Universidade de Vigo

A Semiparametric Estimator of Survival for Doubly Truncated Data

Carla Moreira and Jacobo de Uña Álvarez

Report 09/04

Discussion Papers in Statistics and Operation Research

Departamento de Estatística e Investigación Operativa

Facultade de Ciencias Económicas e Empresariales
Lagoas-Marcosende, s/n · 36310 Vigo
Tfno.: +34 986 812440 - Fax: +34 986 812401
http://eioweb.uvigo.es/
E-mail: depc05@uvigo.es

Universidade de Vigo

A Semiparametric Estimator of Survival for Doubly Truncated Data

Carla Moreira and Jacobo de Uña Álvarez

Report 09/04

Discussion Papers in Statistics and Operation Research

Imprime: GAMESAL

Edita: 

Universidade de Vigo
Facultade de CC. Económicas e Empresariales
Departamento de Estatística e Investigación Operativa
As Lagoas Marcosende, s/n 36310 Vigo
Tfno.: +34 986 812440

I.S.S.N: 1888-5756

Depósito Legal: VG 1402-2007

A semiparametric estimator of survival for doubly truncated data

Carla Moreira [1]
carlamgmm@gmail.com

Jacobo de Uña-Álvarez[1]
jacobo@uvigo.es

[1] Department of Statistics and O.R.
Faculty of Economics, University of Vigo
Vigo 36310
Spain

Abstract

Doubly truncated data are often encountered in the analysis of survival times, when the sample reduces to those individuals with terminating event falling on a given observational window. In this paper we assume that some information on the bivariate distribution function (df) of the truncation times is available. More specifically, we represent this information by means of a parametric model for the joint df of the truncation times. Under this assumption, a new semiparametric estimator of the lifetime df is derived. We obtain asymptotic results for the new estimator, and we show in simulations that it may be more efficient than the Efron-Petrosian NPMLE. Data on the age at diagnosis of childhood cancer in North Portugal are analyzed with the new method.

Key Words: Double truncation, Observational bias, Nonparametric estimation, Survival analysis.

1 Introduction

Randomly truncated data appear in a number of fields, including Epidemiology, survival analysis, economics and Astronomy. Under left-truncation, only individuals with lifetime exceeding a random time (the truncation time) are observed, and ignoring this issue results in a severe bias in estimation. A typical example involving left-truncation is that of cross-sectional sampling, for which the sample reduces to the individuals in progress at a given date. Right-truncation is a similar phenomenon, and studies on AIDS incubation (among others) are known to suffer from this type of selection issue. This is because typically AIDS databases consist only on those individuals diagnosed prior some specific date.

Nonparametric methods for one-sided (left or right) truncated data were introduced in the seminal paper [1], see also [2] and [3]. Some authors have pointed out that available information on the truncation time allows for the construction of more efficient estimators. See for example [4], [5] and [6]. This is an interesting feature in truncated data analysis which is not true for other types of observational biases, as those associated to random censoring. Indeed, in the random censoring model, information on the distribution of the censoring time is useless, since it does not allow for less variance in estimation.

Doubly truncated data are sometimes encountered. Double truncation means that only lifetimes falling on an observable random interval will be recruited. Efron and Petrosian [7] motivated this problem with quasar data, and they introduced the nonparametric maximum likelihood estimator (NPMLE) under double truncation too. Also, Bilker and Wang [8] noticed that time from HIV infection to AIDS diagnosis can be treated as doubly truncated because, besides the right-truncation issue discussed above, HIV was unknown before 1982, thus leading to a left-truncated setup. The NPMLE for doubly truncated data was revisited in [9], who formally established its uniform consistency and convergence to a normal. Moreira and de Uña-Álvarez [10] introduced a bootstrap approximation for the NPMLE.

For further motivation and in order to help intuition, let us introduce in some detail the childhood cancer data analyzed in Section 3. These data concern all the cases of childhood cancer diagnosed in North Portugal between January 1st 1999 and December 31st 2003. Put X^* for the age at diagnosis, and let V^* be the time from birth to the end of recruitment. Then, the sampling information reduces to those (X^*, V^*) values satisfying $U^* \leq X^* \leq V^*$, where $U^* = V^* - 5$ and where time is measured in years. Here, the truncation interval $[U^*, V^*]$ varies from individual to individual, since it is determined by the (random) individual's birthdate. The same truncation pattern will be found in any application in which the sample reduces to those subjects with terminating event in a fixed observational window. Now, if we are interested in the estimation of the distribution function (df) of X^*, computation of the ordinary empirical df can not be recommended in general. This is because each value x of X^* is observed with a probability given by

$$P(U^* \leq X^* \leq V^* | X^* = x) = P(U^* \leq x \leq V^*)$$

(assuming independence between (U^*, V^*) and X^*), which is clearly influenced by the x-value. Roughly speaking, relatively small and large values of X^* may be less frequently observed, hence inducing a sampling bias which is immediately transferred to a systematic bias of the ordinary empirical df. Of course, for specific choices of the distribution of (U^*, V^*) this observational bias can be more or less severe, and this will be explored in much more detail in Section 3. Similar problems arise when the interest is focused on the df of V^*, which for the childhood cancer data is linked to the so-called birth process of the diseased

population. Note that, since $U^* = V^* - 5$, the pair $(X^*, X^* + 5)$ plays the role of truncation interval for V^*.

In this paper we introduce a new estimator under double truncation which makes use of some available information on the df of the pair of truncation times. The definition of the estimator and its main properties are given in Section 2. This new estimator is semiparametric, since some parametric family of df's is assumed for the truncation times, while nothing is assumed about the lifetime df. This situation can be realistic in practice; for example, for the childhood cancer data, the birth dates of the individuals developing the disease could be assumed in principle to follow a stationary process, thus leading to a uniform distribution of U^* (resp. of V^*). Interestingly, unlike the NPMLE, the new estimator has explicit form, which facilitates its usage and analysis. More specifically, we introduce an empirical approximation of the standard error of the estimator and we use it to construct confidence limits. Simulations reported in Section 3 show that the new estimator may behave more efficiently than the Efron-Petrosian NPMLE. Section 4 is devoted to the analysis of the childhood cancer data, while the technical details are deferred to Section 5.

2 The estimator. Asymptotic results

First we introduce some notation. Let X^* be the lifetime of ultimate interest and let (U^*, V^*) be the pair of truncation times, so we observe the triplet (U^*, X^*, V^*) if and only if $U^* \leq X^* \leq V^*$. We assume that (U^*, V^*) is independent of X^*. Let F denote the df of X^*, and let K be the joint df of (U^*, V^*). We denote by (U_i, X_i, V_i), $1 \leq i \leq n$, the observed data. Throughout this paper we assume that K belongs to a parametric family of df's, $\{K_\theta\}_{\theta \in \Theta}$ say, where θ is a vector of parameters and Θ stands for the parametric space. Also, θ_0 will denote the 'true' parameter.

Under the described double truncation scenario, the relative probability of observing a lifetime x is proportional to

$$G(x;\theta) = \int_{\{u \leq x \leq v\}} K_\theta(du, dv).$$

In order to see this, let F^* be the df of the observed lifetimes. Then,

$$F^*(x) = P(X^* \leq x | U^* \leq X^* \leq V^*) = P(\theta)^{-1} \int_0^x G(t;\theta) dF(t),$$

where $P(\theta) = P(U^* \leq X^* \leq V^*) = \int_0^\infty G(t;\theta) dF(t)$. As a consequence, the quotient $F^*(dx)/F(dx)$ (i.e. the relative probability of sampling x) reduces to $P(\theta)^{-1} G(x;\theta)$. Assume that $G(x;\theta)$ is strictly positive on the support of X^*. One immediately obtains the reversed equation

$$F(x) = P(\theta) \int_0^x \frac{dF^*(t)}{G(t;\theta)} \equiv \int_0^x \frac{dF^*(t)}{G(t;\theta)} \Big/ \int_0^\infty \frac{dF^*(t)}{G(t;\theta)}$$

and hence a natural estimator of F follows:

$$\widehat{F}(x;\widehat{\theta}) = \hat{P}(\hat{\theta}) \int_0^x \frac{dF_n^*(t)}{G(t;\hat{\theta})} \equiv \int_0^x \frac{dF_n^*(t)}{G(t;\widehat{\theta})} \Big/ \int_0^\infty \frac{dF_n^*(t)}{G(t;\widehat{\theta})}, \tag{1}$$

where $\widehat{\theta}$ is a suitable estimator of θ, and where F_n^* stands for the ordinary empirical df of the $X_i's$.

