THE UNIVERSITY OF CALGARY

Polynomial Spline Estimation of Partially Linear Single-Index Proportional Hazards Regression Models

by

Jie Sun

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

July, 2007

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a Dissertation entitled "Polynomial Spline Estimation of Partially Linear Single-Index Proportional Hazards Regression Models" submitted by Jie (Rena) Sun in partial fulfillment of the requirements for the degree of Master of Science.

Supervisor, Dr. Xuewen Lu
Department of Mathematics and Statistics

Dr. Karen A. Kopciuk
Department of Mathematics and Statistics;
Alberta Cancer Board

Dr. JianLi Wang
Departments of Psychiatry and Community Health Sciences

July 5, 2007

Date

Abstract

The Cox proportional hazards model usually assumes linearity of covariates on the log hazard function, which may be violated because linearity can not always be guaranteed. To deal with this issue, several previous researchers proposed a number of unstructured or structured nonparametric models. In this thesis, I will consider an alternative way to model covariate effects which could be linear or nonlinear on the log hazard in the proportional hazards model with a set of covariates. I propose a partially linear single-index proportional hazards regression model and apply a polynomial spline smoothing method to model the structured nonparametric single-index component. This method can reduce the dimensionality of the covariates and more efficient estimates of the covariates' effects can be made.

A two-step iterative algorithm to estimate the nonparametric component and the covariate effects is used which does not involve estimating the baseline hazard function. We employed the command 'coxph' from the *R* package 'survival' into our Newton Raphson iteration to get the same results in a much shorter period of time. The nonparametric component is estimated by B-splines, which is a nonparametric smoothing method. Asymptotic properties of the estimators are derived. Monte Carlo simulation studies are presented to compare the new method with the Cox linear proportional hazards model and some other comparable models. Application to the Veteran's Administration Lung Cancer survival data demonstrates the usefulness of our proposed method for gaining insight into the nonlinearity of covariates.

Acknowledgements

This thesis would not have been possible without the guidance, encouragement, and patience of my supervisor, Professor Xuewen Lu. I thank him for sharing his insight and knowledge with me. He is not only an excellent statistician with deep vision but also, and most importantly, a great person. I feel privileged to have had the opportunity to learn from him.

I am also deeply grateful to my committee members, Professors Karen A. Kopciuk and JianLi Wang, for their suggestions and comments.

In addition, I would like to thank the Department of Mathematics and Statistics at the University of Calgary for their assistance and financial support throughtout my graduate studies.

My thanks also go to my friend, Philip Fust, who has been always supportive and encouraging.

As always, it is impossible to mention everybody who had an impact on this work. I would like to thank all of my colleagues who had supported me throughout the course of the thesis, particularly, Cynthia Zheng and Matt Davis, for sharing their valuable statistical insight with me.

Finally my special thanks go to my wonderful parents, Baosheng Sun and Huifen Su, whose faith in me was the motivation to carry on with this work. I thank them for their encouragement, patience and understanding.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to Survival Analysis

In recent years the statistical analysis of survival times has become a fruitful area both of practical application and of methodological research. The techniques of survival analysis are widely used in medical work, particularly in clinical trials and in follow-up studies, in social and economic sciences, including establishing social security benefits, job terminations and promotions, event-history analysis, duration analysis and stock market crashes, and in engineering for reliability and failure time analysis purposes. It is assumed that the outcome or response for each study subject or experimental unit may be viewed as a point event in time. Usually these events will not occur more than once to the same study subject. Examples include deaths of patients (in a clinical trial), the first recurrence of a disease (in a follow-up study of subjects treated for that disease), and the failure of a mechanical component under stress. For a uniform terminology we use the 'term' failure throughout, although in some applications, for example, in a study of the duration of period of unemployment, the outcome (securing a job) may be a desirable occurence.

In this section, we will give some background on survival analysis, and also introduce a popular classic model – the Cox proportional hazards model.

1.1.1 Background of the Statistical Analysis of Survival Times

Often we are interested in the distribution of the times to failure, so-called survival times. Although there are many well-known methods for estimating unconditional survival distributions, examining the relationship between survival and one or more predictors, which are usually termed covariates or explanatory variables in the survival analysis literature, is often of greater interests. The covariates may be dummy variables, representing contrasts between treatment and control groups, or measurements made on each individual upon entry into the study. Sometimes a covariate may itself take different values over time for the same study subject, for example blood pressure measured at repeated intervals during follow-up studies. In sum, covariates can be measured as binary, discrete or continuous, and fixed or varying.

A distinctive feature of the survival times is that some may be censored, that is, some study subjects may not be observed for the full time to failure. The most common form is right-censoring: the period of observation ends, or an individual is removed from the study, before the event of interest occurs. For example, some individuals may still be alive at the end of a clinical trial, or may drop out of the study for various reasons other than death prior to its termination. The second form is left-censoring where the initial event time of a subject at risk is unknown. Lastly, we have the third form, which occurs when an observation is both right and left-censored; hence, it is termed interval-censoring. Censoring complicates the likelihood function, and hence the estimation, of survival models.

Moreover, since censoring is conditional on the value of any covariates in a survival model and on an individual's survival to a particular time, it is necessary to

evaluate whether censoring is independent of the future value of the hazard for the individual. If this condition is not met, then estimates of the survival distribution can be seriously biased. For example, if individuals tend to drop out of a clinical trial shortly before they die and, therefore, their deaths go unobserved, survival time will be over-estimated. Censoring that meets this requirement is termed non-informative. A common type of non-informative censoring occurs when a study terminates at a predetermined date.

We may loosely distinguish three approaches to the statistical analysis of survival times: parametric, nonparametric and semi-parametric methods.

The first approach is parametric. It is defined by its underlying distributional forms, which are specified parametrically, and include the exponential, Weibull, log normal, log logistic, Pareto, gamma, normal, exponential power, Gompertz or inverse Gaussian distributions. Maximum likelihood methods can be used to fit these models which relate the parameters of these distributions to the covariates. These parametric models can offer insight into the nature of the various parameters and functions discussed above, particularly, the hazard rate. For a detailed description of parametric models, see Lawless (2003).

The second approach is nonparametric, which does not assume any functional form for the risk function and the distribution function. It stems from the product limit estimator of the survival distribution introduced by Kaplan & Meier (1958). Both the theories and the scope of these techniques are developing rapidly, for example, see Nielsen and Linton (1995).

A third battery of techniques, sometimes called semi-parametric method, can be used to synthesize the parametric and nonparametric methods. The effect of

primary interest, namely that of the explanatory variables on survival, is represented parametrically or semiparametrically (with parametric components or nonparametric components), but no specific form is assumed for the distribution of survival times. A seminal paper by Cox (1972) on survival regression models and life-tables has motivated much subsequent work in this area. It is also the main subject of this thesis.

1.1.2 Cox Proportional Hazards Model

This section describes the proportional-hazards regression model, which has played a pivotal role in survival analysis since Cox proposed it in 1972. This model assumes the hazard rate or intensity function is a product of an unspecified function of time common to all individuals and a known link function (usually exponential) of a linear combination of the covariates.

Notation and Definitions

For each of the n subjects, indexed by i, assume there is a given fixed covariate or explanatory function X_i. For each individual i, this function may be independent of time, or a deterministic function of time. For notational simplicity, we take X_i, to be scalar for each individual, i, although in practice it will often be a vector.

Define T as the survival time. Recall that the survival function $S(t)$, which is defined as $S(t) = P(T > t) = 1 - F(t)$, and the density $f(t)$ of a positive random variable with absolutely continuous distribution, can be expressed in terms of the hazard function

$$\lambda(t) = lim_{\Delta h \to 0+} \frac{1}{\Delta h} P(T \leq t + \Delta h | T > t) = -\frac{\partial \log S(t)}{\partial t},$$

by the equations

$$S(t) = \exp\{-\int_0^t \lambda(u)du\}, \quad f(t) = \lambda(t)\exp\{-\int_0^t \lambda(u)du\},$$

respectively.

In the absence of censoring, the conditional hazard function, $\lambda_i(t)$, which depends on covariate $X_i(t)$ for the survival time T_i of the i^{th} subject, is assumed to satisfy

$$\lambda_i(t|X(t)) = \lambda_0(t)\exp\{\beta X_i(t)\}.$$

Here $\lambda_0(t)$ is the unknown baseline hazard function corresponding to $X = 0$, so it is the same for all subjects and the unknown coefficient β expresses the dependence of the distribution of T_i on the covariate $X_i(t)$. An important assumption is that failures of different subjects occur independently, and that the value of the covariate function for one subject does not influence the survival time of any other subject.

To include the possibility of censoring in the model, we suppose that for each studied subject there is a random censoring time C_i, which is generally unknown. We observe $Z_i = \min(T_i, C_i)$ together with an indicator of failure ($T_i \leq C_i$) or censoring ($T_i > C_i$). A censoring indicator is defined as $\delta_i = I\{T_i \leq C_i\}$ for the i^{th} subject.

Partial Likelihood

Now we will discuss about the likelihood function of this standard model first.

Let D be the set of subjects observed to die and $|D|$ is the size of set D. Thus, $D = \{i : T_i \leq C_i\}$. Let $R(u)$, the risk set at time u, denote the set of subjects observed to survive until time u. Thus, $R(u) = \{i : T_i \geq u, C_i \geq u\}$, let $|R(u)|$ be

the size of this set at time u. We write $x_{ji} = x_j(t_i)$ for the j^{th} covariate value of the i^{th} subject at their failure time, and R_i be the corresponding risk set $R(t_i)$.

The ordered observed failure times, t_i, will be denoted by $\tau_1 < \cdots < \tau_m$ with $\tau_0 = 0$. For notational simplicity we consider only the case where $|R(\tau_m+)| = 0$, that is where no subjects survive beyond the last observed failure.

If $\lambda_0(t)$ were known, a full likelihood could be derived. A subject censored at $C_i = c_i$ would contribute a term

$$\exp\left\{-\int_0^{c_i} \lambda_0(u) e^{\beta x_i(u)} du\right\},$$

which is the probability of survival past c_i, and a term for a subject who failed at t_i

$$\lambda_0(t_i) e^{\beta x_i(t_i)} \exp\left\{-\int_0^{t_i} \lambda_0(u) e^{\beta x_i(u)} du\right\},$$

which is the probability of failure at t_i.

The product of all such terms may be written as

$$Lik = \prod_{i=1}^{m} \left\{ \left[\exp\{-\int_{\tau_i - 1}^{\tau_i} \lambda_0(u) \sum_{j \in R(u)} e^{\beta x_j(u)} du\} \lambda_0(\tau_i) \sum_{j \in R(\tau_i)} e^{\beta x_j(\tau_i)} \right] \frac{\exp\{\beta x_{J_i}(\tau_i)\}}{\sum_{j \in R(\tau_i)} e^{\beta x_j(\tau_i)}} \right\},$$

where J_i is the index of the subject who fails at τ_i.

If $\lambda_0(t)$ is unknown, the terms in square brackets will provide little information about β, and Cox's (1972) suggestion amounts to basing inferences about β on the remaining terms which together constitute the partial likelihood,

$$PartialLik = \prod_{i \in D} \frac{e^{\beta x_{ii}}}{\sum_{j \in R_i} e^{\beta x_{ji}}}.$$

The corresponding log likelihood is

$$l = \log(PartialLik) = \beta \sum_{i \in D} x_{ii} - \sum_{i \in D} \log\{\sum_{j \in R_i} e^{\beta x_{ji}}\}.$$

Cox (1975) indicates that under suitable regularity conditions, these results imply that *PartialLik* enjoys all the usual asymptotic properties of a likelihood function. Rigorous proofs appear however to have been derived only under rather restrictive conditions (Tsiatis, 1981; Liu and Crowley, 1978).

Discussion

An important assumption of the Cox model is that the covariate variables have a linear effect on the log-hazard function. However, this assumption could be violated and a misleading conclusion could be drawn. As a remedy, nonparametric function estimation has been proposed to estimate the conditional hazard function. Much research has been devoted to investigating this area, including O'Sullivan (1993), Gentleman and Crowley (1991), Tibshirani and Hastie (1987), Fan, Gijbels and King (1997), and Gu (1996), among others. These authors based their research on the following common form for the hazard function:

$$\lambda(t|x) = \lambda_0(t) \cdot \exp\{\varphi(x)\},$$

where $\varphi(x)$ is assumed to be a smooth function of x and is unknown.

But unstructured nonparametric function estimation is subject to the 'curse of dimensionality' and, thus, is not practically useful when the covariate vector x has many components. The curse of dimensionality is a term coined by Bellman (1961) to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to a mathematical space. For example, 100 evenly-spaced sample points suffice to sample a unit interval with no more than 0.01 distance between points; an equivalent sampling of a 10-dimensional unit hypercube with a lattice in spacing of 0.01 between adjacent points would require 10^{20} sample points.

Thus, in some sense, the 10-dimensional hypercube can be said to be a factor of 10^{18} 'larger' than the unit interval.

Avoiding the 'curse of dimensionality' is an issue that concerns many statisticians. Much research has been done using structured nonparametric models. Hastie and Tibshirani (1990) proposed the 'Generalized Additive Model' (GAM), which features an additive term of some unspecified smooth functions of the covariates in place of the usual linear predictor form of the covariates. They estimated the unspecified smooth function using scatterplot smoothers. Sleeper and Harrington (1990) also used additive models to model the nonlinear covariate effects in the Cox PH model. However, they modeled the log-hazard as an additive function of each covariate and then approximated each of the additive components using a polynomial spline. Similar research has been done by Huang (1999). Gray (1992) applied penalized splines to additive models and time-varying coefficient models of the log-hazard function. Using functional ANOVA decompositions, Huang *et al.* (2000) studied a general class of structured models for proportional hazards regression that includes additive models as a specific case; polynomial splines are used as the building blocks for fitting the functional ANOVA models. Other approaches have used a Single Index Model (SIM) to reduce dimension, in order to avoid the 'curse of dimensionality'. We will expand on this approach in the next section.

1.2 Introduction to Single Index Model

The popularity of modeling the covariates using a linear form in an empirical analysis is based on the ease with which the results can be easily interpreted. This tradition

influenced the modeling of various nonlinear regression relationships, where the mean response variable is assumed to be a nonlinear function of a weighted sum of the predictor variables (Härdle and Stoker, 1989; Powell, Stock and Stoker, 1989). This is the so called Single Index Model (SIM). As in linear modeling, this feature is attractive because the coefficients, or their weighted sum, gives a simple picture of the relative impacts of the individual predictor variables on the response variable.

SIM summarizes the effects of the explanatory variables $X_1, \cdots, X_p$ on the conditional mean of Y within a single variable called the index through the mean regression,

$$E(Y|X) = m(X) = g(\beta^T X).$$

It generalizes linear regression by replacing the linear combination $\beta^T X$ with a nonparametric component, $g(\beta^T X)$, where g is an unknown univariate link function.

Because a nonlinear link function is applied to the coefficient vector β, interactions between the covariates can be modeled. Thus, SIM is a useful alternative to the additive model, which also reduces dimensionality but can not incorporate interactions. Applications of SIM lie in a variety of fields, such as discrete choice analysis in econometrics and dose-response models in biometrics, where high-dimensional regression models are often employed (Härdle, Hall and Ichimura, 1993; Powell, Stock and Stoker, 1989; Carroll *et al.*, 1997; Xia *et al.*, 2002). For more examples which motivated the single index model, see Ichimura (1993).

Over the last two decades, many authors had devised various clever estimators of the single-index coefficient vector $\beta = (\beta_1, \cdots, \beta_p)^T$. Usually estimation for SIM is carried out in two steps. First, the coefficients vector β was estimated, and then second, the unknown link function g was estimated by ordinary univariate nonpara-

metric regression of Y on $\beta^T X$ using the index values for the observations.

Discussed above is the single-index model applied to mean regression for uncensored response variables. Some research has been conducted to apply Single Index Model (SIM) with hazard regression models (Nielsen *et al.*, 1998; G⊘rgens, 2004; Wang, 2004; Lu *et al.*, 2006). Recently, Huang and Liu (2006) proposed a model with the conditional hazard function specified as

$$\lambda(t|x) = \lambda_0(t) \cdot \exp\{\varphi(\beta^T x)\},$$

where $\varphi(\cdot)$, is the link function, which is an unknown smooth function. Since $\varphi(\cdot)$ is not specified, the relative risk function has a flexible form. It can model possible departures from the standard Cox PH model that can not be captured by an additive, time-varying coefficient, or a more general functional ANOVA model.

