Parametric estimation of menarcheal age distribution based on recall data

Technical Report No.ASD/2013/3
Dated 7 June, 2013

Sedigheh Mirzaei Salehabadi
and
Debasis Sengupta

Indian Statistical Institute

Applied Statistics Unit
Kolkata 700 108

Parametric estimation of menarcheal age distribution based on recall data

Sedigheh Mirzaei Salehabadi and Debasis Sengupta
Applied Statistical Unit
Indian Statistical Institute Kolkata
Email:sedigheh_r@isical.ac.in, sdebasis@isical.ac.in.

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 MLE 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 based on 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 MLE based on recall data over that based on status quo data.

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 (see Frisch (1985); Eveleth (1986); Anderson and Must (2005)). It is also widely used as a demographic indicator of population fecundity (see Udry and Cliquet (1982)). Menarcheal age distribution has been used to assess reproductive risks (see Sandler, Wilcox and Horney (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)). 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, (see 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 (see Lawless (1982)). The non-parametric maximum likelihood estimator (MLE) is the well-known product limit estimator proposed by Kaplan and 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., six 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 and 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 nonparametric and parametric methods have been developed over the years for the analysis of interval censored data (Turnbull (1976); Miller (1981); Frydman (1994); Aggarwala (2001) and Lee and Wang (2003)). Interval censoring is typically assumed to be noninformative, 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. One may seek to use Baysian methods for informative censoring (see Scharfstein et al. (2001); Scharfstein and Robins (2002); Kaciroti, Raghunathan and Taylor (2012)) for such data. Alternatively, one may seek an estimator, based on 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 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_θ, where θ is a vector of parameters. The i^{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 $\delta_i = I_{(T_i \leqslant S_i)}$, the indicator of the event $(T_i \leqslant S_i)$. The likelihood is

$$\prod_{i=1}^{n} [F_\theta(S_i)]^{\delta_i} [\bar{F}_\theta(S_i)]^{1-\delta_i}, \tag{2.1}$$

where $\bar{F}_\theta(S_i) = 1 - F_\theta(S_i)$. Most researchers use MLE of θ based on the above likelihood, (see Lee and 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 ε_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 noninformative, then the likelihood is

$$\prod_{i=1}^{n}\left[(F_\theta(S_i))^{1-\varepsilon_i}(f_\theta(T_i))^{\varepsilon_i}\right]^{\delta_i}[\bar{F}_\theta(S_i)]^{1-\delta_i}, \tag{2.2}$$

where f_θ is the probability density function corresponding to the distribution F_θ. Aggarwala (2001) proposed the use of the MLE of θ based on an extension of the above likelihood.

It has been pointed out in the previous section that noninformativeness of censoring is difficult to justify in the present context. In particular, the non-recall probability, $P(\varepsilon_i = 0|\delta_i = 1)$ may depend on the time elapsed since menarche, $S_i - T_i$. We model this non-recall probability by $\pi_\eta(S_i - T_i)$, where π_η is a family of increasing functions indexed by the parameter η. According to this model, the likelihood is

$$\prod_{i=1}^{n}\left[\left(\int_0^{S_i} f_\theta(u)\pi_\eta(S_i-u)du\right)^{1-\varepsilon_i}\left[f_\theta(T_i)(1-\pi_\eta(S_i-T_i))\right]^{\varepsilon_i}\right]^{\delta_i}[\bar{F}_\theta(S_i)]^{1-\delta_i}. \tag{2.3}$$

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

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

When π_η is a constant, (2.3) becomes a constant multiple of (2.2). As a further special case, if $\pi_\eta = 1$, then (2.3) reduces to (2.1). When $\pi_\eta = 0$, i.e., all recalls are perfect, the product likelihood (2.3) reduces to

$$\prod_{i=1}^{n}[f_\theta(T_i)]^{\delta_i}[\bar{F}_\theta(S_i)]^{1-\delta_i}, \tag{2.4}$$

which is the same as the likelihood for prospective data obtained from continuous monitoring. Thus the model leading to the likelihood (2.3) is more general than

the standard censoring models.

3. Large Sample Property

The factors in the product likelihood (2.3) 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_θ 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

$$Z_i = (S_i - T_i)\varepsilon_i\delta_i, \tag{3.1}$$

where ε_i and δ_i are as defined in the previous section. Note that the vector

$$Y_i = (S_i, Z_i, \delta_i) \tag{3.2}$$

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

We now show that the i^{th} factor in the product likelihood (2.3) is in fact proportional to the density of Y_i. We prove this result below, 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 both ϑ_1 and ϑ_3 are the counting measure and ϑ_2 is the sum of the counting and the Lebesgue measures (see Ash (2000)). We presume that G is a discrete distribution, with probability mass function g.

