Nonparametric estimation of time-to-event distribution based on recall data in observational studies

Technical Report No. ASU/ 2013/ 7
Dated: 11 December 2013

Sedigheh Mirzaei Salehabadi
and
Debasis Sengupta

Indian Statistical Institute

Applied Statistics Unit
Kolkata 700 108

Nonparametric estimation of time-to-event distribution based on recall data in observational studies

Sedigheh Mirzaei Salehabadi and Debasis Sengupta

Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108, India.

Email: sedigheh_r@isical.ac.in, sdebasis@isical.ac.in

Abstract

In a cross-sectional observational study, time-to-event distribution can be estimated from data on current status or from recalled data on the time of occurrence. In either case, one can treat the data as having been interval censored, and use the nonparametric maximum likelihood estimator proposed by Turnbull (1976). However, the chance of recall may depend on the time span between the occurrence of the event and the time of interview. In such a case the underlying censoring would be informative, rendering the Turnbull estimator inappropriate. In this article, we provide a nonparametric maximum likelihood estimator of the distribution of interest based on a model that makes use of the special nature of the data at hand. Monte Carlo simulations indicate that the proposed estimator has smaller bias than the Turnbull estimator based on incomplete recall data, smaller variance than the Turnbull estimator based on current status data, and smaller mean squared error than both of them. The method is applied to menarcheal data from a recent Anthropometric study on adolescent and young adult females in Kolkata, India.

KEYWORDS: Interval censoring, Informative censoring, Nonparametric maximum likelihood estimator, Self consistency algorithm, Turnbull estimator.

1. INTRODUCTION

Observational data on time to occurrence of a landmark event occur in various fields of biological and social sciences. Some examples of landmark events are onset of menarche in adolescent and young adult females (Bergsten-Brucefors 1976; Chumlea, Schubert, Roche, Kulin, Lee, Himes and Sun 2003; Mirzaei and Sengupta 2013), dental development (Demirjian, Goldstien and Tanner 1973; Eveleth and Tanner 1990), breast development (Cameron 2002; Aksglaede, Sorensen, Petersen, Skakkebak and Juul 2009), beginning of criminal career (Hosmer and Lemeshow 1999), marriage and birth of the first child (Allison 1982), end of a work career (LeClere 2005) and end of a strike (Hosmer and Lemeshow 1999). The probability distribution of the time till occurrence of the event is useful for comparing two populations, setting benchmarks for individuals, setting policy objectives and so on. Estimation of that distribution is therefore an important inferential issue. Ideally one would like to observe a number of individuals continuously or periodically until the occurrence of the landmark event (Korn, Graubard and Midthune 1997; McKay, Bailey, Mirwald, Davison and Faulkner 1998). However, researchers often opt for cross-sectional studies in order to save time and cost.

Cross-sectional studies can produce dichotomous data on the current status of an individual (whether or not the landmark event has occurred till the day of observation). A binary data regression model such as probit or logistic model, with time as the covariate, is often used for estimating the probability distribution function (Hediger and Stine 1987; Ayatollahi, Dowlatabadi and Ayatollahi 2002). It is also possible to estimate the distribution nonparametrically, by regarding the current status data as either left or right censored observations. The nonparametric maximum likelihood estimator (NPMLE) proposed by Turnbull (1976) for interval censored data has occasionally been used in this set-up (Keiding, Begtrup, Scheike and Hasibeder 1996).

In some cross-sectional studies, a subject is asked to recall the time of the landmark event, in case it has already taken place. Such retrospective data are often incomplete. In many cases (e.g., when the event has not happened or the subject cannot recall when it had happened) one can specify only a range for the requisite time. Thus, data arising from retrospective studies are also interval-censored. In this situation, it is tempting to use the NPMLE obtained by Turnbull (1976). In fact, there are instances when this estimator has been used in estimating the distribution

of age at reaching a developmental landmark (see, e.g., Aksglaede et al. (2009)). However, the censoring mechanism in this set-up is likely to depend on the time-to-event, thereby making the censoring informative. This is because of the fact that memory generally fades with time. As an example, for two subjects interviewed at the same age, the one with later onset of menarche is more likely to remember the date. It may be recalled that the Turnbull estimator, is not meant for informatively censored data, and can have large bias when the censoring is informative, as confirmed by simulations reported in this paper.

We propose in Section 2 a new approach for estimating the distribution of the time-to-event by using the recall information through an informative censoring model. Under this model, the time of observation is assumed to be independent of the time-to-event, and the recall probability is regarded as a function of the time gap between the event and the observation. In Section 3, we derive the NPMLE under the model, and establish its existence and uniqueness (for large sample size) and provide a self-consistency algorithm for computing it. Results of Monte Carlo simulations and an illustrative data analysis are reported in Section 4 and 5, respectively. The data analysis is based on a study on menarcheal age of adolescent and young adult females, undertaken by the Indian Statistical Institute, Kolkata, where the landmark event is the onset of menarche. The conditions chosen for simulations are also in line with this application.

Proofs of all the results are given in the Appendix.

2. MODEL AND LIKELIHOOD

Consider a set of subjects having time of the occurrence of landmark event $T_1, \ldots, T_n$, which are samples from a common distribution F with density f and support $[t_{min}, t_{max}]$. Let these subjects be interviewed at times $S_1, \ldots, S_n$, respectively and δ_i be the indicator of $T_i \leq S_i$, i.e., the event having had occurred that event on or before the time of interview.

In the case of current status data, one only observes $(S_i, \delta_i), (i = 1, 2, \ldots, n)$. The corresponding likelihood, conditional on the time of interview, is

$$\prod_{i=1}^{n} [F(S_i)]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}, \tag{1}$$

where $\bar{F} = 1 - F$.

In a retrospective study, the subject may remember the exact time of the event. Let ε_i be the indicator of recalling that time. Thus, for the ith subject, it can be said that $T_i \in A_i$, where the set A_i is defined as follows in three cases.

Case (i). When $\delta_i = 0$, i.e., the landmark event for the ith subject did not occur till the time of interview, we have $A_i = [S_i \vee t_{min}, t_{max}]$.

Case (ii). When $\delta_i = 1$ and $\varepsilon_i = 1$, i.e., the landmark event for the ith subject did occur and the subject can exactly recall the date, we have $A_i = \{T_i\}$.

Case (iii). When $\delta_i = 1$ and $\varepsilon_i = 0$, i.e., the landmark event for the ith subject did occur and the subject cannot recall the exact date, we have $A_i = [t_{min}, t_{max} \wedge S_i]$

Note that there can be a gray area between cases (ii) and (iii), when the subject can recall an approximate date or a range of dates of the landmark event. In view of the possibility of recall error, noted by several researchers (Rabe-Hesketh, Yang and Pickles 2001; Beckett, DaVanzo, Sastry, Panis and Peterson 2001) we ignore such incomplete information and treat these as cases of no recall.

