Scandinavian
Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics
doi: 10.1111/sjos.12107


# Parametric Estimation of Menarcheal Age Distribution Based on Recall Data

SEDIGHEH MIRZAEI SALEHABADI and DEBASIS SENGUPTA
*Applied Statistical Unit, Indian Statistical Inistitute*

RITUPARNA DAS
*Biological Anthropology Unit, Indian Statistical Inistitute*

**ABSTRACT. Menarche, the onset of menstruation, is an important maturational event of female childhood. Most of the studies of age at menarche make use of dichotomous (status quo) data. More information can be harnessed from recall data, but such data are often censored in a informative way. We show that the usual maximum likelihood estimator based on interval censored data, which ignores the informative nature of censoring, can be biased and inconsistent. We propose a parametric estimator of the menarcheal age distribution on the basis of a realistic model of the recall phenomenon. We identify the additional information contained in the recall data and demonstrate theoretically as well as through simulations the advantage of the maximum likelihood estimator based on recall data over that based on status quo data.**

*Key words:* age at menarche, informative censoring, interval censoring, maximum likelihood estimator, status quo data, Weibull distribution

## 1. Introduction

Age at menarche is an important aspect of female growth. The average age at menarche is a widely used as an indicator of population health, timing of maturation and nutritional status (Frisch, 1985; Eveleth, 1986; Anderson & Must, 2005). It is also widely used as a demographic indicator of population fecundity (Udry & Cliquet, 1982). Menarcheal age distribution has been used to assess reproductive risks (Sandler *et al.*, 1984; Parazzini *et al.*, 1997). Most of the attempts at estimating the menarcheal age distribution has been on the basis of dichotomous data, also known as 'status quo' data (see, e.g. Teilmann *et al.* (2009)) or 'current status data' (see, e.g. Betensky (2000) & Dunson & Dinse (2002)). Dichotomous responses (whether menarche has occurred till the day of observation) are easy to obtain by asking young or adult women if they have experienced menarche. When observations take place at designed ages, it is possible to make parametric inference based on a binomial type likelihood, where the probability of occurrence of menarche is determined by the presumed distribution. Improved inference may be possible on the basis of menarcheal age information, recorded prospectively or retrospectively.

In a prospective study, the subjects are tracked over a period of time, and the age at the menarcheal event is recorded (McKay *et al.*, 1998). Some subjects may be lost to follow up. Such a study leads to randomly right censored survival data. The likelihood for this type of censored data can be used for both non-parametric and parametric inference (Lawless, 1982). The non-parametric maximum likelihood estimator (MLE) is the well-known product limit estimator proposed by Kaplan & Meier (1958). However, continuous monitoring is a logistically difficult exercise, and periodic visits lead to grouping of data. When the grouping interval is not too small (e.g. 6 months as in (Towne *et al.*, 2005)), accuracy of inference may be affected.

In a retrospective study, respondents are generally asked to recall at what age they began menstruating. The recall data are prone to be censored (Roberts, 1994; Padez, 2003;

Morabia & Costanza, 1998). In case the subject fails to recall, it follows that the age at menarche lies within the interval ranging from the earliest possible age and the age on the day of interview. Many non-parametric and parametric methods have been developed over the years for the analysis of interval censored data (Turnbull, 1976; Miller, 1981; Frydman, 1994; Aggarwala, 2001; Lee & Wang, 2003). Interval censoring is typically assumed to be non-informative, in which case there is a notional non-observation window that is independent of the quantity being observed. If the observed quantity falls inside this window, one only observes the window. In the case of recall data arising out of cross-sectional studies, the non-observation window is likely to depend on the age at menarche. Rather, it is the age of the subject on the day of observation that may be assumed to be independent of the age at menarche. When menarche is found to have already occurred by that day, the chance of recall may be less for smaller ages at menarche. Thus, the censoring times would not be independent of the age at menarche, and the censoring would be informative. While there have been several approaches to handle informative censoring for various types of data (Scharfstein *et al.*, 2001; Scharfstein & Robins, 2002; Frisch, 1985; Finkelstien *et al.*, 2002; Dunson & Dinse, 2002; Kaciroti *et al.*, 2012), the models and methods proposed, there are specific to the emergent mechanism of censoring, which are different from the nature of censoring in the present case. One may seek an estimator, on the basis of a likelihood that makes use of the special nature of the data at hand.

We propose a new approach for estimating distribution of age at menarche, which uses the recall information through a realistic censoring model. Under this model, the non-recall probability is regarded as a function of the time since menarche. We demonstrate that the new approach produces more precise estimates than what can be achieved through status quo data, while the usual approach based on interval censoring can lead to biased and inconsistent estimates.

## 2. Model and estimation

Let the age at menarche of $n$ subjects, $T_i, (i = 1, 2, \ldots, n)$ be samples from the distribution $F_\theta$, where $\theta$ is a vector of parameters. The $i^{\text{th}}$ subject is visited at age $S_i$. It is assumed that the $S_i$'s are samples from another distribution and are independent of the $T_i$'s.

In the case of status quo data, one observes $(S_i, \delta_i), (i = 1, 2, \ldots, n)$ where $\delta_i = I_{(T_i \le S_i)}$, the indicator of the event $(T_i \le S_i)$. The likelihood is

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

where $\bar{F}_\theta(S_i) = 1 - F_\theta(S_i)$. Most researchers use MLE of $\theta$ on the basis of the aforementioned likelihood (Lee & Wang, 2003).

In a retrospective study, the subject may not recall clearly the age at menarche. Here, we ignore the possibility of the subject recalling an approximate age and regard such occurrence as a non-recall event. Let $\varepsilon_i$ be the indicator of recalling the age at menarche. Note that whenever $\delta_i = 1$ and $\varepsilon_i = 0$, it is known that $T_i < S_i$. If the underlying censoring mechanism is presumed to be non-informative, then the likelihood is

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

where $f_\theta$ is the probability density function corresponding to the distribution $F_\theta$. Aggarwala (2001) proposed the use of the MLE of $\theta$ on the basis of an extension of the aforementioned likelihood.

It has been pointed out in the previous section that non-informativeness of censoring is difficult to justify in the present context. The non-recall probability may depend on the age at interview and the age at menarche. We model this non-recall probability by the function $\pi(S,T) = P(\varepsilon = 0|\delta = 1)$. The likelihood according to this model is

$$\prod_{i=1}^{n}\left[\left(\int_0^{S_i} f_\theta(u)\pi(S_i,u)du\right)^{1-\varepsilon_i}[f_\theta(T_i)(1-\pi(S_i,T_i))]^{\varepsilon_i}\right]^{\delta_i}[\bar{F}_\theta(S_i)]^{1-\delta_i}. \tag{3}$$

In particular, the non-recall probability may depend on the time elapsed since menarche, $S_i - T_i$. We model $\pi(S,T)$ by $\pi_\eta(S-T)$, where $\pi_\eta$ is a family of increasing functions indexed by the parameter $\eta$. According to this model, the likelihood is

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

The MLE based on the aforementioned likelihood is expected to harness the information in the recall data without making unrealistic assumptions about censoring. The parameter $\eta$, which can be a vector, would have to be regarded as a nuisance parameter in the present context.