Theorem 3.1. The density of $Y = (S, Z, \delta)$ with respect to the measure μ is

$$f(s,z,\delta) = \begin{cases} g(s)\bar{F}_\theta(s) & \text{if } z = 0 \text{ and } \delta = 0, \\ g(s)\int_0^s f_\theta(u)\pi_\eta(s-u)du & \text{if } z = 0 \text{ and } \delta = 1, \\ g(s)f_\theta(s-z)(1-\pi_\eta(z)) & \text{if } z > 0 \text{ and } \delta = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

Proof. See the Appendix. □

The likelihood (2.3) can be written in terms of S_i, Z_i and δ_i as

$$\prod_{i=1}^{n}\left[\left(\int_0^{S_i} f_\theta(u)\pi_\eta(S_i-u)du\right)^{I_{(Z_i=0)}}\left[f_\theta(S_i-Z_i)(1-\pi_\eta(Z_i))\right]^{I_{(Z_i>0)}}\right]^{\delta_i}[\bar{F}_\theta(S_i)]^{1-\delta_i}$$
$$=\frac{\prod_{i=1}^{n} f(S_i,Z_i,\delta_i)}{\prod_{i=1}^{n} g(S_i)}. \tag{3.4}$$

The numerator is a product of densities of the form (3.3), while the denominator does not contain any information about θ. This likelihood can also be interpreted as a product of conditional densities of (Z_i,δ_i) given S_i, for $i=1,2,\ldots,n$. Further, this conditional likelihood is free from g, i.e., inference for θ can proceed by ignoring any parameter of g.

Once the likelihood (2.3) 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 triplet (S_i,T_i,ε_i), which completely determine the observable vector (S_i,Z_i,δ_i). Since the likelihood involves only the conditional density of (Z_i,δ_i) given S_i, it suffices to look for conditions on the distribution of (T_i,ε_i) only. Specifically, the conditions would involve the density f_θ, the density of T_i, and the function π_η, which defines the conditional density of the binary random variable ε_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 θ is identifiable with respect to the family of densities f_θ of the menarcheal age, and the parameter η is identifiable with respect to the family of functions π_η representing non-recall probability. In other words, $\theta_1\neq\theta_2$ implies $f_{\theta_1}\neq f_{\theta_2}$, and $\eta_1\neq\eta_2$ implies $\pi_{\eta_1}\neq\pi_{\eta_2}$.

(C2) The parameter spaces for θ and η are compact.

(C3) The random variables T_i, $i=1,2,\ldots,n$ are samples from the density f_θ, and ε_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

θ and η respectively.

(C5) The function $f_\theta(t)$ is differentiable with respect to θ for all t such that the derivative is absolutely bounded by a μ-integrable function $h_1(t)$, and the function $\pi_\eta(z)$ is differentiable with respect to η for all z such that the derivative is absolutely bounded by a μ-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,δ_i) given S_i with respect to θ and η, 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 fulfillment of Condition C5 of Theorem 7.1.1 of Lehman (1999).

The additional conditions for asymptotic normality relate to the log-likelihood obtained from (2.3),

$$\ell(\theta,\eta)=\sum_{i=1}^{n}\Bigg[\delta_i(1-\varepsilon_i)\log\left(\int_0^{S_i} f_\theta(u)\pi(S_i-u)du\right) + \delta_i\varepsilon_i\log\left(f_\theta(T_i)(1-\pi(S_i-T_i))\right)+(1-\delta_i)\log\left(\bar{F}_\theta(S_i)\right)\Bigg]. \tag{3.5}$$

The following conditions, together with C1-C5, ensure asymptotic normality of the MLE of θ and η (see Theorem 7.3.1. of Lehman (1999)).

(C6) First and second derivatives of $\ell(\theta,\eta)$ are defined.

(C7) The Fisher information matrix

$$I(\theta,\eta)=\begin{bmatrix} E\left[\frac{\partial}{\partial\theta}\ell(\theta,\eta)\right]^2 & E\left[(\frac{\partial}{\partial\theta}\ell(\theta,\eta))(\frac{\partial}{\partial\eta}\ell(\theta,\eta))\right] \\ E\left[(\frac{\partial}{\partial\theta}\ell(\theta,\eta))(\frac{\partial}{\partial\eta}\ell(\theta,\eta))\right] & E\left[\frac{\partial}{\partial\theta}\ell(\theta,\eta)\right]^2 \end{bmatrix}$$

is non-zero and continuous with respect to the parameters θ and η.

