Ann. Inst. Statist. Math.
Vol. 56, No. 4, 667–681 (2004)
©2004 The Institute of Statistical Mathematics

NONPARAMETRIC ESTIMATION UNDER LENGTH-BIASED SAMPLING AND TYPE I CENSORING: A MOMENT BASED APPROACH*

JACOBO DE UÑA-ÁLVAREZ

Facultad de Ciencias Económicas y Empresariales, Universidad de Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain, e-mail: jacobo@correo.uvigo.es

(Received October 9, 2002; revised January 21, 2004)

Abstract. Observation of lifetimes by means of cross-sectional surveys typically results in left-truncated, right-censored data. In some applications, it may be assumed that the truncation variable is uniformly distributed on some time interval, leading to the so-called length-biased sampling. This information is relevant, since it allows for more efficient estimation of survival and related parameters. In this work we introduce and analyze new empirical methods in the referred scenario, when the sampled lifetimes are at risk of Type I censoring from the right. We illustrate the method with real economic data.

Key words and phrases: Censoring, cross-sectional, length-biased sampling, stationarity, truncation.

1. Introduction

Observation of lifetime data is often affected by sampling issues, such as truncation and censoring. Much literature is devoted to the problem of estimating a distribution function (df) from left-truncated and right-censored data. Under left-truncation and right-censoring, one observes the random vector (T, Z, δ) iff $Z \geq T$. Here, $Z = \min(Y, C)$ and $\delta = 1_{\{Y \leq C\}}$, Y is the lifetime of ultimate interest, C is the right-censoring time, and T is the truncation time.

Put F for the df of Y and introduce the associated cumulative hazard

$$\Lambda_F(y) = \int_0^y \frac{F(du)}{1 - F(u-)}.$$

Put (T_i, Z_i, δ_i), $1 \leq i \leq n$, for independent observations equally distributed as (T, Z, δ) conditionally on $Z \geq T$. Under the model assumption

$$(1.1) \qquad Y \text{ independent of } (T, C)$$

the conditional nonparametric maximum-likelihood estimator (NPMLE) of $\Lambda_F(y)$ is known to equal

$$\Lambda_{F,n}(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{C_n(Z_i)} 1_{\{Z_i \leq y\}},$$

*Work supported by the Grants PGIDIT02PXIA30003PR and BFM2002-03213.

where

$$C_n(u) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{T_i \le u \le Z_i\}}.$$

The (unique) empirical df associated to $\Lambda_{F,n}$ is given by

$$F_n(y) = 1 - \prod_{i=1}^{n}\left[1 - \frac{\delta_i 1_{\{Z_i \le y\}}}{nC_n(Z_i)}\right], \tag{1.2}$$

the NPMLE of F under (1.1). See Tsai *et al.* (1987), Lai and Ying (1991), Gijbels and Wang (1993), Zhou (1996), and Zhou and Yip (1999) for further motivation and results on these empiricals. Without censoring, the study of F_n goes back to Woodroofe (1985) (see also Stute (1993)). Without truncation, F_n reduces to the Kaplan-Meier estimator (Kaplan and Meier (1958)).

Wang (1991) showed that (1.2) is indeed a conditional NPMLE, in the sense that the function that F_n maximizes is the conditional likelihood that would have been obtained had each subject's censoring time been available even had failure occurred before censoring. The referred paper also provides some interesting discussion on the limitation as well as the interpretation of assumption (1.1). In contrast, the unconditional approach, pioneered by Vardi (1982), proceeds under (partial) knowledge on the truncation distribution. This information on the truncation variable is available in special applications. Conditional and unconditional approaches for estimating F under left-truncation are reviewed and well-discussed in Asgharian *et al.* (2002).

As mentioned, in some applications, the df L of the truncation variable T may be assumed to take a given form. Indeed, some authors have found motivation in renewal processes, Economics, and Epidemiology, for the stationarity (or length-bias) assumption

$$L \sim \text{Uniform}(0, \tau_L)$$

for some $\tau_L > 0$. See for example Winter and Földes (1988), Lancaster (1990), Wang (1991), and van Es *et al.* (2000). The connection between left-truncated data and cross-sectional data (as considered by Wang (1991)) is clearly seen by noting that, in applications with cross-sectional sampling, the truncation variable will be defined as the time elapsed from onset to the sampling date. More details are discussed below.

Knowledge of L is relevant, since the NPMLE of F is no longer (1.2) and estimates more efficient than F_n become available (Wang (1989); de Uña-Álvarez (2001)). When $(0, \tau_L)$ contains the support of F, it is easily seen that the df of Y conditionally on $Z \ge T$ equals the so-called length-biased df (of F)

$$F^*(y) = P(Y \le y \mid Z \ge T) = \mu_F^{-1}\int_0^y uF(du), \tag{1.3}$$

where μ_F denotes the expectation of Y (assumed to exist). Derivation of (1.3) requires not only the uniformity of L, but also some extra assumption such as the independence between Y and T, together with the restriction $P(C \ge T) = 1$. The nice thing regarding (1.3) is that it leads to the "reverse equality"

$$F(y) = \mu_F\int_0^y u^{-1}F^*(du), \qquad \mu_F = \left[\int u^{-1}F^*(du)\right]^{-1}. \tag{1.4}$$

Then, estimation of F is easily introduced by means of an appropriate estimator for its length-biased version F^*.