In an unpublished technical report, Stine & Small (1986) used MLE based on a special case of the aforementioned likelihood, where $\pi_\eta$ is presumed to be a piecewise constant function. They did not study the statistical properties of the estimator.

When $\pi_\eta$ is a constant, (4) becomes a constant multiple of (2). As a further special case, if $\pi_\eta = 1$, then (4) reduces to (1). When $\pi_\eta = 0$, that is, all recalls are perfect, the product likelihood (4) reduces to

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

which is the same as the likelihood for prospective data obtained from continuous monitoring. Thus, the model leading to the likelihood (4) is more general than the standard censoring models.

## 3. Large sample properties

The factors in the product likelihood (4) have different forms in different cases. For example, $T_i$ is used only when $\delta_i = 1$ and $\varepsilon_i = 1$. In order for the standard asymptotic results to be applicable, each factor of this likelihood has to be expressed as the density of some random vector in a suitable probability space.

We have already assumed that the $T_i$'s (menarcheal ages) are samples from the distribution $F_\theta$ and the $S_i$'s (ages on interview date) are samples from another distribution. Let $G$ be the common distribution of the $S_i$'s. Let

$$Z_i = (S_i - T_i)\varepsilon_i\delta_i, \tag{6}$$

where $\varepsilon_i$ and $\delta_i$ are as defined in the previous section. Note that the vector

$$Y_i = (S_i, Z_i, \delta_i), \tag{7}$$

is observed in all cases and contains all the requisite information.

We now show that the $i^{\text{th}}$ factor in the product likelihood (4) is in fact proportional to the density of $Y_i$. We prove this result in the succeeding text, after dropping the subscript $i$ for simplicity. The dominating probability measure used for defining this density is $\mu = \vartheta_1 \times \vartheta_2 \times \vartheta_3$ where $\vartheta_1$ is the counting or the Lebesgue measure, depending on whether $G$ is discrete or continuous, $\vartheta_2$ is the sum of the counting and the Lebesgue measures, and $\vartheta_3$ is the counting measure (Ash, 2000).

**Theorem 1.** *The density of* $\boldsymbol{Y} = (S, Z, \delta)$ *with respect to the measure* $\mu$ *is*

$$f(s,z,\delta) = \begin{cases} g(s)\bar{F}_\theta(s) & \text{if } z = 0 \text{ and } \delta = 0, \\ g(s)\int_0^s f_\theta(u)\pi_\eta(s-u)du & \text{if } z = 0 \text{ and } \delta = 1, \\ g(s)f_\theta(s-z)(1-\pi_\eta(z)) & \text{if } z > 0 \text{ and } \delta = 1, \\ 0 & \text{otherwise}. \end{cases} \tag{8}$$

*Proof.* See the Appendix. □

The likelihood (4) can be written in terms of $S_i$, $Z_i$ and $\delta_i$ as

$$\prod_{i=1}^{n}\left[\left(\int_0^{S_i} f_\theta(u)\pi_\eta(S_i-u)du\right)^{I(Z_i=0)}[f_\theta(S_i-Z_i)(1-\pi_\eta(Z_i))]^{I(Z_i>0)}\right]^{\delta_i}[\bar{F}_\theta(S_i)]^{1-\delta_i},$$
$$= \frac{\prod_{i=1}^n f(S_i,Z_i,\delta_i)}{\prod_{i=1}^n g(S_i)}. \tag{9}$$

The numerator is a product of densities of the form (8), while the denominator does not contain any information about $\theta$. This likelihood can also be interpreted as a product of conditional densities of $(Z_i, \delta_i)$ given $S_i$, for $i = 1, 2, \ldots, n$. Further, this conditional likelihood is free from $g$; that is, inference for $\theta$ can proceed by ignoring any parameter of $g$.

Once the likelihood (4) is identified as a product of densities, standard results for consistency and asymptotic normality of the MLE become applicable. We would look for conditions on the variables $S_i$, $T_i$ and $\varepsilon_i$, which completely determine the observable triplet $(S_i, Z_i, \delta_i)$. Because the likelihood involves only the conditional density of $(Z_i, \delta_i)$ given $S_i$, it suffices to look for conditions on the distribution of $(T_i, \varepsilon_i)$ only. Specifically, the conditions would involve the density $f_\theta$, the density of $T_i$ and the function $\pi_\eta$, which defines the conditional density of the binary random variable $\varepsilon_i$ given $T_i$ and $S_i$.

It may be verified that the following conditions imply the sufficient conditions for consistency given in theorem 7.1.1 of Lehman (1999).

(C1) The parameter $\theta$ is identifiable with respect to the family of densities $f_\theta$ of the menarcheal age, and the parameter $\eta$ is identifiable with respect to the family of functions $\pi_\eta$ representing non-recall probability. In other words, $\theta_1 \neq \theta_2$ implies that $f_{\theta_1} \neq f_{\theta_2}$, and $\eta_1 \neq \eta_2$ implies that $\pi_{\eta_1} \neq \pi_{\eta_2}$.

(C2) The parameter spaces for $\theta$ and $\eta$ are open.

(C3) The random variables $T_i$, $i = 1, 2, \ldots, n$ are samples from the density $f_\theta$, and $\varepsilon_i$'s are independent with
$P(\varepsilon_i = 1 | T_i = t, S_i = s, t < s) = \pi_\eta(s-t)$.

(C4) The sets $A_1 = \{t : f_\theta(t) > 0\}$ and $A_2 = \{z : \pi_\eta(z) > 0\}$ are independent of $\theta$ and $\eta$, respectively.

(C5) The function $f_\theta(t)$ is differentiable with respect to $\theta$ for all $t$ such that the derivative is absolutely bounded by a $\mu$-integrable function $h_1(t)$, and the function $\pi_\eta(z)$ is differentiable with respect to $\eta$ for all $z$ such that the derivative is absolutely bounded by a $\mu$-integrable function $h_2(z)$,

It can be easily seen that Conditions C1–C4 imply Conditions C1–C4 of theorem 7.1.1 of Lehman (1999) in the present case. The Condition C5 implies that the quantities $\int_0^s \frac{\partial}{\partial\theta} f_\theta(u)\pi_\eta(s-u)du$ and $\int_0^s f_\theta(u)\frac{\partial}{\partial\eta}\pi_\eta(s-u)du$ are well defined, and are the derivatives of the conditional density of $(Z_i, \delta_i)$ given $S_i$ with respect to $\theta$ and $\eta$, respectively, in the case $z = 0$ and $\delta = 1$. It is easier to establish the corresponding implications in the other cases, which lead to the fulfilment of Condition C5 of theorem 7.1.1 of Lehman (1999).

