Ann Inst Stat Math (2010) 62:323–341
DOI 10.1007/s10463-008-0178-0

Nonparametric estimation of a conditional distribution from length-biased data

Jacobo de Uña-Álvarez · M. Carmen Iglesias-Pérez

Received: 6 April 2006 / Revised: 31 January 2008 / Published online: 25 April 2008

Abstract In this paper we consider the problem of estimating a conditional distribution function in a nonparametric way, when the response variable is nonnegative, and the observational procedure is length-biased. We propose a proper adaptation of the estimate to right-censoring provoked by limitation in following-up. Large sample analysis of the introduced estimator is given, including rates of convergence, limiting distribution, and efficiency results. We show that the length-bias model results in less variance in estimation, when compared to methods based on observed truncation times. Practical performance of the proposed estimator is explored through simulations. Application to unemployment data analysis is provided.

Keywords Cross-sectional sampling · Left-truncation · Regression · Unemployment duration

1 Introduction

Let (X, Y) be a bivariate random variable, where Y stands for a lifetime (or survival time) and X is a covariate. In a number of applications, observation of (X, Y) is performed via cross-sectional sampling. This means that only individuals "in progress" at a single time point (the cross-section date) are observed. As a result, the sampled information is biased with respect to the length of Y, in the sense that the longer

J. de Uña-Álvarez · M. C. Iglesias-Pérez
Department of Statistics and Operations Research, University of Vigo, Vigo, Spain

J. de Uña-Álvarez (✉)
Facultad de Ciencias Económicas y Empresariales, Universidad de Vigo,
Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain
e-mail: jacobo@uvigo.es

the lifetime, the greater the probability of observing it. In this paper we address the problem of estimating the conditional distribution function (df) of Y given X from such length-biased observations.

Specifically, we will propose nonparametric (kernel) estimation of the conditional df

$$F_x(y) = P(Y \leq y \mid X = x) = E\left[1_{\{Y \leq y\}} \mid X = x\right]$$

on the basis of the biased (X, Y)'s. Our method will be adapted to censoring effects due to time limitations in the following-up after cross-section. After all, we will deal with a censored sample of the so-called length-biased (conditional) df of Y, namely

$$F_x^*(y) = \mu_{F_x}^{-1} \int_0^y s \mathrm{d} F_x(s), \quad y \geq 0,$$

where μ_{F_x} stands for the conditional expectation of Y given $X = x$. The model proposed in Sect. 2 motivates this sampling issue through the inclusion of a uniformly distributed left-truncation variable, which is defined as time elapsed from onset to the cross-section date. In this manner the underlying assumptions, which validate the length-bias model in a cross-sectional scenario as long as the practical limitations of the methods to be introduced will be clear.

In order to illustrate the main ideas behind the paper, consider the following application coming from labour economics, which will be more thoroughly discussed in Sect. 4. Let the Y denote unemployment time, and let X represent the individual's age when entering the unemployment stock. Assume that we get information on the pair (X, Y) from the unemployment stock at a given date; that is, the surveyed data correspond to those individuals who are searching for a job at some specific time. Assume that one is interested in the estimation of the unemployment duration df for a given age at onset. In this application, ordinary statistical methods lead to an overestimation of the unemployment duration. This is because the length-bias induced by the cross-sectional sampling, see Lancaster (1990, pp. 88–97), for a deeper discussion. Hence, proper corrections of the length-bias issue are needed. Besides, assume that some of the unemployment spells are not completely observed because the individual following-up stops some time after recruitment. As a consequence, one has to face the problem of accounting for these censoring effects at the right tail of the distribution too.

Many contributions to solving the length-bias problem have appeared in the statistical literature during the last two decades. Starting with Vardi (1982) (see also Vardi 1985), nonparametric estimation of the marginal df of Y is performed by attaching to each lifetime a weight which is inversely proportional to its size. This method leads to the nonparametric maximum-likelihood estimator (NPMLE) of the lifetime df. Smoothing the NPMLE was considered by Jones (1991) for the purpose of estimating a density. In the context of regression with biased responses, Sköld (1999); Wu (2000); de Uña-Álvarez (2003a), and Cristóbal et al. (2004) (among others) made significant contributions. As in the marginal case, consistent estimation of a regression function under length-biased is obtained by attaching to each pair $(X, Y) = (x, y)$ a weight proportional to y^{-1}, see the mentioned references. These ideas can be adapted to the problem of estimating $F_x(y)$, since this target can be regarded as the conditional expectation of the indicator $1_{\{Y \leq y\}}$.

New interesting problems arise when both length-bias and right-censoring phenomena appear in a given data set. For the length-biased and censored scenario, the first relevant contributions in nonparametric df estimation go back to de Uña-Álvarez (2002) and Asgharian et al. (2002), who adapted Vardi's NPMLE to the censored case. The basic model proposed in de Uña-Álvarez (2002) assumes the independence between the Y and the potential censoring time. This assumption may be inappropiated in most applications, since the length-bias issue induces dependencies between lifetimes and censoring times. Asgharian et al. (2002) overcame this difficulty by assuming the independence between the *residual* (length-biased) lifetimes and censoring times (which are defined as forward times from the cross-section date), an assumption that will be more realistic in applications. One drawback of the NPMLE in Asgharian et al. (2002) is that it remains unclear how the covariates can be incorporated in the estimator. As an interesting alternative, under the residual independence model in Asgharian et al. (2002); de Uña-Álvarez (2004) proposed a moment based estimator of the marginal df. This estimator is easily adapted to the presence of covariates, as we will show in this paper. The efficiency of the moment based estimator relative to that of the NPMLE in Asgharian et al. (2002) was investigated in de Uña-Álvarez and Rodríguez-Casal (2006), who concluded that the former estimator may be fairly competitive for light censoring, and that it may be less biased than the NPMLE (depending on the shape of the underlying df).

The rest of the paper is organized as follows. In Sect. 2 we introduce the model and the estimator for F_x, as a proper adaptation of the methods in de Uña-Álvarez (2004) to the covariate setup. In Sect. 3, the asymptotic properties of the estimator are investigated, including rates of convergence, limiting distribution, and efficiency results. In particular, we show that the length-bias assumption results in less variance in estimation, when compared to methods based on observed truncation times (Iglesias-Pérez and González-Manteiga 1999). In Sect. 4, the practical performance of the proposed estimator is explored through simulations. Finally, Sect. 5 illustrates the possibilities of the proposed estimator through unemployment data analysis for the Galician labour market (Spain).