In the uncensored case, estimation of F^* is given by the ordinary empirical df of the recorded lifetimes, say F_n^*, and the natural estimator of F becomes

$$F_n^V(y) = \frac{\int_0^y u^{-1} F_n^*(du)}{\int u^{-1} F_n^*(du)}, \tag{1.5}$$

see Vardi (1982, 1985) and Horváth (1985). However, under right-censoring estimation of F^* is no longer such a simple issue. Since, conditionally on $Z \geq T$, each lifetime will heavily depend on its corresponding censoring time (just because Y_i and C_i will share the truncation or backwards recurrence time T_i), Kaplan-Meier estimator based on the available (Z_i, δ_i)'s will not converge to F^*.

In this work we propose an extension of (1.5) under censoring from the right. We consider a cross-sectional sampling scenario, as defined in Wang (1991). The truncation variable is the time elapsed from onset to the sampling point, so individuals failing "too soon" cannot be observed. The new estimator is derived in the special case of Type I censoring (see Lawless (1982)), that is, the case in which $C = T + \tau$ for a known fixed positive constant τ. This τ represents the duration of the follow-up period after recruitment. Note that this model is suitable when censoring is uniquely provoked by the end of the follow-up. This happens to be true in many data sets; Section 2 reports an application with unemployment data for which the censoring variable is of the given form. We mention that our proposal is also suitable for the renewal process described in Winter and Földes (1988), who did not incorporate the length-bias relation under right-censoring in the construction of their estimator (thus resulting in less efficiency).

The organization of the paper is as follows. In Section 2, the new estimator is introduced and illustrated with real data. Section 3 is devoted to the main theoretical results concerning the proposed estimator. Consistent estimation of the limiting variance is given, and efficiency of the new empirical relative to that of (1.2) is discussed. In particular, it is shown that the proposed estimator has asymptotic variance less than that corresponding to (1.2). Comments on the extension of the estimate in the presence of covariates are briefly reported in Remark 3. Proofs are deferred to Section 4. Finally, Section 5 summarizes the main conclusions of the paper.

Asgharian *et al.* (2002) derived the maximum-likelihood estimate of F and the accompanying aysmptotic results. These authors based their proposal, rather than on (1.1), on the model assumption

$$C - T \text{ independent of } (Y - T, T) \text{ conditionally on } Z \geq T.$$

Uniformity on T was also assumed. The estimate in the referred work has no explicit form, and must be obtained *via* an iterative algorithm. This results in complicated asymptotics. For example, uniform strong consistency is established under a condition which is difficult to interpret. Furthermore, adaptation of the theory to the covariate setup is unexplored, and does not seem to be an easy task.

2. The estimator

2.1 *Definition*

Put τ_F for the upper bound of the support of F. Assume that

(i) Y is independent of T;
(ii) $T \sim \text{Uniform}(0, \tau_L)$ for some $\tau_L \geq \tau_F$; and
(iii) $C = T + \tau$.

Assumption (i) is typical in left-truncated scenarios. Stationarity (or length-biasing) is incorporated in the model by means of assumption (ii). Finally, (iii) states a Type I censoring scheme. The uncensored case is obtained for $\tau = \infty$ (no limit in following-up). Note that, as mentioned in Introduction, under (i)–(iii) the lifetime df F and the truncated df F^* are connected through the length-bias relation (1.3), respectively (1.4). So, rather than on F, we initially focus on the problem of estimating F^*.

Introduce the function

$$p(y) = E[\delta \mid Y = y, Z \geq T] = P(\delta = 1 \mid Y = y, Z \geq T). \tag{2.1}$$

Under (i)–(iii) it is easily seen that this function satisfies

$$p(y) = 1_{\{y \leq \tau\}} + \frac{\tau}{y} 1_{\{y > \tau\}}.$$

In particular, the probability of uncensoring equals 1 for the recruited individuals with lifetimes taking values on $[0, \tau]$, and decreases towards zero as the lifetime gets larger. This is in accordance with the assumed truncation-censoring scheme, under which lifetimes shorter than τ cannot be censored. The function (2.1) will play a crucial role in the following.

Write

$$\begin{aligned} F^*(y) &= E[1_{\{Y \leq y\}} \mid Z \geq T] = E\left[\frac{E(\delta \mid Y, Z \geq T)1_{\{Y \leq y\}}}{p(Y)} \,\middle|\, Z \geq T\right] \\ &= E\left[\frac{\delta 1_{\{Y \leq y\}}}{p(Y)} \,\middle|\, Z \geq T\right] = E\left[\frac{\delta 1_{\{Z \leq y\}}}{p(Z)} \,\middle|\, Z \geq T\right]. \end{aligned}$$

These equalities suggest estimating $F^*(y)$ through

$$\widehat{F}^*(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i 1_{\{Z_i \leq y\}}}{p(Z_i)} \equiv \sum_{i=1}^{n} W_i 1_{\{Z_i \leq y\}},$$

where

$$W_i = \frac{\delta_i}{np(Z_i)} = \frac{1}{n} 1_{\{Z_i \leq \tau\}} + \frac{\delta_i Z_i}{n\tau} 1_{\{Z_i > \tau\}}, \qquad 1 \leq i \leq n.$$

Note that, for $0 \leq y \leq \tau$, $\widehat{F}^*(y)$ takes the form of the ordinary empirical df of the Z_i's. In other words, the weight attached by $\widehat{F}^*$ to those Z_i falling on $[0, \tau]$ is just $1/n$. This is expected, since (by (iii)) censoring cannot act on that time interval. On the other hand, the weight attached under $\widehat{F}^*$ to each recorded time satisfying $Z_i > \tau$ equals $\delta_i Z_i / n\tau$. As a result, $\widehat{F}^*$ does not jump on the censored times; also, the jump size of $\widehat{F}^*$ at the uncensored times exceeding τ is proportional to Z_i. An heuristic explanation for this can be found in the fact that, under (ii) and (iii), the censoring variable turns out to be uniformly distributed on $(\tau, \tau + \tau_L)$.