Huang and Liu (2006) used spline smoothing for the unknown link function. The greatest advantages of spline smoothing are its simplicity and fast computation, at least when equally spaced knots are used. Hence, the spline estimator of $\varphi(\cdot)$ possesses not only the usual strong consistency and $\sqrt{n}$-rate convergency to an asymptotic normal distribution, but is also fast to compute for large sample size, n, and high dimension covariate vector.

However, in their model, they created an inference procedure by incorporating all of the covariates into one single-index term, which may be arbitrary, since some of them could have linear effects on the response variable while others could have nonlinear effects. In this thesis, I separate the covariate vector into two groups, so that the model has one parametric component and one nonparametric component. The parametric component keeps the linear form of some covariates and simplicity

of the original Cox model, while the nonparametric component adds some flexibility to the model. We adopt the uniform B-spline smoothing on the nonparametric component. More details on our model will be given in Chapter 2.

1.3 Introduction to B-splines

In this section, we provide some basic concepts and several important theorems about B-splines (Schumaker, 1981).

Definition 1.3.1. Let $\cdots \leq y_{-1} \leq y_0 \leq y_1 \leq y_2 \leq \cdots$ be a sequence of real numbers. Given integers i and $m > 0$, we define

$$Q_i^m(x) = \begin{cases} (-1)^m [y_i, \cdots, y_{i+m}](x-y)_+^{m-1} & if\ y_i < y_{i+m} \\ 0 & otherwise, \end{cases}$$

for all real x.

We call Q_i^m the m^{th} order B-spline associated with the knots $y_i, \cdots, y_{i+m}$, where

$$[y_i, \cdots, y_{i+m}]f(x) = \frac{D\begin{pmatrix} y_i, \ldots, y_{i+m} \\ 1, x, \ldots, x^{m-1}, f \end{pmatrix}}{D\begin{pmatrix} y_i, \ldots, y_{i+m} \\ 1, x, \ldots, x^m \end{pmatrix}}.$$

Theorem 1.3.2. *(Recursion Formula)*

Let $m \geq 2$ and suppose $y_i < y_{i+m}$, then for all $x \in \mathbb{R}$,

$$Q_i^m(x) = \frac{(x-y_i)Q_i^{m-1}(x) + (y_{i+m}-x)Q_{i+1}^{m-1}(x)}{y_{i+m}-y_i}.$$

This provides a recursion relation whereby B-splines of order m can be related to B-splines of order $m-1$.

Definition 1.3.3. (Normalized B-spline)

Let $N_i^m(x) = (y_{i+m} - y_i)Q_i^m(x)$. We call N_i^m the normalized B-spline associated with the knots $y_i, \cdots, y_{i+m}$.

Theorem 1.3.4. *(Recursion Formula for Normalized B-spline)*

The B-splines form a partition of the unit; that is,

$$\sum_{i=j+1-m}^{j} N_i^m(x) = 1 \quad for\ all\ \ y_j \le x < y_{j+1},\ i = (j+1-m), \cdots, j.$$

Also, we have

$$N_i^m(x) = \frac{x - y_i}{y_{i+m-1} - y_i} N_i^{m-1}(x) + \frac{y_{i+m} - x}{y_{i+m} - y_{i+1}} N_{i+1}^{m-1}(x).$$

Theorem 1.3.5. *(Derivatives)*

Let $y_i < y_{i+m}$ and suppose D_+ is the right derivative operator. Then

$$D_+ Q_i^m(x) = (m-1)\frac{[Q_i^{m-1}(x) - Q_{i+1}^{m-1}(x)]}{(y_{i+m} - y_i)},$$

$$D_+ N_i^m(x) = (m-1)\left(\frac{N_i^{m-1}(x)}{y_{i+m-1} - y_i} - \frac{N_{i+1}^{m-1}(x)}{y_{i+m} - y_{i+1}}\right).$$

The above-mentioned definitions and theorems are the general forms for any distribution of the knots. The following are the special case formulas for equally spaced knots, which are usually called uniform B-splines.

Theorem 1.3.6. *(Recursion Formula for Uniform B-spline)*

Let $m \ge 2$, then for all $x \in \mathbb{R}$,

$$Q^m(x) = \frac{xQ^{m-1}(x) + (m-x)Q^{m-1}(x-1)}{m}.$$

$$Since\ \ N^m(x) = mQ^m(x),$$

$$\begin{aligned} N^m(x) &= xQ^{m-1}(x) + (m-x)Q^{m-1}(x-1) \\ &= \frac{x}{m-1}N^{m-1}(x) + \frac{m-x}{m-1}N^{m-1}(x-1). \end{aligned}$$

Theorem 1.3.7. *(Derivatives for Uniform B-spline)*

$$D_+N^m(x) = N^{m-1}(x) - N^{m-1}(x-1).$$

See the text by L. L. Schumaker (1981) 'Spline Functions: Basic Theory' for additional definitions and the theorem proofs.

Chapter 2

A Partially Linear Single-Index Proportional Hazards Regression Model

2.1 Motivation

Our motivation to consider a partially linear regression model with many covariates has become at least some of them can have nonlinear effects on the response variable. In this situation, none of the traditional linear models, kernel smoothing methods (Nielsen, Linton and Bickel, 1998; Wang, 2004) and single-index models with spline smoothing (Huang and Liu, 2006) are able to incorporate both linear and nonlinear covariate effects. In addition, when a large number of covariates have nonlinear effects, the multivariate kernel smoothing method suffers from the 'curse of dimentionality'.

We propose a model with log relative risk function or log-hazard function:

$$\alpha^T V + \psi(\beta^T X),$$

which is referred to as the link function and where $\psi(\cdot)$ is an unknown function. We will use spline smoothing only for the nonlinear part including X but not the linear part including V. The covariate effects in this model are addressed in a semiparametric fashion, which offers improved flexibility over the existing methods in modeling the relationship between the failure time and the covariates. Unlike the models considered by Nielsen *et al.* (1998) and Lu *et al.* (2006), our model does

not assume a parametric baseline hazard function. When $p = 1$ (p is the number of covariates in X), β is a scalar, the model becomes the partially linear model studied by Heller (2001); when $p = 0$, the nonparametric component disappears, and the model is the classical Cox PH model; and when $q = 0$ (q is the number of covariates in V), the linear parametric component disappears, and the model is just a single-index model with an unknown link investigated by Huang and Liu (2006). Hence, the properties of this proposed model require a full investigation. The main focus of this paper is the estimation and inference of the parameters β_0 and α_0 under random censoring.

To apply this partially linear single-index model in practice, a practical issue arises: How to partition the covariates into the nonparametric component X and parametric component V? There are several strategies that can be applied to determine the division of the available covariates. The first approach is to utilize subject-matter knowledge related to the data collected in the experiment (or study) and the underlying physical mechanism. In this model, the X vector serves primarily in the role of dimension reduction, while the V vector may contain the key covariates of interest in the study. From this perspective, the selection of V and X can be readily made from the context of the study itself. For example, in a clinical study, the treatment effect of a medicine is of interest and could be coded as a categorical variable. It should be included in the V vector, whereas the other covariates such as patient's age and blood pressure can be included in the X vector. Although a categorical type V is defined here as a categorical variable, V can also be a continuous type variable or a mixture of the two types. In our case studies, we will rationalize this idea of selecting covariates. The second approach is to carry out some simple

analysis of all covariates, which can determine which covariates should be in the V or X component vector. For example, for each covariate, we perform a simple regression analysis based on a kernel smoothing or a spline smoothing method such as a univariate nonparametric regression or partially linear models. If the fitted curve appears to be linear or approximately linear, we then assign this covariate to V, otherwise, we assign it to X.

In the proportional hazards model framework with multi-dimensional covariates, this partially linear single-index model allows flexible modeling of the covariate effects and at the same time retains the feature of being parsimonious and easy to interpret like the Cox model and the single-index model.

2.2 Model Description

Suppose we have an i.i.d. sample with sample size n from a population with data vector of the form (V, X, T, C), where $T_1, \cdots, T_n$ represent survival times, $C_1, \cdots, C_n$ are the corresponding censoring times, and (V, X) the covariate vectors, as defined above. We assume the censoring is non-informative, in other words, T and C are independent given X and V. Suppose we observe for the i^{th} subject, an event time $Z_i = \min(T_i, C_i)$, a censoring indicator $\delta_i = I\{T_i \leq C_i\}$, as well as the q-variable covariate vector, V_i, and the p-variable auxiliary covariate vector, X_i. We denote the observed data for $i = 1, \cdots, n$ as $(V_i, X_i, Z_i, \delta_i)$.

We propose to approximate the derivative of the unknown function $\psi(\cdot)$ as a spline function. Such an approximation can be represented by a basis expansion

$$\psi'(\beta^T X) = \sum_{j=1}^{k} \gamma_j B_j(\beta^T x) = \gamma^T B(\beta^T X),$$

where B_j, $j = 1, \cdots, k$ are the B-spline basis functions (de Boor, 1978) and k is the degree of freedom for the B-spline.. Other choices of a basis function can be used here as well, but B-splines are preferable since they are numerically stable and computationally efficient.

Since any constant in the link function can be absorbed into the baseline function $\lambda_0(\cdot)$, the link function $(\alpha^T V + \psi(\beta^T X))$ is not identifiable. Thus, we specify that when $(V, X) = 0$, the link function is 0 for identifiability reasons and so $\psi(0) = 0$. Then

$$\psi(\beta^T X) = \int_0^{\beta^T x} \sum_{j=1}^{k} \gamma_j B_j(t) dt = \sum_{j=1}^{k} \gamma_j \widetilde{B}_j(\beta^T X) = \gamma^T \widetilde{B}(\beta^T X),$$

where $\widetilde{B}_j(u) = \int_0^u B(s)ds$, $j = 1, \cdots, k$, are the integrals of the B-spline basis functions; $B(u) = (B_1(u), \cdots, B_k(u))^T$; $\widetilde{B}(u) = (\widetilde{B}_1(u), \cdots, \widetilde{B}_k(u))^T$; and $\gamma = (\gamma_1, \cdots, \gamma_k)^T$.

The link function will then become

$$\alpha^T V + \gamma^T \widetilde{B}(\beta^T X).$$

We use quadratic B-splines in the basis expansion of $\psi'(\cdot)$ so that $\psi(\cdot)$ will be a cubic spline.

Let $\tau_1 < \cdots < \tau_m$ be m distinctive ordered event times and $(V_{(i)}, X_{(i)})$ be the ith covariate associated with the individual whose failure time is τ_i. Define D_i as the index set of units failing at time point τ_i and, thus, $D_i = \{j : Z_j = \tau_i\,,\, \delta_j = 1\}$, and $|D_i|$ denotes the size of D_i. The partial likelihood is defined as

$$PL = \prod_{i=1}^{m} \frac{\exp\{v_{(i)}^T \alpha + \gamma^T \widetilde{B}(x_{(i)}^T \beta)\}}{\left[\sum_{l \in R_i} \exp\{v_l^T \alpha + \gamma^T \widetilde{B}(x_l^T \beta)\}\right]^{|D_i|}}, \tag{2.1}$$

where $R_i = \{l : Z_l \geq \tau_i\}$ is the risk set at event time τ_i, $i = 1, \cdots, m$.

Therefore, the log partial likelihood is

$$\begin{aligned} l(\beta, \gamma, \alpha) &= \log(PL) \\ &= \sum_{i=1}^{m} \sum_{j \in D_i} \{v_j^T \alpha + \gamma^T \widetilde{B}(x_j^T \beta)\} \\ &\quad - \sum_{i=1}^{m} |D_i| \log \left[\sum_{l \in R_i} \exp\{v_l^T \alpha + \gamma^T \widetilde{B}(x_l^T \beta)\} \right]. \end{aligned}$$

Denote $\hat{\beta}$, $\hat{\gamma}$ and $\hat{\alpha}$ as the values which maximize this partial likelihood. Then the spline estimate of the unknown function is $\hat{\psi}(u) = \hat{\gamma}^T \widetilde{B}(u)$ and the regression parameter estimates are $\hat{\beta}$ and $\hat{\alpha}$. When there exist ties among event times, standard procedures (Kalbfleisch and Prentice, 2002) for handling ties for the Cox PH model can be used. We use Breslow's approximation in our implementation.

The joint score vector is given by $S_{(\beta,\gamma,\alpha)} = (S_\beta^T, S_\gamma^T, S_\alpha^T)^T$, while the full Hessian matrix is given by

$$H_{(\beta,\gamma,\alpha)} = \begin{pmatrix} H_{\beta,\beta} & H_{\beta,\gamma} & H_{\beta,\alpha} \\ H_{\beta,\gamma}^T & H_{\gamma,\gamma} & H_{\gamma,\alpha} \\ H_{\beta,\alpha}^T & H_{\gamma,\alpha}^T & H_{\alpha,\alpha} \end{pmatrix}.$$

It is quite straightforward to get the score function, S_β, and the Hessian matrix, $H_{\beta,\beta}$, of log likelihood l with respect to β:

$$S_\beta = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} \left(\sum_{j \in D_i} (\gamma^T B(x_j^T \beta) x_j) - |D_i| \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) x_l \right),$$

and $H_{\beta,\beta} = H_1 - H_2$, where $H_1 = \sum_{i=1}^{m} \sum_{j \in D_i} (\gamma^T B'(x_j^T \beta) x_j x_j^T)$ and

$$\begin{aligned} H_2 &= \sum_{i=1}^{m} |D_i| \left[\sum_{l \in R_i} w_{li} \left\{ \gamma^T B'(x_i^T \beta) + \{\gamma^T B(x_l^T \beta)\}^2 \right\} x_l x_l^T \right] \\ &\quad - \sum_{i=1}^{m} |D_i| \left[\sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) x_l \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) x_l^T \right], \end{aligned}$$

with

$$w_{li} = \frac{\exp\left\{v_l^T\alpha + \gamma^T\widetilde{B}(x_l^T\beta)\right\}}{\sum_{j\in R_i}\exp\left\{v_j^T\alpha + \gamma^T\widetilde{B}(x_j^T\beta)\right\}},$$

and $B'(u) = (B_1'(u), \cdots, B_k'(u))^T$.

Similarly, we can get the score function and the Hessian matrix for γ:

$$S_\gamma = \frac{\partial l}{\partial\gamma} = \sum_{i=1}^{m}\left(\sum_{j\in D_i}\widetilde{B}(x_j^T\beta) - |D_i|\sum_{l\in R_i}w_{li}\widetilde{B}(x_l^T\beta)\right),$$

$$\begin{aligned} H_{\gamma,\gamma} &= \frac{\partial^2 l}{\partial\gamma\partial\gamma^T} \\ &= -\sum_{i=1}^{m}|D_i|\left\{\sum_{l\in R_i}w_{li}\widetilde{B}(x_l^T\beta)\widetilde{B}^T(x_l^T\beta) - \sum_{l\in R_i}w_{li}\widetilde{B}(x_l^T\beta)\sum_{l\in R_i}w_{li}\widetilde{B}^T(x_l^T\beta)\right\}, \end{aligned}$$

and the score function, and the Hessian matrix of l with respect to α:

$$S_\alpha = \frac{\partial l}{\partial\alpha} = \sum_{i=1}^{m}\left(\sum_{j\in D_i}v_j - |D_i|\sum_{l\in R_i}w_{li}v_l\right),$$

$$H_{\alpha,\alpha} = \frac{\partial^2 l}{\partial\alpha\partial\alpha^T} = -\sum_{i=1}^{m}|D_i|\left\{\sum_{l\in R_i}w_{li}\cdot v_l\cdot v_l^T - \sum_{l\in R_i}w_{li}\cdot v_l\sum_{l\in R_i}w_{li}\cdot v_l^T\right\}.$$

Also we can easily get $H_{\beta,\gamma} = \partial^2 l/\partial\beta\partial\gamma$, $H_{\beta,\alpha} = \partial^2 l/\partial\beta\partial\alpha$, and

$$H_{\gamma,\alpha} = \frac{\partial l^2}{\partial\gamma\partial\alpha} = -\sum_{i=1}^{m}|D_i|\left\{\sum_{l\in R_i}w_{li}\widetilde{B}(x_l^T\beta)v_l^T - \sum_{l\in R_i}w_{li}\widetilde{B}(x_l^T\beta)\sum_{l\in R_i}w_{li}\cdot v_l^T\right\}.$$

It is easily seen that $H_{\gamma,\gamma}$ and $H_{\alpha,\alpha}$ are negative semi-definite, which implies that the log partial likelihood l is a concave function of γ for fixed β and α, also l is a concave function of α for fixed β and γ. If we combine γ, α into one vector (γ,α), then l should be a concave function of (γ,α) for fixed β. The joint score vector is $S_{(\gamma,\alpha)} = (S_\gamma^T, S_\alpha^T)^T$, and the joint Hessian matrix is given as follows:

$$H_{(\gamma,\alpha)} = \begin{pmatrix} H_{\gamma,\gamma} & H_{\gamma,\alpha} \\ H_{\gamma,\alpha}^T & H_{\alpha,\alpha} \end{pmatrix}.$$

We apply an iterative alternating optimization procedure to calculate the maximum partial likelihood estimate which adopts the specific structure of the problem and is numerically stable. Note that for fixed β, the partial likelihood $l(\beta, \gamma, \alpha)$ is a concave function of (γ, α) whose maximum is uniquely defined, if it exists. We solve the maximization problem by iteratively maximizing $l(\beta, \gamma, \alpha)$ over β and (γ, α). More specifically, for the fixed current value $\hat{\beta}_c$ of β, we update the estimate of (γ, α) by maximizing $l(\hat{\beta}_c, \gamma, \alpha)$, and for the fixed current values of $\hat{\gamma}_c$ and $\hat{\alpha}_c$ of γ and α, we update the estimate of β by maximizing $l(\beta, \hat{\gamma}_c, \hat{\alpha}_c)$. The process is iterated until some specified convergence criterion is met. We find that the proposed procedure is easy to implement and the algorithm usually converged very quickly in our simulation studies.

