=1}^{n} X_i \mathcal{F}_i(\theta) X_i^\top, \tag{3.9}$$

where each $\mathcal{F}_i(\theta)$ is a 3×3 symmetric random matrix with the (r, s)-element equal to $f_i^{rs}(\theta) = -\partial s_{ir}(\theta)/\partial \xi_{is}$, $r, s = 1, 2, 3$, and $\xi_i = (\xi_{i1}, \xi_{i2}, \xi_{i3}) = (\psi, k_i, p)$, where $k_i = \beta^\top z_i$. The information matrix D_n of $l(\theta)$, which has order $(q + 2) \times (q + 2)$, is given by

$$D_n = E[\mathcal{F}_n(\theta_0)] = \sum_{i=1}^{n} X_i \mathcal{D}_i X_i^\top,$$

where

$$\mathcal{D}_i = \begin{bmatrix} d_i^{11} & d_i^{12} & d_i^{13} \\ d_i^{21} & d_i^{22} & d_i^{23} \\ d_i^{31} & d_i^{32} & d_i^{33} \end{bmatrix} \tag{3.10}$$

are 3×3 symmetries with elements $d_i^{rs} = E[f_i^{rs}(\theta_0)]$, $r, s = 1, 2, 3$.

Provided that the expectations in (3.10) are finite under regularity conditions, the general likelihood theory (Cox and Hinkley, 1974, pp. 107-108) suggests that

$$E[S_n(\theta_0)] = 0 \quad \text{and} \quad E[S_n(\theta_0) S_n^\top(\theta_0)] = E[F_n(\theta_0)] = D_n. \tag{3.11}$$

Define

$$\tau_- = \inf \{t > 0 : F_0(t, \psi) > 0\} \quad \text{and} \quad \tau_+ = \sup \{t > 0 : F_0(t, \psi) < 1\}$$

to be the left and right extremes of $F_0(t, \psi)$.

We make the following assumptions as regularity conditions, which place restrictions on the covariates and the relation between the censoring and survival distributions. For $A > 0$, define

$$N_n(A) = \left\{ \theta \in \Theta : (\theta - \theta_0)^\top D_n (\theta - \theta_0) \le A^2 \right\}. \tag{3.12}$$

Assumptions F:

(F1) For each i and $\psi \in (0, \infty)$, $\Pr(c_i \geq t_i^* > \tau_-) > 0$, i.e., $F_0(\tau_{G-}, \psi) > 0$, where τ_{G-} is the left extreme of the censoring distribution G.

(F2) The matrix $\sum_{i=1}^n z_i z_i^\top$ is positive definite for some $n \geq k$.

(F3) $\lim_{n\to\infty} \sum_{i=1}^n M_i \{\mathrm{tr}(X_i D_n^{-1} X_i^\top)\}^{3/2} = 0$, where M_i is defined in Lemma 3.4 (cf. (3.32) below).

(F4) For each i, $\Pr(c_i < \tau_+) = 1$, i.e., $F_0(\tau_{G+}, \psi) < 1$, where τ_{G+} is the right extreme of the censoring distribution G.

Lemma 3.2 below will show that D_n is positive definite for large n under Assumptions (F1)-(F3).

Remark 3.1. The condition $\Pr(c_i \geq t_i^* > \tau_-) > 0$ in (F1) ensures that *uncensored* observations will be observed with positive probability, and represents a minimal requirement that "follow-up" is sufficient in the sample. In our model, if $F_0(t, \psi)$ is exponential, then (F1) is satisfied provided that the censoring distribution G does not degenerate at 0. Condition (F2) simply ensures that the covariates do not degenerate to a lower dimensional subspace for a large sample. Condition (F3) is a "uniform asymptotic negligibility" type of requirement on the covariates which also incorporates some interplay between censoring and survival distribution. (F4) is natural in the boundary case since no $c_i > \tau+$ can be observed when $p_0 = 1$. In practice, the censoring random variable c_i will be bounded, so that $\Pr(c_i < \tau_+) = 1$ holds for any $F(t, \psi)$ with $\tau_+ = \infty$.

Remark 3.2. As a consequence of assumption (F4), if G_i attributes no mass to τ_+ or to large points if $\tau_+ < \infty$, then the log-likelihood will be finite a.s.

Remark 3.3. Maller and Zhou (2002) pointed out that (F4) can be replaced by Assumption (F5) below when $F_0(t, \psi)$ is exponential.

(F5) For some $\zeta > 0$, $E[e^{(\psi_0+\zeta)c}] < \infty$.

Note that in our model, $\tau_+ = \infty$, so that (F4) implies $\tau_{G+} < \infty$. Under (F5), on the other hand, we allow $\tau_{G+} \le \infty$.

3.3. Preliminary Results

3.3.1. Preliminary Results

We state some preliminary results and begin with some expressions for the first and second derivatives of $l(\theta)$.

Lemma 3.1. *We have the following expressions for $s_{ir}(\theta)$ and $f_i^{rs}(\theta)$ $(r, s = 1, 2, 3)$ (see (3.5) and (3.9)).*

$$
\begin{aligned}
s_{i1}(\theta) &= \delta_i \left\{ \frac{1}{\psi} - t_i - [\eta_i - 1] \frac{p t_i [1 - F_0(t_i, \psi)]}{1 - pF_0(t_i, \psi)} \right\} - \frac{(1-\delta_i)\eta_i p t_i [1 - F_0(t_i, \psi)]}{1 - pF_0(t_i, \psi)};\\
s_{i2}(\theta) &= \delta_i \left\{1 + \eta_i \log(1 - pF_0(t_i, \psi))\right\} + (1-\delta_i)\eta_i \log(1 - pF_0(t_i, \psi));\\
s_{i3}(\theta) &= \delta_i \left\{ \frac{1}{p} - \frac{(\eta_i - 1)F_0(t_i, \psi)}{1 - pF_0(t_i, \psi))} \right\} - \frac{(1-\delta_i)\eta_i F_0(t_i, \psi)}{1 - pF_0(t_i, \psi)};\\
f_i^{11}(\theta) &= \frac{\delta_i}{\psi^2} - \{(\eta_i - 1)\delta_i + (1-\delta_i)\eta_i\} \frac{p(1-p)t_i^2[1 - F_0(t_i, \psi)]}{(1 - pF_0(t_i, \psi))^2};\\
f_i^{22}(\theta) &= -\{\delta_i + (1-\delta_i)\} \log(1 - pF_0(t_i, \psi));\\
f_i^{21}(\theta) &= \{\delta_i + (1-\delta_i)\} \frac{\eta_i p t_i [1 - F_0(t_i, \psi)]}{1 - pF_0(t_i, \psi)};\\
f_i^{33}(\theta) &= \frac{\delta_i}{p^2} + \{(\eta_i - 1)\delta_i + (1-\delta_i)\eta_i\} \frac{F_0^2(\psi, t_i)}{(1 - pF_0(t_i, \psi))^2};\\
f_i^{31}(\theta) &= \{\delta_i(\eta_i - 1) + (1-\delta_i)\eta_i\} \frac{t_i[1 - F_0(t_i, \psi)]}{(1 - pF_0(t_i, \psi))^2};\\
f_i^{32}(\theta) &= \{\delta_i\eta_i + (1-\delta_i)\eta_i\} \frac{F_0(t_i, \psi)}{1 - pF_0(t_i, \psi)}.
\end{aligned}
$$

In Section 3.2, (3.11) shows that the positive definiteness of D_n is essential. This in turn depends on the positive definiteness of the $\mathcal{D}_i$, which will be investigated in Lemma 3.2 below. Define

$$
\tilde{l}_i(\xi_i) = \log(p) + \log(\psi) - \psi t_i + k_i + [\exp(k_i) - 1] \log(1 - pF_0(t_i, \psi)),
$$

(see (3.5)), and let $\xi_i = (\psi, k_i, p)$. We also need the notation

$$g_i(y,u,a) = u^\top \frac{\partial \log \tilde{l}_i(y, \xi_i^{(0)})}{\partial \xi_i} + a, \quad y \geq 0,\ u \in R^3,\ a \in R,$$

where $\xi_i^{(0)} = (\psi_0, k_{i0}, p_0) = (\psi_0, \beta_0^\top z_i, p_0)$.

Lemma 3.2.

(i) *Assume that Assumption (F1) holds in the interior case, or Assumptions (F1) and (F4) hold in the boundary case hold. Then the matrix $\mathcal{D}_i$ defined by (3.10) is positive definite for each $i = 1, \ldots, n$.*

(ii) *Suppose that (F2) holds. If $\lambda_{\min}(\mathcal{D}_i) > 0$ for $i \geq 1$, then*

$$\lambda_{\min}(D_n) > 0 \quad \textit{for some} \quad n \geq q+2, \tag{3.13}$$

or equivalently, $\lambda_{\min}(D_n) > 0$ for large enough n. Where $\lambda_{\min}(\mathcal{D}_i)$ denotes the minimal eigenvalue of matrix $\mathcal{D}_i$.

The next lemma gives necessary and sufficient conditions for (F3).

Lemma 3.3. *If Assumptions (F1)-(F2) hold, then for each $A > 0$ and n large enough,*

$$\sup_{\theta \in N_n(A)} \left|X_i^\top (\theta - \theta_0)\right|^2 \leq A^2 tr(X_i^\top D_n^{-1} X_i), \tag{3.14}$$

and for some constant $B > 0$ not depending on A and n,

$$\sup_{\theta \in N_n(A)} \left|\theta - \theta_0\right|^2 \leq BA^2 \max_{1 \leq i \leq n} tr(X_i^\top D_n^{-1} X_i). \tag{3.15}$$

Furthermore, $\lambda_{min}(D_n) \to \infty$ as $n \to \infty$ if in addition (F3) holds.

The following Lemma is important to the proofs of our main results.

Lemma 3.4. *Assume that (F1)-(F3) hold for the interior case or (F1)-(F4) hold for the boundary case, and that covariates z_i, $i = 1, \ldots, n$, are bounded so that $\exp(\beta_0^\top z_i) = \exp(k_{i0}) \leq B_0$ for some constant $B_0 > 0$. Then there exists a constant K (may be dependent on A) such that for $r, s = 1, 2, 3$,*

$$(C1) \quad E\left\{\sup_{\theta\in N_n(A)} |f_i^{rs}(\theta)-f_i^{rs}(\theta_0)|\right\} \le KM_i\sqrt{tr(X_i^\top D_n^{-1}X_i)}; \tag{3.16}$$

$$(C2) \quad E\left\{|f_i^{rs}(\theta_0)|^{3/2}\right\} \le KM_i; \tag{3.17}$$

$$(C3) \quad E\left\{|S_i(\theta_0)|^3\right\} \le KM_i. \tag{3.18}$$

Let $D_n^{-1/2}X_i = w_i = (w_{i1}, w_{i2}, w_{i3})$, where $w_{ir} \in R^{q+2}$, $r = 1, 2, 3$, so that

$$tr\left(X_i^\top D_n^{-1}X_i\right) = |w_{i1}|^2 + |w_{i2}|^2 + |w_{i3}|^2. \tag{3.19}$$

For any unit vector $u \in R^{q+2}$ and $r, s = 1, 2, 3$, let $a_{in}^{rs} = u^\top w_{ir}w_{is}^\top u$. Then for $r, s = 1, 2, 3$,

$$|a_{in}^{rs}| = |u^\top w_{ir}w_{is}^\top u| \le \frac{|w_{ir}|^2 + |w_{is}|^2}{2} \le |w_{ir}|^2 + |w_{is}|^2 \le tr\left(X_i^\top D_n^{-1}X_i\right). \tag{3.20}$$

Lemma 3.5. *If (F3) and* $(C1)-(C2)$ *hold, then*

$$\sum_{i=1}^n a_{in}^{rs}\left(f_i^{rs}(\theta_0) - d_i^{rs}\right) \xrightarrow{p} 0 \quad as \quad n\to\infty. \tag{3.21}$$

3.3.2. Proofs of Preliminary Results

Proof of Lemma 3.1. The formulae for $s_{ir}(\theta)$ and $f_i^{rs}(\theta)$ are verified by differentiation of (3.5). We should mention that the formulae for $d_i^{rs} = E\{f_i^{rs}(\theta_0)\}$ $(r, s = 1, 2, 3)$ are a straightforward application of Lemma 2 of Ghitany, Maller and Zhou (1994), i.e., let $F(x)$ is the distribution of the fail times, c is the censoring random variable and δ is the censoring index, then for any positive measurable function $Q(t) : R \to R$,

$$E[\delta Q(t)] = E\left[\int_0^c Q(x)dF(x)\right]$$

and

$$E[(1-\delta)Q(t)] = E[(1-F(c))Q(c)].$$

∎

Proof of Lemma 3.2. Note that by the definition of $g_i(y,u,a)$ and $s_{ij},\ j=1,2,3$, in Lemma 3.1,

$$\frac{\partial \log \tilde{l}_i(y,\xi_i^{(0)})}{\partial \xi_i} = \begin{bmatrix} \dfrac{1}{\psi_0} - y_i - [\eta_{0i}-1]\dfrac{p_0 y_i \exp(-\psi_0 y_i)}{1-p_0F_0(y_i,\psi_0)} \\ 1+\eta_{0i}\log(1-p_0F_0(y_i,\psi_0)) \\ \dfrac{1}{p_0} - \dfrac{(\eta_{0i}-1)F_0(y_i,\psi_0)}{1-p_0F_0(y_i,\psi_0))} \end{bmatrix}.$$

Thus for $(u,a)\in R^4-\{0\},\ y>0$,

$$\begin{aligned} g_i(y,u,a) = {} & u_1\left\{\frac{1}{\psi_0} - y_i - [\eta_{0i}-1]\frac{p_0 y_i \exp(-\psi_0 y_i)}{1-p_0F_0(y_i,\psi_0)}\right\} \\ & + u_2\left\{1+\eta_{0i}\log(1-p_0F_0(y_i,\psi_0))\right\} \\ & + u_3\left\{\frac{1}{p_0} - \frac{(\eta_{0i}-1)F_0(y_i,\psi_0)}{1-p_0F_0(y_i,\psi_0))}\right\} + a. \end{aligned}$$

Clearly, $g_i(y,u,a)=0$ has no zero (when $(u_1,u_2,u_3)=0$, in which case $a\neq 0$), or only isolated zeroes (when $u\neq 0$), so that $\Pr\{g_i(t^*,u,a)=0\}=0,\ i=1,\dots,n$. Thus $\mathcal{D}_i$ is positive definite by Lemma 7.2 in Maller and Zhou (2002). This proves part(i) of Lemma 3.2.

For part(ii), suppose that $\lambda_{\min}(\mathcal{D}_i)>0,\ i\geq 1$, and that (F3) holds. Define a matrix C_n by

$$C_n = \sum_{i=i}^{n} X_i X_i^\top = \begin{bmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & \sum_{1=i}^{n} z_i z_i^\top & & \\ & & & \ddots & \\ 0 & & & & 1 \end{bmatrix}_{(q+2)\times(q+2).} \tag{3.22}$$

Thus (F3) is obviously equivalent to

$$\lambda_{\min}\{C_{n_0}\}>0, \quad \text{for some} \quad n_0\geq q+2. \tag{3.23}$$

Let ν be any $(q+2)-$dimensional unit vector. Suppose that $\lambda(D_n)=0$ for some $n\geq n_0$. Since

$$\nu^\top D_n \nu = \sum_{i=1}^{n} \nu^\top X_i \mathcal{D}_i X_i^\top \nu \geq \sum_{i=1}^{n} \lambda_{\min}(\mathcal{D}_i)\nu^\top X_i X_i^\top \nu$$

and $\lambda_{\min}(\mathcal{D}_i) > 0$ for $i \geq 1$, this means $\nu^\top X_i X_i^\top \nu = 0$, for some $\nu(n)$ and $1 \leq i \leq n$. This contradicts (3.23) as $n_0 \leq n$. Hence (F3) implies that (3.13) holds for all values of $n \geq n_0$, i.e., for all values of n large enough. ∎

Proof of Lemma 3.3. Note that (3.14) is the same as (4.14) of Ghitany *et al.* (1994), and (3.15) follows just as in the working after (4.24) of Ghitany *et al.* (1994). Furthermore, let v_n be a $(q+2)$-dimensional eigenvector of D_n associated with $\lambda_{\min}(D_n)$, i.e., $D_n v_n = \lambda_{\min}(D_n) v_n$, define the $(q+2)$-dimensional vector $\theta_n = \theta_0 + \sqrt{A^2 / v_n{}^\top D_n v_n}\; v_n$. Since $\theta_n \in N_n(A)$, it follows from (3.15) that

$$A^2 = \lambda_{\min}(D_n)|\theta_n - \theta_0| \leq \lambda_{\min}(D_n) B A^2 \max_{1 \leq i \leq n} tr(X_i^\top D_n^{-1} X_i). \tag{3.24}$$

Thus (F4) implies that $\lambda_{\min}(D_n) \to \infty$ as $n \to \infty$. ∎

Proof of Lemma 3.4. By Taylor expansion of $f_i^{rs}(\theta)$ about θ_0, and using the chain rule of differentiation, we obtain, for $r, s = 1, 2, 3$,

$$\begin{aligned} |f_i^{rs}(\theta) - f_i^{rs}(\theta_0)| &= \left| (\theta - \theta_0)^\top \frac{\partial}{\partial \theta} f_i^{rs}(\tilde{\theta}) \right| \\ &= \left| (\psi - \psi_0) \frac{\partial}{\partial \psi} f_i^{rs}(\tilde{\theta}) + (\beta - \beta_0)^\top z_i \frac{\partial}{\partial k_i} f_i^{rs}(\tilde{\theta}) + (p - p_0) \frac{\partial}{\partial p} f_i^{rs}(\tilde{\theta}) \right|, \end{aligned} \tag{3.25}$$

where $\tilde{\theta}$ is on the segment between θ and θ_0, thus $\tilde{\theta} \in N_n(A)$. We must now bound the quantities in the above. For any $\theta \in N_n(A)$ we have $|D^{1/2}(\theta - \theta_0)|^2 \leq A^2$ by the definition in (3.12), hence by Lemma 3.3, when $\theta = (\psi, \beta^\top, p)^\top$, we have

$$|\psi - \psi_0|^2 + |z_i^\top (\beta - \beta_0)|^2 + |p_0 - p|^2 = |X_i^\top (\theta - \theta_0)|^2 \leq A^2 M, \tag{3.26}$$

where $M = tr(X_i^\top D_n^{-1} X_i)$. From (3.22) it follows that for $\theta \in N_n(A)$,

$$|f_i^{rs}(\theta) - f_i^{rs}(\theta_0)| \leq A\sqrt{M} \left\{ \left| \frac{\partial}{\partial \psi} f_i^{rs}(\tilde{\theta}) \right| + \left| \frac{\partial}{\partial k_i} f_i^{rs}(\tilde{\theta}) \right| + \left| \frac{\partial}{\partial p} f_i^{rs}(\tilde{\theta}) \right| \right\}. \tag{3.27}$$

So we need to derive bounds for the quantities in (3.27). We illustrate this derivation for $r = s = 1$. Note that, by Lemma 3.1,

$$\frac{\partial}{\partial \psi} f_i^{11}(\theta) = \frac{-2\delta_i}{\psi^3} + \{(\eta_i - 1)\delta_i + (1 - \delta_i)\eta_i\} p(1-p) t_i \frac{\tilde{f}_1(\xi, t_i) \tilde{f}_2(\xi, t_i)}{(1 - pF_0(t_i, \psi))^3}; \tag{3.28}$$

$$\frac{\partial}{\partial k_i} f_i^{11}(\theta) = -\left\{\eta_i \delta_i + (1-\delta_i)\eta_i\right\} p(1-p) \frac{\tilde{f}_1(\xi, t_i)}{\left(1 - pF_0(t_i, \psi)\right)^2}; \tag{3.29}$$

$$\frac{\partial}{\partial p} f_i^{11}(\theta) = -\left\{(\eta_i - 1)\delta_i + (1-\delta_i)\eta_i\right\} \frac{\tilde{f}_1(\xi, t_i)\tilde{f}_2(\xi, t_i)}{\left(1 - pF_0(t_i, \psi)\right)^3}, \tag{3.30}$$

where $\tilde{f}_1(\xi, t_i) = t_i^2[1 - F_0(t_i, \psi)]$ and $\tilde{f}_2(\xi, t_i) = pF_0(t_i, \psi) - 2p + 1$. Note that $\tilde{f}_2(\xi, t_i) \leq \exp(-\psi t_i) + 2$, by differentiation we obtain

$$\left|\frac{\partial}{\partial \psi} f_i^{11}(\theta)\right| \leq \frac{2\delta_i}{\psi^3} + \left\{(1-\delta_i) + B_1\right\} \frac{t_i^3(2e^{-\psi t_i} + e^{-2\psi t_i})}{4\left(1 - pF_0(t_i, \psi)\right)^3}; \tag{3.31}$$

$$\left|\frac{\partial}{\partial k_i} f_i^{11}(\theta)\right| \leq \frac{B_0 t_i^2 e^{-\psi t_i}}{4\left(1 - pF_0(t_i, \psi)\right)^2}; \tag{3.32}$$

$$\left|\frac{\partial}{\partial p} f_i^{11}(\theta)\right| \leq \left\{(1-\delta_i) + B_1\right\} \frac{t_i^2(2e^{-\psi t_i} + e^{-2\psi t_i})}{\left(1 - pF_0(t_i, \psi)\right)^3}, \tag{3.33}$$

where $B_1 = B_0 + 1$. It suffices to prove the following inequality from (3.31):

$$E\left\{\sup_{\theta \in N_n(A)} \left[\frac{2\delta_i}{\psi^2} + [(1-\delta_i) + B_1] \frac{t_i^3(2e^{-\psi t_i} + e^{-2\psi t_i})}{4\left(1 - pF_0(t_i, \psi)\right)^3}\right]\right\} \leq M_{i1}^{(11)} < \infty. \tag{3.34}$$

By Lemma 2 in Ghitany *et al.* (1994), the first term on the left side of (3.34) is equivalent to

$$E\left\{\sup_{\theta \in N_n(A)} \left\{\frac{2F_{\theta_0}(c_i)}{\psi^2} + \frac{c_i^3(2e^{-\psi c_i} + e^{-2\psi c_i})}{4\left(1 - pF_0(\psi, c_i)\right)^3} S_{\theta_0}(c_i) + \frac{B_1 t_i^3(2e^{-\psi t_i} + e^{-2\psi t_i})}{4\left(1 - pF_0(t_i, \psi)\right)^3}\right\}\right\}$$
$$\leq E\left\{\sup_{\theta \in N_n(A)} \left\{\frac{2}{\psi^2} + \frac{c_i^3(2e^{-\psi c_i} + e^{-2\psi c_i})}{4\left(1 - pF_0(\psi, c_i)\right)^3} + \frac{B_1 t_i^3(2e^{-\psi t_i} + e^{-2\psi t_i})}{4\left(1 - pF_0(t_i, \psi)\right)^3}\right\}\right\}.$$

Note from (3.26) that $\psi_0 - A\sqrt{M} \leq \psi \leq \psi_0 + A\sqrt{M}$ and $p_0 - A\sqrt{M} \leq p \leq p_0 + A\sqrt{M}$. Thus (3.34) follows from (F1) for the interior case, or (F1) and (F4) for the boundary case, together with $2/\psi^2$ and $u^3(e^{-\psi u} + 2e^{-2\psi u})/4\left(1 - pF_0(\psi, u)\right)^3$ are integrable with respect to $G(du)$, and $B_0 v^3(e^{-\psi v} + 2e^{-2\psi v})/4\left(1 - pF_0(\psi, v)\right)^3$ is integrable with respect to $H_i(dv)$, where $H_i(v) = 1 - [1 - F_i(v)][1 - G(v)]$.

From (3.27) and a similar bound for the terms on the right-hand side of (3.32)-(3.33), we obtain $M_{i2}^{(11)}$ and $M_{i3}^{(11)}$.

The details of the proofs for (C2) and (C3) are fairly standard and thus omitted here. If $E\left\{|f_i^{rs}(\theta_0)|^{3/2}\right\} \leq KN_i^{(rs)}$ and $E\left\{|s_i(\theta_0)|^3\right\} \leq KQ_i^{(r)},\ r, s = 1, 2, 3.$

For each i, let M_i be the maximum value over any of the quantities

$$\left\{\sum_{j=1}^{3} M_{ij}^{(rs)}, N_i^{(rs)}, Q_i^{(r)}, r, s = 1, 2, 3\right\}. \tag{3.35}$$

This completes the proof of lemma 3.4. ∎

Proof of Lemma 3.5. See Theorem 6.2 in Vu *et al.* (1998). ∎

3.4. Main Results

3.4.1. Main Results

Let I_{q+2} and $\mathcal{N}(0, I_{q+2})$ denote the identity matrix and a standard normal random variable in q+2 dimension. We now state the major results of this chapter. We say that a sequence of events $\{A_n\}$ occurs with probability approaching 1 (WPA1) if $\Pr\{A_n\} \to 1$ as $n \to \infty$.

Theorem 3.1. *If Assumptions (F1)-(F3) for the interior case or conditions (F1)-(F4) for the boundary case are satisfied, then a MLE $\hat{\theta}_n$ of θ_0 exists, is locally unique WPA1, and is consistent in probability for θ_0.*

From now on, $\hat{\theta}_n$ denotes the estimator obtained in Theorem 3.1. $\hat{\theta}_n$ is generally not unique on Θ, even WPA1, though it is uniquely defined WPA1 on the neighborhood $N_n(A)$ of θ_0 for each $A > 0$.

Next, we consider the interior case. For this case, we want to test the hypothesis $H_0 : p = p_0 < 1$, against an unrestricted alternative $H_1 : p \in (0, 1)$. Let $L_f(\theta)$ be the likelihood function and $l(\theta) = \log(L_f(\theta))$. The likelihood ratio test statistic for H_0 is defined by $L_n = L_f(\tilde{\theta}_n)/L_f(\hat{\theta}_n)$, where $\tilde{\theta}_n$ is a local maximum estimator of $l(\theta)$ under H_0. Now defined the "deviance" of the restricted model from the unrestricted model by

$$d_n = -2 \log L_n = 2[l(\hat{\theta_n}) - l(\tilde{\theta_n})]. \tag{3.36}$$

Small values of L_n or large values of d_n indicate that H_0 is unlikely to be true. Denote by χ_v^2 a chi-square random variable with v degrees of freedom. Let $D_n^{1/2}$ and $D_n^{\top/2}$ be any left and right square roots of D_n, i.e., any square matrices such that $D_n^{1/2} D_n^{\top/2} = D_n$.

Theorem 3.2. *(The interior case) If conditions (F1)-(F3) are satisfied, then for every $y \geq 0$, as $n \to \infty$,*

$$\Pr\left\{d_n \leq y\right\} \to \Pr\left\{\chi_{q+1}^2 \leq y\right\}, \tag{3.37}$$

and

$$D_n^{\top/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_{q+1}). \tag{3.38}$$

We now turn to the boundary case. We wish to test the hypothesis $H_0 : p_0 = 1$ against an unrestricted alternative $H_1 : p \in (0, 1]$. Thus no individuals are immune to failure under the null hypothesis H_0.

Let $\dot{\theta}_n = (\dot{\vartheta}_n^\top, 1)^\top$ be a local maximum of $l(\theta)$ under H_0, where $\dot{\vartheta}_n$ is a MLE of $\vartheta = (\psi, \beta^\top)^\top$ under the restriction $p = 1$. Again define the "deviance" of the restricted model from the unrestricted model by

$$d_n = 2\left[l(\hat{\theta}_n) - l(\dot{\theta}_n)\right]. \tag{3.39}$$

Partition the expected information matrix D_n as

$$D_n = \begin{bmatrix} \bar{D}_n & g_n \\ g_n^\top & a_n \end{bmatrix}, \tag{3.40}$$

where $\bar{D}_n$ is $(q+1) \times (q+1)$ and g_n is $(q+1) \times 1$. Let

$$X \sim \mathcal{N}(0, I_{q+1}), \quad Y \sim \mathcal{N}(0, 1) \quad \text{and} \quad Z = (X^\top, Y)^\top \sim \mathcal{N}(0, I_{q+2}).$$

Our next theorem shows that d_n has a non-standard asymptotic distribution.