Now we use (1.4) in order to introduce

$$\widehat{F}(y) = \frac{\int_0^y u^{-1} \widehat{F}^*(du)}{\int u^{-1} \widehat{F}^*(du)} \equiv \sum_{i=1}^{n} \widetilde{W}_i 1_{\{Z_i \leq y\}}, \tag{2.2}$$

where

$$\widetilde{W}_i = \frac{W_i Z_i^{-1}}{\sum_{j=1}^n W_j Z_j^{-1}}, \qquad 1 \leq i \leq n.$$

Clearly, $\widehat{F}$ is a proper empirical df. In Theorem 3.1 we show its consistency under (i)–(iii). The curve (2.2) jumps at the uncensored data, but the censored ones receive no mass. The normalizing factor $\widetilde{\mu}_F = [\sum_{j=1}^n W_j Z_j^{-1}]^{-1}$ can be regarded as an estimator for the mean lifetime μ_F. We see that the weights of $\widehat{F}$ are obtained from those of $\widehat{F}^*$ by including a factor $\widetilde{\mu}_F Z_i^{-1}$; this is because of the length-biasing, which provokes an overrepresentation of relatively large times in the sample. An alternative expression for the $\widetilde{W}_i$'s is given by

$$\widetilde{W}_i = \frac{\widetilde{\mu}_F}{n}\left[\frac{1}{Z_i} 1_{\{Z_i \leq \tau\}} + \frac{\delta_i}{\tau} 1_{\{Z_i > \tau\}}\right], \qquad 1 \leq i \leq n.$$

We mention that the $\widetilde{W}_i$'s can be computed without knowledge on the truncation times (the T_i's). Of course, assumption (ii) is responsible for this.

In the uncensored situation ($\tau = \infty$), $\widehat{F}^*$ reduces to the ordinary empirical df of the recorded times, and hence $\widehat{F}$ coincides with (1.5). In such a case, $\widehat{F}$ is known to outperform F_n (this is discussed in de Uña-Álvarez (2001)). In the following section, we show that this property holds true under Type I censoring, and several large sample results on $\widehat{F}$ are presented. Now, we give an illustration of the practical possibilities of this estimator by means of real data analysis.

2.2 *Illustration*

For illustration purposes, we consider data concerning unemployment spells (in months) of 1009 Spanish women. The observations correspond to married women living in Galicia, a small geographic region at the Northwest of Spain. These data, obtained from the I. N. E. (the Spanish Institute for Statistics), were collected by means of inquiries at the individuals' homes from 1987 to 1997. The sampled unemployment spells correspond to those women being unemployed at the inquiry time, which varies from individual to individual. In Economic literature, this is typically referred as stock sampling. As discussed in Introduction, this kind of cross-section results in left-truncation. Here, the truncation variable is defined as time elapsed from beginning of unemployment to the inquiry date. When the unemployed population size may be assumed to be constant, suitable modellization is given by the length-bias model (1.3), see Lancaster (1990). Besides, because of the design of the inquiries, each individual was followed during no more than 1.5 years. Indeed, right-censoring was of Type I, with $\tau = 18$ months. 563 spells were censored at the end of the period of observation, giving a censoring percentage of 56%.

In order to check the stationarity assumption that leads to (1.3), we have computed the conditional NPMLE of the truncation df, see Wang (1991). This estimator is displayed in Fig. 1, and strongly suggests the uniformity of the truncation distribution.

Figure 2 provides the survival function associated to estimator (2.2) for these 1009 data. 95% pointwise confidence bands are included. These bands are computed by using the asymptotic normal distribution of $\widehat{F}$, resulting in intervals of the form

$$\left(1 - \widehat{F}(y) \pm 1.96\frac{\widehat{\sigma}(y)}{\sqrt{n}}\right),$$

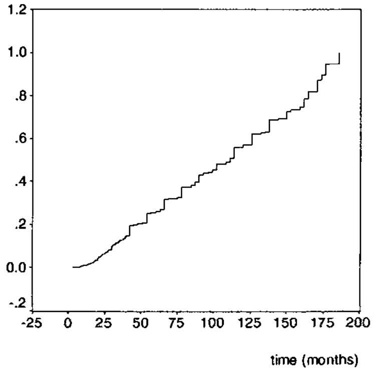

Fig. 1. NPMLE of the truncation distribution for the Galician unemployment data.

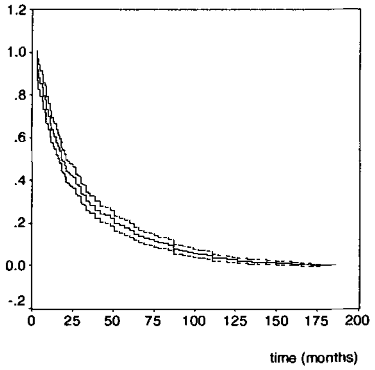

Fig. 2. Estimated survival function (solid line) for the Galician unemployment data, with 95% pointwise confidence bands (dashed lines).