While it is possible to maximize the log-partial likelihood simultaneously with respect to β, γ and α, we find the iterative alternating optimization more appealing. The iterative procedure is numerically more stable and computationally simpler, as there is no need to calculate a larger Hessian matrix as would do if simultaneous optimization were implemented with the Newton-Raphson algorithm. Also, when we fix β to update (γ, α), the partial likelihood $l(\beta, \gamma, \alpha)$ is a concave function of (γ, α) whose maximum is uniquely defined, if it exists. We use the step-halving method to avoid downhill steps and guarantee that each step increases the likelihood. Also, during each step of the Newton-Raphson iteration, we will keep the norm of β the same as the norm of β_0. Here we just use the standardized β_0, in other words, $||\beta_0|| = 1$, to ensure it is identifiable, where $||\cdot||$ denotes the Euclidean norm. The constraint $||\beta|| = 1$ on the single-index coefficient parameters is then required for parameter identifiability. We also keep the direction of the first element of β in the

same direction of the first element of β_0 during iteration. To further simplicity of the maximization process, we require the first component of β_0 to be positive for identifiability purposes.

2.3 Inference

Since we used a constraint, $\|\beta\| = 1$, for identifiability purpose, in order to obtain the variance-covariance matrix of $(\hat{\beta}, \hat{\gamma}, \hat{\alpha})$, we reparameterize $\beta = ((1-\|\sigma\|^2)^{1/2}, \sigma_1, \cdots, \sigma_{p-1})^T$ with $\sigma = (\sigma_1, \cdots, \sigma_{p-1})$.

Let $G^{-1}: \quad (\beta, \gamma, \alpha) \rightarrow (\sigma, \gamma, \alpha)$, and then $G: \quad (\sigma, \gamma, \alpha) \rightarrow (\beta, \gamma, \alpha)$.

The observed Fisher information, $I(\sigma, \gamma, \alpha)$, of this reparametrization vector (σ, γ, α) equals $-H(\sigma, \gamma, \alpha)$, the negative of the Hessian of (σ, γ, α). By a standard application of the martingale theory of partial likelihood (e.g. an extension of the arguments by Prentice and Self, 1983; also see Section 2.6 for the proof of asymptotic normality), we can show that $(\hat{\sigma}, \hat{\gamma}, \hat{\alpha})$ is asymptotically normal with an estimated asymptotic variance-covariance matrix $\{I(\hat{\sigma}, \hat{\gamma}, \hat{\alpha})\}^{-1}$.

By the delta method, $(\hat{\beta}, \hat{\gamma}, \hat{\alpha})$ are asymptotically normal with an estimated asymptotic variance-covariance matrix

$$\begin{aligned} var(\hat{\beta}, \hat{\gamma}, \hat{\alpha}) &= var(G(\hat{\sigma}, \hat{\gamma}, \hat{\alpha})) \\ &= G'(\hat{\sigma}, \hat{\gamma}, \hat{\alpha}) var(\hat{\sigma}, \hat{\gamma}, \hat{\alpha}) [G'(\hat{\sigma}, \hat{\gamma}, \hat{\alpha})]^T, \end{aligned}$$

where G' denotes the derivative of G with respect to (σ, γ, α) and has the estimated

form,

$$
\begin{aligned}
G'(\hat{\sigma},\hat{\gamma},\hat{\alpha}) &= (\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p,\hat{\gamma}_1,\cdots,\hat{\gamma}_k,\hat{\alpha}_1,\cdots,\hat{\alpha}_q)' \\
&= (\hat{\beta}_1,\hat{\sigma}_1,\cdots,\hat{\sigma}_{p-1},\hat{\gamma}_1,\cdots,\hat{\gamma}_k,\hat{\alpha}_1,\cdots,\hat{\alpha}_q)' \\
&= \begin{pmatrix}
\frac{\partial\hat{\beta}_1}{\partial\hat{\sigma}_1} & \frac{\partial\hat{\beta}_1}{\partial\hat{\sigma}_2} & \cdots & \frac{\partial\hat{\beta}_1}{\partial\hat{\sigma}_{p-1}} & 0 & \cdots & 0 \\
\frac{\partial\hat{\sigma}_1}{\partial\hat{\sigma}_1}=1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & \frac{\partial\hat{\sigma}_2}{\partial\hat{\sigma}_2}=1 & \cdots & 0 & 0 & \cdots & 0 \\
 & & & \cdots & & & \\
0 & \cdots & & \frac{\partial\hat{\sigma}_{p-1}}{\partial\hat{\sigma}_{p-1}}=1 & 0 & & \\
0 & \cdots & & & \frac{\partial\hat{\gamma}_1}{\partial\hat{\gamma}_1}=1 & 0 & \\
 & & & \cdots & & &
\end{pmatrix} \\
&= \begin{pmatrix}
\frac{\hat{\beta}_2}{\hat{\beta}_1} \cdots \frac{\hat{\beta}_p}{\hat{\beta}_1} & 0 \cdots 0 \\
I_{p+k+q-1}
\end{pmatrix}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
var(\hat{\beta},\hat{\gamma},\hat{\alpha}) &= \begin{pmatrix}
\frac{\hat{\beta}_2}{\hat{\beta}_1} \cdots \frac{\hat{\beta}_p}{\hat{\beta}_1} & 0 \cdots 0 \\
I_{p+k+q-1}
\end{pmatrix} \\
&\quad \times\{-H(\hat{\sigma},\hat{\gamma},\hat{\alpha})\}^{-1} \begin{pmatrix}
\frac{\hat{\beta}_2}{\hat{\beta}_1} \cdots \frac{\hat{\beta}_p}{\hat{\beta}_1} & 0 \cdots 0 \\
I_{p+k+q-1}
\end{pmatrix}^T,
\end{aligned}
$$

where $I_{p+k+q-1}$ is the $(p+k+q-1)\times(p+k+q-1)$ identity matrix, and $H(\hat{\sigma},\hat{\gamma},\hat{\alpha})$ is the Hessian matrix of $(\hat{\sigma},\hat{\gamma},\hat{\alpha})$.

The Hessian matrix, $H_{(\sigma,\gamma,\alpha)}$, can be expressed as:

$$
H_{(\sigma,\gamma,\alpha)} = \begin{pmatrix}
H_{\sigma,\sigma} & H_{\sigma,\gamma} & H_{\sigma,\alpha} \\
H_{\sigma,\gamma}^T & H_{\gamma,\gamma} & H_{\gamma,\alpha} \\
H_{\sigma,\alpha}^T & H_{\gamma,\alpha}^T & H_{\alpha,\alpha}
\end{pmatrix},
$$

where $H_{\gamma,\gamma}$, $H_{\gamma,\alpha}$ and $H_{\alpha,\alpha}$ are the same as given in Section 2.2.

Denote the covariate vector for individual i as $\widetilde{x}_i^T = (x_{i2}, \cdots, x_{ip})^T$. Using σ to substitute for β, we have this new form of the $\log(PL)$

$$\begin{aligned} l(\sigma,\gamma,\alpha) &= \log(PL) \\ &= \sum_{i=1}^{m}\sum_{j\in D_i}\left\{v_j^T\alpha + \gamma^T\widetilde{B}(\widetilde{x}_j^T\sigma + x_{j1}(1-\|\sigma\|^2)^{1/2})\right\} \\ &\quad -\sum_{i=1}^{m}|D_i|\log\sum_{l\in R_i}\exp\left\{v_l^T\alpha + \gamma^T\widetilde{B}(\widetilde{x}_l^T\sigma + x_{l1}(1-\|\sigma\|^2)^{1/2})\right\}. \end{aligned}$$

The score function of σ is as follows

$$\begin{aligned} S_\sigma &= \frac{\partial l}{\partial\sigma} \\ &= \sum_{i=1}^{m}\sum_{j\in D_i}\left[\gamma^T B(\widetilde{x}_l^T\sigma + x_{l1}(1-\|\sigma\|^2)^{1/2})(\widetilde{x}_j^T - \frac{x_{j1}\sigma}{(1-\|\sigma\|^2)^{1/2}})\right] \\ &\quad -\sum_{i=1}^{m}|D_i|\sum_{l\in R_i}w_{li}\gamma^T B(\widetilde{x}_l^T\sigma + x_{l1}(1-\|\sigma\|^2)^{1/2})(\widetilde{x}_l^T - \frac{x_{l1}\sigma}{(1-\|\sigma\|^2)^{1/2}}). \end{aligned}$$

Let $\xi_i = -x_{i1}\sigma/(1-\|\sigma\|^2)^{1/2} + \widetilde{x}_i^T$, $i = 1, \cdots, n$. Then we will have

$$S_\sigma = \frac{\partial l}{\partial\sigma} = \sum_{i=1}^{m}\sum_{j\in D_i}\gamma^T B(x_j^T\beta)\xi_j - \sum_{i=1}^{m}|D_i|\sum_{l\in R_i}w_{li}\gamma^T B(x_l^T\beta)\xi_l.$$

Then it is straightforward to derive $H_{\sigma,\sigma}$ from the score function S_σ,

$$\begin{aligned}
H_{\sigma,\sigma} &= \frac{\partial S_\sigma}{\partial \sigma} \\
&= \sum_{i=1}^{m}\sum_{j\in D_i} \gamma^T B(\widetilde{x}_j^T\sigma + x_{j1}(1-\|\sigma\|^2)^{1/2})x_{j1}\left(-\frac{I_{p-1}+\sigma\sigma^T/(1-\|\sigma\|^2)}{\sqrt{1-\|\sigma\|^2}}\right) \\
&\quad + \sum_{i=1}^{m}\sum_{j\in D_i} \gamma^T B'(\widetilde{x}_j^T\sigma + x_{j1}(1-\|\sigma\|^2)\xi_j\xi_j^T \\
&\quad - \sum_{i=1}^{m}|D_i|\sum_{l\in R_i} w_{li}\left[\gamma^T B(x_l^T\beta)x_{l1}\left(-\frac{I_{p-1}+\sigma\sigma^T/(1-\|\sigma\|^2)}{\sqrt{1-\|\sigma\|^2}}\right)\right] \\
&\quad - \sum_{i=1}^{m}|D_i|\sum_{l\in R_i} w_{li}\left[\gamma^T B'(x_l^T\beta)\xi_l\xi_l^T\right] \\
&\quad - \sum_{i=1}^{m}|D_i|\sum_{l\in R_i} w_{li}\left[\{\gamma^T B(x_l^T\beta)\}^2\xi_l\xi_l^T\right] \\
&\quad + \sum_{i=1}^{m}|D_i|\sum_{l\in R_i} w_{li}\gamma^T B(x_l^T\beta)\xi_l \sum_{l\in R_i} w_{li}\gamma^T B(x_l^T\beta)\xi_l^T.
\end{aligned}$$

Let $A=(a_{ij})$ be a $(p-1)\times(p-1)$ matrix with entries $a_{ii}=1+\sigma_i^2/(1-\|\sigma\|^2)$ and $a_{ij}=\sigma_i\sigma_j/(1-\|\sigma\|^2)$, $i\neq j$, $i,j=1,\cdots,p-1$, in other words, $A=I_{p-1}+\sigma\sigma^T/(1-\|\sigma\|^2)$. Then we will have

$$\begin{aligned}
H_{\sigma,\sigma} &= \sum_{i=1}^{m}\sum_{j\in D_i} \gamma^T B(x_j^T\beta)(-\frac{x_{j1}}{\sqrt{1-\|\sigma\|^2}})A + \sum_{i=1}^{m}\sum_{j\in D_i} \gamma^T B'(x_j^T\beta)\xi_j\xi_j^T \\
&\quad - \sum_{i=1}^{m}|D_i|\sum_{l\in R_i} w_{li}\left[\gamma^T B(x_l^T\beta)(-\frac{x_{l1}}{\sqrt{1-\|\sigma\|^2}})A\right] \\
&\quad - \sum_{i=1}^{m}|D_i|\sum_{l\in R_i} w_{li}\left[\gamma^T B'(x_l^T\beta)\xi_l\xi_l^T + \{\gamma^T B(x_l^T\beta)\}^2\xi_l\xi_l^T\right] \\
&\quad + \sum_{i=1}^{m}|D_i|\left\{\sum_{l\in R_i} w_{li}\gamma^T B(x_l^T\beta)\xi_l \sum_{l\in R_i} w_{li}\gamma^T B(x_l^T\beta)\xi_l^T\right\}.
\end{aligned}$$

Recall

$$w_{li} = \frac{\exp\left\{v_l^T\alpha + \gamma^T\widetilde{B}(x_l^T\beta)\right\}}{\sum_{j\in R_i}\exp\left\{v_j^T\alpha + \gamma^T\widetilde{B}(x_j^T\beta)\right\}},$$

and so we obtain

$$\begin{aligned}
H_{\sigma,\gamma} &= \frac{\partial S_\sigma}{\partial\gamma} \\
&= \sum_{i=1}^{m}\sum_{j\in D_i}\xi_j B^T(x_j^T\beta) \\
&\quad -\sum_{i=1}^{m}|D_i|\sum_{l\in R_i}\frac{\exp\{v_l^T\alpha+\gamma^T\widetilde{B}(x_l^T\beta)\}}{\sum_{j\in R_i}\exp\{v_j^T\alpha+\gamma^T\widetilde{B}(x_j^T\beta)\}}\xi_l B^T(x_l^T\beta) \\
&\quad -\sum_{i=1}^{m}|D_i|\sum_{l\in R_i}\frac{\exp\{v_l^T\alpha+\gamma^T\widetilde{B}(x_l^T\beta)\}}{\sum_{j\in R_i}\exp\{v_j^T\alpha+\gamma^T\widetilde{B}(x_j^T\beta)\}}\gamma^T B(x_l^T\beta)\xi_l\widetilde{B}^T(x_l^T\beta) \\
&\quad +\sum_{i=1}^{m}|D_i|\sum_{l\in R_i}\frac{\exp\{v_l^T\alpha+\gamma^T\widetilde{B}(x_l^T\beta)\}}{[\sum_{j\in R_i}\exp\{v_j^T\alpha+\gamma^T\widetilde{B}(x_j^T\beta)\}]^2}\gamma^T B(x_l^T\beta)\xi_l \\
&\qquad \cdot\sum_{l\in R_i}\exp\{\gamma^T\widetilde{B}(x_l^T\beta)+v_l^T\alpha\}\widetilde{B}^T(x_l^T\beta) \\
&= \sum_{i=1}^{m}\sum_{j\in D_i}\xi_j B^T(x_j^T\beta) \\
&\quad -\sum_{i=1}^{m}|D_i|\sum_{l\in R_i}w_{li}\left\{\xi_l B^T(x_l^T\beta)+\gamma^T B(x_l^T\beta)\xi_l\widetilde{B}^T(x_l^T\beta)\right\} \\
&\quad +\sum_{i=1}^{m}|D_i|\left\{\sum_{l\in R_i}w_{li}\gamma^T B(x_l^T\beta)\xi_l\sum_{l\in R_i}w_{li}\widetilde{B}^T(x_l^T\beta)\right\},
\end{aligned}$$

$$
\begin{aligned}
H_{\sigma,\alpha} &= \frac{\partial S_\sigma}{\partial \alpha} \\
&= -\sum_{i=1}^{m} |D_i| \sum_{l \in R_i} \frac{\exp\{v_l^T\alpha + \gamma^T \widetilde{B}(x_l^T\beta)\}}{\sum_{j \in R_i} \exp\{v_j^T\alpha + \gamma^T \widetilde{B}(x_j^T\beta)\}} \gamma^T B(x_l^T\beta)\xi_l v_l^T \\
&\quad + \sum_{i=1}^{m} |D_i| \sum_{l \in R_i} \frac{\exp\{v_l^T\alpha + \gamma^T \widetilde{B}(x_l^T\beta)\}}{[\sum_{j \in R_i} \exp\{v_j^T\alpha + \gamma^T \widetilde{B}(x_j^T\beta)\}]^2} \gamma^T B(x_l^T\beta)\xi_l \\
&\quad \cdot \sum_{l \in R_i} \exp\{v_l^T\alpha + \gamma^T \widetilde{B}(x_l^T\beta)\} v_l^T \\
&= -\sum_{i=1}^{m} |D_i| \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T\beta)\xi_l v_l^T \\
&\quad + \sum_{i=1}^{m} |D_i| \left\{ \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T\beta)\xi_l \sum_{l \in R_i} w_{li} v_l^T \right\}.
\end{aligned}
$$