2 Estimation

As announced in the Introduction, we motivate the length-bias model by considering a left-truncation variable (or backwards recurrence time) T, defined as time elapsed from onset to the cross-section date. Under cross-sectional sampling, the Y value is observed if and only if $Y \geq T$ (otherwise, some truncation from the left occurs). In such a case, the T value is also available. We assume that:

H1. Y and T are conditionally independent, given the X;

H2. For each x, the conditional df of T is uniform on an interval containing the support of F_x, and the lower bound of the support of T is zero

Let F_x^* be the conditional df of the observed Y; under H1–H2, we get

$$F_x^*(y) = P(Y \leq y \mid Y \geq T, X = x) = \mu_{F_x}^{-1} \int_0^y s\mathrm{d}F_x(s), \quad y \geq 0,$$

where

$$\mu_{F_x} = \int_0^\infty s\mathrm{d}F_x(s) = E(Y \mid X = x)$$

(assumed to exist). That is, the conditional df of the observed time response Y is the so-called length-biased distribution of F_x (see e.g. Vardi (1982)). Hypothesis H1 is typical in left-truncated scenarios, for which appropriate methods were developed in the eighties (Tsai et al. 1987). On the other hand, H2 implies that the relative probability of observing a spell of a given duration $Y = y$ is proportional to the size of the spell (that is: the sample procedure is length-biased). The convenience of such an assumption when sampling the data *via* cross-sections is discussed in *e.g.* Lancaster (1990) and Wang (1991). Under length-bias, the observed truncation times are useless for inference, since the information on the truncation distribution allows for the construction of more efficient estimators (see Wang 1991; Asgharian et al. 2002; de Uña-Álvarez 2004). Assumption H2 is somehow implicitly assumed in most literature devoted to inference from length-biased data. As a technical remark, we mention that H2 implies that the support of F_x is bounded.

Assume that a given duration Y satisfies $Y \geq T$, so the corresponding spell enters the sample. In order to incorporate censoring effects in the model, we assume in this case that, rather than (T, Y), one observes (T, Z, δ), where $Z = \min(Y, C)$ and $\delta = 1_{\{Y \leq C\}}$. Here, C denotes the potential right-censoring time, and δ is an indicator variable which takes the value 1 for uncensored durations. In this work we deal with censoring induced by limitation in following-up, so

H3. $C = T + \tau$ with probability 1, for a known positive constant τ

This assumption H3 states that individuals still in progress τ time units (the following-up duration) after interception will be automatically censored. This is a special censorship model under which all the censoring times are known in advance. The relevance of such a model was discussed in a number of papers, including Oakes (1993); Chiu (1999) and Wang and Li (2005). Lawless (1982) termed this type of censoring as "Type I censoring" (see Sect. 1.4.1b), see also Kalbfleisch and Prentice (1980, p. 40). We also mention that, in the context of left-truncation and right-censoring, Wang (1991) considered this censoring scheme in her Sect. 2.1, as a special case with some relevance. Under H3, the events $\{Y \geq T\}$ and $\{Z \geq T\}$ coincide with probability 1. This hypothesis was used to analyze unemployment data and survival data in previous papers of the authors, see de Uña-Álvarez (2003b, 2004) and de Uña-Álvarez and Rodríguez-Casal (2007). Despite of its interest in applications, assumption H3 could be relaxed to include possible lost to follow-up cases. This will be discussed later on (see Remark 1).

The sample information will be represented by n independent vectors

$$(X_1, T_1, Z_1, \delta_1), \ldots, (X_n, T_n, Z_n, \delta_n)$$

with the same distribution of (X, T, Z, δ) conditionally on $Z \geq T$. For each i, we have $Z_i = \min(Y_i, C_i) = \min(Y_i, T_i + \tau)$, and $\delta_i = 1_{\{Y_i \leq C_i\}} = 1_{\{Y_i \leq T_i + \tau\}}$. In order to derive a consistent estimator of the conditional df F_x, we introduce the (conditional) subdistribution function of the uncensored, sampled times

$$H_x^{1*}(y) = P\left(Z \leq y, \delta = 1 \mid Z \geq T, X = x\right).$$

Under H1–H3 it is easily seen that

$$H_x^{1*}(y) = \mu_{F_x}^{-1} \int_0^y \left(s 1_{\{s<\tau\}} + \tau 1_{\{s \geq \tau\}}\right) \mathrm{d}F_x(s).$$

Hence,

$$dH_x^{1*}(y) = \mu_{F_x}^{-1} w(y)\, \mathrm{d}F_x(y) \tag{1}$$

where $w(y) = y 1_{\{y<\tau\}} + \tau 1_{\{y \geq \tau\}}$.

Equation (1) leads to

$$F_x(y) = \mu_{Fx} \int_0^y \frac{\mathrm{d}H_x^{1*}(s)}{w(s)}, \tag{2}$$

where besides μ_{Fx} admits the representation

$$\mu_{Fx} = \left(\int_0^\infty \frac{\mathrm{d}H_x^{1*}(s)}{w(s)}\right)^{-1}. \tag{3}$$

Then, estimation of F_x can be performed through a proper estimator of H_x^{1*} by using these identifiability equations. We construct an estimator for F_x by replacing the conditional subdistribution H_x^{1*} in (2) and (3) by the Nadaraya–Watson type smoother

$$\hat{H}_x^{1*}(y) = \sum_{i=1}^n 1_{\{Z_i \leq y, \delta_i = 1\}} B_{ni}(x)$$

where

$$B_{ni}(x) = K\left(\frac{x - X_i}{h}\right)\left(\sum_{j=1}^n K\left(\frac{x - X_j}{h}\right)\right)^{-1},$$

K is a kernel function, and h is the bandwidth (which controls the smoothing level). For simplicity, we assume that X is a continuous univariate random variable. Estimation in the general situation of a multivariate X follows similar lines, although practical problems related to dimension may (and will) appear. From (2) and (3), we come up with

$$\hat{F}_x(y) = \left(\int_0^\infty \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)} \right)^{-1} \int_0^y \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)}$$
$$= \sum_{i=1}^n 1_{\{Z_i \le y\}} \frac{\delta_i B_{ni}(x)}{w(Z_i)} \left(\sum_{j=1}^n \frac{\delta_j B_{nj}(x)}{w(Z_j)} \right)^{-1} \quad (4)$$

where $w(Z_j) = Z_j$ if $Z_j < \tau$ and $w(Z_j) = \tau$ if $Z_j \ge \tau$. We see that the truncation times $T_1, \ldots, T_n$ are not needed for computing $\hat{F}_x$; this is because of the length-bias assumption. Estimation methods conditional on the observed truncation times are possible in our context (Iglesias-Pérez and González-Manteiga 1999), but (as we show in Sect. 3) the pertaining standard errors exceed those of (4).

In the uncensored case ($\tau = \infty$), each time response is weighted by the inverse of its size, plus a factor which measures the closeness of the covariate to the x value. Interestingly, in the uncensored case, this weighting of the data by the inverse of the lifetime is also obtained through maximization of the full likelihood of the biased (X_i, Y_i)'s. As noticed by a referee, in the length-biased scenario, consistent estimation can not be achieved through maximization of a partial likelihood, since the marginal distribution of the covariates contains relevant information on the conditional df. We also point out that $\hat{F}_x$ collapses to the estimator proposed in de Uña-Álvarez (2004) in absence of covariates. The asymptotic properties of (4) are investigated in the next section.