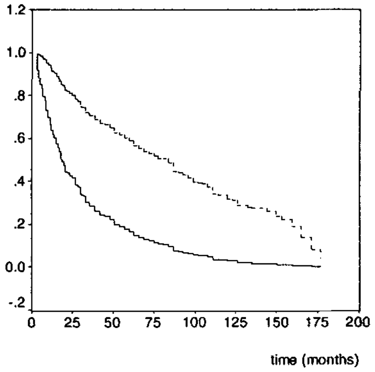

Fig. 3. Estimator of the unbiased survival function (solid line) and Kaplan-Meier survival curve (dashed line) for the Galician unemployment data.

where $\widehat{\sigma}(y)$ stands for the plug-in type estimator of the limiting standard deviation of $\widehat{F}(y)$ to be introduced below, see Section 3.

A major mistake that has occurred several times by users of statistics is to naively ignore the length-bias factor. In Fig. 3 we depict both the estimator for the unbiased survival function and the Kaplan-Meier estimator. Note that the latter estimate does not cope with the length-bias issue. Hence, the Kaplan-Meier curve severely overestimates survival on the entire range of unemployment durations. This figure illustrates in a practical framework how misleading a naive statistical analysis can be.

3. Main results

In this section we present some asymptotic results for the estimator $\widehat{F}$ defined in (2.2). These results guarantee the strong consistency of $\widehat{F}$ and of related curves, such as the empirical mean residual lifetime function (introduced below). Also, an asymptotic representation of (2.2) as a sum of independent, identically distributed random variables is easily obtained, the remainder being negligible at an in-probability rate n^{-1}. This representation immediately gives (a) the asymptotic (normal) distribution of (2.2) and (b) its limit variance. A plug-in type estimator for the variance is introduced. Comparison with the limit variance of (1.2) is performed. Finally, we establish the weak convergence of the empirical process associated to $\widehat{F}$. Throughout this section, hypotheses (i)–(iii) above are assumed to hold. See Section 4 for the proofs of the included results.

THEOREM 3.1. *For each y, we have*

$$\widehat{F}(y) \to F(y) \qquad \textit{with probability } 1.$$

A uniformity argument, immediately gives the following result.

COROLLARY 3.1. *We have*

$$\sup_{0\le y\le \tau_F} |\widehat{F}(y) - F(y)| \to 0 \qquad \textit{with probability } 1.$$

Importantly for applications, Theorem 3.1 may be generalized in the following fashion.

THEOREM 3.2. *For each F-integrable φ, we have*

$$\int \varphi d\widehat{F} \to \int \varphi dF \qquad \textit{with probability } 1.$$

Theorem 3.2 ensures the strong consistency of, e.g., the empirical mean residual time:

$$\widehat{r}(y) = \frac{\int_y^\infty (u-y)\widehat{F}(du)}{1-\widehat{F}(y)} \to E[Y-y \mid Y \ge y]$$

with probability 1. Now, introduce the function

$$\varphi^0(u) = u^{-1}[1_{\{u\le y\}} - F(y)].$$

THEOREM 3.3. *Under* $\int u^{-1}F(du) < \infty$, *we have*

$$\begin{aligned}\widehat{F}(y) - F(y) &= \mu_F \int \varphi^0 d\widehat{F}^* + R_n(y)\\ &= \frac{\mu_F}{n}\sum_{i=1}^{n} \frac{\delta_i}{p(Z_i)Z_i}[1_{\{Z_i \le y\}} - F(y)] + R_n(y),\end{aligned}$$

where $R_n(y) = O_P(n^{-1})$.

Representation in Theorem 3.3, together with the Central Limit Theorem, gives the asymptotic distributional law of the estimate:

$$\sqrt{n}[\widehat{F}(y) - F(y)] \to N(0, \sigma^2(y)) \quad \text{in distribution,}$$

where

$$\begin{aligned}\sigma^2(y) &= \mu_F^2 \operatorname{Var}\left[\frac{\delta_1}{p(Z_1)Z_1}(1_{\{Z_1 \le y\}} - F(y))\right]\\ &= \mu_F[1 - 2F(y)]\int_0^y \frac{F(du)}{p(u)u} + \mu_F F^2(y)\int \frac{F(du)}{p(u)u}.\end{aligned}$$

In the uncensored case, $p \equiv 1$ and we obtain the variance in Vardi (1985). We see that the function $\sigma^2(\cdot)$ depends on F and τ, being independent of the particular choice of $\tau_L(\ge \tau_F)$. At first sight this could be surprising, because the truncation proportion depends on the τ_L value, and an increasing truncation percentage should result in greater variance. But note that the truncation issue results in a smaller sample size, so the standard error of $\widehat{F}(y)$ increases as the truncation proportion approaches to one. Explicitly, under (i)–(iii), censoring and truncation probabilities are given by

$$pc = \frac{1}{\mu_F}\int_\tau^{\tau_F}(1 - F(y))dy = \frac{1}{\mu_F}\int_\tau^{\tau_F}(y - \tau)F(dy) \quad \text{and} \quad pt = 1 - \frac{\mu_F}{\tau_L},$$

respectively.

