Estimation of menarcheal age distribution from imperfectly recalled data

Technical Report No. ASU/2016/4
Dated 20 September, 2016

Sedigheh Mirzaei Salehabadi,
Debasis Sengupta
and
Rahul Ghosal

Indian Statistical Institute

Applied Statistics Unit
Kolkata 700 108

Estimation of menarcheal age distribution from imperfectly recalled data

Sedigheh Mirzaei Salehabadi [1], Debasis Sengupta[2] and Rahul Ghosal [2]

[1] Duke-NUS medical school, National University of Singapore

[2] Indian Statistical Institute, Kolkata, India.

Email: sedigheh.mirzaei@duke-nus.edu.sg[1], sdebasis@isical.ac.in[2],rahulghosal3@gmail.com[2]

Abstract

In a cross-sectional study on menarcheal age, subjects who had experienced menarche were asked to recall the time of menarche. Some respondents were able to recall the date exactly, some recalled only the month or the year in which the event had happened, and some were not able to recall the date at all. The objective was to estimate the menarcheal age distribution from this data, which are interval censored. The censoring is informative, as there is evidence of memory fading with time. Moreover, the censoring involves the time-to-event since birth and the calendar time in a complicated way. We propose a model for this type of data in order to make full use of the information contained in it. In this model, the probabilities of various types of recall are assumed to depend on the time since menarche, through a multinomial regression function. The ensuing likelihood contains different types of factor for the cases with different degrees of recall and those where the menarche did not happen. The structure of the data also varies from case to case. For this reason, the usual asymptotic theory is not readily applicable to the maximizer of this parametric likelihood. We express the observables as one-to-one functions of a vector of continuous, discrete and mixed variables. This transformed vector is observed and has the same size in all the cases. Moreover, the likelihood turns out to be a product of probability densities of this vector in a suitable probability space. We provide a set of regularity conditions on the time-to-event distribution, subject to which the consistency and the asymptotic normality of the maximum likelihood estimator are established. We study the small sample performance of the estimator through Monte Carlo simulations. We also provide a graphical check for the assumption of the multinomial regression model for the recall probabilities. The assumption appears to hold

for the menarcheal data set. Its analysis shows that the use of the imperfectly recalled part of the data in the proposed manner indeed leads to smaller confidence intervals of the survival function.

KEYWORDS: Interval censoring, Informative censoring, Maximum likelihood estimator, Retrospective study, Current status data, Weibull distribution.

1. INTRODUCTION

In a recent survey conducted by the Indian Statistical Institute (ISI) in and around the city of Kolkata (Dasgupta 2015), over four thousand randomly selected individuals, aged between 7 and 21 years, were sampled. In this retrospective and cross-sectional study, the subjects were interviewed on or around their birthdays. The data set on female subjects contains age, menarcheal status, some physical measurements and information on some socioeconomic variables. Those who had already experienced menarche, were asked to recall the time of the onset of their menarche. Among the 2195 females represented in the data set, 775 individuals did not have menarche, 443 individuals recalled the exact date of the onset of menarche, 276 and 209 individuals recalled the calendar month and the calendar year of the onset, respectively, and 492 individuals could not recall any range of dates. Thus, the data are interval-censored. A major goal of this study is to estimate the distribution of the age at onset of menarche.

Many other instances of incompletely recalled time-to-event data exist in the literature. The variables of interest in these studies include age at onset of menarche in adolescent and young adult females (Koo and Rohana 1997), time-to-pregnancy (Joffe, Villard, Li, Plowman and Vessey 1995), time-to-weaning from breastfeeding (Gillespie, dÁrcy, Schwartz, Bobo and Foxma 2006), time-to-injury for victims injured during a year (Harel, Overpeck, Jones, Scheidt, Bijur, Trumble and Anderson 1994) and time-to-employment (Mathiowetza and Ouncanb 1988). In all these studies, the estimation of the probability distribution of time-to-event is an important problem for building a standard for individuals, comparing two populations or assessing the effect of a covariate on this distribution. There is a possibility that the recalled time-to-event may be inaccurate (Koo and Rohana 1997; Mathiowetza and Ouncanb 1988). One of the ways of avoiding this problem is to allow the respondent to provide a range of dates when he or she is unable to recall the exact date. In the ISI study this option was given to the respondents, and the recalled ranges of dates

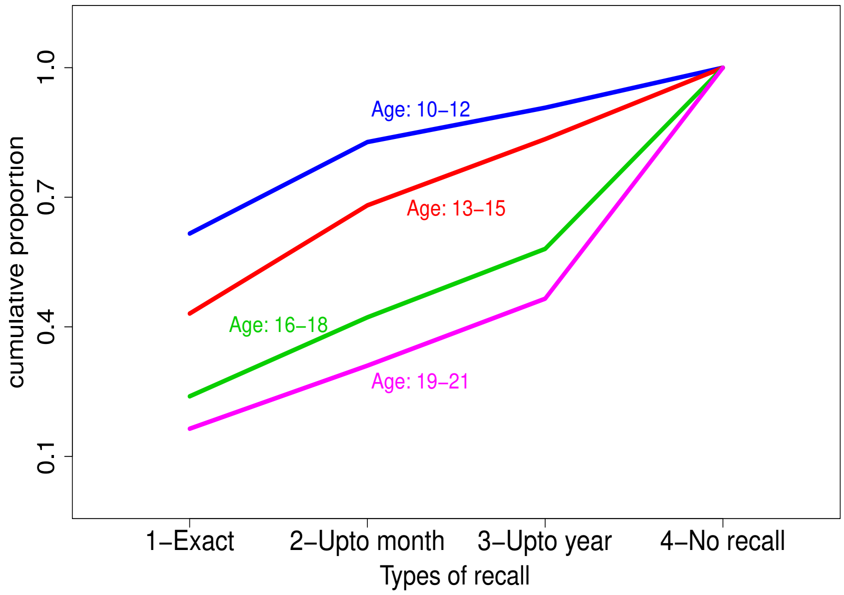

Figure 1: Cumulative proportion of decreasing degrees of recall for different age ranges in menarcheal data.

generally happened to be in terms of calendar months and years. We refer to this special type of incompleteness as partial recall.

Figure 1 shows the cumulative proportion of successively less precise recall in different groups of ages at interview, for the respondents of the ISI study. It is seen that the lines do not cross and the age order is preserved. Also for lower age, there is greater precision of recall. This finding indicates that memory fades with time, i.e., two subjects interviewed at the same age would have different chances of recalling their age at menarche, depending on which of them had experienced the event earlier. It transpires that the censoring mechanism underlying such recall-based data is inherently informative. The natural question is: how can one model the different degrees of partial recall, so that the distribution of menarcheal age can be estimated?

Even though there is abundance of partially recalled data from which the distribution of the time-to-event needs to be estimated, there is no suitable model and method for it. The problem is complicated by two facts. First, the censoring is likely to be informative, as pointed out above. Secondly, the data involves two scales of time, viz. the respondent-specific starting time (e.g., birth) for measuring the time-to-event, and the calendar time through which the partial recall information is expressed. Mirzaei, Sengupta and Das (2015) addressed the first issue by proposing a model for the specific type of informative censoring found in the data set considered here. However, they bypassed the second issue, as they clubbed all the cases of partial recall with the cases of no recall.

In this paper, we propose in Section 2 a new approach for estimating the distribution of the time-to-event, which uses the recall information through a realistic censoring model that makes use of calendar time. Under this model, the time of observation is assumed to be independent of the time-to-event, and a multinomial regression set up is used to represent the chances of no recall, exact recall and recalls up to the calendar month or year. We derive the appropriate likelihood under the proposed model, the corresponding maximum likelihood estimator (MLE) and its asymptotic properties. In Sections 3 and 4, we report the results of Monte Carlo simulations of small sample performance of the MLE and present some diagnostic checks of adequacy of the model. We return to the main data set and analyze it in Section 5. The Proof of a theorem is given in the Appendix.