Remark 1 As mentioned, assumption H3 may be relaxed in order to incorporate a random residual censoring time C–T. Assume that (H3*) C–T and (T, Y) are independent conditionally on $T \le Y$, $X = x$. Under H1, H2 and H3* one gets

$$H_x^{1*}(y) = \mu_{F_x}^{-1} \int_0^y \psi_x(s) \mathrm{d}F_x(s)$$

where

$$\psi_x(y) = \int_0^y (1 - R_x^*(y-t))\mathrm{d}t$$

and R^* stands for the conditional df of the residual censoring time $C - T$ given $X = x$. As a result, we get the representation

$$F_x(y) = \int_0^y \frac{\mathrm{d}H_x^{1*}(s)}{\psi_x(s)} \left[\int_0^\infty \frac{\mathrm{d}H_x^{1*}(s)}{\psi_x(s)} \right]^{-1}$$

and one may introduce an estimator for F_x as above, by previously approximating the function ψ_x. An empirical approximation of ψ_x is defined as

$$\widehat{\psi}_x(y) = \int_0^y (1 - \widehat{R}_x^*(y-t))\mathrm{d}t,$$

where $\widehat{R}_x^*$ denotes the conditional product-limit estimator (Dabrowska 1989) of the residual censoring time df. The statistical properties of the estimator corresponding to this extended model are unexplored at the present date. The empirical function $\widehat{\psi}_x(.)$ collapses to $w(.)$ whenever $P(C - T = \tau) = 1$.

3 Large sample results

Before we state the main results concerning the estimator $\hat{F}_x$, we need some extra notation and technical regularity conditions. We assume

(i) X is a real-valued, continuous random variable
Let M^* denote the biased df of the covariate, $M^*(x) = P(X \leq x \mid T \leq Z)$ and let m^* denote the pertaining probability density function. We will refer to the following assumptions too:

(ii.a) There exists and interval $I = [x_1, x_2]$ contained in the support of m^* such that, for some $\delta > 0$,

$$0 < \gamma = \inf \left[m^*(x) : x \in I_\delta\right] < \sup \left[m^*(x) : x \in I_\delta\right] = \Gamma < \infty,$$

with $0 < \delta\Gamma < 1$, where $I_\delta = [x_1 - \delta, x_2 + \delta]$

(ii.b) H1–H2 hold for $x \in I_\delta$, H3 holds, and there exists $a > 0$ such that the support of F_x is contained in $[a, \infty)$ for $x \in I_\delta$

(iii) The functions m^*, $x \mapsto \mu_{F_x}$, and $x \mapsto F_x(y)$ are continuously differentiable on I_δ, and the first derivative of $x \mapsto F_x(y)$ is bounded on I_δ, uniformly for $y > 0$.

(iv) The functions m^*, $x \mapsto \mu_{F_x}$, and $x \mapsto F_x(y)$ are twice continuously differentiable on I_δ, and the second derivative of $x \mapsto F_x(y)$ is bounded on I_δ, uniformly for $y > 0$.

(v) The support of K is contained in $[-1, 1]$, and the total variation of K is less than some $\lambda < \infty$

(vi) $h = (h_n)$ is a deterministic sequence of positive real number satisfying $h \to 0$, $\ln n/ (nh) \to 0$ and $nh^5/ \ln n = O(1)$.

The order of convergence of $\hat{F}_x$ is given in our first Theorem.

Theorem 1 *Under* (i)–(vi)*,*

$$\sup_{y \geq a, x \in I} \left|\hat{F}_x(y) - F_x(y)\right| = O\left(\left(\frac{\ln n}{nh}\right)^{1/2}\right) \quad \textit{with probability 1.}$$

Proof The result is obtained by considering the following decomposition

$$\hat{F}_x(y) - F_x(y) = \left(\int_0^\infty \frac{\mathrm{d}H_x^{1*}(s)}{w(s)}\right)^{-1} \left[\int_0^y \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)} - \int_0^y \frac{\mathrm{d}H_x^{1*}(s)}{w(s)}\right]$$
$$+ \int_0^y \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)} \left[\left(\int_0^\infty \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)}\right)^{-1} - \left(\int_0^\infty \frac{\mathrm{d}H_x^{1*}(s)}{w(s)}\right)^{-1}\right]$$

and by noting that $0 < a \leq w(y) < \tau$ for $y \geq a$, and

$$\sup_{y \geq 0, x \in I} \left| \hat{H}_x^{1*}(y) - H_x^{1*}(y) \right| = O\left(\left(\frac{\ln n}{nh} \right)^{1/2} \right) \quad \text{with probability 1.} \tag{5}$$

Equation (5) is, in its turn, a consequence of Lemma 5 in Iglesias-Pérez and González-Manteiga (1999). □

The asymptotic distributional law of the estimator is investigated now. Our next result states the convergence of the standardized process to a Gaussian distribution.

Theorem 2 *Under* (i)–(vi)*,*

$$\sqrt{nh}\left[\hat{F}_x(y) - F_x(y)\right] - b_x(y) \xrightarrow{d} N\left(0, \sigma_x^2(y)\right) \quad \textit{for } x \in I \quad \textit{and} \quad y \geq a,$$

where

$$b_x(y) = \left(\left(nh^5\right)^{1/2}\right) \frac{\mu_{F_x}}{2m^*(x)} \left(\int u^2 K(u)\, \mathrm{d}u\right) \left(2\Phi'(x) m^{*\prime}(x) + \Phi''(x) m^*(x)\right),$$

where $\Phi(u) = \mu_{F_u}^{-1}\left[F_u(y) - F_x(y)\right]$, *and*

$$\sigma_x^2(y) = \frac{\mu_{F_x}}{m^*(x)} \left(\int K^2(z)\, \mathrm{d}z\right) \left[(1 - 2F_x(y)) \int_0^y \frac{\mathrm{d}F_x(s)}{w(s)} + F_x^2(y) \int_0^\infty \frac{\mathrm{d}F_x(s)}{w(s)}\right].$$

Proof Write

$$\begin{aligned}
\hat{F}_x(y) - F_x(y) &= \left(\int_0^\infty \frac{\mathrm{d}H_x^{1*}(s)}{w(s)}\right)^{-1} \left[\int_0^y \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)} - \left(\int_0^\infty \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)}\right) F_x(y)\right] \\
&\quad + \left[\int_0^y \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)} - \left(\int_0^\infty \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)}\right) F_x(y)\right] \\
&\quad \times \left[\left(\int_0^\infty \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)}\right)^{-1} - \left(\int_0^\infty \frac{\mathrm{d}H_x^{1*}(s)}{w(s)}\right)^{-1}\right] \\
&= S_1 + S_2
\end{aligned}$$

We have $\sqrt{nh} S_2 = o(1)$ with probability 1, since (from Theorem 1)

$$\begin{aligned}
S_2 &= \left[\hat{F}_x(y) - F_x(y)\right] \left(\int_0^\infty \frac{\mathrm{d}H_x^{1*}(s)}{w(s)}\right)^{-1} \left[\int_0^\infty \frac{\mathrm{d}H_x^{1*}(s)}{w(s)} - \int_0^\infty \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)}\right] \\
&= O\left(\frac{\ln n}{nh}\right) \quad \text{with probability 1.}
\end{aligned}$$

