Accounting for length-bias and selection effects in estimating the distribution of menstrual cycle length

KIRSTEN J. LUM
Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS, Rockville, MD 20852, USA and Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

RAJESHWARI SUNDARAM
Division of Intramural Population Health Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, DHHS, Rockville, MD 20852, USA

THOMAS A. LOUIS
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA and U.S. Census Bureau, Suitland, MD 20746, USA
tlouis@jhu.edu

SUMMARY
Prospective pregnancy studies are a valuable source of longitudinal data on menstrual cycle length. However, care is needed when making inferences of such renewal processes. For example, accounting for the sampling plan is necessary for unbiased estimation of the menstrual cycle length distribution for the study population. If couples can enroll when they learn of the study as opposed to waiting for the start of a new menstrual cycle, then due to length-bias, the enrollment cycle will be stochastically larger than the general run of cycles, a typical property of prevalent cohort studies. Furthermore, the probability of enrollment can depend on the length of time since a woman’s last menstrual period (a backward recurrence time), resulting in selection effects. We focus on accounting for length-bias and selection effects in the likelihood for enrollment menstrual cycle length, using a recursive two-stage approach wherein we first estimate the probability of enrollment as a function of the backward recurrence time and then use it in a likelihood with sampling weights that account for length-bias and selection effects. To broaden the applicability of our methods, we augment our model to incorporate a couple-specific random effect and time-independent covariate. A simulation study quantifies performance for two scenarios of enrollment probability when proper account is taken of sampling plan features. In addition, we estimate the probability of enrollment and the distribution of menstrual cycle length for the study population of the Longitudinal Investigation of Fertility and the Environment Study.

Keywords: Backward recurrence time; Length-bias; Menstrual cycle length; Renewal process; Selection effects; Survival analysis

*To whom correspondence should be addressed.

Published by Oxford University Press 2014. This work is written by (a) US Government employee(s) and is in the public domain in the US.
1. Introduction

We consider the setting of longitudinal prospective studies of cyclic events forming a renewal process, where the primary objective is to estimate the distribution of interarrival times. A necessary consideration when making inferences about the study population based on the observed times is the sampling plan; in particular, one in which enrollment can occur at any time in a gap interval including its start or end, and the enrollment date may differ among participants. After enrollment, generally participants are followed with cycle renewals recorded until censoring at the end of the study, after a fixed number of cycles, loss to follow-up, or occurrence of a competing event. Also, the gap time can be truncated.

For example, the process underlying the periodic shedding of the lining of the uterus known as menstruation is a stationary renewal process in that the occurrence of bleeding events is cyclic over time. The interarrival time of menstrual bleeding events, or menstrual cycle length, is an important determinant of risk for several chronic diseases afflicting females including heart disease and breast cancer; and it is also an important component in assessing female fecundity (see e.g. Jensen and others, 1999). It is typically defined as the number of days from the first day of menstrual bleeding (the day marking the last menstrual period, LMP) to the day preceding the next menstrual cycle, with care taken to distinguish menstrual bleeding from episodic bleeding. Cycle length is commonly measured in time-to-pregnancy studies either via retrospective recall or more reliably in prospective designs using daily diaries of intensity of bleeding along with hormonal monitoring. In this paper, we will analyze cycle lengths from a comprehensive prospective pregnancy study, the Longitudinal Investigation of Fertility and the Environment (LIFE) Study (Buck Louis, Schisterman and others, 2011) in which repeated measurements of menstrual cycle length and couple’s preconception exposure to persistent environmental chemicals were collected. Research goals include estimating the distribution of menstrual cycle length accounting for covariates and residual variation, and using the estimated distribution to predict time-to-pregnancy. An initial step is to estimate the distribution of menstrual cycle length, both in a population of women and for an individual woman.