Theorem 3.3. *(The boundary case) If Assumptions (F1)-(F4) are satisfied, then for every $y \geq 0$, as $n \to \infty$,*

$$\Pr\{d_n \leq y\} \to \frac{1}{2} + \frac{1}{2}\Pr(\chi_1^2 \leq y), \tag{3.41}$$

$$\Pr\left\{\sqrt{a_n - g_n^\top \bar{D}_n^{-1} g_n}(\hat{p}_n - 1) \leq x\right\} \to \Pr(Y \leq x), \tag{3.42}$$

and

$$\Pr\{\hat{p}_n < 1\} \to \frac{1}{2}. \tag{3.43}$$

In addition, if $D_n^{\top/2}$ is the right Cholesky square root of D_n, then for any Borel set $W \subseteq R^{q+2}$ as $n \to \infty$,

$$\Pr\left\{D_n^{\top/2}(\hat{\theta}_n - \theta_0) \in W, \hat{p}_n < 1\right\} \to \Pr(Z \in W, Y \leq 0). \tag{3.44}$$

Theorem 3.1-3.3, just as the works of Vu, Maller and Zhou (1998), generalize the results of Zhou and Maller (1995), which deal with the exponential distribution in the case without covariate information. For an example of application with the "50-50" chi-squared distribution in (3.41), see Maller and Zhou (1996, Chapter 5) and Zhou and Maller (1995). Ghitany *et al.* (1994) discuss how to use the results like Theorem 3.1-3.3 to analyze exponentially distributed data that are classified into different groups (i.e., with one covariate to specify the group each individual belongs to).

3.4.2. Proofs of Main Results

Proofs of Theorems 3.1-3.2 and (3.41) of Theorem 3.3:

We apply the general results of Vu and Zhou (1997). In our case the $g(Y_i, \theta)$ of Vu and Zhou (1997) is the log-likelihood of the ith observation with Y_i replaced by t_i. Obviously, Conditions (A1)-(A2) of Vu and Zhou (1997) are satisfied with our covariate function of the form

$$\eta_i = \exp(\beta^\top z_i) = \exp(k_i).$$

Note that (3.11) is their condition (B1) and the matrix V in (B4) is simply I_{q+2} in our case. Since $\lambda_{\min}(D_n) \to \infty$ as $n \to \infty$, by Lemma 3.3, then their (B2) holds. Thus we can apply the results in that paper provided that their (B3) and (B5) hold, i.e., the observed information matrix can be approximated by the expected information matrix in the sense that

$$\sup_{\theta^* \in [N_n(A)]^9} \|D_n^{-1/2} F_n^*(\tilde{\theta}) D_n^{-\top/2} - I_{q+2}\| \xrightarrow{p} 0 \quad \text{for each} \quad A > 0, \tag{3.45}$$

and the score function is asymptotically normal in the sense that for any unit vector ζ_n in R^{q+2},

$$\zeta_n^\top D_n^{-1/2} S_n(\theta_0) \xrightarrow{d} \mathcal{N}(0,1). \tag{3.46}$$

In (3.45), $\|\cdot\|$ denotes the sum of the absolute values of the elements of a matrix,

$$[N_n(A)]^9 = \left\{ \tilde{\theta} = (\theta^{11}, \ldots, \theta^{33}) : \theta^{rs} \in N_n(A),\ r, s = 1, 2, 3 \right\}, \tag{3.47}$$

and

$$F_n^*(\tilde{\theta}) = \sum_{i=1}^n X_i \begin{bmatrix} f_i^{11}(\theta^{11}) & f_i^{12}(\theta^{12}) & f_i^{13}(\theta^{13}) \\ f_i^{21}(\theta^{21}) & f_i^{22}(\theta^{22}) & f_i^{23}(\theta^{23}) \\ f_i^{31}(\theta^{31}) & f_i^{32}(\theta^{32}) & f_i^{33}(\theta^{33}) \end{bmatrix} X_i^\top. \tag{3.48}$$

Then Theorem 3.1 and the asymptotic distribution of d_n (cf. (3.37) and (3.41)) in Theorems 3.2-3.3 follow from Theorems 2.1-2.2 in Vu and Zhou (1997). For the asymptotic distribution of $\hat{\theta}_n$ given in (3.38), let u be a unit vector in R^{q+2}. Recall that $\hat{\theta}_n$ maximizes the log-likelihood function on $N_n(A)$ WPA1. Thus $S_n(\hat{\theta}_n) = 0$ WPA1 as the log-likelihood function is concave on $N_n(A)$ for all n large enough by (3.45). Hence there exists, by Taylor expansion, a $\bar{\theta}_n$ on the line segment between $\hat{\theta}_n$ and θ_0 such that

$$\begin{aligned} u^\top D_n^{-1/2} S_n(\theta_0) &= u^\top D_n^{-1/2} S_n(\theta_0) - u^\top D_n^{-1/2} S_n(\hat{\theta}_n) \\ &= u^\top D_n^{-1/2} F_n(\bar{\theta}_n)(\hat{\theta}_n - \theta_0) + o_p(1) \\ &= u^\top \left\{ I_{q+2} + D_n^{-1/2}[F_n(\bar{\theta}_n) - D_n] D_n^{-\top/2} \right\} D_n^{1/2}(\hat{\theta}_n - \theta_0) + o_p(1) \\ &= u^\top \left\{ I_{q+2} + o_p(1) \right\} D_n^{1/2}(\hat{\theta}_n - \theta_0) + o_p(1). \end{aligned} \tag{3.49}$$

Thus $\hat{\theta}_n$ is asymptotically normal by (3.46) and (3.49).

It remains to prove (3.45)-(3.46). Let $\tilde{\theta} \in [N_n(A)]^9$ and write

$$D_n^{-1/2} F_n^*(\tilde{\theta}) D_n^{-\top/2} = I_{k+2} + \varepsilon_n^{(1)}(\theta_0) + \varepsilon_n^{(2)}(\tilde{\theta}),$$

where

$$\varepsilon_n^{(1)}(\theta_0) = D_n^{-1/2} \{F_n(\theta_0) - D_n\} D_n^{-\top/2}$$

and

$$\varepsilon_n^{(2)}(\tilde{\theta}) = D_n^{-1/2} \left\{F_n^*(\tilde{\theta}) - F_n(\theta_0)\right\} D_n^{-\top/2}.$$

Let u be any unit vector in R^{q+2}. Observe that

$$u^\top \varepsilon_n^{(1)}(\theta_0) u = \sum_{i=1}^n \sum_{1 \le r,s \le 3} a_{in}^{rs} \left(f_i^{rs}(\theta_0) - d_i\right) \tag{3.50}$$

and

$$\sup_{\tilde{\theta} \in [N_n(A)]^9} \left|u^\top \varepsilon_n^{(2)}(\tilde{\theta}) u\right| \le \sum_{i=1}^n |a_{in}^{rs}| \sup_{\theta^{rs} \in N_n(A)} |f_i^{rs}(\theta^{rs}) - f_i^{rs}(\theta_0)|. \tag{3.51}$$

Thus (3.50) tends to 0 in probability by Lemma 3.5, and (4.51) tends to 0 in probability by Markov inequality together with (3.20), (C1) and (F3). This proves (3.45).

To prove (3.46), let ζ_n be any unit vector in R^{q+2} and define

$$Y_{in} = \zeta_n^\top D_n^{-1/2} X_i S_i(\theta_0), \quad 1 \le i \le n,$$

$$\sigma_{in}^2 = Var(Y_{in}) = \zeta_n^\top D_n^{-1/2} X_i \mathcal{D}_i X_i^\top D_n^{-\top/2} \zeta_n.$$

Then for each n, $Y_{1n}, \dots, Y_{nn}$ are mutually independent with mean $E(Y_{in}) = 0$ and sum of variance $\sigma_{1n}^2 + \cdots + \sigma_{nn}^2 = 1$. It follows from (C3) that

$$\begin{aligned}
\sum_{i=1}^n E(|Y_{in}|^3) &\le \sum_{i=1}^n E\left(\left[\lambda_{\max}(S_i(\theta_0) S_i^\top(\theta_0)) \zeta_n^\top D_n^{-1/2} X_i X_i^\top D_n^{-\top/2} \zeta_n\right]^{3/2}\right) \\
&\le \sum_{i=1}^n E(|S_i(\theta_0)|^3) \left(tr(X_i^\top D_n^{-1} X_i)\right)^{3/2} \\
&\le K \sum_{i=1}^n M_i \left(tr(X_i^\top D_n^{-1} X_i)\right)^{3/2}.
\end{aligned}$$

By (F3), the last expression tends to 0 as $n \to \infty$. Thus (3.46) follows from the Lyapounov Theorem in Billingsley (1968, p.44). ∎

Proofs of (3.42) and (3.43) of Theorem 3.3:

Next we show (3.42) and (3.43) of Theorem 3.3, while the proof of (3.44) can be found in Vu and Zhou (1998, p.652).

Note that $D_n = \sum_{i=1}^n X_i \mathcal{D}_i X_i^\top$ and (3.40) implies

$$\bar{D}_n = E(\bar{F}_n(\theta_0)) \quad \text{with} \quad \bar{F}_n(\theta_0) = \sum_{i=1}^n \bar{X}_i \begin{bmatrix} f_i^{11}(\theta_0) & f_i^{12}(\theta_0) \\ f_i^{21}(\theta_0) & f_i^{22}(\theta_0) \end{bmatrix} \bar{X}_i^\top, \tag{3.52}$$

where $\bar{X}_i$ is of order $(q+1) \times 2$ and $X_i = \begin{bmatrix} \bar{X}_i & 0 \\ 0 & x_{q+2} \end{bmatrix}$.

Partition a $(q+2) \times (q+2)$ real symmetric matrix F_n and a $(q+2) \times (q+2)$ real matrix $\tilde{F}_n$ as

$$F_n = \begin{bmatrix} \bar{F}_n & f_n \\ f_n^\top & a_n \end{bmatrix} \quad \text{and} \quad \tilde{F}_n = \begin{bmatrix} \bar{\tilde{F}}_n & \tilde{f}_n \\ \tilde{h}_n^\top & \tilde{a}_n \end{bmatrix},$$

where $\tilde{F}_n, \bar{\tilde{F}}_n, \tilde{h}_n, \tilde{f}_n$ and $\tilde{a}_n$ are defined in (3.63)-(3.65) below, with $\bar{F}_n$ and $\bar{\tilde{F}}_n$ being $(q+1) \times (q+1)$ matrices; and $f_n, \tilde{f}_n$ and $\tilde{h}_n$ being $(q+1)$-vectors.

Denote by $\bar{S}_n(\theta_0)$ the $(q+1)$-vector $(S_{n1}(\theta_0), \ldots, S_{n(q+1)}(\theta_0))^\top$.

The following Lemmas 3.6-3.8 are used to show (3.42) and (3.43). Their proofs are similar to the arguments in Vu and Zhou (1998, pp.649-651).

Lemma 3.6. *Suppose that* $D_n^{-1/2} F_n D_n^{-\top/2} \xrightarrow{p} I_{q+2}$. *Then*

$$\bar{D}_n^{-1/2} \bar{F}_n \bar{D}_n^{-\top/2} \xrightarrow{p} I_{q+1} \quad \textit{and} \quad [F_n^{-1}]_{(q+2)(q+2)} = (1 + o_p(1))[D_n^{-1}]_{(q+2)(q+2)}.$$

Also if $D_n^{-1/2} S_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, I_{q+2})$, *then* $\bar{D}_n^{-1/2} \bar{S}_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, I_{q+1})$.

For a fixed $A > 0$, define a subset of R^{q+2} by

$$\begin{aligned} N_n^{'}(A) &= \left\{\theta : (\theta - \theta_0)^\top D_n (\theta - \theta_0) \le A^2, p = p_0\right\} \\ &= \left\{(\vartheta^\top p_0)^\top : (\vartheta - \vartheta_0)^\top \bar{D}_n (\vartheta - \vartheta_0) \le A^2\right\}. \end{aligned} \tag{3.53}$$

Lemma 3.7. *Suppose that Conditions (F1)-(F4) hold.*

(i) When $\hat{p}_n < p_0$, $\hat{\theta}_n$ is an interior stationary point of $l(\theta)$ WPA1, i.e.,

$$\Pr(S_n(\hat{\theta}_n) \neq 0, \hat{p}_n < p_0) \to 0 \quad (n \to \infty). \tag{3.54}$$

(ii) Let $\dot{\vartheta}_n$ be the maximizer of the log-likelihood function in $N_n^{'}(A)$. We have

$$\Pr(S_{nj}(\dot{\vartheta}_n, p_0) = 0, \forall j \in [1, q+1]) \to 1 \quad (n \to \infty) \tag{3.55}$$

and

$$\Pr(S_{nj}(\hat{\theta}_n) = 0, \forall j \in [1, q+1]) \to 1 \quad (n \to \infty). \tag{3.56}$$

(iii) $\hat{p}_n = p_0$ if and only if $S_{n(k+2)}(\dot{\vartheta}_n, p_0) \geq 0$ WPA1.

Let

$$b_n = \sqrt{a_n - g_n^\top D_n^{-1} g_n}, \tag{3.57}$$

and

$$v_n^\top = b_n [\,0 \quad \cdots \quad 0 \quad 1\,] D_n^{-\top/2} = \sqrt{a_n - g_n^\top D_n^{-1} g_n} [\,0 \quad \cdots \quad 0 \quad 1\,] D_n^{-\top/2}. \tag{3.58}$$

Lemma 3.8. *$v_n^\top$ is a unit vector. Furthermore,*

$$[\tilde{F}_n^{-1}]_{(q+2)(q+2)} = (\tilde{a}_n - \tilde{h}_n^\top \bar{\tilde{F}}_n^{-1} \tilde{f}_n)^{-1},$$

$$[D_n^{-1}]_{(q+2)(q+2)} = (a_n - g_n^\top \bar{D}^{-1} g_n)^{-1} = b_n^{-2}, \tag{3.59}$$

and

$$[\,0 \quad \cdots \quad 0 \quad 1\,] \tilde{F}_n^{-1} S_n(\theta_0) = \frac{S_{n(q+2)}(\theta_0) - \tilde{h}_n^\top \bar{\tilde{F}}_n^{-1} \bar{S}_n(\theta_0)}{\tilde{a}_n - \tilde{h}_n^\top \bar{\tilde{F}}_n^{-1} \tilde{f}_n}. \tag{3.60}$$

In addition, if $D_n^{-1/2} \tilde{F}_n D_n^{-\top/2} \xrightarrow{p} I_{q+2}$, then

$$b_n [\,0 \quad \cdots \quad 0 \quad 1\,] \tilde{F}_n^{-1} S_n(\theta_0) = v_n^\top D_n^{-1/2} S_n(\theta_0) + o_p(1), \tag{3.61}$$

where $\tilde{F}_n, \bar{\tilde{F}}_n, \tilde{h}_n, \tilde{f}_n$ and $\tilde{a}_n$ are defined in (3.63)-(3.65) below.

We now complete the proof of (3.42)-(3.43). We will work throughout on the event $\{\hat{p}_n < p_0\}$. By Lemma 3.7, $S_n(\hat{\theta_n}) = S_n(\hat{\vartheta}_n, \hat{p}_n) = 0$ WPA1 on this event. By Taylor expansion, we have

$$S_n(\theta_0) = S_n(\theta_0) - S_n(\hat{\vartheta}_n, p_0) + S_n(\hat{\vartheta}_n, p_0) - S_n(\hat{\vartheta}_n, \hat{p}_n)$$

$$= \sum_{i=1}^{n} \begin{bmatrix} f_i^{11}(\tilde{\theta}_n^{11})(\hat{\psi}_n - \psi_0) + f_i^{12}(\tilde{\theta}_n^{12}) z_i^\top (\hat{\beta}_n - \beta_0) \\ z_i f_i^{21}(\tilde{\theta}_n^{21})(\hat{\psi}_n - \psi_0) + z_i f_i^{22}(\tilde{\theta}_n^{22}) z_i^\top (\hat{\beta}_n - \beta_0) \\ f_i^{31}(\tilde{\theta}_n^{31})(\hat{\psi}_n - \psi_0) + f_i^{32}(\tilde{\theta}_n^{32}) z_i^\top (\hat{\beta}_n - \beta_0) \end{bmatrix}$$
$$+ \sum_{i=1}^{n} \begin{bmatrix} f_i^{13}(\tilde{\theta}_n^{13})(\hat{p}_n - p_0) \\ z_i f_i^{23}(\tilde{\theta}_n^{23})(\hat{p}_n - p_0) \\ f_i^{33}(\tilde{\theta}_n^{33})(\hat{p}_n - p_0) \end{bmatrix}, \tag{3.62}$$

where $\tilde{\theta}_n^{rs}, r = 1, 2, 3, s = 1, 2$, lie between θ_0 and $(\hat{\vartheta}_n, p_0)$, and $\tilde{\theta}_n^{rs}, r = 1, 2, 3, s = 3$, lie between $(\hat{\vartheta}_n, p_0)$ and $\hat{\theta}_n$. Define

$$\bar{\tilde{F}}_n = \sum_{i=1}^{n} \bar{X}_i \begin{bmatrix} f_i^{11}(\tilde{\theta}_n^{11}) & f_i^{12}(\tilde{\theta}_n^{12}) \\ f_i^{21}(\tilde{\theta}_n^{21}) & f_i^{22}(\tilde{\theta}_n^{22}) \end{bmatrix} \bar{X}_i^\top, \tilde{h}_n$$
$$= \sum_{i=1}^{n} x_{q+2} \left[f_i^{31}(\tilde{\theta}_n^{31}) \quad f_i^{32}(\tilde{\theta}_n^{32}) \right] \bar{X}_i^\top, \tag{3.63}$$

$$\tilde{f}_n = \sum_{i=1}^{n} x_{q+2} \bar{X}_i \left[f_i^{13}(\tilde{\theta}_n^{13}) \quad f_i^{23}(\tilde{\theta}_n^{23}) \right]^\top, \tilde{a}_n = \sum_{i=1}^{n} x_{q+2}^2 f_i^{33}(\tilde{\theta}_n^{33}), \tag{3.64}$$

and

$$\tilde{F}_n = \begin{bmatrix} \bar{\tilde{F}}_n & \tilde{f}_n \\ \tilde{h}_n^\top & \tilde{a}_n \end{bmatrix} = \sum_{i=1}^{n} X_i \begin{bmatrix} f_i^{11}(\tilde{\theta}^{11}) & f_i^{12}(\tilde{\theta}^{12}) & f_i^{13}(\tilde{\theta}^{13}) \\ f_i^{21}(\tilde{\theta}^{21}) & f_i^{22}(\tilde{\theta}^{22}) & f_i^{23}(\tilde{\theta}^{23}) \\ f_i^{31}(\tilde{\theta}^{31}) & f_i^{32}(\tilde{\theta}^{32}) & f_i^{33}(\tilde{\theta}^{33}) \end{bmatrix} X_i^\top, \tag{3.65}$$

where $\bar{\tilde{F}}_n$ are $(q+1) \times (q+1)$ matrices, and $f_n, \tilde{f}_n$ and $\tilde{h}_n$ are $(q+1)$-vectors. Thus $S_n(\theta_0) = \tilde{F}_n(\hat{\theta}_n - \theta_0)$, and then

$$\hat{\theta}_n - \theta_0 = \tilde{F}_n^{-1} S_n(\theta_0). \tag{3.66}$$

Since $D_n^{-1/2}\tilde{F}_n D_n^{-\top/2} \xrightarrow{p} I_{q+2}$, it follows from (3.61) and (3.66) that

$$
\begin{aligned}
\sqrt{a_n - g_n^\top \bar{D}_n^{-1} g_n}(\hat{p}_n - p_0) &= b_n[0 \quad \cdots \quad 0 \quad 1][\hat{\theta}_n - \theta_0] \\
&= b_n[0 \quad \cdots \quad 0 \quad 1]\tilde{F}^{-1} S_n(\theta_0) \\
&= v_n^\top D_n^{-1/2} S_n(\theta_0) + o_p(1) \qquad (3.67)
\end{aligned}
$$

on $\{\hat{p}_n < p_0\}$, where v_n is defined by (3.58). Therefore, by (3.67),

$$
\Pr\left\{\hat{p}_n < p_0, v_n^\top D_n^{-1/2} S_n(\theta_0) \geq 0\right\} \to 0 \quad \text{as } n \to \infty.
$$

If we assume that

$$
\Pr\left\{\hat{p}_n = p_0, v_n^\top D_n^{-1/2} S_n(\theta_0) < 0\right\} \to 0, \quad n \to \infty, \qquad (3.68)
$$

then it follow from (3.46) that, for $x < 0$,

$$
\lim_{n\to\infty} \Pr\left\{\sqrt{a_n - g_n^\top \bar{D}_n^{-1} g_n}(\hat{p}_n - 1) \leq x\right\} = \Pr(Y \leq x).
$$

Thus (3.42)-(3.43) follow from (3.68). It remains to show (3.68).

Since $\bar{S}_n(\dot{\theta}_n) = 0$ WPA1 by (3.55), by Taylor expansion, there exist $\tilde{\theta}_{n1}^{rs}$, $r, s = 1, 2$, and $\tilde{\theta}_{n1}^{rs}, r = 3, s = 1, 2$, between $\dot{\theta}_n$ and θ_0 such that

$$
\begin{aligned}
\bar{S}_n(\theta_0) &= \bar{S}_n(\theta_0) - \bar{S}_n(\dot{\theta}_n) \\
&= \sum_{i=1}^n \begin{bmatrix} f_i^{11}(\tilde{\theta}_{n1}^{11})(\dot{\psi}_n - \psi_0) + f_i^{12}(\tilde{\theta}_{n1}^{12}) z_i^\top (\dot{\beta}_n - \beta) \\ z_i f_i^{21}(\tilde{\theta}_{n1}^{21})(\dot{\psi}_n - \psi_0) + z_i f_i^{22}(\tilde{\theta}_{n1}^{22}) z_i^\top (\dot{\beta}_n - \beta_0) \end{bmatrix} \\
&= \sum_{i=1}^n \bar{X}_i \begin{bmatrix} f_i^{11}(\tilde{\theta}_{n1}^{11}) & f_i^{12}(\tilde{\theta}_{n1}^{12}) \\ f_i^{21}(\tilde{\theta}_{n1}^{21}) & f_i^{22}(\tilde{\theta}_{n1}^{22}) \end{bmatrix} \bar{X}_i^\top (\dot{\vartheta}_n - \vartheta_0) \\
&= \dot{\bar{F}}_n(\dot{\vartheta}_n - \vartheta_0) \qquad (3.69)
\end{aligned}
$$

and

$$
\begin{aligned}
S_{n(q+2)}(\dot{\theta}_n) &= S_{n(q+2)}(\theta_0) - \sum_{i=1}^n \left\{ f_i^{31}(\tilde{\theta}_{n1}^{31})(\dot{\psi}_n - \psi_0) + f_i^{32}(\tilde{\theta}_{n1}^{32}) z_i^\top (\dot{\beta}_n - \beta_0) \right\} \\
&= S_{n(q+2)}(\theta_0) - \sum_{i=1}^n x_{q+2} \, [\, f_i^{31}(\tilde{\theta}_{n1}^{31}) \quad f_i^{32}(\tilde{\theta}_{n1}^{32}) \,] \, \bar{X}_i^\top (\dot{\vartheta}_n - \vartheta_0) \\
&= S_{n(q+2)}(\theta_0) - \dot{h}_n^\top (\dot{\vartheta}_n - \vartheta_0), \qquad (3.70)
\end{aligned}
$$

where

$$\dot{F}_n = \begin{bmatrix} \bar{\dot{F}}_n & \dot{h}_n \\ \\ \dot{h}_n^\top & a_n \end{bmatrix}, \quad \bar{\dot{F}}_n = \sum_{i=1}^n \bar{X}_i \begin{bmatrix} f_i^{11}(\tilde{\theta}_{n1}^{11}) & f_i^{12}(\tilde{\theta}_{n1}^{12}) \\ \\ f_i^{21}(\tilde{\theta}_{n1}^{21}) & f_i^{22}(\tilde{\theta}_{n1}^{22}) \end{bmatrix} \bar{X}_i^\top,$$

and

$$\dot{h}_n^\top = \sum_{i=1}^n x_{q+2}[f_i^{31}(\tilde{\theta}_{n1}^{31}) \quad f_i^{32}(\tilde{\theta}_{n1}^{32})]\bar{X}_i^\top.$$

Thus (3.45) implies that $D_n^{-1/2}\dot{F}_n D_n^{-\top/2} \xrightarrow{p} I_{q+2}$. By substituting $\dot{F}_n, \bar{\dot{F}}_n, \dot{h}_n$ and a_n for $\tilde{F}_n, \bar{\tilde{F}}_n, \tilde{h}_n$ and $\tilde{a}_n$ into (3.60) and (3.61), it follows form (3.69)-(3.70) that

$$\begin{aligned} \frac{b_n S_{n(q+2)}(\dot{\theta}_n)}{a_n - \dot{h}_n^\top \bar{\dot{F}}_n^{-1} \dot{h}_n} &= \frac{b_n(S_{n(q+2)}(\theta_0) - \dot{h}_n^\top \bar{\dot{F}}_n^{-1} \bar{S}_n(\theta_0))}{a_n - \dot{h}_n^\top \bar{\dot{F}}_n^{-1} \dot{h}_n} \\ &= v_n^\top D_n^{-1/2} S_n(\theta_0) + 0_p(1). \end{aligned} \tag{3.71}$$

By Lemma 3.7, the event $\{\hat{p}_n = p_0\}$ occurs if and only if $\{S_{n(q+2)}(\dot{\theta}_n) \geq 0\}$ occurs WPA1. Thus (3.68) follows from (3.71). This completes the proof. ∎

Chapter 4

Semiparametric Transformation Models for Survival Data with Long-term Survivors

4.1. Introduction

Semiparametric transformation models have recently attracted considerable attention and efforts in the analyses of survival data. A semiparametric transformation model has the form

$$h(T^*) = -z^\top \beta + \varepsilon, \tag{4.1}$$

where T^* is the response variable, $h(\cdot)$ is a smooth, invertible and strictly increasing function on R^1, z is a $q \times 1$ covariate vector, β is a $q \times 1$ coefficient vector, and ε is the random error with distribution function W and density $w > 0$ on R^1. The response T^* is continuous and z and β are bounded. For a general form of transformation models (4.1), if $h(\cdot)$ is completely known or its form is known but with some unknown parameters, then the distribution for error ε may be relaxed to be completely unknown. With a completely unknown $h(\cdot)$, however, we need at least a parametric form of the error distribution. In this chapter, we allow that the distribution of ε may depend on z and/or some unknown parameters and $h(\cdot)$ is completely unknown. The data $\{(T_i^*, z_i), i = 1, \ldots, n\}$ form independent replicates of (T^*, z). Our focus is on the estimation of β based on $\{(T_i^*, z_i), i = 1, \ldots, n\}$.

Model (4.1) covers a number of important models in survival analysis. If the error ε has an extreme value distribution $W(s) = 1 - \exp\{-\exp(s)\}$, then (4.1) becomes the well-known proportional hazards model; if W is the standard logistic or normal distribution, then (4.1) is the proportional odds model or the probit model respectively.

Different approaches have been proposed in the literature to estimate the semiparametric transformation model. In particular, when $h(\cdot)$ is totally unspecified, the *rank likelihood*, which uses the ranks of the responses and estimates β separately from h, has been proposed and studied by Cox (1972), Pettitt (1982), Clayton and Cuzick (1985), Dabrowska and Doksum (1988). In general, however, the rank likelihood is intractable and may not attain the same precision as the full likelihood. Other estimators requiring kernel smoothing have been studied by Horowitz (1996) and Wang and Ruppert (1996), among others. Recently, the *profile likelihood*, which estimates $h(\cdot)$ at observed failure times, was applied to the proportional odds model by Murphy *et al.* (1997), and the *sieve likelihood* was used to estimate the proportional odds model by Shen (1998). These procedures, however, may need to select a smoothing parameter, are computationally intensive, and involve complex distributional theory for the estimators. As a result, inference using these methods may be impractical (Cheung, *et al.* 2001).

While semiparametric transformation models have been extensively applied to the analysis of survival data, they have been seldom considered for long-term survivors, except in June (1996) and Subramanian (2001), which estimated the long-term survival rate using estimating equations. On the other hand, Maller and Zhou (1996) raised the issue of extending the Cox Proportional Hazards (PH) model to the analysis of survival data with long-term survivors, but they did not follow it through. Recently, Hu (1998) successfully extended the *pseudo maximum likelihood* approach, proposed by Gong and Samaniego (1981) for the parametric models, to *pseudo M-estimator and Z-estimator* approaches for semiparametric models. In this chapter, based on the ideas of Maller and Zhou (1996) on the PH model and of the pseudo Z-estimator approach, we propose a simple and explicit approach to estimate the unspecified transformation function $h(\cdot)$, the coefficients of covariates, and the proportion of long-term survivors in semiparametric transformation models. The large-sample properties of the estimators are derived based on the work of Hu (1998). Our proposed method has some merits over other types of methods. For example, it utilizes all the data to estimate the transformation function, so that the statistical inferences of the pseudo Z-estimators can overcome the drawbacks of intractability and imprecision of using rank likelihood, and the computational intensity of the kernel smoothing.