The first p elements on the diagonal of matrix $var(\hat{\beta}, \hat{\gamma}, \hat{\alpha})$ will give $var(\hat{\beta}_i)$, $i = 1, \cdots, p$ and the last q elements on the diagonal of matrix $var(\hat{\beta}, \hat{\gamma}, \hat{\alpha})$ will give $var(\hat{\alpha}_i)$, $i = 1, \cdots, q$. Thus, an approximate 95% confidence interval for β_i is $\hat{\beta}_i \pm 1.96\{var(\hat{\beta}_i)\}^{1/2}$ and, similarly, a 95% confidence interval for α_i is $\hat{\alpha}_i \pm 1.96\{var(\hat{\alpha}_i)\}^{1/2}$. For a fixed u, the variance of the estimated function evaluated at u can be estimated as $var(\hat{\psi}(u)) = \widetilde{B}(u)^T var(\hat{\gamma}) \widetilde{B}(u)$, where $var(\hat{\gamma})$ is given by the appropriate submatrix of $var(\hat{\beta}, \hat{\gamma}, \hat{\alpha})$. An approximate 95% confidence interval for $\psi(u)$ is $\hat{\psi}(u) \pm 1.96\{var(\hat{\psi}(u))\}^{1/2}$.

2.4 Implementation

The 3-step iterative Newton Raphson algorithm for this model are as follows:

Step 1. Start with initial values for $\hat{\beta}^{(0)}$, $\hat{\gamma}^{(0)}$ and $\hat{\alpha}^{(0)}$.

Step 2. Given the values of $\hat{\beta}^{(d)}$, $\hat{\gamma}^{(d)}$ and $\hat{\alpha}^{(d)}$, update the estimate of β using

one iteration of the Newton-Raphson method, using this expression to get $\hat{\beta}^{(d+1)}$,

$$\hat{\beta}^{(d+1)} = \hat{\beta}^{(d)} - \{H_{\hat{\beta}^{(d)},\hat{\beta}^{(d)}}(\hat{\beta}^{(d)}, \hat{\gamma}^{(d)}, \hat{\alpha}^{(d)})\}^{-1} S_{\hat{\beta}^{(d)}}(\hat{\beta}^{(d)}, \hat{\gamma}^{(d)}, \hat{\alpha}^{(d)}).$$

Then standardize $\hat{\beta}^{(d+1)}$, such that $\|\hat{\beta}^{(d+1)}\| = 1$ and ensure its first component is positive.

Step 3. Given the values of $\hat{\beta}^{(d+1)}$, $\hat{\gamma}^{(d)}$ and $\hat{\alpha}^{(d)}$, update the estimates of γ and α simultaneously to obtain $\hat{\gamma}^{(d+1)}$ and $\hat{\alpha}^{(d+1)}$, using one iteration of the Newton-Raphson method with step-halving, as follows:

$$\begin{aligned}&(\hat{\gamma}^{(d+1)}, \hat{\alpha}^{(d+1)}) \\ &= (\hat{\gamma}^{(d)}, \hat{\alpha}^{(d)}) - 2^{-k}\{H_{(\hat{\gamma}^{(d)},\hat{\alpha}^{(d)})}(\hat{\beta}^{(d+1)}, \hat{\gamma}^{(d)}, \hat{\alpha}^{(d)})\}^{-1} S_{(\hat{\gamma}^{(d)},\hat{\alpha}^{(d)})}(\hat{\beta}^{(d+1)}, \hat{\gamma}^{(d)}, \hat{\alpha}^{(d)}),\end{aligned}$$

where k is the smallest nonnegative integer such that

$$l(\hat{\beta}^{(d+1)}, \hat{\gamma}^{(d+1)}, \hat{\alpha}^{(d+1)}) \geq l(\hat{\beta}^{(d+1)}, \hat{\gamma}^{(d)}, \hat{\alpha}^{(d)}).$$

Repeat Steps 2 and 3 until some specified convergence criterion is met.

We note here that the algorithm may not converge if the initial values are far from the maximum partial likelihood values, but it usually converges within a few steps for initial values reasonably close. In our simulation study, we ran the program using different random parameter values and the program converged for 95% initial values chosen.

Since the *coxph* function in the *R* package 'Survival' uses essentially the same idea to get the regression coefficient estimates, we can employ this function in our iterative algorithm. By using the *coxph* function in place of Step 3, the algorithm is modified as follows:

Step 3 (alternative).

ss $\leftarrow$ coxph(Surv(Z, δ) $\sim \widetilde{B}(X^T\hat{\beta}^{(d+1)}) + \mathrm{V}$)

$(\hat{\gamma}^{(d+1)}, \hat{\alpha}^{(d+1)}) \leftarrow$ coef(ss)

After comparing these two methods for Step 3, we obtained identical results, but the second method, which combines the Cox PH and the Newton-Raphson method, takes half the time of the first method. Hence, we recommend the use of the second method for implementation of our approach.

2.5 Number and knot position selection

Since each set of knots determines an approximating model, information based criteria are natural choices for selecting their number and position. In our implementation, we use uniform B-splines as smoothing functions. In other words, for a given number of knots, we put the knots at equally-spaced locations between the smallest and the largest values of $\beta^T X$. In our algorithm, we chose to use the method proposed by Akaike (1973) to select the knots. It is a decision-making strategy which uses a natural criterion for ordering alternative statistical models for the data.

Denote the total number of parameters to be estimated as $P = nknots + ds + p + q - 2$, where $nknots + ds - 2$ is the degree of freedom of B-splines, we need to minimize this quantity:

$$AIC = -2\log(PL(X|\,\theta)) + 2P,$$

where $PL(X|\,\theta)$ is the partial likelihood evaluated at the parameter $\theta = (\beta, \gamma, \alpha)$, P is the total independent parameters in θ, $nknots$ is the number of knots including two boundary points, ds is the degree of the spline, p is the number of X covariates,

and q is the number of V covariates. We can easily see the length of β is p, the length of γ, in other words, the degree of freedom of the B-spline, is $nknots + ds - 2$, and the length of α is q.

Schwarz (1978) proposed an alternative to the AIC called the BIC which uses a Bayesian information criterion. It minimizes

$$BIC = -2\log((PL(X|\,\theta)) + \log(n)P,$$

where n is the number of subjects or objects in a study.

To choose the number and position of the knots, we varied the number of knots in a relatively large range and choose the set which minimizes both the AIC and BIC. The differences between AIC and BIC was very minimal for sample sizes between 200 and 300. Tighter control on the number of knots was obtained with the BIC when the sample size was very large. In addition, we checked the sensitivity of these results for different numbers of knots. The knot positions change in each iteration of the program since β is updated, and so $\beta^T X$ is updated as a consequence. However, results were stable if the program converged.

2.6 Asymptotic Normality and Consistency

We will adopt the standard counting process formulation for survival analysis (Andersen and Gill, 1982; Liu, 2004) to show asymptotic normality and consistency of our proposed estimators. The multivariate counting process $N = (N_1, \cdots, N_n)$ is such that N_i counts failures on the i^{th} subject at times $t \in [0, 1]$ when the subject is under observation.

Defining the censoring process $Y(t) = (Y_1(t), \cdots, Y_n(t))$ as $Y_i(t) = I\{T_i \geq t, C_i \geq t\}$, so that $Y_i(t) = 1$ if the i^{th} subject is under observation at time t and $Y_i(t) = 0$ otherwise. Most questions of interest concern the relationship between failure rate and the histories of some covariate process. Let $X(t) = (X_1(t), \cdots, X_n(t))^T$ denote the covariate processes such that $X_i^T(t) = \{X_{i1}(t), \cdots, X_{ip}(t)\}$ denotes the counting process histories up to time t for subject i.

The counting process formulation permits each N_i to be uniquely decomposed into the sum of its cumulative intensity process Λ_i and a local square integrable martingale M_i, so that

$$N_i(t) = \Lambda_i(t) + M_i(t),$$

for all (t, i), $i = 1, \cdots, n$.

The increasing process Λ_i is, for convenience, taken to be absolutely continuous, giving

$$\Lambda_i(t) = \int_0^t \lambda_i(s)ds.$$

The intensity process $\lambda = (\lambda_1, \cdots, \lambda_n)^T$, under some regularity (e.g. each λ_i bounded by an integrable random variable, Aalen, 1978), in the manner of Cox (1972), can be written as

$$\lambda_i(t) = Y_i(t)\lambda_0(t)\exp\{g(X_i(t); \theta_0)\},\ i = 1, 2, \cdots, n$$

where $g(\cdot; \theta_0)$ is a function with known form $g(X_i(t); \theta_0) = \theta_0^T X_i(t)$ and θ_0 is a vector of unknown parameters. In our context, we replace $X_i(t)$ by $\{V_i(t), X_i(t)\}$ and obtain

$$g(V_i(t), X_i(t); \theta) = V_i(t)^T\alpha + \sum_{j=1}^{k} \gamma_j B_j(X_i(t)^T\beta),$$

where $X_i^T = (X_{i1}, \cdots, X_{ip})$, $V_i^T = (V_{i1}, \cdots, V_{iq})$, $\theta = (\beta, \gamma, \alpha)$, $\gamma = (\gamma_1, \cdots, \gamma_k)$, B_j, $j = 1, \cdots, k$, are B-spline basis functions. Since $\exp\{g(V_i(t), X_i(t); \theta)\} > 0$ is always continuous and twice continuously differentiable, we can adopt the proof of Prentice and Self (1983) for the asymptotic distribution theory in our case.

Note that we focus our attention on the interval $[0, 1]$ for simplicity. As discussed in Andersen and Gill (1982), the argument can be easily extended to the interval $[0, \infty)$.

The log partial likelihood can be written as

$$\begin{aligned} \log L(\theta, t) &= \sum_{i=1}^{n} \int_0^t g(V_i(s), X_i(s); \theta) dN_i(s) \\ &\quad - \int_o^t \log \left[\sum_{i=1}^{n} Y_i(s) \exp\{g(V_i(s), X_i(s); \theta)\} \right] d\overline{N}(s), \end{aligned}$$

where $\overline{N} = \sum_{i=1}^{n} N_i$. The maximum partial likelihood estimate solves the partial likelihood equation $\partial \log L(\theta, 1)/\partial\theta = 0$. By the Doob-Meyer decomposition,

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(s) ds, \quad i = 1, 2, \cdots, n, \ t \in [0, 1]$$

are local martigales on the interval $[0, 1]$.

For a column vector a, we denote $a^{\otimes 2}$ for the matrix aa^T, $\|a\| = \sup_i |a_i|$, and $|a| = (a'a)^{1/2}$. For a matrix A, denote $\|A\| = \sup_{i,j} |a_{ij}|$. For a function $g(v, x; \theta)$, let $\dot{g}(v, x; \theta)$ and $\ddot{g}(v, x; \theta)$ denote the gradient and Hessian of g relative to θ. We define

$$S^{(0)}(\theta, t) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp\{g(V_i(t), X_i(t); \theta)\},$$

$$S^{(1)}(\theta, t) = \partial S^{(0)}(\theta, t)/\partial\theta = \frac{1}{n} \sum_{i=1}^{n} \dot{g}(V_i(t), X_i(t); \theta) Y_i(t) \exp\{g(V_i(t), X_i(t); \theta)\},$$

$$S^{(2)}(\theta,t)=\frac{1}{n}\sum_{i=1}^{n}\dot{g}(V_i(t),X_i(t);\theta)\dot{g}^T(V_i(t),X_i(t);\theta)Y_i(t)\exp\{g(V_i(t),X_i(t);\theta)\},$$

$$\begin{aligned}S^{(3)}(\theta,t) &= \partial^2 S^{(0)}(\theta,t)/\partial\theta^2\\ &= \frac{1}{n}\sum_{i=1}^{n}\{\ddot{g}(V_i(t),X_i(t);\theta)+\dot{g}(V_i(t),X_i(t);\theta)\dot{g}^T(V_i(t),X_i(t);\theta)\}\\ &\quad\cdot Y_i(t)\exp\{g(V_i(t),X_i(t);\theta)\},\end{aligned}$$

$$S^{(4)}(\theta,t)=\frac{1}{n}\sum_{i=1}^{n}\{g(V_i(t),X_i(t);\theta)-g(V_i(t),X_i(t);\theta_0)\}Y_i(t)\exp\{g(V_i(t),X_i(t);\theta_0)\},$$

$$S^{(5)}(\theta,t)=\partial S^{(4)}(\theta,t)/\partial\theta=\frac{1}{n}\sum_{i=1}^{n}\dot{g}(V_i(t),X_i(t);\theta)Y_i(t)\exp\{g(V_i(t),X_i(t);\theta_0)\},$$

$$S^{(6)}(\theta,t)=\partial^2 S^{(4)}(\theta,t)/\partial\theta^2=\frac{1}{n}\sum_{i=1}^{n}\ddot{g}(V_i(t),X_i(t);\theta)Y_i(t)\exp\{g(V_i(t),X_i(t);\theta_0)\},$$

and further

$$E(\theta,t)=S^{(1)}(\theta,t)/S^{(0)}(\theta,t),$$

$$Var(\theta,t)=S^{(2)}(\theta,t)/S^{(0)}(\theta,t)-E(\theta,t)^{\otimes 2}.$$

Note that $E(\theta_0,t)$ and $Var(\theta_0,t)$ can be thought of as the expected covariate vector at time t and corresponding covariace matrix for a study subject failing at t.

We assume the following conditions:

(1). (Finite integral). $\int_0^1 \lambda_0(t)dt<\infty$.

(2). (Asymptotic stability). There exists a compact neighborhood Θ of θ_0 and functions $s^{(k)}$, $k = 0, 1, \cdots, 6$, defined on $\Theta \times [0,1]$ such that for $k = 0, 1, \cdots, 6$

$$\sup_{t\in[0,1],\theta\in\Theta} \|S^{(k)}(\theta,t) - s^{(k)}(\theta,t)\| \xrightarrow{P} 0.$$

(3). (Lindeberg condition). Let $W_i(t) = (V_i^T(t), X_i^T(t))^T$,

$$\int_0^1 n^{-1} \sum_{i=1}^{n} [\dot{g}(W_{ij}(t);\theta_0) - E(\theta_0,t)_j]^2 Y_i(t) \exp\{g(W_{ij}(t);\theta_0)\}$$
$$I\{n^{-1/2}|\dot{g}(W_{ij}(t);\theta_0) - E(\theta_0,t)_j| > \epsilon\}\lambda_0(t)dt \xrightarrow{P} 0,$$

for any $\epsilon > 0$ and $j = 1, 2, \cdots, q+p$.

