Computational Statistics and Data Analysis 93 (2016) 294–307

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Nonparametric regression with doubly truncated data

C. Moreira[a,b,*], J. de Uña-Álvarez[a], L. Meira-Machado[b]

[a] *Statistical Inference, Decision, and OR group & Centro de Investigaciones Biomédicas (CINBIO), University of Vigo, Lagoas - Marcosende, 36310 Vigo, Spain*

[b] *Centre of Mathematics and Department of Mathematics and Applications, University of Minho - Campus de Azurém, 4800-058 Guimarães, Portugal*

ARTICLE INFO

Article history:
Received 12 December 2013
Received in revised form 20 March 2014
Accepted 27 March 2014
Available online 4 April 2014

Keywords:
Local polynomial regression
Kernel smoothing
Bandwidth selection
Random truncation
Biased data
Mean squared error

ABSTRACT

Nonparametric regression with a doubly truncated response is introduced. Local constant and local linear kernel-type estimators are proposed. Asymptotic expressions for the bias and the variance of the estimators are obtained, showing the deterioration provoked by the random truncation. To solve the crucial problem of bandwidth choice, two different bandwidth selectors based on plug-in and cross-validation ideas are introduced. The performance of both the estimators and the bandwidth selectors is investigated through simulations. A real data illustration is included. The main conclusion is that the introduced regression methods perform satisfactorily in the complicated scenario of random double truncation.

1. Introduction

Random truncation is a well-known phenomenon which may be present when observing time-to-event data. For example, recruitment of lifetimes through a cross-section induces left-truncation, so larger event times are observed with higher probability. Another example is found when analyzing data which correspond to events taking place before some specific date; in this case, the time-to-event is right-truncated and, therefore, small lifetimes are over-sampled. These two forms of truncation are one-sided, and relatively simple estimators exist. See e.g. Klein and Moeshberger (2003). Nonparametric estimation methods suitable for one-sided random truncation were developed in the last three decades, see for example Woodroofe (1985), Tsai et al. (1987) or Stute (1993) for the estimation of a cumulative distribution function, and for nonparametric regression, Gross and Lai (1996); Iglesias-Pérez and González-Manteiga (1999), Akritas and LaValley (2005) or Ould-Saïd and Lemdani (2006).

In some applications, two-sided (rather than one-sided) random truncation appears. This occurs, for example, when the sample restricts to those individuals with event falling between two particular dates. This is the case of the sample provided by Moreira and de Uña-Álvarez (2010a), who reported data corresponding to children diagnosed of cancer between 1999 and 2003; in this case, the age at cancer diagnosis is doubly truncated, the truncation times have been determined by the two specific limiting dates of observation. The AIDS Blood Transfusion data in Kalbfleisch and Lawless (1989) is another example of such a situation. These data are restricted to those cases diagnosed of AIDS prior to January 1987. For this data

[*] Corresponding author at: Statistical Inference, Decision, and OR group & Centro de Investigaciones Biomédicas (CINBIO), University of Vigo, Lagoas - Marcosende, 36310 Vigo, Spain.

E-mail addresses: carlamgmm@gmail.com, carla@uvigo.es (C. Moreira), jacobo@uvigo.es (J. de Uña-Álvarez), lmachado@math.uminho.pt (L. Meira-Machado).

http://dx.doi.org/10.1016/j.csda.2014.03.017

set, the induction times are doubly truncated because HIV was unknown before 1982, so any case of transfusion-related AIDS before this time would not have been properly classified. See Section 4 for more details. In these two examples, double truncation affects a retrospective time (time from onset to diagnosis). Therefore, right-censoring issues are not present.

Under double truncation, the observational bias is not so evident as in the one-sided truncated setup. Generally speaking, one may say that, under double truncation, large and small inter-event times will be less probably observed. Unlike for one-sided truncation, the nonparametric maximum-likelihood estimator (NPMLE) of the lifetime distribution has no explicit form under double truncation; this complicates the practice and the theoretical developments. We mention that censoring is a problem different from random truncation, because with censored data the researcher has at least some partial information on the censored lifetimes.

Compared to the huge literature devoted to one-sided truncation, there are only few papers devoted to the random double truncation model. Efron and Petrosian (1999) introduced the NPMLE of a cumulative distribution function (df) under double truncation. The asymptotic properties of this NPMLE were further investigated by Shen (2010). Moreira and de Uña-Álvarez (2010b) introduced a semiparametric estimator of a doubly truncated df, while Moreira et al. (2010) presented an R package to compute the NPMLE and confidence bands. Methods for testing a quasi-independence assumption between the lifetime of interest and the truncation times were investigated by Martin and Betensky (2005). Despite the existence of these papers, random double truncation is a phenomenon which is still quite unknown nowadays. In some applications, the goal is the estimation of a smooth curve such as the density function, the hazard rate function, or the regression function. The estimation of these curves crucially depends on the selected bandwidth or smoothing parameter (Wand and Jones, 1995). To the best of our knowledge, the only paper dealing with smoothing methods under double truncation is Moreira and de Uña-Álvarez (2012), who considered kernel density estimation. In this paper we rather focus on nonparametric kernel regression.