The additional conditions for asymptotic normality relate to the log-likelihood obtained from (4),

$$\ell(\theta,\eta) = \sum_{i=1}^{n}\left[\delta_i(1-\varepsilon_i)\log\left(\int_0^{S_i} f_\theta(u)\pi(S_i-u)du\right) + \delta_i\varepsilon_i\log\left(f_\theta(T_i)(1-\pi(S_i-T_i))\right) + (1-\delta_i)\log\left(\bar{F}_\theta(S_i)\right)\right]. \tag{10}$$

The following conditions, together with C1–C5, ensure asymptotic normality of the MLE of $\theta$ and $\eta$ (Ferguson, 1996).

(C6) Second partial derivatives of $\ell(\theta,\eta)$ with respect to $\theta$ and $\eta$ exist and are continuous, and may be passed under the integral sign in $\int \ell(\theta,\eta)d\mu$.

(C7) The elements of the matrix

$$A(\theta,\eta) = \begin{bmatrix} \frac{\partial^2}{\partial\theta\partial\theta^T}\ell(\theta,\eta) & \frac{\partial^2}{\partial\theta\partial\eta^T}\ell(\theta,\eta) \\ \frac{\partial^2}{\partial\eta\partial\theta^T}\ell(\theta,\eta) & \frac{\partial^2}{\partial\eta\partial\eta^T}\ell(\theta,\eta) \end{bmatrix},$$

are bounded in absolute value, uniformly in some neighbourhood of the true value of the parameter $(\theta,\eta)$, by some function $K(x)$ such that $E_{(\theta_0,\eta_0)}K(X) < \infty$.

(C8) The Fisher information matrix

$$I(\theta,\eta) = E\begin{bmatrix} \left(\frac{\partial}{\partial\theta}\ell(\theta,\eta)\right)\left(\frac{\partial}{\partial\theta}\ell(\theta,\eta)\right)^T & \left(\frac{\partial}{\partial\theta}\ell(\theta,\eta)\right)\left(\frac{\partial}{\partial\eta}\ell(\theta,\eta)\right)^T \\ \left(\frac{\partial}{\partial\eta}\ell(\theta,\eta)\right)\left(\frac{\partial}{\partial\theta}\ell(\theta,\eta)\right)^T & \left(\frac{\partial}{\partial\theta}\ell(\theta,\eta)\right)\left(\frac{\partial}{\partial\theta}\ell(\theta,\eta)\right)^T \end{bmatrix},$$

is non-singular.

## 4. Theoretical comparison of estimates

### *4.1. Bias of maximum likelihood estimator based on interval likelihood*

If one ignores the informative nature of censoring, then the likelihood (2) would appear to be appropriate. We now show that an MLE based on that likelihood may be inconsistent under the general censoring model of Section 2. Inconsistency is established if the bias can be shown not to go to zero as the sample size goes to infinity. As the MLE based on (2) is not generally available in closed form, we avoid computing the asymptotic bias and compute instead the expected value of the score function obtained from the likelihood (2), computed under the general model.

Let $f_\theta(t) = \frac{1}{\theta}e^{-\frac{t}{\theta}}$ and $\pi_\eta(u) = 1 - e^{-\frac{u}{\eta}}$. The derivative of the log-likelihood obtained from (2) with respect to $\theta$ is

$$\sum_{i=1}^{n}\left[\delta_i(1-\varepsilon_i)\left(\frac{\frac{s_i}{\theta^2}e^{\frac{-s_i}{\theta}}}{1-e^{\frac{-s_i}{\theta}}}\right)+\delta_i\varepsilon_i\left(\frac{-1}{\theta}+\frac{t_i}{\theta^2}\right)+(1-\delta_i)\frac{s_i}{\theta^2}\right]. \tag{11}$$

The expectation of (11) with respect to the general model of Section 2 is

$$E_S\left[\frac{S}{\theta^2}\bar{F}_\theta(S)+\int\left(\frac{-S}{\theta}+\frac{t}{\theta^2}\right)(1-\pi_\eta(S-t))f_\theta(t)dt+\frac{\frac{S}{\theta^2}e^{\frac{-S}{\theta}}}{1-e^{\frac{-S}{\theta}}}\int\pi_\eta(S-t)f_\theta(t)dt\right].$$

In the further special case $\eta = \theta$, the aforementioned expression reduces to

$$E_S\left[\frac{1}{2\theta}\frac{\frac{S}{\theta}e^{-\frac{S}{\theta}}}{1-e^{-\frac{S}{\theta}}}\left(2-2e^{-\frac{S}{\theta}}-\frac{S}{\theta}-\frac{S}{\theta}e^{-\frac{S}{\theta}}\right)\right].$$

For the expectation to be equal to zero, the function in square brackets should be orthogonal to the probability function of $S$, which would not hold in general. One can design infinitely many distribution of $S$, which would violate this condition. If the expected value of the score function obtained from (2) is not zero, the asymptotic bias of the corresponding 'MLE' is also not zero.

### *4.2. Additional information from recall data*

In order to identify the additional information arising from recall data, we return to the expression of the likelihood in terms of the joint density of $(S, Z, \delta)$. We presume that the distribution of $S$ does not involve any unknown parameter. Then the joint density of the observed triplet can be written as

$$f_{\theta,\eta}(s,z,\delta) = f_\theta(s,\delta)f_{\theta,\eta}(z|s,\delta).$$

Thus, the log-likelihood for a single sample is

$$\log(f_{\theta,\eta}(s,z,\delta)) = \log(f_\theta(s,\delta)) + \log(f_{\theta,\eta}(z|s,\delta)),$$

and consequently, information for the two parameters is of the form

$$I_R(\theta,\eta) = I_S(\theta,\eta) + I_A(\theta,\eta), \tag{12}$$

where the matrices $I_R$, $I_S$ and $I_A$ are the information arising from recall data, status quo data and recall data conditioned on status quo data, respectively.

Because the likelihood of status quo data is free from $\eta$, $I_S(\theta,\eta)$ is a function of $\theta$ alone and can be written as

$$I_S(\theta,\eta) = \begin{bmatrix} I_1 & 0 \\ 0 & 0 \end{bmatrix},$$

where

$$I_1 = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log(f_\theta(s,\delta))\right].$$

On the other hand, the additional information obtain from the recall data is

$$I_A(\theta,\eta) = \begin{bmatrix} I_2 & I_3 \\ I_3^T & I_4 \end{bmatrix},$$

where

$$I_2 = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log(f_{\theta,\eta}(z|s,\delta))\right],$$
$$I_3 = -E\left[\frac{\partial^2}{\partial\theta\partial\eta^T}\log(f_{\theta,\eta}(z|s,\delta))\right],$$
$$I_4 = -E\left[\frac{\partial^2}{\partial\eta\partial\eta^T}\log(f_{\theta,\eta}(z|s,\delta))\right].$$

In particular, the additional information of $\theta$, the parameter of interest, is

$$I_2 - I_3 I_4^{-1} I_3^T.$$

When $\eta$ is known, the additional information reduces to $I_2$.