(4). (Asymptotic regularity conditions). Let $e = s^{(1)}/s^{(0)}$ and $var = s^{(2)}/s^{(0)} - e^{\otimes 2}$. For each $k = 0, 1, \cdots, 6$, $s^{(k)}(\cdot, t)$ are continuous functions of $\theta \in \Theta$, uniformly in $t \in [0,1]$. Also, $s^{(k)}$, $k = 0, \cdots, 6$, are bounded on $\Theta \times [0,1]$, $s^{(0)}$ is bounded away from zero and the matrix

$$\Sigma = \int_0^1 var(\theta_0,t) s^{(0)}(\theta_0,t)\lambda_0(t)dt,$$

is positive definite. Also, $s^{(0)}(\theta,t)$ and $s^{(4)}(\theta,t)$ are assumed to be twice differentiable with respect to θ on $\Theta \times [0,1]$.

(5). $\dot{g}(W_{ij}(t);\theta_0)$ is locally bounded for $i = 1, \cdots, n$, $j = 1, \cdots, q+p$.

(6).

$$\sup_{\theta\in\Theta} \int_0^t n^{-2} \sum_{i=1}^{n} \|\ddot{g}(V_i(t), X_i(t);\theta)\|^2 Y_i(s)$$
$$\exp\{g(V_i(t), X_i(t);\theta_0)\}\lambda_0(s)ds \xrightarrow{P} 0.$$

Remark 2.6.1. In the special case of the standard Cox model, $g(X;\theta_0) = X^T\theta_0$, these conditions reduce precisely to those given by Andersen and Gill (1982). In

view of the term n^{-2}, condition(6) is rather weak. It vanishes when $g(X;\theta_0) = X^T\theta_0$ since $\ddot{g}(X;\theta) = 0$.

We need the following lemma in our proofs of the asymptotic results.

Lemma 2.6.2. *(Lenglart Inequality).*

Let N be a univariate counting process with continuous compensator A, let $M = N - A$, and let H be a locally bounded, predictable process. Then, for all $\delta, \rho > 0$,

$$P\{\sup_{0\le t\le 1} |\int_0^t H(x)dM(x)| \ge \rho\} \le \frac{\delta}{\rho^2} + P\{\int_0^1 H^2(x)dA(x) \ge \delta\}.$$

Proof: See Fleming and Harrington (1991), p.291.

Now we will show the consistency of $\hat{\theta}$, $\hat{\theta}_n \xrightarrow{P} \theta_0$ in the following content.

Theorem 2.6.3. *There exists a sequence of roots $\hat{\theta}_n$ of the partial likelihood equation such that $\hat{\theta}_n \xrightarrow{P} \theta_0$.*

Proof:

Let

$$\begin{aligned} Z_n(\theta,t) &= \frac{1}{n}[\log L(\theta,t) - \log L(\theta_0,t)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \{g(V_i(s),X_i(s);\theta) - g(V_i(s),X_i(s);\theta_0)\}dN_i(s) \\ &\quad -\frac{1}{n}\int_0^t \log\frac{S^{(0)}(\theta,s)}{S^{(0)}(\theta_0,s)}d\overline{N}(s), \end{aligned}$$

and

$$\begin{aligned} A_n(\theta,t) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \{g(V_i(s),X_i(s);\theta) - g(V_i(s),X_i(s);\theta_0)\}\lambda_i(s)ds \\ &\quad -\frac{1}{n}\int_0^t \log\frac{S^{(0)}(\theta,s)}{S^{(0)}(\theta_0,s)}\overline{\lambda}(s)ds, \end{aligned}$$

where $\overline{\lambda} = \sum_{i=1}^{n} \lambda_i$. Then the process

$$Z_n(\theta,t) - A_n(\theta,t) = \frac{1}{n}\left[\sum_{i=1}^{n}\int_0^t \{g(V_i(s),X_i(s);\theta) - g(V_i(s),X_i(s);\theta_0) - \log\frac{S^{(0)}(\theta,s)}{S^{(0)}(\theta_0,s)}\}dM_i(s)\right],$$

is a locally square integrable martingale for each θ, with predictable variation process at t given by

$$\begin{aligned}
&\langle Z_n(\theta,\cdot) - A_n(\theta,\cdot), Z_n(\theta,\cdot) - A_n(\theta,\cdot)\rangle(t) \\
&= \frac{1}{n^2}\int_0^t \sum_{i=1}^{n} \{g(V_i(s),X_i(s);\theta) - g(V_i(s),X_i(s);\theta_0) - \log\frac{S^{(0)}(\theta,s)}{S^{(0)}(\theta_0,s)}\}^2\lambda_i(s)ds \\
&= \frac{1}{n^2}\int_0^t \sum_{i=1}^{n}\{g(V_i(s),X_i(s);\theta) - g(V_i(s),X_i(s);\theta_0)\}^2 Y_i(s) \\
&\quad \exp\{g(V_i(s),X_i(s);\theta_0)\}\lambda_0(s)ds \\
&\quad -\frac{2}{n^2}\int_0^t \sum_{i=1}^{n}\{g(V_i(s),X_i(s);\theta) - g(V_i(s),X_i(s);\theta_0)\} \\
&\quad \log\frac{S^{(0)}(\theta,s)}{S^{(0)}(\theta_0,s)} Y_i(s)\exp\{g(V_i(s),X_i(s);\theta_0)\}\lambda_0(s)ds \\
&\quad +\frac{1}{n}\int_0^t \{\log\frac{S^{(0)}(\theta,s)}{S^{(0)}(\theta_0,s)}\}^2 S^{(0)}(\theta_0,s)\lambda_0(s)ds \\
&= I_1 - I_2 + I_3.
\end{aligned}$$

By conditions (1), (2) and (4), the third term in the above equation, I_3, converges to zero in probability. The second term I_2 will also converge to zero in probability by the Cauchy-Schwarz inequality if we can show that the first term I_1 converges to zero in probability. It remains to be shown that the first integral in the expression above converges to zero.

By the Taylor expansion of $g(V_i(s), X_i(s); \theta)$ at θ_0, we have

$$\begin{aligned} & g(V_i(s), X_i(s); \theta) - g(V_i(s), X_i(s); \theta_0) \\ &= (\theta - \theta_0)' \dot{g}(V_i(s), X_i(s); \theta_0) + \frac{1}{2}(\theta - \theta_0)' \ddot{g}(V_i(s), X_i(s); \theta^*)(\theta - \theta_0), \end{aligned}$$

where θ^* is bounded between θ and θ_0. Hence,

$$I_1 = T_1 + T_2 + T_3,$$

where

$$\begin{aligned} T_1 &= \frac{1}{n^2} \int_0^t \sum_{i=1}^n (\theta - \theta_0)' \dot{g}(V_i(s), X_i(s); \theta_0) \dot{g}^T(V_i(s), X_i(s); \theta_0)(\theta - \theta_0) Y_i(s) \\ &\quad \cdot \exp\{g(V_i(s), X_i(s); \theta_0)\} \lambda_0(s) ds, \end{aligned}$$

$$\begin{aligned} T_2 &= \frac{1}{n^2} \int_0^t \sum_{i=1}^n (\theta - \theta_0)' \dot{g}(V_i(s), X_i(s); \theta_0)(\theta - \theta_0)' \ddot{g}^T(V_i(s), X_i(s); \theta_0)(\theta - \theta_0) Y_i(s) \\ &\quad \cdot \exp\{g(V_i(s), X_i(s); \theta_0)\} \lambda_0(s) ds, \end{aligned}$$

$$\begin{aligned} T_3 &= \frac{1}{4n^2} \int_0^t \sum_{i=1}^n \{(\theta - \theta_0)' \ddot{g}(V_i(s), X_i(s); \theta_0)(\theta - \theta_0)\}^2 Y_i(s) \\ &\quad \exp\{g(V_i(s), X_i(s); \theta_0)\} \lambda_0(s) ds. \end{aligned}$$

By conditions (1), (2) and (4), T_1 converges to zero in probability. Note that for any vector α and matrix A, we have $\alpha^T A \alpha \leq \|\alpha\|^2 \|A\|$. We observe that

$$T_3 \leq \frac{1}{4n^2} \int_0^t \sum_{i=1}^n \|\theta - \theta_0\|^4 p^4 \|\ddot{g}(V_i(s), X_i(s); \theta^*)\|^2 Y_i(s) \exp\{g(V_i(s), X_i(s); \theta_0)\} \lambda_0(s) ds.$$

So by condition (6), T_3 converges to zero in probability. Using the Cauchy-Schwarz inequality, T_2 converges to zero in probability. Hence, I_1 converges to zero

in probability. We conclude that $\langle Z_n(\theta,\cdot) - A_n(\theta,\cdot), Z_n(\theta,\cdot) - A_n(\theta,\cdot)\rangle$ converges to zero in probability. Thus, by Lemma 1,

$$\lim_{n\to\infty}\{Z_n(\theta,t) - A_n(\theta,t)\} = 0,$$

in probability for all $\theta \in \Theta$. Since Θ is a compact set, we have that $Z_n(\theta,t)$ converges to $A_n(\theta,t)$ in probability uniformly for $\theta \in \Theta$.

On the other hand, from conditions (1), (2) and (4)

$$\begin{aligned} A_n(\theta, t=1) &= \int_0^1 \left[S^{(4)}(\theta,s) - \log\frac{S^{(0)}(\theta,s)}{S^{(0)}(\theta_0,s)}S^{(0)}(\theta_0,s)\right]\lambda_0(s)ds \\ &\xrightarrow{P} \int_0^1 [s^{(4)}(\theta,s) - \log\frac{s^{(0)}(\theta,s)}{s^{(0)}(\theta_0,s)}s^{(0)}(\theta_0,s)]\lambda_0(s)ds \\ &= A(\theta,1). \end{aligned}$$

Note that

$$\frac{\partial A(\theta,1)}{\partial\theta} = \int_0^t [s^{(5)}(\theta,s) - \frac{s^{(0)}(\theta_0,s)}{s^{(0)}(\theta,s)}s^{(1)}(\theta,s)]\lambda_0(s)ds,$$

which equals zero at $\theta = \theta_0$ since $s^{(5)}(\theta_0,s) = s^{(1)}(\theta_0,s)$. Furthermore,

$$\frac{\partial^2 A(\theta,1)}{\partial\theta\partial\theta^T} = \int_0^t [s^{(6)}(\theta,s) - \{\frac{s^{(3)}(\theta,s)}{s^{(0)}(\theta,s)} - \frac{s^{(1)}(\theta,s)^{\otimes 2}}{s^{(0)}(\theta,s)^2}\}s^{(0)}(\theta_0,s)]\lambda_0(s)ds,$$

which equals to $-\Sigma$ at $\theta = \theta_0$ since $s^{(6)}(\theta_0,s) - s^{(3)}(\theta_0,s) = s^{(2)}(\theta_0,s)$. By condition (5), this is a negative definite matrix. Since $s^{(0)}$, $s^{(1)}$, $s^{(3)}$ and $s^{(6)}$ are continuous for $\theta \in \Theta$ uniformly for $t \in [0,1]$, by condition(4) there exists a compact neighborhood $\Theta_1 \subset \Theta$ such that the formula above is negative definite for $\theta \in \Theta_1$. Thus θ_0 is a local maximizer of $A(\theta,1)$. Note that $\lim_{n\to\infty}\{Z_n(\theta,1) - A(\theta,1)\} = 0$ in probability uniformly for $\theta \in \Theta_1$, thus the maximizer $\hat{\theta}$ of $Z_n(\theta,1)$ on Θ_1 will converge to θ_0. Since $\hat{\theta}$ should lie in the interior of Θ_1 for larger n, it solves the partial likelihood

equation. Thus, we have shown the existence of a sequence of consistent roots of the partial likelihood equation. The proof is completed.

Theorem 2.6.4. *(Asymptotic normality of $\hat{\theta}$).*

There exists a sequence of roots $\hat{\theta}_n$ of the partial likelihood equation such that

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0, \Sigma^{-1}).$$

Proof:

Recall that the score function of θ is

$$U(\theta, t) = \sum_{i=1}^{n} \int_0^t \dot{g}(V_i(s), X_i(s); \theta) dN_i(s) - \int_0^t \frac{S^{(1)}(\theta, t)}{S^{(0)}(\theta, t)} d\overline{N}(s).$$

By the Taylor expansion about θ_0,

$$U(\theta, 1) - U(\theta_0, 1) = -H(\theta^*, 1)(\theta - \theta_0),$$

where θ^* lies between θ and θ_0 and

$$H(\theta, t) = \int_0^t \sum_{i=1}^{n} \left\{ -\ddot{g}(V_i(s), X_i(s); \theta) + \frac{S^{(3)}(\theta, s)}{S^{(0)}(\theta, s)} - \left(\frac{S^{(1)}(\theta, s)}{S^{(0)}(\theta, s)} \right)^{\otimes 2} \right\} dN_i(s).$$

Hence

$$n^{-1} H(\theta^*, 1) \sqrt{n}(\hat{\theta} - \theta_0) = n^{-1/2} U(\theta_0, 1).$$

We will show that

$$n^{-1/2} U(\theta_0, 1) \xrightarrow{D} N(0, \Sigma)$$

and

$$n^{-1} H(\theta^*, 1) \xrightarrow{P} \Sigma.$$

Recall that

$$n^{-1/2} U(\theta_0, t) = \int_0^t n^{-1/2} \sum_{i=1}^{n} \{ \dot{g}(V_i(s), X_i(s); \theta_0) - E(\theta_0, s) \} dM_i(s).$$

Denote $H_{ij}(t) = n^{-1/2}\{\dot{g}(W_{ij};\theta_0) - E_j(\theta_0,s)\}$ and $H_i(t) = (H_{i1},\cdots,H_{i(p+q)})^T$. Then by conditions (2) and (5), $H_i(s)$ is a predictable and locally bounded process and

$$n^{-1/2}U(\theta_0,t) = \int_0^t n^{-1/2}\sum_{i=1}^n H_i(s)dM_i(s).$$

Therefore, by condition (2),

$$\begin{aligned}
\left\langle n^{-1/2}U(\theta_0,\cdot), n^{-1/2}U(\theta_0,\cdot)\right\rangle(t) &= \int_0^t n^{-1/2}\sum_{i=1}^n H_i(s)^{\otimes 2}Y_i(s) \\
&\quad \cdot\exp\{g(V_i(s),X_i(s);\theta_0)\}\lambda_0(s)ds \\
&= \int_0^t \left\{S^{(2)}(\theta_0,s) - \frac{S^{(1)}(\theta_0,s)^{\otimes 2}}{S^{(0)}(\theta_0,s)}\right\}\lambda_0(s)ds \\
&= \int_0^t Var(\theta_0,s)S^{(0)}(\theta_0,s)\lambda_0(s)ds \\
&\xrightarrow{P} \int_0^t var(\theta_0,s)s^{(0)}(\theta_0,s)\lambda_0(s)ds \stackrel{def}{=} \Sigma_t.
\end{aligned}$$

By condition (3) and Theorem 5.3.5 of Fleming and Harrington (1991), the local square integrable martingale $n^{-1/2}U(\theta_0,t)$ converges to a Gaussian process with mean zero and covariance function Σ_t. As a consequence, $n^{-1/2}U(\theta_0,1) \xrightarrow{D} N(0,\Sigma)$, since the limiting variance function $\Sigma_t = \int_0^t var(\theta_0,s)s^{(0)}(\theta_0,s)\lambda_0(s)ds$ evaluated at $t=1$ is just Σ.

Now we show that the scaled observed information matrix $n^{-1}H(\theta^*,1)$ converges to Σ in probability. Write

$$\begin{aligned}
n^{-1}H(\theta,t) &= \int_0^t \sum_{i=1}^n \left\{-\ddot{g}(V_i(s),X_i(s);\theta) + \frac{S^{(3)}(\theta,s)}{S^{(0)}(\theta,s)} - \left(\frac{S^{(1)}(\theta,s)}{S^{(0)}(\theta,s)}\right)^{\otimes 2}\right\} dN_i(s) \\
&= \int_0^t n^{-1}\sum_{i=1}^n \{W(\theta,s) - \ddot{g}(V_i(s),X_i(s);\theta)\}\, dN_i(s),
\end{aligned}$$

where $W(\theta,s) = S^{(3)}(\theta,s)/S^{(0)}(\theta,s) - (S^{(1)}(\theta,s)/S^{(0)}(\theta,s))^{\otimes 2}$.

