Journal of Multivariate Analysis 151 (2016) 14–36

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Partially linear single-index proportional hazards model with current status data

Xuewen Lu [a,*], Pooneh Pordeli [a], Murray D. Burke [a], Peter X.-K. Song [b]

[a] Department of Mathematics and Statistics, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada

[b] Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, United States

ARTICLE INFO

Article history:
Received 26 July 2015
Available online 18 July 2016

AMS subject classifications:
62B10
62G20
62N01

Keywords:
B-splines
Counting process
Empirical process
Interval censored data
Monotonicity constraints
Semiparametric efficiency bound

ABSTRACT

A partially linear single-index proportional hazards model with current status data is introduced, where the cumulative hazard function is assumed to be nonparametric and a nonlinear link function is assumed to take a parametric spline function. Efficient estimation and effective algorithm are established. Polynomial spline smoothing is invoked for the estimation of the cumulative baseline hazard function with monotonicity constraint on the functional, while a simultaneous sieve maximum likelihood (SML) estimation is proposed to estimate regression parameters. The proposed SML estimator for the parameter vector is shown to be asymptotically normal and semiparametric efficient. The spline estimator of the functional of the cumulative hazard function is shown to achieve the optimal nonparametric rate of convergence. A simulation study is conducted to examine the finite sample performance of the proposed estimators and algorithm, and an analysis of renal function recovery data is presented.

1. Introduction

Let T be time to occurrence of a certain event (e.g., tumor onset), and let C be censoring time (e.g., a random examination time). Observations of T are called current status data (or case I interval-censored data) when the occurrence of an event is only recorded as it has happened before or after the censoring time C, instead of being observed exactly. Examples for current status data can be found in many biomedical applications, where each subject is observed only once and the only information observed is that the failure of interest has occurred before or after the examination time, so the failure time is either left- or right-censored instead of being recorded exactly.

Suppose $V = (V_1, \ldots, V_q)^\top$ is a q-dimensional covariate vector. The semiparametric proportional hazards (PH) model introduced by Cox [3] is widely used in the analysis of survival data, including the analysis of current status data. The Cox model takes the following form:

$$\lambda(t|V) = \lambda_0(t) \exp(\alpha^\top V), \tag{1}$$

where $\lambda_0(\cdot)$ is an unknown baseline hazard function and α is a q-dimensional regression coefficient vector. As the baseline hazard function is fully unspecified, the model provides great flexibility to address various baseline characteristics, for example, patient's baseline heterogeneity. A key advantage of the PH model is its capacity of assessing predictors for the survival distribution through the conditional hazard function $\lambda(\cdot)$. Such model and various extensions have been extensively

* Corresponding author.
E-mail address: lux@math.ucalgary.ca (X. Lu).

http://dx.doi.org/10.1016/j.jmva.2016.06.004
0047-259X/

discussed in the literature for right-censored data; see for example Andersen and Gill [2], Andersen et al. [1], Kalbfleisch and Prentice [12] and Fleming and Harrington [5] and more references therein. The maximum partial likelihood is the choice of inference approach with unspecified baseline hazard functions. However, the partial likelihood estimation method for right-censored data is not directly applicable to current status data nor to more general interval-censored data, because the unknown baseline hazard function cannot be profiled out in the estimation approach. As a result, the baseline hazard function has to be estimated together with the all other model parameters. Thus, estimation of the regression coefficients and derivation of its asymptotic properties have proven to be a real technical challenge.

Finkelstein [4] developed a method to estimate the parameters in model (1) when the data contain case II interval-censored observations. She assumed that the baseline survival function is discrete and has finitely many known mass points, which reduced the problem to a finite-dimensional parametric estimation problem. In the case of current status data (case I interval-censored data), Huang [9] studied the efficient maximum likelihood estimator for the PH model. He showed that the MLE for the regression parameter is asymptotically normal with $\sqrt{n}$ rate of convergence and achieves the semiparametric information bound, even though the MLE for the baseline cumulative hazard function only converges at $n^{1/3}$ rate. Estimation of the asymptotic variance–covariance matrix for the MLE of the regression parameter was also considered. McMahan et al. [20] proposed new EM algorithms to analyze current status data under both the PH and proportional odds (PO) models. They used monotone splines to approximate the baseline cumulative hazard function in the PH model and the baseline odds function in the PO model. Their method is efficient and provides a variance estimate in a closed form.

