Computational Statistics and Data Analysis 67 (2013) 199–212

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Partially linear single index Cox regression model in nested case-control studies

Shulian Shang [a,*], Mengling Liu [a], Anne Zeleniuch-Jacquotte [a], Tess V. Clendenen [a], Vittorio Krogh [b], Goran Hallmans [c], Wenbin Lu [d]

[a] Departments of Population Health and Environmental Medicine, New York University School of Medicine, New York, USA
[b] Nutritional Epidemiology Unit, National Cancer Institute, Milan, Italy
[c] Public Health and Clinical Medicine/Nutritional Research, Umeå University, Umeå, Sweden
[d] Department of Statistics, North Carolina State University, Raleigh, USA

ARTICLE INFO

Article history:
Received 19 June 2012
Received in revised form 13 May 2013
Accepted 14 May 2013
Available online 22 May 2013

Keywords:
Nested case-control study
Risk-set sampling
Nonparametric regression
Nonlinear effect
Single index model

ABSTRACT

The nested case-control (NCC) design is widely used in epidemiologic studies as a cost-effective subcohort sampling method to study the association between a disease and its potential risk factors. NCC data are commonly analyzed using Thomas' partial likelihood approach under the Cox proportional hazards model assumption. However, the linear modeling form in the Cox model may be insufficient for practical applications, especially when there are a large number of risk factors under investigation. In this paper, we consider a partially linear single index proportional hazards model, which includes a linear component for covariates of interest to yield easily interpretable results and a nonparametric single index component to adjust for multiple confounders effectively. We propose to approximate the nonparametric single index function by polynomial splines and estimate the parameters of interest using an iterative algorithm based on the partial likelihood. Asymptotic properties of the resulting estimators are established. The proposed methods are evaluated using simulations and applied to an NCC study of ovarian cancer.

1. Introduction

Large cohort studies are precious resources to the study of disease etiology. However, it is costly to measure all the risk factors for the entire cohort, especially when disease is rare. As an alternative to the full-cohort design, the nested case-control (NCC) design (Thomas, 1979) has been widely used as a cost-effective subcohort sampling method. In this design, cases are ascertained within a large cohort. At the failure time of each case, a small number of controls are sampled among subjects who are still at risk, possibly matched to the case by some known confounders. Then covariates of interest are only measured on the cases and selected controls. The NCC design maintains the attractive feature of the full-cohort design to analyze biological specimens collected before the disease onset, providing an appropriate time sequence for a cause–effect relationship. In addition, both absolute risk and relative risk can be estimated under the NCC design (Langholz and Borgan, 1997).

NCC data are commonly analyzed using Thomas' partial likelihood approach under the Cox proportional hazards (PH) model (Thomas, 1979; Oakes, 1981), for which the hazard function is specified as $\lambda(t|x) = \lambda_0(t)\exp\{x^T\beta\}$, where $\lambda_0(t)$ is the unknown baseline hazard function and x is a p-dimensional covariate vector. A major assumption of the Cox PH model is that covariates have linear effects on the log of hazard. In epidemiologic studies where a large number of covariates are considered, covariates often exhibit more complex effects than the log–linear format and there may exist interactions

* Corresponding author. Tel.: +1 212 263 0371; fax: +1 212 263 8570.
E-mail addresses: shulian.shang@gmail.com, ss4577@nyu.edu (S. Shang).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.05.011

between them. Flexible models which could handle potential nonlinear effects of covariates with high dimensionality are greatly desired.

The single index model (Stoker, 1986; Hardle and Stoker, 1989; Ichimura, 1993) is a semiparametric model which achieves dimension reduction and avoids the "curse of dimensionality". In the linear regression setting, the single index model is an extension of the generalized linear model, with link function unspecified (Yu and Ruppert, 2002). In the survival analysis context, the single index model has been incorporated into the multiplicative hazard model (Wang, 2004; Huang and Liu, 2006):

$$\lambda(t|x) = \lambda_0(t)\exp\{\psi(x^T\beta)\}, \tag{1}$$

where $\psi(\cdot)$ is an unknown univariate smooth function. The multi-dimensional covariates are reduced to a linear combination as $x^T\beta$, namely a single index, and the coefficient β characterizes the relative importance of x. Note that if $\psi(\cdot)$ is monotone, β has a similar interpretation as the coefficient in the Cox PH model. Researchers have proposed various methods for fitting the single index model, such as the kernel smoothing technique (Ichimura, 1993; Hardle et al., 1993; Wang, 2004), the average derivatives method (Stoker, 1986; Hardle and Stoker, 1989) and polynomial spline approximation (Yu and Ruppert, 2002; Huang and Liu, 2006).