4. Theoretical Comparison of Estimates

4.1. Bias of MLE based on Interval Likelihood

If one ignores the informative nature of censoring, then the likelihood (2.2) 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 (2.2) 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 (2.2), 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 (2.2) with respect to θ is

$$\sum_{i=1}^{n}\left[\delta_i(1-\varepsilon_i)\left(\frac{\frac{s_i}{\theta^2}e^{\frac{-s_i}{\theta}}}{1-e^{\frac{-s_i}{\theta}}}\right)+\delta_i\varepsilon_i\left(\frac{-1}{\theta}+\frac{t_i}{\theta^2}\right)+(1-\delta_i)\frac{s_i}{\theta^2}\right]. \tag{4.1}$$

The expectation of (4.1) with respect to the general model of Section 2 is

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

In the further special case $\eta = \theta$, the above 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 (2.2) 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,δ). 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

$$I_R(\theta,\eta) = I_S(\theta,\eta) + I_A(\theta,\eta), \tag{4.2}$$

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.

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

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

where

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

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

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

In particular, the additional information of θ, the parameter of interest, is

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

When η 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 η $(I_1 + I_2)$, for different values of η and a range of values of θ. It can be seen that, when η 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 η is minimal compared to the gain from recall data. On the other hand, for a small value of η (i.e., menarcheal age forgotten quickly), recall data does not augment the information noticeably.

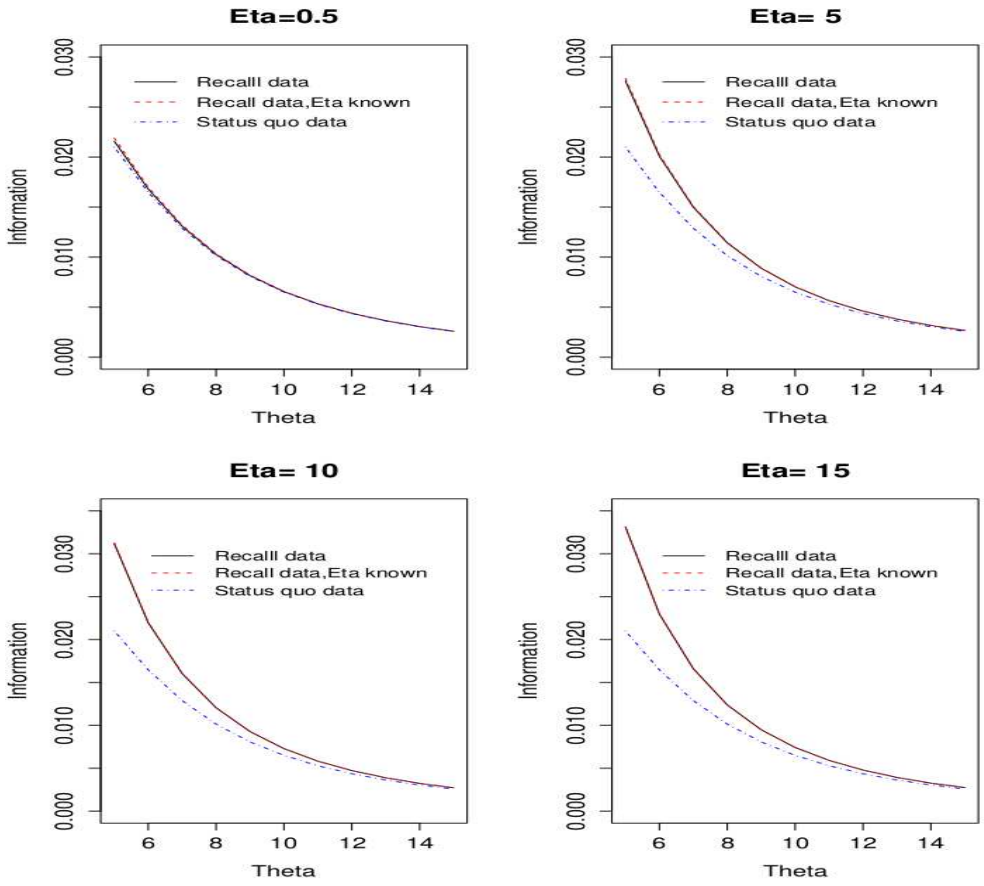

Figure 1: Information based on recall data and status quo likelihoods.