A complicating feature of the LIFE and other such studies of renewal processes is that in order to efficiently sample couples attempting to become pregnant, enrollments occur between renewals as opposed to waiting for the next bleeding event. Furthermore, the interarrival time is potentially right censored by pregnancy or exit from the study resulting in four different combinations of truncation and censoring relevant to the observed enrollment menstrual cycle lengths (more generally, the gap times) as described by Gill and Keiding (2010). In modeling the distribution of (potentially censored) enrollment lengths, we account for two features of this sampling process; length-bias and selection effects. Length-bias relates to the process by which an announcement of a study has a probability of “intersecting” a menstrual cycle length that is proportional to its length. Due to length-biased sampling, the distribution of enrollment lengths is biased in the positive direction compared with the population distribution. Selection relates to the possible dependence of the probability that the couple will decide in favor of enrollment on the time from LMP until they learn of the study (the backward recurrence time), resulting in effects that could be positive or negative; thus, adding to or canceling out the length-bias.

A common and possibly unbiased approach for estimating the menstrual time distribution is to circumvent these challenges by excluding the enrollment cycle lengths and analyzing only post-enrollment cycle lengths (see, e.g. Murphy and others, 1995), conceding some loss of information. However, in some applications the additional information provided by the enrollment data can be important in detecting subtle covariate effects and the selection effects may in themselves be of interest. Moreover, in prospective pregnancy studies, the most fecund women may have very few cycles under observation, if any, post-enrollment.

Several authors have developed methods for estimating the unbiased distribution of survival times from length-biased or left-truncated data (see, e.g. Asgharian and others, 2002; Cox, 1969; Huang and Qin, 2011; Vardi, 1989; Wang, 1989, 1991; Wang and others, 1986; Winter and Foldes, 1988), which would
allow for inclusion of the enrollment menstrual cycle lengths. Building upon this body of work, we propose an approach for estimating both the probability of enrollment as a function of time from LMP and the distribution of menstrual cycle length for the study population, adjusting for length-bias and selection effects. In Sections 2–5, we formalize the issue and develop our estimation approach with a focus on enrollment data. In Sections 6 and 7, we present results from a simulation study and analysis of the distribution of menstrual cycle length for the LIFE study population, respectively. We conclude with a discussion of remaining challenges and impact on analysis and design of longitudinal studies of recurrent events.

2. Estimating the distribution of menstrual cycle length

Our aim is to estimate the distribution of menstrual cycle length for a population focusing on the role of the enrollment cycle data, as the distribution estimated from observations of this cycle can be influenced by the way in which these cycles are ascertained. We assume that the enrollment cycle lengths for different individuals are independent and identically distributed. The sampling (enrollment process) induces a distribution for the initial observed cycle length that could potentially be different from that for the population. Specifically, we account for length-bias and selection effects on the enrollment cycle length distribution so that inferences relate to a reference population. We first consider the uncensored case and then generalize.

2.1 Notation, sampling plan and consequences

Let $\tilde{Y} \sim f(\cdot)$, $0 < \tilde{Y} < \infty$ with associated survivor function $S(y) = \int_{y}^{\infty} f(u) \, du$ and cumulative distribution function (CDF) $F(y) = \int_{0}^{y} f(u) \, du$. Without loss of generality, assume $\tilde{Y} < V < \infty$, where $V$ is a finite upper bound. Assume that the sequence $(\tilde{Y}_1, \tilde{Y}_2, \ldots)$, indexed by $m$, starts at some origin ($T_0$) and continues over time with $T_m = T_0 + \tilde{Y}_1 + \tilde{Y}_2 + \cdots + \tilde{Y}_m$. Consider the following sampling plan for a longitudinal study:

- An announcement of study availability is issued at calendar time $t^*$ and it “intersects” cycle length $\tilde{Y}_{m(t^*)}$ with $m(t^*) = \{ m : T_{m-1} < t^* \leq T_m \}$ at $\tilde{B}$, the backwards recurrence time. That is, $\tilde{B}$ is the time from $T_{m-1}$ to $t^*$ and can take on values in the range $0 \leq \tilde{B} < V$. In addition, let,
  - $\lambda(\tilde{Y}) = \text{pr}(\text{the announcement “intersects” a cycle of length } \tilde{Y})$.
  - $\bar{A}$ be the residual time from enrollment to the next event, so $\tilde{Y}_{m(t^*)} = \tilde{B} + \bar{A}$. We consider the case where both $\tilde{B}$ and $\bar{A}$ are known. As examples, Keiding and others (2002) and Keiding and others (2012) studied estimation of the distribution of time-to-pregnancy in the current-duration design when only $\tilde{B}$ was known; and Brookmeyer and Gail (1987) studied bias in estimation of the distribution of time to disease when $\tilde{B}$ was unknown, but $\bar{A}$ was known.
  - The potential participant (the couple) decides on enrolling in the study. Let $R = 1$ if the participant enrolls and $R = 0$ otherwise. If the participant enrolls, the enrollment cycle length is $\tilde{Y}_{m(t^*)}$.
    - Assume that $t^*$ is stochastically independent of the $T_m$ sequence. As the underlying process $\tilde{Y}$ is a renewal process, using the theorem proved by Winter (1989) and details included in Section 1.1 of supplementary material available at Biostatistics online, $[\tilde{B} \mid \tilde{Y}_{m(t^*)}] \sim \mathcal{U}(0, \tilde{Y}_{m(t^*)})$, $0 \leq \tilde{B} < \tilde{Y} < V$.
    - A potential participant knows only the $\tilde{B}$ value and not $\tilde{Y}_{m(t^*)}$, and so define the probability of enrollment, $\pi(\tilde{B}) = \text{pr}(R = 1 \mid \tilde{Y}_{m(t^*)}, \tilde{B}) = \text{pr}(R = 1 \mid \tilde{B})$. 

A $\lambda(Y) \propto Y$ produces standard length-bias analogous to the waiting time paradox (see, Feller, 1971) or, for example, in Aalen and Husebye (1991) pertaining to the cyclic movements in the small bowel during the fasting state. Just as longer bus waiting times are more likely to be intersected by an individual arriving at a bus stop, longer cycle lengths are more likely to be intersected by the announcement of the study than are shorter. This length-biased sampling results in a distribution of enrollment cycle lengths that is stochastically larger than the population distribution. Further, a non-constant $\pi(\cdot)$ induces additional selection effects and, depending on the shape of $\pi(\cdot)$, can induce a stochastically larger or smaller distribution for an enrollment cycle.

In modeling the enrollment cycle length, one must account for the aforementioned length-bias and selection effects on the first observation of the underlying renewal process. Our focus is on estimating the distribution of $Y$ using a likelihood that adjusts for the enrollment sampling process and henceforth we consider only enrollment lengths.

### 2. Modeling

If $\lambda(Y) \equiv \bar{\lambda}$ and $\pi(\tilde{B}) \equiv \bar{\pi}$, then neither length-bias nor selection effects operate and standard modeling is valid. If $\lambda(Y) \propto Y$ and $\pi(\tilde{B}) \equiv \bar{\pi}$, then $\tilde{Y}_{m(r^*)}$ is distributed according to the length-biased distribution: $f(y)\mu^{-1}$, where $\mu = E(\tilde{Y})$ (see, Cox, 1962). If, in addition, $\pi(\tilde{B})$ depends on $\tilde{B}$, then $\tilde{Y}_{m(r^*)}$ is distributed as a more generally weighted version of $f(\cdot)$. Consequently, statistical models must account for the enrollment sampling plan by conditioning on $R = 1$. To simplify notation, henceforth we drop the "$\cdot$" notation and subscript $m(t^*)$, let $[Y, B] \overset{D}{=} [\tilde{Y}_{m(r^*)}, \tilde{B} \mid R = 1]$ and denote this enrollment distribution by $g$. We have the joint distribution

$$g_{Y,B}(y, b) = f_{Y, B}(y, b \mid R = 1) = \frac{f(y)\bar{\lambda}(y)\pi(b) (1/y)}{\int_{0}^{\infty} f(u)\bar{\lambda}(u)(1/u)[\int_{0}^{u} \pi(t) \, dt] \, du}, \quad 0 \leq b < y < V < \infty.$$  