We would also point out that, while the idea of the "two-step" method is similar to Cheung and Fine (2001), our proposed method is different from theirs. More details on the difference are discussed in Section 4.2 below.

In Section 4.2, we specify semiparametric transformation models for the analysis of survival data with long-term survivors. Estimators of the transformation function h are given in Section 4.3. Section 4.4 presents pseudo Z-estimators of the coefficient vector β and the parameter p for the proportion of susceptibles. Large sample properties of the estimators are provided in Section 4.5. Section 4.6 reports some simulation results. An example of application is discussed in Section 4.7, followed by concluding remarks in Section 4.8.

4.2. Model Specification

Following the usual formulation in survival analysis, we postulate a "true" survival time T_i^* for each individual i, which is only observed if it does not exceed the censoring time c_i of individual i; otherwise, we observe c_i. We also know whether T_i^* is censored or not, through a censoring indicator $\delta_i = I(T_i^* \leq c_i)$. That is, $\delta_i = 1$ if T_i^* is an actual failure time (uncensored) and $\delta_i = 0$ if T_i^* is censored. The observable survival time T_i, possibly censored, is then given by $T_i = T_i^* \wedge c_i = \min(T_i^*, c_i),\ i = 1, \ldots, n.$

In general, both T_i^* and c_i are random variables, so are T_i and δ_i. We further assume that T_i^* is independent of c_i for each i, and $(T_i^*, c_i), i = 1, \ldots, n$, are independent pairs. In addition, c_i are assumed to have the same cumulative distribution function G, which is referred to as an independent and identically distributed (i.i.d.) censoring model. The distribution of T_i^*, on the other hand, are not necessarily identical, and may depend on such covariates as age, gender, treatment method, etc.

Let T^* denote a nonnegative random variable representing the survival time of an individual, with cumulative distribution function (cdf) $F(t, z)$, and z a $q \times 1$ vector of covariates associated with T^*. The Cox Proportional Hazards (PH) model specifies the survival function of T^* with covariate vector z by

$$S(t, z) = (1 - F_0(t))^{\exp(z^\top \beta)}, \tag{4.2}$$

where $F_0(t)$ is a baseline cdf, independent of covariates, and $\beta = (\beta_1, \ldots, \beta_q)^\top$ is an unknown vector of regression parameters (coefficients of covariates) to be estimated.

As discussed in Maller and Zhou (1996), an individual with survival time T^* is referred to as a "long-term survivor" (or a "cured" or "immune" individual) if the cdf $F(t, z)$ of T^* is *improper*, i.e., $F(\infty) = \mathrm{P}(T^* < \infty) < 1$. To incorporate possible existence of long-term survivors into the Cox PH model, we allow the baseline cdf of an individual's survival time to be improper with the form $pF_0(t)$. The parameter p can be interpreted as the proportion of "susceptible" individuals (who are not long-term survivors) when the covariates have no effects on the survival times. Therefore we propose, as suggested in Maller and Zhou (1996), to model the hazard function of a survival time with covariate z by

$$S(t, z) = (1 - pF_0(t))^{\exp(z^\top \beta)}. \tag{4.3}$$

Let $F_i(t) = F(t, z_i) = 1 - (1 - pF_0(t))^{\exp(z_i^\top \beta)}$, which yields

$$\log\{-\log[1 - F_i(t)]\} = z_i^\top \beta + \log\{-\log[1 - pF_0(t)]\}. \tag{4.4}$$

If T_i^* has distribution function $F_i(t)$, then we can write

$$h(T_i^*) = -z_i^\top \beta + \varepsilon_i, \tag{4.5}$$

where $h(t) = \log\{-\log[1 - pF_0(t)]\}$ and $\varepsilon_i = \log\{-\log[1 - F_i(T_i^*)]\}$. And $\{\varepsilon_i\}$ are independent random variables with cdf

$$V_i(x) = \begin{cases} 1 - \exp\{-\exp(x)\}, & \text{if } x < z^\top \beta + \log\{-\log(1-p)\}, \\ 1 & \text{otherwise,} \end{cases} \tag{4.6}$$

which is also an improper distribution function. From the definition of $h(t)$ in model (4.5) we see that $h(t)$ is an increasing and upper bounded function (as $h(t) \leq \log\{-\log[1-p]\}$).

We still call (4.5) a *semiparametric transformation model.* But with an improper cdf $pF_0(x)$, it appears to have been seldom investigated, and the distribution of error variables ε_i may depend on covariate z_i and/or some unknown parameters. In this chapter, we focus on the case of an unspecified (nonparametric) $h(t)$ in model (4.5). We first give a consistent estimator of $h(\cdot)$ via an empirical process, then apply the *pseudo Z-estimator approach*, inspired by Gong and Samaniego (1981) and further studied by Hu (1998), to obtain the estimators of covariate coefficients. Chueng and Fine (2001) also investigated semiparametric transformation model by using a "two-step" (or "three-step") method based on the works of Cheng *et al.* (1995,1997). But our proposed method is a new "two-step" method. We devise an estimator of the monotone transformation $h(\cdot)$ via an empirical process rather than rank estimators in Cheung and Fine (2001). More importantly, the method of Cheung and Fine was based on the works of Cheng *et al.* (1995, 1997), which rely on a completely specified distribution of the error variable ε_i, hence it is not applicable to our model.

To begin, following Doksum (1987), we assume that $T_1^*, \ldots, T_n^*$ follow model (4.5) with covariates satisfying

$$\sum_{i=1}^n z_{ij} = 0 \quad \text{and} \quad z = (z_{ij})_{n \times q} \text{ having rank } q. \tag{4.7}$$

4.3. Estimation of the Transform Function

In order to obtain the pseudo Z-estimator $\hat{\beta}$, we shall first give a consistent estimator of $h(t)$. Following Doksum (1987) we write $h(T_i^*) = -\mu_i + \varepsilon_i$ and $\mu_i = \sum_{j=1}^{q} z_{ij}\beta_j$, which are assumed not all zero. Let F_i and W denote the cdf's of T_i^* and ε_i respectively. Since h is an increasing function, we have

$$F_i(t) = \mathrm{P}(T_i^* \leq t) = \mathrm{P}(h(T_i^*) \leq h(t)) = V_i(h(t) + \mu_i)$$

and so $h(t) = V_i^{-1}\{F_i(t)\} - \mu_i$. By (4.7), $\sum_{i=1}^{n} \mu_i = 0$, hence

$$h(t) = \frac{1}{n}\sum_{i=1}^{n} V_i^{-1}(F_i(t)). \tag{4.8}$$

Note that in pseudo score function (see (4.16) in Section 4.4 below), we focus on $h(t)$ for $t < \tau_{F_0}$ and our consistent estimator for $h(t)$ is defined for $t < \tau_{F_0}$. By (4.6), $V_i(x)$ has a common form $1-\exp(-\exp(x))$ when $x < z_i^\top\beta + \log(-\log(1-p))$. For any observed time $t_i < \tau_{F_0}$, $h(t_i) + \mu_i < z_i^\top\beta + \log(-\log(1-p))$. Hence (4.8) can be rewritten as

$$h(t) = \frac{1}{n}\sum_{i=1}^{n} W^{-1}(F_i(t)), \quad t < \tau_{F_0}. \tag{4.9}$$

where $W(x) = 1 - \exp(-\exp(x)),\ x \in R$. As a result, the theorems in Doksum's (1987) for i.i.d. case can be applied and extended to our case of $h(t)$ in (4.9).

Doksum (1987) gave two consistent estimators of $h(t)$ with uncensored data. We now extend them to the case with censored data and long-term survivors.

4.3.1. Fixed parameters case

We consider the non-local case (see Doksum 1987) with fixed β_j and μ_i as the sample size increases. From (4.9), we can first estimate F_i and then $h(t)$. This can be done in the analysis of variance (ANOVA) model with multiple observations per cell. The ANOVA model can be written as $h(t_{jk}) = \theta_j + \varepsilon_{jk}$, $k = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, q$, where θ_j and n_j are the mean and sample size in cell j, respectively. Let $\lambda_{jn} = n_j/n$, where $n = n_1 + \cdots + n_q$, F_j be the cdf of t_{jk}, and $\hat{F}_j$ the Kaplan-Meier estimator of F_j in cell j. Assume $\lim_{n\to\infty} \lambda_{jn} = \lambda_j$, $0 < \lambda_j < 1$, $j = 1, 2, \ldots, q$. We can write $h(t) = \sum_{j=1}^{q} \lambda_{jn} W^{-1}(F_j(t))$ and estimate it by $\dot{h}(t) = \sum_{j=1}^{q} \lambda_{jn} W^{-1}(\hat{F}_j(t))$.

Throughout the rest of this chapter, we will assume that

$$\tilde{F}(t) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} F_i(t)$$

exists for all $t \in [0, \tau_{F_0}]$, where $\tau_{F_0} = \sup\{t : F_0(t) < 1\}$ is the right extreme point of F_0. Let $\tilde{T} = \max_{1\le i\le n} T_i$ and denote $h(\cdot \wedge \tilde{T})$ by $h_{\tilde{T}}(\cdot)$. Define $S_j(t) = 1 - F_j(t)$,

$$C_j(t) = \int_0^t \frac{dF_j(u)}{[1 - F_j(u)]^2[1 - G(u)]} \quad \text{and} \quad U(t) = \int_0^t \frac{dG(u)}{[1 - G(u)]^2[1 - \tilde{F}(u)]}, \tag{4.10}$$

where G is the censoring distribution function. We establish weak convergence of $\hat{h}_{\tilde{T}}(t)$ on the space $D(0, \tau_{F_0})$.

Theorem 4.1. *Suppose that W has a continuous derivative w bounded away from 0 and ∞ on $[W^{-1}(F_j(0)), W^{-1}(F_j(\tau_{F_0}))]$, F_j is continuous on $[0, \tau_{F_0}]$, $\tau_{F_0} \le \tau_G$, and $\sup_i \int_0^{\tau_{F_0}} [1 - G(u-)]^{-1} dF_i(u) < \infty$. Then the process $\sqrt{n}[\dot{h}_{\tilde{T}}(t) - h_{\tilde{T}}(t)]$ converges weakly on $D(0, \tau_{F_0})$ to the Gaussian process*

$$\sum_{j=1}^{q} \frac{\sqrt{\lambda_j} S_j(t) B_j(C_j(t))}{w\left\{W^{-1}(F_j(t))\right\}},$$

where $B_1(\cdot), \ldots, B_n(\cdot)$ are independent standard Brownian Motion processes.

Note that $W^{-1}(F_j(t)) = h(t) + \mu_j$, hence by Theorem 4.1, $\dot{h}(t)$ is approximately normally distributed with mean $h(t)$ and variance

$$\sum_{j=1}^{q} \frac{\lambda_j S_j(t)^2 C_j(t)}{w^2(h(t) + \mu_j)}.$$

4.3.2. Local parameter set

The method in Section 4.3.1 requires multiple observations per cell. If that is not available for a given data set, we may adopt an alternative approach described below. Following the notations of Doksum (1987), if $E(\varepsilon_i)$ exists, without loss of generality we can view $\mu_i = \sum_{j=1}^{q} z_{ij}\beta$ as the mean of $h(y)$. Moreover, by (4.7), $\bar{\mu} = n^{-1}\sum_{i=1}^{n} \mu_i = 0$. We will assume that β belongs to the set

$$\Omega_n = \left\{ \beta : \sum_{i=1}^{n} \mu_i^2 \le C^2, \max_{1\le i\le n} |\mu_i| \to 0 \right\}, \tag{4.11}$$

where C^2 is a constant independent of n, while β and μ_i may depend on n although this is suppressed in the notations.

Define $\bar{L}(\cdot) = 1 - L(\cdot)$ for any distribution function $L(\cdot)$. Define $F^{(n)}(y) = n^{-1}\#[T_i > y]$ to estimate each $\bar{H}_i(y)$, and let $\hat{\bar{G}}(y)$ be the Kaplan-Meier estimator for the censoring survival function $\bar{G}$. Because for each y,

$$\bar{W}(h(y) + \mu_i) = \bar{F}_i(y) = \frac{\bar{H}_i(y)}{\bar{G}(y)},$$

where $\bar{H}_i(y) = \mathrm{P}(T_i > y)$, $h(y)$ can be estimated naturally by

$$\dot{h}(y) = \bar{W}^{-1}\left(F^{(n)}(y)/\hat{\bar{G}}(y)\right).$$

Theorem 4.2 *Suppose that W has a continuous derivative w bounded away from 0 and ∞ on R^1, G is continuous on $[0, \tau_G]$, $\tau_{F_0} \le \tau_G$, $\int_0^{\tau_G} [1 - \tilde{F}(u-)]^{-1} dG(u) < \infty$, and $R(t_1, t_2) = \lim_{n\to\infty} n^{-1} F_i(t_1)F_i(t_2)$ exists for any $t_1, t_2 \ge 0$. Then the process $\sqrt{n}[\dot{h}_{\tilde{T}}(y) - h_{\tilde{T}}(y)]$ converges weakly on $(0, \tau_{F_0})$ to a Gaussian process.*

4.3.3. Proofs of Theorems 4.1-4.2.

Proof of Theorem 4.1. Write $u_{jn} = \hat{F}_j(t), u = F_j(t)$ and

$$D_{jn} = \sqrt{n}[W^{-1}(\hat{F}_j(t)) - W^{-1}(F_j(t))] = \frac{W^{-1}(u_{jn}) - W^{-1}(u)}{u_{jn} - u}\sqrt{n}[u_{jn} - u].$$

Note that $\mathrm{P}(T_i \geq \tau_{F_0}) > 0$ for the long-term survival time T_i^*, hence the conditions of Theorem 7.3.1 in Shorach and Wellner (1986, pp. 304-306) hold . Then for any $t \in (0, \tau_{F_0})$, this together with Theorem 3.1 of Gill (1983b) implies that $D_{jn}(t \wedge \tilde{T})$ converges weakly to $S_j(t)B_j(C_j(t))/\sqrt{\lambda_j} wW^{-1}(F_j(t))$. As a result,

$$\sqrt{n}[\hat{h}_{\tilde{T}}(t) - h_{\tilde{T}}(t)] = \sum_{j=1}^{q} D_{jn}(t \wedge \tilde{T}) \xrightarrow{d} \sum_{j=1}^{q} \frac{\sqrt{\lambda_j} S_j(t) B_j(C_j(t))}{w\{W^{-1}(F_j(t))\}}. \qquad \blacksquare$$

Proof of Theorem 4.2. Note that $\mathrm{P}(T_i \geq \tau_{F_0}) > 0$ for the long-term survival time T_i^*, hence for any $t \in (0, \tau_{F_0})$, $\sqrt{n}(\hat{\bar{G}}_{\tilde{T}} - \bar{G}_{\tilde{T}})$ converges weakly to a Gaussian process $[1 - G(t)]B(U(t))$ (Zheng, 1987, p.72, also Shorach and Wellner, 1986, pp. 327-328), where B is a standard Brownian process and $U(t)$ is defined in (4.9). Let $v_n = F^{(n)}(y)/\hat{\bar{G}}(y)$, $v = \bar{W}\,(h(y))$, and

$$D_n(y) = \sqrt{n}[\dot{h}(y) - h(y)] = \frac{\bar{W}^{-1}(v_n) - \bar{W}^{-1}(v)}{v_n - v} \sqrt{n}[v_n - v].$$

Then

$$\begin{aligned} v_n - v &= \frac{\hat{W}_n(h(y), \mu)}{\hat{\bar{G}}(y)} - \bar{W}\,(h(y)) \\ &= n^{-1} \sum_{i=1}^{n} \left[\frac{I[\varepsilon_i > (h(y) + \mu_i)]}{\hat{\bar{G}}(y)} - \bar{W}(h(y) + \mu_i) \right] \\ &+ n^{-1} \sum_{i=1}^{n} \left[\bar{W}(h(y) + \mu_i) - \bar{W}(h(y)) \right]. \end{aligned}$$

where the first term converges uniformly to a zero mean Gaussian process by the results in Appendix B of Cheng *et al.* (1997, p.234), and the second term converges uniformly to zero by the Taylor expansion $|\bar{W}(h(y) + \mu_i) - \bar{W}(h(y)| = w(h(y_0))\mu_i$ with $|h(y) - h(y_0)| \leq |\mu_i|$ (see Doksum,1987, p. 341). Finally, note that $[W^{-1}(v_n) - W^{-1}(v)]/(v_n - v)$ converges to $1/wW^{-1}(\tilde{F}(y))$ as in the proof Theorem 4.1. The proof is thus complete. $\blacksquare$

Remark 4.1. In any sample, $\dot{h}(y) = -\infty$ for $y \leq T_{(1)}$, $\dot{h}(y) = +\infty$ for $y \geq T_{(n)}$, where $T_{(1)}, \ldots, T_{(n)}$ are the order statistics of $T_1, \ldots, T_n$. To ensure that the estimator is finite outside the range of the data with small n, Cheung and Fin (2001) proposed a modified estimator of $h(\cdot)$, say $\hat{h}(\cdot)$, i.e.,

$$\hat{h}(y) = \{\dot{h}(T_{(1)}^{+}), y \leq T_{(1)}; \dot{h}(y), T_{(1)} < y < T_{(n)}; \dot{h}(T_{(n)}^{-}), y \geq T_{(n)}\}. \tag{4.12}$$

Since with probability 1 the interval $(0, \tau_{F_0})$ contains $(T_{(1)}, T_{(n)})$ as $n \to \infty$. Hence Theorems 4.1-4.2 but with $\dot{h}(\cdot)$ replaced by $\hat{h}(\cdot)$ also hold for $y \in (0, \tau_{F_0})$.

4.4. Pseudo Z-Estimators

Consider a set of random variables $T_i,\ i = 1, 2, \ldots, n$, to be observed from probability density functions p_{θ_0, w_0} belonging to the following class of semiparametric models:

$$\mathbf{P} = \{P_{\theta,w} : \theta \in \Theta \subset R^q, w \in W \subset R^l\},$$

which is indexed by two sets of unknown parameters:

(i) the parameters of interest $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$ representing the scientific objective; and

(ii) the nuisance parameters $w = (w_1, w_2, \ldots, w_l)$ that are needed in order to fully specify the probability model.

The parameters θ and w take values in some subsets of the q- and l-dimensional Euclidean spaces, respectively. As usual, let θ_0 and w_0 denote the true but unknown values of the parameters, and P_{θ_0, w_0} the true probability distribution that generates the observed data.

In order to find a value $\hat{\theta}$ from the set Θ that most likely represents θ_0 based on the observed data, we may use an estimation procedure called the pseudo-likelihood approach, proposed by Gong and Samaaniego (1981). Its key idea is to replace the true (but unknown) nuisance parameter w_0 with a consistent estimator $\hat{w} = \hat{w}(T_1, T_2, \ldots, T_n)$ in the likelihood function $L(\theta, w) = L(\theta, w | T_1, T_2, \ldots, T_n)$, where $\hat{w}$ is developed by some ad-hoc approach other than the maximum likelihood estimator (MLE). Then treat $L(\theta, \hat{w})$, which is called the *pseudo likelihood function*, as a usual likelihood function of θ only. The estimator $\hat{\theta}$ is then obtained by applying the usual method of maximum likelihood, but on the pseudo likelihood function $L(\theta, \hat{w})$. Such a $\hat{\theta}$ is called a *pseudo M-estimator.*

Instead of restricting to the likelihood-based approach, Hu (1998) considered more general estimation procedures. Let $\Psi : \Theta \times W \to R^{q+l}$ be a suitable deterministic function with values in Euclidean space R^{q+l}. The "true value" $(\theta_0, w_0) \in \Theta \times W$ is the solution of $\Psi(\theta, w) = 0$. A natural way to estimate θ_0 is to find a sequence $\{\Psi_n : \Theta \times W \to R^{q+l}\}$ (which need not be measurable) such that Ψ is the asymptotic version of $\{\Psi_n\}$ as $n \to \infty$, then replace w by some consistent ad-hoc estimator $\hat{w}$, and finally find $\hat{\theta}$ such that $\Psi_n(\hat{\theta}, \hat{w})$ is as close as possible to zero. This estimator $\hat{\theta}$ is called the *pseudo Z-estimator*. In some cases, however, the pseudo Z-estimator $\hat{\theta}$ may only satisfy: $\Psi_n(\hat{\theta}, \hat{w}) = 0_{p^*}(1/\sqrt{n})$. Hu (1998, p.41) also considered such cases and show that the asymptotic properties of $\hat{\theta}$ remain intact.

Hu (1998) further studied the pseudo M-estimator and Z-estimator approaches and successfully extended the work from parametric models to semiparametric models by allowing some components of the nuisance parameter $w = (w_1, w_2, \ldots, w_l)$ to be in an infinite-dimensional space, such as a class of uniformly bounded real-valued functions endowed with a seminorm $\| \cdot \|$. Large-sample properties of the pseudo M-estimator and Z-estimator are extensively discussed in his paper.

In this chapter, inspired by Gong and Samaniego (1981), Hu (1998), and Doksum (1987), we construct the pseudo Z-estimators of regression parameters in model (2.4).

Similar to the Z-estimators defined in Van der Vaart and Wellner (1996, Chap.

3.3), we can find pseudo Z-estimators $\hat{\beta}$ as follows. Suppose that a set of independent random vectors $(T_1^*, z_1), (T_2^*, z_2), \ldots, (T_n^*, z_n)$ are observed from the semiparametric transformation model (4.5). Note that $h(T_i^*) + z_i^\top \beta$, $i = 1, \ldots, n$, are independent random variables with a distribution function W_i defined by (4.6). Let $\beta \in \Theta_n$ be a finite (q-) dimensional parameter space and $h \in H \subset B$, where B is an infinite-dimensional Banach space. We also say that ε_i is censored if and only if T_i^* is censored.

By (4.5)–(4.6), given the covariate value z, the survival function of T^* can be rewritten as $S(t|z) = \exp\{-\exp[h(t) + z^\top \beta]\}$ for $h(t) \leq \log\{-\log(1-p)\}$. Hence the conditional density function is

$$f(t|z) = h'(t) \exp\left\{h(t) + z^\top \beta - \exp[h(t) + z^\top \beta]\right\}. \tag{4.13}$$

For possibly censored observations $\{(t_1, z_1, \delta_1), \ldots, (t_n, z_n, \delta_n)\}$, the likelihood function for (h, β) is then given by

$$L(h, \beta|T, \delta) = \prod_{i=1}^{n} [f(t_i|z_i)(1 - G(t_i|z_i))]^{\delta} [S(t_i|z_i) g(t_i|z_i)]^{1-\delta}, \tag{4.14}$$

where $G(\cdot|z)$ and $g(\cdot|z)$ are the censoring cdf and density function given z (generally, we assume that the censorship is independent of covariate z). By (4.13),

$$L(h, \beta|T, \delta) \propto \prod_{i=1}^{n} \left\{h'(t_i) \exp(h(t_i) + z_i^\top \beta)\right\}^{\delta_i} \exp\left\{-\exp(h(t_i) + z_i^\top \beta)\right\}, \tag{4.15}$$

which is valid for any $t_i \geq 0$ such that $h(t_i) \leq \log\{-\log(1-p)\}$.

Remark 4.2. If we replace $h(t)$ with some consistent ad-hoc estimator $\hat{h}(t)$ defined in (3.5) and $h'(t)$ with some consistent estimator, then we can find pseudo M-estimator $\hat{\beta}$ such that $L(\hat{\beta}, \hat{h}|T, \delta)$ nearly maximizes $L(\beta, \hat{h}|T, \delta)$ (see Hu (1998, p.56) and Van der Vaart (1995)). The estimation of $h'(t)$, however, is more difficult. Although a reasonable estimator of $h'(t)$ may be given by $\hat{h'}(t) = \Delta\hat{h}(t) = \hat{h}(t+) - \hat{h}(t)$, there is no guarantee that it is consistent. Fortunately, this difficulty can be sidestepped by an alternative approach, pseudo Z-estimator approach, which we will discuss below.

By (4.15), the score function of the likelihood function $L(h, \beta|T, \delta)$ is given by

$$\Psi(h, \beta|T, \delta) = \sum_{i=1}^{n} \dot{l}_\beta(\beta, h|T, \delta) = \sum_{i=1}^{n} \left\{ z_i \left[\delta_i - \exp(h(t_i) + z_i{}^\top \beta) \right] \right\}. \tag{4.16}$$

Then the *pseudo score function* is $\Psi(\hat{h}, \beta|T, \delta)$ and we can find pseudo Z-estimator $\hat{\beta}$ such that $\Psi(\hat{\beta}, \hat{h}|T, \delta)$ is as close to zero as possible. In Section 4.5, we will derive the asymptotic properties of the pseudo Z-estimator.

Finally, we mention the estimation of the susceptible proportion parameters p and $p_i = 1 - (1-p)^{\exp(z_i^\top \beta)}$, $i = 1, \ldots, n$. Recall $h(t) = \log\{-\log(1 - pF_0(t))\}$, so that $p = 1 - \exp\{-\exp(h(\infty))\}$. Thus a natural estimator of p is given by

$$\hat{p} = 1 - \exp\{-\exp(\hat{h}(T_{(n)}))\}, \tag{4.17}$$

where $T_{(n)} = \max_{1 \le i \le n} T_i$, and p_i can then be estimated by

$$\hat{p}_i = 1 - (1 - \hat{p})^{\exp(z_i^\top \hat{\beta})}, \quad i = 1, \ldots, n. \tag{4.18}$$

4.5. Asymptotic Properties of Parameter Estimators

4.5.1. Estimators of covariate coefficients

In this subsection, we investigate some asymptotic properties of estimators. We should mention that our results are based on the case of independence but with non-identical distribution (cf. Hu (1998, p.5) together with chap. 3.9 in Shorach and Wellner (1996)). We first introduce some notations which are convenient in the theory of empirical process. Throughout the rest of this section, we will denote, for every $(\beta, h) \in \Theta_n \times H$, the distribution function of $\tilde{X} = (T, \delta, z)$ by $P_{\beta,h}$ and its density function by $p_{\beta,h}$. Let P_n be the empirical distribution of $\tilde{X}_i, i = 1, \ldots, n$, and $P_nM = \int M dP_n = n^{-1}\sum_{i=1}^n M(\tilde{X}_i)$ for any function $M(x)$. Similarly, $PM = n^{-1}\sum_{i=1}^n \int M(x) dP^i_{\beta,h}(x)$ for the true distribution of $\tilde{X}_i$. Then the log-likelihood function for single observation is defined as $l(\beta, h|\tilde{X}) \equiv \log p_{\beta,h}(\tilde{X})$, and the score function for β is denoted by $\dot{l}_\beta(\beta, h|\tilde{X}) = \frac{\partial}{\partial\beta} l(\beta, h|\tilde{X})$. The true values of (β, h) is denoted by (β_0, h_0), which is in $\Theta_n \otimes H \subset R^q \times B$. The true distribution and density function P_{β_0,h_0} and p_{β_0,h_0} are sometimes simplified to P_0 and p_0, respectively. And throughout the rest of this chapter, we assume that $P_0\left[\dot{l}_\beta(\beta, h|\tilde{X})\right]^j$, $j = 1, 2$, exist in sense that

$$P_0\left[\dot{l}_\beta(\beta, h|\tilde{X})\right]^j = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \int \left[\dot{l}_\beta(\beta, h|x)\right]^j dP^i_{\beta,h}(x) < \infty, \quad j = 1, 2. \tag{4.19}$$

Let $\ddot{l}_{\cdot\cdot}$ be the second derivative of β or h. Also assume that $P_0\ddot{l}_{\cdot\cdot}(\beta, h|\tilde{X}) < \infty$ and Lyapunov condition holds for random variables $\dot{l}_\beta(\beta_0, h_0|\tilde{X})$. Hence for independent observations $\tilde{X}_1, \ldots, \tilde{X}_n$, the score function can be written as

$$\frac{1}{n}\sum_{i=1}^n \frac{\partial \log L(\beta, h|\tilde{X}_i)}{\partial\beta} = P_n\dot{l}_\beta(\beta, h|\cdot) = \frac{1}{n}\sum_{i=1}^n \dot{l}_\beta(\beta, h|\tilde{X}_i).$$

We define the parameter space as follows. Denote the set of all monotone functions $f : R \to [0, 1]$ by Q. For some large $A > 0$ and small $\eta > 0$, let

$$C_0 = \{\beta : \beta^\top\beta < A\}, \quad C_\eta = \{h \in H : \|h - h_0\| \le \eta\}, \tag{4.20}$$

where $H = \{h : h = \log\{-\log(1 - pF_0)\},\ F_0 \in Q,\ p \in (0, 1)\}$ and $\|\cdot\|$ is the usual supremum norm.

