Computational Statistics and Data Analysis 80 (2014) 140–152

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust estimation for survival partially linear single-index models

Xiaoguang Wang*, Xinyong Shi

School of Mathematical Sciences, Dalian University of Technology, Dalian, China

ARTICLE INFO

Article history:
Received 29 May 2013
Received in revised form 11 June 2014
Accepted 27 June 2014
Available online 5 July 2014

Keywords:
Local linear regression
M-estimation
Right censoring
Semiparametric models
Synthetic data

ABSTRACT

The partially linear single-index model is an interesting semiparametric model extended by the partially linear model and the single-index model, which supply a good balance between flexibility and parsimony. A robust estimation is proposed to fit the partially linear single-index model in case outliers may occur in the right censored response. This method provides a flexible way for modeling survival data. It is a profile M-estimation version and the estimation procedure involves transforming the censored data into synthetic data at first, then it results in fitting the common partially linear single-index models by a robust loss function. Asymptotic properties for the estimators of the linear and single-index coefficients and the optimal rate of convergence for the estimator of the nonparametric function are established. The finite sample performance of the proposed method is assessed by Monte Carlo simulation studies, and demonstrated by the analyses of PBC data and NCCTG lung cancer data.

1. Introduction

In survival analysis, difficulties arise when the traditional regression models are used to analyze survival data in practice, since survival data are often not completely observed. Let Y be the completely observed variable of interest, such as the survival time or some transformation of the survival time. In many applications, especially in biomedical studies, Y cannot be completely observed due to possibly censoring, for instance, withdrawal of patients from the study, or death from a cause unrelated to the specific disease of being studied, etc. In this paper, we focus on the right censoring mechanism rather than the left censoring although the methodology can be directly extended to that. Let C denote the censoring variable. Due to the right censoring mechanism, we only observe (V, Δ), where $V = \min(Y, C)$ and $\Delta = I(Y \leq C)$ are the observed (possibly censored) response variable and the censoring indicator, respectively.

When a set of predictors are considered, the accelerated failure time (AFT) model is an attractive alternative to the commonly-used Cox proportional hazards model and it is framed as a usual linear model with the response variable as the logarithm of the failure time (Kalbfleisch and Prentice, 2001). It postulates that the effect of explanatory variables is to multiply the predicted event time by some constant and that is more appealing in many ways. Semiparametric estimation in the AFT model with an unspecified error distribution has been studied extensively in the literature for common right censored data. In particular, some approaches in this aspect have received special attention including the Buckley–James iterative method, the rank-based method, and the inverse probability weighting (IPW) method (Miller, 1976; Buckley and James, 1979; Koul et al., 1981; Ritov, 1990; Tsiatis, 1990; Ying, 1993; Zhou, 2006). Another approach to fit the right censored data is to replace the possibly unavailable response Y with surrogate values by an appropriate mapping of the observed data. The key requirement is that the mapping is approximately, what Fan and Gijbels (1994) termed, a censoring

* Correspondence to: 2 Linggong Road, Ganjingzi District, Dalian 116024, China. Tel.: +86 411 84708351 8123.
E-mail addresses: wangxg@dlut.edu.cn (X. Wang), sxy74@163.com (X. Shi).

http://dx.doi.org/10.1016/j.csda.2014.06.020
0167-9473/© 2014 Elsevier B.V. All rights reserved.

unbiased transformation. Also see Koul et al. (1981), Leurgans (1987), and Zhou (1992). Also see Leiva et al. (2007) and Shang et al. (2013). Robust methods are also proposed for the AFT model. Powell (1984) proposed the least absolute deviation method for the fixed censoring times. For random censoring, Ying et al. (1995) modified quantile estimating equations by assuming independence between survival and censoring times. Portnoy (2003) proposed censored quantile regression by redistributing censored data. By using the martingale structure of right censoring time, Peng and Huang (2008) developed a quantile estimating equation. Due to the analytical difficulty and computational complexity of quantile regressions, we propose an M-estimation to fit the AFT model.

For more flexible situation, a nonlinear structure of covariate is employed in the AFT model, that is a semiparametric regression model. It can relax some restrictive assumptions on parametric models and is flexible enough to capture the true underlying relationships between explanatory variables and response in dealing with complex real data. One of the most popular and important semiparametric models is the partial linear single-index model, which is an important generalization of the traditional linear model and the single-index model. It is reduced to a partially linear model (Engle et al., 1986; Härdle et al., 2000), or the single-index model (Härdle et al., 1993; Ichimura, 1993) as special cases. Because of recent advances in the semiparametric regression, Jin et al. (2003, 2006) and Zou et al. (2011) developed the estimation method in the partial linear AFT model. Lu et al. (2006) and Lu and Cheng (2007) proposed estimation for the partial linear single-index model. However, the existing estimation procedures may be sensitive to outliers and their efficiency may be significantly deduced for non-normal errors.