However, when a large number of covariates exist, they often display more complex effects than the linear format and there may exist interactions between them. In this case, flexible models which could handle potential nonlinear effects of covariates with high dimensionality are greatly desired. To incorporate possible nonlinear covariate effects for right-censored data, Huang [10] extended the partly linear proportional hazards model proposed by Sasieni [21], and considered a partly linear additive proportional hazards model as follows:

$$\lambda(t|Z) = \lambda_0(t) \exp\left\{\alpha^\top V + \phi_1(X_1) + \cdots + \phi_p(X_p)\right\}, \tag{2}$$

where $Z = (V^\top, X^\top)^\top$ is a $(q+p)$ dimensional covariate vector in which $V = (V_1, \ldots, V_q)^\top$ is a q-dimensional linear component and $X = (X_1, \ldots, X_p)^\top$ is a p-dimensional nonlinear component, both V and X are time independent, α is a q-dimensional regression coefficient vector, and $\phi_j(\cdot)$'s are unknown smooth functions, $j = 1, \ldots, p$. The partly linear additive Cox model is an extension of the linear Cox model and allows flexible modeling of covariate effects semiparametrically. Ma and Kosorok [17] studied partially linear transformation models with current status data where the partially linear PH model was one of their special cases. Ma [16] considered linear and partly linear PH cure models with current status data.

Another approach in dealing with high-dimensional nonlinear covariates is to use single-index models for dimension reduction. Huang and Liu [11] considered a single-index PH model for analyzing right-censored data. They used polynomial splines estimation along with a partial likelihood approach to estimate the parameters of the model. In a single-index model, the objective of interest depends on a single-index term through an unknown function $\psi(\cdot)$, which is termed as the link function in the rest of the paper. Sun et al. [27] suggested a partially linear single-index proportional hazards (PLSI-PH) model, further generalizing the method of Huang and Liu [11]. Recently, Shang et al. [23] proposed PLSI-PH model to analyze nested case-control data. All of these published works used a B-splines smoothing method to estimate the unknown link function of the single-index term.

In this paper, we consider a PLSI-PH model to analyze current status data, assuming that the log-hazard function given covariate history depends linearly on a covariate vector V and nonlinearly on a covariate vector X through a single-index $\beta^\top X$ and a nonlinear link function. We assume that the baseline hazard function is a nonparametric function and the link function takes a parametric spline function. This single-index allows us to combine many covariates (e.g., biomarkers) to achieve a parsimonious model. The new methodology challenge lies in the fact that the proposed model requires estimating the unspecified baseline hazard function together with the unknown parameters in the model. Using B-splines approximation to the log baseline cumulative hazard function, we propose a simultaneous sieve maximum likelihood estimator for the parametric vector. Asymptotic properties of the estimators are established under some regularity conditions by the means of counting processes and empirical processes.

The paper is organized as follows. Section 2 concerns polynomial splines smoothing estimation and implementation. Section 3 presents regularity conditions, efficient estimator and asymptotic properties. Section 4 exhibits a simulation study. Section 5 illustrates an application of the proposed model in the analysis of a renal function recovery dataset. Section 6 includes some concluding remarks. Major technical proofs and lemmas are relegated to Appendices A and B.

2. Estimation methods and implementation

To analyze current status data, suppose T is time to event of interest and $Z = (V^\top, X^\top)^\top$ is a covariate vector. Conditional on the covariate vector Z, we propose the following PLSI-PH model. The hazard function of T is modeled as follows:

$$\lambda(t|Z) = \lambda_0(t) \exp\{\alpha^\top V + \psi_\gamma(\beta^\top X)\}, \tag{3}$$