Figure 4 reports three variance curves $\sigma^2(\cdot)$ for the special df $F(y) = y^2 1_{\{0 \le y \le 1\}}$ and τ values 1, 2/3 and 1/3. For $\tau_L = 1$, we get 33% of truncation and censoring percentages 0% ($\tau = 1$), 15% ($\tau = 2/3$) and 52% ($\tau = 1/3$), but the figure remains valid for each $\tau_L \ge \tau_F$. Note that the case $\tau = 1$ (no censoring) gives the asymptotic variance for (1.5). As expected, the variance $\sigma^2(y)$ increases with censoring. Recall also that a larger value of τ implies more following-up after recruitment.

Estimation of $\sigma^2(y)$ is obtained when replacing the unknown df F by the estimator $\widehat{F}$. This gives

$$\widehat{\sigma}^2(y) = \widetilde{\mu}_F[1 - 2\widehat{F}(y)]\int_0^y \frac{\widehat{F}(du)}{p(u)u} + \widetilde{\mu}_F\widehat{F}^2(y)\int \frac{\widehat{F}(du)}{p(u)u}.$$

Note that Theorem 3.2 guarantees consistency of $\widehat{\sigma}^2(y)$ under $\sigma^2(y) < \infty$.

Now, set $\sigma_0^2(y)$ for the limit variance of (1.2). The following result ensures that efficiency is gained when including the model information (ii) and (iii) in the construction of the estimate.

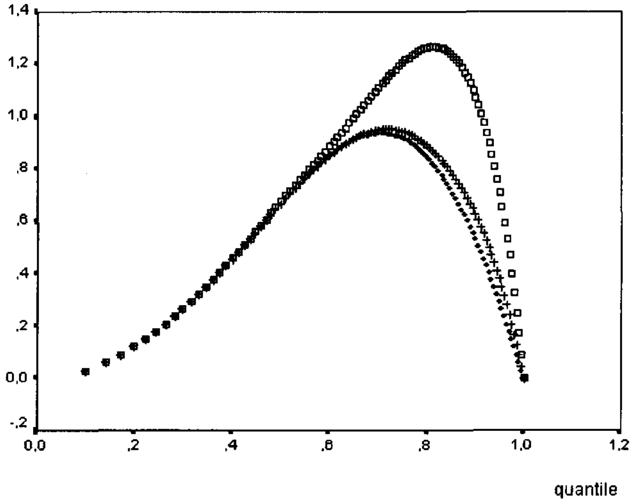

Fig. 4. Asymptotic variance of $\widehat{F}$ as a function of the F-quantiles, for the special case $F(y) = y^2 1_{\{0\le y\le 1\}}$ and τ values 1 (dotted line), 2/3 (crosses) and 1/3 (squares).

Theorem 3.4. *For each y, we have $\sigma^2(y) \le \sigma_0^2(y)$.*

A quantification of how much efficiency is gained through the use of $\widehat{F}(y)$ is given by the asymptotic efficiency of (1.2) relative to $\widehat{F}(y)$, which is defined as

$$ARE(F_n(y), \widehat{F}(y)) = \frac{\sigma^2(y)}{\sigma_0^2(y)}.$$

As claimed by Theorem 3.4, this quantity does not exceed the unity. The ARE, as a function of y, depends on the special form of F and on the following-up period duration τ. But, interestingly, the ARE is not influenced by the particular choice of τ_L, and then situations with different truncation levels may result in the same relative efficiency rate (similarly as above for $\sigma^2(\cdot)$).

We have computed the relative efficiency values for the case in which $F(y) = (y/\tau_F)^2 1_{\{0\le y\le \tau_F\}}$ and $\tau < \tau_F$. This example is important because it shows that, in special situations, using $\widehat{F}(y)$ instead of $F_n(y)$ may result in much more efficiency. In this example, the truncation proportion is given by $1 - 2/3\tau_L$. Also, it can be seen that

$$\lim_{y\to\tau_F} ARE(F_n(y), \widehat{F}(y)) = \frac{1-\tau/\tau_F}{2},$$

and the relative performance of (1.2) at large quantiles gets poorer as $\tau \to \tau_F$. Table 1 reports ARE values for several choices of τ/τ_F (approaching to one) and selected large quantiles of F (75, 85 and 95%).

Representation in Theorem 3.3 is useful for establishing weak convergence too. Clearly, by Theorem 3.3 and the multivariate Central Limit Theorem, the finite dimensional distributions of the process $\sqrt{n}[\widehat{F}(\cdot) - F(\cdot)]$ converge to a multivariate normal distribution, with covariance structure as that given below. In Section 4 we prove that this process is tight.

Table 1. *ARE* values for the model $F(y) = (y/\tau_F)^2 1_{\{0\le y\le\tau_F\}}$. We consider large quantiles of F (75, 85 and 95%) and several rates τ/τ_F.

