Recent Advances in the Statistical Analysis of Retrospective Time-to-Event Data

Sedigheh Mirzaei Salehabadi and Debasis Sengupta

Abstract In a cross-sectional observational study on time-to-event, the probability distribution of that time is often estimated from data on current status. Recall data on the time of occurrence of the landmark event can provide more information in this regard. Even so, the subjects may not be able to recall the time precisely. This type of incompleteness is a peculiarity of recall data, which poses a challenge to analysis. Valid likelihood-based procedures for inference have emerged in a number of papers published only recently. In this article, we review these papers and show how one can estimate the time-to-event distribution parametrically or nonparametrically, and also assess the effect of covariates, by using current status data or incompletely recalled data. The methods are illustrated through the analysis of menarcheal data from a recent anthropometric study of adolescent and young adult females in Kolkata, India.

Keywords Current status data · Informative censoring · Interval censoring
Relative risk regression model · Retrospective study · Turnbull estimator

1 Introduction

Time to occurrence of an event is an object of interest in various fields. Observational studies have been carried out to study the time until onset of menarche of females (Bergsten-Brucefors 1976; Chumlea et al. 2003; Mirzaei and Sengupta 2015), breast development of females (Cameron 2002; Aksglaede et al. 2009), dental development of infants (Demirjian et al. 1973; Eveleth and Tanner 1990), birth of the first child of a woman (Allison 1982), beginning of a criminal career (Hosmer and Lemeshow 1999), end of a work career (LeClere 2005), end of a strike (Hosmer and Lemeshow 1999), and so on. In anthropometric studies, the age of passing various

S. M. Salehabadi
E. K. Shriver National Institute of Child Health and Human Development, Bethesda, MD, USA

D. Sengupta (✉)
Applied Statistical Unit, Indian Statistical Institute, Kolkata, India
e-mail: sdebasis@isical.ac.in

R. Dasgupta (ed.), *Advances in Growth Curve and Structural Equation Modeling*,
https://doi.org/10.1007/978-981-13-1843-6_9

developmental landmarks is examined by their own right, and also as useful covariates for body dimensions used for obtaining growth curves (Salsberry et al. 2009; Vizmanos et al. 2001). One may wish to estimate the probability distribution of the time to occurrence of a particular event in order to compare two populations. Such estimates may also be useful in setting benchmarks for individuals or setting policy objectives. Most of the observational studies on time-to-event are cross-sectional in nature, though there are some instances of study designs for observing a number of individuals continuously or periodically until the occurrence of the landmark event (Korn et al. 1997; McKay et al. 1998).

There are many parametric models for the probability distribution of the time-to-event, viz. exponential, Weibull, lognormal, gamma, Gompertz, log-logistic, Pareto, generalized gamma. Once a parametric model for the time-to-event has been chosen, standard techniques for parametric inference become applicable. However, these techniques are meant for complete data. Cross-sectional time-to-event data may be incomplete in many ways. For example, the time would not be known in the case of individuals who did not experience the event. If one records only the current status of the individual in terms of the happening of the event, the time-to-event is not recorded even for those who have experienced the event. If the interviewed individual is asked to recall the time of occurrence of the event, there may be occasional cases of complete failure to remember. This would result in another form of incompleteness. Yet more complex forms of incompleteness would arise if some individuals are only able to recall a range of time when the event had occurred.

Most of the data arising from these situations can be broadly referred to as censored data. There are modified versions of likelihood-based techniques, which work for censored data. However, the nature of modification depends on the nature of censoring. One has to make certain assumptions about the censoring mechanism in order to be able to specify an appropriate likelihood. A key assumption which is often made is that the mechanism of censoring is independent of the time-to-event. This assumption essentially means that a particularly long or particularly short time-to-event does not have any more or any less chance of being censored, compared to other cases. It can be shown that this assumption can be particularly problematic for data obtained through recall. The event of recall induces a special type of dependent censoring that has been specifically modeled in recently published literature.

This article is intended to provide an up-to-date overview of the methods of inference available to those who aspire to analyze time-to-event data collected from a cross-sectional study, without going deeply into the technical details, which can always be obtained from the original sources cited here. We focus on methods that are based on likelihood. Consequently, many popular methods, such as those based on probit model for the event of menarche before a specific age (Hediger and Stine 1987), are excluded from the purview of our discussion.