As an example, consider the special case, where $f_\theta(t) = \frac{1}{\theta}e^{-\frac{t}{\theta}}$ and $\pi_\eta(z) = 1 - e^{-z/\eta}$. Figure 1 shows plots of the information arising from status quo data ($I_1$), from recall data ($I_1 + I_2 - I_3 I_4^{-1} I_3^T$) and from recall data with known $\eta$ ($I_1 + I_2$), for different values of $\eta$ and a range of values of $\theta$. It can be seen that when $\eta$ is large, there is a considerable gap between the first two, while there is not much gap between the second and the third curves. Thus, in this case, the price for not knowing the nuisance parameter $\eta$ is minimal compared with the gain from recall data. On the other hand, for a small value of $\eta$ (i.e. menarcheal age forgotten quickly), recall data do not augment the information noticeably.

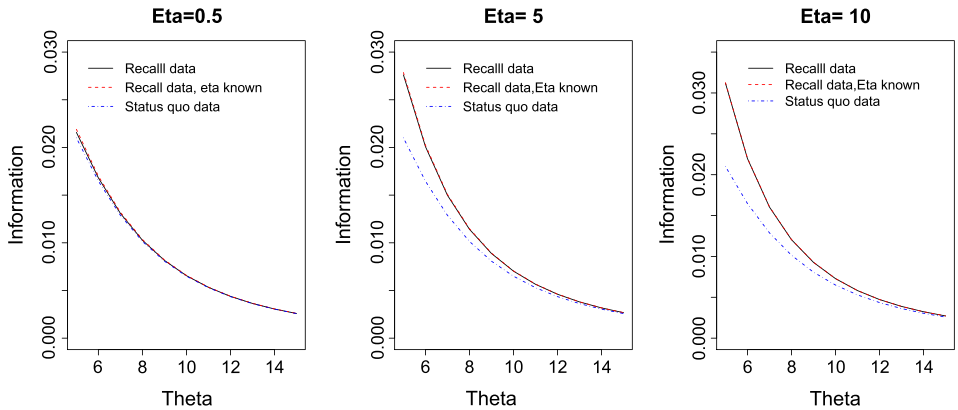


*Fig. 1.* Information based on recall data and status quo likelihoods.

## 5. Simulation results

For the purpose of simulation, we assume that 'age at menarche' follows the Weibull distribution with shape and scale parameters $\alpha$ and $\beta$, respectively. Thus, $\theta = (\alpha, \beta)$. Further, we assume that 'age at interview' follows the discrete uniform distribution over [7,21] and that

$$\pi_\eta(x) = 1 - e^{-\frac{x}{\eta}}. \tag{13}$$

We use the following values of the parameters:

(i) $\alpha = 11$, $\beta = 13$ and $\eta = 3$ ;
(ii) $\alpha = 10$, $\beta = 12$ and $\eta = 5$.

The two choices correspond to median ages at menarche of about 11.57 and 12.58 years, and inter-quantile ranges of about 1.78 and 1.80 years, respectively. The mean times to forget are 3 and 5 years, respectively.

We compare the performance of MLEs based on the status quo likelihood (1), the interval censoring likelihood (2) and the recall data likelihood (4) for our model. Computation of MLEs in all the cases is carried out through numerical optimization of likelihood using 'quasi-Newton' method (Nocedal & Wright, 2006).

We run 1000 simulations for each of the aforementioned combinations of parameters, for sample sizes $n = 50$, 500 and 1000.

Table 1 shows the bias, the standard deviation, the mean squared error (MSE) and the Cramer–Rao lower bound for the MLEs of the three parameters based on the three likelihoods, for the combination of parameter values in cases (i) and (ii).

In both cases, it is found that the bias for the MLE based on interval censoring likelihood stabilizes around a positive constant when the sample size increases. When the sample size is small, there is bias in the MLE from status quo data, but it reduces as the sample size increases. The standard deviation of the MLE based on our model is smaller than that based on status quo data and is also in line with the Cramer–Rao lower bound—particularly when the sample size is large.

In order to check the robustness of the proposed method against departure from the assumed form of the non-recall probability function $\pi_\eta$, we use the following non-recall function for data generation.

$$\pi_\eta(x) = 0.05I(0 < x \leq 2.5) + 0.35I(2.5 < x \leq 4.5) + 0.95I(4.5 < x < \infty). \tag{14}$$

We generate the data from two different models.

(iii) The 'age at menarche' from Weibull distribution with parameters $\alpha = 11$ and $\beta = 13$ and the $\pi_\eta$ function defined in (14),
(iv) The 'age at menarche' from Weibull distribution with parameters $\alpha = 10$ and $\beta = 12$ and the $\pi_\eta$ function defined in (14).

We run 1000 simulations for each of the above combinations of parameters, for sample sizes $n = 50$, 500 and 1000. Table 2 shows the performance of MLEs based on the status quo likelihood (1), the interval censoring likelihood (2) and the recall data likelihood (4) based on the incorrect model (13). We compute the bias, the standard deviation and the MSE for the MLEs of the parameters of interest, on the basis of the three likelihoods, for the combination of parameter values in cases (iii) and (iv).

Table 1. *Bias, Stdev, MSE and CRLB of estimated parameters in case (i)* $\alpha = 11$, $\beta = 13$, $\eta = 3$ *and case (ii)* $\alpha = 10$, $\beta = 12$ ,$\eta = 5$