If the underlying censoring mechanism is presumed to be noninformative, then the likelihood arising from the above retrospective data, conditional on the time of interview, is

$$\prod_{i=1}^{n} \left[\{F(S_i)\}^{1-\varepsilon_i}\{f(T_i)\}^{\varepsilon_i}\right]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}, \tag{2}$$

However, it has been mentioned in Section 1 that the censoring mechanism is likely to be informative. Specifically, the non-recall probability, $P(\varepsilon_i = 0|\delta_i = 1)$ may depend on the time elapsed since the time of that event, $S_i - T_i$. We model the conditional probability of forgetting the date as an unspecified function of the elapsed time,

$$\pi(s-t) = P(\varepsilon_i = 0|T_i = t, S_i = s), \qquad s > t. \tag{3}$$

According to this model, the likelihood, conditional on the ages at interview, is

$$\prod_{i=1}^{n} \left[\left(\int_0^{S_i} f(u)\pi(S_i - u)du \right)^{1-\varepsilon_i} \{f(T_i)(1 - \pi(S_i - T_i))\}^{\varepsilon_i} \right]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}. \tag{4}$$

Here, the informativeness of the censoring mechanism is captured through the function π. When π is a constant, the likelihood (4) becomes a constant multiple of (2). As a further special case, if

$\pi = 1$, then the likelihood (4) reduces to (1). On the other hand, when $\pi = 0$, i.e., the landmark event times are perfectly recalled, the product likelihood (4) reduces to

$$\prod_{i=1}^{n} [f(T_i)]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}, \tag{5}$$

which is the likelihood for randomly right-censored data. These reductions follow from the fact that the model (3) leading to the likelihood (4) is more general than the usual censoring models.

In order to simplify the integral in (4) we presume that π is a piecewise constant function of the form

$$\pi(x) = b_1 I(x_1 < x \leq x_2) + b_2 I(x_2 < x \leq x_3) + \ldots + b_k I(x_k < x < \infty), \tag{6}$$

where $0 < x_1 < x_2 < \cdots < x_k$; $0 < b_1, b_2, \ldots, b_k \leq 1$. Note that it is possible to arrange the parameters $b_1, b_2, \ldots, b_k$ in increasing order, to make π a non-decreasing function, which corresponds to the general perception that memory fades with time. We do not use such a constraint in this paper.

In view of (6), the likelihood (4) reduces to

$$\begin{aligned} L = \prod_{i=1}^{n} \Bigg[& \left\{ \sum_{l=1}^{k} b_l \big(F(W_l(S_i)) - F(W_{l+1}(S_i)) \big) \right\}^{1-\varepsilon_i} \\ & \left\{ f(T_i) \left(1 - \sum_{l=1}^{k} b_l I\big(W_{l+1}(S_i) < T_i \leq W_l(S_i) \big) \right) \right\}^{\varepsilon_i} \Bigg]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}. \end{aligned} \tag{7}$$

where $W_l(S_i) = t_{max} \wedge ((S_i - x_l) \vee t_{min})$. Note that

$$t_{min} = W_{k+1}(S_i) \leq W_k(S_i) \leq W_{k-1}(S_i) \leq \cdots \leq W_1(S_i) = t_{max}. \tag{8}$$

Depending on the value of S_i, some of the above inequalities may in fact be equalities. Specifically, if l is an index such that $S_i - x_{l+1} \leq t_{min} < S_i - x_l$ then $t_{min} = W_{k+1}(S_i) = \cdots = W_{l+1}(S_i)$. Further, if l is an index such that $S_i - x_{l+1} < t_{max} \leq S_i - x_l$, then we have $W_l(S_i) = \cdots = W_1(S_i) = t_{max}$. The remaining equalities would be strict.

The likelihood (7) can be rewritten as

$$L = \prod_{i=1}^{n} \left[\left\{ \sum_{l=1}^{k} b_l \int_{A_{il}} f(u)du \right\}^{1-\varepsilon_i} \left\{ f(T_i) \left(1 - \sum_{l=1}^{k} b_l I\big(T_i \in A_{il}\big) \right) \right\}^{\varepsilon_i} \right]^{\delta_i} \big[\bar{F}(S_i)\big]^{1-\delta_i}, \tag{9}$$

where $A_{il} = [W_{l+1}(S_i), W_l(S_i))$ for $2 \leq l < k$ and $A_{i1} = [W_2(S_i), t_{max}]$. Note from (8) that some of the A_{il}'s can be empty. The simple form of the likelihood, which can be written without any integral, paves the way for estimation.

3. NONPARAMETRIC ESTIMATION

3.1 Reduction of The Problem

Assuming that k and $x_1, x_2, \ldots, x_k$ are known, we attempt to narrow down the domain over which the likelihood needs to be maximized. Recall from Section 2 that for retrospective data with possible non-recall, the actual time of the landmark event of the ith subject can be said to belong to a set A_i. In case (i), A_i is an interval. In case (ii), it is a singleton set. In case (iii), it may be written as a union of the disjoint intervals used in the likelihood (9), i.e., $A_i = \bigcup_{l=1}^{k} A_{il}$. In summary, if we regard a singleton set as a special case of an interval, then each of the sets A_i, $i = 1, \ldots, n$, can be said to be constituted of a union of one or more intervals.

Consider the collection of all these intervals. Let $\{B_1, \ldots, B_M\}$ be the set of all unique members of this collection, disregarding replications. Following (Turnbull 1976), we look for intersections of these intervals. Denote the non-empty subsets of the index set $\{1, 2, \ldots, M\}$ by $s_1, s_2, \ldots, s_{2^M-1}$. Define

$$I_r = \Big\{ \bigcap_{i \in s_r} B_i \Big\} \bigcap \Big\{ \bigcap_{i \notin s_r} B_i^c \Big\} \qquad \text{for } r = 1, 2, \ldots, 2^M - 1.$$

Some of the I_r's may be empty sets, denoted here by ϕ. Let

$$\mathcal{C} = \{s_r : I_r \neq \phi,\, 1 \leq r \leq 2^M - 1\}.$$

While the members of $\mathcal{C}$ are distinct, the collection of intervals I_r such that $s_r \in \mathcal{C}$ may have multiplicities. Let $\mathcal{A}$ be the set of distinct intervals I_j such that $s_r \in \mathcal{C}$. Let $p_r = P(I_r)$, for all $I_r \in \mathcal{A}$. By using the definition of I_r, we can rewrite the contribution of individual i to the likelihood in the three cases of censoring as follows.