The marginal distribution of $Y$ is

$$g_{Y}(y) = \frac{f(y)\bar{\lambda}(y)(1/y) \int_{0}^{y} \pi(t) \, dt}{\int_{0}^{\infty} f(u)\bar{\lambda}(u)(1/u)[\int_{0}^{u} \pi(t) \, dt] \, du}, \quad 0 < y < V < \infty; \quad (2.1)$$

a weighted distribution with weight $\omega(\cdot, y) \propto \bar{\lambda}(y)(1/y) \int_{0}^{y} \pi(t) \, dt$ and CDF $G_Y(y) = \int_{0}^{y} g_{Y}(u) \, du$. The marginal distribution for $B$ is

$$g_{B}(b) = \pi(b) \frac{\int_{b}^{\infty} f(u)\bar{\lambda}(u)(1/u) \, du}{\int_{0}^{\infty} f(u)\bar{\lambda}(u)(1/u)[\int_{0}^{u} \pi(t) \, dt] \, du}, \quad 0 \leq b < V < \infty; \quad (2.2)$$

which is proportional to $\pi(b)$, and if $\bar{\lambda}(Y) \propto Y$, to $S(b)$.

### 3. Parameter estimation

Assume that $f(\cdot) = f(\cdot \mid \theta)$, and that recruitment for the longitudinal study continues until $n$ participants are enrolled. At enrollment, participants report the LMP of the current cycle, so observed data for participant $i$ are $(b_i, y_i)$. The full likelihood is the product of the conditional likelihood ($[y \mid b]$) and the marginal
likelihood for $B$. Specifically, with $f(\cdot) = f(\cdot \mid \theta)$, assuming that the functions $\lambda(\cdot)$ and $\pi(\cdot)$ are known,

$$
\mathcal{L}(\theta; y \mid b, \lambda) = \mathcal{L}_c(\theta; y \mid b, \lambda) \times \mathcal{L}_m(\theta; b, \lambda),
$$
(3.1)

$$
\mathcal{L}_c(\theta; y \mid b, \lambda) = \prod_{i=1}^{n} g_{yi} | b_i(y_i \mid b_i) = \prod_{i=1}^{n} \frac{f(y_i)\lambda(y_i)(1/y_i)}{\int_{b_i}^{\infty} f(u)\lambda(u)(1/u)\,du},
$$
(3.2)

and from (2.2), the marginal likelihood of $B$ is

$$
\mathcal{L}_m(\theta \mid b, \pi, \lambda) = \prod_{i=1}^{n} \frac{\pi(b_i)\int_{b_i}^{\infty} f(u)\lambda(u)(1/u)\,du}{\int_{0}^{\infty} f(u)\lambda(u)(1/u)\left\{\int_{0}^{u} \pi(t)\,dt\right\}\,du}.
$$
(3.3)

Note that $\mathcal{L}_m$ depends only on the shape of $\pi(\cdot)$ and not on its magnitude (is invariant to multiplication of $\pi(\cdot)$ by a positive constant).

We focus on estimating the survivor (tail) function $S(y) = \int_{y}^{\infty} f(u)\,du$ or in the parametric case $S(y \mid \theta) = \int_{y}^{\infty} f(u \mid \theta)\,du$. Under the renewal process sampling plan in Section 2.1, we henceforth assume that $\lambda(y) \propto y$. If $\pi(\cdot)$ can be estimated, it is straightforward to use the full likelihood. The conditional likelihood ($\mathcal{L}_c$) does not depend on $\pi(\cdot)$ and so if this selection function is not of interest, inferences can be based on (3.2). Including the marginal likelihood ($\mathcal{L}_m$) may provide a notable increment in information on $\theta$, but it does depend on $\pi(\cdot)$. As (2.2) shows, $\pi(\cdot)$ and $f$ interact, and a recursive approach is effective.

### 3.1 Recursive estimation

The following recursive algorithm requires $g_B$ (equivalently the CDF $G_B$). Using the observed backward recurrence times, $g_B$ can be estimated by the empirical probability mass function (equivalently $G_B$ by the empirical distribution function). Alternatively, a smoothed estimate of $g_B$, can be obtained using adaptive bandwidth kernel density estimation (see, e.g. Wand and Jones, 1994). A two-stage algorithm for estimating $\pi(\cdot)$ and $S(\cdot)$ proceeds as follows:

1. Conditioning on $B$, let $\hat{S}^{(0)}(\cdot)$ be a non-parametric estimate of $S(\cdot)$ based on the conditional likelihood (see, e.g. Winter and Foldes, 1988). For parametric estimation, equivalently let $\hat{\theta}^{(0)}$ be the estimate of $\theta$ based on maximizing $\mathcal{L}_c(\theta; y \mid b, \lambda)$ in (3.2).
2. Using the relation between $\pi(\cdot)$ and $g_B$ in (2.2), obtain $\hat{\pi}^{(1)}(\cdot)$ (only the shape is needed), with $S(\cdot) = \hat{S}^{(0)}(\cdot)$.
3. Let $\hat{S}^{(v)}(\cdot)$ be the non-parametric maximum likelihood estimate using the full likelihood, with $\pi(\cdot) = \hat{\pi}^{(v)}(\cdot)$. Equivalently, for parametric estimation let $\hat{\theta}^{(v)}$ be the maximum likelihood estimate (MLE) using $\mathcal{L}(\theta; y \mid b, \lambda)$ in (3.1).
4. Using relation (2.2), update $\hat{\pi}^{(v+1)}(\cdot)$ (up to a constant of proportionality) with $S(\cdot) = \hat{S}^{(v)}(\cdot)$.
5. Iterate between steps 3 and 4 until convergence criterion is met.

In Sections 1.2 and 1.3 of supplementary material available at Biostatistics online, we provide a proof of the existence and uniqueness of the solution for $\pi$.

### 3.2 Non-parametric estimation of $S(y)$

Assuming $\lambda(y) \propto y$ and $\pi(b) \equiv \tilde{\pi}$, the inverse probability weighted empirical estimate of $S(y)$ derived by Cox (1969) for uncensored lengths $(y_i; i = 1, \ldots, n)$ sampled from $G_Y$ is

$$
\hat{S}(y) = \sum_{i=1}^{n} w_i \mathbb{1}\{y_i \geq y\}; \quad y > 0, \quad \text{where} \quad w_i = \frac{1/y_i}{\sum_{j=1}^{n} 1/y_j}.
$$
(3.4)
Under the sampling plan in Section 2.1, we have a single sample of uncensored enrollment cycle lengths and backward recurrence times \(\{ (y_i, b_i); i = 1, \ldots, n \}\) and we allow for a non-constant \(\pi(b)\). In this case, \(Y\) has the marginal density, \(g_Y(\cdot)\), given in (2.1) and it is straightforward to derive the relation:

\[
E \left[ \left\{ \lambda(Y_i) \frac{1}{Y_i} \int_0^{Y_i} \pi(t) \, dt \right\}^{-1} \mathbb{1}\{Y_i \geq y\} \right] = \frac{S(y)}{\int_0^\infty f(u) \lambda(u)(1/u) [\int_0^u \pi(t) \, dt] \, du}.
\]

Since this expectation is proportional to \(S(y)\), we estimate \(S(y)\) as in (3.4), but with modified weights:

\[
w_i = \frac{[\lambda(y_i)/(1/y_i)] \int_0^{y_i} \pi(t) \, dt}{{\sum}_{j=1}^n [\lambda(y_j)/(1/y_j)] \int_0^{y_j} \pi(t) \, dt}^{-1}.
\]

The weights in (3.5) depend on \(\pi(\cdot)\) up to a multiplicative constant; therefore, the recursive algorithm presented in Section 3.1 can be amended to include an additional step prior to step 3 updating \(w_i^{(0)}\) using (3.5) with \(\pi(\cdot) = \hat{\pi}(\cdot)\).

For right-censored data, assuming \(\pi(\cdot) \equiv \bar{\pi}\), several non-parametric approaches for estimating \(S(\cdot)\) accounting for length-bias or left-truncation have been developed (Asgharian and others, 2002; Huang and Qin, 2011; Vardi, 1989). In order to relax the assumption on \(\pi(\cdot)\), we note that conditioning on \(B\) the likelihood does not depend on \(\pi(\cdot)\). Methods for estimating \(S(\cdot)\) based on \(L_c\) have been developed by many including Keiding and Gill (1990), Wang (1991), Wang and others (1986), and Winter and Foldes (1988). When estimation of \(\pi(\cdot)\) is not of interest, these methods allow for non-parametric estimation of \(S(\cdot)\), but may lead to a loss of efficiency when compared with using the full likelihood. In the simulation study (see Section 6), we do not find evidence of an efficiency loss.