| | | | $n = 50$ | | | $n = 500$ | | | $n = 1000$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Estimator | Property | Case | $\alpha$ | $\beta$ | $\eta$ | $\alpha$ | $\beta$ | $\eta$ | $\alpha$ | $\beta$ | $\eta$ |
| MLE from status quo | Bias | (i) | 7.302 | −0.103 | — | 0.647 | 0.002 | — | 0.499 | 0.001 | — |
| | Stdev | | 13.67 | 0.519 | — | 1.361 | 0.149 | — | 0.923 | 0.104 | — |
| | MSE | | 239.9 | 0.280 | — | 2.271 | 0.022 | — | 1.101 | 0.011 | — |
| | CRLB | | 14.135 | 0.217 | — | 1.672 | 0.021 | — | 0.850 | 0.010 | — |
| MLE from interval censoring | Bias | | 2.788 | 0.193 | — | 1.410 | 0.223 | — | 1.332 | 0.220 | — |
| | Stdev | | 4.528 | 0.318 | — | 0.854 | 0.102 | — | 0.589 | 0.069 | — |
| | MSE | | 28.26 | 0.138 | — | 2.721 | 0.062 | — | 2.122 | 0.053 | — |
| | CRLB | | 1.933 | 0.043 | — | 0.255 | 0.004 | — | 0.126 | 0.003 | — |
| MLE from our method | Bias | | 1.563 | −0.016 | 0.058 | 0.325 | 0.008 | −0.005 | 0.250 | 0.006 | −0.0002 |
| | Stdev | | 4.100 | 0.310 | 0.875 | 0.777 | 0.100 | 0.242 | 0.545 | 0.069 | 0.169 |
| | MSE | | 19.239 | 0.096 | 0.769 | 0.709 | 0.010 | 0.058 | 0.360 | 0.005 | 0.028 |
| | CRLB | | 5.266 | 0.043 | 0.619 | 0.566 | 0.004 | 0.062 | 0.288 | 0.004 | 0.030 |
| MLE from status quo | Bias | (ii) | 8.589 | −0.047 | — | 1.083 | 0.048 | — | 0.938 | 0.032 | — |
| | Stdev | | 9.529 | 0.507 | — | 1.239 | 0.148 | — | 0.873 | 0.100 | — |
| | MSE | | 164.6 | 0.259 | — | 2.707 | 0.024 | — | 1.641 | 0.011 | — |
| | CRLB | | 68.83 | 0.121 | — | 1.571 | 0.019 | — | 0.757 | 0.010 | — |
| MLE from interval censoring | Bias | | 2.391 | 0.198 | — | 1.369 | 0.217 | — | 1.317 | 0.213 | — |
| | Stdev | | 2.667 | 0.287 | — | 0.614 | 0.088 | — | 0.431 | 0.061 | — |
| | MSE | | 12.83 | 0.121 | — | 2.250 | 0.055 | — | 1.919 | 0.049 | — |
| | CRLB | | 2.327 | 0.027 | — | 0.195 | 0.003 | — | 0.096 | 0.003 | — |
| MLE from our method | Bias | | 1.581 | 0.093 | 0.166 | 0.619 | 0.046 | 0.009 | 0.570 | 0.042 | 0.003 |
| | Stdev | | 2.500 | 0.281 | 1.392 | 0.585 | 0.086 | 0.391 | 0.411 | 0.060 | 0.284 |
| | MSE | | 8.747 | 0.088 | 1.963 | 0.726 | 0.009 | 0.152 | 0.493 | 0.005 | 0.081 |
| | CRLB | | 4.366 | 0.014 | 7.062 | 0.369 | 0.002 | 0.693 | 0.166 | 0.003 | 0.080 |

Stdev, standard deviation; MSE, mean squared error; CRLB, Cramer–Rao lower bound; MLE, maximum likelihood estimator.

Table 2. *Bias, Stdev and MSE of estimated parameters in case (iii) $\alpha = 11$ and $\beta = 13$ and the $\pi_\eta$ function defined in (14) and case (iv) $\alpha = 10$ and $\beta = 12$ and the $\pi_\eta$ function defined in (14)*

| | | | $n = 50$ | | $n = 500$ | | $n = 1000$ | |
|---|---|---|---|---|---|---|---|---|
| Estimator | Property | Case | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ |
| MLE from status quo | Bias | (iii) | 8.666 | −0.100 | 0.692 | −0.009 | 0.484 | 0.005 |
| | Stdev | | 14.945 | 0.515 | 1.330 | 0.152 | 0.928 | 0.102 |
| | MSE | | 298.2 | 0.275 | 2.259 | 0.023 | 1.095 | 0.010 |
| MLE from interval censoring | Bias | | 2.544 | 0.262 | 1.401 | 0.254 | 1.311 | 0.237 |
| | Stdev | | 3.221 | 0.306 | 0.734 | 0.096 | 0.547 | 0.068 |
| | MSE | | 16.839 | 0.162 | 2.502 | 0.074 | 2.019 | 0.061 |
| MLE from our method | Bias | | 2.014 | 0.117 | 0.899 | 0.111 | 0.812 | 0.101 |
| | Stdev | | 3.091 | 0.293 | 0.706 | 0.093 | 0.523 | 0.066 |
| | MSE | | 13.602 | 0.099 | 1.308 | 0.021 | 0.933 | 0.014 |
| MLE from status quo | Bias | (iv) | 9.039 | −0.066 | 1.158 | 0.045 | 0.881 | 0.040 |
| | Stdev | | 14.846 | 0.514 | 1.292 | 0.151 | 0.814 | 0.103 |
| | MSE | | 301.9 | 0.268 | 3.010 | 0.024 | 1.439 | 0.012 |
| MLE from interval censoring | Bias | | 2.644 | 0.287 | 1.598 | 0.286 | 1.486 | 0.269 |
| | Stdev | | 2.677 | 0.301 | 0.675 | 0.096 | 0.463 | 0.066 |
| | MSE | | 14.15 | 0.173 | 3.011 | 0.091 | 2.423 | 0.077 |
| MLE from our method | Bias | | 2.169 | 0.351 | 1.147 | 0.142 | 1.038 | 0.140 |
| | Stdev | | 2.583 | 0.289 | 0.655 | 0.092 | 0.450 | 0.063 |
| | MSE | | 11.37 | 0.207 | 1.745 | 0.029 | 1.281 | 0.024 |

Stdev, standard deviation; MSE, mean squared error; MLE, maximum likelihood estimator.

In both cases, the MSE of the MLEs based on our method is generally smaller than the same obtained from the two other methods but somewhat larger than the MSE reported in Table 1.

We now check the robustness of the method against the basic assumption that the non-recall probability function depends only on the time since menarche. In view of the possibility that some subjects having early menarche may remember the date even after a long time, we consider the alternative form of the non-recall probability function as follows.

$$\pi(S,T) = \begin{cases} 0.5\left(1 - e^{-\frac{S-T}{\eta}}\right) & \text{if } T < 9, \\ \left(1 - e^{-\frac{S-T}{\eta}}\right) & \text{if } T \geq 9. \end{cases} \tag{15}$$

Under the aforementioned model, very early menarcheal ages would be remembered more often, making these cases account for a larger share of exact recall cases, as compared with the model (13).

We generate data from two different models.

(v) The 'age at menarche' from Weibull distribution with parameters $\alpha = 11$ and $\beta = 13$, and the $\pi$ function defined in (15) when $\eta = 3$;
(vi) The 'age at menarche' from Weibull distribution with parameters $\alpha = 10$ and $\beta = 12$, $\eta = 5$ and the $\pi$ function defined in (15) when $\eta = 5$.

