Computational Statistics and Data Analysis 93 (2016) 257–269

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Partially linear transformation cure models for interval-censored data

Tao Hu [a], Liming Xiang [b,*]

[a] School of Mathematical Sciences & BCMIIS, Capital Normal University, Beijing 100048, China

[b] School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Singapore

ARTICLE INFO

Article history:
Received 4 March 2014
Received in revised form 11 August 2014
Accepted 16 August 2014
Available online 27 August 2014

Keywords:
Cure rate
Interval censoring
Semiparametric efficiency
Sieve likelihood ratio test
Sieve maximum likelihood estimator
Transformation

ABSTRACT

There has been considerable progress in the development of semiparametric transformation models for regression analysis of time-to-event data. However, most of the current work focuses on right-censored data. Significantly less work has been done for interval-censored data, especially when the population contains a nonignorable cured subgroup. A broad and flexible class of semiparametric transformation cure models is proposed for analyzing interval-censored data in the presence of a cure fraction. The proposed modeling approach combines a logistic regression formulation for the probability of cure with a partially linear transformation model for event times of susceptible subjects. The estimation is achieved by using a spline-based sieve maximum likelihood method, which is computationally efficient and leads to estimators with appealing properties such as consistency, asymptotic normality and semiparametric efficiency. Furthermore, a goodness-of-fit test can be constructed for the proposed models based on the sieve likelihood ratio. Simulations and a real data analysis are provided for illustration of the methodology.

1. Introduction

Interval-censored data commonly arise in clinical and pubic health sciences, where the event time is never observed precisely but known at intermittent follow-up visits. Right and left censoring are special cases of interval censoring. A number of statistical methods for handling such data can be found, e.g., in Sun (2006). Due to advances in modern medical techniques and health care, survival rates have substantially improved. This is often indicated in some studies by the presence of a significant "cured" proportion of subjects who are not at risk to experience the event of interest, such as failure or relapse. As a consequence, there may be a considerable number of subjects with large right-censored times in a sample.

Basically, there are two classes of cure rate models developed in the literature to allow for such a subgroup of subjects to be cured. One class is promotion time (non-mixture) cure models proposed by Yakovlev and Tsodikov (1996), in which a single model is used to describe the survival function of the whole population with right-censored data. For interval-censored data, Liu and Shen (2009) studied the promotion time cure model using a semiparametric maximum likelihood approach, while Hu and Xiang (2013) developed a spline-based sieve likelihood method for the same model, leading to more efficient estimators.

Models we consider in this paper belong to the other class of cure rate models, named as mixture cure models, which assume that the study population is a mixture of susceptible and non-susceptible subjects, and then model the effects of

* Corresponding author. Tel.: +65 6513 7451; fax: +65 6515 8213.
E-mail address: lmxiang@ntu.edu.sg (L. Xiang).

http://dx.doi.org/10.1016/j.csda.2014.08.014
0167-9473/

covariates on the cure rate of the population and the survival function of non-susceptible subjects through the logistic regression and survival model components respectively. This modeling approach is straightforward and has attracted much attention in the literature (e.g. Farewell, 1982, Kuk and Chen, 1992, Li et al., 2001 and Tsodikov et al., 2003). Extensions of the mixture cure models to interval-censored data have been discussed in recent years. Banerjee and Carlin (2004) proposed parametric cure rate models for interval-censored smoking relapse times from a Bayesian perspective. With current status data, known as a special case of interval-censored data, Lam and Xue (2005) discussed a semiparametric accelerated failure time (AFT) model based cure mixture using the sieve maximum likelihood estimation. Ma (2010) studied a semiparametric Cox cure model for mixed case interval-censored data using the maximum likelihood approach and weighted bootstrap for estimation and inference. More recently, Li and Ma (2010) applied a fully parametric AFT model and Xiang et al. (2011) and Lam and Wong (2014) considered the semiparametric Cox model for the susceptible group in the analysis of clustered interval-censored data, where the correlated nature of the data was accommodated through random effects in both the logistic regression and survival model components.