In order to prove our asymptotic properties Theorems 4.3-4.5, we need some Lemmas 4.1-1.4 and these lemmas follow from the following Conditions 1-5, thus we first state some conditions from Hu (1998, pp.43-52). Note that $p^*(1)$ in the following representations indicates that the left-hand side convergence to zero in outer probability in case that the term on the left is not Borel measurable.

Condition 1. (Stochastic Equicontinuity Condition)

$$\frac{|\sqrt{n}(P_n - P_0)\dot{l}_\beta(\hat{\beta}, \hat{h}) - \sqrt{n}(P_n - P_0)\dot{l}_\beta(\beta_0, h_0)|}{1 + \sqrt{n}|\hat{\beta} - \beta_0|} = o_{p^*}(1), \tag{4.21}$$

where $|\hat{\beta} - \beta_0| = o_{p^*}(1)$ and $\|\hat{h} - h_0\| = o_{p^*}(1)$.

Condition 2. $\sqrt{n} P_n \dot{l}_\beta(\beta_0, h_0) = O_{p^*}(1)$.

For i.i.d. observations and $\Theta_n \subset R^q$, the $P_n \dot{l}_\beta(\beta_0, h_0)$ in Condition 2 will be an average of the form $P_n \dot{l}_\beta(\beta_0, h_0)$. So Condition 2 holds automatically if $P_0 \dot{l}_\beta^2(\beta_0, h_0) < \infty$ by the central limit theorem (Hu, 1998, p.45). For independent but not identically distributed observations, Condition 2 holds if Lyapunov Condition holds for $\dot{l}_\beta(\beta_0, h_0)$.

Condition 3. (Smoothness Condition) Assume that $P_0 \dot{l}_\beta(\beta, h)$ is differentiable at (β_0, h_0) in the sense that there exists a continuous and nonsingular $q \times q$ matrix $P_0 \ddot{l}_{\beta\beta}(\beta_0, h_0) : (\Theta_n - \beta_0) \to R^q$ and a continuous linear function $P_0 \ddot{l}_{\beta h}(\beta_0, h_0) : \dot{H} \to R^q$ such that

$$
\begin{aligned}
&|P_0 \dot{l}_\beta(\beta, h) - P_0 \dot{l}_\beta(\beta_0, h_0) - P_0 \ddot{l}_{\beta\beta}(\beta_0, h_0)(\beta - \beta_0) - P_0 \ddot{l}_{\beta h}(\beta_0, h_0)(h - h_0)| \\
&\qquad = o(|\beta - \beta_0| + O(\|h - h_0\|)
\end{aligned}
$$

for $(\beta, h) \in D = \left\{(\beta, h) : |\beta - \beta_0| \leq \eta_n \downarrow 0, \|h - h_0\| \leq c n^{-1/2}\right\}$, where c is a constant.

Condition 4. $\sqrt{n} P_0 \ddot{l}_{\beta h}(\beta_0, h_0) |\hat{h} - h_0| = O_p(1)$.

Condition 5. under the true probability $P_{\beta_0, h_0} \equiv P_0$,

$$
\sqrt{n} \begin{bmatrix} (P_n - P_0) \dot{l}_\beta(\beta_0, h_0) \\ \hat{h} - h_0 \end{bmatrix} \xrightarrow{d} \Lambda = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix} \tag{4.22}
$$

where $\Lambda_1 = (\Lambda_{11}, \ldots, \Lambda_{1q})^\top \sim N_q(0, \Sigma_{11})$ with Σ_{11} being a $q \times q$ positive definite matrix, and $\Lambda_2 = \Lambda_2(t)$ is a zero-mean Gaussian process on T with auto-covariance function $\Sigma_{22}(t, t') = Cov(\Lambda_2(t), \Lambda_2(t'))$ for any $t, t' \in T$. The cross-covariance function between Λ_{1i} and $\Lambda_2(t)$ for $t \in T$ is denoted by $\Sigma_{1i2} = Cov(\Lambda_{1i}, \Lambda_2(t))$.

Remark 4.3. Conditions 1-4 are used to prove the rate of convergence for estimators and Condition 5 is for the proof of asymptotic normality of estimators.

In the work of Hu (1998, p.5, pp.48-49), it is allowed the rate of convergence for $\hat{h}$ to be n^{ϵ} for some $\epsilon > 0$. Thus it often requires a lot of work to verify Condition 4 for certain estimator $\hat{h}$ which converges at a slower rate than $\sqrt{n}$. However for our problems in this Chapter, we have established the $\sqrt{n}$-consistent estimator $\hat{h}$ in Theorems 4.1-4.2, then just as Hu (1998, p.49) show that this condition holds automatically for $\sqrt{n}-$consistent estimator, since the linear operate $\ddot{l}_{\beta h}(\beta_0, h_0)$ is continuous, which implies that $\sqrt{n} \mid \ddot{l}_{\beta h}(\beta_0, h_0)[\hat{h} - h_0] \mid \leq \sqrt{n}c \parallel \hat{h} - h_0 \parallel = O_{p^*}(1)$.

The following Lemmas 4.1–4.4 are due to Hu (1998, pp.42-56).

Lemma 4.1. (Consistency) *Suppose that* β_0 *is the unique solution to* $P_0\dot{l}_\beta(\beta, h_0) = 0$ *and* $\hat{h}$ *is an estimator of* h_0 *such that* $\|\hat{h} - h_0\| = o_{p^*}(1)$. *If*

$$\sup_{\beta\in\Theta_n, \|h-h_0\|\leq\eta_n} \frac{|P_n\dot{l}_\beta(\beta, h) - P_0\dot{l}_\beta(\beta, h_0)|}{1 + |P_n\dot{l}_\beta(\beta, h)| + |P_0\dot{l}_\beta(\beta, h_0)|} = o_{p^*}(1) \tag{4.23}$$

for every sequence $\{\eta_n\} \downarrow 0$, *then* $\hat{\beta}$ *almost solving equation* $P_n\dot{l}_\beta(\hat{\beta}, \hat{h}) = o_{p^*}(1)$ *converges in outer probability to* β_0.

Lemma 4.2. *Suppose that the class of functions*

$$\{\psi(\beta, h) : |\beta - \beta_0| < \gamma, \|h - h_0\| < \gamma\}$$

is P_0*-Donsker for some* $\gamma > 0$, *and that* $P_0|\psi(\beta, h|X) - \psi(\beta_0, h_0|X)|^2 \to 0$, *as* $|\beta - \beta_0| \to 0$ *and* $\|h - h_0\| \to 0$. *If* $\hat{\beta} \xrightarrow{p^*} \beta_0$ *and* $\|\hat{h} - h_0\| \xrightarrow{p^*} 0$, *then*

$$|\sqrt{n}(P_n - P_0)\left(\psi(\hat{\beta}, \hat{h}) - \psi(\beta_0, h_0)\right)| = o_{p^*}(1).$$

Note that the conditions of Lemma 4.2 provide a set of simple sufficient conditions for Condition 1 to hold, so we will only verify these conditions in Theorem 4.4 below.

Lemma 4.3. (Rate of Convergence) *Suppose that* $\hat{\beta}$ *satisfies* $P_n\dot{l}_\beta(\hat{\beta}, \hat{h}) = o_{p^*}(\frac{1}{\sqrt{n}})$ *and is a consistent estimator of* β, *which is the unique point for which* $P_0\dot{l}_\beta(\beta, h_0)$, *and $\hat{h}$ is an estimator of h_0 satisfying $\|\hat{h}-h_0\| = O_{p^*}(n^{-1/2})$ .Then under conditions 1-4, $\sqrt{n}(\hat{\beta}-\beta_0) = O_{p^*}(1)$.*

Lemma 4.4. (Normality) *Suppose that β_0 is the unique solution to $P_0\dot{l}_\beta(\beta, h_0) = 0$ and $\hat{h}$ is an estimator of h_0 satisfying $\|\hat{h}-h_0\| = O_{p^*}(1)$. Then under Condition 1 and Conditions 3-5, we have $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} \left(-P_0\ddot{l}_{\beta\beta}(\beta_0, h_0)\right)^{-1} N_q(0, V)$, where $V = Var(\Lambda_1 + P_0\ddot{l}_{\beta h}(\beta_0, h_0)\Lambda_2)$.*

Now we return to the asymptotic properties of the pseudo Z-estimator $\hat{\beta}$. From (4.15), we can see that the log-likelihood function for a single observation is given by $l(\beta, h|\tilde{x}) = \delta\left\{\log(h^{'}(t)) + h(t) + z^\top\beta\right\} - \exp\{h(t) + z^\top\beta\}$, and the first derivative of $l(\beta, h|\tilde{x})$ with respect to β is

$$\dot{l}_\beta(\beta, h|\tilde{x}) = z\left\{\delta - \exp(h(t) + z^\top\beta)\right\}. \tag{4.24}$$

Furthermore, since $E\left\{\exp(2h(T))\right\} < \infty$ (which holds for exponential $F_0(t)$) and

$$E(\delta^2) = E(\delta) = \mathrm{P}(T^* \le c) = \int_R [1 - G(u)]dF(u) < \infty, \tag{4.25}$$

then taking the population average in (4.24), we have

$$P_0\dot{l}_\beta(\beta, h|\tilde{X}) < \infty. \tag{4.26}$$

Also by Lebesgue's Dominated Convergence Theorem and (4.7),

$$P_0|\dot{l}_\beta(\beta, h)) - \dot{l}_\beta(\beta_0, h_0))|^2 = o_p(1), \tag{4.27}$$

where $|\beta-\beta_0| \le \eta_n \downarrow 0$ and $\|h-h_0\| \le cn^{-1/2}$. Note that β_0 is the unique point such that $\dot{l}_\beta(\beta, h_0|\tilde{x}) = 0$, we obtain $\hat{\beta}$ by solving $\dot{l}_\beta(\beta, \hat{h}|\tilde{x}) = 0$.

Theorem 4.3. *Suppose that covariate $|z| \le C$, C is a positive constant; $p \in (0, 1)$, parameter spaces are as in (4.20); $\hat{h}$ is a consistent estimator of h given in Section 4.3. Then $\hat{\beta}$ solves equation $P_n\dot{l}_\beta(\hat{\beta}, \hat{h}) = o_{p^*}(1/\sqrt{n})$ almost surely and converges in outer probability to β_0, where $\dot{l}_\beta(\beta, h)$ is defined in (4.24).*

Theorem 4.4. *Under the conditions of Theorem 4.3,* $\sqrt{n}(\hat{\beta} - \beta_0) = O_{p^*}(1)$.

Theorem 4.5. *If* β *is univariate, then under the conditions of Theorem 4.3 and (4.7),* $\sqrt{n}(\hat{\beta} - \beta_0)$ *is asymptotically normal with mean 0 and variance* V.

Remark 4.4. For the asymptotic variance V of $\sqrt{n}(\hat{\beta}-\beta)$ in Theorem 4.5, a precise representation can be found in Corollary 3.1.4 of Hu (1998, p.54) for our i.i.d. setup since there exists a zero-mean $\alpha(\tilde{X}, \beta_0, h_0)$ such that $\sqrt{n}P_0\ddot{l}_{\beta\beta}(\beta_0, h_0)[\hat{h} - h] = \sqrt{n}P_n\alpha(\cdot, \beta_0, h_0) + o_p(1)$, where $\alpha(\cdot, \beta_0, h_0)$ is defined by (3.1.21) in Hu (1998, p55). In such a case

$$V = Var[\dot{l}_\beta(\beta_0, h_0|\tilde{X})] + Var[\alpha(\tilde{X}, \beta_0, h_0)].$$

Without such an $\alpha(\tilde{X}, \beta_0, h_0)$, however, the close form of V is not available. In that case we can estimate V by bootstrap method, which will be discussed in Section 4.5.3 below.

4.5.2. Estimator of susceptible proportion

We next investigate the asymptotic properties of the estimators for susceptible proportion parameters p and p_i, which are defined in Section 4.4 as

$$\hat{p} = 1 - \exp\{-\exp(\hat{h}(T_{(n)})\}$$

and

$$\hat{p}_i = 1 - (1 - \hat{p})^{\exp(z_i\hat{\beta})}, \quad i = 1, \ldots, n,$$

where $T_{(n)} = \max_{1\leq i\leq n} T_i$.

T_i's are independent but not necessarily identically distributed. The cdf of T_i is given by $H_i = 1 - (1 - F_i)(1 - G)$, where $1 - F_i(t) = S(t, z_i)$ is defined in (4.4). Mimicking Theorems 3.1–3.2 in Maller and Zhou (1996), we have the following Theorems.

Theorem 4.6. *If* $\tau_{F_0} \le \tau_G$, $\sup_i |z_i| < C$, *and* β *is given in (4.20), then* $T_{(n)} \uparrow \tau_G$ *a.s.*

Theorem 4.7. *Suppose that* h *is continuous at* τ_G *in case* $\tau_G < \infty$. *Under the conditions of Theorems 4.2 and 4.6,* $\hat{p} \to p$ *a.s. and* $\sqrt{n}(\hat{p} - p)$ *converges in distribution to a zero mean normal variable.*

Similarly we can discuss the asymptotic properties of the estimators of $\hat{p}_i$, $i = 1, \dots, n$. The details are omitted here.

4.5.3. Bootstrap methods for standard errors of pseudo Z-estimators

In Subsections 4.5.1–4.5.2, we discussed the asymptotic properties of estimators. For the purpose of statistical inference, however, we also need the asymptotic variances of $\hat{\beta}$ and $\hat{p}$, which appear to be intractable due to lack of explicit forms. Therefore we resort to bootstrap methods to estimate the variances of the estimators.

The bootstrap technique was introduced by Efron (1979) originally as a tool for "estimating" ad-hoc-estimators that cannot be calculated explicitly due to lack of a closed presentation using a Monte-Carlo study (see also Efron, 1982, and Efron and Tibshirani, 1986). Meanwhile, results of Singh (1981), Bickel and Freedan (1981) and others show that bootstrap as an estimator for the distribution of a (standardized) estimator often gives an even better approximation to the true distribution than the limiting distribution. It may be expected that similar results would also hold for our model as well.

We emphasize that our model (4.5) is essentially from the Cox proportional hazards model. Since the PH model allows an "improper" baseline function, any bootstrap method for the PH model should naturally apply to our model as well. For the bootstrap method related to the PH model, Efron and Tibshirani (1986) simply re-sampled from the triples $(T_i, z_i, \delta_i),\ i = 1, \ldots, n$, ignoring the special structure provided by the PH model. This plan is likely to be "model-robust" (cf. Hjort, 1992) in that it is less sensitive to departures from the PH model. In this chapter, we follow Efron and Tibshirani (1986) and use the Monte Carlo algorithm in two steps to obtain our bootstrap estimators as follows.

(i) We sample with replacement from the triples $\{(T_1, z_1, \delta_1), \ldots, (T_n, z_n, \delta_n)\}$, which is given a probability mass of $1/n$ at each point $(T_i, z_i, \delta_i),\ i = 1, \ldots, n$. Hence we have bootstrap samples, say $d^*(1), \ldots, d^*(B)$. For each of the bootstrap data set $d^*(b)$, evaluate $\hat{h}^*(b)$ (hence $\hat{p}^*(b)$ and $\hat{p}_i^*(b)$) and $\hat{\beta}^*(b)$ for $b = 1, \ldots, B$;

(ii) Calculate the sample mean and standard deviation of the statistics of interest. For example, the mean and standard deviation of $\hat{\beta}^*$ are given respectively by

$$\hat{\beta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}^*(b) \quad \text{and} \quad \hat{\sigma}_\beta = \left(\frac{1}{B-1} \sum_{b=1}^{B} \{\hat{\beta}^*(b) - \hat{\beta}^*(\cdot)\}^2 \right)^{1/2}. \quad (4.28)$$

Before closing up this section, we would also mention other bootstrap schemes for the PH model. Recently, Karrison (1990), Loughin (1995, 1998) and Loughin and Koeler (1996) investigated alternative bootstrap techniques for the PH models, including a residual bootstrap (Loughin, 1995) and a semiparametric bootstrap (Loughin and Koehler, 1996). We may follow Loughin and Koehler (1996) to bootstrap data $(T_i^{*b}, c_i^*),\ i = 1, \ldots, n$, from $\hat{S}_i$ and $\hat{G}$, respectively while keeping $z_i^* = z_i$ and $\hat{S}_i(t) = \exp\left\{-\exp(\hat{h}(t) + z_i^\top \hat{\beta})\right\}$. $\hat{h}(t)$, $\hat{\beta}$ are reasonable estimators of h and β (cf. Hjort, 1992, p.384) and the Kaplan-Meier estimator $\hat{G}$ is a suitable estimator for the censoring distribution of G. This is a natural semiparametric bootstrap technique for the PH models. According to Karrison (1990), Efron and Gong (1983) remarked that these two methods are "asymptotically equivalent but can perform quite differently in small-sample situation". This bootstrap scheme may be of interest for further research work.

4.5.4. Proofs

Proof of Lemma 4.1. See Hu (1998, p.42) Theorem 3.1.1. ∎

Proof of Lemma 4.2. See Hu (1998, p.44) Lemma 3.1.1. ∎

Proof of Lemma 4.3. See Hu (1998, p.49) Theorem 3.1.3. ∎

Proof of Lemma 4.4. See Hu (1998, p52) Corollary 3.1.2. ∎

Proof of Theorem 4.3. To prove the consistency of pseudo Z-estimator $\hat{\beta}$, by Lemma 4.1 we only to show that

$$\sup_{\beta \in \Theta_n, \|h-h_0\| \leq \eta_n} |P_n \dot{l}_\beta(\beta, h) - P_0 \dot{l}_\beta(\beta, h_0)| = o_{p^*}(1) \tag{4.29}$$

for every sequence $\{\eta_n\}$ such that $\eta_n \downarrow 0$. Since

$$|P_n \dot{l}_\beta(\beta, h) - P_0 \dot{l}_\beta(\beta, h_0)| \leq |(P_n - P_0)\dot{l}_\beta(\beta, h)| + |P_0(\dot{l}_\beta(\beta, h) - \dot{l}_\beta(\beta, h_0))|,$$

and by (4.26) the second term obviously tends to zero when $|h - h_0| \leq \eta_n \downarrow 0$, it suffices to show that the class of functions $F_\eta \equiv \{\dot{l}_\beta(\beta, h) : \beta \in C_0 \subset R, |h - h_0| \leq \eta\}$ is a VC-class for some $\eta > 0$, where C_0 is defined in (4.20). This implies that the uniform strong law of large numbers holds, i.e., $\sup_{f \in F_\eta}(P_n - P_0)f \xrightarrow{p} 0$ (see Van der Vaart and Wellner (1996, Chap. 2.6–2.7) for details). We first note the following facts:

(1) The set Q of all monotone functions is a Donsker for every probability measure by Van der Vaart and Wellner (1996, Example 2.6.21). Also, similar to the proof of Lemma 5.1.1 in Hu (1998, p.110), we can verify that $Q_p = \{p_{F_0} : F_0 \in Q,\ p \in (0,1)\}$ is a VC-class by the boundedness of $p \in (0,1)$.

(2) As $\phi(x) = \log(-\log x)$ is a monotone function, $H = \{h = \log\{-\log(1 - p_{F_0})\} : F_0 \in Q,\ p \in (0,1)\}$ is a VC-class by Lemma 2.6.18 of Van der Vaart and Wellner (1996, p. 147). Hence the C_η defined in (4.20) is a VC-class.

These imply that F_η is a VC-class and so (4.29) holds. The consistency of $\hat{\beta}$ then follows from Lemma 4.1. ∎

Proof of Theorem 4.4. First note that

$$\dot{l}_\beta(\beta, h) - \dot{l}_\beta(\beta_0, h_0) = z\left\{\exp(h_0)\exp(z^\top\beta_0) - \exp(h)\exp(z^\top\beta\right\}. \tag{4.30}$$

To verify the stochastic equicontinuity condition:

$$|\sqrt{n}(P_n - P_0)(\dot{l}_\beta(\hat{\beta}, \hat{h}) - \dot{l}_\beta(\beta_0, h_0)| = o_{p^*}(1), \tag{4.31}$$

let $F_\gamma = \{z(\exp(h_0)\exp(z^\top\beta_0) - \exp(h)\exp(z^\top\beta)) : |\beta - \beta_0| \leq \gamma, \|h - h_0\| \leq \gamma\}$. Since $\phi(x) = \exp(x)$ is a monotone function, the class of functions F_γ is VC-class by Lemma 2.6.18 in Van der Vaart and Wellner (1996, p.147). Thus (4.31) follows from (4.27) and Lemma 4.2.

Next, the Condition 3 holds by (4.26)-(4.27). Moreover $P_n\dot{l}_\beta(\beta_0, h_0)$ converges in distribution to a normal random variable by the central limit theorem (under Lyapunov Condition). Thus $\sqrt{n}|\hat{\beta} - \beta| = O_{p^*}(1)$ by Lemma 4.3. ∎

Proof of Theorem 4.5. By Theorems 4.1 and 4.2 in Section 4.3 together with Slutsky's theorem and the central limit theorem, we can see that (4.22) holds with Λ_1 being normally distributed with mean zero and positive variance, and $\sqrt{n}(\hat{h}-h_0)$ converges weakly to Gaussian process Λ_2. Hence by Lemma 4.4, $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normal with mean 0 and variance $\{P_0\ddot{l}_{\beta\beta}(\beta_0, h_0)\}^{-2}V$, where $V = Var(\Lambda_1 + P_0\ddot{l}_{\beta h}(\beta_0, h_0)\Lambda_2)$. ∎

Proof of Theorem 4.6. Note that $\sup_i \tau_{H_i} = \tau_G$ and $\mathrm{P}(T_{(n)} \leq y) = \prod_{i=1}^n H_i(y)$. Under the assumptions of Theorem 4.7 we have, for any $y < \tau_G$,

$$\sum_{n=1}^{\infty} \mathrm{P}(T_{(n)} \leq y) = \sum_{n=1}^{\infty}\prod_{i=1}^{n} H_i(y) < \infty.$$

So $\sum \mathrm{P}(Y_{(n)} \leq y)$ converges for $y < \tau_G$ and $T_{(n)} \leq \tau_G$ a.s. It then follows from the Borel-Cantelli lemma that $T_{(n)} \to \tau_G$ a.s. ∎

Proof of Theorem 4.7. Note that

$$\hat{h}(T_{(n)}) - h(\tau_G) = \{\hat{h}(T_{(n)}) - h(T_{(n)})\} + \{(h(T_{(n)}) - h(\tau_G)\},$$

and $h(T_{(n)}) \to h(\tau_G)$ a.s. by Theorem 4.6, where $\hat{h}$ is defined in Theorems 4.1–4.2. Using Slutsky theorem we can prove that $\hat{p} \to p$ a.s., while the convergence of $\sqrt{n}(\hat{p} - p)$ to a normal distribution follows from the standard delta-method. ∎

4.6. Some Simulation Results

In the simulation study, we compare the performance of the pseudo Z-esti mators (PZEs) with the true values. The calculations can be seen more clearly in special cases. We consider a two-sample problem with exponentially distributed lifetimes. The two samples are of sizes, say, n_1 and n_2, respectively, $n_1 + n_2 = n$, with sample membership being indicated by the dummy variable

$$z_i = \begin{cases} -1, & \text{if individual } i \text{ is in sample 1} \\ 1, & \text{if individual } i \text{ is in sample 2} \end{cases}, \quad i = 1, 2, \ldots, n.$$

Data are generated from the survival functions $S_1(t) = (1 - pF_0(t))^{\exp(-\beta)}$ (with respect to sample 1) and $S_2(t) = (1 - pF_0(t))^{\exp(\beta)}$ (with respect to sample 2). Let $F_0(t)$ be an exponential distribution with parameter $\psi = 0.058$. The proportion of susceptible without covariate effects is $p = 0.90$ and the coefficient of covariates is $\beta = 0.3581$. Censoring times c are generated from a uniform distribution between 0 to 100. For this simulation, $\hat{h}$ is given in Theorem 4.2 and samples of sizes $n_1 = n_2 = 100$ and $n_1 = n_2 = 400$ were replicated 10000 times.

Let p_1 and p_2 denote the susceptible proportions in samples 1 and 2 respectively. Then $p_1 = 1 - (1-p)^{\exp(-\beta)} = 1 - 0.1^{\exp(-0.3581)} = 0.8000$ and $p_2 = 1 - (1-p)^{\exp(\beta)} = 1 - 0.1^{\exp(0.3581)} = 0.9629$. Let $(\hat{p}^{(1)}, p_1^{(1)}, p_2^{(1)}, \beta^{(1)})$ and $(\hat{p}^{(2)}, p_1^{(2)}, p_2^{(2)}, \beta^{(2)})$ denote the estimators of (p, p_1, p_2, β) under the pseudo Z-estimator approach with sample sizes $n_1 = n_2 = 100$ and $n_1 = n_2 = 400$ respectively. The means and the standard deviations of the simulated estimates are displayed in Table 4.1 below.

Table 4.1. Simulation results on the estimators of $\beta = 0.3581$ and $p = 0.9$

	$\beta^{(1)}$	$\beta^{(2)}$	$p_1^{(1)}$	$p_2^{(1)}$	$p_1^{(2)}$	$p_2^{(2)}$	$\hat{p}^{(1)}$	$\hat{p}^{(2)}$
$Mean$	0.3555	0.3565	0.7691	0.9483	0.7725	0.9504	0.8761	0.8788
STD	0.072	0.0388	0.0551	0.0201	0.0286	0.0101	0.0331	0.0168

From Table 4.1, we see that the pseudo Z-estimates are reasonably close to the true values of the parameters respectively, and the accuracy improves as the sample sizes increase from $n_1 = n_2 = 100$ to $n_1 = n_2 = 400$.

We also performed bootstrap analysis to estimate our parameters for comparing its effects with the above Monte Carlo simulations. The bootstrap samples are re-sample from one set of data $\{(T_i, z_i, \delta_i), i = 1, \ldots, n\}$ with sample size $n_1 = n_2 = 100$ and another set with $n_1 = n_2 = 400$, generated by the Monte Carlo simulations described above. In each case the number of resamples is $B = 200$. We then use (4.23) to calculate the means and standard deviations from the resamples. The results are reported in Table 4.2 below, where the notations are of the same meanings as in Table 4.1.

Note that the standard deviations in Table 4.1 are from replications with known parameter values, which cannot be obtained in practice with one sample. The standard deviations in Table 4.2, on the other hand, are calculated from the resamples of a single sample, and hence the same can be done in practical situations with one set of data.

Table 4.2 shows that the bootstrap pseudo Z-estimators are also close to the true values of the parameters. It is interesting to note that the bootstrap estimators give somehow better approximations to the true values of parameters than the averages of a large number (10,000) of replicates from ordinary Monte Carlo simulations, and the standard deviations obtained from the bootstrap are close to those from Monte Carlo simulations.

Table 4.2. Summary of bootstrap estimators of $\beta = 0.3581$ and $p = 0.9$

	$\beta^{(1)}$	$\beta^{(2)}$	$p_1^{(1)}$	$p_2^{(1)}$	$p_1^{(2)}$	$p_2^{(2)}$	$\hat{p}^{(1)}$	$\hat{p}^{(2)}$
Mean	0.3499	0.3517	0.7701	0.9484	0.7992	0.9579	0.8761	0.8928
STD	0.0831	0.0364	0.0521	0.0098	0.0285	0.0091	0.0228	0.0167

4.7. An Example of Application in Criminology

An important issue of interest in criminology is the "recidivism" of individuals, who return to prison (or are rearrested) some time after being released from the last imprisonment. Of interest is the proportion of recidivists among those released after their first imprisonment, and the rate of return to prison (or re-arrest).

If we view the time that an individual remains out of prison, or the time elapsed before a rearrest, as the "survival time", then we can fit the investigation into the area of survival analysis. The return times are censored by the need to restrict follow-up of released prisoners to a finite time, such as a "cutoff" date for the records. Since prisoners were released at different time points, however, the recidivism times may be censored at any point from the time of their initial release till the cutoff time. A proportion of released prisoners may never commit another crime – they are the "long-term survivors", also referred to as "immune" individuals, in our terminology.

We now demonstrate the application of the methodology in this chapter on a set of recidivism data consisting of 5324 individuals released from Western Australian prisons. The data record the times of successive arrests of these individuals with a number of covariates. For simplicity of illustration, we consider the "survival time" to the second arrests, and the effects of one covariate: "bail" or "no bail and the others". Individuals are classified into two groups: Group 1 consists of 2374 individuals of "bail" and Group 2 has 2950 individuals with "no bail and the others".

Figure 4.1 shows the Kaplan-Meier estimators for the second rearrest times, respectively, of the two groups, with the proportions 0.7673 for Group 1 and 0.8391 for Group 2.