2. ESTIMATION

2.1 Model and Likelihood

Consider a set of n subjects having ages at occurrence of landmark events $T_1, \ldots, T_n$, which are samples from the distribution F_θ, with density f_θ, where θ is a vector of parameters. Let these subjects be interviewed at ages $S_1, \ldots, S_n$, respectively. Suppose the S_i's are samples from another distribution and are independent of the T_i's. Let δ_i be the indicator of $T_i \le S_i$. This inequality means that the event for the ith subject had occurred 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 times of interview, is

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

where $\bar{F}_\theta(S_i) = 1 - F_\theta(S_i)$. For properties of the MLE based on the above likelihood, see Lee and Wang (2003).

The structure of recalled data is generally more complicated. Mirzaei et al. (2015) proposed a simplistic model where the subject may either recall the time of the event exactly or not remember it at all. They used another indicator, ε_i, to record whether an exact recall is possible. As the chance of recall may depend on the time elapsed since the event, they modeled the non-recall probability as a function of this time. According to this model,

$$P(\varepsilon_i = 0|S_i = s, T_i = t) = \pi_\eta(s-t) \qquad \text{for} \quad 0 < t < s,$$

where π_η is a family of functions indexed by the parameter η. Thus, the likelihood is

$$\prod_{i=1}^{n} \left[\left(\int_0^{S_i} f_\theta(u)\pi_\eta(S_i-u)du \right)^{1-\varepsilon_i} [f_\theta(T_i)(1-\pi_\eta(S_i-T_i))]^{\varepsilon_i} \right]^{\delta_i} [\bar{F}_\theta(S_i)]^{1-\delta_i}. \tag{2}$$

When π_η is a constant, this likelihood becomes a constant multiple of the likelihood corresponding to non-informatively interval censored data.

Let us now consider the possibility that the ith subject can recall the date of the event only up to a calendar month or a calendar year. Let ε_i be a variable that indicates how the ith subject recalls the time of his/her landmark event.

$$\varepsilon_i = \begin{cases} 0 & \text{if the exact date is recalled,} \\ 1 & \text{if the date is recalled up to the calendar month,} \\ 2 & \text{if the date is recalled up to the calendar year,} \\ 3 & \text{if the event has not happened or the date is not recalled.} \end{cases} \tag{3}$$

Note that in the previous case, the variable ε_i could assume only two values, 0 and 1. We regard the multiple possibilities as outcomes of a multinomial selection. The allocation probabilities are modeled as functions of the time elapsed since the occurrence of the event. Thus, for $0 < t < s$, we model the allocation probabilities as

$$\begin{aligned} P(\varepsilon_i = 0|S_i = s, T_i = t) &= \pi_\eta^{(0)}(s-t), \\ P(\varepsilon_i = 1|S_i = s, T_i = t) &= \pi_\eta^{(1)}(s-t), \\ P(\varepsilon_i = 2|S_i = s, T_i = t) &= \pi_\eta^{(2)}(s-t), \\ P(\varepsilon_i = 3|S_i = s, T_i = t) &= \pi_\eta^{(3)}(s-t). \end{aligned} \tag{4}$$

where $\sum_{k=0}^{3} \pi_\eta^{(k)}(s-t) = 1$, and η is a vector of parameters.

We refer to the set-up described in the first paragraph of this section, together with (3) and (4) as the proposed model. According to this model, there would be five cases for an individual i, with different contributions to the likelihood.

Case (i) When $\delta_i = 0$ (the event has not occurred till the time of observation), the contribution of the ith individual to the likelihood is $\bar{F}_\theta(S_i)$.

Case (ii): When $\delta_i = 1$ and $\varepsilon_i = 0$ (the event has occurred and the ith individual can remember the time), the contribution of the individual to the likelihood is $f_\theta(T_i)\pi_\eta^{(0)}(S_i - T_i)$.

Case (iii): When $\delta_i = 1$ and $\varepsilon_i = 1$ (the event has occurred but the ith individual can only recall the calendar month of the event), the contribution of the individual to the likelihood is $\int_{M_{i1}}^{M_{i2}} f_\theta(u)\pi_\eta^{(1)}(S_i - T_i)du$, where M_{i1} and M_{i2} are the ages of the individual at the beginning and the end of the calendar month recalled by the individual. These limits can be computed if the date of birth is known.

Case (iv): When $\delta_i = 1$ and $\varepsilon_i = 2$ (the event has occurred but the ith individual can only recall the calendar year of the event), the contribution of the individual to the likelihood is $\int_{Y_{i1}}^{Y_{i2}} f_\theta(u)\pi_\eta^{(2)}(S_i - T_i)du$, where Y_{i1} and Y_{i2} are the ages of the individual at the beginning and the end of the calendar year recalled by the individual. The date of birth is needed for computing these ages also.

Case (v): When $\delta_i = 1$ and $\varepsilon_i = 3$ (the event has occurred but the ith individual cannot recall the time at all), the contribution of the individual to the likelihood is $\int_0^{S_i} f_\theta(u)\pi_\eta^{(3)}(S_i - T_i)du$.

Therefore, the overall likelihood is

$$\prod_{i=1}^{n}[\bar{F}_\theta(S_i)]^{1-\delta_i}\Bigg[\left(f_\theta(T_i)\pi_\eta^{(0)}(S_i - T_i)\right)^{I_{(\varepsilon_i=0)}} \times \left(\int_{M_{i1}}^{M_{i2}} f_\theta(u)\pi_\eta^{(1)}(S_i - T_i)du\right)^{I_{(\varepsilon_i=1)}}\left(\int_{Y_{i1}}^{Y_{i2}} f_\theta(u)\pi_\eta^{(2)}(S_i - T_i)du\right)^{I_{(\varepsilon_i=2)}}\left(\int_{0}^{S_i} f_\theta(u)\pi_\eta^{(3)}(S_i - T_i)du\right)^{I_{(\varepsilon_i=3)}}\Bigg]^{\delta_i}. \tag{5}$$

Note that when the probabilities of an individual recalling only the month or year of the event is zero, i.e., $\pi_\eta^{(1)} = \pi_\eta^{(2)} = 0$, the likelihood (5) reduces to (2).

The maximum likelihood estimator (MLE) of θ and η are obtained by maximizing the above likelihood by setting the partial derivatives equal to zero. A Newton-Raphson iteration may be used to compute the MLEs.

2.2 Large sample properties

The factors in the product likelihood (5) have different forms in different cases. For example, T_i is used only when $\delta_i = 1$ and $\varepsilon_i = 0$, while M_{i1} and M_{i2} are used only when $\delta_i = 1$ and $\varepsilon_i = 1$. For the standard asymptotic results to be applicable, each factor of this likelihood has to be expressed as the density of some random vector with respect to a suitable dominating measure.

The main challenge to obtaining a common format of the data lies in the fact that M_{i1}, M_{i2}, Y_{i1} and Y_{i2} are the age of the ith individual at specified calendar times. In order to overcome this difficulty, we make use of the fact that those observables are functions of T_i and the date of birth of the ith individual. Specifically, for the ith subject, let m_i be the serial number of the month of birth within the year of birth and d_i be the time (measured in years) from the beginning of the month of birth till the event of birth. These variables are observed for every individual, together with the other observables. For the sake of simplicity, we assume that every month has duration 1/12 and every year has duration 1.