The remainder of this article is organized as follows. Section 2 reviews the current status data on time-to-event and likelihood-based inference procedures available for it. Section 3 deals with perfectly recalled time-to-event data and the relevant procedures. Section 4 dwells on parametric and nonparametric inference in the case where some of the time-to-event are not recalled at all. Section 5 shows how one can incorporate the effect of covariates on time-to-event distribution through regression models for various types of cross-sectional data. In Sect. 6, there is a brief discussion on partial recall and recall error. An illustrative data analysis is reported in Sect. 7. The data analysis is based on a study of menarcheal age of adolescent and young adult females, undertaken by the Indian Statistical Institute, Kolkata. Some concluding remarks are given and areas for future work are identified in Sect. 8.

2 Current Status Data

Current status data, also known as status quo data (Teilmann et al. 2009), consist of the value of a binary status variable that indicates whether or not the landmark event has occurred till the day of observation.

Consider a set of n subjects with the landmark event occurring at times $T_1, \ldots, T_n$, which are samples from a common distribution F with density f and support $[t_{min}, t_{max}]$. Let these subjects be observed at times $S_1, \ldots, S_n$, respectively, chosen from a finite set $\mathcal{S}$. Let, for $i = 1, \ldots, n$, δ_i be the indicator of $T_i \leq S_i$, i.e., the event having had occurred on or before the time of interview.

Current status data arise from the observation consisting only of (S_i, δ_i), $(i = 1, 2, \ldots, n)$. The corresponding likelihood, conditional on the time of interview, is

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

where $\bar{F} = 1 - F$. If the distribution F is assumed to be a member of a parametric family characterized by the parameter θ (which may be a vector of parameters), then the parametric MLE is obtained by maximizing the above likelihood with respect to θ. There has been considerable interest in the parametric analysis of current status data (Shiboski and Jewell 1992; Sun and Kalbfleisch 1993). For properties of the MLE based on the above likelihood, see Lee and Wang (2003).

It is also possible to estimate the distribution nonparametrically, that is, without assuming any particular functional form of the distribution. Note that if the ith respondent is observed to have experienced the event of interest, then it is known that the time-to-event T_i belongs to the interval $[t_{min}, S_i]$. If the event has not been experienced, then T_i belongs to the interval $[S_i, t_{max}]$. In either case, T_i is known to belong to an interval. This is a special case of interval censoring, sometimes referred to as Case I interval censoring (Sun 2006).

In general, interval censoring refers to the situation where one only knows that the time-to-event lies in a certain window of time; i.e., T_i belongs to an observed interval $[L_i, R_i]$. The case of no censoring $(L_i = R_i)$ can be indicated by the binary variable η_i. When the data contain instances of no censoring $(L_i = R_i)$, censoring from the right $(R_i = t_{max})$, censoring from the left $(L_i = t_{min})$, and censoring from both sides ($t_{min} < L_i < R_i < t_{max}$), the censoring is called mixed interval censoring (Sun 2006, Chap. 2).

If the censoring mechanism is independent of the time T_i (an assumption that usually holds for current status data), the general likelihood for interval-censored data is

$$\prod_{i=1}^{n} [f(T_i)]^{\eta_i} [\bar{F}(L_i) - \bar{F}(R_i)]^{1-\eta_i}, \tag{2}$$

where f is the probability density function corresponding to the distribution F. Note that in the case of current status data, $\eta_i = 0$ for every i, and $[L_i, R_i]$ is constrained to be either $[t_{min}, S_i]$ or $[S_i, t_{max}]$, so that the likelihood (2) reduces to (1). A nonparametric maximum likelihood estimator (NPMLE) of F for general interval-censored data would be the distribution function that maximizes the above likelihood. This NPMLE was derived by Ayer et al. (1955). Turnbull (1976) worked on it further and gave a computational algorithm. This algorithm consists of partitioning the range $[t_{min}, t_{max}]$ into disjoint subintervals, such that every observed interval $[L_i, R_i]$ can be expressed as a union of these subintervals. There can only be a finite number of such subintervals. Once this partitioning is done, the task of identifying the NPMLE reduces to allocating optimum probabilities to these subintervals so that the total probability is 1 and the above likelihood is maximized. See Keiding et al. (1996) for details of this estimator, generally known as the Turnbull estimator.