QUANTILES

		75%	85%	95%
	0.75	.2902	.2052	.1476
τ/τ_F	0.85	.3686	.2049	.1088
	0.90	.5111	.2296	.0921
	0.95	.6834	.4813	.0823

THEOREM 3.5. *Assume that F is continuous. Under $\int u^{-1}F(du) < \infty$, we have that $\sqrt{n}[\widehat{F}(\cdot) - F(\cdot)]$ converges in distribution to a zero-mean Gaussian process with covariance structure given by*

$$\sigma(y,z) = \mu_F \int \frac{[1_{\{u\le y\}} - F(y)][1_{\{u\le z\}} - F(z)]}{p(u)u} F(du).$$

Remark 1. An important question is that of the bias of $\widehat{F}(y)$. Derivation of $E[\widehat{F}(y)]$ is not obvious, because the denominator $\int u^{-1}\widehat{F}^*(du)$ appearing in (2.2) is a random quantity. For the best of our knowledge, this problem has not been addressed so far even in the uncensored case, in which the proposed estimator reduces to (1.5). In de Uña-Álvarez and Saavedra (2004), some simulations show that (1.5) tends to slightly underestimate the target $F(y)$, the bias being more serious for small quantiles. This makes sense, since the truncation issue results in less information on these points.

Remark 2. Under (i)–(iii), the NPMLE of F^* is the maximizer of

$$\prod_{i=1}^{n}\left\{[dF^*(Z_i)]^{\delta_i}\left[\int_{r>Z_i}\frac{dF^*(r)}{r}\right]^{1-\delta_i}\right\}.$$

The NPMLE of F is obtained from this maximizer *via* equation (1.4). Asymptotic properties of this estimator were investigated by Asgharian *et al.* (2002). Numerical methods are required in order to compute the NPMLE. As mentioned in Introduction, our moment-based approach has some advantages when compared to the maximum likelihood criterion. The simple explicit form of estimator (2.2) results in nice asymptotics. Direct comparison to the purely nonparametric estimator in Tsai *et al.* (1987) for left-truncated right-censored data is possible, see Theorem 3.4. Besides, consistent plug-in estimation of (asymptotic) standard errors is easily introduced. A deeper comparison between the NPMLE and (2.2) would be of great interest, but that is out of the scope of the present work.

Remark 3. In many practical cases, a p-dimensional covariate vector X is attached to each individual. In this framework, efforts usually focus on the estimation of the $(p+1)$-variate df $F_{X,Y}(x,y) = P(X\le x, Y\le y)$. When (X,Y) is independent of the truncation variable, a simple extension of model (i)–(iii) and estimate (2.2) is possible. The extended model is particularly useful for regression analysis, when the (time) response is subject to length-biasing and Type I censoring. In a technical report, de

Uña-Álvarez (2003a) showed that weighting through the $\widetilde{W}_i$'s in (2.2) allows for consistent estimation in the presence of covariates. In the uncensored case, related problems under general sampling bias were investigated in de Uña-Álvarez (2003b).

4. Proofs

Since Theorem 3.1 can be regarded as a particular case of Theorem 3.2, we state the proof for the latter result.

Proof of Theorem 3.2. Apply the Strong Law to get

$$\int u^{-1}\varphi(u)\widehat{F}^*(du) \to \int u^{-1}\varphi(u)F^*(du)$$

with probability 1, provided that the limit exists. But the length-bias relation (1.3) gives

$$\int u^{-1}\varphi(u)F^*(du) = \mu_F \int \varphi dF,$$

so F-integrability of φ (together with the existence of μ_F) is enough for consistency purposes. Similarly,

$$\int u^{-1}\widehat{F}^*(du) \to \int u^{-1}F^*(du)$$

with probability 1, and

$$\int \varphi d\widehat{F} = \frac{\int u^{-1}\varphi(u)\widehat{F}^*(du)}{\int u^{-1}\widehat{F}^*(du)} \to \frac{\int u^{-1}\varphi(u)F^*(du)}{\int u^{-1}F^*(du)} = \int \varphi dF$$

holds almost surely. □

Proof of Theorem 3.3. Direct algebra gives

$$\widehat{F}(y) - F(y) = \mu_F \int \varphi^0 d\widehat{F}^* + R_n(y)$$

where

$$R_n(y) = -\mu_F[\widehat{F}(y) - F(y)]\left[\frac{1}{\widetilde{\mu}_F} - \frac{1}{\mu_F}\right].$$

The Central Limit Theorem ensures

$$\sqrt{n}\left[\frac{1}{\widetilde{\mu}_F} - \frac{1}{\mu_F}\right] = \sqrt{n}\left[\int u^{-1}\widehat{F}^*(du) - \int u^{-1}F^*(du)\right] = O_P(1)$$

under $\int u^{-2}F^*(du) < \infty$. This condition is equivalent to $\int u^{-1}F(du) < \infty$. The existence of $\int u^{-1}F(du)$ together with the delta method give

$$\sqrt{n}[\widehat{F}(y) - F(y)] = \sqrt{n}\left[\frac{\int_0^y u^{-1}\widehat{F}^*(du)}{\int u^{-1}\widehat{F}^*(du)} - \frac{\int_0^y u^{-1}F^*(du)}{\int u^{-1}F^*(du)}\right] = O_P(1),$$

and hence the theorem follows. □

Proof of Theorem 3.4. The asymptotic variance of (1.2) is known to equal

$$\sigma_0^2(y) = [1 - F(y)]^2 \int_0^y \frac{H_1^*(du)}{C(u)^2},$$

where

$$H_1^*(u) = P(Z \le u, \delta = 1 \mid Z \ge T) \quad \text{and} \quad C(u) = P(T \le u \le Z \mid Z \ge T).$$

Now, under (i)–(iii) it can be checked that

$$H_1^*(u) = \mu_F^{-1} \int_0^u p(v)vF(dv) \quad \text{and} \quad C(u) = \mu_F^{-1}p(u)u[1 - F(u-)].$$