When $\epsilon_i = 1$, i.e., the month of the event is recalled, we can write

$$\begin{aligned} M_{i1} &= \lfloor 12(d_i + T_i) \rfloor / 12 - d_i, \\ M_{i2} &= M_{i1} + 1/12, \end{aligned} \tag{6}$$

where $\lfloor \cdot \rfloor$ is the integer part of its argument. In other words, knowledge of $\lfloor 12(d_i + T_i) \rfloor$, m_i and d_i is equivalent to the knowledge of M_{i1}, M_{i2}, m_i and d_i. Likewise, when $\epsilon_i = 2$, i.e., the year of the event is recalled, we can write

$$\begin{aligned} Y_{i1} &= \lfloor \big(T_i + d_i + (m_i - 1)/12\big) \rfloor - \big(d_i + (m_i - 1)/12\big), \\ Y_{i2} &= Y_{i1} + 1. \end{aligned} \tag{7}$$

In other words, knowledge of $\lfloor \big(T_i + d_i + (m_i - 1)/12\big) \rfloor$, m_i and d_i is equivalent to the knowledge of Y_{i1}, Y_{i2}, m_i and d_i.

We now define

$$W_i = \begin{cases} T_i & \text{if } \varepsilon_i = 0 \text{ , } \delta_i = 1, \\ \lfloor 12(d_i + T_i) \rfloor / 12 & \text{if } \varepsilon_i = 1 \text{ , } \delta_i = 1, \\ \lfloor \big(T_i + d_i + (m_i - 1)/12\big) \rfloor & \text{if } \varepsilon_i = 2 \text{ , } \delta_i = 1, \\ 0 & \text{if } \varepsilon_i = 3, \ \delta_i = 1, \text{ or if } \delta_i = 0. \end{cases} \tag{8}$$

The variable W_i captures the essential part of the various forms of data, viz. T_i, M_{i1}, M_{i2}, Y_{i1} and Y_{i2}, that are observable in some cases but not in others. Therefore, we define the observable vector

$$Y_i = (S_i, W_i, \varepsilon_i, \delta_i, m_i, d_i), \tag{9}$$

which contains all the information available in various forms in different cases. In fact, all the observed variables can be retrieved from this vector.

We have already assumed that the T_i's (time-to-event) are samples from the distribution F_θ and the S_i's (ages on interview date) are samples from another distribution. We now denote by G_1, G_2 and G_3 the distributions of S_i, m_i and d_i, respectively, for every i. The distribution G_2 is defined over the set $\{1, 2, \ldots, 12\}$, and G_3 is defined over the interval $[0, 1/12]$. The latter assumption disregards the fact that d_i is known only up to days (measured as fixed fractions of a year). We make this assumption in order to keep the description simple.

It turns out that the i^{th} factor in the product likelihood (5) is in fact proportional to the probability density of Y_i. This follows from Theorem 1 presented below, after the subscript i is dropped for simplicity. The dominating probability measure used for defining this density is $\mu = \vartheta_1 \times \vartheta_2 \times \vartheta_3 \times \vartheta_4 \times \vartheta_5 \times \vartheta_6$ where ϑ_1 is the measure with respect to which G_1 has a density (e.g., the counting or the Lebesgue measure, depending on whether G_1 is discrete or continuous), ϑ_2 is the sum of the counting and the Lebesgue measures, each of ϑ_3, ϑ_4 and ϑ_5 is the counting measure and ϑ_6 is the Lebesgue measure (Ash 2000).

Theorem 1 *The density of $Y = (S, W, \varepsilon, \delta, m, d)$ with respect to the measure μ is*

$$h(s, w, \varepsilon, \delta, m, d) = \begin{cases} g_1(s)g_2(m)g_3(d)\bar{F}_\theta(s) & \text{if } \delta = 0, \\ g_1(s)g_2(m)g_3(d)f_\theta(w)\pi_\eta^{(0)}(s-w)I_{(w<s)} & \text{if } \varepsilon = 0 \text{ and } \delta = 1, \\ g_1(s)g_2(m)g_3(d)\int_{w-d}^{w+\frac{1}{12}-d} f_\theta(u)\pi_\eta^{(1)}(s-u)du & \text{if } \varepsilon = 1 \text{ and } \delta = 1, \\ g_1(s)g_2(m)g_3(d)\int_{w-d-\frac{m-1}{12}}^{w+1-d-\frac{m-1}{12}} f_\theta(u)\pi_\eta^{(2)}(s-u)du & \text{if } \varepsilon = 2 \text{ and } \delta = 1, \\ g_1(s)g_2(m)g_3(d)\int_0^s f_\theta(u)\pi_\eta^{(3)}(s-u)du & \text{if } \varepsilon = 3 \text{ and } \delta = 1, \end{cases} \tag{10}$$

where g_1, g_2 and g_3 are the densities of G_1, G_2 and G_3 with respect to the measures ϑ_1, ϑ_5 and ϑ_6, respectively.

The likelihood (5) can be written in terms of S_i, W_i, ε_i, δ_i, m_i and d_i as

$$\prod_{i=1}^{n}[\bar{F}_\theta(S_i)]^{1-\delta_i}\left[\left(f_\theta(W_i)\pi_\eta^{(0)}(S_i-W_i)\right)^{I_{(\varepsilon_i=0)}}\left(\int_{W_i-d_i}^{W_i-d_i+\frac{1}{12}} f_\theta(u)\pi_\eta^{(1)}(S_i-u)du\right)^{I_{(\varepsilon_i=1)}}\times\right.$$
$$\left.\left(\int_{W_i-d_i-\frac{m_i-1}{12}}^{W_i-d_i-\frac{m_i-1}{12}+1} f_\theta(u)\pi_\eta^{(2)}(S_i-u)du\right)^{I_{(\varepsilon_i=2)}}\left(\int_0^{S_i} f_\theta(u)\pi_\eta^{(3)}(S_i-u)du\right)^{I_{(\varepsilon_i=3)}}\right]^{\delta_i},$$
$$=\frac{\prod_{i=1}^n h(S_i,W_i,\varepsilon_i,\delta_i,m_i,d_i)}{\prod_{i=1}^n g_1(S_i)g_2(m_i)g_3(d_i)}. \tag{11}$$

The numerator is a product of densities of the form (10), while the denominator does not contain any information about θ. This likelihood can also be interpreted as a product of conditional densities of $(W_i,\varepsilon_i,\delta_i)$ given S_i, m_i and d_i, for $i=1,2,\ldots,n$. Further, this conditional likelihood is free from g_1, g_2 and g_3, i.e., inference for θ can proceed by ignoring any parameter of g_1, g_2 and g_3.

Since the likelihood (11) is identified as a product of densities, standard results for consistency and asymptotic normality of the MLE become applicable. However, while the usual conditions for these results are specified in terms of the density of Y_i, we would prefer conditions that involve the density f_θ (the density of T_i) and the functions $\pi_\eta^{(0)}, \pi_\eta^{(1)}, \pi_\eta^{(2)}$ and $\pi_\eta^{(3)}$, which define the conditional probability distribution of the random variable ε_i given T_i and S_i.

It may be verified that the following conditions on the model proposed in Section 2.1 imply the sufficient conditions for consistency given in Theorem 7.1.1 of Lehman (1999).

(C1) The parameters θ and η are identifiable with respect to the family of densities f_θ of the time-to-event and the family of functions $\pi_\eta^{(k)}, k=1,2,3$. In other words, $f_{\theta_1}=f_{\theta_2}$ implies $\theta_1=\theta_2$ and congruence of $\pi_{\eta_1}^{(k)}$ and $\pi_{\eta_2}^{(k)}$ for $k=1,2,3$ implies $\eta_1=\eta_2$.