This paper focuses on the robust estimation approach for the partially linear single-index AFT model. We consider the following randomly censored survival partially linear single-index model

$$Y = g(\mathbf{X}^T\boldsymbol{\beta}) + \mathbf{Z}^T\boldsymbol{\alpha} + \varepsilon, \tag{1}$$

where Y is the survival time or the logarithm survival time, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ is a vector of unknown parameters in $\mathbb{R}^{p+q}$ with $\|\boldsymbol{\beta}\| = 1$ (where $\|\cdot\|$ denotes the Euclidean norm), $\mathbf{X} \in \mathbb{R}^p$, $\mathbf{Z} \in \mathbb{R}^q$, $g(\cdot)$ is an unknown univariate link function, and ε is a random error with mean zero. The restriction $\|\boldsymbol{\beta}\| = 1$ on the single-index coefficients is required for the parameter identifiability. We assume without loss of generality that the first component of $\boldsymbol{\beta}$ is positive, i.e., the parametric space of $\boldsymbol{\beta}$ is $\boldsymbol{\Theta} = \{\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T, \|\boldsymbol{\beta}\| = 1, \beta_1 > 0\}$ for ensuring identifiability. Assume that $(\mathbf{X}, \mathbf{Z})$ and ε are independent and C is independent of $(Y, \mathbf{X}, \mathbf{Z})$. Remind $V = \min(Y, C)$ and $\Delta = I(Y \le C)$. The observations $\{(V_i, \mathbf{X}_i, \mathbf{Z}_i, \Delta_i), i = 1, \ldots, n\}$ is a random sample from the population $\{(V, \mathbf{X}, \mathbf{Z}, \Delta)\}$.

To resist the potential outliers, we focus on the M-estimation (Huber, 1964) to fit the partially linear single index survival model. The basic idea of our robust fit is to replace the least square loss function with a Huber type loss function. Hence, one of the building blocks of our proposal is the robust estimation. After transforming the censored data into synthetic data unbiasedly, we fulfill the M-estimation for the semiparametric model by an iterative method. The local linear regression technique is applied to approximate the nonparametric single index function given the parametric part. Then we plug in the function estimator and compute the parametric part. These two steps are both obtained by a Huber type loss function. Via this iterative procedure, the nonparametric component and the parametric component can be estimated.

The rest part of the paper is as follows. Section 2 describes our robust estimation procedure. The asymptotic properties of these estimators are studied. Section 3 provides a computation algorithm and illustrates the estimation performance with Monte Carlo simulation results. In Section 4, we evaluate the proposed method by analyzing PBC data and NCCTG lung cancer data. Section 5 ends the paper with a brief discussion. All the conditions and technical proofs of the main results are deferred to the Appendix.

2. Estimation procedure and main results

In this section, we propose a robust estimation procedure and establishing its asymptotic properties. The main motivation is to generalize the robust estimating method for partially linear single-index models to survival time data. We adopt the Fan and Gijbels (1994) method to transform the right censored response to the unbiased synthetic data. In order to estimate the extra nonparametric component g, we use the local linear regression technique to construct a local objective function. Then we develop the estimation for parametric parts. The new profile estimation procedure naturally combines the survival analysis and the partially linear single-index model in one unified framework, which greatly facilitates the complicated data analysis and model inferences.

2.1. Transformation of the data

We construct the synthetic data in this subsection. Let $\bar{G} = 1 - G$ and $\bar{F} = 1 - F$ are the survival functions of the right censoring variable C and the survival time variable Y, respectively, where $G(t) = P(C \le t)$ and $F(t) = P(Y \le t)$. When $\bar{G}$ is unknown, we can estimate it by the Kaplan–Meier product-limit estimator. Denote $\tau_G = \inf\{t : G(t) = 1\}$ and $\tau_F = \inf\{t : F(t) = 1\}$. We suppose $\tau_F \le \tau_G$ throughout this paper. Denote $T_1 = V/(1 - \hat{G}(V-))$ and $T_2 = \int_{-\infty}^{\infty}\{I[V \ge s]/(1 - \hat{G}(s-)) - I[s < 0]\}ds$, where $1 - \hat{G}(\cdot-)$ is the left-continuous version of the Kaplan–Meier product-limit estimator defined as

$$1 - \hat{G}(t) = \prod_{i=1, V_{(i)} \le t}^{n} \left[\frac{n-i}{n-i+1}\right]^{I[\Delta_{(i)}=0]},$$