The covariate z is taken as $z = 1$ for Group 1 and $z = -0.8047$ (cf. (4.7)) for Group 2. Based on our semiparametric transformation model, the pseudo Z-estimators of the parameters p, p_1, p_2, and the covariate coefficient β, are $\hat{p} = 0.8044$, $\hat{p}_1 = 0.7562$, $\hat{p}_2 = 0.8449$ and $\hat{\beta} = -0.1451$, which show that $\hat{p}_1$ and $\hat{p}_2$ are close to the Kaplan-Meier estimates 0.7673 and 0.8391.

Table 4.3 below gives the bootstrap pseudo Z-estimates and their standard deviations for the parameters.

Table 4.3. Bootstrap pseudo Z-estimates and standard deviations for the recidivism data

	$\hat{p}$	$\hat{p}_1$	$\hat{p}_2$	$\hat{\beta}$
Mean	0.8056	0.7586	0.8451	-0.1424
STD	0.0229	0.0257	0.0213	0.0182

4.8. Concluding Remarks

In this chapter we propose a new methodology to analyze survival data with long-term survivors and covariates, which in particular covers the widely applied proportional hazards models. Although semiparametric transformation models have been extensively studied for censored samples, the previous methods are mainly based on the likelihood functions and little has been done on semiparametric regression. Our approach is based on a semiparametric transformation of the response variable. We first use semiparametric techniques to estimate the unknown transform function, then use the pseudo Z-estimator method to estimate the coefficients of covariates. The asymptotic properties of the estimators are investigated, and the estimation of susceptible proportions is discussed based on the estimator of the semiparametric transformation. This new "two-step" estimation procedure provides a simple and effective methodology to analyze survival data with long-term survivors and covariates.

The simulation study indicates that the proposed procedure can produce efficient estimators for both semiparametric and parametric components of the model. An application of the proposed model is also demonstrated with a set of crime data on the issue of recidivism.

The semiparametric transformation models we have considered assume a linear relationship between the transformed response variable and the covariates. In practice, however, this may not always be adequate to model the relation between the response and covariates. Hence a useful extension to our work may allow a non-linear function $g(z)$ of covariates to replace the linear form $z^\top \beta$. The function $g(z)$ may be either parameterized or nonparametric. Alternatively, if the information suggests that $h(T)$ is linearly dependent on some covariate vector z, but unknown on other covariate w, then a partially linear model $h(T) = z^\top \beta + g(w) + \varepsilon$ may be considered. These extensions are certainly interesting topics that require further research efforts. Another issue worth for further study is whether and how alternative approaches, such as the Class-k method proposed by Zheng (1984), can be applied to the semiparametric transformation models considered in this chapter.

Acknowledgment Data for this chapter have been supplied by the Western Australia Police Service. The release and use of police data are subject to a number of caveats (see http://www.crc.law.uwa.edu.au/police_caveats.html).

We are grateful to Professor Ross A. Maller of Australian National University for providing us with the data in Section 4.7.

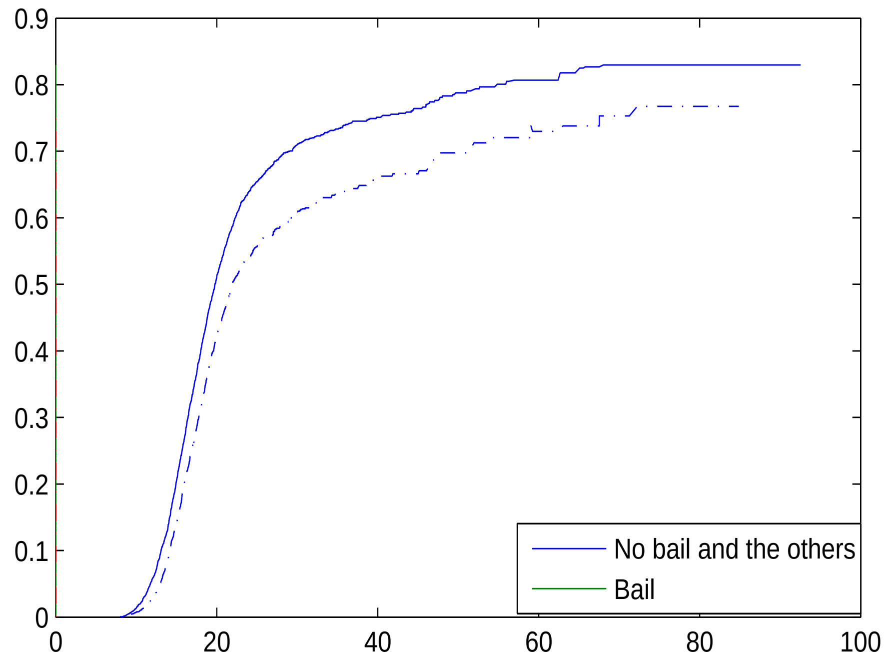


Figure 4.1. KMEs for the survival time to the second rearrest

Chapter 5

Discrete-time Survival Analysis for Survival Data with Long-term Survivors

5.1. Introduction

Conventional event history models typically assume that the entire population is at risk of experiencing the event of interest throughout the observation period. There are, however, many situations where some individuals will never experience such events as death from a particular cancer or arrest for a crime, who are commonly referred to as *cured individuals*, *immunes*, or *long-term survivors*. For continuous-time survival analysis, two types of models have been prevalent in handling survival data with long-term survivors. One is the *mixture model*, which was developed by Farewell (1982) and further researched by, among others, Kuk and Chen (1992), Peng and Dear (2000), Sy and Taylor (2000), Yamaguchi (1992), which assume a latent subpopulation that can be considered a priori to have a zero risk of experiencing the event. In such a situation, it may be preferable to model jointly the probability of long-term survivors and the timing of event occurrence. Another model, which will be referred to as the *non-mixture model*, incorporates the probability of cure into the proportional hazards model by assuming a bounded cumulative hazard. This model has been studied by Chen, Ibrahim and Sinha (1999), Tsodikov (1998) and Yakovlev and Tsodikov (1996) and so on, and is different from the mixture model in many aspects.

When covariates are included, however, mixture models do not have a proportional hazards structure even if the survival function for the timing of event occurrence does. For non-mixture models, on the other hand, Chen *et al.* (1999)

showed that if the survival function for the timing of event occurrence has proportional hazards, then the distribution for the entire population no longer does so. In order to retain the proportional hazards structure, Maller and Zhou (1996) proposed an alternative model via a different motivation from other cure models, but it has not been further investigated in the literature.

Although continuous-time survival models are suitable and frequently used in many settings, discrete-time survival models are often more natural in social and behavioral science applications, where the data typically possess three features:

(i) The event histories are discrete times. This is because in some situations events can only occur at regular, discrete time points. In other cases, the available data may only record particular time intervals in which each event occurs (Allison, 1982).

(ii) There are often ties among the uncensored event histories (failure times).

(iii) Survival data may contain concomitant information, such as demographics, account balances and payments associated with economic values for customer databases, and the occurrence of other events such as the acquisition of new products and services. The covariates are often time-dependent as present single irreversible events may occur at some point in the customer life time such as paying off an installment loan.

Therefore, discrete-time survival analysis models may be more appropriate to model such data as the economic values for customer retention, and are often well suited to analyze these kinds of data.

Discrete-time survival analysis techniques have recently been used to analyze *customer history data* extracted from operational customer databases, which contain histories of vital events such as cancelation of products and services (churn), downgrading, acquiring add-on products or upgrading, product return, loan prepayment etc. (Potts, 2004). The purpose was to build predictive models for customer retention, cross-selling and other database marketing endeavors; cf. Linoff (2004), Mosler (2003), Rosset *et al.* (2002), Baesens *et al.* (2004), Van den poel and Lariviere (2004), Lariviere and Van den poel (2004, 2005), Stepanova and Thomas (2000), among others. In order to ensure the success of a customer retention management (CRM) strategy, it is crucial that customers remain loyal at least to a certain extent to the company in case. This is a latest research topic of long-life customer within the group of customers in terms of spending and its evolution (Reinartz and Kumar, 2000). The economic value of long-life customers has been widely recognized in the literature, which includes: (a) Long-life customers buy more (Paulin *et al.*, 1998), and, if satisfied, may provide new referrals through positive word-of-mouth for the company (Colgate *et al.*, 1996); (b) Long-life customers become less costly to serve due to the bank's great knowledge of the existing customer and to decreased servicing costs (Paulin *et al.*, 1998). These economic values for customer retention with long-life customers can be naturally modeled by discrete-time model for survival data with long-term survivors - representing long-life customers.

There has been plenty of literature to investigate survival data with long-term survivors, such as Kuk and Chen (1992), Yakovlev (1992), Yakovlev *et al.* (1993), Maller and Zhou (1996), Chen *et al.* (1999), and Tsodikov *et al.* (2003), among others. However, apart from Stephen (2001), little has been done to model such discrete-time observations as the customer economic values with long-life customers by using existing models. Inspired by the works of Linoff (2004) and Potts (2004), in this chapter we will first review some existing discrete-time survival models that have been used to analyze survival data from social and behavioral science, and then extend them to accommodate long-term survivors. The proposed models and approaches can be directly applied to analyze survival data from social and behavioral science such as the economic values for customer retention with long-life customers.

In Sections 5.2-5.3 next, we elaborate on specifications of existing discrete-time survival models and their extensions to allow long-term survivors. Estimation problems are discussed in Sections 5.4-5.5, and large-sample properties of the proposed estimators are presented in Section 5.6. Section 5.7 reports some simulation results, and an example of application is shown in Section 5.8. Finally, Section 5.9 concludes.

5.2. Existing Model Specifications

In this section, we specify some discrete-time survival analysis models, which will be extended to accommodate long-term survivors in Section 5.3. Models in Subsections 5.2.1-5.2.3 have been discussed by many authors. The grouped related risk model was given by Kalbfleisch and Prentice (1973). The related risk model was introduced by Cox (1972), which has been discussed extensively with regard to inferential problems, and relevant references are given in Chapter 4 of Kalbfleisch and Prentice (2002). The discrete related risk model was proposed by Prentice and Kalbfleisch (2003). Cox (1972) suggested the linear logistic model for the discrete-time model. The piecewise exponential model can be found, for example, in Holford (1976), and was used by Potts (2004) to model customer event history data.

Let T_0 be a discrete random variable taking values $a_1 < a_2 < \cdots$ with associated probability function

$$f_0(a_i) = \Pr(T_0 = a_i), \quad i = 1, 2, \ldots, \tag{5.1}$$

and survival function

$$S_0(t) = \sum_{l:a_l>t} f_0(a_l). \tag{5.2}$$

The hazards at a_i, denoted by λ_{i0}, is defined as the conditional probability of failure at a_i given that the individual has survival to a_i, i.e.,

$$\lambda_{i0} = \Pr(T_0 = a_i | T_0 \geq a_i) = \frac{f_0(a_i)}{S_0(a_i-)}, \quad i = 1, 2, \ldots, \tag{5.3}$$

where $S_0(a-)$ denotes the left limit of $S_0(t)$ at a. Then the survival function and the probability function can be expressed respectively by (see Kalbfleisch and Prentice, 2002, pp.8-9)

$$S_0(t) = \prod_{i:a_i\leq t} (1 - \lambda_{i0}) \tag{5.4}$$

and

$$f_0(a_i) = \lambda_{i0} \prod_{l=1}^{i-1} (1 - \lambda_{l0}). \tag{5.5}$$

As in the continuous-time case, the discrete hazards function $(\lambda_{i0}, i = 1, 2, \ldots)$ uniquely determines the distribution of the failure time variable T_0.

5.2.1. Grouped related risk model

A discrete analog of the related risk model with fixed covariates can be obtained by applying the survival function relationship $S^s(t) = [S_0(t)]^{\exp(z^\top \beta)}$ directly to a discrete model. Let the failure time T^s given basic covariates z have a discrete distribution with masses at $0 \leq a_1 < a_2 < \cdots$. Let $z^\top = (z_1, z_2, \ldots, z_q)$ be a vector of covariates and β a vector of regression coefficients to be estimated, where $^\top$ denotes the transpose, and $S_0(t)$ represent the baseline survival function for $z = 0$. The corresponding survival function for covariates z is

$$S^s(t) = \prod_{i:a_i\leq t} (1 - \lambda_{i0})^{\exp(z^\top \beta)}. \tag{5.6}$$

The hazards function at a_i for covariate z is then

$$\lambda_i^s = 1 - (1 - \lambda_{i0})^{\exp(z^\top \beta)}. \tag{5.7}$$

It is of interest to note that the discrete model (5.7) can also be obtained by grouping the continuous model with $\lambda_0(t)$ being a continuous function of t,

$$\lambda^s(t, z) = \lambda_0(t) \exp(z^\top \beta). \tag{5.8}$$

Thus if continuous failure times arising from the related risk model (5.8) are grouped into disjoint intervals $[0, c_1), [c_1, c_2), \ldots, [c_{K-1}, c_K = \infty)$, the hazards of failure in the ith interval for an individual with covariate z is

$$Pr\left\{T^s \in [c_{i-1}, c_i) | T^s \geq c_{i-1}\right\} = 1 - (1 - \lambda_{i0})^{\exp(z^\top \beta)}, \tag{5.9}$$

where $\lambda_{i0} = \exp\left[\int_{c_{i-1}}^{c_i} \lambda_0(u) du\right]$. This discrete model is then the uniquely appropriate one for grouped data from the continuous related risk model. If the discrete baseline cumulative hazards function is written as $\Lambda_0(t) = \sum_{a_i \leq t} \lambda_{i0}$, then model (5.7) can be rewritten as

$$d\Lambda^s(t, z) = 1 - [1 - d\Lambda_0(t)]^{\exp(z^\top \beta)}. \tag{5.10}$$

Model (5.10) is grouped related risk model proposed by Kalbfleisch and Prentice (1973).

5.2.2. Discrete related risk model

In terms of hazards relationship, a simplest discrete, mixed or continuous model proposed and discussed by Kalbfleisch and Prentice (2003) is given by

$$d\Lambda^s(t, z) = d\Lambda_0(t) \exp(z^\top \beta), \tag{5.11}$$

which retains the multiplicative hazards relationship of the proportional hazards structure. Kalbfleisch and Prentice (2003) refer to this model as the discrete and continuous related risk model. In the continuous case, model (5.11) reduces to (5.8) and its survival function is given by $S^s(t) = [S_0(t)]^{\exp(z^\top \beta)}$. For the discrete case, by (5.4), the corresponding survival function can be depicted by

$$S^s(t) = \prod_{i: a_i \leq t} \left[1 - \lambda_{i0} \exp(z^\top \beta)\right], \tag{5.12}$$

where λ_{i0} is the hazards function corresponding to the baseline survival function $S_0(t)$. The model has the advantage of retaining the relative risk interpretation of multiplicative factor $\exp(z^\top \beta)$ for both the continuous and discrete cases (Kalbfleisch and Prentice, 2002, 2003).

5.2.3. Discrete logistic model

The discrete-time logistic model was proposed by Cox (1972), which specifies a linear log odds model for the hazards probability at each potential failure time. Thus if $\Lambda_0(t)$ is an arbitrary discrete or continuous cumulative hazards function, then the hazards for an arbitrary z is $d\Lambda^s(t,z)$, which satisfies

$$\frac{d\Lambda^s(t,z)}{1 - d\Lambda^s(t,z)} = \frac{d\Lambda_0(t)}{1 - d\Lambda_0(t)} \exp(z^\top \beta). \tag{5.13}$$

This is a linear logistic model with an arbitrary location parameter corresponding to each discrete failure time point. For a continuous $\Lambda_0(t)$, model (5.13) also reduces to (5.8). The effect of the covariates is to act multiplicatively, not on the hazards but on the discrete odds (Kalbfleisch and Prentice, 2003). Thus the interpretation of $\exp(z^\top \beta)$ is as an odds ratio rather than a related risk. This model was considered by many authors, such as Kalbfleisch and Prentice (2002, 2003) and Potts (2004).

After some algebra, we can find from (5.13) that for the discrete-time case,

$$\lambda_i^s = \frac{\exp(d_{i0} + z^\top \beta)}{1 + \exp(d_{i0} + z^\top \beta)}.$$

Hence the survival function is given by

$$S^s(t) = \prod_{i|a_i \le t} \left[1 - \frac{\exp(d_{i0} + z^\top \beta)}{1 + \exp(d_{i0} + z^\top \beta)} \right], \tag{5.14}$$

where $d_{i0} = \log[\lambda_{i0}/(1 - \lambda_{i0})]$.

5.2.4. Piecewise exponential model

The piecewise exponential model (Holford, 1976) allows for a wide variety of hazard rate shapes. Assume that the period of follow-up is divided into K intervals $(b_0 = 0, b_1], (b_1, b_2], \ldots, (b_{K-1}, b_K = \infty)$. It is further assumed that the baseline hazard rate $\lambda_{i0}(t)$ is constant over each interval $(b_{i-1}, b_i]$. If the hazards rate is approximated by a step function

$$\lambda_i^s(t) = \lambda_{i0}(t)\exp(z^\top\beta), \quad t \in (b_{i-1}, b_i], \tag{5.15}$$

then the survival times within each interval are exponentially distributed. The distribution function $F^s(t) = 1 - S^s(t)$ is then given by

$$F^s(t) = 1 - \exp\left\{-\lambda_i^s(t - b_{i-1}) - \sum_{l=1}^{i-1}\lambda_l^s(b_j - b_{j-1})\right\} \tag{5.16}$$

for $t \in (b_{i-1}, b_i]$, $i = 1, 2, \ldots, K$. We note that when $K = 1$, $F^s(t)$ reduces to the parametric exponential model. This model was also considered by Chen *et al.* (1999) (see, also Ibrahim *et al.*, 2001), where model (5.16) is referred to as a semiparametric model.

Before closing this section, we should mention that models (5.10), (5.11) and (5.13) may be encompassed by the following formulation

$$h[d\Lambda^s(t,z)] = h[d\Lambda_0(t)] + z^\top\beta, \tag{5.17}$$

where h is a monotone-increasing and twice-differentiable function mapping from [0,1] into $[-\infty, \infty]$ with $h(0) = -\infty$ (Kalbfleisch and Prentice, 2002). Obviously these models all reduce to the Cox model for the continuous case, and very similar if all of the discrete hazard contributions λ_{i0} are small (Kalbfleisch and Prentice, 2002). Furthermore, Steele (2003) also considered the logistic model in which the hazard rate in (5.4) is modeled as $\lambda_{i0} = \exp(z^\top\beta)/(1 + \exp(z^\top\beta))$ in the sense that the effect of covariates is on the baseline hazard rate λ_{i0}, which is different from the above models introduced in Sections 5.2.1-5.2.4.

5.3. Cure Models with Covariates

A cure model is applicable when there are "long-term survivors" present in survival data. As a result, cured subjects must be censored since they never fail. In contrast, "susceptible" subjects will eventually experience the event if they are followed for long enough. For continuous survival times it is common to model survival data with long-term survivors using mixture cure models. Thus the population is a mixture of two latent subpopulations: one consists of *susceptibles* who have a positive risk of experiencing the event, even though this may not be observed during the study period; and the other consists of *long-term survivors* (or *cured subjects*) who are not subject to the event of interest, hence will be observed up to the end of study and so always appear as right-censored.

Recently, non-mixture cure models have been proposed to model survival data with long-term survivors. In this section, we review the approaches for cure models with covariates. Three different types of approaches are prevalent in the literature, which are reviewed in Subsections 5.3.1–5.3.3. We will extend these models to accommodate long-term survivors.

5.3.1. Mixture cure models with covariates

A mixture model formulation is an attractive approach to analyzing such data, in that it contains two parts which can be interpreted separately by adding structure to the standard survival model. The model can be formulated as follows. Assume that the failure time can be decomposed as

$$T^* = \eta T^s + (1 - \eta)\infty, \tag{5.18}$$

where $T^s < \infty$ denotes the failure time of a susceptible subject and η indicates, by the value 1 or 0, whether the sampled subject is susceptible or not. If we assume the proportion of the susceptibles to be $\Pr(\eta = 1) = p$, where $p \in (0, 1]$, then the distribution function of T^* is given by

$$F(x) = \Pr(T^s \leq x)\Pr(\eta = 1) + \Pr(\infty \leq x)\Pr(\eta = 0) = pF_0(x) + 0 = pF_0(x),$$

where $F_0(\cdot)$ is the latent distribution function for T^s (susceptible group), or equivalently,

$$S(t) = pS_0(t) + (1 - p). \tag{5.19}$$

Common parametric choices for $F_0(t)$ are exponential and Weibull distributions. Nonparametric choices for $F_0(t)$ have also been considered. The effects of some independent covariates on both the incidence probability p and the survival function $S_0(t)$ for the susceptible group can be modeled. The incidence model is typically given by

$$p(x) = \frac{\exp(x^\top \gamma_1)}{1 + \exp(x^\top \gamma_1)}, \tag{5.20}$$

where x is a vector of covariates and γ_1 is a parameter to be estimated.

In the continuous case for $S_0(t)$, Farewell (1977, 1982) assumed a Weibull distribution

$$S_0(t) = \exp[-\phi t^\alpha], \tag{5.21}$$

where α is a parameter to be estimated and ϕ is modeled as $\phi = \exp(x^\top \gamma_2)$. Different formulations can also be used in the above setting, especially in the survival function (5.21) for the susceptible group. Yamaguchi (1992) applied a cure model with a logistic mixture probability model and an accelerated failure time model with a generalized gamma distribution. Maller and Zhou (1996) studied the cure model extensively, including nonparametric failure time models for one sample and parametric failure time regression models. Recent work has focused on nonparametric failure models. Taylor (1995) assumed a model with a logistic mixture probability and a completely unspecified failure time process, estimated by a Kaplan-Meier type estimator. In this section, we are particularly interested in the works of Kuk and Chen (1992), Sy and Taylor (2000) and Peng and Dear (2000), in which they considered a semiparametric Cox proportional hazards model for the failure time process.

For the discrete case of $S_0(t)$, Steele (2003) modeled it as

$$S_0(t) = \prod_{i:a_i \le t} (1 - \lambda_{i0})$$

with λ_{i0} of a logistic form

$$\lambda_{i0} = \frac{\exp(x_i^\top \gamma)}{1 + \exp(x_i^\top \gamma)}.$$

Note that for the continuous case of $S^s(t)$ our models (5.6), (5.12) and (5.14) all reduce to Cox model, thus model (5.19) reduces to the case of Kuk and Chen (1992). For the discrete case of $S^s(t)$, however, this is no longer the case and no literature other than Steele (2003) accommodates a cure model with a discrete-time susceptible distribution. In this section, we will model a cure model with a discrete-time $S^s(t)$. Following Kuk and Chen (1996), our idea is to model (5.19) with those $S^s(t)$ defined in Section 5.2, i.e., $S_0(t)$ is replaced by $S^s(t)$ in (5.19). As preliminary results, we now elaborate on the survival function $S(t)$ and the related hazard rate λ. Note from (5.19), (5.5) and (5.3) that the survival function $S(t)$, the probability function $f(a_i)$ and the hazard rate λ_i can be defined as below:

$$S(t) = pS^s(t) + (1 - p) = 1 - pF^s(t), \tag{5.22}$$

$$f(a_i) = S(a_i-) - S(a_i) = pf^s(a_i) = p\lambda_i^s \textstyle\prod_{l=1}^{i-1}(1 - \lambda_l^s), \tag{5.23}$$

$$\lambda_i = \frac{f(a_i)}{S(a_i-)} = \frac{pf^s(a_i)}{1 - pF^s(a_i-)} = \frac{pf^s(a_i)}{pS^s(a_i-) + (1 - p)}. \tag{5.24}$$

Then we obtain corresponding discrete-time cure models by inserting $S^s(t)$ (or $F^s(t)$), $f^s(a_i)$ and λ_i^s given in Subsections 5.2.1-5.2.4 into models (5.22)-(5.24). For example, for grouped related risk cure model we have

$$S(t) = p\left\{\prod_{i:a_i \le t} (1 - \lambda_{i0})^{\exp(z^\top \beta)}\right\} + (1 - p), \tag{5.25}$$

$$f(a_i) = p\left\{1-(1-\lambda_{i0})^{\exp(z^\top\beta)}\right\}\prod_{l=1}^{i-1}(1-\lambda_{l0})^{\exp(z^\top\beta)}. \tag{5.26}$$

The hazards function at a_i for covariate z is then given by $\lambda_i = f(a_i)/S(a_i-)$. Finally we should note that our structure for discrete-time cure models is different from that of Steele (2003), in which the effect on covariates is added to (5.19) by $S(t) = p\prod_{i:a_i\le t}(1-\lambda_{i0}) + (1-p)$ with $\lambda_{i0} = \exp(z_i^\top\beta)/(1+\exp(z_i^\top\beta))$.

5.3.2. An alternative mixture model with covariates

It is easy to see that in the presence of covariates, the aforesaid mixture cure models $S(t)$ does not have a proportional hazards structure even if $S^s(t)$ does. In this subsection, we study an alternative mixture model with covariates, which does have a proportional hazards structure and was proposed by Maller and Zhou (1996) but has not further investigated so far. Their idea is to extend Cox model $S(t) = [S_0(t)]^{\exp(z^\top\beta)}$ with a parametric or completely unspecified baseline $S_0(t)$ to "improper" (semiparametric baseline) Cox model defined by

$$S(t) = [S_0(t)]^{\exp(z^\top\beta)} = [1-pF_0(t)]^{\exp(z^\top\beta)}, \tag{5.27}$$

where $F_0(t)$ is a proper distribution function which may be parameterized or completely unspecified. We term this model as an *improper* proportional hazards model.

Assume that $F_0(t)$ is a discrete-time distribution function, its survival function is defined in (5.4) as $F_0(t) = 1-\prod_{l:a_i\le t}(1-\lambda_{l0})$. Then model (5.27) is further specified by

$$S(t) = \left[1-p+p\prod_{l:a_l\le t}(1-\lambda_{l0})\right]^{\exp(z^\top\beta)}. \tag{5.28}$$

The probability function $f(a_i)$ and the hazards rate function λ_i can be derived by the formulations of $f(a_i) = S(a_i-) - S(a_i)$ and $\lambda_i = f(a_i)/S(a_i-) = 1 - S(a_i)/S(a_i-)$. Thus the hazard function may be given by

$$\lambda_i = 1-\left[\frac{1-p+p\prod_{l=1}^{i}(1-\lambda_{l0})}{1-p+p\prod_{l=1}^{i-1}(1-\lambda_{l0})}\right]^{\exp(z^\top\beta)}. \tag{5.29}$$

Note that model (5.29) reduces to model (5.7) when $p = 1$.

We close this section by introducing anther generalization of the continuous failure model (5.27) rather than model (5.28). Note that (5.27) for continuous $\lambda_0(t)$ is equivalent to

$$\lambda(t) = \lambda_0(t) \exp(z^\top \beta),$$

where $\lambda_0(t) = pf_0(t)/(1-pF_0(t))$ and $F_0(t)$ is a continuous distribution with density function $f_0(t)$. Now if $F_0(t)$ (hence $f_0(t)$) is a discrete time distribution defined as before, then following (5.4) and (5.5) we have

$$\lambda_i = \frac{p\lambda_{i0} \prod_{l=1}^{i-1}(1-\lambda_{l0})}{1-p+p\prod_{l=1}^{i-1}(1-\lambda_{l0})} \exp(z^\top \beta). \tag{5.30}$$

Model (5.30) does have a proportional hazards structure for discrete-time variables, and its hazard rate is different from the one in (5.29) for model (5.28). Note that model (5.30) reduces to model (5.11) as $p = 1$ and is equivalent to model (5.29) if $\beta = 0$ (i.e., no effects of covariates).

5.3.3. Non-mixture cure models with covariates

In order to retain the proportional hazards structure, Yakovlev and Tsodikov (1996), Tsodikov (1998) and Chen, Ibrahim and Sinha (1999) proposed a model termed as "non-mixture" models, which can accommodate long-term survivors and also have the proportional hazards structure, and are different from model (5.27). In these models, the probability of cure is incorporated into the proportional hazards model by assuming a bounded cumulative hazard $\tilde{G}_0(t)$ with $\tilde{G}_0(\infty) = \zeta$. One way to enforce this is to write $\tilde{G}_0(t) = \zeta G_0(t)$, where $G_0(t)$ is the distribution function of a nonnegative random variable. Then the survival distribution $S(t)$ for the population can be written as

$$S(t) = \exp\{-\zeta G_0(t)\}. \tag{5.31}$$

We can see from (5.31) that the cured proportion is $\lim_{t\to\infty} S(t) = e^{-\zeta}$. Chen *et al.* (1999) showed that if $S(t)$ is taken to have a proportional hazards structure, then the conditional survival function for the susceptible group no longer has a proportional hazards structure. Hence in the non-mixture model, the survival distribution $S(t)$ for the entire population is modeled as a proportional hazard model, whereas in the mixture cure models, the non-cured group is often modeled as a proportional hazards model.