The Turnbull estimator has an undesirable characteristic. When a subinterval is of positive length (i.e., left and right end-points do not coincide), the probability allocated to that interval can be distributed in any manner within the interval, without affecting the value of the likelihood. In other words, the NPMLE is not unique. Two different distribution functions that allocate identical probabilities to each subinterval (while distributing the probability within the intervals in different ways) can happen to be NPMLEs. In the case of current status data, every single subinterval is likely to be of positive length. Therefore, the ambiguity about the NPMLE prevails everywhere, except at the boundaries of the subintervals! Practically speaking, the NPMLE specifies a distribution only at a finite number of points and is silent about how they should be interpolated to obtain the full description of a distribution function.

A desirable property of an estimator is that when the sample size is increased, it should be probabilistically very close to the quantity being estimated. This property is called consistency. Consistency of an estimator, under appropriate conditions, needs to be established for it to be credible. This holds for estimators of single parameters, vector parameters, and even functions. In particular, when a distribution function is estimated by a function computed from the data, it should converge to the true distribution function, under appropriate conditions, as the sample size goes to infinity. In the case of an NPMLE of a distribution function obtained from interval-censored data, this requirement poses a conceptual problem, since the NPMLE is only the set of values of a function at a few points and not a fully specified function. Gentleman and Geyer (1994) brought in the requisite formalism to establish the consistency of the NPMLE of F from independently interval-censored data.

Various methods of inference for interval-censored data have been explained in books such as Sun (2006), Kalbfleisch and Prentice (2002), and Lee and Wang (2003).

3 Perfectly Recalled Time Data

In some cross-sectional studies, a subject is asked to recall the time of occurrence of the landmark event, in case it has already taken place. Such retrospective data are usually incomplete (Roberts 1994; Padez 2003). The subject may not be able to recall the time at all or may be able to specify only a range for the requisite time. Even if there is no difficulty of recall (which may happen, for instance, if there is a formal record of the time of occurrence), there would be incompleteness in the data in respect of those individuals who did not experience the event yet. In this section, we only consider the latter situation, where there is no problem of recall and the only incompleteness arises from the possible nonoccurrence of the event at the time of data collection.

Going by the notations used in the previous section, the observable quantities in this situation are T_i when $\delta_i = 1$ and S_i when $\delta_i = 0$. The censoring involved here is from the right, in the sense that the time-to-event is longer than the time of observation (censoring time). This is a special case of interval censoring, with $R_i = L_i = T_i$ when $\delta_i = 1$ and $L_i = S_i,\ R_i = t_{max}$ when $\delta_i = 0$. The simplified form of the likelihood (2) is

$$\prod_{i=1}^{n} [f(T_i)]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}. \tag{3}$$

Assuming S_i is random and independent of T_i, we essentially have randomly right-censored data, which has been dealt with extensively in the literature. Usual large sample properties of many parametric likelihood-based techniques have been shown to hold for randomly right-censored data, under appropriate conditions (Lawless 2003). Modifications of goodness-of-fit tests for randomly right-censored data have also been proposed (Lawless 2003, Chap. 10). If one does not assume the functional form of the distribution, the above likelihood can be maximized with respect to the function $\bar{F}$ to obtain the nonparametric MLE. This NPMLE happens to be the well-known Kaplan–Meier estimator. For properties of this estimator, two-sample tests, and other related procedures, see Kalbfleisch and Prentice (2002), Hosmer et al. (2008), Lawless (2003), and Klein and Moeschberger (2003).

4 Recalled Time Data with Occasional Failure to Recall

Let us now consider the situation where a subject may not be able to recall the time of the event of interest. Non-recall necessarily means that the time-to-event T_i can have any value smaller than the time till observation (S_i), which corresponds to left censoring. Here, we ignore the possibility of the subject recalling an approximate date and regard such occurrence as a non-recall event. Thus, the entire data set would consist of only three types of cases: complete data arising from the cases of perfect recall, left-censored data arising from the case of non-recall, and right-censored data arising from the cases where the event did not take place yet. These cases can be described by the binary variable δ_i, which indicates whether the event happened till the time of observation ($T_i \leq S_i$), and another binary variable ε_i, which indicates whether the time of the event is recalled at all (assuming that it has happened). Specifically, the three cases correspond to $\delta_i \epsilon_i = 1$, $\delta_i(1 - \epsilon_i) = 1$, and $\delta_i = 0$.