The estimator $\widehat{F}(x;\widehat{\theta})$ can be alternatively motivated as a MLE. As noted by Shen [8], the likelihood of the (U_i, X_i, V_i)'s ($\mathcal{L}$) can be factorized as a product of the conditional likelihood of the (U_i, V_i)'s given the X_i's ($\mathcal{L}_c$), and the marginal likelihood of the X_i's ($\mathcal{L}_m$):

$$\mathcal{L} = \mathcal{L}_c \times \mathcal{L}_m \equiv \prod_{i=1}^n \frac{g(U_i, V_i; \theta)}{G(X_i; \theta)} \times \prod_{i=1}^n \frac{G(X_i;\theta)}{P(\theta)}$$

where $g(u, v; \theta) = P(U^* = u, V^* = v) = K_\theta(du, dv)$ stands for the joint density of (U^*, V^*) (assumed to exist). For each θ, $\mathcal{L}_m$ is maximized by

$$\widehat{F}(x;\theta) = \hat{P}(\theta) \int_0^x \frac{dF_n^*(t)}{G(t;\theta)} \equiv \int_0^x \frac{dF_n^*(t)}{G(t;\theta)} \Big/ \int_0^\infty \frac{dF_n^*(t)}{G(t;\theta)},$$

and the maximum is a constant [4]. Hence, the maximizer of the full-likelihood $\mathcal{L}$ is given by $\left(\widehat{\theta}, \widehat{F}(x;\widehat{\theta})\right)$, where $\widehat{\theta}$ stands for the maximizer of $\mathcal{L}_c$.

The estimator (1) reduces to the semiparametric estimator in [4] when there is no truncation from the right (i.e. $P(V^* = \infty) = 1$). Note that, unlike with left-truncated data, the function $G(x;\theta)$ does not need to be (and, in general, it will not be) monotone here.

Remark 1. In some instances the random vector (U^*, V^*) will fall on a line w. p. 1, $V^* = U^* + \tau$ say, see Section 4 for motivation. In this case, we rather have (whenever $v = u + \tau$) $g(u, v; \theta) = P(U^* = u) = L(du; \theta)$ and $G(x;\theta) = L(x;\theta) - L(x - \tau; \theta)$, where $L(.;\theta)$ is the df of the parametric model assumed for U^*. In particular, if U^* follows a uniform distribution on an interval which contains $(a_F - \tau, b_F)$, where (a_F, b_F) stands for the support of X^*, we have that $G(x;\theta)$ is constant. In this case, there is no observational bias, and the ordinary empirical df of the X_i's is a consistent estimator of F.

Remark 2. Unlike the Efron-Petrosian NPMLE (which must be computed in an iterative way), the new semiparametric estimator $\widehat{F}(x;\widehat{\theta})$ has explicit form. This immediately leads to simpler asymptotic expressions for the limiting distribution. As it will be seen, a simple plug-in method to estimate the standard error of (1) can be introduced.

Now we state the main results for both the estimated parameter $\widehat{\theta}$ and the semiparametric estimator $\widehat{F}(x;\widehat{\theta})$. Since we are mainly interested in testing

problems about θ and the construction of confidence limits for $F(x)$, we only report here the results concerning the distributional convergence of these estimators. Of course, formal results of consistency can be also obtained following similar arguments to those in [4]. Also, in order to favour the reading of the manuscript, we only refer to the needed, notationally involved assumptions as "under regularity": These assumptions mainly impose smoothness on $\log \mathcal{L}_c$ and the convergence of the solution of the maximum likelihood equation. Similarly, all the limit variances are assumed to be finite. See Section 5 for further details.

Theorem 2.1. Let $\widehat{\theta}$ a solution to the maximum likelihood equation, that is,

$$\frac{\partial}{\partial \theta} \log \mathcal{L}_c(\widehat{\theta}) = 0.$$

Under regularity, we have $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \to N(\underline{0}, I(\theta_0)^{-1})$ in law, where $I(\theta_0)$ is the Fisher information matrix

$$I(\theta_0) = -E_{\theta_0}\left[\frac{\partial^2}{\partial \theta^2} \log \frac{g(U_i, V_i; \theta)}{G(X_i; \theta)}\bigg|_{\theta=\widehat{\theta}}\right].$$

As usual, in practice one will rather use the empirical Fisher information

$$\widehat{I}(\widehat{\theta}) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log \frac{g(U_i, V_i; \theta)}{G(X_i; \theta)}\bigg|_{\theta=\widehat{\theta}}$$

to compute the standard errors and covariances of the estimated parameters.

Example 1. As an illustrative example, we consider a situation in which $U^* \sim Beta(\theta_1, 1)$ and $V^* \sim Beta(1, \theta_2)$ are independent. Then, we have $g(u, v, \theta) = \theta_1\theta_2 u^{\theta_1 - 1}(1 - v)^{\theta_2 - 1}$,$0 < u, v < 1$. Assume that the support of X^*, (a_F, b_F) say, is contained in the interval $(0, 1)$, so we have

$$G(x; \theta) = P\left(U^* \leq X^* \leq V^* \mid X^* = x\right) = P(U^* \leq x)P(V^* \geq x) = x^{\theta_1}(1-x)^{\theta_2}, a_F < x < b_F.$$

The conditional likelihood becomes $\mathcal{L}_c\left(\theta\right) = \mathcal{L}_{c,1}\left(\theta_1\right)\mathcal{L}_{c,2}\left(\theta_2\right)$, where

$$\mathcal{L}_{c,1}\left(\theta_1\right) = \prod_{i=1}^{n} \frac{\theta_1 U_i^{\theta_1 - 1}}{X_i^{\theta_1}}, \qquad \mathcal{L}_{c,2}\left(\theta_2\right) = \prod_{i=1}^{n} \frac{\theta_2\left(1 - V_i\right)^{\theta_2 - 1}}{\left(1 - X_i\right)^{\theta_2}},$$

which are respectively maximized by

$$\widehat{\theta}_1 = \left[-\frac{1}{n}\sum_{i=1}^{n} \log \frac{U_i}{X_i}\right]^{-1}, \qquad \widehat{\theta}_2 = \left[-\frac{1}{n}\sum_{i=1}^{n} \log \frac{1 - V_i}{1 - X_i}\right]^{-1}.$$

It is straightforward to obtain the second order derivatives of the log-likelihood, these are:

$$\frac{\partial^2}{\partial\theta_1^2}\log\frac{g(u,v,\theta)}{G(t,\theta)} = -\frac{1}{\theta_1^2}, \qquad \frac{\partial^2}{\partial\theta_2^2}\log\frac{g(u,v,\theta)}{G(t,\theta)} = -\frac{1}{\theta_2^2}, \qquad \frac{\partial^2}{\partial\theta_1\partial\theta_2}\log\frac{g(u,v,\theta)}{G(t,\theta)} = 0,$$

so the Fisher information matrix becomes $I(\theta) = diag(1/\theta_1^2, 1/\theta_2^2)$. This example will be of further use in the simulations section below.

For our next result we need to introduce the following matrix:

$$\begin{aligned} W(x,\theta) &= P(\theta)^2 \int_0^\infty \frac{\partial G(t;\theta)}{\partial\theta}\frac{dF^*(t)}{G(t;\theta)^2}\int_0^x \frac{dF^*(t)}{G(t;\theta)} - P(\theta)\int_0^x \frac{\partial G(t;\theta)}{\partial\theta}\frac{dF^*(t)}{G(t;\theta)^2} \\ &= \int_0^\infty \frac{\partial G(t;\theta)}{\partial\theta}\frac{1}{G(t;\theta)}\left[F(x) - I(t\le x)\right]dF(t). \end{aligned}$$

Theorem 2.2. Under regularity, we have $\sqrt{n}\left(\widehat{F}(x;\widehat{\theta}) - F(x)\right) \to N(0,\sigma^2(x))$ in law, where $\sigma^2(x) = \sigma_1^2(x) + \sigma_2^2(x)$, with $\sigma_1^2(x) = W(x,\theta_0)^T I(\theta_0) W(x,\theta_0)$ and

$$\sigma_2^2(x) = P(\theta_0)\left[\int_0^x \frac{dF(t)}{G(t;\theta_0)} + F^2(x)\int_0^\infty \frac{dF(t)}{G(t;\theta_0)} - 2F(x)\int_0^x \frac{dF(t)}{G(t;\theta_0)}\right].$$

Remark 3. The first term in the limit variance $\sigma^2(x)$ comes from the variance when estimating θ_0, while the second term is directly related to the correction of the observational bias. Indeed, the limit variance of (1) reduces to $\sigma_2^2(x)$ in the case of perfect knowledge on the biasing function $G(.;\theta_0)$. This would be the case, for example, when sampling individuals with terminating events in a fixed, given observational window of length τ ($V^* = U^* + \tau$), if one knew the distributional form of the 'incidence process' U^*. Also interestingly, in the case of no observational bias (i.e. a constant function $G(.;\theta_0)$), $\sigma_2^2(x)$ reduces to $F(x)(1-F(x))$ which is just the asymptotic variance of the ordinary empirical df.