Table 3. *Bias, Stdev and MSE of estimated parameters of interest in case (v)* $\alpha = 11$ *and* $\beta = 13$*, and the* $\pi$ *function defined in (15) when* $\eta = 3$*, and (vi)* $\alpha = 10$ *and* $\beta = 12$*,* $\eta = 5$ *and the* $\pi$ *function defined in (15) when* $\eta = 5$

| | | | $n = 50$ | | | $n = 500$ | | | $n = 1000$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Estimator | Property | Case | $\alpha$ | $\beta$ | $P(T < 9)$ | $\alpha$ | $\beta$ | $P(T < 9)$ | $\alpha$ | $\beta$ | $P(T < 9)$ |
| MLE from status quo | Bias | (v) | 8.306 | −0.112 | 0.001 | 2.449 | −0.006 | 0.0004 | 0.136 | 0.0002 | −0.0001 |
| | Stdev | | 15.46 | 0.511 | 0.026 | 1.349 | 0.147 | 0.008 | 0.854 | 0.101 | 0.005 |
| | MSE | | 304.9 | 0.273 | 0.0007 | 1.878 | 0.022 | 0.00006 | 0.747 | 0.010 | 0.00003 |
| MLE from interval censoring | Bias | | 2.384 | 0.175 | −0.005 | 0.789 | 0.206 | −0.006 | 0.764 | 0.206 | −0.006 |
| | Stdev | | 4.852 | 0.341 | 0.012 | 0.806 | 0.103 | 0.004 | 0.546 | 0.072 | 0.002 |
| | MSE | | 29.21 | 0.147 | 0.0002 | 1.272 | 0.053 | 0.00005 | 0.883 | 0.048 | 0.00004 |
| MLE from our method | Bias | | 1.313 | −0.035 | 0.003 | −0.209 | −0.023 | 0.002 | −0.191 | −0.019 | 0.002 |
| | Stdev | | 4.951 | 0.335 | 0.019 | 0.743 | 0.103 | 0.005 | 0.498 | 0.071 | 0.003 |
| | MSE | | 26.213 | 0.114 | 0.0003 | 0.595 | 0.011 | 0.00004 | 0.284 | 0.005 | 0.00001 |
| MLE from status quo | Bias | (vi) | 7.425 | −0.103 | −0.002 | 0.286 | 0.023 | −0.001 | 0.129 | 0.010 | −0.001 |
| | Stdev | | 14.59 | 0.519 | 0.058 | 1.288 | 0.151 | 0.019 | 0.845 | 0.102 | 0.013 |
| | MSE | | 267.7 | 0.280 | 0.003 | 1.739 | 0.023 | 0.0004 | 0.731 | 0.010 | 0.0002 |
| MLE from interval censoring | Bias | | 1.211 | 0.222 | −0.012 | 0.528 | 0.155 | −0.012 | 0.478 | 0.155 | −0.012 |
| | Stdev | | 2.220 | 0.294 | 0.025 | 0.569 | 0.091 | 0.008 | 0.401 | 0.064 | 0.006 |
| | MSE | | 6.391 | 0.136 | 0.0007 | 0.603 | 0.032 | 0.0002 | 0.389 | 0.028 | 0.0002 |
| MLE from our method | Bias | | 0.558 | −0.046 | 0.004 | −0.135 | −0.022 | 0.003 | −0.088 | −0.021 | 0.003 |
| | Stdev | | 2.094 | 0.291 | 0.032 | 0.540 | 0.089 | 0.010 | 0.383 | 0.062 | 0.007 |
| | MSE | | 4.694 | 0.087 | 0.001 | 0.309 | 0.008 | 0.0001 | 0.154 | 0.004 | 0.00006 |

Stdev, standard deviation; MSE, mean squared error; MLE, maximum likelihood estimator.

We run 1000 simulations for each of the aforementioned combinations of parameters, for sample sizes $n = 50$, 500 and 1000. Table 3 shows the performance of MLEs based on the status quo likelihood (1), the interval censoring likelihood (2) and the recall data likelihood (4) under the model (13). In addition to the original parameters $\alpha$ and $\beta$, we consider the derived parameter $P(T < 9)$ representing the probability of very early menarche, which is expected to be overestimated when the exponential model (13) is assumed instead of (15). We compute the bias, the standard deviation and the MSE for the MLEs based on the three likelihoods, for the combination of parameter values in cases (v) and (vi).

In both cases, the bias, the standard deviation and the MSE of the MLEs based on our method are smaller than the same, computed from the two other methods. Further, the proposed method of $P(T < 9)$ is found to have a positive bias as expected. The amount of bias is not very large. Performances of the MLEs of $\alpha$ and $\beta$ are in line with that reported in Table 1, where there was no specification error in the non-recall probability function.

## 6. Adequacy of model

In order to check how well the assumed parametric model actually fits the data, one can use the chi-square goodness-of-fit test (Gibbons & Chakraborti, 2003). For this purpose, the data may be transformed to the trivariate vector $\boldsymbol{Y} = (S, Z, \delta)$, and the support of the joint distribution of this vector may be appropriately partitioned, depending on the availability of data. An example is given in the next section.

Modelling of the non-recall function can be a critical issue. There would be a trade-off between a flexible model with many parameters (nuisance parameters in the present context) on the one hand, and a parsimonious but restrictive model on the other. The following exploratory technique may be used as a guideline for selecting the functional form of the non-recall probability $\pi$. Assume $\pi$ has the form

$$\pi(x) = b_1 I(x_1 < x \le x_2) + b_2 I(x_2 < x \le x_3) + \ldots + b_k I(x_k < x < \infty), \tag{16}$$

where $k$ is large integer, $x_1, x_2, \ldots, x_k$ are a chosen set of time-points in increasing order and $b_1, b_2, \ldots, b_k$ are unspecified parameters taking values in the range $[0, 1]$. In view of (16), the likelihood (4) reduces to

$$L = \prod_{i=1}^{n} \left[ \left\{ \sum_{l=1}^{k} b_l \left( F_\theta(S_i - x_l) - F_\theta(S_i - x_{l+1}) \right) \right\}^{1-\varepsilon_i} \left\{ f_\theta(T_i) \left( 1 - \sum_{l=1}^{k} b_l I\left( S_i - x_{l+1} < T_i \le S_i - x_l \right) \right) \right\}^{\varepsilon_i} \right]^{\delta_i} [\bar{F}_\theta(S_i)]^{1-\delta_i}. \tag{17}$$

If the distribution of $T$ is known, one can obtain the MLE of the parameters $b_1, b_2, \ldots, b_k$. The Hessian matrix with respect to the $b_l$'s ($l = 1, 2, \ldots, k$) can easily be shown to be non-negative definite. Therefore, there is a unique maximum of the likelihood function for these parameters. One can use Newton–Raphson iterative steps to determine the conditional MLE of the piecewise constant function $\pi$, for any given $F_\theta$. While using a parametric form $\pi_\eta$, one can first estimate the MLEs $\hat{\theta}$ and $\hat{\eta}$ and then compare the plot of $\pi_{\hat{\eta}}$ with the plot of the conditional MLE of the piecewise constant version of $\pi$ with large $k$, with $F_\theta$ held fixed at $F_{\hat{\theta}}$. This graphical comparison can be used to judge the suitability of the function $\pi_\eta$.

