Journal of Multivariate Analysis 130 (2014) 45–63

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Analyzing right-censored and length-biased data with varying-coefficient transformation model

Cunjie Lin [a,*], Yong Zhou [a,b]

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

[b] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China

ARTICLE INFO

Article history:
Received 25 January 2013
Available online 17 May 2014

AMS subject classification:
62H12

Keywords:
Length-biased data
Local linear
Right-censored
Transformation model
Varying-coefficient model

ABSTRACT

Right-censored and length-biased data arise in many applications, including disease screening and epidemiological cohort studies. It is challenging to analyze such data, since independent censoring assumption is violated in the presence of biased sampling. In this paper, we study the varying-coefficient transformation models with right-censored and length-biased data. We use the local linear fitting technique and propose estimators of varying coefficients by constructing the local inverse probability weighted estimating equations. We have shown that the proposed estimators are consistent and asymptotically normal and their variances can be estimated consistently. We pay special attention to the case where the censoring variable depends on the covariates. We conduct simulation studies to assess the performance of the proposed method and demonstrate its application on a real data example.

1. Introduction

Length-biased data arise in many applications, including disease screening and epidemiological cohort studies. The prevalent sampling design, with which only studying subjects with a condition or disease are sampled, is generally more efficient in practice. The observations collected via this sampling scheme are usually subject to right-censoring due to the loss of follow-up. In addition, the observations are also left truncated, as those who failed before the recruitment cannot be observed. When the truncation time follows a uniform distribution, that is, the incidence of disease onset follows a stationary Poisson process, the observations from a prevalent cohort study are length-biased [31]. For example, in the Canadian study of Health and Aging (CSHA), about 10,000 Canadians over 65 years old were recruited and screened for dementia. For the individuals who were found to have dementia in the study population, the approximate date of onset of dementia and the subsequent time of death or censoring were recorded. Length-biased sampling occurred here because those individuals who had dementia and did not survive to the examination time for the CSHA were excluded from the study. Only those individuals who had dementia and lived out the examination time for the CSHA could have been observed (see [2]).

The analysis of right-censored and length biased data has attracted a lot of attention in the literature. Many articles focus on estimating the unbiased distribution under length-biased sampling. See, for example, [30,32,33,20,2]. One challenge in analyzing such data is the dependence between the right censoring time and the failure time, i.e. informative censoring. Some authors avoided this difficulty by prohibiting right censoring [32–34] or ignoring it.

* Corresponding author.
E-mail addresses: cunjie1024@yahoo.cn, lincunjie@amss.ac.cn, cunjie1024@gmail.com (C. Lin).

http://dx.doi.org/10.1016/j.jmva.2014.05.003

Another challenge is that the observed length-biased data change the model structure assumed for the target population. For example, under the Cox proportional hazards model, the standard partial likelihood approach is not suitable for the length-biased data because its likelihood function cannot be decomposed in the usual way. Qin and Shen [25] developed weighted estimating equation methods for estimating the Cox model, taking into account the change of the model structure. The more recent work has discussed statistical inference for both parametric and non-parametric regression models for the length-biased data. For example, Shen, Ning and Qin [28] studied the covariate effects on failure times using the semiparametric transformation and the accelerated failure time (AFT) models. Ning, Qin and Shen [23] proposed a Buckley–James-type estimator for estimating the covariate effects under the AFT models.

In this paper, we consider the varying-coefficient transformation model for right-censored and length-biased data. The varying-coefficient model is a simple and useful extension of the standard regression model. This model has found applications in many areas, such as time series analysis [7,6,5], longitudinal data analysis [18,12], and survival analysis [11]. Many authors have studied this model for right-censored survival data [8,14,17,27].

However, to the best of our knowledge, there has been no work on this model for right-censored and length-biased data. The existing method cannot directly apply to such data because they involve informative censoring. In addition, length-biased sampling changes the distribution of the observed data. We propose a method that combines the main model with an auxiliary model for dependent censoring and adopt the local linear fitting technique in estimating the main model. In most of the existing studies, it is assumed that the censoring variable is independent of the covariates. However, this assumption is not satisfied in the situations where the survival time and censored time depend on some common covariates. We develop an inverse cumulative-probability weighted approach to taking into account dependent censoring.