In practice, one will be interested in the construction of confidence limits for $F(x)$. Theorem 2.2 suggests the following $100(1-\alpha)\%$ confidence interval:

$$I_{1-\alpha} = \left(\widehat{F}(x;\widehat{\theta}) \pm z_{\alpha/2}\frac{\widehat{\sigma}(x)}{\sqrt{n}}\right) \tag{2}$$

with

$$\begin{aligned} \widehat{\sigma}^2(x) &= \widehat{W}(x,\widehat{\theta})^T\widehat{I}(\widehat{\theta})\widehat{W}(x,\widehat{\theta}) + \\ &\quad + \hat{P}(\widehat{\theta})\left[\int_0^x \frac{d\widehat{F}(t;\widehat{\theta})}{G(t;\widehat{\theta})} + \widehat{F}(x;\widehat{\theta})^2\int_0^\infty \frac{d\widehat{F}(t;\widehat{\theta})}{G(t;\widehat{\theta})} - 2\hat{F}(x;\hat{\theta})\int_0^x \frac{d\widehat{F}(t;\widehat{\theta})}{G(t;\widehat{\theta})}\right] \end{aligned}$$

where

$$\widehat{W}(x,\theta) = \widehat{P}(\theta)^2 \int_0^\infty \frac{\partial G(t;\theta)}{\partial \theta} \frac{dF_n^*(t)}{G(t;\theta)^2} \int_0^x \frac{dF_n^*(t)}{G(t;\theta)} - \widehat{P}(\theta) \int_0^x \frac{\partial G(t;\theta)}{\partial \theta} \frac{dF_n^*(t)}{G(t;\theta)^2}$$

and where $z_{\alpha/2}$ stands for the $(1-\alpha)$-th quantile of the standard normal distribution.

3 Simulations

In this section we illustrate the finite sample behaviour of the semiparametric estimator (1) through simulation studies. Results corresponding to the (conditional) maximum likelihood estimator of θ_0 will be reported too. One of the main goals of our simulations will be the comparison between the new estimator and the Efron-Petrosian NPMLE, say $F_n^{EP}(x)$. For this comparison, we will use the quotient of mean squared errors (MSEs) as a measure of relative efficiency. Note that if

$$RE(F_n^{EP}(x), \widehat{F}(x;\widehat{\theta})) = \frac{MSE(\widehat{F}(x;\widehat{\theta}))}{MSE(F_n^{EP}(x))}$$

attains a value of $\rho < 1$, then the semiparametric estimator performs better; more specifically, it is meant that the efficiency of the NPMLE is just the $100\rho\%$ that of the semiparametric estimator, or that the new estimator is $1/\rho$ times more efficient that the NPMLE.

We have considered two different situations of double truncation. In Case 1, the pair (U^*, V^*) has a joint density $g(u,v) = g_1(u)g_2(v)$, where g_1 and g_2 denote the marginal densities (so U^* and V^* are independently generated). In Case 2, we simulate U^* and then we take $V^* + \tau$ for some fixed constant $\tau > 0$, so the joint density of (U^*, V^*) does not exist. Note that Case 2 follows the spirit of the childhood cancer data discussed in the Introduction.

For Case 1, we take $U^* \sim U(0,1)$ and $V^* \sim U(0,1)$, while X^* was generated according to a $U(a_F, b_F)$ (Models 1.1-1.3) or $Beta(1/2,1)$ adapted to the support (a_F, b_F), that is $X^* = (b_F - a_F)U(0,1)^2 + a_F$ (Models 1.4-1.6), where the values of $0 < a_F < b_F < 1$ were chosen as follows: $(a_F, b_F) = (0.25, 1)$ for Models 1.1 and 1.5, $= (0.1, 0.9)$ for Models 1.2 and 1.4, $= (0.5, 0.75)$ for Model 1.3, and $= (0, 0.5)$ for Model 1.6. All these Models 1.1-1.6 fall under the umbrella of our Example 1 in Section 2. For Case 2, we take $U^* \sim U(0, 0.75)$, $X^* \sim U(0,1)$ in Model 2.1, and $U^* \sim U(0,1)$, $X^* \sim 0.75Beta(3/4,1) + 0.25$ in Model 2.2.

In Figure 1 we illustrate the observational bias induced by each of these eight models. Note that the situations range from no or almost no observational bias (Models 2.2, 1.3), to strongly biased situations (Models 2.1, 1.4, 1.5, 1.6). We included in the simulations sampling biases in favour of small lifetimes (e.g. Models 1.1 and 1.5) and large lifetimes too (Model 1.6); besides, situations with an observable s-shaped curve (when the true curve is either linear or concave)

were also considered (Models 2.1, 1.2 and 1.4). As parametric information on the pair (U^*, V^*) we always consider a $Beta(\theta_1, 1)$ for U^* and a $Beta(1, \theta_2)$ for V^* in Case 1. For each Model, we simulate 1000 trials with final sample size $n = 50$, 250 or 500. This means that, for each trial, the number of simulated data is much larger than n, actually $N \approx nP(\theta_0)^{-1}$ were needed on average, where recall that $P(\theta_0)$ stands for the proportion of no truncation. For the simulated models, the proportion of truncation ranged between 75% and 88%; however, since (as usual with truncated data) we worked on the basis of reaching a given n, the type of observational bias as depicted in Figure 1 is more informative than the probability of truncation itself.

-Insert Figure 1-

In Tables 1-8 we report the MSEs of the semiparametric estimator and of the NPMLE for each Model and sample size, evaluated at each of the nine deciles $x_{0.1}, ..., x_{0.9}$ of F. For comparison purposes, we also include in these Tables 1-8 the MSE pertaining to the 'ideal' estimator which makes use of the true biasing function $G(.; \theta_0)$, say $\widehat{F}(x; \theta_0)$. We also report in Table 9 the bias and standard deviations of the estimated parameter θ_0. In all the cases it is seen that the estimators converge to their respective targets. As expected, the more efficient estimator was that based on the true biasing function; in this case, the term $\sigma_1^2(x)$ in the variance of the semiparametric estimator, see Theorem 2.2, just vanishes. When comparing $\widehat{F}(x; \widehat{\theta})$ to $F_n^{EP}(x)$, the most relevant result in that in all the cases the relative efficiency of the NPMLE was below 1, with the only exception of Model 2.2, medium and large sample size ($n = 250, n = 500$), and Model 2.1, large sample size ($n = 500$). In this latter case, a systematic bias in the estimation of θ_1 is appreciated (see Table 9), which could probably explain the relative poor behaviour of the new estimator. Indeed, even in these exceptions the estimator $\widehat{F}(x; \theta_0)$ outperforms the NPMLE. In general, we can conclude that the new estimator is more efficient, with a MSE which can be up to 13% that of the NPMLE.

-Insert Tables 1-9-

Another consequence of the simulations is that the efficiency of the semiparametric estimator relative to the NPMLE tends to increase at the right tail. An exception to this are those situations with no much observational bias (e.g. Models 1.2, 1.3 and 2.2), for which the maximum deficiency of the NPMLE is found around the median. As regards the influence of the sample size on the relative performance of the estimators, we can see from Tables 1-8 that, in general, a larger n leads to a slightly worse relative behaviour of the NPMLE. Finally, it is interesting to compare the MSEs of the 'ideal' estimator under double truncation, $\widehat{F}(x; \theta_0)$, to those corresponding to the ordinary empirical df based on the same sample size, which is known to be

$$MSE_{ord}(x) = \frac{F(x)(1 - F(x))}{n}.$$

This comparison allows to investigate the relative difficulties in estimation which are directly implied by the sampling bias. For example, in Model 1.3 it is seen that the MSE of $\widehat{F}(x;\theta_0)$ is close to $MSE_{ord}(x)$ along all the x deciles; this is in agreement with the absence of a strong sampling bias (Figure 1). On the contrary, in Model 1.6 (under which there is a significative observational bias) we see that the MSE of $\widehat{F}(x;\theta_0)$ is up to one order of magnitude grater than that associated to the ordinary empirical df.