5. Simulation Results

For the purpose of simulation, we assume that 'age at menarche' follows the Weibull distribution with shape and scale parameters α and β, respectively. Thus, $\theta = (\alpha, \beta)$. Further, we assume that 'age at interview' follows the discrete uniform distribution over [7,21] and that π_η has the exponential distribution function with mean η. 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 MLE's based on the status quo likelihood (2.1) the interval censoring likelihood (2.2) and the recall data likelihood (2.3) for our model. Computation of MLE's in all the cases is done through numerical optimization of likelihood (Nocedal and Wright (2006)).

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

Table 1 and 2 show the bias, the standard deviation (Stdev), the mean squared error (MSE) and the Cramer-Rao Lower Bound (CRLB) for the MLE's of the three parameters based on the three likelihoods, for the combination of parameter values (i) and (ii) respectively.

		n=50			n=500			n=1000		
Estimator	Property	α	β	η	α	β	η	α	β	η
MLE	Bias	7.760	–0.076	–	0.720	0.004	–	0.488	0.008	–
from	Stdev	6.678	0.447	–	1.469	0.141	–	0.894	0.105	–
Status	MSE	104.8	0.270	–	2.690	0.020	–	1.036	0.011	–
quo	CRLB	43.410	0.145	–	2.140	0.017	–	0.872	0.011	–
MLE	Bias	3.280	0.180	–	1.460	0.230	–	1.373	0.230	–
from	Stdev	2.998	0.332	–	0.787	0.100	–	0.592	0.071	–
Interval	MSE	19.75	0.140	–	2.760	0.065	–	2.235	0.059	–
censoring	CRLB	5.991	0.016	–	0.608	0.095	–	0.327	0.003	–
MLE	Bias	2.080	0.120	–0.038	0.361	0.020	0.007	0.202	0.010	-0.005
from	Stdev	2.629	0.300	0.825	0.721	0.089	0.247	0.545	0.055	0.167
Our	MSE	11.236	0.010	0.700	0.660	0.009	0.063	0.338	0.003	0.028
Method	CRLB	4.541	0.018	0.592	0.514	0.008	0.061	0.327	0.002	0.028

Table 1: Bias, Stdev, MSE and CRLB of estimated parameters in case (i)

		n=50			n=500			n=1000		
Method	Property	α	β	η	α	β	η	α	β	η
MLE	Bias	9.60	–0.073	–	1.070	0.044	–	0.967	0.039	–
from	Stdev	3.673	0.510	–	1.245	0.155	–	0.841	0.098	–
Status	MSE	107.2	0.262	–	2.730	0.026	–	1.642	0.011	–
quo	CRLB	7.821	0.260	–	1.210	0.023	–	0.700	0.010	–
MLE	Bias	3.040	0.260	–	1.450	0.240	–	1.430	0.230	–
from	Stdev	3.106	0.306	–	0.671	0.095	–	0.458	0.071	–
Interval	MSE	18.81	0.130	–	2.560	0.068	–	2.255	0.058	–
censoring	CRLB	1.282	0.052	–	0.400	0.006	–	0.190	0.004	–
MLE	Bias	1.930	0.230	0.075	0.581	0.060	-0.004	0.420	0.044	0.004
from	Stdev	2.584	0.300	1.068	0.860	0.095	0.332	0.418	0.062	0.281
Our	MSE	11.236	0.010	0.700	0.660	0.009	0.063	0.351	0.006	0.079
Method	CRLB	2.383	0.026	1.076	0.612	0.008	0.11	0.200	0.005	0.079

Table 2: Bias, Stdev, MSE and CRLB of estimated parameters in case (ii)

It is found that the bias for the MLE based on interval censoring likelihood stabilizes around a positive constant when 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.

6. 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 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 3 gives a summary of the findings. Figure 2 shows the plot of the survival functions corresponding to the three sets of estimates.

Estimator	Estimate (standard error)			Median	95% Confidence Interval
	α	β	η		of Median
MLE from	10.74	12.17		11.76	(11.62,11.90)
Status quo	(0.320)	(0.005)			
MLE from	11.80	12.65		12.25	(12.20,12.30)
Interval censoring	(0.061)	(0.001)			
MLE from	10.19	12.21	3.47	11.78	(11.72,11.84)
Our Method	(0.090)	(0.001)	(0.140)		

Table 3: Estimated parameters and median age at menarch from different methods

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

Figure 3 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.