Covariates can be incorporated into the non-mixture cure model through ζ. One example is to use $\zeta(x) = \exp(x^\top \gamma)$. Tsodikov (1998) treated $G_0(t)$ as nuisance and used marginal likelihood to estimate the cure rate $\zeta(x)$. Chen *et al.* (1999) specified a parametric discrete form for $G_0(t)$ and used a Bayesian approach. Brown and Ibrahim (2003) extended this non-mixture cure model to include longitudinal covariates.

Inspired by Kuk and Chen (1999), in this section we propose an alternative form of $G_0(t)$ as $S^s(t) = [S_0(t)]^{\exp(z^\top \beta)}$, where $S^s(t) = 1 - G_0(t)$ is the survival function of $G_0(t)$, in which the covariates can also be modeled through $G_0(t)$. Then non-mixture model (5.31) can be re-specified as

$$S(t) = \exp\left\{\zeta S^s(t) - \zeta\right\}, \quad S^s(t) = [S_0(t)]^{\exp(z^\top \beta)}. \tag{5.32}$$

Model (5.32) was also proposed and studied by Tsodikov (2002) for the continuous distribution $S_0(t)$. When $S_0(t)$ is a discrete survival function defined in Section 5.2, proceeding along the line of Subsection 5.3.1, we can obtain the survival function $S(t)$ and the related hazard rate λ. This discrete distribution for $S(t)$ appears to have not studied so far.

Note from (5.32), (5.5) and (5.3) that the survival function $S(t)$ can be defined, such as by (5.6), by

$$S(t) = \exp\left\{\zeta\left[\prod_{i|a_i\le t}(1-\lambda_{i0})^{\exp(z^\top\beta)}\right]-\zeta\right\}. \tag{5.33}$$

Thus the probability function $f(a_i)$ and the related hazards rate function λ_i can be derived by the similar formulations in Subsection 5.3.1.

Other discrete-time non-mixture cure models can also be obtained by inserting $S^s(t)$ or $F^s(t)$, $f^s(a_i)$ and $\lambda_i{}^s$ defined in Section 5.2 into model (5.32).

5.4. Parametric Estimation for Discrete-time Models

We consider survival studies in which n items or individuals are put on test and data of form $\{T_j, \delta_j, z_j\}$, $j = 1, 2, \ldots, n$, are collected. Here δ_j is an indicator variable ($\delta_j = 0$ if the jth item is censored; $\delta_j = 1$ if the jth item failed), T_j is the corresponding failure or censoring time, and z_j is a vector of covariates that will be incorporated into the failure time model, which is presumed specified up to an unknown parameter vector β.

To obtain the likelihood function of β, it is necessary to consider the nature of the censoring mechanism. For most cases in survival analysis, we assume that the censoring times c_j's are i.i.d. random variables with survival function G, and are independent of the failure times T_j's and covariates z_j's. We only observed $T_j = \min(T_j^*, c_j)$ and $\delta_j = I(T_j^* \le c_j)$. This is called the i.i.d. random censoring mechanism.

Consider the discrete regression models discussed in Section 5.2, the inferences about the regression parameter β and the baseline cumulative hazards function Λ_0 based on the i.i.d. random censoring mechanism sample have been investigated by Kalbfleisch and Prentice (2002, Section 4.8), where they assume Λ_0 to be a discrete baseline cumulative hazards function with masses $\lambda_{10}, \lambda_{20}, \ldots, \lambda_{K0}$ at the discrete time points $a_1, a_2, \ldots, a_K$, where $0 < a_1 < a_2 < \cdots < a_K$, so that $\Lambda_0 = \sum_{j=1}^{K} \lambda_{j0} I(a_j \le t)$. Note that the number of possible mass points K included in the study is fixed so that the baseline hazards will be specified in terms of a finite number K of parameters. This allows straightforward asymptotic arguments applied to the maximum likelihood estimator (Kalbfleisch and Prentice, 2002).

A full maximum likelihood analysis of models (5.6), (5.12) and (5.14) were considered in Section 4.8.1 of Kalbfleisch and Prentice (2002) and they also considered some specific analyses available for models (2.12) and (2.14) in their Sections 4.8.2-4.8.3, where Breslow-Peto approximate partial likelihood were used to infer parameter β. Large-sample properties were also investigated by using discrete time martingale in Chap. 5 of Kalbfleisch and Prentice (2002). In this section, we sketch the estimation procedures proposed by Kalbfleisch and Prentice (2002) for the discrete-time models in Section 5.2 under right censoring (and possibly left truncation as well).

5.4.1. Maximum likelihood estimation

As in Kalbfleisch and Prentice (2002), for a fixed K we assume Λ_0 to be a discrete baseline cumulative hazards function with masses $\lambda_{10}, \lambda_{20}, \ldots, \lambda_{K0}$ at discrete times $a_1, a_2, \ldots, a_K$, where $0 < a_1 < a_2 < \cdots < a_K$, so that $\Lambda_0 = \sum_{j=1}^{K} \lambda_{j0} I(a_j \le t)$. Let D_i represent the set of labels attached to individuals failing at a_i and R_i the set of labels attached to individuals censored at a_i or observed to survive past a_i. Then the log-likelihood function of $(\lambda_0^\top, \beta^\top)$ can be written as (cf. Kalbfleisch and Prentice, 2002, p.137, (4.35))

$$\log L_g(\theta) = \sum_{i=1}^{K} \left\{ \sum_{j_i \in D_i} \log \lambda_{j_i}^{s} + \sum_{j_i \in R_i} \log\left(1 - \lambda_{j_i}^{s}\right) \right\}, \tag{5.34}$$

where $\lambda_{j_i}^{s}$ is the hazard function of the jth individual with covariates z_j and failure time point a_i, which is pertaining to the discrete-time survival functions $S^s(t)$

defined in Section 5.2. For the grouped related risk cure model,

$$\lambda_{j_i}^s = 1 - (1 - \lambda_{i0})^{\exp(z_j^\top \beta)}, \tag{5.35}$$

for discrete related risk model,

$$\lambda_{j_i}^s = \lambda_{i0}\exp(z_j^\top \beta), \tag{5.36}$$

and for discrete logistic model with $d_{i0} = \log[\lambda_{i0}/(1 - \lambda_{i0})]$,

$$\lambda_{j_i}^s = \frac{\exp(d_{i0} + z_j^\top \beta)}{1 + \exp(d_{i0} + z_j^\top \beta)}). \tag{5.37}$$

Denote $\lambda_0 = (\lambda_{10}, \lambda_{20}, \ldots, \lambda_{K0})^\top$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_q)^\top$. The components of the score vector $\theta = (\lambda_0^\top, \beta^\top)$ are

$$\frac{\partial \log L_g(\theta)}{\partial \lambda_{i0}}, i = 1, 2, \ldots, K, \quad \text{and} \quad \frac{\partial \log L_g(\theta)}{\partial \beta}.$$

Then the maximum likelihood estimator $(\hat{\lambda}_0^\top, \hat{\beta}^\top)$ is a solution to

$$\left(\frac{\partial \log L_g(\theta)}{\partial \lambda_{10}}, \ldots, \frac{\partial \log L_g(\theta)}{\partial \lambda_{k0}}, \frac{\partial \log L_g(\theta)}{\partial \beta_1}, \ldots, \frac{\partial \log L_g(\theta)}{\partial \beta_q}\right)^\top = 0.$$

Calculations of $(\hat{\lambda}_0^\top, \hat{\beta}^\top)$ by a Newton-Raphson iteration require second derivative of $\log L_g$. The Fisher's information matrix can be written as

$$H = -\begin{bmatrix} \frac{\partial^2 \log Lg}{\partial \lambda_0 \partial \lambda_0} & \frac{\partial^2 \log Lg}{\partial \lambda_0 \partial \beta} \\ \frac{\partial^2 \log Lg}{\partial \beta \partial \lambda_0} & \frac{\partial^2 \log Lg}{\partial \beta \partial \beta} \end{bmatrix}_{2\times 2}. \tag{5.38}$$

5.4.2. Approximate partial likelihood estimation for model (5.11)

We consider model (5.11) in Section 5.2. Just as mentioned above, model (5.11) does have a proportional hazards structure, hence Breslow-Peto approximate partial likelihood for tied observations can be applied to draw statistical inferences. To estimate β in this model, express the likelihood function as

$$L = \prod_{i=1}^{k} \frac{\exp[s_i^\top (a_i)\beta]}{\{\sum_{l\in R_i} \exp[z_l^\top (a_i)\beta]\}^{d_i}}. \tag{5.39}$$

where $s_i(a_i) = \sum_{l=1}^{d_i} z_l(a_i)$, d_i is the number of individuals failed at time a_i, and R_i is the set of individuals at risk just prior to time a_i. The likelihood in (5.39) gives rise to the following estimating equation for β:

$$U(\beta) = \sum_{i=1}^{K} [s_i(a_i) - d_i \mathcal{E}(\beta, a_i)] = 0, \tag{5.40}$$

where $\mathcal{E}(\beta, a_i) = \sum_{l \in R_i} z_l(a_i) \mathcal{P}_l(\beta, a_i)$ with

$$\mathcal{P}_l(\beta, a_i) = \frac{\exp(z_l^\top(a_i)\beta)}{\sum_{j \in R_i} \exp(z_j^\top(a_i)\beta)}.$$

As with the partial likelihood estimating equation, under suitable regularity conditions, a central limit theorem for standard version of the score $U(\beta)$ might be expected to apply as the sample size becomes large. Prentice and Kalbfleisch (2003) show that the estimator $\hat{\beta}$ that solves (5.40) is consistent for β, and further, that $\sqrt{n}(\hat{\beta} - \beta)$ is asymptotically normal with zero mean and covariance matrix that can be estimated using a standard sandwich type of estimator. Prentice and Kalbfleisch (2003) also show that an unbiased estimate of the variance of $U(\beta)$ is given by

$$V_U(\beta) = I(\beta) - \sum_{j=1}^{K} \sum_{l \in R(t_j)} [z_l(t_j) - \mathcal{E}(\beta, t_j)]^{\otimes 2} \exp[2 z_l^\top(t_j)\beta] \hat{\alpha}_0(\beta, t_j), \tag{5.41}$$

where

$$\hat{\alpha}_0(\beta, t_j) = \frac{d_j(d_j - 1)}{\{\sum_{l \in R(t_j)} \exp[z_l^\top(t_j)\beta]\}^2 - \sum_{l \in R(t_j)} \exp[2 z_l^\top(t_j)\beta]}, \tag{5.42}$$

and

$$I(\beta) = \sum_{j=1}^{K} d_j \mathcal{V}(\beta, t_j) \tag{5.43}$$

is the observed information arising from the Breslow-Peto approximate partial likelihood (5.40), and

$$\mathcal{V}(\beta, t_j) = \sum_{l \in R_j} [z_l(t_j) - \mathcal{E}(\beta, t_j)]^{\otimes 2} \mathcal{P}_l(\beta, t_j).$$

The asymptotic covariance of $\hat{\beta} - \beta$ is estimated by

$$I(\hat{\beta})^{-1} V_U(\hat{\beta}) I(\hat{\beta})^{-1}. \tag{5.44}$$

Prentice and Kalbfleisch (2003) present simulations that suggest this variance estimator to perform better than the naive estimator $I(\hat{\beta})^{-1}$ in yielding confidence intervals with more actuate converge.

A natural estimate for $\Lambda_0 = \sum_{i=1}^{K} \lambda_{i0}$ with given β is the Nelson-Aalen type estimator,

$$\hat{\Lambda}_0(\beta, a_i) = \sum_{j=1}^{i} \frac{d_j}{\sum_{l \in R(a_j)} \exp(z_l^\top(a_j)\hat{\beta})}, \quad i = 1, 2, \ldots, K. \tag{5.45}$$

Thus λ_{i0} can be estimated by $d_i \left\{ \sum_{l \in R(a_i)} \exp(z_l^\top(a_i)\hat{\beta}) \right\}^{-1}$. The asymptotic properties for $\hat{\Lambda}_0(\beta, a_i)$ were also considered in Prentice and Kalbfleisch (cf. pp. 115-116, Prentice and Kalbfleisch, 2002). Then the asymptotic properties for $\hat{\lambda}_{i0}$ can be obtained. We omit the details here.

We can see that (5.45) gave a reasonable estimator for λ_{i0}, which does, however, has some unsatisfactory feature, for example, the estimated hazard contribution can exceed one, this contradict the constraint that the hazard do not exceed one. Nonetheless, it is a reasonable estimate when hazards are relatively small and one is looking at events relatively early in the failure time distribution (Prentice and Kalbfleisch , 2002.)

Finally we mention that the partial likelihood for the discrete logistic model is just as (5.39), but with the parameter β measuring the effect on a logit scale. Moreover, the piecewise exponential model (5.16) has also been considered by Holford (1976).

5.5. Parametric Estimation for Discrete-time Cure Models

Consider customer event history data, which is observed under a different censoring and possibly truncation mechanism rather than the i.i.d random censoring. They are represented by an observed event time $T_j = \min(T_j^*, a - B_j), j = 1, 2, \ldots,$ where T_j^* is the target event time of the jth customer, and its survival function $S(t)$ is defined in Section 5.3. Since at the time the data was extracted for analysis, all customers usually have not experienced the event, only target event time $T_j^* \leq a - B_j$ can be observed; otherwise, it is considered to be (right) censored. The event indicator is $\delta_j = I(T_j^* \leq a - B_j)$. The data of origin, B_j, can vary among customers. Typically, B_j represents the data that an account was opened. In this censoring scheme (generalized type I censoring), there is a fixed point a, when the extracted data was current (for each customer the variables are measured from the moment they became customer until the moment of lapsing or censoring). Another possible cause of censoring is the occurrence of an independent and mutually exclusive competing event. For example, if event of interest is the cancelation of a service, then a customer that moves out of service area might be considered as being censored (Potts 2004).

The data used for mining customer histories consist of retrospective samples extracted from large operational databases. In some applications, the available data consist of a cross-sectional snapshot of customers that were active as of some fixed data c. Such a sample is considered to be truncated on the left. The sample is length-biased because, for given start data, B_j, only the lengthier event times appear in the sample (Potts 2004).

In many situation of survival analysis, however, we may encounter left and right censored (LCRC) data. In Section 5.5.1, we first discuss maximum likelihood estimation for three models for left and right censored data. For the case of left truncated and right censored (LTRC) data, we will sketch its maximum likelihood estimation in Remark 5.2 below. Then in Section 5.5.2, some specified technique, namely the Breslow-Peto approximate partial likelihood, will be applied to our model proposed in Subsection 5.3.2 under the LTRC mechanism.

Now assume that the event time T_j^*, which is distributed as $S(t)$ in Section 5.3, a discrete-time cure survival function with covariates z. However, our interests focus on the estimation for the proportion of the susceptible population. To start, we begin by constructing our likelihood function for proposed models in Section 5.3.

Here for brevity we only present three full likelihood functions for the grouped related risk cure mode in (5.25), the improper proportional hazards cure model in (5.29), and the discrete-time proportional hazards model in (5.30), respectively. The other combinations of discrete-time cure models proposed in Section 5.3 can be dealt with in the same way.

5.5.1. Maximum likelihood estimation

As in Section 5.4, we assume Λ_0 to be a discrete baseline cumulative hazards function with masses $\lambda_{10}, \lambda_{20}, \ldots, \lambda_{K0}$ at the discrete times $a_1, a_2, \ldots, a_K$, where $0 < a_1 < \cdots < a_K$, so that $\Lambda_0 = \sum_{j=1}^{K} \lambda_{j0} I(a_j \leq t)$. Assume that independent random variables T_j are distributed with a discrete survival function $S(t)$ defined in Section 5.3. T_j^* may be right censored at $c_j = a - B_j$. We only observe $T_j = \min(T_j^*, c_j)$ and $\delta_j = I(T_j^* \leq c_j)$. If the data are also left censored at $l_j = c - B_j$, then we only observe $X_j = \max\{\min(T_j^*, c_j), l_j\}$ and δ_j defined below:

$$X_j = \begin{cases} l_j, & T_j^* < l_j, \\ T_j^*, & l_j \leq T_j^* \leq c_j, \\ c_j, & T_j^* > c_j, \end{cases}, \quad \delta_j = \begin{cases} -1, & T_j^* < l_j, \\ 0, & l_j \leq T_j^* \leq c_j, \\ 1, & T_j^* > c_j. \end{cases} \tag{5.46}$$

It follows that the likelihood for the full sample is $L = \prod_{j=1}^{n} L_j$ where

$$L_j = f(x_j)^{(1-\delta_j^2)}[S(x_j)]^{(\delta_j^2+\delta_j)/2}[1 - S(x_j-)]^{(\delta_j^2-\delta_j)/2}. \tag{5.47}$$

Now let x_j denote the failure time at which the jth individual experiences the event in period a_i, or the censoring time of the jth individual if it is right-censored. Then we can rewrite (5.47) by using the hazards function as follow:

$$L_j = \lambda_{j_i}^{(1-\delta_j^2)} \prod_{k=1}^{j_i-1} [1-\lambda_{j_k}]^{[1+(\delta_j-\delta_j^2)/2]} [1-\lambda_{j_i}]^{(\delta_j^2+\delta_j)/2} \left[1-\prod_{k=1}^{j_i-1}(1-\lambda_{j_k})\right]^{(\delta_j^2-\delta_j)/2} \tag{5.48}$$

where λ_{j_i} is the hazard function of the jth individual with covariate z_j and failure time point a_i, corresponding to the discrete-time cure survival functions $S(t)$ defined in Section 5.3. For the grouped related risk cure model (cf. (5.24) and (5.7)),

$$\lambda_{j_i} = \frac{p\left[1-(1-\lambda_{i0})^{\exp(z_j^\top \beta)}\right]\prod_{l=1}^{i-1}(1-\lambda_{l0})^{\exp(z_j^\top \beta)}}{p\left[\prod_{l=1}^{i-1}(1-\lambda_{l0})^{\exp(z_j^\top \beta)}\right]+(1-p)}; \tag{5.49}$$

for the improper proportional hazards model (cf. (5.28) and (5.29)),

$$\lambda_{j_i} = 1-\left[\frac{1-p+p\prod_{l=1}^{i}(1-\lambda_{l0})}{1-p+p\prod_{l=1}^{i-1}(1-\lambda_{l0})}\right]^{\exp(z_j^\top \beta)}; \tag{5.50}$$

and for discrete-time proportional hazards cure model (cf. (5.30))

$$\lambda_i = \frac{p\lambda_{i0}\prod_{l=1}^{i-1}(1-\lambda_{l0})}{1-p+p\prod_{l=1}^{i-1}(1-\lambda_{l0})}\exp(z^\top \beta). \tag{5.51}$$

Thus the full likelihood estimator for $(p, \lambda_0^\top, \beta^\top)$ can be obtained by maximizing the likelihood function given by (5.47) or (5.48).

Remark 5.1: Assume that the duration of study is made up of K time *periods* (cf. (5.16)). A single nonrepeated event is considered so that data collection (and the observation of risk) is discontinuous for individual j in time period a_{j_i} for one of three reasons: 1) the jth individual experiences the event in a_{j_i}; 2) the individual drops out of the study in a_{j_i}; and 3) the individual experiences the event before a_{j_i} (cf. Muthen and Masyn, 2005, p. 32). In the first case, $T_j = a_{j_i}$. In the second case, it is only known that $T_j > a_{j_i-1}$, since the individual drops out of *during* the period a_{j_i}, whether $T_j > a_{j_i}$ is not known. In the last case, it is only known that

$T_j \leq a_{j_i}$, since the individual experiences the event *before* a_{j_i}, it is not known if $T_i \leq a_{j_i-1}$. For the failure time $T_j = a_{j_i}$, the likelihood may be expressed in terms of the hazard as

$$\Pr(T_j = a_{j_i}) = f_{j_i}(\theta) = \lambda_{j_i} \prod_{k=1}^{j_i-1} (1 - \lambda_{j_k}). \tag{5.52}$$

For individual with $T_j > a_{j_i-1}$, the likelihood is

$$\Pr(T_j > a_{j_i-1}) = S_{j_i-1}(\theta) = \prod_{k=1}^{j_i-1} (1 - \lambda_{j_k}). \tag{5.53}$$

For individual with $T_j \leq a_{j_i}$, the likelihood is

$$\Pr(T_j \leq a_{j_i}) = 1 - S_{j_i}(\theta) = 1 - \prod_{k=1}^{j_i} (1 - \lambda_{j_k}). \tag{5.54}$$

Thus the likelihood function is given by

$$L_j = f_{j_i}(\theta)^{(1-\delta_j^2)} [S_{j_i-1}(\theta)]^{(\delta_j^2+\delta_j)/2} [1 - S_{j_i}(\theta)]^{(\delta_j^2-\delta_j)/2}. \tag{5.55}$$

Remark 5.2: If the data are left-truncated and right-censored (LTRC), then we observe $X = \min(T, c)$, l and η, where $\eta = 1$ if $l \leq T \leq c$ and $\eta = 0$ if $c < T$, *conditional on* $T \geq l$. Nothing is observed if $T < l$. In effect, observations are made from the distribution of (X, l, η) conditional on $T \geq l$ (cf. Tsai, 1988). Then the conditional likelihood of the observed data $\{(x_j, l_j, \eta_j), j = 1, 2, \ldots, n\}$, given $T_j \geq l_j$, is

$$\mathcal{L} \sim \prod_{j=1}^{n} \frac{[f(x_j)]^{\eta_j} [S(x_j)]^{1-\eta_j}}{S(l_j-)} = \prod_{j=1}^{n} \left\{ [\lambda_{j_i}]^{\eta_j} \prod_{k: a_k \geq l_j}^{j_i} (1 - \lambda_{j_k}) \right\}, \tag{5.57}$$

where T_{j_i} denotes that the jth individual with covariate z_j fails at point a_i, $f(\cdot)$ and $S(\cdot)$ are defined in Section 5.2 (for the case $p = 1$). The conditional likelihood function defined in (5.57) is also used by Potts (2004) (see also Tsai, 1988, and Allison, 1982, page 74). For the case of $p \leq 1$ (with long-term survivors), we also have (5.56) (hence (5.57)) with $S(\cdot)$ defined in Section 5.2 replaced by $S(\cdot)$ defined in Section 5.3 (cf. (5.62) below).

Note from Section 5.4.1, where left truncation is allowed, that the large sample properties for the estimators solved from (5.57) are straightforwardly derived from that of Kalbfleisch and Prentice (2002). In fact (5.57) is equivalent to (5.34) as $p = 1$.

5.5.2. Approximate partial likelihood estimation for model (5.30)

Approximate partial likelihood estimators for model (5.30) are similar to that for model (5.11). The hazards function for the jth individual is not changed by left truncation/censoring at duration a_{j_i} because it is defined already conditional on "survival" until duration d and therefore it makes no difference to also condition on "survival" until a_{j_i} ($a_{j_i} < d$) (Hougaard (2000), Karlson (2005)). Then the inference procedure for model (5.11) can apply to model (5.30), so do the large-sample properties for the estimators.

Furthermore, since (cf.(5.30))

$$\lambda_i^s = \frac{p\lambda_{i0}\prod_{l=1}^{i-1}(1-\lambda_{l0})}{1-p+p\prod_{l=1}^{i-1}(1-\lambda_{l0})}, \tag{5.58}$$

a natural estimate for $\Lambda_0 = \sum_{i=1}^{K}\lambda_i^s$ for given β is the Nelson-Aalen type estimator,

$$\hat{\Lambda}_0(\beta, a_i) = \sum_{j=1}^{i}\frac{d_j}{\sum_{l\in R(a_i)}\exp\{z_l^\top(a_i)\hat{\beta}\}}, \quad i = 1, 2, \ldots, K. \tag{5.59}$$

Thus λ_{i0} and p can be estimated by (5.58). The asymptotic properties for $\hat{\Lambda}_0(\beta, a_i)$ and $\hat{\lambda}_i^s$ can also be obtained in a similar way as Prentice and Kalbfleisch (cf. pp. 115-116, Prentice and Kalbfleisch, 2002). Then the asymptotic properties for $\hat{\lambda}_{i0}$ and $\hat{p}$ would follow.

5.6. Asymptotic Properties for MLEs for LCRC Case

In this section, we discuss some asymptotic properties of maximum likelihood estimators for proposed models in Subsection 5.1. The sample properties for full maximum likelihood estimator for survival data subject to left and right censoring have been investigated by Liu (1996a,1996b), where the censoring are assumed to be random (but may be non-identically distributed) and the failure times are assumed to be continuous and i.i.d. random variables. In this chapter, we extend the existing results to the case where

(i) the survival functions for the failure time are assumed to be independent but not identically distributed with discrete-time distributions; and

(ii) survival times are subject to fixed but unequal left censoring; and generalized type I right censoring.

The following Lemma 5.1 for fixed (left and/or right) censoring case is straightforwardly derived from Lemma 2 in Ghitany *et al.* (1994), where the right censoring is assumed to be random. We will omit its proof. Following the notations defined above, we have

Lemma 5.1. *For $1 \leq j \leq n$, and any measurable function $Q : R \to R$,*

$$E\{(1-\delta_j^2)Q(X_j)\} = \int_{l_j}^{c_j} Q(s)dF_j(s) = \sum_{k|l_j \leq a_k \leq c_j} Q(x_k)f(a_{j_k}),$$

$$E\left\{\frac{\delta_j^2-\delta_j}{2}Q(X_j)\right\} = Q(l_j)[1-S(l_j)],$$

$$E\left\{\frac{\delta_j^2+\delta_j}{2}Q(X_j)\right\} = Q(c_j)S(c_j),$$

where $F_j(s) = 1 - S_j(s)$ and $S_j(s)$ is defined in, such as (5.25), but with z replaced by z_j; and $f(a_{j_i})$, $S(c_j)$ and $S(c_j)$ are defined in (5.61), (5.62) and (5.63) below.

Next we list some conditions (referred as L-Condition) below.

L-Condition: For any $i = 1, 2, \ldots, K$; $j = 1, 2, \ldots, n$,

(A) $f_{j_i}(\theta) > 0$ are continuous for $\theta = (p, \lambda_0^\top, \beta^\top)$.

(B) $\dfrac{\partial f_{j_i}(\theta)}{\partial \theta_s}$, $\dfrac{\partial S_{j_i}(\theta)}{\partial \theta_s}$, $\dfrac{\partial^2 f_{j_i}(\theta)}{\partial \theta_s \partial \theta_t}$, $\dfrac{\partial^2 S_{j_i}(\theta)}{\partial \theta_s \partial \theta_t}$ $(s, t = 1, 2, 3)$ are continuous for θ.

(C) For any $\theta_0 \in \Theta$, there exists $\mu_{\theta_0} = \{\theta : \|\theta - \theta_0\| \le \eta_{\theta_0}\} \subset \Theta$ such that for $\theta \in \mu_{\theta_0}$,

$$\left|\frac{\partial f_{j_i}(\theta)}{\partial \theta_s}\right| \le H_s^j, \quad \left|\frac{\partial^2 f_{j_i}(\theta)}{\partial \theta_s \partial \theta_s}\right| \le H_{st}^j, \quad \left|\frac{\partial^2 \log f_{j_i}(\theta)}{\partial \theta_s \partial \theta_s}\right| \le \Phi_{st}^j,$$

$$\left|\frac{\partial^2 \log S_{j_i}(\theta)}{\partial \theta_s \partial \theta_s}\right| \le \Phi_{st}^j, \quad \left|\frac{\partial^2 \log[1 - S_{j_i}(\theta)]}{\partial \theta_s \partial \theta_s}\right| \le \Psi_{st}^j,$$

where

$$\sum_{j=1}^{\infty} H_s^j < \infty, \quad \sum_{j=1}^{\infty} H_{st}^j < \infty, \quad \sum_{j=1}^{\infty} \left[\Phi_{st}^j\right]^2 f_{j_i}(\theta_0) < \infty,$$

$$\sup_j \left[\Phi_{st}^j\right]^2 S_{j_i}(\theta_0) < \infty, \quad \sup_j \left[\Psi_{st}^j\right]^2 [1 - S_{j_i}(\theta_0)] < \infty,$$

$f_{j_i}(\theta)$ and $S_{j_i}(\theta)$ are defined in (5.52) and (5.53).

Let $f(a_{j_i})$ denote the probability function that individual j with covariates z_j fails at time point a_i, such as (5.25) and (5.26) in grouped related risk cure model. Then we have

$$S(a_{j_i}) = p\left\{\prod_{k=1}^{i} (1 - \lambda_{k0})^{\exp(z_j^\top \beta)}\right\} + (1 - p), \tag{5.60}$$

$$f(a_{j_i}) = p\left\{1 - (1 - \lambda_{i0})^{\exp(z_j^\top \beta)}\right\} \prod_{k=1}^{i-1} (1 - \lambda_{k0})^{\exp(z_j^\top \beta)}. \tag{5.61}$$

(D) There exist $(T_i, \delta_i) \neq (T_j, \delta_j)$ for $i \neq j$, $i, j = 1, 2, \ldots, n$ and $n \ge 2$, such that an unique solving from (5.50).