4 The childhood cancer data

In this section we report our analysis of the childhood cancer data. As mentioned in the Introduction, these data concern all the cases of childhood cancer diagnosed in North Portugal between January 1st 1999 and December 31st 2003. 406 individuals reported complete information on the age at diagnosis X^* (we take years as time scale) and the date of birth D^*. As discussed in the Introduction, the age at diagnosis is doubly truncated by $(U^* = V^* - 5, V^*)$, where V^* stands for the elapsed time between birth and end of study (December 31st 2003); in other words, V^* represents the age of the individual at the closing date. Then, following our Remark 1 in Section 2, we have $G(x;\theta) = L(x;\theta) - L(x-\tau;\theta)$ for the biasing function, where $L(.;\theta)$ stands for the df of U^* and τ is the length of the observational window (5 years). In Figure 2, left, the semiparametric estimator of the df of X^* is depicted, together with the 95% pointwise confidence band, which was calculated according to (2). As parametric information on U^*, we have taken the $Beta(\theta, 1)$ distribution, adapted to the support $(-5, 15)$. It is important to remark that, by definition of childhood cancer, the support of X^* is $(0, 15)$, and hence our sampling information reduces to the births which took place in 1984 and afterwards. For comparison purposes, we also included in Figure 2, left, the Efron-Petrosian NPMLE.
In Figure 2, right, we depict the NPMLE of the df of U^* (which is doubly truncated by $(X^* - 5, X^*)$) together with the fitted $Beta(\theta, 1)$ model, for which we obtained $\widehat{\theta} = 1.19$ (standard error: 0.1817). Note that the estimators in the left panel are constructed by inverse-weighting the data $X_i$according to their counterparts in the right panel. We also point out that the null hypothesis of a uniform distribution for U^* (that is, $\theta = 1$) is accepted at a 5% level; this can be interpreted in terms of the stationarity of the birth dates for the individuals who will develop the disease. As a consequence, there is no much observational bias on the age at diagnosis in this case (see Remark 1).
When analyzing subgroups of individuals, however, we have found situations in which there exists a clear sampling bias. This was the case for example for the 38 reported cases of neuroblastoma cancer (results not shown). Similarly, we have confirmed the existence of a remarkable bias on the $V_i's$ for the whole data set, resembling the situation of simulated Model 2.1, see Figure 1. Hence, accounting for an eventual observational bias may be a matter of much importance in applications.

-Insert Figure 2-

In order to investigate the performance of the confidence limits in Figure 2, left, we have compared along 1000 trials the Monte Carlo standard deviation $s_{MC}(x)$ of the semiparametric estimator and the values of the asymptotic approximation $s_A(x) = \widehat{\sigma}(x)/\sqrt{n}$, and the mean and standard deviation of $s_{MC}(x)/s_A(x)$ for the nine deciles are reported in Table 10. We have taken Model 2.2 and $n = 500$ for these simulations since it corresponds almost perfectly to the situation in Figure 2 (see Figure 1). From this Table 10 we see that the asymptotic formula provides a good approximation of the standard error of the semiparametric estimator at least up to quantile 0.70. At the right tail of the distribution, however, some overestimation of the actual standard is appreciated. As a consequence, the confidence band in Figure 2 could be somehow inflated at the far right tail.

-Insert Table 10-

5 Technical proofs

Technical proofs of Theorems 2.1 and 2.2 follow standard arguments. For the sake of completeness, here we provide the key steps of the proofs.

Proof to Theorem 2.1.

Under regularity, we have:

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\log\mathcal{L}_c(\widehat{\theta}) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log\frac{g(U_i,V_i;\theta)}{G(X_i;\theta)}\Big|_{\theta=\widehat{\theta}} = \\ &= \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log\frac{g(U_i,V_i;\theta)}{G(X_i;\theta)}\Big|_{\theta=\theta_0} + \sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log\frac{g(U_i,V_i;\theta)}{G(X_i;\theta)}\Big|_{\theta=\widetilde{\theta}}\left(\widehat{\theta}-\theta_0\right) \end{aligned}$$

where we have used the mean value theorem, and where $\widetilde{\theta}$ is between $\widehat{\theta}$ and θ_0. Introduce the matrix $A_n = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log g(U_i,V_i;\theta)G(X_i;\theta)^{-1}\big|_{\theta=\widetilde{\theta}}$ so we have

$$A_n\left(\widehat{\theta}-\theta_0\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log\frac{g(U_i,V_i;\theta)}{G(X_i;\theta)}\Big|_{\theta=\theta_0} \equiv \frac{1}{n}\sum_{i=1}^{n}\xi(U_i,V_i,X_i).$$

Since the conditional density of (U_i,V_i) given X_i is $g(u,v;\theta_0)G(X_i;\theta_0)^{-1}I(u \le X_i \le v)$ we have

$$\begin{aligned} E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log\frac{g(U_i,V_i;\theta)}{G(X_i;\theta)}\Big|_{\theta=\theta_0} \mid X_i\right] &= \int\int_{u\le X_i\le v}\frac{\partial}{\partial\theta}\frac{g(u,v;\theta)}{G(X_i;\theta)}\Big|_{\theta=\theta_0}dudv \\ &= \frac{\partial}{\partial\theta}\int\int_{u\le X_i\le v}\frac{g(u,v;\theta)}{G(X_i;\theta)}dudv\Big|_{\theta=\theta_0} = \underline{0}, \end{aligned}$$

where we have assumed interchangeability of differentiation and integration. Hence, $E\xi(U_i, V_i, X_i) = \underline{0}$.

Besides, assuming interchangeability of differentiation and integration,

$$\begin{aligned}
& -E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\log\frac{g(U_i,V_i;\theta)}{G(X_i;\theta)}\bigg|_{\theta=\theta_0} \mid X_i\right] \\
= & -\int\int_{u\le X_i\le v}\frac{\partial^2}{\partial\theta^2}\frac{g(u,v;\theta)}{G(X_i;\theta)}\bigg|_{\theta=\theta_0}dudv \\
& +\int\int_{u\le X_i\le v}\frac{\partial}{\partial\theta}\frac{g(u,v;\theta)}{G(X_i;\theta)}\bigg|_{\theta=\theta_0}\left[\frac{\partial}{\partial\theta}\frac{g(u,v;\theta)}{G(X_i;\theta)}\bigg|_{\theta=\theta_0}\right]^T\frac{G(X_i;\theta_0)}{g(u,v;\theta_0)}dudv \\
= & -\frac{\partial^2}{\partial\theta^2}\int\int_{u\le X_i\le v}\frac{g(u,v;\theta)}{G(X_i;\theta)}dudv\bigg|_{\theta=\theta_0} \\
& +\int\int_{u\le X_i\le v}\frac{\partial}{\partial\theta}\frac{g(u,v;\theta)}{G(X_i;\theta)}\bigg|_{\theta=\theta_0}\left[\frac{\partial}{\partial\theta}\log\frac{g(u,v;\theta)}{G(X_i;\theta)}\bigg|_{\theta=\theta_0}\right]^T dudv \\
= & \underline{0}+\int\int_{u\le X_i\le v}\frac{\partial}{\partial\theta}\log\frac{g(u,v;\theta)}{G(X_i;\theta)}\bigg|_{\theta=\theta_0}\left[\frac{\partial}{\partial\theta}\log\frac{g(u,v;\theta)}{G(X_i;\theta)}\bigg|_{\theta=\theta_0}\right]^T\frac{g(u,v;\theta_0)}{G(X_i;\theta_0)}dudv \\
= & E_{\theta_0}\left[\left(\frac{\partial}{\partial\theta}\log\frac{g(U_i,V_i;\theta)}{G(X_i;\theta)}\right)\left(\frac{\partial}{\partial\theta}\log\frac{g(U_i,V_i;\theta)}{G(X_i;\theta)}\right)^T\bigg|_{\theta=\theta_0} \mid X_i\right],
\end{aligned}$$

so the information matrix $I(\theta_0) = E_{\theta_0}\left[\xi(U_i, V_i, X_i)\xi(U_i, V_i, X_i)^T\right]$ becomes

$$I(\theta_0) = -E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\log\frac{g(U_i,V_i;\theta)}{G(X_i;\theta)}\bigg|_{\theta=\theta_0}\right].$$

Under regularity, we have $A_n \to I(\theta_0)$, so

$$\widehat{\theta}-\theta_0 \sim I(\theta_0)^{-1}\frac{1}{n}\sum_{i=1}^{n}\xi(U_i, V_i, X_i)$$

and, by the CLT,

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \to N(\underline{0}, I(\theta_0)^{-1}) \qquad \text{in law.}$$

Proof to Theorem 2.2.