Case (i): Let $\delta_i = 0$. If l_i is the index such that $B_{l_i} = A_i$, then

$$P(A_i) = P(B_{l_i}) = \sum_{\substack{r: l_i \in s_r \\ s_r \in \mathcal{C}}} p_r.$$

Case (ii): Let $\delta_i \epsilon_i = 1$. If l_i is the index such that $I_{l_i} = A_i$ then $P(A_i) = p_{l_i}$.

Case (iii): Let $\delta_i(1-\epsilon_i)=1$. If $l_{i1}, l_{i2}, \ldots, l_{ik}$ are indices such that $B_{l_{it}} = A_{it} \quad$ for $1 \le t \le k$, then

$$P(A_{it}) = P(B_{l_{it}}) = \sum_{\substack{r: l_{it} \in s_r \\ s_r \in \mathcal{C}}} p_r.$$

When the likelihood (9) is written in terms of the p_r's, it reduces to

$$L = \prod_{i=1}^{n} \left[\left\{ \sum_{t=1}^{k} b_t \left(\sum_{\substack{r: l_{it} \in s_r \\ s_r \in \mathcal{C}}} p_r \right) \right\}^{1-\varepsilon_i} \left\{ p_r|_{I_r = A_i} \left(1 - \sum_{t=1}^{k} b_t I\big(T_i \in A_{it}\big) \right) \right\}^{\varepsilon_i} \right]^{\delta_i} \left[\sum_{\substack{r: l_i \in s_r \\ s_r \in \mathcal{C}}} p_r \right]^{1-\delta_i}, \tag{10}$$

Thus, maximizing the likelihood (9) is equivalent to maximizing the likelihood (10) with respect to p_r for $s_r \in \mathcal{C}$.

There is a partial order among the members of $\mathcal{C}$ in the sense that some sets are contained in others. Let

$$\mathcal{C}_0 = \{s_j : s_j \in \mathcal{C}; s_j \subset s_{j'} \text{ does not hold for any } s_{j'} \in \mathcal{C}\}.$$

Thus, $\mathcal{C}_0$ is a subset of $\mathcal{C}$ that retains only those sets that are *not* proper subsets of any other set. Our first result shows that the maximization of the likelihood can be restricted to this smaller set.

Theorem 1 $\displaystyle\max_{\substack{\{p_r: s_r \in \mathcal{C}\} \\ p_r \in [0,1], \sum_{s_r \in \mathcal{C}} p_r = 1}} L = \max_{\substack{\{p_r: s_r \in \mathcal{C}_0\} \\ p_r \in [0,1], \sum_{s_r \in \mathcal{C}_0} p_r = 1}} L$

It follows from the above theorem that the likelihood has the same maximum value with or without the restriction $p_r = 0$ for $s_r \in \mathcal{C} \backslash \mathcal{C}_0$. Therefore, we can replace $\mathcal{C}$ by $\mathcal{C}_0$ in (10).

Let $\mathcal{A}_0$ be the set of distinct intervals I_j such that $s_j \in \mathcal{C}_0$. In order to simplify the notation, let $\mathcal{A}_0 = \{J_1, J_2, \ldots, J_m\}$ and $q_j = P(J_j)$. Let $\boldsymbol{p} = (q_1, q_2, ..., q_m)$. Theorem 1 implies that one can maximize the likelihood subject to the restriction $\sum_{j=1}^{m} q_j = 1$. The problem reduces to maximizing

$$\begin{aligned} L(\boldsymbol{p}) &= L(q_1, ..., q_m) \\ &= \prod_{i=1}^{n} \left[\left\{ \sum_{t=1}^{k} b_t \left(\sum_{j: J_j \subset A_{it}} q_j \right) \right\}^{1-\varepsilon_i} \left\{ \sum_{j: J_j = A_i} q_j \left(1 - \sum_{t=1}^{k} b_t I\big(T_i \in A_{it}\big) \right) \right\}^{\varepsilon_i} \right]^{\delta_i} \left[\sum_{j: J_j \subset A_i} q_j \right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\sum_{j=1}^{m} \beta_{ij} q_j \right], \end{aligned} \tag{11}$$

subject to $\sum_{j=1}^{m} q_j = 1$ and $q_j \geq 0$, where

$$\beta_{ij} = \begin{cases} 1 & \text{if } J_j \subseteq A_i,\ \delta_i = 0, \\ 1 - \sum_{t=1}^{k} b_t I\big(T_i \in A_{it}\big) & \text{if } J_j \subseteq A_i,\ \delta_i \epsilon_i = 1, \\ \sum_{t=1}^{k} b_t & \text{if } J_j \subseteq A_{it},\ \delta_i(1-\epsilon_i) = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The task of identifying a maximum of the above likelihood is simplified through the following result, which is interesting by its own right.

Theorem 2 *The probability of $\mathcal{A}_0$ containing only singletons goes to one, as $n \to \infty$.*

We are now ready for the next result regarding the existence and uniqueness of the NPMLE.

Theorem 3 *The likelihood* (11) *has a maximum and the probability that it has a unique maximum goes to one, as $n \to \infty$.*

3.2 A Self-Consistency Algorithm

Following the work of Efron (1967) on computing the Kaplan-Meier estimator (Kaplan and Meier 1958) through a self consistency algorithm and similar work by Turnbull (1976) in the case of interval censored data, we seek to obtain a self consistency algorithm to calculate the NPMLE described above.

For $i = 1, 2, \ldots, n$, let

$$L_{ij} = \begin{cases} 1 & \text{if } T_i \in J_j, \\ 0 & \text{otherwise,} \end{cases}$$

When $\delta_i \epsilon_i = 1$, the value of L_{ij} is known. Otherwise, its expectation with respect to the probability vector $\boldsymbol{p}$ is given by

$$E(L_{ij}) = \frac{\beta_{ij} q_j}{\sum_{j=1}^{m} \beta_{ij} q_j} = \mu_{ij}(\boldsymbol{p}), \quad \text{say.} \tag{13}$$

Thus, $\mu_{ij}(\boldsymbol{p})$ represents the probability that the i-th observation lies in J_j. The average of these probabilities across the n individuals,

$$\frac{\sum_{i=1}^{n} \mu_{ij}(\boldsymbol{p})}{n} = \pi_j(\boldsymbol{p}), \quad \text{say,} \tag{14}$$

should indicate the probability of the interval J_j. Thus, it is reasonable to expect that the vector $\boldsymbol{p}$ would satisfy the equation

$$q_j = \pi_j(\boldsymbol{p}) \quad \text{for} 1 \leq j \leq m. \tag{15}$$

An estimator of $\boldsymbol{p}$ may be called self consistent if it satisfies the simultaneous equations

$$q_j = \pi_j(\boldsymbol{p})$$

The form of the equations (15) suggests the following iterative procedure.

STEP I. Obtain a set of initial estimates q_j^0 $(1 \leq j \leq m)$.

