Statistics and Probability Letters 134 (2018) 150–158

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

The nonparametric quantile estimation for length-biased and right-censored data

Jianhua Shi [a,b,*], Huijuan Ma [c], Yong Zhou [d,e]

[a] School of Mathematics and Statistics, Minnan Normal University, Zhangzhou, China
[b] Fujian Key Laboratory of Mathematical Analysis and Applications, Fujian Normal University, Fuzhou, China
[c] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, 30322, USA
[d] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China
[e] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 18 April 2017
Received in revised form 21 August 2017
Accepted 31 October 2017
Available online 23 November 2017

Keywords:
Length-biased data
Right-censored data
Quantile estimation
Bahadur representation

ABSTRACT

This paper studies the nonparametric estimator of the quantile function under length-biased and right censored data, with the property of length-bias that the residual lifetime share the same distribution as the truncation time. A nonparametric estimator of the quantile function is proposed based on the improved product-limit estimator of distribution function that takes into account the auxiliary information about the length-biased sampling. Asymptotic properties of the estimator are derived, and numerical simulation studies are conducted to assess the performance of the proposed estimator, an application is also given using the Channing house data.

1. Introduction

In the prevalent study for survival analysis, the right-censored time-to-event data is often collected subject to the left-truncation, since the individuals who have experienced the failure event before the recruitment time are not observable. When the incidence of the disease follows a stationary Poisson process, the left-truncated data is called as 'the length-biased data'. As a special case of the left truncation, the length-biased sampling means that the left-truncation variable is uniformly distributed, and the sampling probability of survival time is proportional to its length. See for example, the dementia survey by the Canadian Study of Health and Aging (1994), where over 10,000 elderly Canadians (65 years or older) living in institutions or at home were screened for dementia in 1991. Such survival times arising from length-biased sampling are left truncated by uniformly distributed random truncation times when the incidence of disease onset follows a stationary Poisson process (Winter and Földes, 1988; De Uña-álvarez, 2004). It has been shown that the residual lifetime shares the same distribution to the truncation time in length-biased data (Huang and Qin, 2011). Another real length-biased example is the Channing house dataset (Hyde, 1980), see Section 5.

Due to the fact that ignoring the length-biased information can lead to substantial overestimation of the survival time, the classical survival analysis methods such as the Kaplan–Meier estimator would fail when the sample is length-biased. In recent years, a lot of literatures on the statistical inferences for the length-biased data have been published, such as Addona and Wolfson (2006), Shen et al. (2009), Qin and Shen (2010), Carone et al. (2012), Huang and Qin (2012), Huang, et al. (2012) Chen et al. (2015), Qiu et al. (2016), etc.

Based on the different type of complex data, various methods for the estimation of the quantile function have been proposed. Under the random right censorship model, Aly et al. (1985) studied the quantile process of the product-limit

* Correspondence to: Minnan Normal University, Zhangzhou, Fujian, 363000, China.
E-mail address: shijianhua2011@163.com (J. Shi).

https://doi.org/10.1016/j.spl.2017.10.020
0167-7152/

(PL) estimator via strong approximation methods. For left-truncation data, Gürler et al. (1993) derived a Bahadur type representation and confidence intervals and bands for the quantile function of the pertaining product-limit estimator. Some other literatures on estimation of quantile function are under the left-truncation and right-censoring model. Zhou (1997) deeply discussed the quantile function estimator by empirical process technology, and obtained some precise Bahadur type representations for the estimator. Using the kernel smoothing method, Zhou et al. (2000) obtained asymptotic normality and a Berry–Esseen type bound for the kernel quantile estimator. Wang et al. (2015) proposed a nonparametric maximum likelihood estimator of quantile residual lifetime, showed the asymptotic properties of the estimator, and illustrated some other good statistical property of the estimator through the simulation studies and a real data analysis. Some other literatures include Takeuchi et al. (2006), Zhao et al. (2011), Liang and de Uña-Álvarez (2011) and Zhang and Tan (2015) among others.