Note that

$$\widehat{F}(x;\widehat{\theta})-F(x) = \left[\widehat{F}(x;\widehat{\theta})-\widehat{F}(x;\theta_0)\right]+\left[\widehat{F}(x;\theta_0)-F(x)\right] \equiv I+II.$$

For II we have:

$$\begin{aligned}\widehat{F}(x;\theta_0)-F(x) &= \widehat{P}(\theta_0)\int_0^x \frac{dF_n^*(t)}{G(t;\theta_0)} - P(\theta_0)\int_0^x \frac{dF^*(t)}{G(t;\theta_0)} = \\ &= \left[\int_0^x \frac{dF_n^*(t)}{G(t;\theta_0)} - \int_0^x \frac{dF^*(t)}{G(t;\theta_0)}\right] P(\theta_0) + \\ &\quad + \left[P(\theta_0)^{-1} - \widehat{P}(\theta_0)^{-1}\right] P(\theta_0)\widehat{P}(\theta_0)\int_0^x \frac{dF_n^*(t)}{G(t;\theta_0)}.\end{aligned}$$

By the SLLN we have that $\widehat{P}(\theta_0)\int_0^x G(t;\theta_0)^{-1}dF_n^*(t) \to P(\theta_0)\int_0^x G(t;\theta_0)^{-1}dF^*(t) = F(x)$ w. p. 1, and hence II is asymptotically equivalent to

$$\begin{aligned}II &\sim P(\theta_0)\frac{1}{n}\sum_{i=1}^n \frac{I(X_i\le x)}{G(X_i,\theta_0)} - \widehat{P}(\theta_0)^{-1}P(\theta_0)F(x) \\ &= P(\theta_0)\frac{1}{n}\sum_{i=1}^n \frac{I(X_i\le x)-F(x)}{G(X_i,\theta_0)} = \frac{1}{n}\sum_{i=1}^n \eta(X_i,x)\end{aligned}$$

where $\eta(t,x) = P(\theta_0)(I(t\le x)-F(x))G(t,\theta_0)^{-1}$. For I introduce the function

$$\begin{aligned}\widehat{W}(x,\theta) &= \frac{\partial}{\partial\theta}\left(\widehat{P}(\theta)\int_0^x \frac{dF_n^*(t)}{G(t;\theta)}\right) \\ &= \widehat{P}(\theta)^2\int_0^\infty \frac{\partial G(t;\theta)}{\partial\theta}\frac{dF_n^*(t)}{G(t;\theta)^2}\int_0^x \frac{dF_n^*(t)}{G(t;\theta)} - \widehat{P}(\theta)\int_0^x \frac{\partial G(t;\theta)}{\partial\theta}\frac{dF_n^*(t)}{G(t;\theta)^2},\end{aligned}$$

By the mean value we have, for some $\widetilde{\theta}$ between $\widehat{\theta}$ and θ_0,

$$\begin{aligned}\widehat{F}(x;\widehat{\theta})-\widehat{F}(x;\theta_0) &= \widehat{P}(\widehat{\theta})\int_0^x \frac{dF_n^*(t)}{G(t;\widehat{\theta})} - \widehat{P}(\theta_0)\int_0^x \frac{dF_n^*(t)}{G(t;\theta_0)} = \\ &= \widehat{W}\left(x,\widetilde{\theta}\right)^T\left(\widehat{\theta}-\theta_0\right).\end{aligned}$$

Note that (under regularity) we have $\widehat{W}\left(x,\widetilde{\theta}\right) \to W(x,\theta_0)$ for each sequence $\widetilde{\theta}\to\theta_0$, where $W(x;\theta)$ is the matrix in Theorem 2.2, namely

$$\begin{aligned}W(x,\theta) &= P(\theta)^2\int_0^\infty \frac{\partial G(t;\theta)}{\partial\theta}\frac{dF^*(t)}{G(t;\theta)^2}\int_0^x \frac{dF^*(t)}{G(t;\theta)} - P(\theta)\int_0^x \frac{\partial G(t;\theta)}{\partial\theta}\frac{dF^*(t)}{G(t;\theta)^2} \\ &= \int_0^\infty \frac{\partial G(t;\theta)}{\partial\theta}\frac{1}{G(t;\theta)}\left[F(x)-I(t\le x)\right]dF(t).\end{aligned}$$

This shows that $I \sim W(x,\theta_0)^T\left(\widehat{\theta}-\theta_0\right)$ and hence

$$I \sim W(x,\theta_0)^T I(\theta_0)^{-1}\frac{1}{n}\sum_{i=1}^n \xi(U_i,V_i,X_i).$$

In sum,

$$\widehat{F}(x;\widehat{\theta}) - F(x) \sim W(x,\theta_0)^T I(\theta_0)^{-1} \frac{1}{n}\sum_{i=1}^{n} \xi(U_i,V_i,X_i) + \frac{1}{n}\sum_{i=1}^{n}\eta(X_i,x)$$

and, by the CLT,

$$\sqrt{n}\left(\widehat{F}(x;\widehat{\theta}) - F(x)\right) \to N(0,\sigma^2(x)) \qquad \text{in law,}$$

where

$$\sigma^2(x) = Var_{\theta_0}\left[W(x,\theta_0)^T I(\theta_0)^{-1}\xi(U_i,V_i,X_i) + \eta(X_i,x)\right].$$

Since $E_{\theta_0}[\eta(X_i,x)] = 0$, $E_{\theta_0}[\xi(U_i,V_i,X_i) \mid X_i] = \underline{0}$, and $E_{\theta_0}\left[\xi(U_i,V_i,X_i)\xi(U_i,V_i,X_i)^T\right] = I(\theta_0)$, it is easily seen that

$$\sigma^2(x) = W(x,\theta_0)^T I(\theta_0) W(x,\theta_0) + E_{\theta_0}\left[\eta(X_i,x)^2\right] = \sigma_1^2(x) + E_{\theta_0}\left[\eta(X_i,x)^2\right],$$

while

$$\begin{aligned} E_{\theta_0}\left[\eta(X_i,x)^2\right] &= P(\theta_0)^2 E_{\theta_0}\left[\frac{(I(X_i \le x) - F(x))^2}{G(X_i;\theta_0)^2}\right] \\ &= P(\theta_0)\left[\int_0^x \frac{dF(t)}{G(t;\theta_0)} + F^2(x)\int_0^\infty \frac{dF(t)}{G(t;\theta_0)} - 2F(x)\int_0^x \frac{dF(t)}{G(t;\theta_0)}\right] \\ &= \sigma_2^2(x). \end{aligned}$$

Ackowledgement. Work supported by the research Grant MTM2008-03129 of the Spanish Ministerio de Ciencia e Innovación, and by the Grant PGIDIT07PXIB300191PR of the Xunta de Galicia. We also thank the IPO (the Portuguese Institute of Cancer at Porto) for kindly providing the childhood cancer data.

6 References

1. Lynden-Bell D. A method of allowing for known observational selection in small samples applied to 3CR quasars.
Monthly Notices of the Royal Astronomical Society 1971; **155**:95-118.

2. Woodroofe M. Estimating a distribution function with truncated data.
Annals of Statistics 1985; **13**:163-177.

3. Stute W. Almost sure representation of the product-limit estimator for truncated data.
Annals of Statistics 1993; **21**:146-156.

4. Wang, MC. A semiparametric model for randomly truncated data.
Journal of the American Statistical Association 1989; **84**:742-748.

5. Asgharian M, MLan CE, Wolfson DB. Length-biased sampling with right-censoring: an unconditional approach.
Journal of the American Statistical Association 2002; **97**:201-209.

6. De Uña-Álvarez J. Nonparametric estimation under length-biased sampling and Type I censoring: a moment based approach.
Annals of the Institute of Statistical Mathematics 2004; **56**: 667-681.

7. Efron B, Petrosian V. Nonparametric methods for doubly truncated data.
Journal of the American Statistical Association 1999; **94**:824-834.

8. Bilker W, Wang MC. A semiparametric extension of the Mann-Whitney test for randomly truncated data.
Biometrics 1996; **52**:10-20.

9. Shen PS. Nonparametric analysis of doubly truncated data.
Annals of the Institute of Statistical Mathematics 2008. DOI: 10.1007/s10463-008-0192-2.

10. Moreira C, De Uña-Álvarez J. Bootstrapping the NPMLE for doubly truncated data.
Journal of Nonparametric Statistics; 2009, conditionally accepted.

Tables and Figures

PT	n	Deciles	$MSE(SP)$	$MSE(EP)$	$MSE(F0)$
		1	0.002061525	0.002501928	0.001763928
		2	0.004360882	0.005616232	0.003535746
		3	0.006532868	0.008791825	0.005062856
		4	0.008465470	0.011780232	0.006492010
81%	50	5	0.010744836	0.015313579	0.008155081
		6	0.012467664	0.018293383	0.009417968
		7	0.013742171	0.020847448	0.010575092
		8	0.013452420	0.022124487	0.010353295
		9	0.012523448	0.022805056	0.009881422
		1	0.0004605006	0.0005834183	0.0003965030
		2	0.0009644717	0.0013419631	0.0007742867
		3	0.0014673954	0.0021546637	0.0011621046
		4	0.0020203275	0.0031082746	0.0016366977
81%	250	5	0.0026204136	0.0042121517	0.0021234243
		6	0.0029918981	0.0052171106	0.0024414605
		7	0.0033648650	0.0062948142	0.0028592649
		8	0.0035624288	0.0072376022	0.0031989463
		9	0.0035199741	0.0080021470	0.0033504125
		1	0.0002155447	0.0002904056	0.0001897584
		2	0.0004951456	0.0007152155	0.0003997071
		3	0.0007966937	0.0012244044	0.0006110297
		4	0.0010717383	0.0017311702	0.0008291346
81%	500	5	0.0013595858	0.0022541476	0.0010573720
		6	0.0016508148	0.0028493750	0.0013230435
		7	0.0019271126	0.0034747819	0.0016089230
		8	0.0021161708	0.0040692894	0.0018204765
		9	0.0021049021	0.0044711624	0.0019353243

Table 1: MSE of the semiparametric estimator (SP), the Efron-Petrosian NPMLE (EP), and the ideal estimator with perfect knowledge on the biasing function ($F0$), along 1000 trials for Model 1.1. (Sample size n, proportion of truncation PT).