STEP II. At the nth stage of iteration, use current estimate, $\boldsymbol{p}^n$, to evaluate $\mu_{ij}(\boldsymbol{p}^n)$ for $i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m$ and $\pi_j(\boldsymbol{p}^n)$ for $j = 1, 2, \ldots, m$ from (13) and (14), respectively.

STEP III. Obtain improved estimates $\boldsymbol{p}^{n+1}$ by setting $q_j^{n+1} = \pi_j(\boldsymbol{p}^n)$.

STEP IV. Return to Step II with $\boldsymbol{p}^{n+1}$ replacing $\boldsymbol{p}^n$.

STEP V. Iterate; stop when the required accuracy has been achieved.

The following theorem shows that the equation (15) defining a self consistency estimator must be satisfied by an NPML estimator of $\boldsymbol{p}$.

Theorem 4 *An NPML estimator of $\boldsymbol{p}$ must be self consistent.*

Let $\hat{\boldsymbol{p}} = (\hat{q_1}, ..., \hat{q_m})$ denote a value of $\boldsymbol{p}$ for which $L(\boldsymbol{p})$ attains its maximum over the set $\Re = \{\boldsymbol{p} | \sum_{j=1}^m q_j = 1, \quad q_j \geq 0\}$. Then a maximum likelihood estimate $\hat{F}$ of F is given by

$$\hat{F}(t) = \hat{q_1} + \hat{q_2} + ... + \hat{q_j} \qquad \forall \quad j : J_j \subseteq [0, t].$$

The variance of this estimator may be estimated through bootstrap resampling.

4. SIMULATION

For the purpose of the simulation, we generate sample time to landmark event from the Weibull distribution with shape and scale parameters $\alpha = 11$ and $\beta = 13$, respectively, and truncate the generated samples to the interval [8,16]. This truncated distribution has median of 11.57.

The corresponding 'time of interview' is generated from the discrete uniform distribution over $\{7, 8, \ldots, 21\}$. These choices are in line with the data analytic example of the next section, where the time to landmark event is the age at menarche in years. As for the forgetting probability, we use (6) with $k = 3$, $x_1 = 0, x_2 = 2.5$ and $x_3 = 4.5$ and two sets of values of the parameters:

Case (a): $b_1 = 0.1, b_2 = 0.4$ and $b_3 = 0.95$,

Case (b): $b_1 = 0.05, b_2 = 0.15$ and $b_3 = 0.35$.

The choice in Case (a) corresponds to rapid forgetting with the passage of time, while the choice in Case (b) indicates better retention.

As stated in Section 3.1, the proposed NPMLE of F, based on the likelihood (7), is implemented by assuming $k, x_1, x_2, \ldots, x_k$ in (6) are known.

We compare the performance of this estimator, described here as the proposed NPMLE, with the two NPMLEs based on (1) and (2), described here as Turnbull estimator (status) and Turnbull estimator (duration), respectively. As a benchmark, we also evaluate the performance of the empirical distribution function (EDF), a hypothetical estimator, computed from the underlying complete data. The results reported here are based on 500 simulation runs for sample sizes $n = 100$, 300 and 1000.

The Turnbull estimator (status) is unique only up to integer age intervals. Therefore, in all the plots, we represent it through a set of unconnected points at integer ages.

Figure 1 shows plots of the bias, the variance and the mean square error (MSE) of the four estimators for different ages, for $n = 100$ and parameters of the forgetting function (6) chosen as in case (a). The proposed NPMLE is found to have smaller bias than the Turnbull estimator (duration), and smaller variance than the Turnbull estimator (status), and much smaller MSE than both the Turnbull estimators.

Figure 2 shows these plots for $n = 100$ and parameters of the forgetting function (6) chosen as in case (b). Even though the bias of the estimators reduce, the overall pattern of performances remains the same. The Turnbull estimator (duration) appears to have much smaller bias when forgetting is less prevalent.

Figures 3 and 4 show plots similar to Figure 1 and 2 for $n = 300$. There is a marked reduction in

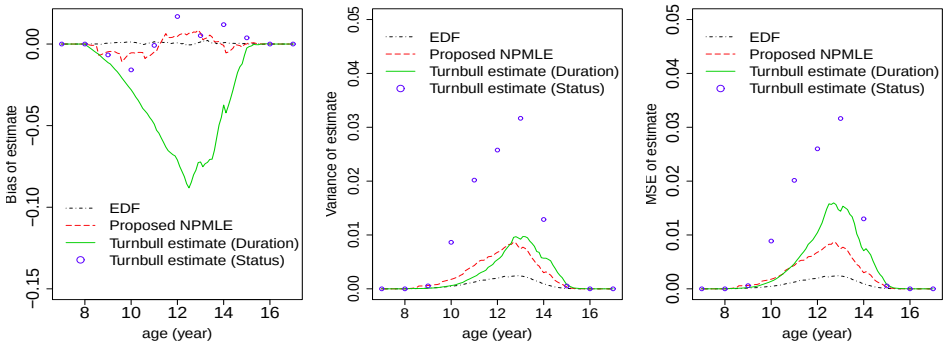

Figure 1: Comparison of bias, variance and MSE of the four estimator in case (a) and $n = 100$

the bias of the proposed NPNLE and the Turnbull estimator (status), while the previously observed pattern of relative performances continues to prevail.

Simulations for $n = 1,000$, leading to Figures 5 and 6, show that the bias and the variance of the proposed NPMLE continues to reduce with sample size. In contrast, the bias of the Turnbull estimator (duration) appears to have stagnated.

We now turn to the performance of the bootstrap estimator of variance. Figure 7 shows the plot of the average (across 500 runs) of the bootstrap estimate of variance of the proposed NPMLE shown in panel (I) and the sample variance (across 500 runs) of the NPMLE in panel (II). The corresponding plots for the other estimators are also shown. Figure 8 shows the standard error (across 500 runs) of the bootstrap estimator of variance, alongside the average (across 500 runs) of the same. It is seen that the standard error is generally much smaller than the average. Thus, the bootstrap estimator of variance appears to be a reasonable one.

These plots are for $n = 1,000$ and the parameters of the forgetting function chosen as in case (b). The two sets of the plots agree. Plots for other sample sizes and other values of parameters, which show similar patterns, are omitted for the sake of brevity.