So, the aim is to study the asymptotic distribution of $\sqrt{nh}S_1$. Because

$$\sqrt{nh}S_1 = \frac{\mu_{F_x}}{\hat{m}^*(x)}\left[\sqrt{nh}\sum_{i=1}^{n}\frac{1}{nh}K\left(\frac{x-X_i}{h}\right)\frac{\delta_i}{w(Z_i)}\left(1_{\{Z_i\leq y\}}-F_x(y)\right)\right], \qquad (6)$$

where $\hat{m}^*(x)$ is the Parzen–Rosenblatt estimator for the density function of the sampled covariate, and $\hat{m}^*(x)$ converges in probability to $m^*(x)$, it suffices to study the limiting distribution of the second term in (6) which is

$$\begin{aligned}&\sqrt{nh}\sum_{i=1}^{n}\frac{1}{nh}\left\{K\left(\frac{x-X_i}{h}\right)\frac{\delta_i}{w(Z_i)}\left(1_{\{Z_i\leq y\}}-F_x(y)\right)\right.\\&\quad\left.-E\left[K\left(\frac{x-X_i}{h}\right)\frac{\delta_i}{w(Z_i)}\left(1_{\{Z_i\leq y\}}-F_x(y)\right)\right]\right\}\\&\quad+\sqrt{nh}\sum_{i=1}^{n}\frac{1}{nh}E\left[K\left(\frac{x-X_i}{h}\right)\frac{\delta_i}{w(Z_i)}\left(1_{\{Z_i\leq y\}}-F_x(y)\right)\right]=I+II.\end{aligned}$$

We analyze I and II similarly as in Iglesias-Pérez and González-Manteiga (1999), Corollary 3 In this manner, we obtain that

$$\begin{aligned}II=&\left(\left(nh^5\right)^{1/2}\right)\frac{1}{2}\left(\int u^2K(u)\,\mathrm{d}u\right)\\&\times\left(2\Phi'(x)m^{*\prime}(x)+\Phi''(x)m^*(x)\right)+o\left(\left(nh^5\right)^{1/2}\right)\end{aligned}$$

where

$$\Phi(u)=E\left[\frac{\delta}{w(Z)}\left(1_{\{Z\leq y\}}-F_x(y)\right)\mid X=u,T\leq Z\right]=\mu_{F_u}^{-1}\left[F_u(y)-F_x(y)\right].$$

And $I=\sum_{i=1}^{n}\eta_{i,n}(y,x)$, where for each n, $\eta_{i,n}(y,x)$ are zero-mean iid random variables. The application of Central Limit Theorem for triangular arrays [Theorem 7.2 in Billingsley 1968] gives the asymptotic normality of I. (It is used that the functions K and $\xi_i(y,x)=\frac{\delta_i}{w(Z_i)}\left(1_{\{Z_i\leq y\}}-F_x(y)\right)$ are bounded for $x\in I$ and $y\geq a$). We point out that

$$\sum_{i=1}^{n}\mathrm{Var}\left[\eta_{i,n}(y,x)\right]=\left(\int K^2(z)\,\mathrm{d}z\right)V(x)m^*(x)+o(1)<\infty$$

and

$$\begin{aligned}V(x) &= \operatorname{Var}\left[\frac{\delta}{w(Z)}\left(1_{\{Z\le y\}} - F_x(y)\right) \mid X = x, T \le Z\right]\\ &= (1-2F_x(y))\int_0^y \frac{\mathrm{d}H_x^{1*}(s)}{w(s)^2} + F_x^2(y)\int_0^\infty \frac{\mathrm{d}H_x^{1*}(s)}{w(s)^2}\\ &= \mu_{F_x}^{-1}\left[(1-2F_x(y))\int_0^y \frac{\mathrm{d}F_x(s)}{w(s)} + F_x^2(y)\int_0^\infty \frac{\mathrm{d}F_x(s)}{w(s)}\right].\end{aligned}$$

All the previous reasonings complete the proof. □

As mentioned, the length-bias assumption allows for the construction of more efficient estimators, when compared to those based on the truncation times (of product-limit type). In de Uña-Álvarez (2004), a comparison between the (limit) variances pertaining to both approaches was performed in order to illustrate this fact. In the following Theorem, we extend this result to the conditional setup. Introduce

$$s_x^2(y) = \frac{(1-F_x(y))^2}{m^*(x)}\left(\int K^2(z)\mathrm{d}z\right)\left(\int_0^y \frac{\mathrm{d}H_x^{1*}(s)}{C_x^*(s)^2}\right),$$

the limit variance of the conditional product-limit estimator for left-truncated, right-censored data in Iglesias-Pérez and González-Manteiga (1999), where

$$C_x^*(s) = P\left(T \le s \le Z \mid T \le Z, X = x\right).$$

Theorem 3 *Under* (i)–(vi)*, for each $x \in I$ and $y \ge a$ we have $\sigma_x^2(y) \le s_x^2(y)$.*

Proof Under our model assumptions it is easily seen that

$$C_x^*(s) = \mu_{F_x}^{-1} w(s)\left(1 - F_x\left(s^-\right)\right).$$

Use (1) to conclude

$$\int_0^y \frac{\mathrm{d}H_x^{1*}(s)}{C_x^*(s)^2} = \mu_{F_x}\int_0^y \frac{\mathrm{d}F_x(s)}{w(s)\left[1-F_x\left(s^-\right)\right]^2}.$$

So, in order to show $s_x^2(y) - \sigma_x^2(y) \ge 0$ it suffices to prove that

$$\begin{aligned}&(1-F_x(y))^2\int_0^y \frac{\mathrm{d}F_x(s)}{w(s)\left[1-F_x\left(s^-\right)\right]^2}\\ &\quad-\left[(1-F_x(y))^2\int_0^y \frac{\mathrm{d}F_x(s)}{w(s)} + F_x^2(y)\int_y^\infty \frac{\mathrm{d}F_x(s)}{w(s)}\right] \ge 0.\end{aligned}$$

But this inequality follows by arguments similar to those in de Uña-Álvarez (2004), by noting that $w(y)$ is a nondecreasing function. □

An issue of much practical interest is that of the selection of the estimate's smoothing parameter. Typically, the smoothing degree is chosen in order to minimize the mean integrated squared error (MISE), defined as

$$\text{MISE}_\rho\left[\hat{F}_x\right] = E\int\left[\hat{F}_x(y) - F_x(y)\right]^2\rho(y)\mathrm{d}y$$

where the weighting function $\rho(y)$ is used to mitigate endpoint effects. Note that, from Theorem 2, the asymptotic MISE is given by

$$\text{AMISE}_\rho\left[\hat{F}_x\right] = h^4\int\overline{b}_x(y)^2\rho(y)\mathrm{d}y + (nh)^{-1}\int\sigma_x^2(y)\,\rho(y)\mathrm{d}y,$$

where

$$\overline{b}_x(y) = \frac{\mu_{F_x}}{2m^*(x)}\left(\int u^2K(u)\,\mathrm{d}u\right)\left(2\Phi'(x)m^{*\prime}(x) + \Phi''(x)m^*(x)\right).$$