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

(C3) The set $A_1=\{t: f_\theta(t)>0\}$ is independent of θ and the set $A_2=\left\{t:\pi_\eta^{(k)}(t)\in(0,1),\ \sum_{k=1}^3\pi_\eta^{(k)}(t)\in(0,1)\right\}$ is independent of η.

(C4) The functions $f_\theta(t)$, $\pi_\eta^{(1)}(t)$, $\pi_\eta^{(2)}(t)$ and $\pi_\eta^{(3)}(t)$ are differentiable with respect to θ and η for all t such that the derivative is absolutely bounded by a μ-integrable function.

The additional conditions for asymptotic normality are conditions 1-5 of Ferguson (1996), where the log-likelihood is

$$\ell(\theta,\eta) = \sum_{i=1}^{n}\Bigg[\delta_i I_{(\varepsilon_i=3)} \log\left(\int_0^{S_i} f_\theta(u)\pi_\eta^{(3)}(S_i-u)du\right) + \delta_i I_{(\varepsilon_i=2)} \log\left(\int_{Y_{i1}}^{Y_{i2}} f_\theta(u)\pi_\eta^{(2)}(S_i-u)du\right)$$
$$+ \delta_i I_{(\varepsilon_i=1)} \log\left(\int_{M_{i1}}^{M_{i2}} f_\theta(u)\pi_\eta^{(1)}(S_i-u)du\right) + \delta_i I_{(\varepsilon_i=0)} \log\left(f_\theta(T_i)\pi_\eta^{(0)}(S_i-T_i))\right) + (1-\delta_i)\log\left(\bar{F}_\theta(S_i)\right)\Bigg]. \tag{12}$$

3. SIMULATION OF PERFORMANCE

We compare the performance of MLE's based on the current status likelihood (1) (described here as Status MLE), the likelihood (2) based on binary recall i.e., whether the date of the event is recalled exactly or not (described here as Binary Recall MLE) and the likelihood (5) based on partial recall, where the recall information may be based on calendar time (described here as Partial Recall MLE). Computation of MLE's in all the cases is done through numerical optimization of likelihood using the Quasi-Newton method (Nocedal and Wright 2006).

For the purpose of simulation, we generate samples of time-to-event from the Weibull distribution with shape and scale parameters θ_1 and θ_2, respectively. Thus, $\theta = (\theta_1, \theta_2)$. We generate the recall probabilities through the multinomial logistic model as

$$\log\left(\pi_\eta^{(k)}(s-t)/\pi_\eta^{(0)}(s-t)\right) = \alpha_k + \beta_k(s-t), \qquad k = 1,2,3.$$

Since $\sum_{k=0}^{3} \pi_\eta^{(k)}(s-t) = 1$, the probabilities can be written as

$$\begin{aligned} \pi_\eta^{(0)}(s-t) &= 1/\big(1+\textstyle\sum_{k=1}^{3} e^{\alpha_k+\beta_k(s-t)}\big), \\ \pi_\eta^{(k)}(s-t) &= e^{\alpha_k+\beta_k(s-t)}/\left(1+\textstyle\sum_{k=1}^{3} e^{\alpha_k+\beta_k(s-t)}\right), \qquad k=1,2,3, \end{aligned} \tag{13}$$

where $\eta = (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3)$. Further, we generate the 'age at interview' from the discrete uniform distribution over [7,21].

We use the following values of the parameters.

(i) $\theta = (11, 13)$ and $\eta = (-0.05, -0.05, -0.05, 0.01, 0.01, 0.01)$,

(ii) $\theta = (11, 13)$ and $\eta = (-2, -1, -0.4, 0.05, 0.3, 0.02)$,

(iii) $\theta = (11, 13)$ and $\eta = (-2, -0.7, -1, 0.5, 0.06, 0.2)$,

(iv) $\theta = (11, 13)$ and $\eta = (-2, -2, -2, 0.3, 0.08, 0.08)$.

Note that for the chosen value of θ, the median of Weibull distribution turns out to be 11.57, which is in line with the median estimated from the same data under a simpler model proposed by Mirzaei et al. (2015). Also, the chosen values of η correspond to the following probabilities of different types of recall, five years after the event.

(i) $\pi_\eta^{(0)}(5) = \pi_\eta^{(1)}(5) = \pi_\eta^{(2)}(5) = \pi_\eta^{(3)}(5) = 0.25$,

(ii) $\pi_\eta^{(0)}(5) = 0.28$, $\pi_\eta^{(1)}(5) = 0.46, \pi_\eta^{(2)}(5) = 0.21, \pi_\eta^{(3)}(5) = 0.05$,

(iii) $\pi_\eta^{(0)}(5) = 0.23$, $\pi_\eta^{(1)}(5) = 0.15, \pi_\eta^{(2)}(5) = 0.23, \pi_\eta^{(3)}(5) = 0.38$,

(iv) $\pi_\eta^{(0)}(5) = 0.5$, $\pi_\eta^{(1)}(5) = 0.1, \pi_\eta^{(2)}(5) = 0.1, \pi_\eta^{(3)}(5) = 0.3$.

Note that choice (iv) is meant to favour the Binary Recall MLE, as the chances of recall up to a month or a year are slim. Choice (ii) should favour the Partial Recall MLE. Choice (iii), with a high probability attached to 'no recall', gives Status MLE its best chance. Choice (i) should not favour any particular method.

While computing the Binary Recall MLE, we assume the following form of the non-recall probability function π_η:

$$\log\Big(\pi_\eta(s-t)/1-\pi_\eta(s-t)\Big) = \alpha + \beta(s-t).$$

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

Tables 1 to 4 show the bias, the standard deviation (Stdev) and the mean squared error (MSE) for the MLE's of the parameter $\theta = (\theta_1, \theta_2)$, the median of time-to-event and the estimated exact recall probability when $s - t = 5$ based on the three likelihoods, for the combination of parameter values in case (i) to case (iv), respectively.

In cases (i)–(iii), it is found that the bias and the standard deviation (and consequently the MSE) of the Partial Recall MLE is less than those of the other two estimators and its performance improves with increasing sample size. The Status MLE, which uses the least amount of information from the data, has the poorest performance even in case (iii), where a substantial proportion of the subjects are designed to have no recollection of the event date. The substantial gap between the

Table 1: Bias, Stdev and MSE of estimated parameters for case (i) $\theta = (11, 13)$ *and* $\eta = (-0.05, -0.05, -0.05, 0.01, 0.01, 0.01)$

Estimator	Property	n	θ_1	θ_2	Median	Probability of exact recall
	Bias		9.391	–0.073	0.036	–
Status	Stdev		15.11	0.522	0.504	–
MLE	MSE		316.4	0.278	0.255	–
Binary	Bias		2.16	0.019	0.076	–0.005
Recall	Stdev	50	2.870	0.350	0.350	0.087
MLE	MSE		12.93	0.122	0.128	0.007
Partial	Bias		1.390	0.031	0.074	0.001
Recall	Stdev		1.730	0.239	0.243	0.086
MLE	MSE		4.910	0.058	0.065	0.007
	Bias		1.110	0.046	0.082	–
Status	Stdev		1.306	0.151	0.148	–
MLE	MSE		2.936	0.024	0.028	–
Binary	Bias		0.740	0.053	0.078	0.0002
Recall	Stdev	500	0.705	0.248	0.120	0.033
MLE	MSE		1.045	0.064	0.021	0.011
Partial	Bias		0.761	0.044	0.072	0.0004
Recall	Stdev		0.457	0.074	0.076	0.024
MLE	MSE		0.788	0.007	0.011	0.0006
	Bias		0.949	0.038	0.071	–
Status	Stdev		0.856	0.110	0.107	–
MLE	MSE		1.634	0.013	0.017	–
Binary	Bias		0.722	0.044	0.070	0.0004
Recall	Stdev	1000	0.494	0.079	0.079	0.017
MLE	MSE		0.765	0.008	0.011	0.0003
Partial	Bias		0.753	0.042	0.069	0.0007
Recall	Stdev		0.354	0.053	0.053	0.017
MLE	MSE		0.693	0.004	0.008	0.0003