In model (1), all components of x are treated equally in the sense that no distinction is made between covariates of primary interest and nuisance ones. In epidemiologic studies, there are usually major risk factors of interest and multiple confounders such as demographics, anthropometric measures and socioeconomic status. Covariates that are most interesting to investigators would be modeled parametrically to render easy interpretation on their effects. Therefore, a partially linear single index (plSI) model has been proposed to extend model (1),

$$\lambda(t|v,x) = \lambda_0(t)\exp\{v^T\alpha + \psi(x^T\beta)\}, \tag{2}$$

where $v \in R^q$, $x \in R^p$ and $\psi(\cdot)$ is the unknown link function as above. In the linear regression setting, researchers have proposed to use the local linear method (Carroll et al., 1997), the kernel smoothing method (Xia et al., 1999) and the penalized spline method (Yu and Ruppert, 2002) to fit the partially linear single index model. In survival analysis, Lu et al. (2006) considered model (2) with a parametric baseline hazard function; Sun et al. (2008) studied this model with the polynomial spline technique; Li and Zhang (2011) extended model (2) to time-varying coefficients.

To the best of our knowledge, the inference of model (2) has not been studied for the NCC design. In this paper, we develop methods for the statistical inference of model (2) for NCC data and establish asymptotic properties of the resulting estimators. We are motivated by an NCC study investigating the association of inflammation-related cytokines and their modulators with the risk of ovarian cancer (Clendenen et al., 2011). This case-control study was nested within three prospective cohorts and for each case, two controls were selected at random from cohort members who fulfilled the risk set criteria. In total, we observed 230 cases and 432 matched controls. The levels of cytokines and cytokine modulators were measured from stored blood samples collected at enrollment. Potential confounders included body mass index and medical history. Our main interest is to estimate the effect of biomarkers on the risk of ovarian cancer while adjusting for confounders. We thus study the partially linear single index proportional hazards model, which allows flexible and parsimonious modeling of nonlinear effects of confounders and easy interpretation on the parameters for covariates of interest.

This paper is organized as follows. In Section 2, we present methods for estimation, inference and implementation of the proposed model. Section 3 includes simulation studies evaluating the finite sample performance of our proposed estimator and the analysis of the NCC data on ovarian cancer as an illustration. We conclude in Section 4 with discussions and provide all the technical details in the Appendix.

2. Methods

2.1. Notation and model

Suppose that we have a cohort with size n. For the ith subject, $i = 1, \ldots, n$, let $z_i = \min(t_i, c_i)$ be the observed survival time subject to censoring, where t_i denotes the survival time and c_i denotes the censoring time. Define $\delta_i = I(t_i \le c_i)$ as the censoring indicator. At a specific time t, let $\tilde{R}(t) = \{i : z_i \ge t\}$ denote the risk set. By the NCC design, subjects with the observed event, i.e. $\delta_i = 1$, are identified as cases. At the failure time of each case, $(M-1)$ controls are randomly sampled without replacement from the risk set, excluding the case itself. For case i, let R_i^* denote the indices of the $(M-1)$ selected controls and define the case-control set $R_i = R_i^* \cup \{i\}$. Then covariate information is assembled for the cases and selected controls, consisting of two components: the q-dimensional vector v denotes the primary risk factors to be modeled parametrically, and the p-dimensional vector x denotes confounders to be included in the nonparametric single index component.

For the purpose of identifiability of the partially linear single index model (2), we impose the constraints that $\psi(0) = 0$, $\|\beta\| = (\beta^T\beta)^{1/2} = 1$ and the first nonzero component of β is positive (Wang, 2004). Following Huang and Liu (2006) and Sun et al. (2008), we use a polynomial spline function to first approximate the derivative of the unknown function $\psi(\cdot)$ by

$$\psi'(x^T\beta) = \sum_{j=1}^{k}\gamma_j B_j(x^T\beta) = \gamma^T\mathbf{B}(x^T\beta), \tag{3}$$