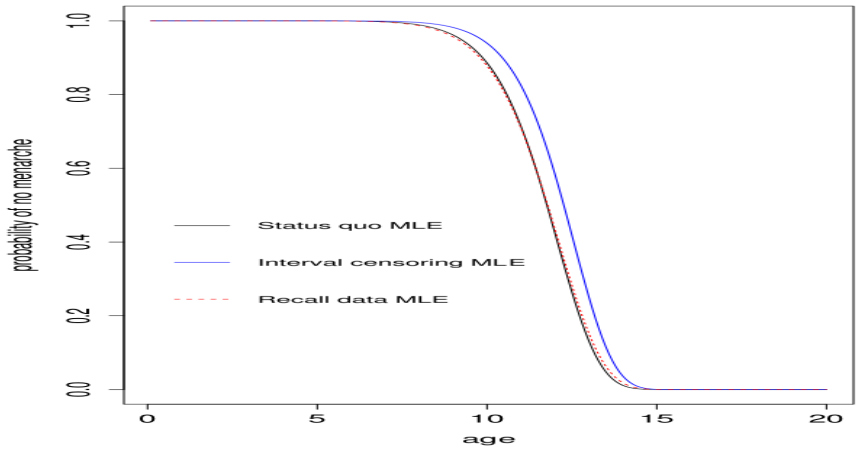

Figure 2: Survival plots for real data based on three methods.

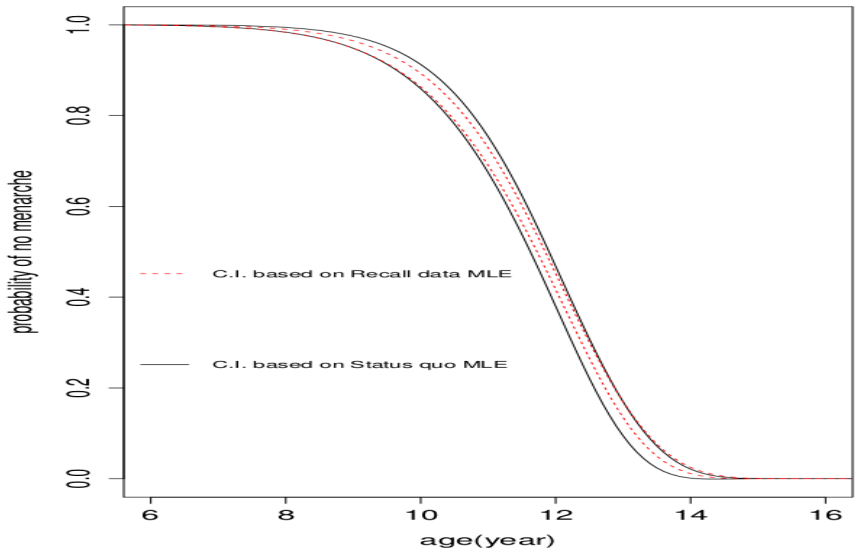

Figure 3: Confidence Interval for Probability of no Menarche based on two methods.

7. 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.

Modeling of the non-remembering 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. In the two foregoing sections, we have opted for an exponential model with a single parameter.

The data set analysed in Section 6 also contains ‘partial’ recall data relating to the week/month/year of menarche. More sophisticated modeling 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 6, 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.

Appendix: Proof of Theorem 3.1.

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

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

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

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

□

References

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

ANDERSON, S. E. and 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.

EVELETH, P. B. (1986). *Timing of menarche: secular trend and population differences*. School age Pregnancy and Parenthood: Biosocial Dimensions.

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.

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.

KACIROTI, N. A., RAGHUNATHAN, T. E. and TAYLOR, J. M. G. (2012). A Bayesian model for time-to-event data with informative censoring. *Biostatistics* **13,** 341–354.

KAPLAN, E. L. and 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. and WANG, J. W. (2003). *Statistical Methods for Survival Data Analysis*. John Wiley.

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. 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.

MILLER, R. G. (1981). *Survival Analysis*. John Wiley.

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

NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical optimization*, second ed. *Springer Series in Operations Research and Financial Engineering*. 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. and 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. and HORNEY, L. F. (1984). Age at menarche and subsequent reproductive events. *Amer. J. Epidemiol.* **119,** 765–774.

SCHARFSTEIN, D. O. and 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. and ROTNITZKY, A. (2001). Inference in randomized studies with informative censoring and discrte time-to-event endpoints. *Biometrics* **57,** 404–413.

STINE, R. A. and SMALL, R. D. (1986). *Estimating the Distribution of Censored Logistic Recall Data* **83,**. Technical Report, Department of Statistic, University of Pennsylvania.

TEILMANN, G., PETERSEN, J. H., GORMSEN, M., DAMGAARD, K., SKAKKEBAEK, N. E. and 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. and SIERVOGEL, 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. and CLIQUET, R. L. (1982). A cross-cultural examination of the relationship between ages at menarche, marriage and first birth. *Demography* **19,** 53–63.