Table 2: Bias, Stdev and MSE of estimated parameters for case (ii) $\theta = (11, 13)$ *and* $\eta = (-2, -1, -0.4, 0.05, 0.3, 0.02)$

Estimator	Property	n	θ_1	θ_2	Median	Probability of exact recall
	Bias		9.220	–0.10	0.020	–
Status	Stdev		14.40	0.516	0.491	–
MLE	MSE		293.8	0.276	0.241	–
Binary	Bias		1.916	–0.096	0.061	–0.0003
Recall	Stdev	50	2.660	0.331	0.334	0.098
MLE	MSE		10.70	0.112	0.116	0.009
Partial	Bias		1.170	0.027	0.064	–0.024
Recall	Stdev		1.500	0.206	0.210	0.088
MLE	MSE		3.620	0.043	0.048	0.008
	Bias		1.140	0.031	0.070	–
Status	Stdev		1.270	0.140	0.138	–
MLE	MSE		2.913	0.021	0.024	–
Binary	Bias		0.698	0.097	0.070	–0.014
Recall	Stdev	500	0.703	0.097	0.102	0.026
MLE	MSE		0.980	0.011	0.015	0.001
Partial	Bias		0.728	0.044	0.071	0.0004
Recall	Stdev		0.414	0.067	0.067	0.027
MLE	MSE		0.702	0.006	0.009	0.001
	Bias		0.905	0.044	0.075	–
Status	Stdev		0.864	0.105	0.103	–
MLE	MSE		1.567	0.013	0.016	–
Binary	Bias		0.590	0.044	0.067	–0.013
Recall	Stdev	1000	0.476	0.073	0.076	0.018
MLE	MSE		0.573	0.008	0.010	0.0005
Partial	Bias		0.071	0.004	0.072	0.0005
Recall	Stdev		0.298	0.047	0.048	0.019
MLE	MSE		0.572	0.004	0.007	0.0003

Table 3: Bias, Stdev and MSE of estimated parameters for case (iii) $\theta = (11, 13)$ *and* $\eta = (-2, -0.7, -1, 0.5, 0.06, 0.2)$

Estimator	Property	n	θ_1	θ_2	Median	Probability of exact recall
	Bias		9.611	–0.069	0.037	–
Status	Stdev		15.75	0.506	0.490	–
MLE	MSE		340.4	0.260	0.242	–
Binary	Bias		2.080	0.024	0.075	–0.033
Recall	Stdev	50	3.120	0.338	0.351	0.084
MLE	MSE		14.06	0.114	0.129	0.008
Partial	Bias		1.440	0.029	0.068	–0.004
Recall	Stdev		2.220	0.260	0.273	0.103
MLE	MSE		7.012	0.069	0.079	0.011
	Bias		1.158	0.045	0.082	–
Status	Stdev		1.270	0.152	0.149	–
MLE	MSE		2.945	0.025	0.031	–
Binary	Bias		0.703	0.045	0.069	–0.019
Recall	Stdev	500	0.740	0.105	0.108	0.024
MLE	MSE		1.041	0.013	0.020	0.001
Partial	Bias		0.699	0.045	0.070	0.0003
Recall	Stdev		0.550	0.080	0.082	0.027
MLE	MSE		0.792	0.008	0.012	0.0007
	Bias		0.929	0.046	0.078	–
Status	Stdev		0.850	0.107	0.105	–
MLE	MSE		1.586	0.013	0.017	–
Binary	Bias		0.600	0.044	0.066	–0.019
Recall	Stdev	1000	0.517	0.077	0.080	0.018
MLE	MSE		0.620	0.007	0.010	0.0007
Partial	Bias		0.655	0.043	0.067	0.0004
Recall	Stdev		0.384	0.059	0.061	0.021
MLE	MSE		0.576	0.005	0.008	0.0004

Table 4: Bias, Stdev and MSE of estimated parameters for case (iii) $\theta = (11, 13)$ *and* $\eta = (-2, -2, -2, 0.3, 0.08, 0.08)$

Estimator	Property	n	θ_1	θ_2	Median	Probability of exact recall
	Bias		9.251	–0.121	–0.015	–
Status	Stdev		14.94	0.650	0.621	–
MLE	MSE		308.7	0.432	0.386	–
Binary	Bias		1.230	0.007	0.047	–0.018
Recall	Stdev	50	2.200	0.440	0.470	0.101
MLE	MSE		6.350	0.193	0.220	0.010
Partial	Bias		1.250	0.017	0.054	0.013
Recall	Stdev		2.000	0.452	0.450	0.110
MLE	MSE		5.330	0.204	0.210	0.012
	Bias		1.090	0.034	0.070	–
Status	Stdev		1.260	0.143	0.141	–
MLE	MSE		2.770	0.021	0.024	–
Binary	Bias		0.634	0.040	0.065	0.205
Recall	Stdev	500	0.505	0.076	0.080	0.029
MLE	MSE		0.656	0.007	0.010	0.043
Partial	Bias		0.647	0.041	0.066	0.220
Recall	Stdev		0.516	0.075	0.081	0.030
MLE	MSE		0.685	0.007	0.011	0.050
	Bias		0.930	0.042	0.073	–
Status	Stdev		0.865	0.105	0.103	–
MLE	MSE		1.614	0.013	0.016	–
Binary	Bias		0.592	0.042	0.064	–0.010
Recall	Stdev	1000	0.415	0.050	0.062	0.020
MLE	MSE		0.520	0.004	0.007	0.0005
Partial	Bias		0.620	0.042	0.067	0.021
Recall	Stdev		0.370	0.051	0.053	0.020
MLE	MSE		0.518	0.004	0.008	0.0007

performance of the Binary Recall MLE and the Partial Recall MLE shows that the later estimator is able to utilize the additional information available from partial recall data. In case (iv) (the case where the parameters are chosen to produce lesser proportion of partial recalls, it is seen that the performance of the Binary Recall MLE is better than that of the proposed MLE.

4. ADEQUACY OF THE 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. For this purpose, the data may be transformed to the vector $Y = (S, W, \varepsilon, \delta, m, d)$, 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.