The AMISE is minimized by the bandwidth

$$h_{\text{AMISE}} = \left[\frac{\int\sigma_x^2(y)\,\rho(y)\mathrm{d}y}{4n\int\overline{b}_x(y)^2\rho(y)\mathrm{d}y}\right]^{1/5}.$$

This bandwidth depends on unknown quantities. A possible approach for bandwidth selection is to compute h_{AMISE} by plugging-in some empirical (e.g. kernel smoothing) counterparts for these unknowns. However, in principle there is no optimal candidate for the estimation of this bandwidth, and preliminary simulations of our own have reveal that the unknowns (particularly that involving $\overline{b}_x(y)$) are not always well approximated by their kernel analogues. We propose instead some cross-validation criterion in the spirit of Bowman et al. (1998). Specifically, introduce

$$\text{CV}(h) = \sum_{i=1}^n\frac{\delta_iB_{ni}(x)}{w(Z_i)}\int\left[1_{\{Z_i\le y\}} - \hat{F}_{x,-i}(y)\right]^2\rho(y)\mathrm{d}y\left(\sum_{j=1}^n\frac{\delta_jB_{nj}(x)}{w(Z_j)}\right)^{-1} \quad (7)$$

where $\hat{F}_{x,-i}(y)$ is the leave-one-out version of $\hat{F}_x(y)$ computed from the sample when deleting the datum (X_i, Z_i, δ_i). Note that this is the cross-validation function in Bowman et al. (1998) with the ordinary weights n^{-1} replaced by those associated to (4). These authors establish, in a setup different although somehow related to ours, that their cross-validation function is a particularly good approximation to MISE. In order to see how (7) is motivated, put $\left(X_i^0, Y_i^0\right)$ for a random variable independent of $\left(X_j, T_j, Z_j, \delta_j\right)$, $j \neq i$, such that, given $X_i^0 = x$, Y_i^0 follows the conditional distribution F_x, and note that

$$\begin{aligned}\mathrm{MISE}_{\rho,n-1}\left[\widehat{F}_x\right] &\equiv \int E\left[\widehat{F}_{x,-i}(y) - F_x(y)\right]^2 \rho(y)\mathrm{d}y \\ &= \int E\left\{\left[1_{\{Y_i^0 \leq y\}} - \widehat{F}_{x,-i}(y)\right]^2 \mid X_i^0 = x\right\} \rho(y)\mathrm{d}y \\ &\quad - \int E\left\{\left[1_{\{Y_i^0 \leq y\}} - F_x(y)\right]^2 \mid X_i^0 = x\right\} \rho(y)\mathrm{d}y\end{aligned}$$

where the first integral can be rewritten as $\int_0^\infty \int E\left[1_{\{s\leq y\}} - \widehat{F}_{x,-i}(y)\right]^2 \rho(y)\mathrm{d}y dF_x(s)$, and where the second integral is free of h. Then, it is reasonable to choose h to minimize an estimator of the first integral. Now, by using arguments similar to those in Theorem 2, it is seen that the expected value of (7) satisfies

$$\begin{aligned}&E\left[\mathrm{CV}(h)\right] \\ &\simeq \frac{\mu_{F_x}}{m^*(x)} E\left[\sum_{i=1}^n \frac{1}{nh} K\left(\frac{x - X_i}{h}\right) \frac{\delta_i}{w(Z_i)} \int \left[1_{\{Z_i \leq y\}} - \hat{F}_{x,-i}(y)\right]^2 \rho(y)\mathrm{d}y\right] \\ &= \frac{\mu_{F_x}}{m^*(x)} E\left[\frac{1}{h} K\left(\frac{x - X_1}{h}\right) \frac{\delta_1}{w(Z_1)} \int \left[1_{\{Z_1 \leq y\}} - \hat{F}_{x,-1}(y)\right]^2 \rho(y)\mathrm{d}y\right] \\ &\simeq \int_0^\infty \int E\left[1_{\{s\leq y\}} - \widehat{F}_{x,-1}(y)\right]^2 \rho(y)\mathrm{d}y dF_x(s) \quad \text{as } h \to 0,\end{aligned}$$

which is just what we want to estimate. This motivates the function $CV(h)$ as a smoothing criterion.

Remark 2 The (asymptotically) optimal MISE bandwidth is of the form $h_{AMISE} = cn^{-1/5}$, giving an optimal rate of convergence for the MISE of $n^{-4/5}$. As in the iid setup, faster convergence rates for the MISE can be obtained by using higher-order kernels, according to the degree of smooth of the conditional distribution to be estimated. Specifically, assume that K is a k th-order kernel (see e.g. Wand and Jones 1995, p. 33) and that the function $\psi(u) = \Phi(u)m^*(u)$ (see the proof to Theorem 2 for notation) has k derivatives. Then, the bias term in Theorem 2 equals

$$\frac{\mu_{F_x}}{m^*(x)}\left(\left(nh^{2k+1}\right)^{1/2}\right)(-1)^k \frac{1}{k!}\left(\int u^k K(u)\,\mathrm{d}u\right)\psi^{(k)}(x) + o\left(\left(nh^{2k+1}\right)^{1/2}\right).$$

From this, we get $h_{AMISE} = O(n^{-1/(2k+1)})$ and the optimal MISE is a $O(n^{-2k/(2k+1)})$.

4 Simulation study

In this section we explore the practical performance of the proposed estimator through simulations. Basically, we are concerned with two major objectives. First, our Theorem 3 reveals that, asymptotically, the model information on the truncation distribution results in less variance in estimation. Our simulations will analyze this issue for finite sample sizes, as long as the impact of the smoothing level on the estimate's mean integrated squared error (MISE). Secondly, it remains the question of how to choose the smoothing level given the sampled information. The performance of the bandwidth which minimizes a cross-validation criterion will be briefly discussed.

For the design of the simulation study we followed de Uña-Álvarez and Rodríguez-Casal (2006), after proper adaptation to the conditional setup. Explicitly, we simulated the following conditional model:

Step 1. Draw X according to a Uniform model on the unit interval: $X \sim U(0, 1)$
Step 2. Given $X = x$, draw Y according to the conditional distribution $F_x(y) = y^{\alpha(x)}$, $0 < y < 1$, where $\alpha(x) = 0.75 + x^2$
Step 3. Draw $T \sim U(0, 1)$. If $T > Y$ then reject the datum (X, Y, T) and go to Step 1. Otherwise,
Step 4. Compute $C = T + \tau$ (for a positive constant τ to be indicated below), $Z = \min(Y, C)$ and $\delta = 1_{\{Y \leq C\}}$

Steps 1 and 2 define a model under which the X and the Y are positively correlated. We will consider the problem of recovering F_x from the sampled data, for the special x values 0.25, 0.50, and 0.75, with corresponding $\alpha(x)$ values 0.8125, 1, and 1.3125. These values result in a concave, uniform, and convex conditional distribution function, respectively.