Define

$$C(\theta,t)=\int_0^t n^{-1}\sum_{i=1}^{n}\{W(\theta,s)-\ddot{g}(V_i(s),X_i(s);\theta)\}\,\lambda_i(s)ds.$$

Then, $n^{-1}H(\theta,t)-C(\theta,t)$ is a local square integrable martingale with

$$\begin{aligned}
&\left\langle n^{-1}H(\theta,\cdot)-C(\theta,\cdot),n^{-1}H(\theta,\cdot)-C(\theta,\cdot)\right\rangle(t)\\
&= n^{-2}\int_0^t\sum_{i=1}^{n}\{W(\theta,s)-\ddot{g}(V_i(s),X_i(s);\theta)\}^2\,\lambda_i(s)ds\\
&= \int_0^t n^{-1}W(\theta,s)^{\otimes 2}S^{(0)}(\theta,s)\lambda_0(s)ds\\
&\quad-\int_0^t 2n^{-1}W(\theta,s)S^{(6)}(\theta,s)\lambda_0(s)ds\\
&\quad+\int_0^t n^{-2}\sum_{i=1}^{n}\ddot{g}(V_i(s),X_i(s);\theta)^{\otimes 2}Y_i(s)\exp\{g(V_i(s),X_i(s);\theta_0)\}\lambda_0(s)ds\\
&= B_1+B_2+B_3.
\end{aligned}$$

By conditions(1), (2) and (4), $B_1\xrightarrow{P}0$, $B_2\xrightarrow{P}0$, while $B_3\xrightarrow{P}0$ follows from condition (6). So we have

$$\lim_{n\to\infty}\{n^{-1}H(\theta,1)-C(\theta,1)\}\xrightarrow{P}0.$$

It follows from conditions (1), (2) and (4) that,

$$\begin{aligned}
&C(\theta,1)\\
&\xrightarrow{P}\int_0^1\left\{\frac{s^{(3)}(\theta,s)}{s^{(0)}(\theta,s)}s^{(0)}(\theta_0,s)-\left(\frac{s^{(1)}(\theta,s)}{s^{(0)}(\theta,s)}\right)^{\otimes 2}s^{(0)}(\theta_0,s)-s^{(6)}(\theta,s)\right\}\lambda_0(s)ds\\
&=\int_0^1\left\{\frac{s^{(2)}(\theta,s)}{s^{(0)}(\theta,s)}-\left(\frac{s^{(1)}(\theta,s)}{s^{(0)}(\theta,s)}\right)^{\otimes 2}\right\}s^{(0)}(\theta_0,s)\lambda_0(s)ds.
\end{aligned}$$

Thus, consistency of $\hat{\theta}$ implies that $\theta^*\xrightarrow{P}\theta_0$ and by conditions (1) and (4), we have

$$C(\theta^*,1)\xrightarrow{P}\int_0^1\left\{\frac{s^{(2)}(\theta_0,s)}{s^{(0)}(\theta_0,s)}-\left(\frac{s^{(1)}(\theta_0,s)}{s^{(0)}(\theta_0,s)}\right)^{\otimes 2}\right\}s^{(0)}(\theta_0,s)\lambda_0(s)ds=\Sigma.$$

Hence, $n^{-1}H(\theta^*, 1) \xrightarrow{P} \Sigma$. The proof of Theorem 2.6.4 is complete.

Chapter 3

Simulation Study

To evaluate the iterative Newton-Raphson algorithm and to assess the finite sample performance of the estimators for the proposed model, we conducted two simulation studies which both employed a constant baseline.

The observed data are $Z = \min\{T, C\}$, where T_i and C_i are chosen to ensure a specified percentage of censored observations, but in a way which preserves the conditional independence between T_i and C_i given X_i and V_i. In order to compare our partially linear single-index model with the standard Cox linear model, two simulation studies were conducted:

1. A sine curve was adopted for the true nonparametric function ψ:

$$\psi(\beta_0^T X) = 5\sin\{\beta_0^T X/2\}.$$

2. A linear line was adopted for the true nonparametric function ψ:

$$\psi(\beta_0^T X) = \beta_0^T X.$$

We used the angle between the true direction and the estimated direction of the parameter vectors to measure the performance of the models. Specifically, the angle between β_0 and $\hat{\beta}$ is

$$\omega(\beta_0, \hat{\beta}) = \arccos\left(\frac{\langle \beta_0, \hat{\beta}\rangle}{\|\beta_0\| \cdot \|\hat{\beta}\|}\right),$$

and the angle between α_0 and $\hat{\alpha}$ is

$$\omega(\alpha_0, \hat{\alpha}) = \arccos\left(\frac{\langle \alpha_0, \hat{\alpha}\rangle}{\|\alpha_0\| \cdot \|\hat{\alpha}\|}\right),$$

here $\langle a, b\rangle$ denotes the inner product of a and b.

3.1 Experiment I (Sine curve model)

In this example, the true log-relative risk function is

$$\alpha_0^T V + \psi(\beta_0^T X) = \alpha_0^T V + 5\sin(\beta_0^T X/2),$$

where X is comprised of five covariates, $X_i \sim N(1,4)$, $i = 1,2,3$; $X_i \sim Uniform(-2, 2)$, $i = 4,5$; V is comprised of three covariates, $V_1 \sim Uniform(-2,2)$, $V_i \sim Binomial(n = 1, p = 0.5)$, $i = 2,3$; $\beta_0 = (1,-1,1,-1,1)^T/\sqrt{5}$ and $\alpha_0 = (0.5,-1,2)^T$.

The distribution of the censoring variable, C, is exponential with hazard function

$$\lambda_c(t|X = x, V = v) = \exp(\mu + x^T\rho_1 + v^T\rho_2).$$

Here $\rho_1 = (0.5,-0.5,0.5,-0.5,0.5)^T$, $\rho_2 = (1,-1,1)^T$, and μ is a constant, taking values -2 and 0, so that censoring rate cr is about 12% and 37%, respectively. The sample size, n, was chosen to be 200 and 300, respectively. The number of simulations was 500.

For this simulation experiment, using both the AIC and BIC criteria, we found that the maximum partial likelihood estimates are not very sensitive to the number of knots used. However, the more knots we used, the longer the program needed to run until the algorithm converged. So we decided to use splines with four equally spaced knots to approximate the $\psi(\cdot)$ in our estimation procedure.

We compared the performance of the proposed method with ψ unknown with that of the standard Cox model and with that of the model with ψ known, respectively.

Table 3.1: Sine curve model: summary statistics for angles between β_0 and $\hat{\beta}$. The regression estimates are obtained by using maximum partial likelihood method based on the identity link, the true known link, and the proposed semi-unknown link.

	cr = 12%; n=300		cr = 37%; n=300	
$\psi(\cdot)$	mean	StDev	mean	StDev
Identity	5.510	2.283	5.251	2.159
Known	2.314	0.944	2.748	1.094
Unknown	2.443	0.925	2.870	1.102

Table 3.2: Sine curve model: summary statistics for angles between α_0 and $\hat{\alpha}$. The regression estimates are obtained by using maximum partial likelihood method based on the identity link, the true known link, and the proposed semi-unknown link.

	cr=12%; n=300		cr=37%; n=300	
$\psi(\cdot)$	mean	StDev	mean	StDev
Identity	6.141	3.755	5.825	3.508
Known	3.079	1.800	3.683	2.173
Unknown	3.235	1.907	3.734	2.264

We note that the true functional form is generally not known in real applications. Here it is used as a benchmark to gauge the performance of our proposed estimators.

Summary statistics for the angles between the true and the estimated directions are reported in Table 3.1 and Table 3.2, for β and α, respectively, when the sample size is 300. These results indicate that when the relative risk form is misspecified, the estimate based on the wrong form, in other words, the Cox identity link in this case, will give a substantially biased estimate with a large standard deviation. On the other hand, the proposed estimate of β with the unknown relative risk form is close to the estimate based on the known true link functional form.

Results for assessing the accuracy of the standard error formula are given in

Table 3.3: Sine curve model: results for β in a simulation study using the proposed method. Avg.: sample average; StDev: sample standard deviation; 95% cov. prob.: empirical coverage probability of the 95% confidence interval. Based on 500 Monte Carlo simulations.

n=200, cr=12%	β_1	β_2	β_3	β_4	β_5
True β	0.4472	-0.4472	0.4472	-0.4472	0.4472
Avg.($\hat{\beta}$)	0.4474	-0.4474	0.4461	-0.4474	0.4477
Bias	0.0002	-0.0002	-0.0011	-0.0002	0.0005
StDev($\hat{\beta}$)	0.0202	0.0210	0.0211	0.0306	0.0293
Avg.{SE($\hat{\beta}$)}	0.0189	0.0202	0.0190	0.0286	0.0286
StDev{SE($\hat{\beta}$)}	0.0019	0.0023	0.0020	0.0027	0.0029
95% cov. prob.	0.9315	0.9435	0.9254	0.9254	0.9476
n=200, cr=37%	β_1	β_2	β_3	β_4	β_5
True β	0.4472	-0.4472	0.4472	-0.4472	0.4472
Avg.($\hat{\beta}$)	0.4482	-0.4468	0.4471	-0.4473	0.4466
Bias	0.0010	0.0004	-0.0001	-0.0001	-0.0006
StDev($\hat{\beta}$)	0.0239	0.0249	0.0247	0.0371	0.0365
Avg.{SE($\hat{\beta}$)}	0.0221	0.0234	0.0222	0.0334	0.0335
StDev{SE($\hat{\beta}$)}	0.0025	0.0029	0.0026	0.0038	0.0036
95% cov. prob.	0.9360	0.9360	0.9200	0.9300	0.9240
n=300, cr=12%	β_1	β_2	β_3	β_4	β_5
True β	0.4472	-0.4472	0.4472	-0.4472	0.4472
Avg.($\hat{\beta}$)	0.4467	-0.4484	0.4456	-0.4475	0.4479
Bias	-0.0005	-0.0012	-0.0016	-0.0003	0.0007
StDev($\hat{\beta}$)	0.0175	0.0164	0.0164	0.0247	0.0249
Avg.{SE($\hat{\beta}$)}	0.0154	0.0162	0.0153	0.0232	0.0231
StDev{SE($\hat{\beta}$)}	0.0014	0.0015	0.0013	0.0019	0.0019
95% cov. prob.	0.9240	0.9400	0.9200	0.9400	0.9400
n=300, cr=37%	β_1	β_2	β_3	β_4	β_5
True β	0.4472	-0.4472	0.4472	-0.4472	0.4472
Avg.($\hat{\beta}$)	0.4470	-0.4485	0.4458	-0.4469	0.4478
Bias	-0.0002	-0.0013	-0.0014	0.0004	0.0006
StDev($\hat{\beta}$)	0.0202	0.0193	0.0196	0.0296	0.0290
Avg.{SE($\hat{\beta}$)}	0.0179	0.0187	0.0179	0.0270	0.0269
StDev{SE($\hat{\beta}$)}	0.0018	0.0018	0.0017	0.0024	0.0025
95% cov. prob.	0.9060	0.9400	0.9200	0.9320	0.9300

Table 3.4: Sine curve model: results for α in a simulation study using the proposed method. Avg.: sample average; StDev: sample standard deviation; 95% cov. prob.: empirical coverage probability of the 95% confidence interval. Based on 500 Monte Carlo simulations.

n=200, cr=12%	α_1	α_2	α_3
True α	0.500	-1.000	2.000
Avg.($\hat{\alpha}$)	0.519	-1.046	2.074
Bias	0.019	-0.046	0.074
StDev($\hat{\alpha}$)	0.083	0.179	0.217
Avg.{SE($\hat{\alpha}$)}	0.080	0.177	0.207
StDev {SE($\hat{\alpha}$)}	0.004	0.006	0.010
95% cov. prob.	0.940	0.944	0.936
n=200. cr=37%	α_1	α_2	α_3
True α	0.500	-1.000	2.000
Avg.($\hat{\alpha}$)	0.527	-1.049	2.081
Bias	0.027	-0.049	0.081
StDev($\hat{\alpha}$)	0.103	0.214	0.254
Avg.{SE($\hat{\alpha}$)}	0.098	0.209	0.240
StDev {SE($\hat{\alpha}$)}	0.006	0.010	0.014
95% cov. prob.	0.938	0.948	0.936
n=300, cr=12%	α_1	α_2	α_3
True α	0.500	-1.000	2.000
Avg.($\hat{\alpha}$)	0.516	-1.023	2.055
Bias	0.016	-0.023	0.055
StDev($\hat{\alpha}$)	0.066	0.148	0.166
Avg.{SE($\hat{\alpha}$)}	0.064	0.140	0.164
StDev {SE($\hat{\alpha}$)}	0.003	0.004	0.006
95% cov. prob.	0.948	0.942	0.934
n=300, cr=37%	α_1	α_2	α_3
True α	0.500	-1.000	2.000
Avg.($\hat{\alpha}$)	0.518	-1.021	2.064
Bias	0.018	-0.021	0.064
StDev($\hat{\alpha}$)	0.077	0.174	0.198
Avg.{SE($\hat{\alpha}$)}	0.078	0.166	0.191
StDev {SE($\hat{\alpha}$)}	0.004	0.006	0.009
95% cov. prob.	0.956	0.938	0.934

Table 3.3 and Table 3.4, for β and α, when $n = 300$ or 200, and the censoring rate is approximately 12% or 37%, respectively. These results are based on 500 Monte Carlo simulation runs per setting. We find that the proposed method works well: the bias between the true parameter and the estimated parameter are really small, less than 0.2% of the true value; the estimated standard deviations of the estimates are reasonably close to the Monte Carlo standard deviations of the estimates; and the Monte Carlo coverage probabilities of the 95% confidence intervals are reasonably close to the nominal level.

Also, from Table 3.3 and Table 3.4, we can see that for a fixed censoring rate, the standard errors and biases of the estimates are decreasing, and the coverage probability is closer to the nominal level 95%, when the sample size is increasing. For a fixed sample size, the standard errors and biases of the estimates are increasing, and the coverage probability is getting further from the nominal level 95%, when the censoring rate is increasing.

Figure 3.1 and Figure 3.2 show that the estimated link function and 95% confidence interval based on one random simulation with a sample size 300 and a censoring rate 37% and a sample size 300 and a censoring rate 12%, respectively. Moreover, the fitted function (the dotted curve) in Figure 3.1 captures the true function (the solid curve) closely, indicating the proposed method works quite well even the censoring rate is relatively high.

Figure 3.3 shows that the estimated link function and 95% confidence interval based on one random simulation with a sample size 200 and a censoring rate 12%. The fitted function (the dotted curve) captures the true function (the solid curve) closely. Also, compared to Figure 3.2, which is based on a single simulation with the

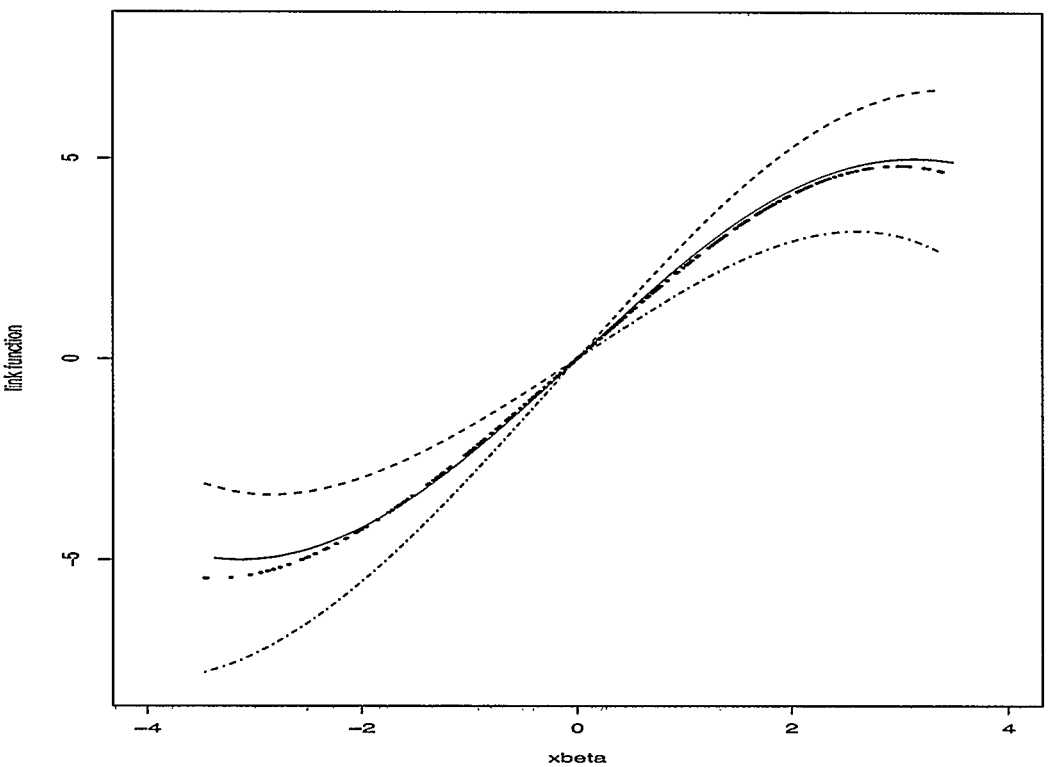

Figure 3.1: A single random simulation for Sine curve model with its corresponding 95% confidence interval using a sample size 300 and a censoring rate of 37%. The solid curve is the true function, the dotted curve is the estimated function, the dashed line is the upper limit of the 95% confidence interval while the dot-dashed line is the lower limit of the 95% confidence interval.