Such a data set can be readily seen to be a special case of interval-censored data, discussed in Sect. 2, where the likelihood (2) reduces to

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

However, this likelihood and the related procedures are applicable only when the censoring mechanism is independent of the time-to-event. Incompleteness in recall data in a cross-sectional study occurs in such a way that this assumption is violated. This is because of the fact that memory often fades with time. Between two persons interviewed at the same age, the one with earlier occurrence of the event of interest has less chance of recalling the time. Mirzaei et al. (2014) and Mirzaei and Sengupta (2016) have shown that the use of the likelihood (2) can lead to biased estimation, both in the parametric and the nonparametric cases, though there are instances when the NPMLE (Turnbull estimator) has been used for studying the distribution of age at reaching a developmental landmark by using recall data (see, e.g., Aksglaede et al. 2009).

In some existing models and methods for dependent censoring (see, e.g., Finkelstein et al. 2002; Scharfstein and Robins 2002), censoring is assumed to occur through duration variables that have the same origin of measurements as that of the duration of interest. Since this assumption does not hold here, these methods are not applicable. Mirzaei et al. (2014) took into account the special type of incompleteness arising from recall data by new modeling. They recognized that the non-recall probability may depend on the observation time and the time-to-event, and modeled it as a function π of the time elapsed since the occurrence of the event till the time of observation,

$$P(\varepsilon_i = 0 | S_i = s, T_i = t) = \pi(s - t),$$

where $s > t > 0$. The three types of data mentioned above would lead to different contributions to the likelihood. By putting these cases together, the likelihood according to this model can be shown to be

$$\prod_{i=1}^{n} \left[\left(\int_0^{S_i} f(u)\pi(S_i - u)du \right)^{1-\varepsilon_i} [f(T_i)(1 - \pi(S_i - T_i))]^{\varepsilon_i} \right]^{\delta_i} [\bar{F}(S_i)]^{1-\delta_i}. \tag{5}$$

When π is a constant, the likelihood (5) becomes a constant multiple of the independent interval censoring likelihood (4). As a further special case, if $\pi = 1$, it reduces to the current status likelihood (1). On the other hand, when $\pi = 0$, the likelihood reduces to the perfect recall likelihood (3). Thus, the model that leads to the likelihood (5) is more general than the models for independent censoring.

Mirzaei et al. (2014) assumed parametric forms of the functions π and F, established consistency and asymptotic normality of the MLE under the above model, subject to suitable regularity conditions. They also suggested a graphical method of guessing the functional form of the non-recall probability π. Mirzaei and Sengupta (2016) allowed the distribution function to be arbitrary and eliminated the integral in the likelihood (5) by assuming a piecewise constant form of π:

$$\pi(x) = \begin{cases} b_1 & \text{if } x_1 < x \leq x_2, \\ b_2 & \text{if } x_2 < x \leq x_3, \\ \vdots & \\ b_k & \text{if } x_k < x < \infty, \end{cases} \tag{6}$$

where $0 = x_1 < x_2 < \cdots < x_k$; $0 < b_1, b_2, \ldots, b_k \leq 1$. They derived the NPMLE of F obtained by maximizing the resulting likelihood, which can be obtained through a self-consistency algorithm. Significantly, they showed that when the sample size is large, the NPMLE tends to have probability concentrated only on the distinct times of exactly recalled events. Accordingly, they proposed an approximate NPMLE (AMLE), which is computationally much simpler and is asymptotically equivalent to the NPMLE. The AMLE is obtained by maximizing the approximate likelihood, written in terms of the probabilities $q_1, \ldots, q_m$ attached to the exactly recalled event times $t_1, \ldots, t_m$, as the product of weighted sums

$$\prod_{i=1}^{n} \left(\sum_{j=1}^{m} \alpha_{ij} q_j \right). \tag{7}$$

The weights α_{ij} are computable from the data as linear functions of $b_1, \ldots, b_k$, which may be regarded as nuisance parameters while maximizing (7) with respect to $q_1, \ldots, q_m$. Mirzaei and Sengupta (2016) discussed how the variance of the AMLE can be estimated. They showed that both the NPMLE and the AMLE are consistent estimators of the underlying distribution under general conditions.