As values for τ, we considered 0.3, 0.5, and 0.7, corresponding to censoring levels (in the observable, truncated population) of 49.9, 25.8, and 9.4%, so the impact of the censoring degree on the estimate's performance could be evaluated. As sample sizes we took $n = 50$ and 100. Since we took three different values for the fixed x (as mentioned), three degrees of censoring, and two sample sizes, up to eighteen different cases are included in our simulations.

For each case, we computed a Monte Carlo approximation of the MISE function, based on M=1000 trials. Specifically, we computed

$$\mathrm{MISE}_\rho(h \mid \hat{F}_x) \equiv \mathrm{MISE}_\rho\left[\hat{F}_x\right] = \frac{1}{M}\sum_{m=1}^{M}\int\left[\widehat{F}_{x,m}(y) - F_x(y)\right]^2 \rho(y)\mathrm{d}y$$

for a grid of bandwidths h (ranging from 0.3 to 1.8 with step 0.05), where $\widehat{F}_{x,m}(y)$ denotes the estimator (4) based on the m -th sample, and $\rho(y)$ is a weighting function which role is to eliminate the endpoint effects (we took $\rho(y) = 1_{\left\{y \in \left[F_x^{-1}(0.05), F_x^{-1}(0.95)\right]\right\}}$ in our computations). In Table 1 we provide, for each situation, the bandwidth $h^*(\hat{F}_x)$ for which $\mathrm{MISE}_\rho(h \mid \hat{F}_x)$ is minimum, as long as the $\mathrm{MISE}_\rho(h^*(\hat{F}_x) \mid \hat{F}_x)$ value. For comparison purposes, the results pertaining to the conditional product-limit estimator in Iglesias-Pérez and González-Manteiga (1999), say $\widetilde{F}_x$, are also reported. The estimator $\widetilde{F}_x$ was proposed as an extension of Dabrowska (1989)'s to cope with the problem of left-truncation, so it provides a consistent estimator of F_x (alternative to $\hat{F}_x$) in our scope. Table 1 includes the efficiency of $\widetilde{F}_x$ relative to $\hat{F}_x$, defined as the

Table 1 Optimal MISE bandwidth and minimum MISE value for the estimators $\hat{F}_x$ and $\widetilde{F}_x$, together with their relative efficiency, along the 1,000 simulated samples

x	τ	MISE($\hat{F}_x$)	$h^*(\hat{F}_x)$	MISE($\tilde{F}_x$)	$h^*(\tilde{F}_x)$	RE($\tilde{F}_x$, $\hat{F}_x$)
$n = 50$						
0.25	0.3	0.0171	0.85	0.0207	0.95	0.8268
	0.5	0.0160	0.80	0.0197	0.95	0.8135
	0.7	0.0157	0.75	0.0195	0.95	0.8050
0.50	0.3	0.0114	0.90	0.0153	1.25	0.7499
	0.5	0.0102	0.85	0.0143	1.25	0.7175
	0.7	0.0098	0.80	0.0140	1.25	0.7283
0.75	0.3	0.0107	0.75	0.0148	0.70	0.6233
	0.5	0.0090	0.70	0.0136	0.70	0.6561
	0.7	0.0083	0.70	0.0134	0.70	0.7183
$n = 100$						
0.25	0.3	0.0108	0.70	0.0137	0.90	0.7870
	0.5	0.0101	0.65	0.0132	0.90	0.7638
	0.7	0.0099	0.65	0.0131	0.90	0.7550
0.50	0.3	0.0067	0.80	0.0091	1.20	0.7405
	0.5	0.0061	0.75	0.0086	1.15	0.7061
	0.7	0.0059	0.75	0.0085	1.15	0.6925
0.75	0.3	0.0066	0.65	0.0089	0.50	0.7383
	0.5	0.0056	0.65	0.0082	0.50	0.6830
	0.7	0.0053	0.65	0.0081	0.50	0.6627

quotient between their respective minimum MISE values:

$$\mathrm{RE}(\widetilde{F}_x, \hat{F}_x) = \frac{\mathrm{MISE}_\rho(h^*(\hat{F}_x) \mid \hat{F}_x)}{\mathrm{MISE}_\rho(h^*(\widetilde{F}_x) \mid \widetilde{F}_x)}.$$

That is, the relative efficiency $\mathrm{RE}(\widetilde{F}_x, \hat{F}_x)$ measures how good is the estimate based on the observed truncation times when compared to that based on the length-bias model, assuming that both of them are computed with their respective optimal smoothing levels.

In Table 1 we see that all the MISE values decrease with an increasing sample size and a decreasing censoring level. Of course, this was expected. The difficulty when estimating the three undelying models F_x ($x = 0.25, 0.50, 0.75$) can be judged from the achieved minimum MISE values. There is an agreement between the two considered estimators in that the hardest distribution to estimate is the concave one ($x = 0.25$). On the other hand, it is clearly seen in Table 1 that the proposed estimator ourperforms that based on the observed truncation times. Hence, as announced by our Theorem 3, there is important information contained in the uniform truncation model when estimating the conditional lifetime df. The relative efficiency of $\widetilde{F}_x$ may be as small as 62.3%, and it takes values below one in all the considered situations.

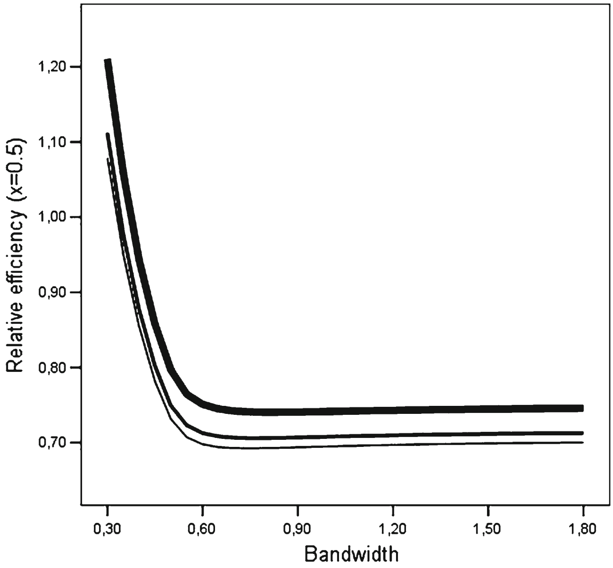

Fig. 1 MISE of $\widehat{F}_x$, normalized by the minimum MISE of $\widetilde{F}_x$, for $x = 0.5$ and $n = 100$, and three different censoring levels: 9% (*thin line*), 26% (*medium*), and 50% (*thick line*)

Figure 1 shows the curve $h \mapsto \text{MISE}_\rho(h \mid \hat{F}_x)$ for $x = 0.50$ and $n = 100$ when normalized by using $\text{MISE}_\rho(h^*(\widetilde{F}_x) \mid \widetilde{F}_x)$ as a denominator (results for $n = 50$ and other x values are similar and they are not displayed here). Hence, the shape of each curve reveals the impact that choosing a wrong bandwidth has in the performance of $\hat{F}_x$, while the achieved values of the curve represent the relative performance of $\widetilde{F}_x$ (based on its optimal smoothing level) when compared to $\hat{F}_x$. We see that there is a serious risk when undersmoothing the estimator, the impact of oversmoothing being less important otherwise. Interestingly, Fig. 1 shows that the proposed estimator $\hat{F}_x$ outperforms $\widetilde{F}_x$ even in the unfair situation in which the latter uses optimal smoothing and the former not. This is a new argument for choosing the estimate pertaining to the lengh-bias model (whenever is true) instead of that based on the observed truncation times.