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

We used the Weibull model for menarcheal age and the exponential model for non-recall probability, as in the previous section, and used the three different methods mentioned in that section to estimate the parameters as well as the median of age at menarche. Table 4 gives a summary of the findings. Figure 2(A) shows the plot of the survival functions corresponding to the three sets of estimates.

The median estimated from our method is close to the median estimated from the status quo likelihood, but the confidence interval based on our estimate is narrower. The standard errors of the distributional parameters are also smaller. It is also seen that the median estimated from the interval censoring likelihood, which ignores the informative nature of censoring, is different from the other two estimates. The corresponding 95% confidence interval does not have any

Table 4. *Estimated parameters and median age at menarche from different methods for real data*

| | Estimate (standard error) | | | | 95% confidence interval |
|---|---|---|---|---|---|
| Estimator | $\alpha$ | $\beta$ | $\eta$ | Median | of median |
| MLE from status quo | 10.74 (0.320) | 12.17 (0.005) | | 11.76 | (11.62, 11.90) |
| MLE from interval censoring | 11.80 (0.061) | 12.65 (0.001) | | 12.25 | (12.20, 12.30) |
| MLE from our method | 10.19 (0.090) | 12.21 (0.001) | 3.47 (0.140) | 11.78 | (11.72, 11.84) |

MLE, maximum likelihood estimator.

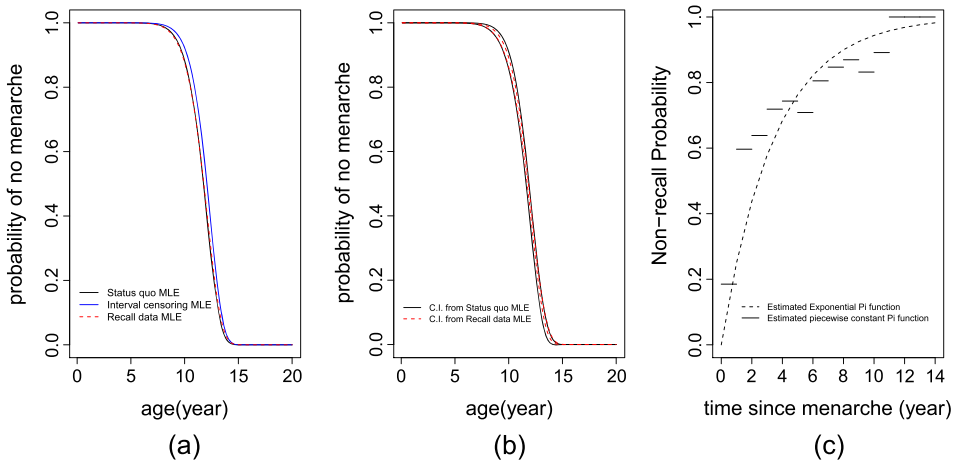


*Fig. 2.* (a) Survival plots for real data based on three methods, (b) confidence interval for probability of no menarche based on two methods and (c) plots of exponential and piecewise constant maximum likelihood estimator of $\pi$.

overlap with other two confidence intervals. The survival functions estimated from the three models, shown in Fig. 2(A), also shows that the MLE based on interval censoring likelihood is very different from the other two MLEs. This occurrence may be attributed to the bias of this MLE, which is expected even when the sample size is large (Sections 4.1 and 5).

Figure 2(B) shows the loci of upper and lower confidence limits for the probability of no menarche based on status quo MLE and recall data MLE. The latter pair of limits correspond to a narrower interval for any given age.

In order to check how well the assumed parametric model fits the data, we use the chi-square goodness-of-fit test, by categorizing the triplet $(S, Z, \delta)$ as follows:

The range of $S$ is split into the sets $\{7, 8, 9, 10, 11\}$ and $\{12, 13, 14, 15, 16, 17, 18, 19, 20, 21\}$;
The range of $Z$ is split into the sets $\{0\}$, $(0, 1.5]$ and $(1.5, 11]$;
The range of $\delta$ has two points, 0 and 1, in any case.

The combinations of these categories produce 12 bins. Further, there are three parameters to estimate. Thus, the null distribution should be $\chi^2$ with 8 degrees of freedom. The $p$-value of the test statistics for the given data happens to be 0.11. Therefore, the model can be said to be appropriate.

As we mentioned in the last section, one can check adequacy of the functional form of $\pi_\eta$ by comparing $\pi_{\hat{\eta}}$ with the conditional MLE of a piecewise constant function (16). We use segments of 1-year duration for this analysis. Note that for the given data, the largest value of $S_i - T_i$ in a perfectly recalled case happens to be 10.88 years. With $F$ chosen as Weibull and $\alpha$ and $\beta$ fixed at the values reported in Table 4, we obtain the conditional MLE of the values of $\pi$ in the different segments. Whenever $x_l \geq 11$, the likelihood (17) is an increasing function of $b_l$ and is maximized at $b_l = 1$. Therefore, the maximization is needed with respect to $b_1, \ldots, b_{11}$ only. Figure 2(C) shows the plot of the exponential $\pi_{\hat{\eta}}$ and the conditional MLE of the piecewise constant $\pi$ in the range 0 to 14 years. The two plots are found to be close to each other. This supports the choice of the exponential form of $\pi_\eta$.

## 8. Concluding remarks

The thrust of this paper has been to offer a realistic model for menarcheal recall data amenable to informative censoring. As the MLE obtained from the usual interval censoring likelihood is not consistent, the MLE under the proposed model should be an attractive alternative.

The data set analysed in Section 7 also contains 'partial' recall data relating to the week/month/year of menarche. More sophisticated modelling would be required for handling data of such complex nature. The work presented in this paper can be used as a point of departure for such models. Another direction of future research could be inclusion of the possibility of error in recall data. The dichotomization of the recall information used in Section 7, where all 'partial' recall data have been ignored and regarded as cases of no recall, reduces the impact of recall error.

It would also be of interest to get rid of any model for the age at menarche and to look for a non-parametric estimator. This problem will be taken up in future.

## Acknowledgements

This research is partially sponsored by the project 'Physical growth, body composition and nutritional status of the Bengal school aged children, adolscents, and young adults of Calcutta, India: Effects of socioeconomic factors on secular trends', funded by the Neys Van Hoogstraten Foundation of the Netherlands, and the project 'Optimization and Reliability

Modeling' funded by the Indian Statistical Institute, Kolkata. The authors thank Professor Parasmani Dasgupta, leader of the first project, for making the data available for this research. They also thank Professor Biswabrata Pradhan, leader of the second project, and Professor Anup Dewanji for helpful discussions. They also gratefully acknowledge useful comment from the associate editor.