Modeling of the recall probability functions is a critical issue. There can be a trade off between a flexible model with many parameters 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 recall probabilities $\pi^{(1)}, \pi^{(2)}, \pi^{(3)}$ and $\pi^{(0)}$. Let us use the generic piecewise constant form of the probabilities:

$$\begin{aligned}
\pi^{(0)}(x) &= b_{01}I(x_1 < x \le x_2) + b_{02}I(x_2 < x \le x_3) + \ldots + b_{0k}I(x_k < x < \infty),\\
\pi^{(1)}(x) &= b_{11}I(x_1 < x \le x_2) + b_{12}I(x_2 < x \le x_3) + \ldots + b_{1k}I(x_k < x < \infty),\\
\pi^{(2)}(x) &= b_{21}I(x_1 < x \le x_2) + b_{22}I(x_2 < x \le x_3) + \ldots + b_{2k}I(x_k < x < \infty),\\
\pi^{(3)}(x) &= b_{31}I(x_1 < x \le x_2) + b_{32}I(x_2 < x \le x_3) + \ldots + b_{3k}I(x_k < x < \infty). \qquad (14)
\end{aligned}$$

where $0 = x_1 < x_2 < \cdots < x_k$ are a chosen set of time-points and $b_{l1}, b_{l2}, \ldots, b_{lk},\ l = 0, 1, 2, 3$ are unspecified parameters taking values in the range $[0, 1]$ such that $\sum_{l=0}^{3} b_{lj} = 1$ for $j = 1, 2, \ldots, k$. Then the likelihood (5) reduces to

$$\begin{aligned}
L(\theta, \eta) = \prod_{i=1}^{n} [\bar{F}_\theta(S_i)]^{1-\delta_i} \Bigg[&\left\{ f_\theta(T_i) \left(\sum_{l=1}^{k} b_{0l} I\big(S_i - x_{l+1} < T_i \le S_i - x_l\big) \right) \right\}^{I_{(\varepsilon_i=0)}} \times \\
&\left\{ \sum_{l=1}^{k} b_{1l} \Big(F_\theta\big(\min(S_i - x_l, M_{i2})\big) - F_\theta\big(\max(S_i - x_{l+1}, M_{i1})\big) \Big) \right\}^{I_{(\varepsilon_i=1)}} \times \\
&\left\{ \sum_{l=1}^{k} b_{2l} \Big(F_\theta\big(\min(S_i - x_l, Y_{i2})\big) - F_\theta\big(\max(S_i - x_{l+1}, Y_{i1})\big) \Big) \right\}^{I_{(\varepsilon_i=2)}} \times \\
&\left\{ \sum_{l=1}^{k} b_{31l} \Big(F_\theta(S_i - x_l) - F_\theta(S_i - x_{l+1}) \Big) \right\}^{I_{(\varepsilon_i=3)}} \Bigg]^{\delta_i}.
\end{aligned}$$

If the distribution of T is known, one can obtain the MLE of the parameters $b_{l1}, b_{l2}, \ldots, b_{lk},\ l = 0, 1, 2, 3$. Newton–Raphson iterative steps may be used to determine the conditional MLE of the piecewise constant functions $\pi^{(1)}, \pi^{(2)}, \pi^{(3)}$ and $\pi^{(0)}$, for any given F_θ. Using a parametric form $\pi_\eta^{(1)}, \pi_\eta^{(2)}, \pi_\eta^{(3)}$ and $\pi_\eta^{(0)}$, one can first estimate the MLEs $\hat{\theta}$ and $\hat{\eta}$ and then compare the plots of $\hat{\pi}_\eta^{(1)}, \hat{\pi}_\eta^{(2)}, \hat{\pi}_\eta^{(3)}$ and $\hat{\pi}_\eta^{(0)}$ with the plots of the conditional MLE of the piecewise constant versions of $\pi^{(1)}, \pi^{(2)}, \pi^{(3)}$ and $\pi^{(0)}$, with F_θ held fixed at $F_{\hat{\theta}}$. This graphical comparison can be used to judge the suitability of the recall probability functions, as we illustrate in the next section.

5. DATA ANALYSIS

For the data set described in Section 1, the landmark event is the onset of menarche in young and adolescent females. We used the Weibull model for menarcheal age and the multinomial logistic model for the recall probabilities $\pi_\eta^{(0)}, \pi_\eta^{(1)}, \pi_\eta^{(2)}$ and $\pi_\eta^{(3)}$, as in Section 3. We used the three different methods mentioned in Section 3 for estimating the parameters θ_1 and θ_2 as well as the median of the age at menarche. Table 5 gives a summary of the findings.

The Partial Recall MLE of the median age at menarche is closer to the Binary Recall MLE of the median. The standard errors of the Partial Recall MLE are smaller than the corresponding standard errors of the other two estimators.

The survival functions estimated from the three models are shown in Figure 2. The Status MLE is found to be different from the other two MLE's. This may have been due to the excessive variance of the Status MLE. Though there appears to be little difference between the Binary Recall MLE and the Partial Recall MLE, their standard errors are different. Figure 3 shows the plot of the width of the asymptotic 95% confidence interval (C.I.) of the estimated survival function based

Table 5: Estimated parameters and median age at menarche from different methods for the menarcheal data

Estimator	$\theta_1(Stdev)$	$\theta_2(Stdev)$	Median$(Stdev)$
Status MLE	10.76(0.51)	12.18(0.07)	11.77(0.031)
Binary Recall MLE	10.86(0.33)	12.33(0.05)	11.92(.018)
Partial Recall MLE	10.37(0.24)	12.39(0.04)	11.96(.0096)

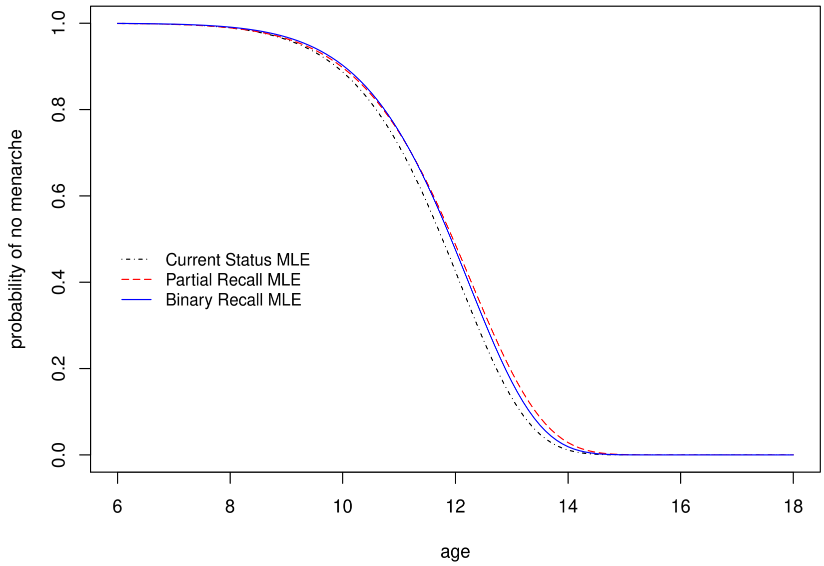

Figure 2: Survival functions for the menarcheal data based on three methods.

on the three likelihoods. It is clear that the C.I. for the Partial Recall MLE has smaller width than those of the other estimators. Therefore, the Partial Recall MLE and its confidence intervals may be preferred here.

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 hexatuple $(S, W, \varepsilon, \delta, m, d)$ as follows.

The range of S is split into the intervals $[6.99, 14]$ and $(14, 21]$,

the range of W is split into the sets $\{0\}$, $(0, 11.84]$ and $(11.84, 15.76]$,

the range of ε has four points, 0, 1, 2 and 3, which are not clubbed,

the range of δ has two points, 0 and 1, which are not clubbed,

d takes value in its whole range $[0, 1/12]$, which are clubbed,

and m takes value in its whole range $\{1, 2, \ldots, 12\}$, which are clubbed.