5. DATA ANALYSIS

In a recent anthropometric study conducted by the Biological Anthropology Unit of the Indian Statistical Institute in and around the city of Kolkata, India from 2005 to 2011 ((ISI 2012), p.108), a total of 2194 randomly selected individuals, aged between 7 and 21 years, were surveyed. The subjects were interviewed on or around their birthdays. The data set contains age, menarcheal

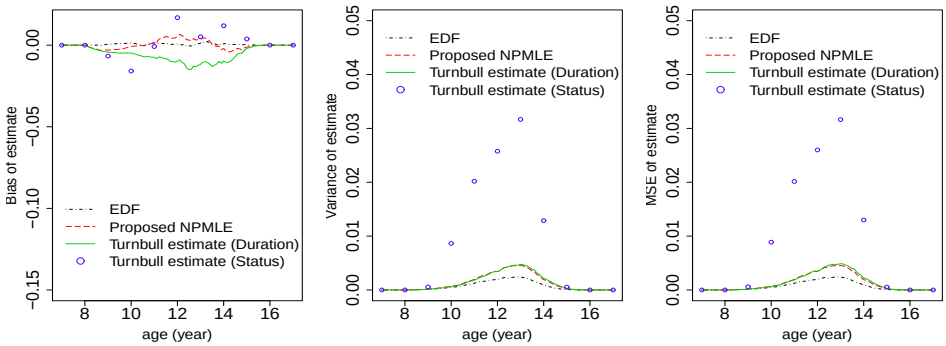

Figure 2: Comparison of bias, variance and MSE of the four estimator in case (b) and $n = 100$

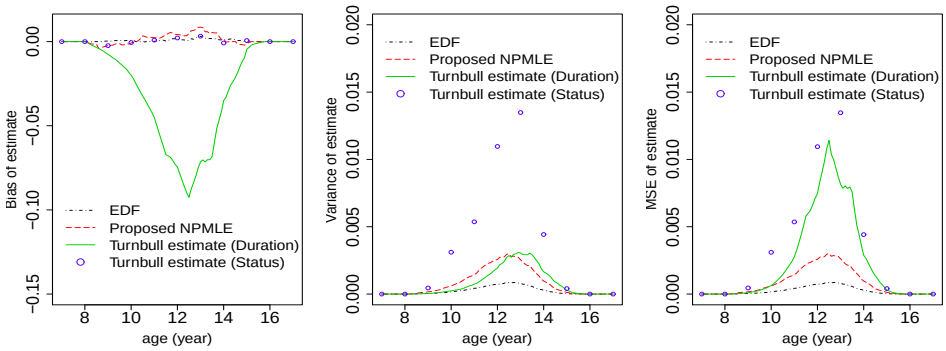

Figure 3: Comparison of bias, variance and MSE of the four estimator in case (a) and $n = 300$

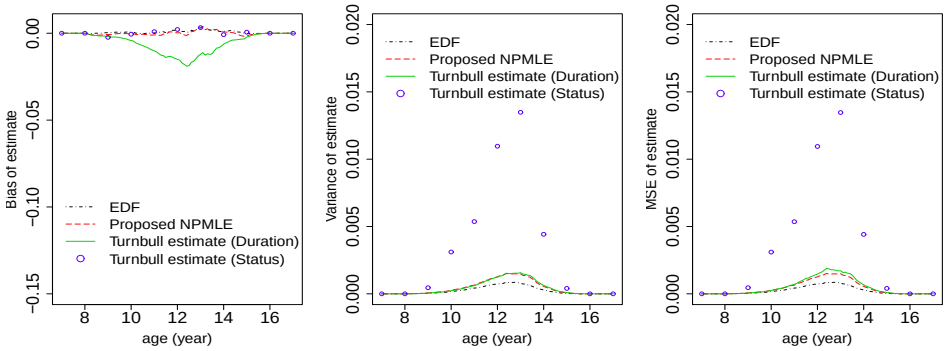

Figure 4: Comparison of bias, variance and MSE of the four estimator in case (b) and $n = 300$

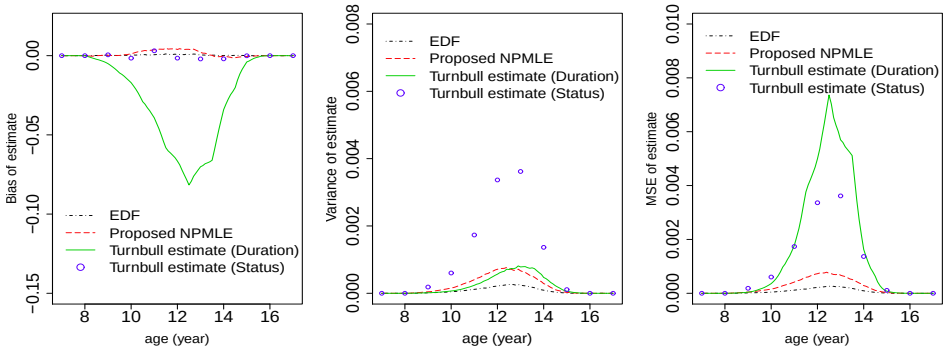

Figure 5: Comparison of bias, variance and MSE of the four estimator in case (a) and $n = 1000$

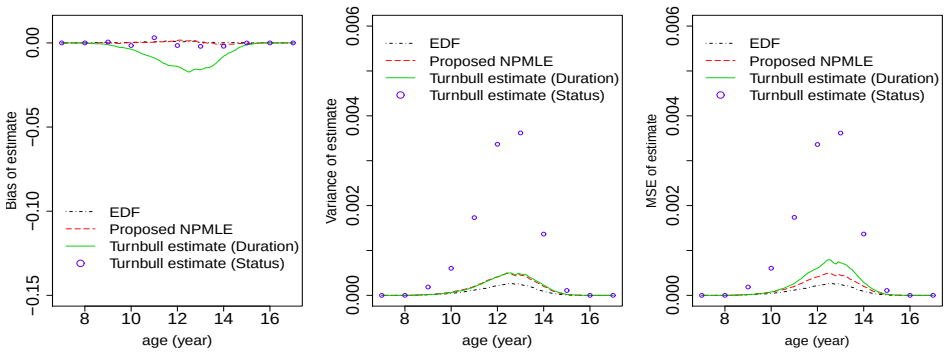

Figure 6: Comparison of bias, variance and MSE of the four estimator in case (b) and $n = 1000$

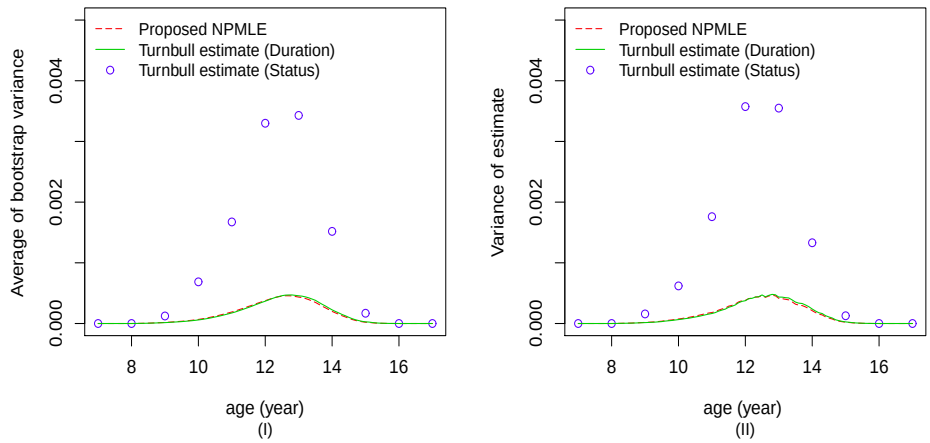

Figure 7: (I) Average of bootstrap variance and (II) Variance of estimated distribution function using three methods.