The two-sample problem for data of this type has not been addressed yet. A solution under the restriction of proportional hazards may be obtained by considering the Cox regression model with a single binary covariate, discussed in the next section.

5 Regression

All the likelihoods presented in the three foregoing sections are based on the assumption that the underlying time-to-event for all the individuals are independent and have a common distribution F with density f. If each individual has a different distribution, the same likelihoods continue to hold after F and f in the factors are replaced by their individual-specific versions: F_i and f_i, respectively.

A parametric regression model provides a functional description of the distribution of T_i given the covariate vector Z_i in terms of the distribution parameters θ and the regression parameters β. Specifically, $F_i(t|Z_i)$ can be written as $F_0(t|Z_i, \theta, \beta)$, where F_0 is a known 'baseline distribution.' This substitution reduces the problem of obtaining the MLEs of the regression parameters as another optimization problem with β and θ (and possibly the parameters of the function π) as optimizing variables. This problem is conceptually similar to parametric estimation. Standard procedures (see, e.g., Lee andWang 2003) with appropriate modification of asymptotic results are applicable.

In recent years, semiparametric regression models have gained popularity. These models deal with covariates parametrically, while keeping a nonparametric flavor as far as the baseline distribution is concerned. They make fewer assumptions than a completely parametric model, but more assumptions than a model that would assign a different time-to-event distribution to every case. This amounts to expressing $F_i(t|Z_i)$ as $F_0(t|Z_i, \beta)$, where F_0 is a completely unspecified distribution function. Examples of semiparametric regression models are Cox's relative risk model (Cox 1972), the accelerated failure time (AFT) model (Wei 1992), the additive hazard regression model (Klein and Moeschberger 2003), the proportional odds model (Dabrowska 1988), and so on. A summary of the methods available for randomly right-censored data may be found in Hosmer et al. (2008). For current status data, Huang (1996) provided consistent estimators of covariate effects under Cox's proportional hazards regression model. See Huang and Wellner (1997), for a review of various methods for other regression models, with special emphasis on current status data. See Sun (2006) for an updated summary of regression models and methods for general interval-censored data under the assumption of independent censoring.

Mirzaei and Sengupta (2015) considered regression under Cox's model for the special type of dependent censoring arising from recall data with the possibility of non-recall. When this model is combined with the likelihood (5), the resulting likelihood becomes

$$\prod_{i=1}^{n}[\bar{F}_i(S_i|Z_i)]^{1-\delta_i}\left[\{f_i(T_i|Z_i)(1-\pi(S_i-T_i))\}^{\varepsilon_i}\left(\int_0^{S_i} f_i(u|Z_i)\pi(S_i-u)du\right)^{1-\varepsilon_i}\right]^{\delta_i}, \tag{8}$$

where

$$\bar{F}_i(t|Z_i) = [\bar{F}_0(t)]^{\exp(\beta Z_i)}, \tag{9}$$

$f_i(t|Z_i)$ being the derivative of $F_i(t|Z_i)$. The above likelihood is meant to be maximized with respect to β, the (possibly vector) parameter used to describe the function π and the unspecified function F_0. Mirzaei and Sengupta (2015) simplified the optimization problem by (a) removing integrals through a piecewise constant form of π and (b) restricting probability allocations of the baseline distribution to the distinct times of precisely recalled events. Their simulation results show that chi-square tests of significance, obtained from the likelihood in the conventional manner after disregarding the nonparametric nature of the likelihood and the approximations involved, produce reasonably reliable p values. An R program for fitting this model is available from the authors on request.