However, either the parametric survival model or the proportional hazards assumption is too restrictive for susceptible subjects in many biomedical applications and may result in unreliable estimates when the assumption is violated. This leads to the development of more accurate modeling approaches. The proportional odds model, as a useful alternative, provides better summary of data when the hazard ratio between two sets of covariate values is not constant but converges over time (Murphy et al., 1997). The proportional odds cure rate models have been studied consequently by Mao and Wang (2010) and Gu et al. (2011). To encompass both the proportional hazards and proportional odds based mixture cure models as special cases, Fine (1999) and Lu and Ying (2004) exploited the mixture cure model with linear transformation models for the susceptible population using estimating equation methods. Their work has been further extended by Othus et al. (2009) to more general situations incorporating dependent censoring as well as time-dependent covariates. In addition, transformation models within the non-mixture cure model framework have been discussed by Zeng et al. (2006) given biological considerations and Yin (2008) for multivariate survival data. However, all these aforementioned studies have focused on right-censored data. Ma and Kosorok (2005) studied partly linear transformation models for current status data without considering the possibility of cure.

In this paper, we develop a semiparametric transformation modeling approach, in which survival time of a susceptible subject is specified by transforming it into a variable linked to covariates through a partially linear regression function while the cure fraction is modeled by a logistic regression. A key feature of semiparametric models with interval-censored data is that the baseline hazard function cannot be eliminated from the likelihood or partial likelihood as that under right censoring. We approximate the unknown functions via a monotone B-spline (Schumaker, 1981) and construct a spline-based sieve semiparametric likelihood (Shen and Wong, 1994; Shen, 1997) to estimate parameters and the nonparametric components simultaneously. Because the sieve likelihood function usually involves fewer parameters to be estimated than the nonparametric likelihood function, the proposed estimation may reduce the computational burden in comparison with the non-parametric maximum likelihood (NPML) estimation techniques used in Ma and Kosorok (2005) and Liu and Shen (2009). After estimation, it is natural to construct hypothesis test for checking the goodness-of-fit of the model. We then develop a sieve likelihood ratio test for selecting significant variables in the parametric component and checking the nonparametric component, in a similar spirit to the testing approach given by Shen and Shi (2005).

The rest of the paper is organized as follows. Section 2 presents the model, sieve maximum likelihood estimation and its computational issues. Section 3 studies asymptotic properties of the resulting estimator and then develops a goodness-of-fit test based on the sieve likelihood ratio. Simulation studies and a real data example are provided in Sections 4 and 5, respectively. Finally, Section 6 concludes the paper with a discussion. Sketches of proofs of theorems are given in the Appendix.

2. Model and inference

2.1. Model and likelihood function

Under the mixture cure modeling approach (Farewell, 1982), a decomposition of the event time is given by

$$T = YT^* + (1 - Y)\infty, \tag{2.1}$$

where $T^* < \infty$ denotes the failure time of a susceptible subject and Y indicates, by the value 1 or 0, whether the study subject is susceptible or not. Conditional on covariates $X \in \mathbb{R}^p$, $W \in \mathbb{R}$ and $Z \in \mathbb{R}^q$, we define the partially linear transformation cure models as follows:

$$\begin{aligned} &H(T^*) = -\boldsymbol{\beta}^\tau X - \phi(W) + \epsilon, \\ &p(Z) = P(Y = 1|Z) = \frac{\exp(\boldsymbol{\alpha}^\tau Z)}{1 + \exp(\boldsymbol{\alpha}^\tau Z)}, \end{aligned} \tag{2.2}$$

where H is an unknown non-decreasing transformation, ϕ is an unknown smooth function, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are unknown regression parameter vectors of p- and q-dimension, respectively. Covariates Z and X may share some common components and Z includes 1 so that $\boldsymbol{\alpha}$ contains an intercept term. ϵ is an error term with a known distribution F. Common choices of F include