(E) $l_i < c_i, i = 1, 2, \ldots, n$ (if the censoring and truncation are to assumed to be random, Condition E should be instead $\Pr(l < c) = 1$ (cf. Tsai, 1988, p.319).

(F) For any j, the following sequences are all linearly independent,

$$\frac{\partial \log f_{j_i}(\theta_0)}{\partial \theta_1}, \frac{\partial \log f_{j_i}(\theta_0)}{\partial \theta_2}, \ldots, \frac{\partial \log f_{j_i}(\theta_0)}{\partial \theta_{K+q+1}};$$

$$\frac{\partial \log S_{j_i}(\theta_0)}{\partial \theta_1}, \frac{\partial \log S_{j_i}(\theta_0)}{\partial \theta_2}, \ldots, \frac{\partial \log S_{j_i}(\theta_0)}{\partial \theta_{K+q+1}};$$

$$\frac{\partial \log[1-S_{j_i}(\theta_0)]}{\partial \theta_1}, \frac{\partial \log[1-S_{j_i}(\theta_0)]}{\partial \theta_2}, \ldots, \frac{\partial \log[1-S_{j_i}(\theta_0)]}{\partial \theta_{K+q+1}},$$

(G) $\sum_{j=1}^{\infty}(\frac{\partial \log f_{j_i}(\theta_0)}{\partial \theta_s})^4 f_{j_i}(\theta_0) < \infty, \quad \sup_j(\frac{\partial \log S_{j_i}(\theta_0)}{\partial \theta_s})^4 S_{j_i}(\theta_0) < \infty$

$$\sup_j(\frac{\partial \log[1-S_{j_i}(\theta_0)]}{\partial \theta_s})^4[1-S_{j_i}(\theta_0)] < \infty, s = 1, 2, \ldots, K+q+1.$$

Now we state our theorem for the asymptotic properties. The rigid proof for this theorem is similar to those in Liu (1996a,b) by using the law of large numbers and the central limit theorem for independent but non-identical distributions. We omit the details here for brevity.

Theorem 5.1. *Assume that Conditions (A)-(E) hold, and the maximum likelihood estimator $\hat{\theta}_n$ for θ is solved from (5.47). Then*

$$\mathrm{P}_{\theta_0}\left\{\lim_{n\to\infty}\hat{\theta}_n = \theta_0\right\} = 1 \ (\theta_0 \in \Theta).$$

Furthermore, if Conditions (F)-(G) also hold, then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, V^{-1}(\theta_0)),$$

where $\xrightarrow{d}$ denotes convergence in distribution, $V(\theta_0) = (v_{st}(\theta_0))_{(K+q+1)\times(K+q+1)}$ with

$$\begin{aligned} v_{st}(\theta) = & \lim_{n\to\infty} n^{-1}\sum_{j=1}^{n}\sum_{i:l_j\le a_i\le c_j}\frac{\partial \log f_{j_i}(\theta)}{\partial \theta_s}\frac{\partial \log f_{j_i}(\theta)}{\partial \theta_t}f_{j_i}(\theta) \\ & + \lim_{n\to\infty} n^{-1}\sum_{j=1}^{n}\frac{\partial \log S(c_j)}{\partial \theta_s}\frac{\partial \log S(c_j)}{\partial \theta_t}S(c_j) \\ & + \lim_{n\to\infty} n^{-1}\sum_{j=1}^{n}\frac{\partial \log[1-S(l_j)]}{\partial \theta_s}\frac{\partial \log[1-S(l_j)]}{\partial \theta_t}[1-S(l_j)], \end{aligned}$$

which can be derived from Lemma 1, and $S(c_j)$ and $S(l_j)$ are given by, from (5.60),

$$S(c_j) = p \prod_{k:a_k \le c_j} (1-\lambda_{k0})^{\exp(z_j^\top \beta)} + (1-p), \tag{5.62}$$

$$S(l_j) = p \prod_{k:a_k \le l_j} (1-\lambda_{k0})^{\exp(z_j^\top \beta)} + (1-p). \tag{5.63}$$

5.7. Simulation Results

We now investigate the performance of the proposed models and methods via some simulations. In this section we report three simulation results. The first is based on model (5.49) by using maximum likelihood estimator; the second is based on models (5.50) and (5.51) for two-sample problem by using maximum likelihood estimator; and the last is to investigate the approximate partial likelihood estimator under model (5.58) for two-sample problem. We simulated data from models (5.49)-(5.51),(5.58), with $p = 0.9$ and $\beta = 0.3581$. The baseline hazards λ_{i0}, valued at 0.2,0.375,0.3,0.7143,1, are from a discrete-time distribution with $K = 5$ failure time points, and whose probability functions are $f(a_i)$, valued at $0.2, 0.3, 0.15, 0.25, 0.1$, respectively. Note that $\lambda_{10} = f(a_1)$ and $\lambda_{K0} = f(a_K) \equiv 1$, hence we only need to estimate $\lambda_{i0}, i = 1, 2, \ldots, K-1$. The censoring distribution is assumed to be discrete uniform on $[3, 7]$ and the left censoring/truncations are discrete uniform on $[1, 3]$. The covariates z_i are generated from the uniform distribution over $[0, 1]$ for model (5.49). For two-sample problem models (5.50)-(5.51) and (5.58), we take $z_i = 0$ for group one and $z_i = 1$ for group two. The sample sizes are taken as $n = 500$ for model (5.49), $n_1 = n_2 = 250$ for model (5.50) and $n_1 = n_2 = 300$ for models (5.51) and (5.58), and the simulations are repeated 500 times. The simulation results are reported in Tables 5.1-5.4 below. From these tables we can see that the proposed models and related methods performed reasonably well.

Table 5.1. Maximum likelihood estimators for model (5.49)

	$\hat{p}$	$\hat{\beta}$	$\hat{\lambda}_{10}$	$\hat{\lambda}_{20}$	$\hat{\lambda}_{30}$	$\hat{\lambda}_{40}$
Mean	0.9072	0.3609	0.1991	0.3780	0.2951	0.7121
STD	0.0960	0.0148	0.0021	0.0068	0.0109	0.0049

Table 5.2. Maximum likelihood estimators for model (5.50)

	$\hat{p}$	$\hat{\beta}$	$\hat{\lambda}_{10}$	$\hat{\lambda}_{20}$	$\hat{\lambda}_{30}$	$\hat{\lambda}_{40}$
Mean	0.8898	0.3747	0.2059	0.3879	0.3102	0.7214
STD	0.0225	0.0769	0.0198	0.0217	0.0268	0.0400

Table 5.3. Maximum likelihood estimators for model (5.51)

	$\hat{p}$	$\hat{\beta}$	$\hat{\lambda}_{10}$	$\hat{\lambda}_{20}$	$\hat{\lambda}_{30}$	$\hat{\lambda}_{40}$
Mean	0.9068	0.3570	0.2011	0.3730	0.2922	0.7154
STD	0.0574	0.0136	0.0095	0.0119	0.0332	0.0136

Table 5.4. Approximate partial likelihood estimators for model (5.58)

	$\hat{p}$	$\hat{\beta}$	$\hat{\lambda}_{10}$	$\hat{\lambda}_{20}$	$\hat{\lambda}_{30}$	$\hat{\lambda}_{40}$
Mean	0.8952	0.3653	0.2025	0.3776	0.3029	0.7157
STD	0.0245	0.0092	0.0563	0.0654	0.0621	0.0256

5.8. An Example of Application with Reanalysis Bladder Tumors Recurrence Data

Byar (1980) discussed a randomized trial, conducted by the Veteran's Administration Cooperative Urological Group, among patients having superficial bladder tumors. One question of interest concerned the effect of the treatment thiotepa on the rate of tumor recurrence. Tumors present at baseline were removed transurethrally prior to randomization. In addition to the effect of the treatment, there was interest in the relation between the recurrence rate and the number of pre-randomization tumors as well as the size of the largest such tumors.

Prentice and Kalbfleisch (2003) investigated these data, and gave the parameter estimators based on the discrete-time survival analysis. However, its Kaplan-Meier estimator (cf. Figures 5.1-5.2) indicates the presence of "long-term survivors" in that data set. In this section, we will reanalyze these data with consideration of long-term survivors.

Table 5.5 below, extracted from Andrews and Herzberg (1985, pp. 254-259), shows a part of data from this trial, including possibly right censored times to first post-randomization recurrence. There were 48 patients assigned to the placebo group of whom 29 experienced at least one recurrence, and 38 patients assigned to thiotepa group of whom 18 experienced at least one recurrence, over a trial with 31 months of follow-up. Recurrence times were recorded monthly, which resulted in some tied recurrence times, including eight tied recurrence times at two and three months each, among others.

Table 5.6 below reports the analysis following a grouping of the recurrence times into six month intervals. We originally considered 10 intervals, but then merged the last four intervals since the last failure occurs in the seventh interval. We take the covariates to be 0 for Placebo Group and 1 for Thiotepa Group. Our analyses are based on models (5.50) and (5.51) for maximum likelihood estimation and model (5.51) for approximate partial likelihood estimation. In Prentice and Kalbfleisch (2003), only estimators for the coefficients of covariates were reported. In this chapter, we report the estimators of the susceptible proportions and the discrete-time baseline function $F_0(t)$, which are determined by the baseline hazards rates $\lambda_{i0}, i = 1, 2, \ldots, 7$. We see that our estimators for λ_{i0} is reasonable in view of (5.45) since the related estimators $\hat{\lambda}_{i0}$ all are less one. The results for the maximum likelihood estimators (MLEs), the estimators for the variances of the MLEs, and

Table 5.5. Bladder tumors recurrence data extracted from Andrews and Herzberg (1985, pp.254-259)

Initial Tumors[1]		Recurrence	Initial Tumors		Recurrence	Initial Tumors		Recurrence[2]
Number	Size	Time	Number	Size	Time	Number	Size	Time
Placebo			3	1	29	1	1	17
1	3	1*	1	2	37*	1	1	22*
2	1	4*	4	1	9	1	3	25*
1	1	7*	5	1	16	1	5	25*
5	1	10*	1	2	41*	1	1	25*
4	1	6	1	1	3	1	1	6
1	1	14*	2	6	6	1	1	6
1	1	18*	2	1	3	2	1	2
1	3	5	1	1	9	1	3	1*
1	1	12	1	1	18	8	3	26
3	3	23*	1	3	49*	1	1	38*
1	3	10	3	1	35*	1	1	22
1	1	3	1	7	17	6	1	4
3	1	3	3	1	3	3	1	24
2	3	7	1	1	59*	3	2	41*
1	1	3	3	2	2	1	1	41*
1	2	26*	1	3	5	1	1	1
8	1	1	2	3	2	1	1	44*
1	4	2	1	1	0*	6	1	2
1	2	25				1	2	45*
1	4	29*				1	4	2
1	2	29*	*Thiotepa*			1	4	46*
4	1	29*	1	2	9*	3	3	49*
1	6	28	1	1	10*	1	1	50*
1	5	2	1	1	13*	4	1	4
2	1	3	2	6	3	3	4	54*
1	3	12	5	3	1	2	1	38
1	2	32*	5	1	18*	1	3	59*
2	1	34*	1	3	17	8	1	5
2	1	36*	5	1	2	1	1	1*

[1]Initial number of tumors of eight denotes 8 or more. Size denotes the diameter of tumor in centimeters.

[2]Recurrence times are measured in months. An asterisk denotes right censoring.

the approximate partial likelihood estimators are given in Table 5.6. It shows that these estimators are reasonable, and the results for the baseline are close between models (5.50) and (5.51).

Table 5.6. Maximum likelihood estimates under models (5.50) and (5.51)

	$\hat{p}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\lambda}_{10}$	$\hat{\lambda}_{20}$	$\hat{\lambda}_{30}$	$\hat{\lambda}_{40}$	$\hat{\lambda}_{50}$	$\hat{\lambda}_{60}$
MLE(50)	0.575	0.141	0.086	−0.768	0.412	0.271	0.322	0.169	0.576	0.455
STD(50)	0.001	0.002	0.004	0.004	0.009	0.031	0.004	0.001	0.004	0.001
MLE(51)	0.573	0.127	0.075	−0.676	0.403	0.274	0.324	0.170	0.599	0.464
STD(51)	0.005	0.008	0.015	0.015	0.036	0.125	0.017	0.002	0.016	0.004
APLE(51)	0.581	0.119	0.073	−0.690	0.409	0.273	0.323	0.173	0.594	0.450

MLE(50): Maximum likelihood estimates for model (5.50)

MLE(51): Maximum likelihood estimates for model (5.51)

APLE(51): Approximate partial likelihood estimates for model (5.51)

where $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ are the estimators of the covariate coefficients corresponding to number of baseline tumors, size (cm) of largest baseline tumor and treatment (0-placebo; 1-thiotepa) respectively, and STD(5.50), STD(5.51) are the estimators of the standard deviations of the MLEs based on models (5.50), (5.51) respectively.

Prentice and Kalbfleisch (2003) analyzed this data set without considering long-term survivors. In this section, however, we fit the data using survival cure models. The first question is, are there really long-terms survivors in the data? This raises the problem of testing the null hypothesis $H_0 : p_0 = 1$. Such a hypothesis leads to a "boundary-value" test in which the parameter p_0 lies on the boundary of is parameter space $(0, 1]$. The large-sample properties of such boundary tests are now well understood. We can apply some recently derived theory of Vu and Zhou (1997) to obtain them easily for model (5.50). Let $\hat{\theta}$ denote the (unrestricted) maximum likelihood estimator of the parameter (vector) θ and $\dot{\theta}$ be the restricted maximum likelihood estimator of θ under H_0. Then the deviance statistic, defined as $d_n = 2(\mathcal{L}_n(\hat{\theta}) - \mathcal{L}_n(\dot{\theta}))$, is a 50-50 mixture of a chi-square random variable with 1 degree of freedom and a point mass at 0 asymptotically, where $\mathcal{L}_n(\hat{\theta})$ and $\mathcal{L}_n(\dot{\theta})$ are the values of the log-likelihood function at $\hat{\theta}$ and $\dot{\theta}$ respectively.

Following the above result, we next test H_0 based on models (5.50) and (5.51). The estimate $\hat{\theta}$ of $\theta = (p_0, \beta_1, \beta_2, \beta_3, \lambda_{10}, \ldots, \lambda_{60})$ is given in Table 5.6. For model (5.51), the restricted maximum likelihood estimators under H_0 is

$$\dot{\theta} = (1, 0.1450, 0.0886, -0.5924, 0.2174, 0.1184, 0.1117, 0.0489, 0.1846, 0.0569).$$

This gives a deviance $d_n = 15.84$ and a p-value less than 0.01, and so the null hypothesis $H_0 : p_0 = 1$ is strongly rejected. Therefore we can conclude with confidence that long-term survivors indeed exist in the data. Similarly for model (5.50), the maximum likelihood estimator of θ under H_0 is

$$\dot{\theta} = (1, 0.1635, 0.0966, -0.7756, 0.2335, 0.1213, 0.1150, 0.0505, 0.1891, 0.0590),$$

with a deviance $d_n = 16.74$. Thus we again obtain strong evidence for the presence of long-term survivors.

The strong evidence for long-term survivors in the above tests suggests that we should not ignore them in the analysis of the data in Table 5.5, and the estimates of $(\beta_1, \beta_2, \beta_3, \lambda_{10}, \ldots, \lambda_{60})$ given in Table 5.6 are more credible than those in $\dot{\theta}$ which ignore long-term survivors.

Furthermore, the estimates of the standard deviations of the approximate partial likelihood estimators $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ under model (5.51) (cf. (5.44)) are calculated as $(0.0824, 0.1089, 0.5309)$. Since $\hat{\beta}_2 = 0.073$, the 95% confidence interval of β_2 is $(-0.1403, 0.2864)$, which includes 0. Hence we would accept $H_0 : \beta_2 = 0$. This result, however, is not supported by the maximum likelihood models (5.50) and (5.51), which produce 95% confidence intervals for β_2 as $(0.0782, 0.0938)$ and $(0.0456, 0.1044)$ respectively, both exclude 0 and indicate $\beta_2 > 0$. This phenomenon may be due to the fact that the maximum likelihood estimation utilizes full information from the data, whereas the partial likelihood utilizes only partial information.

5.9. Concluding Remarks

In this chapter we reviewed four existing discrete-time survival models and then extend them to discrete-time cure survival models for survival data with long-term survivors. It is natural to use these discrete-time cure survival models to analyze the economic values of customer retention with long-life customers whose data typically possess three features: discrete time, ties, and concomitant information. Although discrete-time survival models have been used recently to model the economic values of customer retention, the proposed models for long-term survivors have not been previously addressed. We also investigated the estimation approaches - full maximum likelihood for the proposed discrete-time cure survival models and approximate partial likelihood for some special discrete survival data with long-term survivors subject to fixed but non-identical left truncation and right censoring. Then we discussed the asymptotic properties of the maximum likelihood estimators. The simulation study indicates that the proposed procedure can produce efficient and reasonable estimates. An application on a set of Bladder tumors recurrence data is also demonstrated.

The discrete-time cure survival models we have considered here assume time-independent covariates. In practice, however, this may not always be adequate to model the relation between the response and varying-time covariates or longitudinal data. Hence a useful extension to our work is to consider varying-time covariates to replace the time-independent covariates. Furthermore, the proposed estimating approaches in this paper are available to the interior case where the parameter space Θ is assumed to be an open set. Since the proportion of the susceptibles may be equal to 1, which lies on the boundary of the parameter space, another useful extension to our work in this chapter would include boundary case of the parameter space. Finally we mention that some covariates may be related to the incidence probability p, such as a logistic relationship discussed in Maller and Zhou (1996). These extensions are certainly worthwhile topics for further research efforts.

Figure 5.1.The Kaplan−Meier estimator for bladder tumors recurrence data

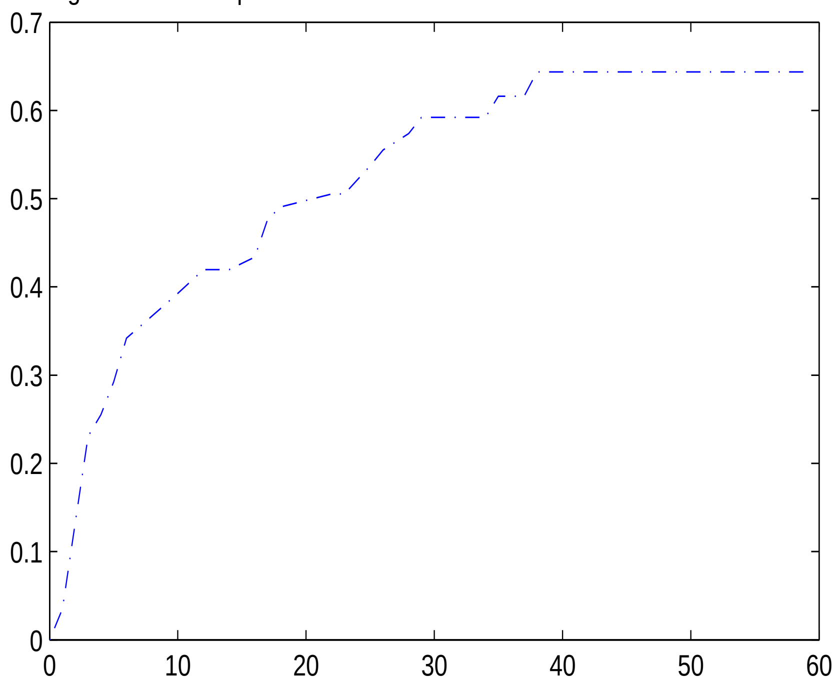
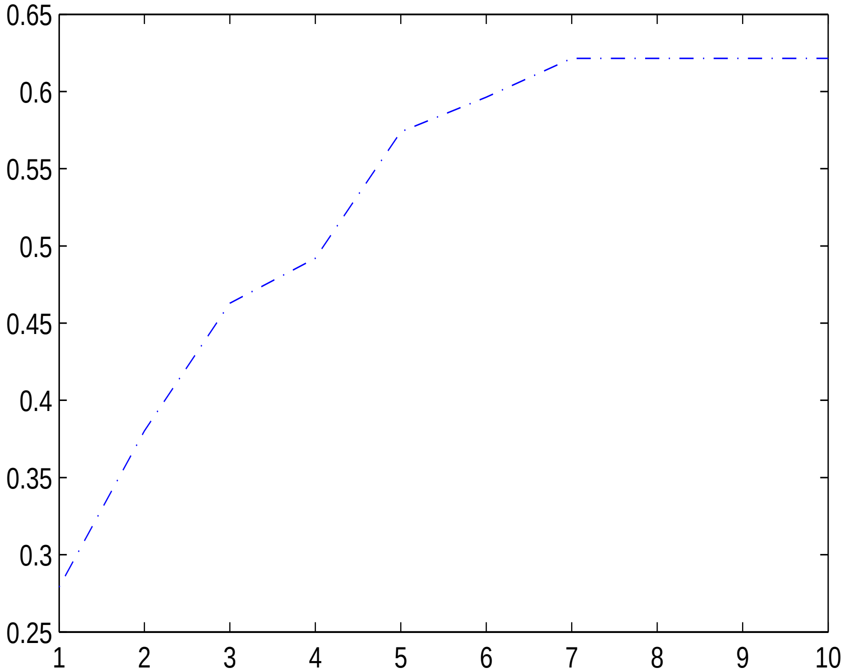


Figure 5.2.The Kaplan−Meier estimator for grouped bladder tumors recurrence data



References

[1] Aalen, O.O. (1975). *Statistical Inference for a Family of Counting Processes.* Ph.D Thesis, University of California, Berkeley.

[2] Allison, P.D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology,* **13**, 61-98.

[3] Andersen, P.K. Borgan, B. Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag Press.

[4] Andersen, P.K. and Gill, R.D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics*, **10**, 1100-1120.

[5] Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker,* Springer-Verlag: New York.

[6] Baesens, B., Verstraeten, G., Van den poel, D., Petersen, M.E., Kenhove, P.V., Vanth- ienen, J. (2004). Bayesian network classifier for identifying the slope of the customer lifecycle of long-life customers.*European Journal of Operational Research*, **156**, 508-523.

[7] Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics,* **9**, 1196-1217.

[8] Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley, New York.

[9] Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics*, **30**, 89-99.

[10] Brown, E.R. and Ibrahim, J.G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics*, **59**, 686-693.

[11] Byar, D.P. (1980). "The Veteran's Administration study of chemoprophylaxis of recurrent stage I bladder tumors:comparisons of placebo, pyridoxine, and topical thiotepa," in *Bladder Tumors and Other Topic in Urological Oncology,*(M. Pav one-Macaluso,P.H.Smith, and F.Edsmyn. eds.) Plenum: New York, pp. 363-370.

[12] Carroll, R. J., Maca, J. D. and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika*, **86**, 541-554.

[13] Chen, M.H., Ibrahim, J.G. and Sinha, D. (2001). *Bayesian survival analysis,* Springer-Verlag Press. New York.

[14] Chen, M.-H., Harrington, D.P. and Ibrahim, J.G. (1999). Bayesian models for high-risk melanoma. a case study of ECOG trial E1690. *Technical Report MS-06-99-22,*

Department of Mathematical Science, Worcester Polytechnic Institute.

[15] Chen, M.H., Ibrahim, J.G. and Sinha, D. (1999). A New Bayesian Model for Survival Data with a Survival Fraction. *Journal of the American Statistical Association*, **94**, 909-919.

[16] Cheng, S.C., Wei, L.J. and Ying, Z. (1997). Predictive survival probabilities with semiparametric transformation model. *Journal of the American Statistical Association*, **92**, 227-235.

[17] Cheng, S.C., Wei, L.J. and Ying, Z. (1995). Analysis transformation models with censored data. *Biometrika*, **82**, 835-845.

[18] Cheung, Y.K. and Fine, J.P. (2001). Likelihood estimation after nonparametric transformation. *Statistics and Probability Letters*, **55**, 1-7.

[19] Clayton, D.G. and Cuzick, J. (1985). The semiparametric Pareto model for regression analysis of survival data. *Bulletin de l'Institut International de Statistique,***51**, 175-180.

[20] Colgate,M., Stewart, K. and Kinsella,R. (1996). Customer defection: a study of the student market in Irelan *International Journal of Bank marketing*, **14**,23-29.

[21] Cox, D.R. and Oakes, D. (1984).Analysis of survival data. *Chapman and Hall.*

[22] Cox, D.R. (1975). Partial Likelihood. *Biometrika.* **62**, 269-276.

[23] Cox, D.R. and Hinkley, H. (1974). *Theoretical Statistics*, Chapman and Hall, London.

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

[25] Dabrowska, D.M. and Doksum, K.A. (1988). Partial likelihood in transformation models with censored data. *Scandinavian Journal of Statistics,* **15**, 1-23.

[26] Daniels, H. E. (1945). The statistical theory of the strength of bundles of threads. I. *Proceedings of the Royal Society. London. Series A*, **183**, 405-435.

[27] Doksum, K.A. (1987). An extension of partial likelihood methods for proportional hazards models to general transformation models. *The Annals of Statistics,* **15**, 325-345.

[28] Eckert, R. S., Carroll, R. J. and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics,* **53**, 262-272.

[29] Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confi-

dence intervals, and other measures of statistical accuracy. *Statistical Science,* **1**, 54-77.

[30] Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician,* **37**, 36–48.

[31] Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa.

[32] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics,* **7**, 1-26.

[33] Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, **13**, 342-368.

[34] Farewell, V.T. (1986). Mixture models in survival analysis: are they worth risk? *The Canadian Journal of Statistics*, **14**, 257-262.

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

[36] Farewell, V.T. (1977). A model for binary variable with time censored observation. *Biometrika*, **64**, 43-46.

[37] Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis*. John Wiley and Sons, Chichester.

[38] Ghitany, M.E., Maller, R.A. and Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. *Journal of Multivariate Analysis* , **49**, 218-241.

[39] Gill, R.D. (1984). Understanding Cox's regression model: martingale approach. *Journal of the American Statistical Association*, **79**, 441-447.

[40] Gill, R.D. (1983a). Discussion of the paper by Helland and Kurtz. *Bulletin of the International Statistical Institute.*, **50**, 239-243.

[41] Gill, R. (1983b). Large sample behavior of the product-limit estimator on the whole line. *The Annals Statistics,* **11**, 49-83.

[42] Goldman, A.I. (1984). Survivorship analysis when cure is a possibility:a Monton Carlo study. *Statistics in Medicine,* **3**, 153-163.

[43] Gong, G and Samaniego, F.J. (1981). Pseudo maximum likelihood estimation: theory and applications . *The Annals of Statistics,* **9**, 861-869.

[44] Hjort, N.L. (1992). On inference in parametric survival data. *International Statistics Review,* **60**, 355-387.

[45] Holford, T.R. (1976). Life tables with concomitant information. *Biometrics,* **32**, 587-597.

[46] Horowitz, J.L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable . *Econometrica,* **64**, 103-137.

[47] Hougaard, P. (2000). *Analysis of multivariate survival data.* Springer-Verlag Press. New York.

[48] Hu, H.L. (1998). *Pseudo maximum likelihood estimation for semiparametric models.* Ph.D. Thesis, University of Washington.

[49] Iturria, S.J., Carroll, R.J. and Firth, D. (1999). Polynomial regression and estimating functions in the presence of multiplicative measurement error. *Journal of the Royal Statistical Society. Series B. Statistical Methodology,* **61**, 547-561.

[50] June, S. (1996). Regression analysis for long-term survival rate. *Biometrika,* **83**, 227-232.

[51] Kalbfeisch, J.D. and Prentice, R.L. (2002). *The Statistical analysis of failure time data* (Second Edition). John Wiley and Sons, Chichester.

[52] Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihood based on Cox's regression and life model. *Biometrika*, **60**, 267–278.

[53] Karlsson, M. (2005). Analysis of interpurchase intervals of cars using truncated and censored data. *Research Reports of Department of Statistics in Umea University,Sweden.*

[54] Karrison, T. (1990). Bootstrapping censored data with covariates. *Journal of Statistical Computation and Simulation,* **36**, 195-207.

[55] Kendall, M. and Stuart, A. (1977). *The advanced theory of statistics. Vol. 1. Distribution theory. Fourth edition.* Macmillan Publishing Co., Inc., New York.

[56] Kong, F.H. and Gu, M.G. (1999). Consistent estimation in Cox proportional hazards model with covariates measurement error. *Statistica Sinica* , **9**, 953-969.

[57] Kong, F.H., Huang, W.H. and Li, X.M. (1998). Estimating survival curves under proportional hazards model with covariate measurement errors. *Scandinavian Journal of Statistics. Theory and Applications*, **25**, 573-587.