6 Imperfect or Erroneous Recall

As mentioned in Sects. 2 and 4, it is possible that respondents may recall only a range of time for the event of interest. Mirzaei et al. (2016) found in the case of a menarcheal data set (partially analyzed in the next section) that, rather than remembering a range of ages for the age at menarche, respondents often remember a range of calendar dates for the occurrence of the event. Thus, the different types of partial recall can be grouped into recalling the month of occurrence, the year of occurrence, and so on, apart from the scenario of no recall at all. They proposed a multinomial logistic model for the recall probabilities and were able to extend the parametric method reported in Sect. 4 to this situation. Work on nonparametric estimation and extension of the Cox regression model is in progress.

It should be noted that the recalled time-to-event can sometimes be erroneous (see Beckett et al. 2001). Grouping of the cases of partial recall as above might reduce the impact of recall error somewhat, but would not address the issue specifically. There have been some attempts to incorporate this fact into the modeling through latent variables (see, e.g., Rabe-Hesketh et al. 2001). However, adapting such modeling to recall data would require further research.

7 Data Analysis

The data we use here are based on an anthropometric study conducted by the Indian Statistical Institute in and around the city of Kolkata, India, from 2005 to 2011 (Dasgupta 2015, p.108). A total of 2195 randomly selected females, aged between 7 and 21 years, were surveyed. The subjects were interviewed on or around their birthdays. The data set contains age, some physical information of each individual, menarcheal status, age at menarche (if recalled), and some socioeconomic information. For this data set, the landmark event is the onset of menarche. Among the 2195 cases 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

Table 1 Estimated parameters and median age at menarche from different methods for real data

Estimator	Estimate (standard error)			Median	95% Confidence interval of median
	α	β	η		
Current status MLE	10.74 (0.320)	12.17 (0.005)		11.76	(11.62, 11.90)
Interval censored MLE	11.80 (0.061)	12.65 (0.001)		12.25	(12.20, 12.30)
Binary recall MLE	10.19 (0.090)	12.21 (0.001)	3.47 (0.140)	11.78	(11.72, 11.84)

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 was to estimate the distribution of the age at onset of menarche and the dependence of age at menarche on socioeconomic variables. For simplicity, we dichotomize the recalled information; i.e., we club the cases of partial and no recall and refer to them as cases of no exact recall.

To illustrate the parametric approach, we used the Weibull model for menarcheal age and the exponential model with scale parameter η for non-recall probability. We compared the performance of MLEs based on the current status likelihood (1) (described here as current status MLE), the likelihood (4) based on interval-censored data with noninformative censoring (described here as interval-censored MLE), and the likelihood (5) based on binary recall information when the censoring mechanism is recognized as informative (described here as binary recall MLE). Computation of MLEs in all the cases is done through numerical optimization of likelihood using the quasi-Newton method (see Nocedal and Wright 2006). Table 1 gives a summary of the findings. The interval-censored MLE of the median is somewhat different from the other two MLEs, which is possibly because of the bias of the former. The binary recall MLE has a narrower confidence interval for the median than the current status MLE.

As another illustration, in Fig. 1, we compare graphically the closeness of the parametric estimator of the time-to-event distribution with the AMLE (see Sect. 4), for the menarcheal data set when a piecewise constant model of π with $k = 8$ is used for the non-recall probability. The jump points of the piecewise constant function are assumed to be evenly distributed over the range 0–13 years (maximum possible separation between menarcheal age and age at observation in the sample). The two estimators are somewhat close to one another.

The age at menarche can potentially be affected by diet and physical activities. These factors can be related to more easily measured socioeconomic variables such as parents' education and monthly family expenditure (Khan et al. 1996; Padez 2003; Aryeetey et al. 2011). We considered the monthly family expenditure in Indian rupees (indexed with respect to 2008 as base year) and a couple of binary variables indicating