PT	n	Deciles	$MSE(SP)$	$MSE(EP)$	$MSE(F0)$
		1	0.003205087	0.004098422	0.002786051
		2	0.005268108	0.006261627	0.004206696
		3	0.006696844	0.007787114	0.005041426
		4	0.007539667	0.008738120	0.005515241
80%	50	5	0.007798335	0.009009446	0.005656342
		6	0.007288578	0.008532197	0.005281088
		7	0.006823107	0.007836684	0.004993707
		8	0.005843062	0.006688943	0.004401733
		9	0.003665437	0.004133689	0.002844762
		1	0.0006612811	0.0007401788	0.0005704673
		2	0.0011054316	0.0012673141	0.0008702240
		3	0.0014132606	0.0016158483	0.0010215962
		4	0.0016154385	0.0018796395	0.0011028547
80%	250	5	0.0015849107	0.0018471389	0.0010887800
		6	0.0014835090	0.0017447277	0.0010803251
		7	0.0013203771	0.0015496812	0.0009956297
		8	0.0010728513	0.0012337628	0.0008595568
		9	0.0006518220	0.0007339265	0.0005764728
		1	0.0003120416	0.0003602504	0.0002766546
		2	0.0005149313	0.0006415407	0.0004045606
		3	0.0006709098	0.0008446623	0.0004912223
		4	0.0007257232	0.0009391137	0.0005099646
80%	500	5	0.0007586053	0.0009860204	0.0005285877
		6	0.0007277505	0.0009313139	0.0005214042
		7	0.0006724494	0.0008398698	0.0005021680
		8	0.0005273762	0.0006345009	0.0004312750
		9	0.0003424828	0.0003995011	0.0003063205

Table 2: MSE of the semiparametric estimator (SP), the Efron-Petrosian NPMLE (EP), and the ideal estimator with perfect knowledge on the biasing function ($F0$), along 1000 trials for Model 1.2. (Sample size n, proportion of truncation PT).

PT	n	Deciles	$MSE(SP)$	$MSE(EP)$	$MSE(F0)$
		1	0.001643361	0.001826674	0.001617875
		2	0.003355181	0.003829595	0.003299569
		3	0.004476626	0.005039243	0.004368294
		4	0.004793375	0.005321404	0.004660045
77%	50	5	0.004961015	0.005571135	0.004782501
		6	0.004945223	0.005437397	0.004709823
		7	0.004373857	0.004716770	0.004248012
		8	0.003448285	0.003633319	0.003323457
		9	0.002102385	0.002182934	0.002042890
		1	0.0003699743	0.0004009675	0.0003674355
		2	0.0006148683	0.0007147792	0.0006022873
		3	0.0008438507	0.0009896280	0.0008299906
		4	0.0009957376	0.0011658171	0.0009813479
77%	250	5	0.0010278453	0.0011676031	0.0010152014
		6	0.0010026198	0.0011193546	0.0009935872
		7	0.0009102337	0.0009967867	0.0009010553
		8	0.0007269902	0.0007931460	0.0007154419
		9	0.0004328618	0.0004545924	0.0004262758
		1	0.0001780884	0.0001920186	0.0001749993
		2	0.0003065206	0.0003423694	0.0002999427
		3	0.0003934577	0.0004459520	0.0003824862
		4	0.0004777986	0.0005382140	0.0004629937
77%	500	5	0.0005015651	0.0005819730	0.0004880664
		6	0.0005055059	0.0005813051	0.0004910243
		7	0.0004660662	0.0005366701	0.0004536832
		8	0.0003672294	0.0004171184	0.0003567809
		9	0.0002039861	0.0002210477	0.0002002135

Table 3: MSE of the semiparametric estimator (SP), the Efron-Petrosian NPMLE (EP), and the ideal estimator with perfect knowledge on the biasing function ($F0$), along 1000 trials for Model 1.3. (Sample size n, proportion of truncation PT).

PT	n	Deciles	$MSE(SP)$	$MSE(EP)$	$MSE(F0)$
		1	0.002756128	0.002843182	0.002504684
		2	0.005741742	0.006169036	0.004858145
		3	0.007566078	0.008418652	0.006068308
		4	0.008116256	0.009188946	0.006159935
82%	50	5	0.007582895	0.009071123	0.005409582
		6	0.007247459	0.008781476	0.005028898
		7	0.006140689	0.007619646	0.004219278
		8	0.004704445	0.005777274	0.003340843
		9	0.003011730	0.003878136	0.002324548
		1	0.0006807418	0.0007117903	0.0006448516
		2	0.0011905385	0.0013069722	0.0010468952
		3	0.0014523878	0.0016876455	0.0011520026
		4	0.0017076320	0.0020382737	0.0012527151
82%	250	5	0.0016394125	0.0020223843	0.0010800891
		6	0.0014859324	0.0018870870	0.0009264248
		7	0.0012353703	0.0016329398	0.0007706206
		8	0.0009740618	0.0012899735	0.0006340005
		9	0.0005863008	0.0007491766	0.0004251185
		1	0.0003660532	0.0003877299	0.0003344946
		2	0.0005866765	0.0006861441	0.0005016258
		3	0.0007175698	0.0008797747	0.0005690123
		4	0.0007953163	0.0010352153	0.0005650214
82%	500	5	0.0008368848	0.0011034784	0.0005613380
		6	0.0007377919	0.0009823821	0.0004790982
		7	0.0006520586	0.0008537107	0.0004394912
		8	0.0004701132	0.0006208146	0.0003270498
		9	0.0002835636	0.0003509468	0.0002314125

Table 4: MSE of the semiparametric estimator (SP), the Efron-Petrosian NPMLE (EP), and the ideal estimator with perfect knowledge on the biasing function ($F0$), along 1000 trials for Model 1.4. (Sample size n, proportion of truncation PT).

PT	n	Deciles	$MSE(SP)$	$MSE(EP)$	$MSE(F0)$
		1	0.002338723	0.002622883	0.002121781
		2	0.004129233	0.004980223	0.003494245
		3	0.005742738	0.007501941	0.004654370
		4	0.007507386	0.010353965	0.005987021
80%	50	5	0.008916953	0.012815505	0.006952554
		6	0.009539072	0.014890884	0.007463607
		7	0.010113182	0.016709066	0.007970725
		8	0.010047823	0.017811123	0.008084181
		9	0.009284608	0.018508854	0.007705560
		1	0.000456180	0.0005249711	0.0004320292
		2	0.000831028	0.0010391416	0.0007375843
		3	0.001176654	0.0015957810	0.0009895059
		4	0.001493068	0.0021930578	0.0011914535
80%	250	5	0.001877504	0.0028716680	0.0014562261
		6	0.002177379	0.0034408979	0.0016677462
		7	0.002293955	0.0038849495	0.0017679650
		8	0.002402818	0.0041641200	0.0019276272
		9	0.002174655	0.0040684373	0.0018668506
		1	0.0002046008	0.0002294443	0.0001852800
		2	0.0003983050	0.0004834413	0.0003383581
		3	0.0005883101	0.0007714166	0.0004719512
		4	0.0007923803	0.0010535405	0.0006189862
80%	500	5	0.0009679968	0.0012965262	0.0007535814
		6	0.0010725547	0.0014963970	0.0008366527
		7	0.0012592570	0.0017692816	0.0010034636
		8	0.0012904023	0.0018442446	0.0011026186
		9	0.0011690711	0.0017159248	0.0010669378

Table 5: MSE of the semiparametric estimator (SP), the Efron-Petrosian NPMLE (EP), and the ideal estimator with perfect knowledge on the biasing function ($F0$), along 1000 trials for Model 1.5. (Sample size n, proportion of truncation PT).

