Journal of Statistical Planning and Inference 156 (2015) 80–89

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Estimation for semiparametric transformation models with length-biased sampling

Xuan Wang [a], Qihua Wang [a,b,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China

[b] Institue of Statistical Science, Shenzhen University, Shenzhen 518006, PR China

ARTICLE INFO

Article history:
Received 20 November 2013
Received in revised form 15 April 2014
Accepted 2 August 2014
Available online 19 August 2014

MSC:
62N01
62N02
62P10

Keywords:
Length-biased data
Right-censored data
Transformation model
Estimating equation

ABSTRACT

For length-biased and right-censored data, we propose an estimation method to assess the effects of risk factors under the semiparametric linear transformation model. Unlike the existing method of Shen et al. (2009) based on the ranks of observed failure times, the new estimators are obtained from counting process-based unbiased estimating equations. Consistency and asymptotic normality for the estimators are derived under suitable regularity conditions. We evaluate the finite sample performance of the proposed method and make a comparison with that of Shen et al. (2009) by simulation studies. A real data example is analyzed to illustrate the proposed method.

1. Introduction

By terminology, length-biased data are left-truncated and right censored data under the stationary assumption that the initial times follow a stationary Poisson process. In observational studies, one often encounters length-biased data. A large number of examples for length-biased data can be found in Qin and Shen (2010) and Shen et al. (2009). Under the length-biased sampling, the observed samples are not randomly sampled from the population of interest but with probability proportional to their lengths, which makes the observed time intervals from initiating to failure in the prevalent cohort tend to be longer than those from the underlying distribution for the general population.

Extensive literature has focused on estimating the unbiased distribution given length-biased data (Wang, 1991; Asgharian and Wolfson, 2005). There are two main difficulties encountered in analyzing length-biased data. One is that when studying the effects of risk factors on the population failure time, the model structure assumed for the target population is often different from the one for the observed length-biased data. The other is that the failure time and right-censoring time are not independent except in trivial cases. To model risk factors on the distribution of the underlying population, Wang (1996) described a proportional hazards model for length-biased data and used a biased-adjusted risk set to construct the pseudo-likelihood for estimation by just ignoring right censoring. Qin and Shen (2010) proposed two estimating equations to estimate covariate coefficients under the Cox model based on two mean zero processes respectively.

* Corresponding author at: Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China.
E-mail address: qhwang@amss.ac.cn (Q. Wang).

http://dx.doi.org/10.1016/j.jspi.2014.08.001

Under semiparametric transformation models, Shen et al. (2009) estimated the unknown covariate coefficients by constructing unbiased estimating equations with an inverse weight chosen to adjust the bias induced by the length-biased data and dependent censoring for length-biased data based on the method of Cheng et al. (1995). Their estimating equations use the ordering information of the observed length-biased times.

In this paper, we propose a new method to assess covariate effects under the semiparametric transformation model for length-biased data subject to right censoring. After correcting for the bias due to length-biased data and dependent censoring, we obtain a mean zero process, which is a generalization of the martingale process of Chen et al. (2002). Estimating equations are then constructed based on the mean zero process.

The remainder of this paper is organized as follows. In Section 2, we give the basic notations, construct the estimating equations and present theoretical properties for the proposed estimators. In Section 3, we evaluate the finite sample performance of the proposed method and compare it with existing method due to Shen et al. (2009) by simulation studies. We also apply our method to the analysis of a real data set. A discussion is provided in Section 4 and the proofs of theorems are given in the Appendix.

2. Methodology

2.1. Data and model

We assume that T_0 is the time measured from the initiating event to failure, A is the time from the initiating event to examination (truncation variable), V is the time from examination to failure (residual survival time) and C is the time from examination to censoring (residual censoring time). With length-biased sampling, one can only observe T among those $T_0 > A$, where $T = A + V$ is the observed failure time. Let $\tilde{T}_i = T_i \wedge (A_i + C_i)$, $\delta_i = I(V_i \le C_i)$ and $\mathbf{Z}_i$ be a vector of covariate for the ith subject for $i = 1, 2, \ldots, n$. It is reasonable to assume C and (A, V) are independent given $\mathbf{Z}$ and we firstly assume C is independent of $\mathbf{Z}$.

Let f_0 denote the density for the unbiased data T_0, called the unbiased density and g represent the length-biased density for the length-biased data T (conditional on $T_0 > A$). Then the following relationship holds,

$$g(t) = \frac{tf_0(t)}{\mu}, \qquad \mu = \int_0^\infty sf_0(s)\,\mathrm{d}s.$$

Given covariates $\mathbf{Z} = \mathbf{z}$, similar equation holds,

$$g(t|\mathbf{z}) = \frac{tf_0(t|\mathbf{z})}{\mu(\mathbf{z})}, \qquad \mu(\mathbf{z}) = \int_0^\infty sf_0(s|\mathbf{z})\,\mathrm{d}s. \tag{2.1}$$

To make (2.1) meaningful, we require $\mu(\mathbf{z}) < \infty$.

The linear transformation model assumes that the true underlying failure time T_0 is related to the covariates linearly. That is

$$H(T_0) = -\beta'\mathbf{Z} + \varepsilon, \tag{2.2}$$

where H is an unknown increasing function, ε is a random variable independent of $\mathbf{Z}$ with a known distribution and β is an unknown p-dimensional regression parameter of interest. The proportional hazards model and the proportional odds model are special cases of model (2.2) with ε following the extreme value distribution and the standard logistic distribution respectively.

The usual estimation methods for linear transformation model (Cheng et al., 1995; Chen et al., 2002; Zeng and Lin, 2006) cannot be used here directly. The method of Shen et al. (2009) for length-biased data based on Cheng et al. (1995) takes advantage of the stochastic ordering preserved for length-biased times and unbiased times. Differently from their method, here we construct estimating equations based on a mean zero process.

2.2. Estimating equations approach

By a similar calculation to that of Zelen (2005) and Asgharian and Wolfson (2005), the joint distribution of (A, V) given $\mathbf{Z}$ is

$$f_{A,V}(a, v|\mathbf{Z} = \mathbf{z}) = f_0(a + v|\mathbf{z})I(a > 0, v > 0)/\mu(\mathbf{z}). \tag{2.3}$$

Let $q(t, \delta|\mathbf{z})(t \ge 0, \delta \in \{0, 1\})$ denote the joint conditional density of $\tilde{T}$ and δ given covariates $\mathbf{Z} = \mathbf{z}$, then,

$$q(t, 1|\mathbf{z})\,\mathrm{d}t = Pr(\tilde{T} \in (t, t + dt), \delta = 1|\mathbf{z}). \tag{2.4}$$

Since it is assumed that the residual censoring time C is independent of (A, V) given covariates $\mathbf{Z} = \mathbf{z}$, it follows

$$Pr(\tilde{T} \in (t, t + dt), \delta = 1|\mathbf{z}) = \int_0^t f_{A,V}(t - v, v|\mathbf{z})S_C(v)\,\mathrm{d}v\,\mathrm{d}t = f_0(t|\mathbf{z})w(t)/\mu(\mathbf{z})\,\mathrm{d}t, \tag{2.5}$$

where $w(t) = \int_0^t S_C(v)\,\mathrm{d}v$ and $S_C(\cdot)$ is the survival function for the residual censoring variable C.