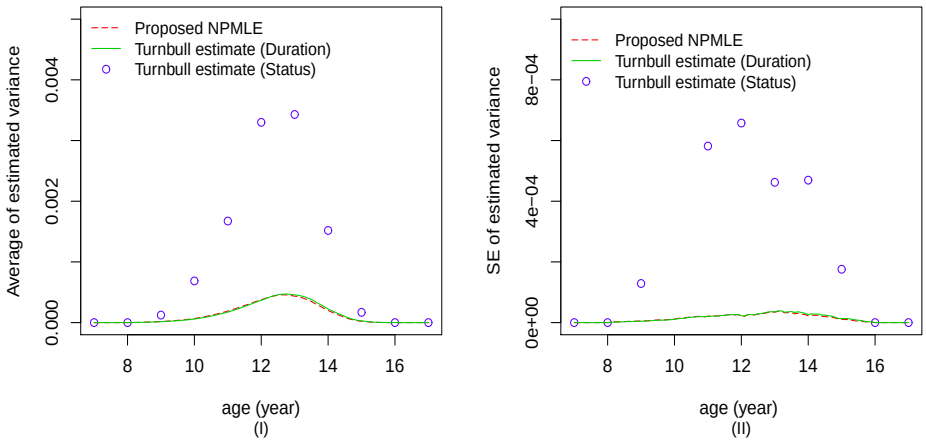

Figure 8: (I) Average and (II) Standard error of bootstrap variance using three methods.

status, age at menarche (if recalled), and some other information.

We used the same model for forgetting probability, as in the previous section. Figure 9 shows the estimated distribution function using the proposed NPMLE, the Turnbull estimator (duration) and the Turbull estimator (Status). As can be seen, the proposed NPMLE and the Turnbull (status) estimator are close to each other but Turnbull (durarion) estimator, which is expected to be biased, is far from them. Since the Turnbull estimator (status) is not uniquely defined except at integer ages, the proposed NPMLE may be preferred.

Plots of the bootstrap estimators of variance of the three estimators, shown in Figure 10, reveal that Turnbull (status) estimator has a much larger variance compared to the proposed NPMLE. The proposed method appears to produce a more accurate and precise estimate than the other two methods.

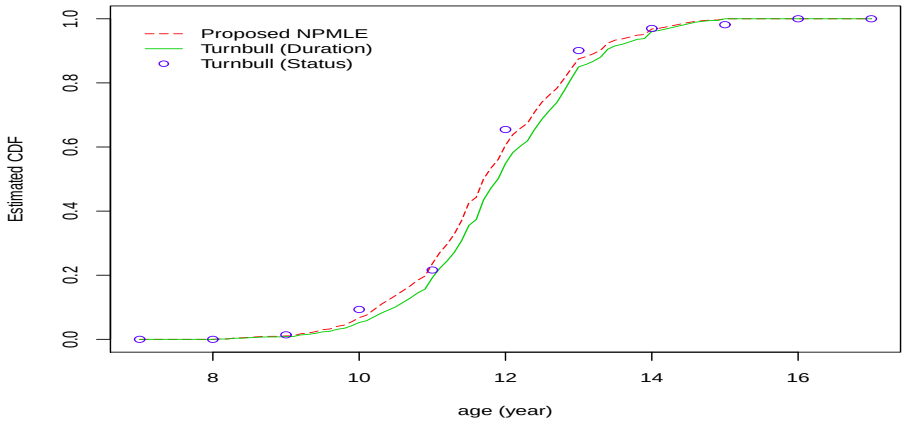

Figure 9: Estimated distribution function of data using three methods.

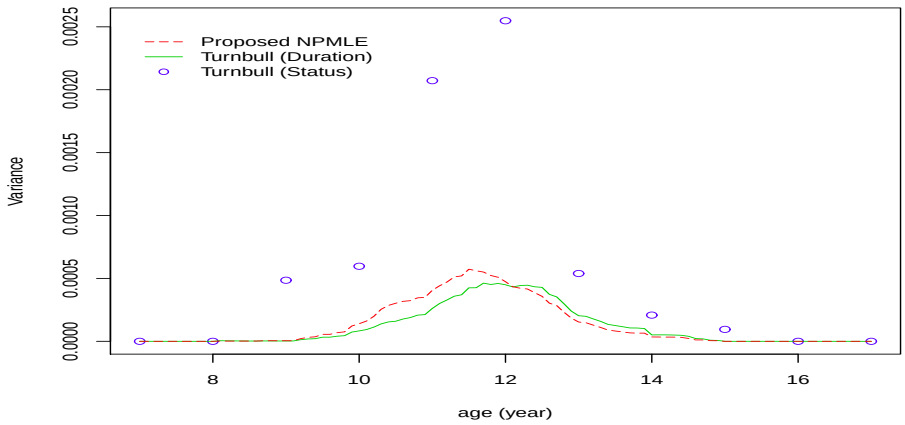

Figure 10: Variance of estimated distribution function of data using three methods.

APPENDIX

A. PROOF OF THEOREM 1

By definition of $\mathcal{C}$ and $\mathcal{C}_0$ we can rewrite the likelihood (10) as follows.

$$L = \prod_{i=1}^{n} \left[\left\{ \sum_{t=1}^{k} b_t \left(\sum_{\substack{r: l_{it} \in s_r \\ s_r \in \mathcal{C}_0}} p_r + \sum_{\substack{r: l_{it} \in s_r \\ s_r \in \mathcal{C} \setminus \mathcal{C}_0}} p_r \right) \right\}^{1-\varepsilon_i} \left\{ p_{l_i} \left(1 - \sum_{t=1}^{k} b_t I\big(T_i \in A_{it}\big) \right) \right\}^{\varepsilon_i} \right]^{\delta_i} \left[\sum_{\substack{r: l_i \in s_r \\ s_r \in \mathcal{C}_0}} p_r + \sum_{\substack{r: l_i \in s_r \\ s_r \in \mathcal{C} \setminus \mathcal{C}_0}} p_r \right]^{1-\delta_i} . \tag{A.1}$$

For any $s \in \mathcal{C} \setminus \mathcal{C}_0$, let $\mathcal{A}_s = \{I_{s'} : s' \in \mathcal{C}_0, s \subset s'\}$. By construction of $\mathcal{C}_0$, $\mathcal{A}_s$ is a non-empty set. The elements of $\mathcal{A}_s$ are disjoint sets consisting of unions of intervals, which are subsets of $[t_{min}, t_{max}]$. Let I_{s^*} be that member of $\mathcal{A}_s$ which satisfies the condition 'there is $\alpha \in I_{s^*}$ such that $\alpha < \beta$ whenever $\beta \in I_{s^+}$ for any $I_{s^+} \in \mathcal{A}_s$'. We shall show that by shifting mass from any I_{s_r} to I_{s^*}, there will be no reduction in the contribution of any individual to the likelihood (A.1).