In this paper, a nonparametric estimator for the quantile of distribution function (d.f.) for the length-biased and right-censored (LBRC) data is proposed. The proposed quantile estimate function is based on the improved product-limit estimator of distribution function, which takes the auxiliary information about the length-biased sampling scheme into account. The rest of the paper is organized as follows. Section 2 introduces some notations and the proposed quantile estimator. Section 3 presents some large sample properties, such as the consistency, two Bahadur type representations and the asymptotic normality. In Section 4, some simulation studies are conducted to evaluate the behavior of the quantile estimator. In Section 5, a real dataset is analyzed to illustrate the application of our approach. Finally, proofs of the theoretical results are provided in Appendix A, however, several related lemmas and their proofs are postponed to Supplementary materials.

2. Nonparametric quantile estimation

Denote (A^0, T^0, C^0) as a random vector, where T^0 is the survival time of interest with continuous distribution function $F(\cdot)$ and density function $f(\cdot)$, A^0 is the left truncation time with distribution function $F_A(\cdot)$, C^0 is the total censored time. Without loss of generality, it is assumed that they are all nonnegative random variables with C^0 and (A^0, T^0) being mutually independent, but T^0 and C^0 may be dependent.

Define $Y^0 = T^0 \wedge C^0 = \min(T^0, C^0)$ and the indicator of censoring status $\Delta^0 = I(T^0 \le C^0)$. Then in the setting of LBRC, nothing is observed if $Y^0 < A^0$, only the individuals with $Y^0 \ge A^0$ can be observed. Moreover, it is assumed that the survival time T^0 is independent of the calendar time of the disease onset W^0. In this model, a reasonable assumption is $\alpha = P(Y^0 \ge A^0) > 0$, and the calendar time of the disease onset is uniformly distributed on the interval between zero and the sampling time. For simplicity, the superscript ^{0} of letter is dropped to indicate the observed time without special statement, for example, write T as the observed survival time, $V = T - A$ as the corresponding residual time, etc. Then for each individual, the observed random vector is (A, Y, Δ), whose n independent and identically distributed copies are denoted as $\{(a_i, y_i, \delta_i), i = 1, 2, \ldots, n\}$.

Define

$$\tilde{Q}(t) = \frac{1}{n}\sum_{i=1}^{n}[I(a_i \le t) + \delta_i I(y_i - a_i \le t)]$$

and

$$\tilde{K}(t) = \frac{1}{n}\sum_{i=1}^{n}[I(a_i \ge t) + I(y_i - a_i \ge t)].$$

Then the survival d.f. of A can be consistently estimated by Kaplan–Meier type estimator $\tilde{S}_A(t) = \prod_{u\in[0,t]}\{1 - \frac{d\tilde{Q}(u)}{\tilde{K}(u)}\}$. And a nonparametric composite likelihood estimator (Huang and Qin, 2011) for $F(\cdot)$ combining with LBRC information was constructed by

$$1 - \tilde{F}_n(t) = \prod_{u\in[0,t]}\{1 - d\tilde{\Lambda}(u)\} \tag{2.1}$$

where

$$\tilde{\Lambda}(t) = \int_0^t \frac{d\bar{N}(u)}{\tilde{R}_n(u)}, \tilde{R}_n(t) = n^{-1}\sum_{j=1}^{n} I(y_j \ge t) - \tilde{S}_A(t),\ \ \bar{N}(t) = n^{-1}\sum_{j=1}^{n}\delta_j I(y_j \le t).$$

Recently, Shi et al. (2015) successfully established an almost sure representation for the estimator $\tilde{F}_n(t)$, which help us to study the properties of quantile function. As well known, the quantile function for a d.f. $G(\cdot)$ is defined as $G^{-1}(p) = \inf\{u : G(u) \ge p\}, p \in (0, 1)$.

This paper focuses on estimating the quantile function for $F^{-1}(p)$ for some constant $0 < p < 1$ based on LBRC data. A natural estimator defined via the product-limit estimator (2.1) is proposed by

$$\tilde{F}_n^{-1}(p) = \inf\{u : \tilde{F}_n(u) \ge p\}, p \in (0, 1), n = 1, 2, \ldots. \tag{2.2}$$

This is in contrast to the nonparametric maximum likelihood estimator (Vardi, 1989) that has not closed-form in the presence of censoring.