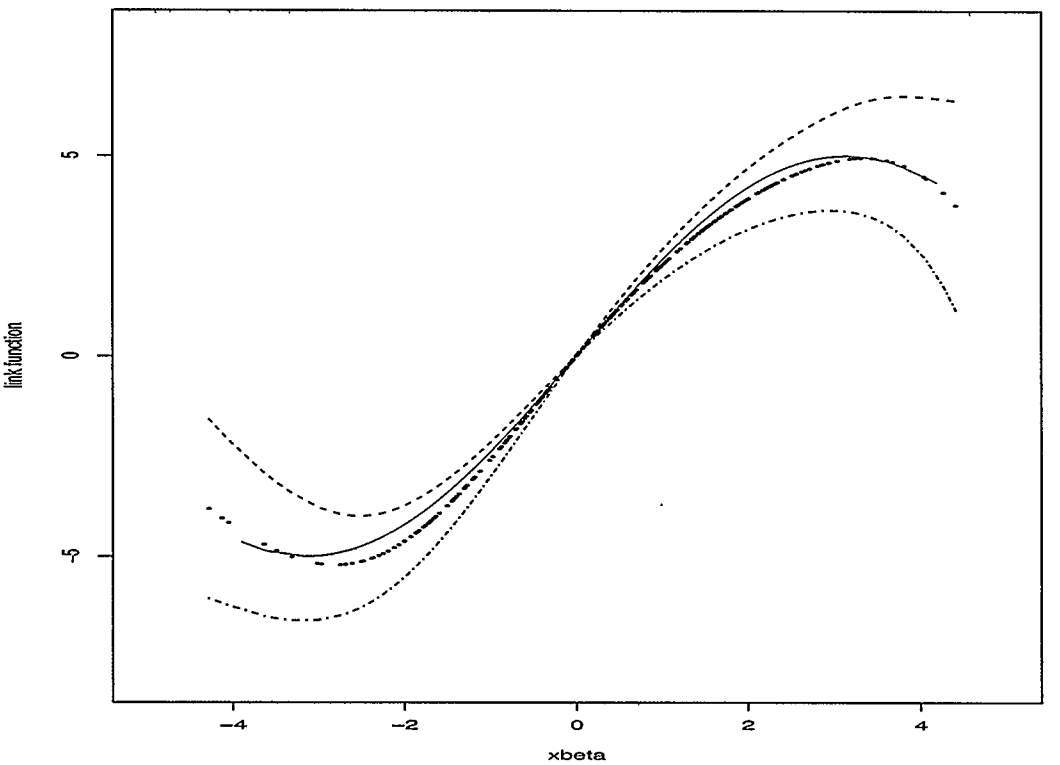

Figure 3.2: A single random simulation for Sine curve model with its corresponding 95% confidence interval using a sample size 300 and a censoring rate of 12%. The solid curve is the true function, the dotted curve is the estimated function, the dashed line is the upper limit of the 95% confidence interval while the dot-dashed line is the lower limit of the 95% confidence interval.

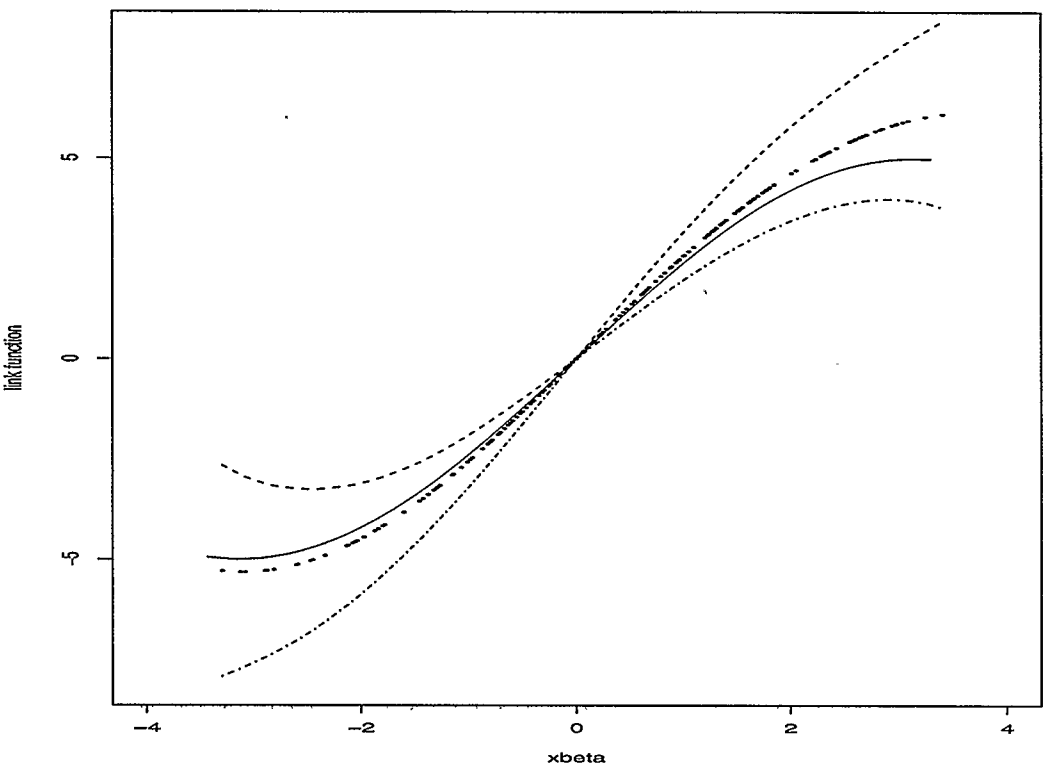

Figure 3.3: A single random simulation for Sine curve model with its corresponding 95% confidence interval using a sample size 200 and a censoring rate of 12%. The solid curve is the true function, the dotted curve is the estimated function, the dashed line is the upper limit of the 95% confidence interval while the dot-dashed line is the lower limit of the 95% confidence interval.

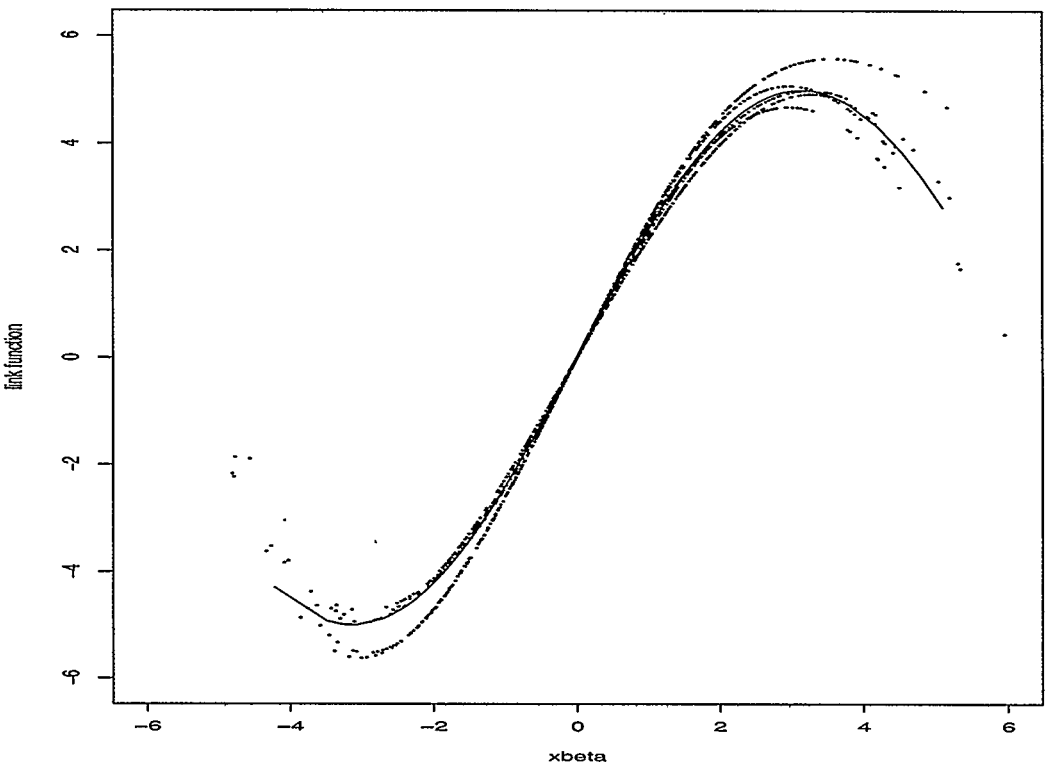

Figure 3.4: 5 random simulations for Sine curve model at sample size 300 and censoring rate 12%. The solid curve is the true function, while other five dotted curves are the estimated functions of five random simulations.

same censoring rate 12% but a sample size 300, the estimated curve in Figure 3.3 is not as close to the true curve as the one in Figure 3.2 with a larger sample size 300. In other words, for a fixed censoring rate, the performance of the proposed model is getting better, when the sample size is increasing.

Figure 3.4 shows that the fitted function based on five random simulations, from which we can tell the estimated function is generally very close to the true function, even just based on a single simulation.

3.2 Experiment II (linear model)

In this example, the true log-relative risk function used is

$$\alpha_0^T V + \psi(\beta_0^T X) = \alpha_0^T V + \beta_0^T X,$$

in other words, it is linear. We want to evaluate, when the underlying log relative risk function is linear, how much efficiency is lost if our method is used instead of the standard Cox model. We generated covariates V and X, censoring variable C, as in the previous simulation experiment. The number of simulation runs is still 500. Based on both the AIC and BIC criteria, we again found that the estimates were not very sensitive to the number of knots, but the more knots we used, the longer the program took to run. So once again we used splines with four equally spaced knots to approximate the $\psi(\cdot)$ in our estimation procedure.

We compared the performance of the proposed method with the standard Cox model to estimate the linear relative risk function, which is, indeed, the link function of the standard Cox model.

Summary statistics for the angles between the true and the estimated directions are reported in Table 3.5 and Table 3.6, for β and α, respectively. When the censoring rate is low, the results indicate that our proposed method works almost equally as good as the Cox linear model, which is reflected by the proximity of the estimates obtained from these two models. Hence, efficiency loss is negligible in our example. When the censoring rate is high, the estimates of β are worse than that of α.

Results for comparing the standard error and the 95% coverage probability of the standard Cox model with those of our proposed model are given in Table 3.7 and Table 3.8, for β or α, respectively. The sample size is 300 or 200, and the censoring rate is about 10% or 40%, respectively. They are all based on 500 Monte Carlo simulation runs. We find that: the estimated standard errors of $\hat{\beta}$ of the proposed model are generally even smaller than those of the Cox model, the standard errors of $\hat{\alpha}$ of the proposed model are slightly larger than those of the Cox model, and the Monte Carlo coverage probabilities of the 95% confidence intervals in the Cox model are reasonably closer to the nominal level than those in the proposed model. Since the differences between these two models are relatively small, we conclude that the resultant efficiency loss by using the proposed method is negligible if the underlying model is in fact the standard Cox linear PH model.

Table 3.5: Linear model: summary statistics for angles between β_0 and $\hat{\beta}$. The regression estimates are obtained by using maximum partial likelihood method based on the identity link, and the proposed semi-unknown link. ARE: Asymptotic Relative Efficiency.

	cr=10%; n=300		cr=40%; n=300		cr=10%; n=200		cr=40%; n=200	
$\psi(\cdot)$	mean	StDev	mean	StDev	mean	StDev	mean	StDev
Identity	4.702	1.804	5.672	2.313	5.688	2.358	6.976	2.891
Unknown	4.873	1.862	6.119	3.446	5.979	2.476	8.272	9.309
ARE		93.9%		45.1%		90.7%		9.6%

Table 3.6: Linear model: summary statistics for angles between α_0 and $\hat{\alpha}$. The regression estimates are obtained by using maximum partial likelihood method based on the identity link, and the proposed semi-unknown link. ARE: Asymptotic Relative Efficiency.

	cr=10%; n=300		cr=40%; n=300		cr=10%; n=200		cr=40%; n=200	
$\psi(\cdot)$	mean	StDev	mean	StDev	mean	StDev	mean	StDev
Identity	3.148	1.871	3.798	2.274	3.852	2.329	4.763	2.906
Unknown	3.190	1.877	3.836	2.329	3.942	2.348	4.975	3.018
ARE		99.4%		95.3%		98.4%		92.7%

Table 3.7: Linear model: results for β in a simulation study using the proposed method. Avg.: sample average; 95% cov. prob.: empirical coverage probability of the 95% confidence interval. Based on 500 Monte Carlo simulations.

n=200, cr=10%	β_1	β_2	β_3	β_4	β_5
Avg.{SE($\hat{\beta}_{Cox}$)}	0.0491	0.0493	0.0493	0.0746	0.0747
Avg.{SE($\hat{\beta}_{new}$)}	0.0384	0.0383	0.0384	0.0575	0.0573
95% cov. prob. (Cox)	0.9420	0.9400	0.9320	0.9440	0.9260
95% cov. prob. (new)	0.9420	0.9200	0.9340	0.9260	0.9360
n=200, cr=40%	β_1	β_2	β_3	β_4	β_5
Avg.{SE($\hat{\beta}_{Cox}$)}	0.0616	0.0617	0.0614	0.0928	0.0928
Avg.{SE($\hat{\beta}_{new}$)}	0.0472	0.0477	0.0487	0.0713	0.0717
95% cov. prob. (Cox)	0.9260	0.9300	0.9580	0.9440	0.9340
95% cov. prob. (new)	0.9360	0.9020	0.9080	0.9200	0.9380
n=300, cr=10%	β_1	β_2	β_3	β_4	β_5
Avg.{SE($\hat{\beta}_{Cox}$)}	0.0394	0.0394	0.0393	0.0598	0.0597
Avg.{SE($\hat{\beta}_{new}$)}	0.0311	0.0309	0.0311	0.0464	0.0463
95% cov. prob. (Cox)	0.9360	0.9480	0.9560	0.9480	0.9460
95% cov. prob. (new)	0.9200	0.9220	0.9420	0.9340	0.9200
n=300, cr=40%	β_1	β_2	β_3	β_4	β_5
Avg.{SE($\hat{\beta}_{Cox}$)}	0.0489	0.0487	0.0490	0.0739	0.0738
Avg.{SE($\hat{\beta}_{new}$)}	0.0385	0.0389	0.0393	0.0576	0.0578
95% cov. prob. (Cox)	0.9400	0.9520	0.9520	0.9480	0.9340
95% cov. prob. (new)	0.9197	0.9197	0.9438	0.9217	0.9277

Table 3.8: Linear model: results for α in a simulation study using the proposed method. Avg.: sample average; 95% cov. prob.: empirical coverage probability of the 95% confidence interval. Based on 500 Monte Carlo simulations.

n=200, cr=10%	α_1	α_2	α_3
Avg.{SE($\hat{\alpha}_{Cox}$)}	0.077	0.171	0.200
Avg.{SE($\hat{\alpha}_{new}$)}	0.079	0.174	0.204
95% cov. prob. (Cox)	0.948	0.948	0.926
95% cov. prob. (new)	0.944	0.932	0.948
n=200. cr=40%	α_1	α_2	α_3
Avg.{SE($\hat{\alpha}_{Cox}$)}	0.102	0.211	0.239
Avg.{SE($\hat{\alpha}_{new}$)}	0.104	0.216	0.245
95% cov. prob. (Cox)	0.926	0.948	0.928
95% cov. prob. (new)	0.938	0.928	0.926
n=300, cr=10%	α_1	α_2	α_3
Avg.{SE($\hat{\alpha}_{Cox}$)}	0.062	0.137	0.160
Avg.{SE($\hat{\alpha}_{new}$)}	0.063	0.138	0.162
95% cov. prob. (Cox)	0.950	0.956	0.942
95% cov. prob. (new)	0.928	0.940	0.946
n=300, cr=40%	α_1	α_2	α_3
Avg.{SE($\hat{\alpha}_{Cox}$)}	0.081	0.169	0.191
Avg.{SE($\hat{\alpha}_{new}$)}	0.082	0.171	0.194
95% cov. prob. (Cox)	0.944	0.960	0.952
95% cov. prob. (new)	0.952	0.936	0.946

Chapter 4

Case Study – Veteran's Administration Lung Cancer Data

4.1 Description of the Data

In this chapter we present results from applying our partially linear single-index model to a dataset from the Veteran's Administration Lung Cancer Study Clinical Trial (Kalbfleisch and Prentice, 2002).

