Computational Statistics and Data Analysis 55 (2011) 389–400

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# Partially varying coefficient single index proportional hazards regression models[☆]

Jianbo Li[a,b], Riquan Zhang[a,c,*]

[a] Department of Statistics, East China Normal University, Shanghai, 200241, PR China

[b] Department of Statistics, The Chinese University of Hong Kong, Hong Kong

[c] Department of Mathematics, Shanxi Datong University, Datong, 037009, PR China

ARTICLE INFO

*Article history:*
Received 4 December 2009
Received in revised form 5 April 2010
Accepted 9 May 2010
Available online 24 May 2010

*Keywords:*
Partially varying coefficient single index proportional hazards models
Polynomial *B*-spline
Asymptotic normality
Consistency

ABSTRACT

In this paper, the partially varying coefficient single index proportional hazards regression models are discussed. All unknown functions are fitted by polynomial *B* splines. The index parameters and *B*-spline coefficients are estimated by the partial likelihood method and a two-step Newton–Raphson algorithm. Consistency and asymptotic normality of the estimators of all the parameters are derived. Through a simulation study and the VA data example, we illustrate that the proposed estimation procedure is accurate, rapid and stable.



## 1. Introduction

The proportional hazards (PH) regression model, proposed by Cox (1972, 1975), is one of the primary tools in biomedical studies involving survival times. Let $T$ be a lifetime variable, then its corresponding hazards function is given by

$$\lambda(t|X) = \lambda_0(t)e^{\beta^T X}, \tag{1}$$

where $X \in R^p$ is a covariate vector. The function $\lambda_0(t)$ is the conditional hazard function of $T$ when $X = \mathbf{0}$, called baseline function. $\lambda_0(t)$ is usually assumed to be unknown. Under model (1), the conditional failure rates associated with any two values of covariate $X$ are proportional. $\beta$ is regression coefficient vector, which is used to assess the dependence of the distribution of survival time $T$ on $X$. The extension of model (1) to time-dependent covariates are easily dealt with by the counting process and the martingale approaches (Andersen and Gill, 1982). The detailed parametric and nonparametric inferences can be found in the books by Cox and Oakes (1984), Andersen et al. (1993) and Fan and Gijbels (1996) etc.

The classical assumption of Cox PH model is that the covariates have linear effect on the log hazard function. However, in real applications, that assumption is not always met and may lead to wrong conclusions. To overcome this obstacle, many statisticians have extended the Cox PH model (1). Fan et al. (1997), Gentleman and Crowley (1991), Gu (1996), O'Sullivan (1993) and Tibshirani and Hastie (1987) discussed the nonparametric hazards models:

$$\lambda(t|X) = \lambda_0(t)e^{\psi(X)} \tag{2}$$

☆ The research was supported in part by National Natural Science Foundation of China (10871072) and Doctoral Fund of Ministry of Education of China (20090076110001).

* Corresponding author at: Department of Statistics, East China Normal University, Shanghai, 200241, PR China. Tel.: +86 138 3525 5996.
*E-mail addresses:* ljianb66@gmail.com (J. Li), zhangriquan@163.com (R. Zhang).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.05.007

where $\psi(\cdot)$ is an unspecified smooth function. However, the estimation of unstructured nonparametric function may suffer from the so-called "curse of dimensionality" (Bellman, 1961) and, thus, is not practically suitable when the dimension of the covariate vector $X$ is high. Structured instead of nonstructured nonparametric models were introduced by many authors. For example, Sleeper and Harrington (1990) modeled the nonlinear covariate effects in the Cox model by additive approach and polynomial splines; Gray (1992) considered additive and time-varying coefficient Cox regression models by use of penalized splines method; Using functional ANOVA decompositions, Huang et al. (2000) studied a general class of structured models including additive models as a specific case. Besides these, some authors have applied partially linear additive or single index models to hazard regression (Gorgens, 2004; Huang, 1999; Lu et al., 2006; Nielsen et al., 1998; Wang, 2004). Wang (2004) proposed a type of single index models with hazards function $\lambda(t|X) = \lambda_0(t)\psi(\beta^T X)$ and Huang and Liu (2006) considered another type of single index hazards model with conditional hazards function of the form

$$\lambda(t|X) = \lambda_0(t)e^{\psi(\beta^T X)}, \tag{3}$$

where $\psi(\cdot)$ is similar to the one in (2). They approximated this unknown link function by spline smoothing method. The inference procedures for the link function and index parameters were given. However, some of covariates may have linear effects on the log hazards. Therefore, Sun et al. (2008) discussed partially linear single index hazards models specified as

