Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT
Journal of Multivariate Analysis 92 (2005) 186–204
Journal of Multivariate Analysis
http://www.elsevier.com/locate/jmva

Covariate selection for semiparametric hazard function regression models

Florentina Bunea* and Ian W. McKeague

Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, USA

Received 22 September 2002

Abstract

We study a flexible class of nonproportional hazard function regression models in which the influence of the covariates splits into the sum of a parametric part and a time-dependent nonparametric part. We develop a method of covariate selection for the parametric part by adjusting for the implicit fitting of the nonparametric part. Asymptotic consistency of the proposed covariate selection method is established, leading to asymptotically normal estimators of both parametric and nonparametric parts of the model *in the presence of* covariate selection. The approach is applied to a real data set and a simulation study is presented.

AMS 2000 subject classifications: 62N02

Keywords: Additive risk model; Cox model; Penalized partial likelihood; Penalized likelihood; Model selection; Survival analysis

1. Introduction

Covariate selection is a form of model selection in which the class of models under consideration is represented by subsets of covariate components to be included in the analysis. Model selection methods are well developed in parametric settings, and in recent years they have been extended to wide classes of nonparametric models [2]. For applications in survival analysis, however, in which the presence of censoring and the use of complex time-dependent hazard function regression models is

*Corresponding author.
E-mail address: flori@stat.fsu.edu (F. Bunea).

0047-259X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2003.09.006

becoming increasingly popular (see, e.g., [1]), generally applicable and fully validated procedures have not yet been developed.

In this paper we study covariate selection for conditional hazard function models of the form

$$h(t,x,z) = \psi(\beta^T x + f(t)^T z), \tag{1.1}$$

where ψ is a known (nonnegative) link function, (x,z) is a partition of the covariates into a q-vector x and a p-vector z, β is an unknown q-vector of regression parameters and $f(t)$ is an unknown p-dimensional nonrandom function of time. We develop a model selection procedure to find the best subset of x-covariates and study the asymptotic properties of the corresponding regression parameter estimates *after* model selection.

The above model provides a flexible extension of the Cox proportional hazards model $h(t,x) = \exp(\beta^T x + f(t))$, where $f(t)$ is the log-baseline hazard function. Our model is more flexible in the sense that it allows some of the covariates to have a longitudinal (or time-dependent) influence on survival. For the identity link function, the model reduces to the partly parametric additive risk model of McKeague and Sasieni [13]. Recently, Martinussen et al. [12] studied the model in the case of an exponential link function.

Typically, some covariates are known to have a longitudinal influence on survival, so those covariates are placed in z. However, only a small (but fixed) number of covariates can be treated in this way as an additional time-dependent function enters the model for each component of z. The remaining covariates are placed in x. This creates the need for a procedure to select a subset of the x-components that avoids both overfitting and underfitting. With the nonzero components of β corresponding to an unknown subset $I = I_0$ of the x-covariates, the statistical problem is to estimate I_0 and the corresponding components of β.

Numerous covariate selection procedures have been proposed for the Cox model: penalized partial likelihood—henceforth PPL [16], a backwards elimination covariate selection method [9], Bayesian model averaging [14,15], Bayesian variable selection [8], the lasso method for PPL [17], and nonconcave PPL [7]. Large sample properties of these procedures are largely unexplored, with the exceptions of Senoussi [16] and Fan and Runze [7]. All these procedures only require *parametric* model selection techniques because they exploit partial likelihood which does not involve the infinite-dimensional part of the semiparametric model (the baseline hazard function). A more sophisticated PPL procedure was developed by Letué [11] for fitting the general proportional hazards model

$$h(t,x) = \exp(g(x) + f(t)),$$

where $g(x)$ is an unknown function of the covariates x and $f(t)$ is the log-baseline hazard function. This model may be unsuitable, however, when x has high dimension because of the curse of dimensionality. None of the above procedures extends beyond the proportional hazards framework.

To study semiparametric models of form (1.1), in which a partial likelihood for β is not available and I_0 is also regarded as a parameter, we need a different approach.