We can check the effect of shifting mass on any other individual which is in s_r, as follows.

Case (i). Let $\delta_i = 0$. Since all supersets of s_r are in $\mathcal{C}_0$, there will be no change in contribution of any other individual which is in s_r.

Case (ii). Let $\delta_i \varepsilon_i = 1$. In this case singleton sets, consisting of every l_i, must belong to $\mathcal{C}_0$. Therefore the contribution to the likelihood can only increase.

Case (iii). Let $\delta_i(1 - \varepsilon_i) = 1$. Since all supersets of s_r are in $\mathcal{C}_0$, there will be no change in contribution of any other individual which is in s_r.

The foregoing discussion shows maximizing L can be restricted to $\{p_r : s_r \in \mathcal{C}_0\}$.

B. PROOF OF THEOREM 2

Define $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$ as index sets of individuals in the three different cases of censoring. The interview time is a discrete random variable with finite domain and also $x_1, x_2, \ldots, x_k$ are finite. Therefore, even when n is large, there is at most a finite number (say N) of distinct sets of the form

$$A_s = \Big\{ \bigcap_{i \in s} B_i \Big\} \bigcap \Big\{ \bigcap_{i \in \mathcal{S}_1 \cup \mathcal{S}_3 \setminus s} B_i^c \Big\},$$

where $s \subseteq \mathcal{S}_1 \cup \mathcal{S}_3$. Denote $s^{(1)}, s^{(1)}, \ldots, s^{(N)}$, the index sets corresponding to the N distinct set described above.

Consider a member of $\mathcal{A}_0$, say I_s, where s is a subset of $\{1, 2, \ldots, n\}$. If $s \subseteq \mathcal{S}_2$, then it is already a singleton. If not, it can be written as $s^{(j)} \cup (s \backslash s^{(j)})$, with $s^{(j)} \subseteq \mathcal{S}_1 \cup \mathcal{S}_3$ and $s \backslash s^{(j)} \subseteq \mathcal{S}_2$ for some $j \in \{1, 2, \ldots, N\}$. Let us consider three further special cases.

Case (a). Let $s = s^{(j)} \cup \{r\}$ for $r \in \mathcal{S}_2$. In this case, I_s is either a singleton or a null set. If it is a null set, then it cannot be a member of $\mathcal{A}$, and hence of $\mathcal{A}_0$. Thus, Case (a) contributes only singletons to $\mathcal{A}_0$.

Case (b). Let $s = s^{(j)} \cup \{r_1, r_2, \ldots, r_p\}$, for $r_1, r_2, \ldots, r_p \in \mathcal{S}_2$ when $p > 1$. In this case I_s is either a singleton or a null set. Since the absolute continuity of the menarcheal age distribution almost surely precludes coincidence of two sample values (say, T_{r_1} and T_{r_2}), I_s is a null set with probability 1. In summary, Case (b) cannot contribute anything other than a singleton to $\mathcal{A}_0$.

Case (c). Let $s = s^{(j)}$. The probability that a specific individual (say, the i-th one) has menarche at an age contained in $A_{s^{(j)}}$ is

$$P(T_i \in A_{s^{(j)}}, \delta_i \epsilon_i = 1).$$

Since this quantity is strictly positive, the probability that none of the n individuals have had menarche in $A_{s^{(j)}}$ and recalled the date is

$$\left(1 - P(T_i \in A_{s^{(j)}}, \delta_i \epsilon_i = 1)\right)^n,$$

which goes to zero as $n \to \infty$. Thus, the probability that there is $i \in \mathcal{S}_2$ such that $T_i \in A_{s^{(j)}}$ goes to one as $n \to \infty$. Therefore, $I_{s^{(j)} \cup \{i\}} = I_{s^{(j)}} \cap \{T_i\}$ is non-null. It follows that $P[I_s \notin \mathcal{A}_0]$ goes to one.

The statement of the theorem follows by combining the three cases.

C. PROOF OF THEOREM 3

From (11), the log-likelihood is given by

$$l(\boldsymbol{p}) = \sum_{i=1}^{n} \left(\ln \left(\sum_{j=1}^{m} \beta_{ij} q_j \right) \right) \tag{A.2}$$

We shall show that the likelihood (A.2) is strictly concave, and use the facts that 'if there is a point $\boldsymbol{p}^*$ in domain of $\boldsymbol{p}$ that satisfies $\frac{\partial l(\boldsymbol{p}^*)}{\partial \boldsymbol{p}^*} = 0$ when $l(\boldsymbol{p})$ is a C^1 concave function, then $\boldsymbol{p}^*$ is a global maximum' and 'a strictly concave function with convex domain cannot possess more than one global maximum' (Simon and Blume 1994). In view of the facts that $\sum_{j=1}^{m} q_j = 1$ and $q_j \geq 0$, the convexity of domain is easy to see.

Let $\boldsymbol{B}$ be an $n \times m$ matrix with β_{ij} in the ijth position. The first derivative of (A.2) with respect to $\boldsymbol{p}$ is

$$\frac{\partial l(\boldsymbol{p})}{\partial \boldsymbol{p}} = \sum_{i=1}^{n} \frac{\boldsymbol{b}_i}{\boldsymbol{b}_i^T \boldsymbol{p}}$$

where $\boldsymbol{b}_i$ is the ith row of $\boldsymbol{B}$ matrix. The second derivative or the Hessian is

$$\frac{\partial l(\boldsymbol{p})}{\partial \boldsymbol{p} \partial \boldsymbol{p}^T} = -\sum_{i=1}^{n} \frac{\boldsymbol{b}_i \boldsymbol{b}_i^T}{(\boldsymbol{B}_i^T \boldsymbol{p})^2}$$

which is non-positive definite matrix. Hence $l(\boldsymbol{p})$ is concave which is sufficient to have a maximum. We need to show that the probability of the Hessian matrix being negative definite goes to one, that is, for any vector $\boldsymbol{u} \neq 0$,

$$P\left(\sum_{i=1}^{n} \frac{(\boldsymbol{b}_i{}^T \boldsymbol{u})^2}{(\boldsymbol{b}_i^T \boldsymbol{p})^2} = 0\right) \to 0.$$