Let (X^*, Y^*) be the two-dimensional variable of interest, where Y^* is the lifetime or the inter-event time of main interest, and X^* is a one-dimensional continuous covariate. Since Y^* may represent a transformation of the lifetime (such as the logarithm) in applications, or just a different type of response, we just assume that the support of Y^* is contained in the reals. The goal is the estimation of the regression function $m(x) = E[Y^*|X^* = x]$. Due to the presence of random double truncation, we are only able to observe (X^*, Y^*) when $U^* \le Y^* \le V^*$, where (U^*, V^*) are the truncation times; in that case, (U^*, V^*) are also observed. On the contrary, when $U^* \le Y^* \le V^*$ is violated, nothing is observed. As usual with random truncation, we assume that the truncation times are independent of (X^*, Y^*). Let $(U_1, V_1, X_1, Y_1), \dots, (U_n, V_n, X_n, Y_n)$ be the observed sample, these are iid data with the same distribution as (U^*, V^*, X^*, Y^*) given $U^* \le Y^* \le V^*$, and let $m^T(x) = E[Y_1|X_1 = x]$ be the observed regression function. In general, $m^T(x)$ and the target $m(x)$ will differ; see e.g. Fig. 4, in which these two curves are estimated for the AIDS Blood Transfusion data. This is because of the truncating condition which introduces an observational bias. Similar features were reported in the context of length-biasing, in which the relative probability of sampling a given value of (X^*, Y^*) is proportional to the length of Y^*, see e.g. Cristóbal and Alcalá (2000). In the doubly truncated setup, this relative probability of observing $(X^*, Y^*) = (x, y)$ is given by $G(y) = P(U^* \le y \le V^*)$, since (X^*, Y^*) and (U^*, V^*) are independent. This function G can be estimated from the data by maximum likelihood principles, see the iterative algorithm in Section 2.

The rest of the paper is organized as follows. In Section 2 we introduce the relationship between the observed conditional distribution and that of interest. As it will be seen, by downweighting the (X_i, Y_i)s with the largest values of $G_n(Y_i)$ (where G_n is an estimator for G), we are able to obtain a consistent estimator of $m(x)$. Weighted local polynomial type estimators are considered to this end. We give the asymptotic bias and variance of the weighted Nadaraya–Watson (i.e. local constant) estimator and the weighted local linear kernel estimator, and a confidence interval is introduced. We also propose two different methods to choose the bandwidth for these estimators in practice. In Section 3 we investigate the finite-sample performance of the estimators, and the bandwidth selectors through simulations. Section 4 illustrates all the proposed methods by considering AIDS Blood Transfusion data of Kalbfleisch and Lawless (1989). Finally, in Section 5 we report the main conclusions of our investigation. The technical proofs and details are deferred to the Appendix.

2. The estimators

In this Section we introduce the proposed estimators. We also include the asymptotic results (Section 2.1) and the bandwidth selection algorithms (Section 2.2). Firstly we introduce the needed notations. Let $F(.|x)$ be the conditional df of Y^* given $X^* = x$, so $m(x) = \int_{-\infty}^{\infty} tF(dt|x)$, and let $\alpha(x) = P(U^* \le Y^* \le V^*|X^* = x) = \int_{-\infty}^{\infty} G(t)F(dt|x)$ be the conditional probability of no truncation. It is assumed that $\alpha(x) > 0$. Let $F^*(.|x)$ be the observable conditional df, that is $F^*(y|x) = P(Y_1 \le y|X_1 = x)$. We have

$$F^*(y|x) = \alpha(x)^{-1} \int_{-\infty}^{y} G(t)F(dt|x)$$

for every y. This means that, for a fixed value of the covariate, the response Y^* is observed with a relative probability proportional to $G(Y^*)$. Conversely, provided that $G(t) > 0$ for all t, one may write $F(y|x) = \alpha(x) \int_{-\infty}^{y} G(t)^{-1}F^*(dt|x)$, where $\alpha(x) = 1/\alpha^*(x)$ with $\alpha^*(x) = \int_{-\infty}^{\infty} G(t)^{-1}F^*(dt|x) = E\left[G(Y_1)^{-1}|X_1 = x\right]$. Therefore, the target $m(x)$ is written as $m(x) = m^*(x)/\alpha^*(x)$ where $m^*(x) = E\left[Y_1G(Y_1)^{-1}|X_1 = x\right]$.