Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 2653 – 2669

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Nonparametric estimation from length-biased data under competing risks

Jacobo de Uña-Álvarez*, Alberto Rodríguez-Casal[1]

Department of Statistics and Operations Research, University of Vigo, 36310 Vigo, Spain

Received 22 February 2005; received in revised form 23 September 2005; accepted 11 January 2006
Available online 2 February 2006

Abstract

A new model for cross-sectional lifetime data is presented. The model is based on the length-bias assumption, and it is adapted to situations in which several types of censoring may occur. The NPMLE of the survival function is derived. An EM-algorithm to approximate the NPMLE is devised. The performance of the introduced estimator is investigated through simulations. A real set of data collected as part of a study on unemployment duration in Spain is used for illustration purposes.

Keywords: Censored data; Cross-section; EM-algorithm; Nonparametric likelihood; Self-consistency

1. Introduction

Cross-sectional sampling of survival data implies that one is only able to observe lifetimes corresponding to individuals "in progress" at a given time point t_0 (the cross-section date). That is, the individuals entering the sample are those who have already experienced the initiation of an event *prior* to time t_0. Since lifetimes observed in this manner tend to be longer than average, in absence of appropriate corrections statistical analysis results in an overestimation of survival. See Wang (1991), Huang and Wang (1995) and Asgharian et al. (2002) for illustration of theses facts and access to basic references.

When the incidence rate of the "disease" under study (onset, generally speaking) may be regarded as constant over time, cross-sectional sampling leads to the so-called length-bias model (same references). The information of this model is relevant since it allows for more efficient estimation of the survival curve and related parameters (Wang, 1989; de Uña-Álvarez, 2004). Asgharian et al. (2002) provide NPMLE of the survival function when the sampled individuals are at risk of censoring from the right. This censoring may emerge because of time limitation in following-up, or by some losing in follow-up (among other reasons). However, the methods proposed in the referred work are inconsistent when censoring times are initially allowed to be smaller than the corresponding left-truncation times (time from onset to cross-section). This situation is found (for example) in the competing risks setup, where more than one final (absorbing) state for the individuals are possible.

* Corresponding author. Tel.: +34 986812492; fax: +34 986812401.
E-mail address: jacobo@uvigo.es (J. de Uña-Álvarez).

[1] Present address: Department of Statistics and Operation Research, University of Santiago de Compostela, Spain

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.01.007

As an example, assume that one is interested in analyzing the time spent between becoming unemployed and finding a new job. As explained more thoroughly in Section 5, the investigation of the unemployment time is typically based on cross-sectional data (that is: sampling from the unemployed stock). Since some individuals will stop searching for a job after some time, the variable of interest is (potentially) censored from the right. Now, this censoring may precede the cross-section date, and the situation described above follows. In this example, the possible final states are represented by "employed" (individuals who find a job) and "out of the labor force" (individuals who stop their search).

The paper is organize as follows: in Section 2, we present the basic notation an assumptions for the model. In Section 3, we derive the NPMLE of the survival distribution under the proposed model. An EM-algorithm is introduced in order to compute the NPMLE. The convergence of the algorithm to a self-consistent estimator and to the NPMLE is discussed. In Section 4, we report some simulations in order to investigate the bias and the variance of the proposed estimator. In Section 5 we illustrate the method using unemployment data. Section 6 includes some concluding remarks.

2. Preliminaries

The sample information is represented by $\left(T_1, X_1, \gamma_1\right), \ldots, \left(T_n, X_n, \gamma_n\right)$, i.i.d. data with the same distribution as (T, X, γ) given $T \leqslant X$. Here, T is the left-truncation time, X is the (possibly censored) lifetime, and γ stands for a censoring indicator. For cross-sectional data sampled at time t_0, the time of onset equals $t_0 - T$. We distinguish between different types of censoring in the following fashion: $X = \min(Y, C, D)$, where Y is the lifetime of ultimate interest; C is a censoring time inherent to the process under investigation (for example, consider the presence of several risks of failure, then Y stands for the main failure time, while C indicates the minimum among the remaining failure times); and D denotes some censoring time induced by the following-up mechanism. In many practical situations, this D may be represented as $D = T + \tau$, where $\tau = t_1 - t_0$ represents the follow-up period duration (from interception to the closing date t_1), but other situations are possible. Then, γ allows to identify the individual's final state:

$$\gamma = \begin{cases} 0 & \text{if } Z < D \text{ and } \delta = 0 \text{ (and hence } X = C), \\ 1 & \text{if } Z < D \text{ and } \delta = 1 \text{ (and hence } X = Y), \\ 2 & \text{if } Z \geqslant D \text{ (and hence } X = D), \end{cases}$$

where $Z = \min(Y, C)$ and $\delta = \mathbb{I}(Y \leqslant C)$. As we will show, distinguishing between C and D censoring times have a real impact in the final estimation procedure.

The model assumptions on the population variable (T, Y, C, D) are:

H1. T, Y, and C are mutually independent;
H2. $P(D \geqslant T) = 1$;
H3. $D - T$ and $(T, Z - T, \delta)$ are independent conditionally on $T \leqslant X$;
H4. $T \sim$ *Uniform* $(0, \tau_{\mathrm{L}})$ for some $\tau_{\mathrm{L}} \geqslant \tau_{\mathrm{H}}$;

where τ_{H} stands for the upper bound of the support of Z. Assumption H1 is a natural extension of the independent left-truncation model in the presence of censoring inherent to the process (resp., competing risks). It just states that the truncation variable gives no information on the process under investigation, for which an independent competing risks model is considered. See de Uña-Álvarez (2004) for further motivation in a related context. Of course, T is independent of (Z, δ) under H1. Assumption H2 is a natural consequence of the definition of D, since the following-up period includes the observed individuals (for which $X \geqslant T$). Importantly, H2 implies that the events $\{T \leqslant X\}$ and $\{T \leqslant Z\}$ coincide with probability one. H3 claims that the residual censoring time $D - T$ for the observed individuals is independent of everything else. Asgharian et al. (2002) implicitly worked under an assumption similar to H3 in a more restrictive model which did not take competing risks into account. We mention that H3 is automatically satisfied whenever censoring induced by following-up is uniquely provoked by time limitation (this is actually the case in some applications). Finally, H4 introduces the stationarity (length-bias) assumption via a uniform distribution for the truncation times. As a technical remark, we mention that H4 restricts the random variable Z to the family of distributions with bounded support.