## References

Aggarwala, R. (2001). Progressive interval censoring: some mathematical results with application to inference. *Commun. Statist. Theory Meth.* **30**, 1921–1931.

Anderson, S. E. & Must, A. (2005). Interpreting the continued decline in the average age at menarche: results from two nationally representative surveys of U.S. girls studied 10 years apart. *J. Pediatrics* **147**, 753–760.

Ash, R. B. (2000). *Probability and measure theory*, Harcourt/Academic Press, Burlington, MA.

Betensky, R. A. (2000). On nonidentifiability and noninformative censoring for current status data. *Biometrika* **87**, 218–221.

Dunson, D. B. & Dinse, G. E. (2002). Baysian models for multivariate current status data with informative censoring. *Biometrics* **58**, 79–88.

Eveleth, P. B. (1986). Timing of menarche: secular trend and population differences. In *School age pregnancy and parenthood: biosocial dimensions* (eds J. B. Lancaster & B. A. Hamburg), Aldine-De Gruyter Pub., New York; 39–52.

Ferguson, T. S. (1996). *A course in large sample theory*, Chapman and Hall, London.

Finkelstien, D. M., Goggines, W. B. & Schoenfeld, D. A. (2002). Analysis of failure time data with dependent interval censoring. *Biometrics* **58**, 298–304.

Frisch, R. E. (1985). Fatness, menarche and female fertility. *Perspectives Biol. Med.* **28**, 611–633.

Frydman, H. (1994). A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *J. Roy. Statist. Soc. Ser. B* **56**, 71–74.

Gibbons, J. D. & Chakraborti, S. (2003). *Nonparametric statistical inference*, Marcel Dekker, Inc, New York.

ISI. (2012). *Annual Report of the Indian Statistical Institute 2011-12*, Indian Statistical Institute. Available at: http://library.isical.ac.in/jspui/handle/10263/5345?mode=full [accessed on 31 May 2014].

Kaciroti, N. A., Raghunathan, T. E. & Taylor, J. M. G. (2012). A Bayesian model for time-to-event data with informative censoring. *Biostatistics* **13**, 341–354.

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

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

Lee, E. T. & Wang, J. W. (2003). *Statistical methods for survival data analysis*, John Wiley, New York.

Lehman, E. L. (1999). *Elements of large-sample theory*, Springer-Verlag, New York.

McKay, H. A., Bailey, D. B., Mirwald, R. L., Davison, K. S. & 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.

Miller, R. G. (1981). *Survival analysis*, John Wiley, New York.

Morabia, A. & Costanza, M. C. (1998). International variability in ages at menarche, first livebirth, and menopause. *Amer. J. Epidemiol.* **148**, 1195–1205.

Nocedal, J. & Wright, S. J. (2006). *Numerical optimization*, Springer, New York.

Padez, C. (2003). Age at menarche of schoolgirls in Maputo, Mozambique. *Ann. Hum. Biol.* **30**, 487–495.

Parazzini, F., Chatenoud, L., Tozzi, L., Benzi, G., Pino, D. D. & Fedele, L. (1997). Determinants of risk of spontaneous abortions in the first trimester of pregnancy. *Epidemiology* **8**, 681–683.

Roberts, D. F. (1994). Secular trends in growth and maturation in British girls. *Amer. J. Hum. Biol.* **6**, 13–18.

Sandler, D. P., Wilcox, A. J. & Horney, L. F. (1984). Age at menarche and subsequent reproductive events. *Amer. J. Epidemiol.* **119**, 765–774.

Scharfstein, D. O. & Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika* **89**, 617–634.

Scharfstein, D. O., Robins, J. M., Eddings, W. & Rotnitzky, A. (2001). Inference in randomized studies with informative censoring and discrete time-to-event endpoints. *Biometrika* **57**, 404–413.

Stine, R. A. & Small, R. D. (1986). Estimating the distribution of censored logistic recall data. Technical Report, Department of Statistic, University of Pennsylvania. **83**.

Teilmann, G., Petersen, J. H., Gormsen, M., Damgaard, K., Skakkebaek, N. E. & Jensen, T. K. (2009). Early puberty in internationally adopted girls: hormonal and clinical markers of puberty in 276 girls examined biannually over two years. *Hormone Research Paediatrics* **72**, 236–246.

Towne, B., Czrewinski, S. A., Demerath, E. W., Blangero, J., Roche, A. F. & Siervoge, R. M. (2005). Heritability of age at menarche in girls from the Fels longitudinal study. *Amer. J. Phys. Anthropol.* **128**, 210–219.

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

Udry, J. R. & Cliquet, R. L. (1982). A cross-cultural examination of the relationship between ages at menarche, marriage and first birth. *Demography* **19**, 53–63.

*Received July 2013, in final form April 2014*

Sedigheh Mirzaei Salehabadi, Applied Statistical Unit, Indian Statistical Institute, Kolkata, 700108 India.
E-mail: sedigheh_r@isical.ac.in

## Appendix

*Proof of theorem 3.1*

*Proof.* The density in the first two cases can be obtained by considering the corresponding probability masses:

$$\begin{aligned}
f(s,0,0) &= P(Z=0,\delta=0|S=s)g(s),\\
&= P(T>s|S=s)g(s) = (\bar{F}_\theta(s))g(s);\\
f(s,0,1) &= E_T[f(s,0,1)|T],\\
&= E_T[P(S>T|S=s,T)g(s)\pi_\eta(s-T)],\\
&= \int_0^s g(s)\pi_\eta(s-u)\,f_\theta(u)du.
\end{aligned}$$

In the third case, the density can be derived as the derivative of a probability,

$$\begin{aligned}
f(s,z,1) &= g(s)\frac{\partial P(Z<z,\delta=1|S=s)}{\partial z},\\
&= g(s)\lim_{h\to 0}\frac{P(z<Z\le z+h,\delta=1|S=s)}{h},\\
&= g(s)\lim_{h\to 0}\frac{P(z<Z\le z+h|S=s)}{h},\\
&= g(s)\lim_{h\to 0}\frac{P(z<s-T\le z+h,T<s,\varepsilon=1)}{h},\\
&= g(s)\lim_{h\to 0}\frac{P(s-z-h<T\le s-z,\varepsilon=1)}{h},\\
&= g(s)\lim_{h\to 0}\frac{E_T[P(s-z-h<T\le s-z|T)(1-\pi_\eta(s-T))]}{h},\\
&= g(s)\lim_{h\to 0}\frac{\int_{s-z-h}^{s-z} f_\theta(u)(1-\pi_\eta(s-u))du}{h},\\
&= g(s)\,f_\theta(s-z)(1-\pi_\eta(z)).
\end{aligned}$$

□