$$\lambda(t|X, V) = \lambda_0(t)e^{\psi(\beta^T X)+\alpha^T V} \tag{4}$$

with covariates $X \in R^p, V \in R^q$ and unknown link function $\psi(\cdot)$ as above. In their paper, they discussed the inference procedure and its properties for model (4). They also analyzed how to partition the covariates into the nonlinear component, $X$, and linear component, $V$, in details. The model (4) can be seen as the generalization of the model (1) and model (3). But the common assumption of all the three models (1), (3) and (4) is that the covariates have constant effects on the log hazard directly or through unknown link function indirectly. In practice, this may be invalidated. For example, in Veterian's Administration Lung Cancer study comparing a test therapy with a standard therapy for inoperable lung cancer, discussed in Section 3.2, the primary end point is the time to death. As we discussed, the efficiency of therapy is related with patient's age. That is, we cannot say that one therapy works absolutely better than another for all patients. Cai and Sun (2003), Marzec and Marzec (1997), Murphy and Sen (1991), Murphy (1993), Sasieni and Winnett (2003), Tian et al. (2005), Verweij and van Houwelingen (1995) and Zucker and Karr (1990) and others considered time-dependent coefficients (varying coefficient) hazards models:

$$\lambda(t|X) = \lambda_0(t)e^{\beta(t)^T X} \tag{5}$$

with unknown smooth function vector $\beta(t) = (\beta_1(t), \beta_2(t), \dots, \beta_p(t))^T$.

In this paper, motivated by the model (4) and (5), we proposed partially varying coefficient single index proportional hazards models as follows:

$$\lambda(t|X, V) = \lambda_0(t)e^{\psi(\beta^T X)+\alpha(U)^T V}, \tag{6}$$

where $\psi(\cdot)$ and $\lambda_0(t)$ is similar to ones in model (4) and (5); $X \in R^p$ is the nonlinear covariate vector; $V \in R^q$ is varying coefficient component covariate vector; $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T \in R^p$ is regression coefficient vector of $X$ with $\|\beta\| = 1$ and first element larger than 0 for identification without loss of generality; $\alpha(\cdot) = (\alpha_1(\cdot), \alpha_2(\cdot), \dots, \alpha_q(\cdot))^T \in R^q$ is the functional coefficient vector of $V$. The effect of covariate $V$ on the log hazard may vary with $U$. $U$ is a covariate, which may be one entry of $X, V, t$ itself or some other covariate. For instance, U can be the age of subjects or the experiment time. Furthermore, when U is one entry of $X$, the varying coefficient part can be seen as the interaction of $X$ and $V$.

The model (6) is flexible enough to cover a variety of situations. When $\psi(x) = x$ and $\alpha(\cdot) = \mathbf{0}$ or $\psi(\cdot) = 0$ and $\alpha(\cdot)$ is a constant vector, the model becomes model (1); When $\alpha(\cdot) = 0$, or equivalently, there are no effects of the predictors $V$ on $T$, (6) is nonparametric single index hazard model (3); When $\alpha(\cdot)$ is a constant function vector, the model (6) becomes model (4); When $\psi(\cdot) = 0$, the model (6) is reduced to model (5); If $p = 1$, the model will be partially varying coefficient additive hazard model (Huang, 1999). Hence, it is interesting to investigate the properties of model (6). Note that any constant in $\psi(\cdot)$ can be absorbed in $\lambda_0(t)$ and any scale of $\beta$ can be absorbed in $\psi(\cdot)$. So we impose $\psi(0) = 0$ for identification. The main focus of this paper is making inference for the parameters $\beta$ and functional coefficients $\alpha_i(\cdot)$'s under right random censoring.

The rest of paper is organized as follows. In Section 2, we describe our proposed model, including the estimation procedure, consistency, asymptotic normality, inference and implementation. A simulation study and VA data example are given in Section 3, which also serves the purpose of proposed inference procedure and computation algorithm. Section 4 gives some brief conclusions. All proofs and computation details are left in the Appendix.

## 2. The estimation procedure

### 2.1. B-spline estimation

In this section, we present the $B$-spline estimation procedure for unknown functions in model (6) under right censoring scheme. Denote $C$ as the censoring variable, $Z = \min(T, C)$ as the observed event times and $\delta = I(T \le C)$ as the censoring