PT	n	Deciles	$MSE(SP)$	$MSE(EP)$	$MSE(F0)$
		1	0.022404885	0.045594084	0.018490388
		2	0.024503233	0.044865534	0.020270767
		3	0.032486759	0.048896819	0.028455172
		4	0.034998113	0.047797480	0.031139253
88%	50	5	0.027908672	0.037770787	0.024600942
		6	0.019496239	0.026238071	0.016553085
		7	0.012145830	0.016351242	0.010091269
		8	0.006495243	0.008700548	0.005392628
		9	0.002341388	0.003095843	0.002008335
		1	0.0121041067	0.0223848545	0.0117419308
		2	0.0171870249	0.0258557977	0.0164099578
		3	0.0165201139	0.0232662978	0.0154892774
		4	0.0131557457	0.0182123528	0.0118974900
88%	250	5	0.0098292503	0.0134045860	0.0086339675
		6	0.0068030350	0.0091360734	0.0057766857
		7	0.0041079031	0.0054709021	0.0033669327
		8	0.0019914063	0.0026563129	0.0015796922
		9	0.0006210185	0.0008114685	0.0004868327
		1	0.0081730927	0.0175171374	0.0077169826
		2	0.0107087347	0.0184297984	0.0101753062
		3	0.0090210141	0.0150610888	0.0083696209
		4	0.0070218726	0.0114885416	0.0063805859
88%	500	5	0.0051321210	0.0082448627	0.0045158575
		6	0.0035244073	0.0055293217	0.0030031453
		7	0.0021658743	0.0033106562	0.0017893906
		8	0.0010492905	0.0015661037	0.0008573037
		9	0.0003185259	0.0004595646	0.0002616082

Table 6: MSE of the semiparametric estimator (SP), the Efron-Petrosian NPMLE (EP), and the ideal estimator with perfect knowledge on the biasing function ($F0$), along 1000 trials for Model 1.6. (Sample size n, proportion of truncation PT).

PT	n	Deciles	$MSE(SP)$	$MSE(EP)$	$MSE(F0)$
		1	0.019951458	0.05083268	0.008611647
		2	0.026848915	0.05245271	0.010828164
		3	0.028649233	0.05107925	0.010199766
		4	0.027721640	0.05055763	0.010164008
75%	50	5	0.024536855	0.05011307	0.010660798
		6	0.020311702	0.05058149	0.011120568
		7	0.016901747	0.05124272	0.012462514
		8	0.014123425	0.05112158	0.013067435
		9	0.006360175	0.05051990	0.006298227
		1	0.005671855	0.01524855	0.003434277
		2	0.007075935	0.01495085	0.003643953
		3	0.008356141	0.01494414	0.004206298
		4	0.008727914	0.01453400	0.005175000
75%	250	5	0.008802001	0.01397055	0.006402457
		6	0.009423967	0.01344969	0.008386406
		7	0.010680300	0.01279897	0.010656285
		8	0.011727498	0.01195224	0.012072659
		9	0.005905551	0.01144145	0.006012630
		1	0.006666588	0.007032078	0.001722100
		2	0.020947175	0.007110407	0.002215478
		3	0.034679305	0.007493308	0.003065115
		4	0.034917055	0.007663500	0.004265499
75%	500	5	0.026043511	0.007710129	0.005839187
		6	0.013169883	0.007592489	0.007793968
		7	0.002742600	0.007364631	0.010183209
		8	0.001639729	0.006730566	0.011738457
		9	0.003346154	0.006165853	0.005967015

Table 7: MSE of the semiparametric estimator (SP), the Efron-Petrosian NPMLE (EP), and the ideal estimator with perfect knowledge on the biasing function ($F0$), along 1000 trials for Model 2.1. (Sample size n, proportion of truncation PT).

PT	n	Deciles	$MSE(SP)$	$MSE(EP)$	$MSE(F0)$
		1	0.004769624	0.003344468	0.0018248
		2	0.010151958	0.007000935	0.0032980
		3	0.013932303	0.009912258	0.0039784
		4	0.015960945	0.012009022	0.0046872
75%	50	5	0.015804086	0.013167737	0.0048392
		6	0.013632571	0.013639357	0.0045700
		7	0.010423768	0.013382892	0.0038996
		8	0.006592271	0.011517173	0.0028600
		9	0.003097190	0.007558132	0.0018792
		1	0.0009581576	0.0004867772	0.000363568
		2	0.0022491909	0.0009275727	0.000651888
		3	0.0033527567	0.0012455409	0.000850256
		4	0.0039056644	0.0014435148	0.001038560
75%	250	5	0.0037282665	0.0014132206	0.001005504
		6	0.0032595586	0.0013991997	0.000984464
		7	0.0024715950	0.0012702681	0.000864368
		8	0.0015001175	0.0010560455	0.000705072
		9	0.0005944791	0.0005435971	0.000370624
		1	0.0004388828	0.0002269796	0.000186032
		2	0.0011374353	0.0004904380	0.000359636
		3	0.0015451974	0.0005727416	0.000406516
		4	0.0018451360	0.0006748815	0.000495724
75%	500	5	0.0018560337	0.0006790248	0.000523612
		6	0.0016470770	0.0006552353	0.000497228
		7	0.0012445665	0.0005955961	0.000444732
		8	0.0007602557	0.0004673307	0.000328308
		9	0.0003028591	0.0002689372	0.000181004

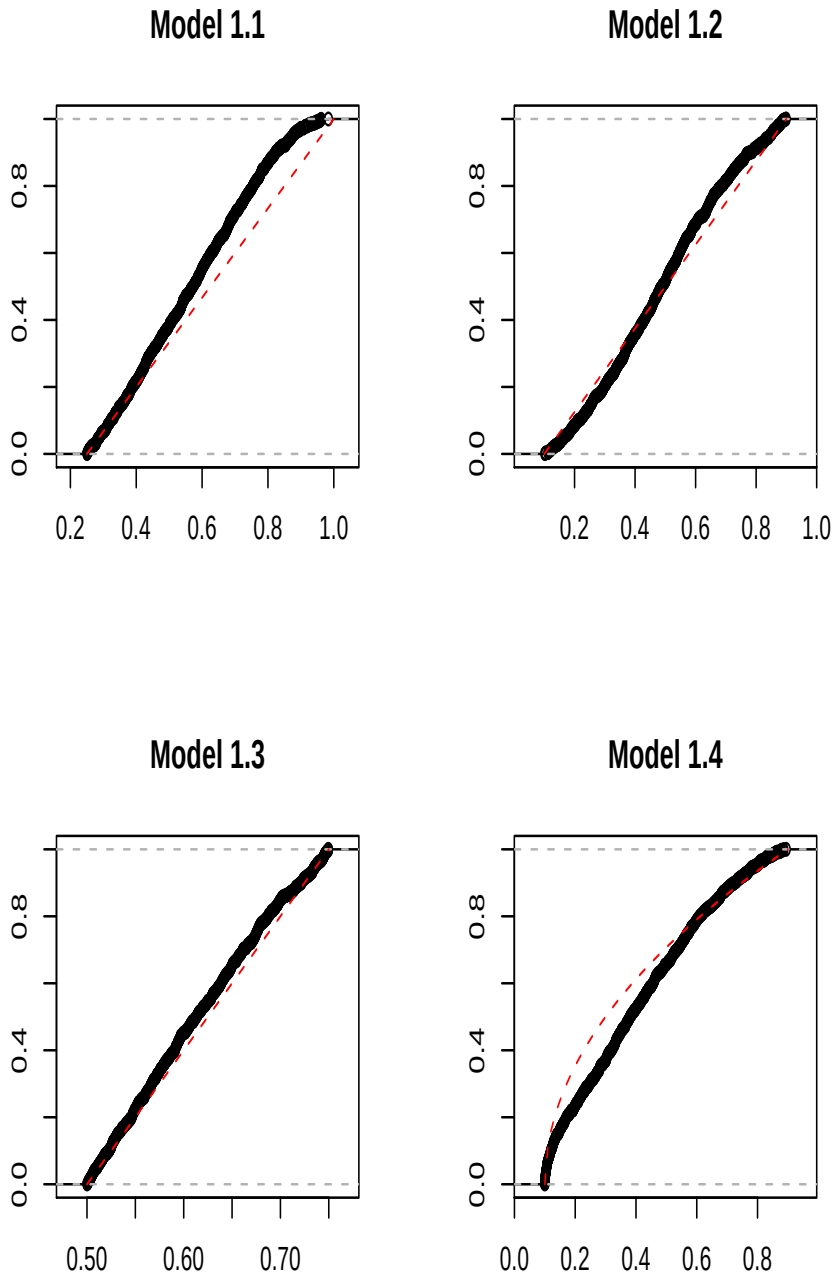
Table 8: MSE of the semiparametric estimator (SP), the Efron-Petrosian NPMLE (EP), and the ideal estimator with perfect knowledge on the biasing function ($F0$), along 1000 trials for Model 2.2.(Sample size n, proportion of truncation PT).