When $\delta = 0$, the value of ε is irrelevant and W can only be zero. Thus, there are only two bins, corresponding to the discretized value of S. Similarly, when $\delta = 1$ and $\varepsilon = 3$, W can only be zero

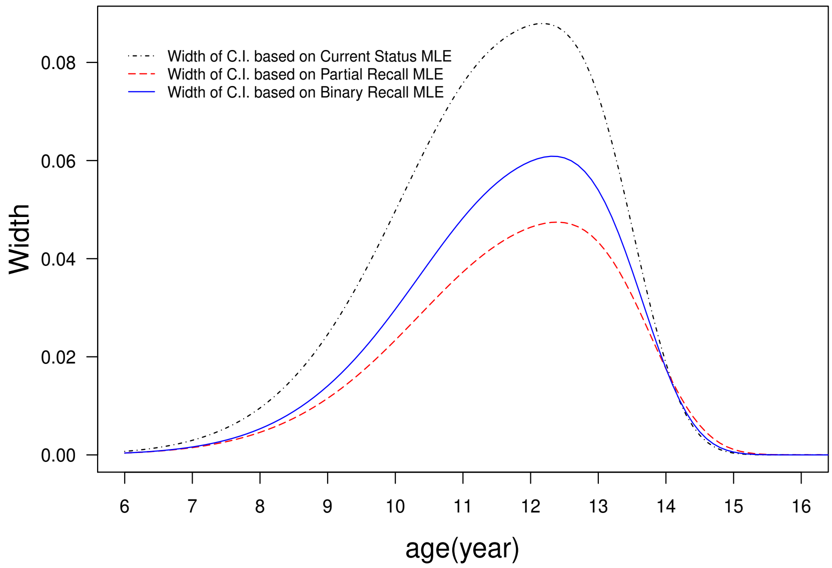

Figure 3: Width of asymptotic 95% C.I. of survival function for the menarcheal data based on three methods.

and again there are only two bins. When $\delta = 1$ and $\varepsilon = 0, 1$ or 2, in each case there are four bins arising from two groups of values of S and two groups of non-zero values of W. Thus we have a total of 16 bins. After merging two bins with small expected frequencies with neighboring bins, we have a reduced total of 14 bins. Further, there are eight parameters to estimate. Thus, the null distribution should be χ^2 with 5 degrees of freedom. The p–value of the test statistic for the given data happens to be 0.188. Therefore, violation of the chosen model is not indicated.

As we mentioned in the last section, one can check the adequacy of the functional form of the $\pi_\eta^{(l)}$'s by comparing the $\pi_{\hat{\eta}}^{(k)}$'s with the conditional MLE of a piecewise constant function (14). We use segments of one 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 θ_1 and θ_2 fixed at the values reported in Table 5, we obtain the conditional MLE of the values of $\pi_\eta^{(0)}$, $\pi_\eta^{(1)}$, $\pi_\eta^{(2)}$, $\pi_\eta^{(3)}$ in the different segments. Figure 4(a) shows the plots of the estimated recall probabilities under the logistic and the piecewise constant models in the range 0 to 12 years, when the number of segments are assumed to be $k = 4$. The estimated functions under the logistic and piecewise constant models are found to be close to each other for $l = 0, 1, 2, 3$. Figure 4(b) shows

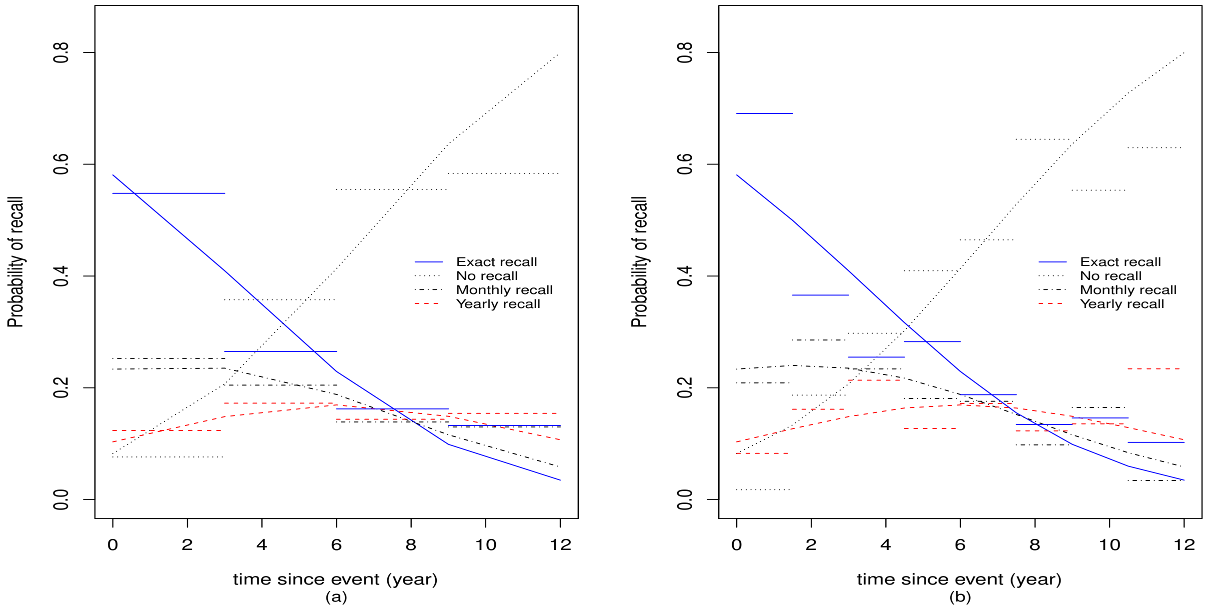

Figure 4: Approximation of estimated logistic recall probabilities with estimated piecewise constant recall probabilities with (a) 4 pieces, (b) 8 pieces.

the same plots when $k = 8$. Similarity of the two sets of estimates is evident yet again. This finding justifies the choice of the logistic form of the recall probability functions.

We have seen the cumulative proportions of decreasing degrees of recall for different age ranges in the case of the menarcheal data in Figure 1. As an additional graphical check for the assumed model, we consider the model based estimates of the same cumulative proportions for ages $s = 11, 14, 17$ and 20 (i.e., at the middle of the respective age intervals). We used the Partial Recall MLE of parameters $\hat{\theta}$ and $\hat{\eta}$ of the menarcheal data set to calculate $f_{\hat{\theta}}$ and $\pi_{\hat{\eta}}^{(j)}$ for $j = 0, 1, 2, 3$ and then computed the requisite probabilities through numerical integration. Figure 5 shows the comparison of the cumulative proportions in different age groups of interview (shown in solid lines) with the corresponding model based estimates (shown in dashed lines). The closeness of the cumulative proportions of recall probability with their model based estimates also support the choice of the overall model.

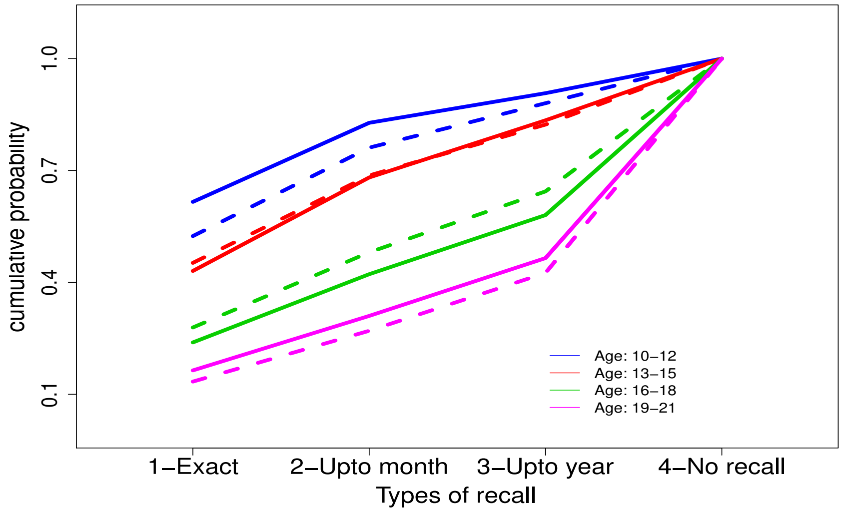

Figure 5: Cumulative proportions of decreasing degrees of recall (solid lines) for different age ranges in menarcheal data and model based estimates of these proportions at the middle of the age interval (dashed lines).