Hence,

$$\sigma_0^2(y) = \mu_F[1 - F(y)]^2 \int_0^y \frac{F(du)}{w(u)[1 - F(u-)]^2},$$

where we put $w(u) = p(u)u$. Write

$$\begin{aligned}&\mu_F^{-1}[\sigma_0^2(y) - \sigma^2(y)]\\ &= (1 - F(y))^2 \left[\int_0^y (1 - F)^{-2}w^{-1}dF - \int_0^y w^{-1}dF\right] - F^2(y)\int_y^\infty w^{-1}dF.\end{aligned}$$

But

$$\begin{aligned}\int_0^y (1 - F)^{-2}w^{-1}dF - \int_0^y w^{-1}dF &= \int_0^y \frac{F(2 - F)dF}{w(1 - F)^2}\\ &\ge \frac{1}{w(y)} \int_0^y \frac{F(2 - F)dF}{(1 - F)^2}\\ &\qquad \text{(because } w \text{ is nondecreasing)}\\ &= \frac{F^2(y)}{w(y)(1 - F(y))}.\end{aligned}$$

It follows

$$\begin{aligned}\mu_F^{-1}(\sigma_0^2(y) - \sigma^2(y)) &\ge \frac{F^2(y)(1 - F(y))}{w(y)} - F^2(y)\int_y^\infty w^{-1}dF\\ &\ge \frac{F^2(y)(1 - F(y))}{w(y)} - \frac{F^2(y)}{w(y)}\int_y^\infty dF\\ &\qquad \text{(because } w \text{ is nondecreasing)}\\ &= 0\end{aligned}$$

as we wanted to show. □

Proof of Theorem 3.5. We will show that the process $\sqrt{n}[\widehat{F}(\cdot) - F(\cdot)]$ is tight. Write

$$\begin{aligned}\sqrt{n}[\widehat{F}(y) - F(y)] &= \widetilde{\mu}_F\sqrt{n}\left[\int_0^y u^{-1}\widehat{F}^*(du) - \int_0^y u^{-1}F^*(du)\right]\\ &\quad + \sqrt{n}[\widetilde{\mu}_F - \mu_F]\int_0^y u^{-1}F^*(du).\end{aligned}$$

Since $\sqrt{n}[\widetilde{\mu}_F - \mu_F]$ is bounded in probability and $y \mapsto \int_0^y u^{-1}F^*(du)$ is continuous, the second term is tight. Since $\widetilde{\mu}_F$ consistently estimates μ_F, it suffices proving tightness for

$$\begin{aligned} P_n(y) &= \sqrt{n}\left[\int_0^y u^{-1}\widehat{F}^*(du) - \int_0^y u^{-1}F^*(du)\right] \\ &= n^{-1/2}\sum_{i=1}^n \left\{\frac{\delta_i 1_{\{Z_i \le y\}}}{p(Z_i)Z_i} - \int_0^y u^{-1}F^*(du)\right\}. \end{aligned}$$

For $y_1 \le y \le y_2$,

$$E[|P_n(y) - P_n(y_1)|^2|P_n(y_2) - P_n(y)|^2] = E\left[\left(n^{-1/2}\sum_{i=1}^n \alpha_i\right)^2\left(n^{-1/2}\sum_{i=1}^n \beta_i\right)^2\right],$$

where

$$\begin{aligned} \alpha_i &= \frac{\delta_i 1_{\{y_1 < Z_i \le y\}}}{p(Z_i)Z_i} - \int_{y_1}^y u^{-1}F^*(du), \\ \beta_i &= \frac{\delta_i 1_{\{y < Z_i \le y_2\}}}{p(Z_i)Z_i} - \int_y^{y_2} u^{-1}F^*(du). \end{aligned}$$

Since $E(\alpha_i) = E(\beta_i) = 0$, we get by Cauchy-Schwarz

$$E\left[\left(n^{-1/2}\sum_{i=1}^n \alpha_i\right)^2\left(n^{-1/2}\sum_{i=1}^n \beta_i\right)^2\right] \le n^{-2}[nE(\alpha_1^2\beta_1^2) + 3n(n-1)E(\alpha_1^2)E(\beta_1^2)].$$

Now, since

$$\begin{aligned} \alpha_1^2 &\le 2\left\{\frac{\delta_1 1_{\{y_1 < Z_1 \le y\}}}{p(Z_1)^2 Z_1^2} + \left[\int_{y_1}^y u^{-1}F^*(du)\right]^2\right\}, \\ \beta_1^2 &\le 2\left\{\frac{\delta_1 1_{\{y < Z_1 \le y_2\}}}{p(Z_1)^2 Z_1^2} + \left[\int_y^{y_2} u^{-1}F^*(du)\right]^2\right\}, \end{aligned}$$

we get

$$\begin{aligned} E(\alpha_1^2\beta_1^2) \le 4\Bigg\{&\left[\int_y^{y_2} u^{-1}F^*(du)\right]^2 \int_{y_1}^y p(u)^{-1}u^{-2}F^*(du) \\ &+ \left[\int_{y_1}^y u^{-1}F^*(du)\right]^2 \int_y^{y_2} p(u)^{-1}u^{-2}F^*(du) \\ &+ \left[\int_{y_1}^y u^{-1}F^*(du)\right]^2 \left[\int_y^{y_2} u^{-1}F^*(du)\right]^2\Bigg\}, \\ E(\alpha_1^2)E(\beta_1^2) \le 4\Bigg\{&\int_{y_1}^y p(u)^{-1}u^{-2}F^*(du) \int_y^{y_2} p(u)^{-1}u^{-2}F^*(du) \\ &+ \int_{y_1}^y p(u)^{-1}u^{-2}F^*(du) \left[\int_y^{y_2} u^{-1}F^*(du)\right]^2 \end{aligned}$$