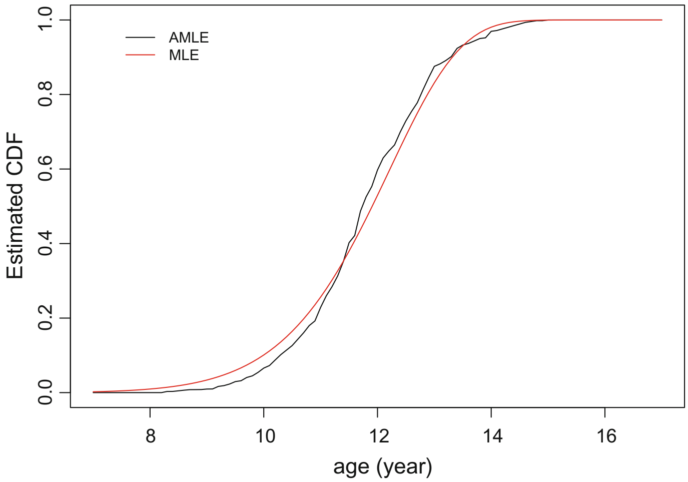

Fig. 1 Comparison of MLE and AMLE of menarcheal age distribution

Table 2 Estimated regression coefficients and their p values

Covariates	Estimated value	p value
Whether father passed high school	0.091	0.0036
Whether mother passed high school	0.249	0.0061
Monthly family expenditure	0.0002	0.0047

whether the father or the mother of the respondent had passed high school. The present analysis concerns a subset of the original data, consisting of respondents who came from a nuclear family and were the only child of their respective parents. Among the total of 673 respondents, 241 individuals did not have menarche, 147 individuals had menarche and recalled the date of its onset, and the remaining 285 individuals had menarche but could not recall the date. The median of monthly family expenditure was Rupees 7808. The fathers of 492 respondents and the mothers of 420 respondents had passed high school.

The estimated regression coefficients and the corresponding p values are reported in Table 2. All the coefficients are found to be significant at the 1% level. The p value of the combined hypothesis of insignificance of all the three regression coefficients is 0.00093.

We now consider four hypothetical subjects with covariate profiles Z described below.

$Z = (0, 0, 7808)$: Monthly family income is Rupees 7808 (median income of the group), neither parent passed high school.

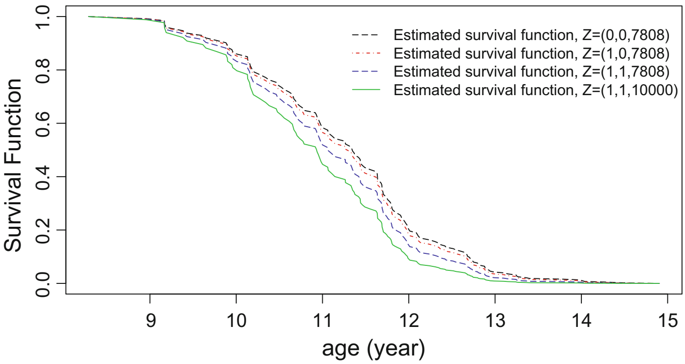

Fig. 2 Estimated survival function in different cases

$Z = (1, 0, 7808)$: Monthly family income is Rupees 7808, only the father passed high school.
$Z = (1, 1, 7808)$: Monthly family income is Rupees 7808, both the parents passed high school.
$Z = (1, 1, 10000)$: Monthly family income is Rupees 10000, both the parents passed high school.

A comparative plot of the estimated survival functions of these four subjects is given in Fig. 2. Father's status of having passed high school is found to be associated with earlier maturation. The mothers having the same qualification are seen to have an even greater impact in the form of earlier maturation. A 28% higher monthly family expenditure is also found to have a considerable impact on the survival function of the age at menarche.

8 Concluding Remarks

Cross-sectional time-to-event data obtained from recall have been found to be surprisingly complex in terms of the nature of incompleteness. Many interesting questions have been answered in recent years through careful modeling, and many more remain to be answered. We have indicated in Sect. 5 how the Cox regression model can be fitted in the case of recall data with the possibility of non-recall. Fitting of other regression models and adapting such models to partial recall data remain to be explored. Further challenges include handling of recall error and of random effects (frailty).

Acknowledgements This research was partially sponsored by the project 'Physical growth, body composition and nutritional status of the Bengal school aged children, adolescents, 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, leader of the project, for making the data available for this research. Also, the first author thanks

Dr. Bibhas Chakrobarty for his financial support through the Duke-NUS start-up grant R-913-200-074-263, the NIH grant 1 R01 DE023072-01 and the Singapore Ministry of Education grant MOE2015-T2-2-056.

References

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

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, *13*, 61–98.

Aryeetey, R., Ashinyo, A., & Adjuik, M. (2011). Age at menarche among basic level school girls in Medina, Accra. *African Journal of Reproductive Health*, *103*, 103–110.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., & Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, *26*, 647–647.