[58] KuK, A.Y.C. and Chen, C.H. (1992). A mixture model combining logistic regression

with proportional hazards regression. *Biometrika* , **79**, 531-541.

[59] Lariviere, B. and Van den poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications,* **29**, 272-284.

[60] Lariviere, B. and Van den poel, D. (2004). Investigating the role of product features in preventing customer churn by using survival analysis and choice modeling: the case of financial services. *Expert Systems with Applications,* **27**, 277-285.

[61] Li, C-S, Taylor, J. M. G. and Sy, J. P. (2001). Identifiability of cure models. *Statistics and Probability Letters* , **59**, 389-395.

[62] Liang, H., Hardle, W. and Carroll, R. J. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *The Annals of Statistics* , **27**, 1519-1535.

[63] Linoff, G.S. (2004). Survival data mining for customer insight. *Intelligente Enterprise, Skokie*, **7**, 28-33.

[64] Liu, H.B. (1996a). On the strong consistency of maximum likelihood estimators for truncated and censored data. *Journal of Central Chan Normal University*, **30**, 393–398.

[65] Liu, H.B. (1996b). Asymptotic normality of maximum likelihood estimators for truncated and censored data. *Journal of HuZhou Teachers Clllege*, **83**, 17–22.

[66] Loughin, T.M. (1998). On the bootstrap and monotone likelihood in the cox proportional hazards regression model. *Lifetime Data Analysis,* **4**, 393-403.

[67] Loughin, T.M. and Koehler, K.J. (1996). A semiparametric bootstrap for proportional hazards models. In *Lifetime Data: Models in Reliability and Survival Analysis,* N.P. Jewell, A.C. Kimber, M.-L.Lee, and G.A. Whitemore, eds., Kluwer Academic Publishers: Norwell,MA,pp. 207-212.

[68] Loughin, T.M. (1995). A residual bootstrap for regression parameters in proportional hazards models. *Journal of Statistical Computation and Simulation,* **52**, 367-384.

[69] Maller, R.A. and Zhou, X. (2002). Analysis parametric models for competing risks. *Statistica Sinica*, **12**, 725-750.

[70] Maller, R.A. and Zhou, X. (1996). *Survival analysis with long-term survivors.* John Wiley and Sons, Chichester.

[71] Mosler, K. (2003). Mixture models in econometric duration analysis. *Applied Stochastic Models in Business and Industry,* **19**, 91-104.

[72] Muthen, B. and Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Educational and Behavioral Statistics,* **30**, 27-58.

[73] 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.

[74] Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, **48**, 829-838.

[75] Nakamura, T. (1990). Corrected score function for error-in-variables models: methodology and application to generalized linear models. *Biometrika*, **77**, 127-137.

[76] Paulin, M. Perrien, J, Ferguson, R.J. Salazar, A.M.A. and Seruya, L.M. (1998). Relational norms and client retention: external effectiveness of commercial banking in Canada and Mexico *International Journal of Bank marketing*, **16**, 24–31.

[77] Peng, Y. and Dear, K.B.G. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, **56**, 237–243.

[78] Pepe, M., Self, S.G. and Prentice, R. L. (1989). Further results on covariate measurement errors in cohort studies with time to response data. *Statistics in Medicine,* **8**, 1167–1178.

[79] Pettitt, A.N. (1982). Inference for the linear model using a likelihood estimation based on rank. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* **44**, 234-243.

[80] Potts, W. (2004). Survival data mining. Technical white paper in Data Miners, http:// www.data-miners.com.

[81] Prentice, R.L. and Kalbfleisch, J.D. (2003). Mixed discrete and continuous Cox regression model. *Liftable Data Analysis*, **9**: 195–210.

[82] Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika,* **69**, 331–342.

[83] Reinartz, W.J. and Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: an empirical investigation and implications for marketing. *Journal of Marketing*, **64**, 17–35.

[84] Rosset, S., Neumann,E., Eick, U., Vatnik, N. and Idan, Y. (2002). Customer lifetime value modeling and its use for customer retention planning. *SIGKDD,*2,23-36.

[85] Self, G.S. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio test under nonstandard conditions. *Journal of the*

American Statistical Association, **82**, 605-610.

[86] Sen, P.K. and Singer, J.M. (1993). *Large sample methods in statistics.* Chapman-Hall, London.

[87] Shen, X.T. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika,* **85**, 165-177.

[88] Shorack, G.R. and Wellner, J.A. (1986). *Empirical processes with applications to statistics.* John Wiley and Sons.

[89] Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics,* **9**, 1187-1195.

[90] Steele, F. (2003). A discrete-time multilevel mixture model for event history data with long-term survivors, with an application to an analysis of contraceptive sterilization in Bangladesh. *Lifetime Data Analysis,* **9**, 155-174.

[91] Stefanski, L.A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics. Theory and Methods,* **18**, 4335–4358.

[92] Stepanova, M. and Thomas, L. (2000). Survival analysis methods for personal loan data. *Operations Research,* **50**, 277-289.

[93] Stephen, P.J. (2001). A discrete time split population survival (cure) model regression with censored data. http://fmwww.bc.edu/RePEc/usug 2001/UKSUGspj.pdf.

[94] Subramanian, S. (2001). Parameter estimation in regression for long-term survival rate from censored data. *Journal of Statistical Planning and Inference,* **99**, 211-222.

[95] Sy, J.P. and Taylor, J.M.G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics,* **56,** 227-236.

[96] Taylor, J.M.G. (1995). Semi-parametric estimation in failure time mixture model. *Biometrics,* **51**, 899-907.

[97] Thomas, A. (2004). An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions. *Scandinavian Journal of Statistics*, **31**, 43-50.

[98] Tsai, W.Y. (1988). Estimating of the survival function with increasing failure rate based on left truncated and right censored data. *Biometrika* , **75**, 319-324.

[99] Tsiatis, A.A. and Davidian, M. (2001). A semiparametric estimator for the propor-

tional hazards model with longitudinal covariates measured with error. *Biometrika,* **88**,447-458.

[100] Tsiatis, A. A. DeGruttola, V. and Wulfsohn, M.S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Application to survival CD4 counts in patients with AIDS. *Journal of the American Statistical Association* , **90**, 27-37.

[101] Tsiatis, A.A. (1981). A large sample study of Cox's regression model. *Annals of Statistics*, **9**, 93-108.

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

[103] Tsodikov, A. (2002). Semi-parametric models of long- and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage. *Statistics in Medicine* , **21**,895-920.

[104] Tsodikov, A.D. (1998). A proportional hazards model taking acount of long-term survivors. *Biometrics*, **54**, 1508-1516.

[105] Van den poel, D. and Lariviere, B. (2004). Customer attrition analysis for financial services using proportional hazards models. *Europeau Journal of Operational Research,* **157**, 196-217.

[106] Van der Vaart, A.W. and Wellner, J.A. (1996). *Weak convergence and empirical processes.* Springer-Verlag Press. New York.

[107] Van der Vaart, A.W. (1995). Efficiency of infinite-dimensional M-estimators. *Statistica Neerlandica*, **49**, 9-30.

[108] Vu, H.T.V., Maller, R.A. and Zhou, X. (1998). Asymptotic properties of a class of mixture models for failure data: the interior and boundary cases. *Annals of the Institute of Statistical Mathematics* , **50**, 627-635.

[109] Vu, H.T.V., Maller, R.A. and Zhou, X. (1997). Generalization of likelihood ratio tests under nonstandard conditions. *The Annals of Statistics* , **25**, 897–916.

[110] Wang, N. and Ruppert, D. (1996). Estimation of regression parameters in a semiparametric transformation model. *Journal of Statistical Planning and Inference,* **52**, 331-351.

[111] Wang, C. Y., Wang, S.J., Gutierrez, R. G. and Carroll, R. J. (1998). Local linear regression for generalized linear models with missing data. *The Annals of Statistics*

, **26**, 1028-1050.

[112] Wei, L.J. (1984). Testing goodness of fit for proportional hazards model with censored observations. *Journal of the American Statistical Association,* **79**, 649-652.

[113] Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint for survival and longitudinal data measured with error. *Biometrics*, **53**, 330-339.

[114] Yakovlev, A.Y. and Tsodikov, A.D. (1996). *Stochastic models of tumor latency and their biostatistical applications*, World Scientific, River Edge, New Jersey.

[115] Yakovlev, A.Y. (1994). "letter to the editor.". *Statistics in Medicine*, **13**, 983-986.

[116] Yakovlev, A.Y., Asselain,B., Bardon, V.J., Fourquet, A., Hoang, T., Rochefediere, A., and Tsodikov, A. D.(1993). "A simple stochastic model of tumor recurrence and its applications to data on premenopausal breast cancer" , in *Biometrie et Analyse de Donnees Spatio-temporlles"* , 12, eds. B. Asselain, M. Boniface,C.Duby,C. Lopez,J.P. Masson, and J. Tranchefort; France, Rennes; Soclete Francaise de Biometrie, NESA, 66-82.

[117] Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of survival fraction: an application to the analysis of " permanent employment" in Japan. *Journal of the American Statistical Association*, **87**, 284-292.

[118] Yu, M.G., Law, N. and Taylor, J.M.C. (2004). Joint longitudinal-survival-cure models and their application to prostate cancer. *Statistica Sinica,* **114**, 853-862.

[119] Zheng, Z.K. (1987). The weak convergence of the Kaplan-Meier estimator on half line under the independent but non-identical censoring. *Applied Mathematics a Journal of Chinese University,* **2**, 6474-351.

[120] Zheng, Z.K. (1984). *Regression analysis with censored data*. Ph.D. dissertation, Columbia University.

[121] Zhou, H. B. and Pepe, M. S. (1995). Auxiliary covariates data in failure time regression. *Biometrika*, **82**,139-149.

[122] Zhou,S. and Maller, R.A. (1995). The likelihood ratio test for the presence of immunes in a censored samples. *Statistics* , **27**, 181-201.

Appendix A

Let $F_0(t,\psi) = 1 - \exp(-\psi t)$. Then

$$\alpha_h(t,\theta) = \frac{p\psi \exp(-\psi t)}{1 - pF_0(t,\psi)} \exp(Z_h\beta).$$

In order to simplify the notations we will omit h of $\alpha_h(t,\theta)$ and of Z_h, these derivatives of $\alpha(t,\theta)$ and $\log \alpha(t,\theta)$ with respect to θ are listed below:

$$\frac{\partial^3}{\partial p^3}\alpha(t,\theta) = \frac{6[-1+\exp(-\psi t)]^2 \psi \exp[-\psi t + z\beta]}{[1 - pF_0(t,\psi)]^4}, \tag{A.1}$$

$$\frac{\partial^3}{\partial \psi^3}\alpha(t,\theta) = (-1+p)\{(p^2\psi t + 3p^2)\exp(-2\psi t) + (4p^2\psi t + 4p\psi t)\exp(-\psi t)$$

$$+(p^2 + -2p + 1)\psi t - (3p^2 - 6p + 3)\}\frac{\psi t^2 \exp[-\psi t + z\beta]}{[1 - pF_0(t,\psi)]^4}, \tag{A.2}$$

$$\frac{\partial^3}{\partial \beta^3}\alpha(t,\theta) = \frac{pz^3\psi \exp[-\psi t + z\beta]}{1 - pF_0(t,\psi)}, \tag{A.3}$$

$$\frac{\partial^3}{\partial p^2 \partial \psi}\alpha(t,\theta) = -2\{(1 - 2p - 2\psi t)\exp(-\psi t) + (p + p\psi t)\exp(-\psi t)$$

$$+(\psi t - p\psi t + p - 1)\}\frac{\exp[-\psi t + z\beta]}{[1 - pF_0(t,\psi)]^4}, \tag{A.4}$$

$$\frac{\partial^3}{\partial p^2 \partial \beta}\alpha(t,\theta) = \frac{-2z(-1+\exp(-\psi t))\psi \exp[-\psi t + z\beta]}{[1 - pF_0(t,\psi)]^3}, \tag{A.5}$$

$$\frac{\partial^3}{\partial p \partial \psi^2}\alpha(t,\theta) = \{(2p^2 + p\psi t)\exp(-2\psi t) + (4p^2\psi t - 4p\psi t)\exp(-\psi t)$$

$$+(1 - 2p + p^2)\psi t + (4p - 2)\}\frac{t\exp[-\psi t + z\beta]}{[1 - pF_0(t,\psi)]^3}, \tag{A.6}$$

$$\frac{\partial^3}{\partial p \partial \beta^2}\alpha(t,\theta) = \frac{z^2\psi \exp[-\psi t + z\beta]}{[1 - pF_0(t,\psi)]^2}, \tag{A.7}$$

$$\frac{\partial^3}{\partial \psi^2 \partial \beta}\alpha(t,\theta) = \frac{zp(p-1)t\{(p\psi t + 2p)\exp(-\psi t)}{1 - pF_0(t,\psi)}\}$$

$$+\frac{zp(p-1)t(p-1)\psi t + (2 - 2p)}{1 - pF_0(t,\psi)}, \tag{A.8}$$

$$\frac{\partial^3}{\partial\psi\partial\beta^2}\alpha(t,\theta) = \frac{pz^2[1-p+p\exp(-\psi t)+(p-1)\psi t]\exp[-\psi t+z\beta]}{[1-pF_0(t,\psi)]^2}, \qquad (A.9)$$

$$\frac{\partial^3}{\partial p^3}\log\alpha(t,\theta) = \frac{2}{p^3} - \frac{2(-1+\exp(-\psi t))^3}{[1-pF_0(t,\psi)]^3}, \qquad (A.10)$$

$$\frac{\partial^3}{\partial\psi^3}\log\alpha(t,\theta) = \frac{pt^3\exp(-\psi t)}{1-pF_0(t,\psi)} - \frac{3pt^3\exp(-2\psi t)}{[1-pF_0(t,\psi)]^2}$$

$$+\frac{2p^3t^3\exp(-3\psi t)}{[1-pF_0(t,\psi)]^3} + \frac{2}{\psi^3}, \qquad (A.11)$$

$$\frac{\partial^3}{\partial p^2\partial\psi}\log\alpha(t,\theta) = \frac{(2-2\exp(-\psi t))t\exp(-\psi t)}{[1-pF_0(t,\psi)]^3}, \qquad (A.12)$$

$$\frac{\partial^3}{\partial p\partial\psi^2}\log\alpha(t,\theta) = \frac{t^2\exp(-\psi t)(p\exp(-\psi t)-1+p)}{[1-pF_0(t,\psi)]^3}, \qquad (A.13)$$

$$\frac{\partial^2}{\partial p\partial\psi}\log\alpha(t,\theta) = \frac{t\exp(-\psi t)}{[1-pF_0(t,\psi)]^2}, \qquad (A.14)$$

$$\frac{\partial^2}{\partial p^2}\log\alpha(t,\theta) = \frac{-1}{p^2} + \frac{(-1+\exp(-\psi t))^2}{[1-pF_0(t,\psi)]^2}, \qquad (A.15)$$

$$\frac{\partial^2}{\partial\psi^2}\log\alpha(t,\theta) = \frac{-1}{\psi^2} - \frac{pt^2\exp(-\psi t)}{1-pF_0(t,\psi)} + \frac{p^2t^2\exp(-2\psi t)}{[1-pF_0(t,\psi)]^2}, \qquad (A.16)$$

$$\frac{\partial}{\partial p}\log\alpha(t,\theta) = \frac{1}{p} + \frac{1-\exp(-\psi t)}{1-pF_0(t,\psi)}, \qquad (A.17)$$

$$\frac{\partial}{\partial\psi}\log\alpha(t,\theta) = \frac{1}{\psi} - t + \frac{pt\exp(-\psi t)}{1-pF_0(t,\psi)}, \qquad (A.18)$$

$$\frac{\partial}{\partial\beta}\log\alpha(t,\theta) = z, \qquad (A.19)$$

and the other derivatives we needed are zero.

Appendix B

It is well known that the survival proportion reflected in the right tail of a standard marginal survival curve at the end of the follow-up period may not adequately estimate the curve fraction. Farewell (1986) showed that it is hard to distinguish a censored individual in the susceptible group from a long-term survivor. This leads to difficulty in distinguishing models with a large proportion of susceptibles and long tails of the latency distribution from those with a small proportion of susceptibles and short tails of the latency distribution. As a result, the identifiability of a cure model becomes very important in order to obtain unique estimates of the model parameters.

In this appendix we will investigate the identifiability of the related cure model. To start we give some notations and definitions.

B.1. Identifiability for SCR and BCH Models

Let T^+ be an arbitrary large time (possibly infinity) beyond which we have no interest. In the general case, the proportion for the susceptible p depends on covariates x through a function $p(x)$, and the "proper" distributions $F_0(t)$ in the SCR model or $G_0(t)$ in the BCH model are independent of any covariate x. Without loss of generality, assume that x is one-dimensional.

Let $\mathcal{F}_0 = \{F_0(t, \psi) : \psi \in \Psi\}$ be the class of proper failure time distributions and $\mathcal{G}_0 = \{G_0(t) : G_0(t) \text{is a proper distribution on } [0, \infty)\}$. Let $\mathcal{X}$ be the design space, for convenience, we assume $\mathcal{X}$ to be the closed interval $[a_0, a_1]$. Denote the space of incidence probability function by $\mathcal{P} = \{p(x) : p(x) \neq 0, 1 \text{ for all } x \in \mathcal{X}\}$ and $\zeta = \{\zeta(x) : \zeta(x) \neq 0, \infty, \text{ for all}\, x \in \mathcal{X}\}$.

Denote the class of SCR models and the class of BCH models by, respectively,

$$\mathcal{H}_1 = \{S(t, p(x), \psi) = 1 - p(x)F_0(t, \psi), t < T^+, p(\cdot) \in \mathcal{P}, F_0(\cdot, \psi) \in \mathcal{F}_0\} \qquad (B.1.1)$$

and

$$\mathcal{H}_2 = \{S(t, \zeta(x)) = \exp(-\zeta(x)G_0(t), t < T^+, x \in \mathcal{X}, \zeta(x) \in \zeta, G_0(t) \in \mathcal{G}_0\}. \quad (B.1.2)$$

Definition B.1.1. The class $\mathcal{H}_1$ of SCR models is identifiable if for any two members $S(t, p(x), \psi) = 1 - p(x)F_0(t, \psi)$ and $S^{\triangle}(t, p(x), \psi) = 1 - p^{\triangle}(x)F_0(t, \psi^{\triangle})$ of $\mathcal{H}_1$, the following two conditions hold:

(i) $S(t, p(x), \psi) \equiv S^{\triangle}(t, p(x), \psi)$ if and only if $p(x) \equiv p^{\triangle}(x)$ for $x \in \mathcal{X}$; and

(ii) $F_0(t, \psi) \equiv F_0(t, \psi^{\triangle})$ for almost all $t \in (0, T^+)$.

Definition B.1.2. The class $\mathcal{H}_2$ of BCH models is identifiable if for any two members $S(t, \zeta(x)) = \exp\{-\zeta(x)G_0(t)\}$ and $S^{\triangle}(t, \zeta(x)) = \exp\{-\zeta^{\triangle}(x)G_0^{\triangle}(t)\}$ of $\mathcal{H}_2$, the following two conditions hold:

(i) $S(t, \zeta(x)) \equiv S^{\triangle}(t, \zeta(x))$ if and only if $\zeta(x) \equiv \zeta^{\triangle}(x)$ for $x \in \mathcal{X}$; and

(ii) $G_0(t) \equiv G_0^{\triangle}(t)$ for almost all $t \in (0, T^+)$.

Thus, similar to the proofs of Theorem 1-5 in Li *et al.* (2001), we can show the following Theorems B.1.1-B.1.5. We only sketch them here, the details can be found in their literature.

Theorem B.1.1. *Let x be a continuous covariate in the design space $\mathcal{X} = [a_0, a_1]$, where $-\infty < a_0 < a_1 < \infty$. Then the model given by*

$$S(t, p(x)) = 1 - p(x)F_0(t), \quad t < T^+, \qquad (B.1.3)$$

is not identifiable in each of the following three cases:

(i) $p(x)$ is unspecified, unless $F_0(T^+) = 1$ and $\max_{x \in \mathcal{X} p(x)} = 1$;

(ii) $p(x)$ is a constant parameter;

(iii) $p(x)$ *is specified as a logistic function*

$$p(x) = \frac{\exp(a + bx)}{1 + \exp(a + ba)}$$

with $b \neq 0$.

Theorem B.1.2. *Let* x *be a continuous covariate in the design space* $\mathcal{X} = [a_0, a_1]$, *where* $-\infty < a_0 < a_1 < \infty$. *Then the model*

$$S(t, p(x)) = 1 - p(x) + p(x)[S_0(t)]^{r(x)}, \quad t < T^+, \tag{B.1.4}$$

is identifiable if $r(x) \in R$, *where* $r(x) > 0$ *is the related hazards function of the conditional survival function and* $S_0(t) = 1 - F_0(t)$.

Theorem B.1.3. *The model*

$$S(t, p(x), \psi) = 1 - p(x)F_0(t, \psi)], \quad t < T^+, \tag{B.1.5}$$

is identifiable regardless of whether $p(x)$ *is parametric or not.*

Theorem B.1.4. *The model*

$$S(t, \zeta(x)) = \exp\{-\zeta(x)G_0(t)\}, \quad x \in \mathcal{X},\ t < T^+, \tag{B.1.6}$$

is not identifiable if

(i) $G_0(t)$ *and* $\zeta(x)$ *are unspecified; or*

(ii) $G_0(t)$ *is unspecified and* $\zeta(x) = \exp(a + bx)$ *with* $a \neq 0$.

Theorem B.1.5. *The model*

$$S(t, \zeta(x)) = \exp\{-\zeta(x)G_0(t, \psi)\}, \quad x \in \mathcal{X},\ t < T^+, \tag{B.1.7}$$

is identifiable if $\zeta(x)$ *is either unspecified or equal to* $\exp(a + bx)$.

B.2. Identifiability for Improper PH Model

For our "improper" PH model, we should re-establish its identifiability for its special structure which is different from that of SCR and BCH models.

Now we assume p to be a constant parameter in the sense that p is independent of any covariate x. Define

$$\mathcal{H}_3 = \{S(t,p,\psi) = [1 - pF_0(t,\psi)]^{r(x)}, t < T^+, x \in \mathcal{X}, p \in (0,1], F_0(t,\psi) \in \mathcal{F}_0\}. \tag{B.2.1}$$

Then following the notations in Section B.1, we can define the identifiability of the "improper" PH model.

Definition B.2.1. The class $\mathcal{H}_3$ of "improper" models is identifiable if for any two members, say $S(t,p) = [1 - pF_0(t)]^{r(x)}$ and $S^{\triangle}(t,p^{\triangle}) = [1 - p^{\triangle}F_0^{\triangle}(t)]^{r^{\triangle}(x)}$, the following two conditions hold:

(i) $S(t,p) \equiv S^{\triangle}(t,p^{\triangle})$ if and only if $r(x) \equiv r^{\triangle}(x)$ for $x \in \mathcal{X}$; and

(ii) $F_0(t) \equiv F_0^{\triangle}(t)$ for almost all $t \in (0,T^+)$,

(ii) $p = p^{\triangle}$.

Definition B.2.2. The class $\mathcal{H}_3$ of "improper" PH models is identifiable if for any two members of $\mathcal{H}_3$, say $S(t,p,\psi) = [1 - pF_0(t,\psi)]^{r(x)}$ and $S^{\triangle}(t,p^{\triangle},\psi^{\triangle}) = [1 - p^{\triangle}F_0(t,\psi^{\triangle})]^{r^{\triangle}(x)}$, $S(t,p,\psi) \equiv S^{\triangle}(t,p^{\triangle},\psi^{\triangle})$ holds if and only if $p \equiv p^{\triangle}$, $r(x) \equiv r^{\triangle}(x)$ and $F_0(t,\psi) \equiv F_0(t,\psi^{\triangle})$ for $x \in \mathcal{X}$ and almost all $t \in (0,T^+)$.

Next we establish our Theorems for the identifiability of the "improper" PH model.

Theorem B.2.1. *Let x be a continuous covariate in the design space $\mathcal{X} = [a_0, a_1]$,* *where $-\infty < a_0 < a_1 < \infty$. Then the model*

$$S(t,p) = [1 - pF_0(t)]^{r(x)}, \quad t < T^+ \tag{B.2.2}$$

is not identifiable.

Theorem B.2.2. *Let x be a continuous covariate in the design space $\mathcal{X} = [a_0, a_1]$, where $-\infty < a_0 < a_1 < \infty$. Assume that exists a interval $d = [d_1, d_2] \in (0, \infty)$ such that $F(t, \psi)$ is continuous for t on d, where $d_1 < d_2$. Then the model*

$$S(t,p,\psi) = [1 - pF_0(t,\psi)]^{r(x)}, \quad t < T^+ \tag{B.2.3}$$

is identifiable.

Proof of Theorem B.2.1. We need to contradict that $S(t,p) \equiv S^{\triangle}(t,p^{\triangle})$ if and only if $p \equiv p^{\triangle}$, $r(x) \equiv r^{\triangle}(x)$ and $F_0(t) \equiv F_0^{\triangle}(t)$ for almost all $t \in (0, T^+)$.

For any given $S(t,p) = [1 - pF_0(t)]^{r(x)}$, take $c \neq 1$ and let

$$r^{\triangle}(x) \equiv \frac{r(x)}{c}, \quad p^{\triangle} = 1 - (1-p)^c$$

and

$$F_0^{\triangle}(t) = \frac{1 - [1 - pF_0(t)]^c}{p^{\triangle}} = \frac{1 - [1 - pF_0(t)]^c}{1 - (1-p)^c}.$$

Then

$$\begin{aligned} S^{\triangle}(t,p^{\triangle}) &= [1 - p^{\triangle}F_0^{\triangle}(t)]^{r^{\triangle}(x)} = [1 - (1 - [1 - pF_0(t)]^c)]^{r(x)/c} \\ &= \{[1 - pF_0(t)]^c\}^{r(x)/c} = [1 - pF_0(t)]^{r(x)} = S(t,p) \end{aligned}$$

for all $t \in (0, T^+)$, whereas $r(x) \not\equiv r^{\triangle}(x)$ (as $c \neq 1$). This shows that the model in (B.2.2) is not identifiable. ∎

Proof of Theorem B.2.2. We need to show that $S(t,p,\psi) \equiv S^{\triangle}(t,p^{\triangle},\psi^{\triangle})$ if and only if $p \equiv p^{\triangle}$, $r(x) \equiv r^{\triangle}(x)$ and $F_0(t,\psi) \equiv F_0(t,\psi^{\triangle})$ for almost all $t \in (0, T^+)$.

The "if" part is clearly true, so we concentrate on the "only if" part. Suppose that $S(t,p,\psi) \equiv S^{\triangle}(t,p^{\triangle},\psi^{\triangle})$. Then

$$\frac{r^{\triangle}(x)}{r(x)} \equiv \frac{\log(1-pF_0(t,\psi))}{\log(1-p^{\triangle}F_0(t,\psi^{\triangle}))}. \tag{B.2.4}$$

Since the left-hand side depends only on x and the right-hand side depends only on t, the ratios in (B.2.4) must be a positive constant, say $1/c$, which does not depend on x and t. Consequently,

$$\frac{r^{\triangle}(x)}{r(x)} \equiv \frac{1}{c} \quad \text{and} \quad 1-p^{\triangle}F_0(t,\psi^{\triangle}) \equiv [1-pF_0(t,\psi)]^c. \tag{B.2.5}$$

The second equality in (B.2.5) holds only if $c=1$ since $F(t,\psi)$ is continuous for t on d. Thus (B.2.5) gives

$$r(x) \equiv r^{\triangle}(x) \quad \text{and} \quad \frac{p^{\triangle}}{p} \equiv \frac{F_0(t,\psi)}{F_0(t,\psi^{\triangle})}. \tag{B.2.6}$$

It follows that

$$F_0(t,\psi^{\triangle}) \equiv c_1 F_0(t,\psi) \quad \text{and} \quad p^{\triangle} = p/c_1 \tag{B.2.7}$$

for some positive constant c_1. Let $t \to T^+$ in the first equality of (B.2.7), we get $c_1 = 1$. Therefore, by (B.2.7), $F_0(t,\psi^{\triangle}) \equiv F_0(t,\psi)$ and $p^{\triangle} = p$, which together with the first equality of (B.2.6) proves the "only if" part, and so completes the proof. ∎

Before close this section, we should mention that, as a special case, our "improper" PH model reduces to the ordinary Cox's PH model when $p=1$. Similar to the proofs of Theorems B.2.1-B.2.2 we can show that model $S(t) = [1-F_0(t)]^{r(x)}$ is not identifiable while model $S(t) = [1-F_0(t,\psi)]^{r(x)}$ is identifiable. The details will be omitted here.