6. CONCLUDING REMARKS

The aim of this paper has been to offer a realistic model for time-to-event based on partial recall information through a realistic informative censoring model, where the range of relevant dates may depend on calendar time (rather than time since the event). The simulations and the data analysis of the menarcheal data set show that there is much to be gained from partial recall information in the form of the event falling in a calendar month or a calendar year. Many other forms of partial recall information may be handled in a similar way. As the simulations reported in Section 3 show, a particular category of partial recall (eg. recall upto a calendar month or year) is justified if that category is not very rare in the data.

The recalled time-to-event can sometimes be erroneous. Grouping of the uncertainly recalled event date by the calendar month or year may reduce the error to some extent. If one adopts this solution, the method presented in this paper provides a viable method of analysis. Skinner and Humphreys (1999), while working with data without instances of non-recall, has modeled erroneously recalled time-to-event as $t_i' = t_i k_i$, where t_i is the correct time-to-event and k_i is a multiplicative error of recall that is independent of t_i. Since k_i's are unobservable, they have used a mixed-effects regression model to account for erroneous recalls. One may investigate whether a similar adjustment in the term $f_\theta(T_i)$ of the likelihood (5), improves the analysis.

It would also be of interest to get rid of any model for the time-to-event, and to look for a non-parametric estimator on the basis of the likelihood (5). 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. The authors thank Professor Parasmani Dasgupta of the Biological Anthropology Unit of ISI, for making the data available for this research.

A. PROOF OF THEOREM 1

In the second case, the density can be derived as,

$$
\begin{aligned}
h(s,w,1,1,m,d) &= g_1(s)g_2(m)g_3(d)\frac{\partial P(W<w,\delta=1,\varepsilon=1|s,m,d)}{\partial w}\\
&= g_1(s)g_2(m)g_3(d)\lim_{h\to 0}\frac{P(w<W\leqslant w+h,\delta=1,\varepsilon=1|s,m,d)}{h}\\
&= g_1(s)g_2(m)g_3(d)\lim_{h\to 0}\frac{P(w<T\leqslant w+h,T\leq s,\varepsilon=1)}{h}\\
&= g_1(s)g_2(m)g_3(d)\lim_{h\to 0}\frac{P(w<T\leqslant w+h,\varepsilon=1)}{h}\\
&= g_1(s)g_2(m)g_3(d)\lim_{h\to 0}\frac{E_T[P(w<T\leqslant w+h|T)\pi_0(s-T)I_{(w\leq s)}]}{h}\\
&= g_1(s)g_2(m)g_3(d)\lim_{h\to 0}\frac{\int_w^{w+h} f_\theta(u)\pi_0(s-u)du I_{(w\leq s)}}{h}\\
&= g_1(s)g_2(m)g_3(d)f_\theta(w)\pi_0(s-w)I_{(w\leq s)}.
\end{aligned}
$$

The density in the other cases can be obtained by considering the corresponding probability masses:

$$\begin{aligned}
h(s,w,\varepsilon,0,m,d) &= P(W=0,\delta=0|s,m,d)g_1(s)g_2(m)g_3(d) \\
&= P(T>S|S=s)g_1(s)g_2(m)g_3(d) = \bar{F}_\theta(s)g_1(s)g_2(m)g_3(d); \\
h(s,w,0,1,m,d) &= E_T[g_1(s)g_2(m)g_3(d)P(T\le s|T,m,d,s)\pi_1(s-T)] \\
&= \int_0^s g_1(s)g_2(m)g_3(d)f_\theta(u)\pi_1(s-u)du \\
&= g_1(s)g_2(m)g_3(d)\int_0^s f_\theta(u)\pi_1(s-u)du;
\end{aligned}$$

$$\begin{aligned}
h(s,w,&2,1,m,d) = g_1(s)g_2(m)g_3(d)P(W=w,\varepsilon=2,\delta=1|s,m,d) \\
&= g_1(s)g_2(m)g_3(d)P(\lfloor 12(d+T)\rfloor/12 = w,\varepsilon=2,\delta=1|s,m,d) \\
&= g_1(s)g_2(m)g_3(d)P(12w\le 12(d+T)<12w+1,\varepsilon=2,\delta=1|s,m,d) \\
&= g_1(s)g_2(m)g_3(d)\int_{w-d}^{w+\frac{1}{12}-d} f_\theta(u)\pi_2(s-u)du; \\
h(s,w,&3,1,m,d) = g_1(s)g_2(m)g_3(d)P(W=w,\varepsilon=3,\delta=1|s,m,d) \\
&= g_1(s)g_2(m)g_3(d)P(\lfloor\big(T+d+(m-1)/12\big)\rfloor = w,\varepsilon=3,\delta=1|s,m,d) \\
&= g_1(s)g_2(m)g_3(d)P(w-d-(m-1)/12\le T<w+1-d-(m-1)/12,\varepsilon=3,\delta=1|s,m,d) \\
&= g_1(s)g_2(m)g_3(d)\int_{w-d-\frac{m-1}{12}}^{w+1-d-\frac{m-1}{12}} f_\theta(u)\pi_3(s-u)du;
\end{aligned}$$

REFERENCES

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

Dasgupta, P. (2015), *Physical Growth, Body Composition and Nutritional Status of Bengali School aged Children, Adolescents and Young adults of Calcutta, India: Effects of Socioeconomic Factors on Secular Trends. (in collaboration with M Nub, Sengupta D and de Onis M),* Available at URL http://www.neys-vanhoogstraten.nl/wp-content/uploads/2015/06/Academic-Report-ID-158.pdf.

Ferguson, T. S. (1996), *A Course in Large Sample Theory.* Texts in Statistical Science Series. Chapman & Hall, London.

Gillespie, B., dÁrcy, H., Schwartz, K., Bobo, J. K., and Foxma, B. (2006), "Recall of age of weaning and other breastfeeding variables," *International Breastfeeding Journal*, pp. 1–4.

Harel, Y., Overpeck, M. D., Jones, D. H., Scheidt, P. C., Bijur, P. E., Trumble, A. C., and Anderson, J. (1994), "The effects of recall on estimating annual nonfatal injury rates for children and adolescents.," *American Journal of Public Health*, 84(4), 599–605.

Joffe, M., Villard, L., Li, Z., Plowman, R., and Vessey, M. (1995), "A time to pregnancy questionnaire designed for long term recall: validity in Oxford, England.," *Journal of Emidemiology and community health*, 49, 314–319.

Koo, M. M., and Rohana, T. E. (1997), "Accuracy of short-term recall of age at menarche," *Annals of Human Biology*, 24, 61–64.

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*, New York: Springer-Verlag.

Mathiowetza, N. A., and Ouncanb, G. J. (1988), "Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment," *Journal of Business & Economic Statistics*, 6(2), 221–229.

Mirzaei, S. S., Sengupta, D., and Das, R. (2015), "Parametric estimation of menarcheal age distribution based on recall data," *Scandinavian Journal of Statistics.*, 42, 290–305.

Nocedal, J., and Wright, S. J. (2006), *Numerical Optimization*, second edn, New York: Springer Series in Operations Research and Financial Engineering. Springer.

Skinner, C. J., and Humphreys, K. (1999), "Weibull regression for lifetimes measured with error," *Lifetime Data Anal.*, 5, 23–37.
URL: *http://dx.doi.org/10.1023/A:1009674915476*