n	Models	$\hat{\theta}_1$	$\hat{\theta}_2$
	1.1	0.01705371(0.1514354)	0.01585983 (0.1443526)
	1.2	0.01726363 (0.1497229)	0.02933369 (0.1495772)
	1.3	0.01749347 (0.1460744)	0.02417042 (0.1514656)
	1.4	0.02186703 (0.1474750)	0.02048321 (0.1430628)
50	1.5	0.02706971 (0.1552804)	0.01079327 (0.1445023)
	1.6	0.02180327 (0.1449595)	0.01996355(0.1423285)
	2.1	0.02899540 (0.3013396)	
	2.2	0.03638238 (0.4554813)	
	1.1	0.004412362 (0.0667142)	0.002932107 (0.0630119)
	1.2	0.004521573 (0.06132066)	0.005129794 (0.0634064)
	1.3	0.004876636 (0.06022638)	0.001160238 (0.06304874)
	1.4	0.001231379 (0.06294079)	0.003617618 (0.0664989)
250	1.5	0.004666796 (0.06252818)	0.003306503 (0.06489206)
	1.6	0.003802629 (0.06437469)	0.0009008375 (0.06324714)
	2.1	-0.02029049 (0.1324506)	
	2.2	0.02308238 (0.1993036)	
	1.1	0.001510942 (0.04522862)	0.00257935 (0.04530109)
	1.2	0.0006533051 (0.04351268)	0.001897150 (0.04442708)
	1.3	0.001790503 (0.04461787)	0.002274069 (0.04564226)
	1.4	0.002067952 (0.04379179)	0.001343955 (0.04421807)
500	1.5	0.004440153 (0.04496469)	-0.0003432533 (0.04393165)
	1.6	0.002948612 (0.04498032)	0.0001128553 (0.04477341)
	2.1	-0.5619712 (0.06346337)	
	2.2	0.009266967 (0.1393680)	

Table 9: Bias and standard deviation (in brackets) of the estimated parameters for the simulated models.

Deciles	Mean $S_{MC}(x)/S_A(x)$	sd $S_{MC}(x)/S_A(x)$
1	1.0550	0.2424487
2	1.1330	0.2526269
3	1.1170	0.2434006
4	1.1470	0.2502776
5	1.1460	0.2532571
6	1.1170	0.2531902
7	1.0360	0.2476896
8	0.8911	0.2390539
9	0.6544	0.2262604

Table 10: Mean and standard deviation of $s_{MC}(x)/s_A(x)$ along 1000 trials of Model 2.2 with $n = 500$.


Model 1.1
Model 1.2
Model 1.3
Model 1.4
0.0
0.4
0.8
0.2
0.4
0.6
0.8
1.0
0.50
0.60
0.70
0.0
0.2
0.4
0.6
0.8

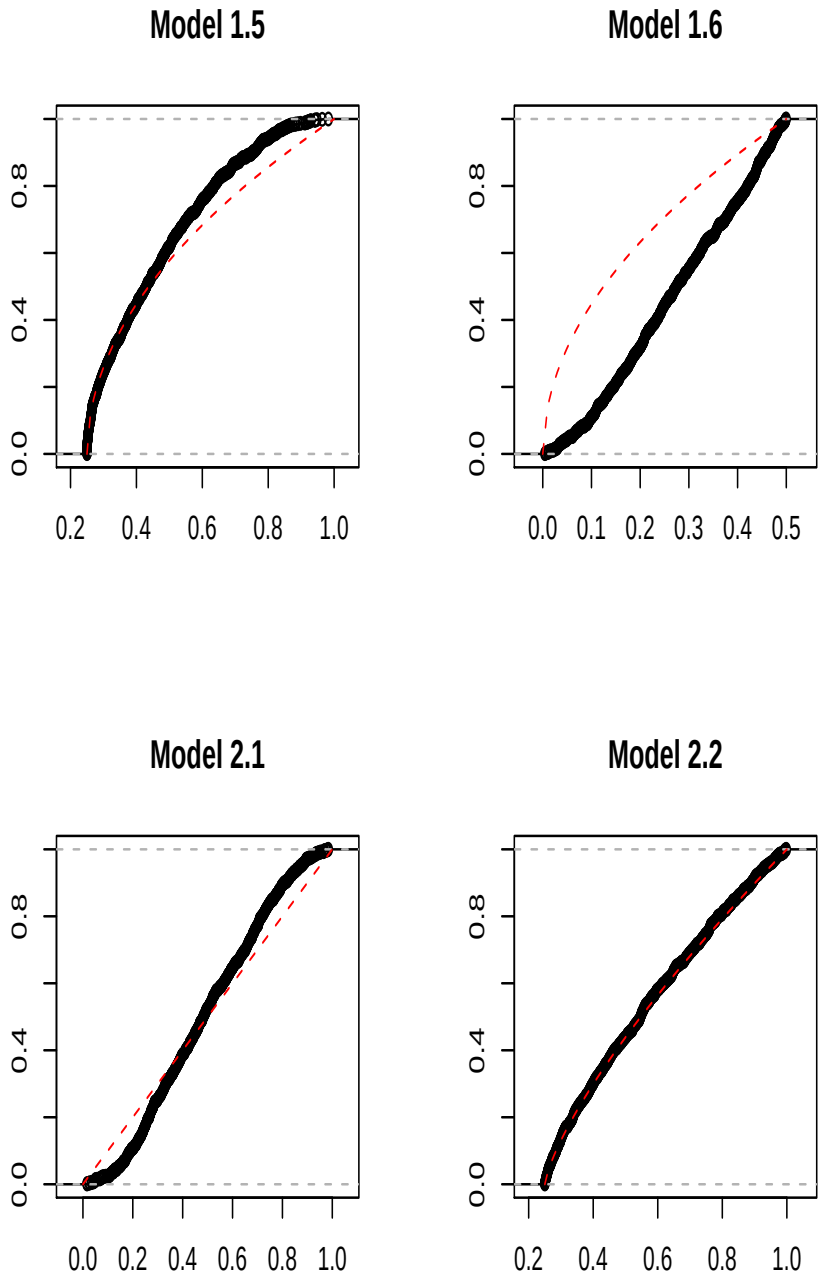


Figure 1: Observational bias for the simulated models: target F (dashed line) and observable lifetime distribution approximated by the empirical df F_n^* of a single sample with n=5000 (solid line).

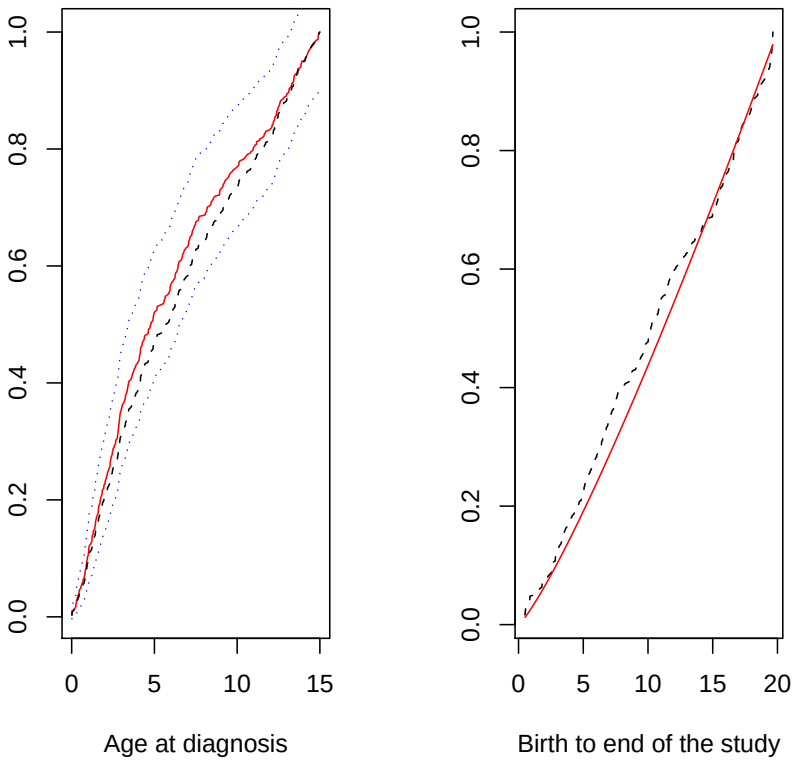


Figure 2: Left: The semiparametric estimator of the cumulative df for the age at diagnosis(solid line) and 95% pointwise confidence band (dotted lines);the Efron-Petrosian NPMLE is indicated. Right: Fitted beta distribution for V^* (solid line) and corresponding Efron-Petrosian NPMLE (dashed line). Childhood cancer data.