As mentioned, the second goal in our simulations is to explore the performance of the bandwidth which minimizes the cross-validation function (7), h_{CV} say. This can be measured in two different "scales": (a) in the sense of the bandwidth's closeness to the optimal smoothing degree $h^*(\hat{F}_x)$, and (b) (more importantly) in the sense of the MISE value corresponding to the selected bandwidth. The function $\text{CV}(h)$ was minimized over a grid of h values, ranging from 0.1 to 1.5 with a step of 0.02, and we took the bandwidth h_{CV} as the largest local minimum. We computed h_{CV} along 1,000 simulated samples for the three models (x =0.25, 0.50 and 0.75), sample sizes $n = 50$ and 100, and $\tau = 0.7$. In Table 2 we report the mean, the median and the standard deviation of the cross-validation bandwidth in each situation. In many cases (about one third), the bandwidth was larger than the distance from the particular x to the interval endpoints (these situations corresponding to cross-validation functions with

Table 2 Mean, median and standard deviation of the cross-validation bandwidth along 1,000 simulated samples, for the case $\tau = 0.7$

x	n	Mean(h_{cv})	Median(h_{cv})	SD(h_{cv})
0.25	50	0.52	0.62	0.2426
	100	0.53	0.63	0.2354
0.50	50	0.38	0.44	0.1426
	100	0.37	0.42	0.1415
0.75	50	0.51	0.56	0.2327
	100	0.51	0.53	0.2178

a global minimum at $h = 1.5$). In such cases, the bandwidth was redefined as such a distance (van Keilegom et al. (2001)); the median bandwidths are unaffected by this issue otherwise. From Table 2 it can be seen that the variance of h_{CV} reduces for an increasing sample size; besides, the cross-validation criterion seems to undersmooth the conditional distribution, although the distance between h_{CV} and the optimal MISE bandwidth (see Table 1) gets smaller for $n = 100$. Regarding the MISE of the final estimator, we have computed the increase of the estimate's error (relative to that based on $h^*(\hat{F}_x)$) when using the cross-validation smoothing degree (taken in median). This increase was always below 8%, except for the case $x = 0.50$ (14.3–23.7% depending on n). These results are not that bad if we consider the high risk associated to undersmoothing for $x = 0.50$, see Fig. 1.

5 Application

In this section, we illustrate the proposed estimation method through unemployment data analysis in the Spanish case. Our data source is the survey *Encuesta de Población Activa* (Labour Force Survey) of the Spanish Institute for Statistics, between 1987 and 1997. This survey collects information on about 60,000 homes in Spain, each three months. Each home is followed for the next 18 months. The available information corresponds to those unemployment spells of married women being unemployed at the time of inquiry (1,009 spells in the case of Galicia). The unemployment origin is known for these subjects, because they are asked to provide the date they started searching for a job. The main variable in our application (the Y) is defined as total time (in months) on unemployment. Then, the final event is the first occurrence between the events "finding a job" or "stop searching for a job". Unfortunately, some spells (56% of the sample size) are right-censored at the end of the follow-up period. Right-censored unemployment times correspond to those women still searching for a job 18 months after the first inquiry took place.

Besides censoring effects, some length-bias is present in this data set. The length-bias comes from the fact that observation is restricted to those individuals being in the unemployed stock at the inquiry time. The truncation time T is defined as time from origin to the inquiry date. For these data, the length-bias assumption was informally checked through a graphical comparison between the empirical truncation df (Wang 1991) and the uniform model, showing a good fit. Then, estimation of F_x through (4) was performed, where the covariate X represents age (in years) when entering the

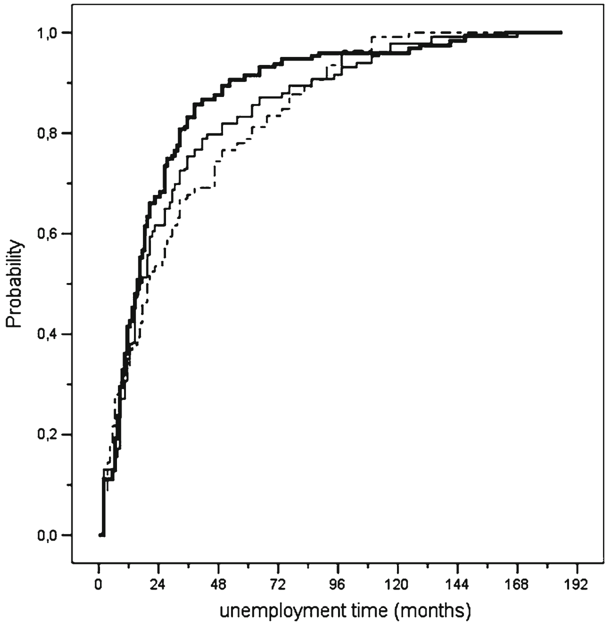

Fig. 2 Nonparametric estimators for the unemployment time distribution conditionally on age: 30 (*thick line*), 35 (*thin line*), and 40 years old (*dashed line*), with common bandwidth $h = 3$

unemployed stock. The value for τ was 18 months (the follow-up period duration), the function K was chosen as the Epanechnikov kernel and (from some preliminary investigation) the bandwidth h was chosen to be 3 years. In practice, a large bandwidth would mask the covariate effects (since all the observations would contribute to the final estimate, regardless the X value), while a too small bandwidth would result in undesirable wiggly estimator for F_x.

Estimators $\hat{F}_x$ for x =30, 35 and 40 years are displayed in Fig. 2. From this figure, it is seen that young unemployed people get out of the unemployment stock earlier. The influence of age on unemployment can be simply measured through the conditional median, defined as

$$\widehat{M}(x) = \hat{F}_x^{-1}(0.5) = \inf\left\{y : \hat{F}_x(y) \geq 0.5\right\}.$$

Median unemployment times computed from these curves report values about 16 ($x = 30$), 17 ($x = 35$) and 20 ($x = 40$) months.