In other words, we need to show that for any arbitrary vector $\boldsymbol{u} \neq 0$,

$$P\left(\boldsymbol{b}_i{}^T \boldsymbol{u} = 0, \quad \forall i\right) = P\left(\boldsymbol{B}\boldsymbol{u} = 0\right) \to 0. \tag{A.3}$$

It is clear from (12) that for an individual (say i) having exactly recalled menarcheal age, $\boldsymbol{b}_i$ has only one non-zero element. In this situation, the equation $\boldsymbol{b}_i{}^T \boldsymbol{u} = 0$ implies that the corresponding element of $\boldsymbol{u}$ is zero. Further, theorem 2 shows that, with probability tending to one, the columns of $\boldsymbol{B}$ correspond only to singleton members of $\mathcal{A}_0$ associated individuals with exact recall of menarcheal age. Therefore, with probability tending to one, the event $\boldsymbol{B}\boldsymbol{u} = 0$ coincides with the event $\boldsymbol{u} = 0$. This completes the proof.

D. PROOF OF THEOREM 4

We can incorporate the constraint $\sum_{j=1}^{m} q_j = 1$, by using the Lagrange multiplier, to maximize

$$\ell = \sum_{i=1}^{n} \left(\ln\left(\sum_{j=1}^{m} \beta_{ij} q_j \right) \right) + \lambda\left(\sum_{j=1}^{m} q_j - 1 \right) \tag{A.4}$$

By setting the derivative of ℓ with respect to λ equal to 0, we have

$$\frac{\partial \ell}{\partial \lambda} = \sum_{j=1}^{m} q_j - 1 = 0. \tag{A.5}$$

On the other hand, by setting the derivative of ℓ with respect to q_j's equal to 0, we obtain

$$\frac{\partial \ell}{\partial q_j} = \sum_{i=1}^{n} \frac{\beta_{ij}}{\sum_{r=1}^{m} \beta_{ir} q_r} - \lambda = 0 \qquad \forall j = 1, 2, \ldots, m. \tag{A.6}$$

By multiplying both sides of (A.6) by q_j and adding them over all values of j, we get

$$\sum_{j=1}^{m} \sum_{i=1}^{n} \frac{\beta_{ij} q_j}{\sum_{r=1}^{m} \beta_{ir} q_r} = \lambda \sum_{j=1}^{m} q_j, \tag{A.7}$$

which simplifies, after interchange of the summations and utilization of (A.5), to

$$\lambda = n. \tag{A.8}$$

By substituting into (A.6) the optimum value of λ obtained above, we have

$$\sum_{i=1}^{n} \frac{\beta_{ij}}{\sum_{r=1}^{m} \beta_{ir} q_r} = n \qquad \text{for } j = 1, \ldots, m,$$

which is equivalent to (15), the equation defining the self consistency estimator.

REFERENCES

Aksglaede, L., Sorensen, K., Petersen, J. H., Skakkebak, N. E., and Juul, A. (2009), "Recent decline in age at breast development: The Copenhagen Puberty Study," *Pediatrics*, 123,(5), 932–939.

Allison, P. D. (1982), "Discrete-Time Methods for the Analysis of Event Histories," *Sociological Methodology*, 13, 61–98.

Ayatollahi, S. M., Dowlatabadi, E., and Ayatollahi, S. A. (2002), "Age at menarche in Iran," *Ann. Hum. Biol.*, 29,(4), 355–362.

Beckett, M., DaVanzo, J., Sastry, N., Panis, C., and Peterson, C. (2001), "The quality of retrospective data: An examination of long-term recall in a developing country," *The Journal of Human Resources*, 36,(3), 593–625.

Bergsten-Brucefors, A. (1976), "A note on the accuracy of recalled age at menarche," *Ann. Hum. Biol.*, 3, 71–73.

Cameron, N. (2002), *Human Growth and Development* Academic Press.

Chumlea, W. C., Schubert, C. M., Roche, A. F., Kulin, H. E., Lee, P. A., Himes, J. H., and Sun, S. S. (2003), "Age at menarche and racial comparisons in US girls," *Pediatrics*, 11,(1), 110–113.

Demirjian, A., Goldstien, H., and Tanner, J. M. (1973), "A new system of dental age assessment," *Annals of Human Biology*, 45, 211–227.

Efron, B. (1967), "The two sample problem with censored data," *In Proc. 5th Berkely Symp. on Math. Statist. Prob.*, pp. 831–853.

Eveleth, P. B., and Tanner, J. M. (1990), *Worldwide Variation in Human Growth*, 2nd edn Cambridge University Press.

Hediger, M. L., and Stine, R. A. (1987), "Age at menarche based on recall data," *Ann. Hum. Biol.*, 14, 133–142.

Hosmer, D. W., and Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data* John Wiley & Sons.

ISI (2012), *Annual Report of the Indian Statistical Institute 2011-12* Indian Statistical Institute. Available at URL http://library.isical.ac.in/jspui/handle/10263/5345?mode=full.

Kaplan, E. L., and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *J. Amer. Statist. Assoc.*, 53, 457–481.

Keiding, N., Begtrup, K., Scheike, T. H., and Hasibeder, G. (1996), "Estimation from current-status data in continuous time," *Lifetime Data Anal.*, 2,(2), 119–129.

Korn, E. L., Graubard, B. I., and Midthune, D. (1997), "Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale," *Am J Epidemiol.*, 145, 72–80.

LeClere, M. J. (2005), "PREFACE Modeling Time to Event: Applications of Survival Analysis in Accounting, Economics and Finance," *Review of Accounting and Finance*, 4, 5–12.

McKay, H. A., Bailey, D. B., Mirwald, R. L., Davison, K. S., and Faulkner, R. A. (1998), "Peak bone mineral accrual and age at menarche in adolescent girls: A 6-year longitudinal study," *J. Pediatrics*, 13, 682–687.

Mirzaei, S. S., and Sengupta, D. (2013), "Parametric estimation of menarcheal age distribution based on recall data," *Thecnical Report No.ASD/2013/3, Applied Statistical Unit, Indian Statistical Institue*, 3. Available at URL www.isical.ac.in/ asu/TR/TechRepASD201303.pdf.

Rabe-Hesketh, S., Yang, S., and Pickles, A. (2001), "Multilevel models for censored and latent responses," *Stat Methods Med Res.*, 10,(6), 409–427.

Simon, C. P., and Blume, L. (1994), *Mathematics for economists*, New York: W W Norton.

Turnbull, B. W. (1976), "The empirical distribution function with arbitrarily grouped, censored and truncated data," *J. Roy. Statist. Soc. Ser. B*, 38, 290–295.