$$+\int_y^{y_2} p(u)^{-1}u^{-2}F^*(du)\left[\int_{y_1}^{y} u^{-1}F^*(du)\right]^2$$
$$+\left[\int_{y_1}^{y} u^{-1}F^*(du)\right]^2\left[\int_{y}^{y_2} u^{-1}F^*(du)\right]^2\Bigg\}.$$

Set $\psi_1(y)=\int_0^y u^{-1}F^*(du)$ and $\psi_2(y)=\int_0^y p(u)^{-1}u^{-2}F^*(du)$, which exist under assumption $\int u^{-1}F(du)<\infty$. We have

$$E[|P_n(y)-P_n(y_1)|^2|P_n(y_2)-P_n(y)|^2]\leq K[\psi_1(y_2)-\psi_1(y_1)]^2+12[\psi_2(y_2)-\psi_2(y_1)]^2,$$

where

$$K=32\psi_2(\infty)+16\psi_1(\infty)^2.$$

Since ψ_1 and ψ_2 are nondecreasing, continuous functions, the result follows from Theorem 15.6 in Billingsley (1968). □

5. Conclusions

This work presents new empirical methods for estimating survival and related parameters from cross-sectional lifetime surveys. This kind of surveys typically suffer from left-truncation and censoring from the right. The proposed estimators are based on a stationarity (or length-bias) assumption, that is, the uniform distribution is assumed to hold for the truncation variable. In this case, simple empirical methods can be proposed, provided that censoring is of Type I (as it happens in many data sets). In applications, stationarity may fail when the period of the study includes recession and/or expansion cycles.

The new methods have actual advantages when compared to other existing ones. First, more efficiency is gained by including the truncation-censoring information in the construction of the estimates. Second, the formal simplicity of the proposed techniques allows for easy asymptotic analysis (including variance estimation), and straightforward extension to the covariate framework. Importantly, both the length-bias and the Type I censoring assumption can be easily checked in practice. As a possible disadvantage, it should be mentioned that the proposed estimator may be less efficient than the NPMLE in Asgharian *et al.* (2002), particularly under heavy censoring. However, this relative efficiency issue is a topic that requires more investigation.

Acknowledgements

Thanks to Gema Álvarez Llorente for providing the economic data used in Section 2. Special thanks to M. Carmen Iglesias-Pérez for her assistance with the illustration. The author also acknowledges comments and suggestions by two anonymous referees.

References

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

Billingsley, P. (1968). *Convergence of Probability Measures*, Wiley, New York.

de Uña-Álvarez, J. (2001). On efficiency under selection bias caused by truncation (unpublished).

de Uña-Álvarez, J. (2003*a*). Empirical estimation under length-bias and Type I censoring, Reports in Statistics and Operations Research, No. 03-04, Department of Statistics and Operations Research, University of Santiago de Compostela.

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

de Uña-Álvarez, J. and Saavedra, A. (2004). Bias and variance of the nonparametric MLE under length-biased censored sampling: A simulation study, *Communications in Statistics-Simulation and Computation*, **33**, 397–413.

Gijbels, I. and Wang, J. L. (1993). Strong representation of the survival function estimator for truncated and censored data with applications, *Journal of Multivariate Analysis*, **47**, 210–229.

Horváth, L. (1985). Estimation from a length-biased distribution, *Statistics and Decisions*, **3**, 91–113.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457–481.

Lai, T. L. and Ying, Z. (1991). Estimating a distribution function with truncated and censored data, *The Annals of Statistics*, **19**, 417–442.

Lancaster, T. (1990). *The Econometric Analysis of Transition Data*, Cambridge University Press, Cambridge.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

Stute, W. (1993). Almost sure representations of the product-limit estimator for truncated data, *The Annals of Statistics*, **21**, 146–156.

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

van Es, B., Klaassen, C. A. J. and Oudshoorn, K. (2000). Survival analysis under cross-sectional sampling: Length bias and multiplicative censoring, *Journal of Statistical Planning and Inference*, **91**, 295–312.

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

Vardi, Y. (1985). Empirical distributions in selection bias models. With discussion by C. L. Mallows, *The Annals of Statistics*, **13**, 178–205.

Wang, M.-C. (1989). A semiparametric model for randomly truncated data, *Journal of the American Statistical Association*, **84**, 742–748.

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

Winter, B. B. and Földes, A. (1988). A product-limit estimator for use with length-biased data, *Canadian Journal of Statistics*, **16**, 337–355.

Woodroofe, M. (1985). Estimating a distribution function with truncated data, *The Annals of Statistics*, **13**, 163–177.

Zhou, Y. (1996). A note on the TJW product-limit estimator for truncated and censored data, *Statistics and Probability Letters*, **26**, 381–387.

Zhou, Y. and Yip, P. S. F. (1999). A strong representation of the product-limit estimator for left truncated and right censored data, *Journal of Multivariate Analysis*, **69**, 261–280.