The rest of the article is organized as follows. In Section 2 we describe the structure of the right-censored and length-biased data. We describe an inverse cumulative-probability weighting method for dependent censoring. In Section 3 we state the main results, including the consistency and asymptotic normality of the proposed estimators. Based on these results, we provide consistent variance estimation for these estimators. In Section 4 we conduct simulation studies to evaluate the finite sample performance of the proposed method and analyze a real data set to illustrate its application. The proofs of the asymptotic results are given in the Appendix.

2. Model and estimating procedure

2.1. Background

Let $\tilde{T}$ be an uncensored variable of interest without length-bias, that is, the time from the initiation event to the failure event. Consider the varying-coefficient model

$$H(\tilde{T}) = \mathbf{a}(U)^T\mathbf{X} + \varepsilon, \tag{2.1}$$

where $H(\cdot)$ is a known function, $\mathbf{a}(U) = (a_1(U), \ldots, a_p(U))^T$ is a $p \times 1$ vector of unknown coefficient functions and $\mathbf{X}$ is a p-dimensional vector of covariates. A well-known form of $H(\cdot)$ is the Box–Cox [4] transformation. For simplicity, we assume that U is a single variable and the error term ε satisfies $E(\varepsilon) = 0$, $E(\varepsilon^2) = \sigma_\varepsilon^2 < \infty$.

Let A be the time from the initiating event to examination, V be the duration measured from the examination to the failure, and C be the time from the examination to censoring. Under length-biased sampling, only those T with $\tilde{T} > A$ can be observed. So $T = A + V$ is the observed survival time. Here A is also referred to as the truncation variable (or backward recurrence time) and V as the residual survival time (or forward recurrence time).

Let f be the unbiased density function of $\tilde{T}$. The biased density function f_b of the length-biased data T given the covariates $\mathbf{X} = \mathbf{x}$, $U = u$ has the form (see [28])

$$f_b(t|\mathbf{x}, u) = \frac{tf(t|\mathbf{x}, u)}{\mu(\mathbf{x}, u)}, \quad \mu(\mathbf{x}, u) = \int_0^\infty sf(s|\mathbf{x}, u)ds,$$

where $f(t|\mathbf{x}, u)$ denotes the unbiased density given the covariates $(\mathbf{x}, u)$, and $\mu(\mathbf{x}, u) < \infty$. It is worthy noting that the above biased density function for the length-biased data T is derived conditionally on $\tilde{T} > A$.

Suppose we have a random sample, $(Y_i, A_i, \delta_i, \mathbf{X}_i, U_i)$, $i = 1, 2, \ldots, n$, where $Y_i = \min(T_i, A_i + C_i)$, $T_i = A_i + V_i$, $\delta_i = I(V_i \leq C_i)$ and n is the sample size. We assume that given the covariates $(\mathbf{X}, U)$, C and (A, V) are independent and the stationarity assumption holds, that is, the initiation times follow a stationary Poisson process. In order to avoid lengthy technical discussion of the tail behavior of the limiting distributions, we assume that $P(C > t_{\mathbf{X},U}|\mathbf{X}, U) > 0$, where $0 < t_{\mathbf{X},U} = \sup\{t : P(V \geq t|\mathbf{X}, U) > 0\}$. By this assumption, the support of C covers the support of V for any covariate value. Furthermore, inference is assumed to be restricted to the interval $[0, \tau]$ to ensure that all regression parameters are estimable, where τ is chosen such that $\inf_{\mathbf{x},u} P(V \geq \tau|\mathbf{x}, u) > 0$.

2.2. Inverse cumulative-probability of censoring weighting estimators

We combine the method of Efron's redistribution-of-mass and that of Shen, Ning and Qin [28] to construct the weighted estimating equations for right-censored and length-biased data. The key idea is to use the inverse cumulative-probability