In Fig. 3 we show a pointwise confidence band with asymptotic nominal level of 95% for the 35 years old unemployment group. The confidence limits were computed from the asymptotic variance given in Theorem 2. Note that a plug-in estimator of $\sigma_x^2(y)$ is given by

$$\widehat{\sigma}_x^2(y) = \frac{\widehat{\mu}_{F_x}}{\widehat{m}^*(x)}\left(\int K^2(z)\,\mathrm{d}z\right)\left[\left(1-2\widehat{F}_x(y)\right)\int_0^y \frac{\mathrm{d}\widehat{F}_x(s)}{w(s)} + \widehat{F}_x^2(y)\int_0^\infty \frac{\mathrm{d}\widehat{F}_x(s)}{w(s)}\right]$$

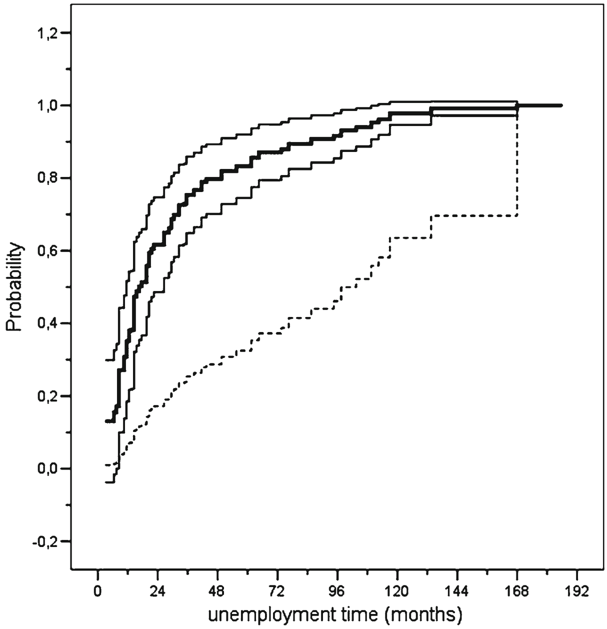

Fig. 3 Proposed nonparametric estimator for the unemployment time distribution conditionally on age: 35 years old (*thick line*), and pointwise 95% confidence band (*thin lines*). For comparison, Dabrowska estimator is included (*dashed line*)

where

$$\widehat{\mu}_{F_x} = \left(\int_0^\infty \frac{\mathrm{d}\hat{H}_x^{1*}(s)}{w(s)} \right)^{-1}.$$

Then, the displayed confidence limits are

$$\widehat{F}_x(y) \pm 1.96 \frac{\widehat{\sigma}_x(y)}{\sqrt{nh}}.$$

Note that the band is wider at the left tail of the distribution; this is a consequence of the length-bias issue, under which small times are less probably observed.

A major mistake that has occurred several times by users of statistics is to naively ignore the length-bias factor. From Fig. 3, one can compare both the estimator for the unbiased survival function ($\hat{F}_x$, solid thick line) and the generalized Kaplan–Meier estimator (Dabrowska 1989, dashed line), for $x = 35$. Note that the latter estimate does not cope with the length-bias issue. Hence, the Kaplan–Meier curve severely underestimates the conditional distribution on the entire range of unemployment durations. This figure illustrates in a practical framework how misleading a naive statistical analysis can be.

Acknowledgments We are very grateful to Gema Álvarez-Llorente and M. Soledad Otero-Giráldez for providing the Galician unemployment data. Thanks also to an Associate Editor and two anonymous referees for suggestions which have improved the presentation of the manuscript. Work supported by the Grant MTM2005-01274 (FEDER funding included) of the Spanish Ministerio de Educación y Ciencia and by the Grants PGIDIT06PXIC300117PN and PGIDIT07PXIB300191PR of the Xunta de Galicia.

References

Asgharian, M., M'Lan, C. E., Wolfson, D. B. (2002). Length-biased sampling with right-censoring: an unconditional approach. *Journal of the American Statististical Association 97*, 201–209.

Billingsley, P. (1968). *Convergence of probability measures*. New York: Wiley.

Bowman, A., Hall, P., Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika*, *85*, 799–808.

Chiu, S. N. (1999). An unbiased estimator for the survival function of censored data. *Communications in Statististics -Theory and Methods*, *28*, 2249–2260.

Cristóbal, J. A., Ojeda, J. L., Alcalá, J. T. (2004). Confidence bands in nonparametric regression with length-biased data. *Annals of the Institute of Statistical Mathematics*, *56*, 475–496.

Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Annals of Statistics*, *17*, 1157–1167.

Iglesias-Pérez, M. C., González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with applications. *Journal of Nonparametric Statistics*, *10*, 213–244.

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika*, *78*, 511–519.

Kalbfleisch, J. D., Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley.

Lancaster, T. (1990). *The econometric analysis of transition data*. Cambridge: Cambridge University Press.

Lawless, F. (1982). *Statistical models and methods for lifetime data*. New York: Wiley.

Oakes, D. (1993). A note on the Kaplan–Meier estimator. *The American Statistician*, *47*, 39–40.

Sköld, M. (1999). Kernel regression in the presence of size-bias. *Journal of Nonparametric Statistics*, *12*, 41–51.

Tsai, W. Y., Jewell, N. P., Wang, M. C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, *74*, 883–886.

de Uña-Álvarez, J. (2002). Product-limit estimation for length-biased censored data. *Test*, *11*, 109–126.

de Uña-Álvarez, J. (2003a). Large sample results under biased sampling when covariables are present. *Statistics & Probability Letters*, *63*, 287–293.

de Uña-Álvarez, J. (2003b). On survival analysis for diseases with stationary incidence. *Proceedings to the ISI International Conference on Environmental Statistics and Health*, Santiago de Compostela, July 16-18, (pp. 1–9). Available at http://isi-eh.usc.es/.

de Uña-Álvarez, J. (2004). Nonparametric estimation under length-biased sampling and Type I censoring: a moment based approach. *Annals of the Institute of Statistical Mathematics*, *56*, 667–681.

de Uña-Álvarez, J., Rodríguez-Casal, A. (2006). Comparing nonparametric estimators for length-biased data. *Communications in Statististics -Theory and Methods*, *35*, 905–919.

de Uña-Álvarez, J., Rodríguez-Casal, A. (2007). Nonparametric estimation from length-biased data under competing risks. *Computational Statistics and Data Analysis*, *51*, 2653–2669.

van Keilegom, I., Akritas, M. G., Veraverbeke, N. (2001). Estimation of the conditional distribution in regression with censored data: a comparative study. *Computational Statistics and Data Analysis*, *35*, 487–500.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics*, *10*, 616–620.

Vardi, Y. (1985). Empirical distributions in selection bias models (with discussion). *Annals of Statistics*, *13*, 178–205.

Wand, M. P., Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall.

Wang, J., Li, Y. (2005). Estimators for survival function when censoring times are known. *Communications in Statistics -Theory and Methods*, *34*, 449–459.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statististical Association*, *86*, 130–143.

Wu, C. O. (2000). Local polynomial regression with selection biased data. *Statistica Sinica*, *10*, 789–817.