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

Bergsten-Brucefors, A. (1976). A note on the accuracy of recalled age at menarche. *Annals of Human Biology*, *3*, 71–73.

Cameron, N. (2002). *Human growth and development*. Academic Press.

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

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, *34*, 187–220.

Dabrowska, D. M., & Doksum, K. A. (1988). Estimation and testing in a two-sample generalized odds-rate model. *Journal of the American Statistical Association*, *83*, 744–749.

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. Nubé, D. Sengupta & M. de Onis). http://www.neys-vanhoogstraten.nl/wp-content/uploads/2015/06/Academic-Report-ID-158.pdf

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

Eveleth, P. B., & Tanner, J. M. (1990). *Worldwide variation in human growth* (2nd ed.). Cambridge University Press.

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

Gentleman, R., & Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, *81*, 618–623.

Hediger, M. L., & Stine, R. A. (1987). Age at menarche based on recall data. *Annals of Human Biology*, *14*, 133–142.

Hosmer, D. W., & Lemeshow, S. (1999). *Applied survival analysis: Regression modeling of time to event data*. John Wiley.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis* (2nd ed.). Hoboken: John Wiley.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics*, *24*, 540–568.

Huang, J., & Wellner, J. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data*. New York: John Wiley.

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

Khan, A. D., & Schroeder, D. G., Reynaldo, M., Haas, J. D., & Rivera, J. (1996). Early childhood determinants of age at menarche in rural Guatemala. *American Journal of Human Biology*, *8*, 717–723.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. New York: Springer-Verlag.

Korn, E. L., Graubard, B. I., & Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale. *American Journal of Epidemiology*, *145*, 72–80.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data* (2nd ed.). New York: John Wiley.

LeClere, M. J. (2005). Modeling time to event: Applications of survival analysis in accounting, economics and finance. *Review of Accounting and Finance*, *4*, 5–12.

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

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. *Journal of Pediatrics*, *13*, 682–687.

Mirzaei, Salehabadi S., & Sengupta, D. (2015). Regression under Coxs model for recall-based time-to-event data in observational studies. *Computational Statistics & Data Analysis*, *92*, 134–147.

Mirzaei, Salehabadi S., & Sengupta, D. (2016). Nonparametric estimation of time-to-event distribution based on recall data in observational studies. *Lifetime Data Analysis*, *22*, 473–503.

Mirzaei, Salehabadi S., Sengupta, D., & Das, R. (2014). Parametric estimation of menarcheal age distribution based on recall data. *Scandinavian Journal of Statistics*, *42*, 290–305.

Mirzaei, S. S., Sengupta, D., & Ghosal, R. (2016). Estimation of menarcheal age distribution from imperfectly recalled data. Applied Statistical Unit, Technical Report No. ASU/2016/4, Indian Statistical Institute. http://www.isical.ac.in/asu/TR/TechRepASU201604.pdf

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

Padez, C. (2003). Age at menarche of schoolgirls in Maputo, Mozambique. *Annals of Human Biology*, *30*, 487–495.

Rabe-Hesketh, S., Yang, S., & Pickles, A. (2001). Multilevel models for censored and latent responses. *Statistical Methods in Medical Research*, *10*, 409–427.

Roberts, D. F. (1994). Secular trends in growth and maturation in British girls. *American Journal of Human Biology*, *6*, 13–18.

Salsberry, P. J., Reagan, P. B., & Pajer, K. (2009). Growth differences by age of menarche in African American and white girls. *Nursing Research*, *58*, 382–390.

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

Shiboski, S. C., & Jewell, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association*, *87*, 360–372.

Sun, J., & Kalbfleisch, J. D. (1993). The analysis of current status data on point processes. *Journal of the American Statistical Association*, *88*, 1449–1454.

Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. New York: Springer.

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.

Turnbull, Bruce W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, *38*, 290–295.

Vizmanos, B., Marti-Henneberg, C., Clivillé, R., Moreno, A., & Fernández-Ballart, J. (2001). Age of pubertal onset affects the intensity and duration of pubertal growth peak but not final height. *American Journal of Human Biology*, *13*, 409–416.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis (with discussion). *Statistics in Medicine*, *11*, 1871–1879.