The Veteran's Administration lung cancer data were used by Kalbfleisch and Prentice (2002) to illustrate the Cox PH model. In this clinical trial, males with advanced inoperable lung cancer were randomized to either a standard or test chemotherapy. The primary end point for therapy comparison was time to death (variable 'Status' is 1 for dead and 0 for censored). Only nine out of the 137 survival times were censored. The censoring rate here is 6.6%. The data set includes six covariates: treatment, age at diagnosis, Karnofsky score, diagnosis time, cell type and prior therapy. Descriptions are as follows:

1. Treatment: the standard (0) or test (1) chemotherapy given to the patient,
2. Age: age in years,
3. Karnofsky score: a subjective measure of how well the patient is doing reported by experienced nurses. It is useful to track it over time, to see the ups and downs in the disease process, as defined below:

99: Normal, no complaints or evidence of disease

90: Able to perform normal activity; minor signs and symptoms of disease

80: Able to perform normal activity with effort; some symptoms of disease

70: Cares for self, unable to perform normal activity or to do active work

60: Requires occasional assistance but is able to care for most of own needs

50: Requires considerable assistance and frequent medical care

40: Requires special care and assistance; disabled

30: Hospitalization indicated, although death not imminent; severely disabled

20: Hospitalization necessary; active supportive treatment required, very sick

10: Fatal processes progressing rapidly; moribund

00: Dead

4. Diagnosis time: months from Diagnosis,
5. Cell type: 1=squamous, 2=small cell, 3=adenocarcinoma, 4=large,
6. Prior therapy: 0=no, 1=yes.

4.2 Comparison of the Partially Linear Single-Index Model and the Cox Model using the VA Lung Cancer Data

Huang and Liu (2006) used this same dataset to fit a single-index model in which they specified the hazard function as

$$\lambda(t|x) = \lambda_0(t) \exp\left\{\psi(\beta_0^T x)\right\},$$

where $\psi(\cdot)$, referred to as the link function, is an unknown smooth function, and all of the covariates were assigned to X. In their analysis, suppose $x = (x_1, x_2^T)^T$, where

$x_1 = 0$ or 1 is a treatment indicator and x_2 represents the other covariates. Then the treatment effect in terms of the log hazard at x_2 is

$$\log \lambda(t|1, x_2) - \log \lambda(t|0, x_2) = \psi(\beta_{01} + \beta_{02}^T x_2) - \psi(\beta_{02}^T x_2),$$

which does not depend on t, but in general depends on the value of x_2.

However, Figure 3 in their paper indicates that the log-hazard ratio for treatment effect may not depend on the value of x_2 either for this dataset. Thus, it might be better to model the treatment effect parametrically and other covariates nonparametrically. Therefore, we applied our proposed method to fit a partially linear single-index model with the 'treatment' covariate assigned to the parametric part V, and other covariates assigned to the nonparametric part X. Then the treatment effect in terms of log hazard at $X = x$ is

$$\log \lambda(t|1, x) - \log \lambda(t|0, x) = \alpha \cdot 1 + \psi(\beta^T x) - \{\alpha \cdot 0 + \psi(\beta^T x)\} = \alpha,$$

which is a constant.

On the other hand, the primary purpose of this clinical trial was to investigate whether the test chemotherapy works better or worse than the standard chemotherapy, so putting variable 'treatment' into V makes the treatment effect easy to interpret.

We also included 'Karnofsky score', 'Diagnosis time' and the other covariates, which we think should be related to the risk of failure, into vector V, along with treatment. However, none of them were statistically significant. There is a possibility that they do have nonnegligible effects after being combined with each other in a specific way, but none of them has a simple linear relationship with the risk directly.

So we decided to still keep them, but include them in the nonparametric covariate part, X.

The unknown function $\psi(\cdot)$ of the logarithm of the relative risk form is fitted as a spline function with 4 knots, which was the number of knots which minimized both the AIC and BIC criteria. The estimated parameters and corresponding standard errors are presented in Table 4.1. To compare the results of the standard Cox model, the single-index model of Huang and Liu (2006) (referred as 'HL' in what follows) and our proposed model, we rescaled the estimates of the model with unknown ψ, such that the coefficient vector has the same norm as that for the Cox model with identity ψ. These results are shown in the last column of Table 4.1. The results from Huang and Liu (2006) are presented in the second column.

Table 4.1: The VA lung cancer data regression. The parameter estimates and corresponding standard errors (in the parentheses) for the standard Cox model, the single-index model of Huang and Liu (2006), and the proposed model and the proposed model after rescaling.

	Identity ψ (Cox)	HL ψ (rescaled)	Unknown ψ (proposed)	Unknown ψ (rescaled)
Treatment	0.290 (0.207)	0.592 (0.140)	0.485 (0.203)	0.504 (0.211)
Age	-0.009 (0.009)	-0.021 (0.007)	-0.012 (0.001)	-0.012 (0.001)
Karnofsky score	-0.033 (0.006)	-0.050 (0.009)	0.000 (0.001)	0.000 (0.001)
Diagnosis time	-0.000 (0.009)	0.013 (0.003)	-0.004 (0.002)	-0.004 (0.002)
Cell Type				
Squamous vs large	-0.400 (0.283)	-0.742 (0.121)	0.808 (0.006)	0.840 (0.006)
Small vs large	0.457 (0.266)	-0.387 (0.124)	0.469 (0.016)	0.488 (0.017)
Adeno vs large	0.789 (0.303)	-0.145 (0.191)	0.356 (0.025)	0.370 (0.026)
Prior therapy	0.072 (0.230)	-0.011 (0.011)	0.039 (0.026)	0.041 (0.027)

The results of the standard Cox model, which agree very closely with those re-

ported in Kalbfleisch and Prentice (2002, p.120), are quite different from those of the proposed unknown model. The angle between the vectors of regression coefficients of X from the two models is 79.91°. By comparing the first and fourth column of Table 4.1, we also find the standard errors of the estimates for the proposed model are generally smaller than those for the Cox model, except for the standard errors for the Treatment variable, which are still reasonably close to each other. All of these results indicate that the Cox model may give biased estimates.

Table 4.2 shows the p-values obtained from the standard Cox model, the single-index HL model and the proposed partially linear single-index model. If we use 5% as the default significance level, we can see that Treatment, Age, Diagnosis time, Squamous vs Large and Small vs Large are all statistically significant based on the proposed model while the standard Cox model suggests they are not. In contrast to these results, the Karnofsky score is not statistically significant based on the proposed model while the standard Cox PH model suggests it is. This table agrees with the conclusions we presented in Table 4.1.

In our partially linear single index model with link function $\alpha^T V + \psi(\beta^T X)$, the V vector only includes one covariate 'Treatment', so we can assess the treatment effect based on the sign of the coefficient α. From Table 4.1, we see that $\hat{\alpha}$ is 0.485, which is a positive value. This indicates that the test treatment (Treatment = 1) increases the log-hazard by 0.485, compared to the standard chemotherapy (Treatment = 0). The exponential of the link function is increased by a factor $\exp(\hat{\alpha}) = 1.624 > 1$, or in other words, the risk of death will increase on the new chemotherapy treatment. Hence, we can conclude that overall the test treatment is worse than the standard treatment for the population sampled from, which is in agreement with the results

Table 4.2: The VA lung cancer data. The p-value of the estmates for the standard Cox model, the single-index model and the proposed model.

P-value	Identity ψ (Cox)	Unknown ψ (HL)	Unknown ψ (proposed)
Treatment	0.160	0.000*	0.019*
Age	0.360	0.000*	0.000*
Karnofsky score	0.000*	0.000*	0.783
Diagnosis time	0.990	0.000*	0.020*
Cell Type			
Squamous vs large	0.160	0.000*	0.000*
Small vs large	0.086	0.000*	0.000*
Adeno vs large	0.009*	0.437	0.000*
Prior therapy	0.760	0.271	0.224

* p-value < 0.05

obtained by Huang and Liu (2006) from their single-index model.

Chapter 5

Discussion and Future Work

In the preceding chapters, the use of a partially linear single-index model was proposed as a useful tool to model covariates which can have possibly linear and nonlinear effects on the log hazard in the proportional hazards model. This approach can reduce the dimensionality of the covariates and obtain efficient estimates of the covariates' effects at the same time.

As in every model which uses B-spline, the question of knot selection is raised. Uniform B-splines are a special case which use a smoothing function for the nonparametric single-index component. In our simulation studies and real example, uniform B-splines with equally spaced knots worked well. In practice, fixed knots are a simple and convenient solution for small to medium-sized datasets, but in general it is also desirable to have a data-dependent method for optimal knot selection and placement. Different adaptive knot selection methods and a bootstrapping method have been investigated by several researchers. As we discussed in Section 2.5, the AIC and BIC criteria can be used to decide the number of knots needed. In practice, we should also consider time efficiency to implement this method, because the more knots used in the spline, the longer the program will take to run. In order to take all of these aspects into consideration, developing a spline knot selection method is thus part of future work I will undertake.

The proposed partially linear single-index model can be widely used for several kinds of datasets in addition to clinical trial survival data. In mathematical finance

applications, our proposed model can be used to quantify what factors influence the duration of the life time of a hedge fund, which exhibits time-to-event features (Baba and Goko, 2006). The idea of treating hedge funds as survival data is still new and rarely investigated. In the future, we will explore the suitability and the performance of our model applied to other datasets besides clinical trial survival data. It is also possible to extend our proposed modelling techniques for survival data to handle other types of censored data such as doubly censored data or interval censored data, which is part of our future work as well.

Bibliography

[1] O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6:701–726, 1978.

[2] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory. Budapest*, pages 267–281, 1973.

[3] P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10:1100–1120, 1982.

[4] N. Baba and H. Goko. Survival analysis of hedge funds. *Bank of Japan Working Paper Series*, (No. 06.-E-05), 2006.

[5] R. Bellman and B. Gluss. On various versions of the defective coin problem. *Information and Control*, 4:118–131, 1961.

[6] R.E. Bellman. *Adaptive Control Processes.* Princeton University Press, Princeton, NJ., 1961.

[7] R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92:477–489, 1997.

[8] D. R. Cox. Regression model and life-table (with discussion). *Journal of the Royal Statistical Society*, Series B 34:187–220, 1972.

[9] D. R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.

[10] C. de Boor. *A Practical Guide to Splines.* New York: Springer, 1978.

[11] J. Fan, I. Gijbels, and M. King. Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, 25:1661–1690, 1997.

[12] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis.* New York: Wiley, 1991.

[13] R. Gentleman and J. Crowley. Local full likelihood estimation for the proportional hazards model. *Biometrics*, 47:1283–1296, 1991.

[14] R. R. Gill. Understanding Cox's regression model: A Martingale approach. *Journal of the American Statistical Association*, 79:441–448, 1984.

[15] T. G∅rgens. Average derivatives for hazard functions. *Econometric Theory*, 20:437–463, 2004.

[16] R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87:942–951, 1992.

[17] C. Gu. Penalized likelihood hazard estimation: A general procedure. *Statistica Sinica*, 6:861–876, 1996.

[18] W. Härdle, P. Hall, and H. Ichimura. Optimal smoothing in single-index models. *The Annals of Statistics*, 21:157–178, 1993.

[19] W. Härdle and T. M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*, 84:986–995, 1989.

[20] T. Hastie and R. Tibshirani. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46:1005–1016, 1990.

[21] G. Heller. The Cox proportional hazards model with a partly linear relative risk function. *Lifetime data analysis*, 7:255–277, 2001.

[22] J. Huang. Efficient estimation of the partly linear additive Cox model. *The Annals of Statistics*, 27:1536–1563, 1999.

[23] J. Z. Huang, C. Kooperberg, C. J. Stone, and Y. K. Truong. Functional ANOVA modeling for proportional hazards regression. *The Annals of Statistics*, 28:960–999, 2000.

[24] J. Z. Huang and L. Liu. Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics*, 62:793–802, 2006.

[25] H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58:71–120, 1993.

[26] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data, Second Edition.* Wiley-Interscience, New York, 2002.

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

[28] J. F. Lawless. *Statistical Models and Methods for Lifetime Data, Second Edition.* John Wiley & Sons Inc., Hoboken, New Jersey, 2003.

[29] O.B. Linton and J.P. Nielsen. Kernel estimation in a nonparametric marker dependent hazard model. *The Annals of Statistics*, 23:1735–1748, 1995.

[30] O.B. Linton, J.P. Nielsen, and P. Bickel. On a semiparametric survival model with flexible covariate effect. *The Annals of Statistics*, 26:215–241, 1998.

[31] L. Liu. *Semiparametric and nonparametric models for survival data.* PhD thesis, University of Pennsylvania, Pennsylvania,United States, 2004.

[32] P. Y. Liu and J. Crowley. Large sample theory of the MLE based on Cox's regression model for survival data. Technical report, 1978.

[33] X. Lu, G. Chen, X.-K. Song, and R. S. Singh. A class of partially linear single-index survival models. *Canadian Journal of Statistics*, 34:97–112, 2006.

[34] X. Lu and T.-L. Cheng. Randomly censored partially linear single-index models. *Journal of Multivariate Anal.*, 2007 in press.

[35] D. Oakes. Survival Times: Aspects of partial likelihood. *Intl. Stat. Review*, 49:235–264, 1981.

[36] F. O'Sullivan. Nonparametric estimation in the Cox model. *The Annals of Statistics*, 21:124–145, 1993.

[37] J. L. Powell, J. H. Stock, and T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57:1403–1430, 1989.

[38] R. L. Prentice and S. G. Self. Asymptotic distribution theory for Cox type regression models with general relative risk form. *The Annals of Statistics*, 11:804–813, 1983.

[39] D. Ruppert. *Discussion of Maximization by Parts in Likelihood Inference (manual)*. Cornell University.

[40] L. L. Schumaker. *Spline Functions: Basic Theory*. Wiley-Interscience, New York, 1981.

[41] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[42] L. A. Sleeper and D. P. Harrington. Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, 85:941–949, 1990.

[43] A. A. Tsiatis. A large sample study of the estimate for the integrated hazard function in Cox's regression model for survival data. *The Annals of Statistics*, 9:93–108, 1981.

[44] A. A. Tsiatis. A large sample study of the estimate for the integrated hazard function in Cox's regression model for survival data. *The Annals of Statistics*, 9:93–108, 1981.

[45] W. Wang. *Proportial hazards regression model with unknown link function and applications to longitudinal time-to-event data*. PhD thesis, University of California, Davis, California,United States, 2001.

[46] W. Wang. Proportial hazards regression model with unknown link function and time-dependent covariates. *Statistica Sinica*, 14:885–905, 2004.

[47] Y. Xia, H. Tong, W. K. Li, and L. Zhu. An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64:363–410, 2002.