### 3.3 Semi-parametric estimation of \(S(y \mid \theta)\)

In the semi-parametric approach, we assume \(\bar{Y} \sim f(\cdot; \theta)\); however, we do not make any parametric assumptions about \(\pi(\cdot)\). The recursive algorithm given in Section 3.1 may be used to estimate \(\pi(\cdot)\) and \(S(y \mid \theta) = \int_y^\infty f(u \mid \theta) \, du\) with the adaptation in step 3 of producing \(\hat{\theta}^{(0)}\) from the full likelihood written in terms of the current estimate of \(\pi(\cdot) = \hat{\pi}^{(0)}(\cdot)\). As we will show in Section 4.1, the semi-parametric approach can accommodate censoring. Bootstrap bias-corrected and accelerated (BCa) confidence intervals (Efron, 1987) can be constructed for \(\theta\) that account for the uncertainty in estimating \(\pi(\cdot)\).

### 4. Censoring

To accommodate censoring, let \(\tilde{C}\) denote the censoring time defined from the most recent event (LMP) to the censoring date. Under the sampling plan detailed in Section 2.1, length-bias operates on the enrollment cycle length and also on the enrollment cycle censoring time, when defined from LMP. Consequently, length-bias operates on the available length: \(\bar{Y} \wedge \tilde{C}\). For the enrollment cycle, we eliminate the “−” notation and write \(C = B + V\), where \(V\) is the residual time from enrollment date to censoring date, and write \(X = Y \wedge C\) as the length from LMP to the next menstrual period or censoring, with event indicator \(\delta = I(Y \wedge C)\). As Asgharian and others (2002), Huang and Qin (2011), and Wang (1991) among others noted, \(B\) is the initial segment of both \(Y\) and \(C\), so in general \(Y\) and \(C\) are dependent. We assume the standard relations for left truncated, right-censored random variables, that \([A \perp V \mid B]\) (conditional independence) and the conditional distribution of \(V\) given \(B\) is non-informative regarding \(f\), and so \(f_{V \mid B}\) does not depend on \(\theta\).

As before, recruitment continues until \(n\) couples are enrolled, with enrollment data \((b_i, x_i, \delta_i); i = 1, \ldots, n\). Under the conditional independence and non-informative assumptions, the conditional
likelihood of $X$ given $B$ is (leaving out $\theta$ on the right-hand side)

$$
\mathcal{L}_c(\theta \mid \mathbf{b}, \lambda, \mathbf{x}, \delta) \propto \prod_{i=1}^{n} \frac{f(x_i)\lambda(x_i)(1/x_i)^{\delta}}{\int_{b_i}^{\infty} f(u)\lambda(u)(1/u) \, du} \left\{ \int_{u_i}^{\infty} f(u)\lambda(u)(1/u) \, du \right\}^{(1-\delta)}.
$$

The marginal likelihood retains the form from (3.3), and so the full likelihood is

$$
\mathcal{L}(\theta \mid \pi, \mathbf{b}, \lambda, \mathbf{x}, \delta) \propto \prod_{i=1}^{n} \frac{f(x_i)\lambda(x_i)(1/x_i)^{\delta}}{\int_{b_i}^{\infty} f(u)\lambda(u)(1/u) \, du} \left\{ \int_{u_i}^{\infty} f(u)\lambda(u)(1/u) \, du \right\}^{(1-\delta)}.
$$

4.1 Semi-parametric estimation of $S(y \mid \theta)$ in the censored case

We find the MLE $\hat{\theta}$ and use $S(y \mid \hat{\theta}) = \int_{y}^{\infty} f(u; \hat{\theta}) \, du$, $y > 0$. For known $\pi(\cdot)$, maximum likelihood methods can be used to estimate $\theta$, $S(\cdot \mid \theta)$ and other functions of it, and the likelihood ratio can be used to calculate confidence regions. However, $\pi(\cdot)$ is usually not known and a recursive algorithm similar to that in Section 3.1 is necessary. Starting with an estimate of $g_B(b)$, proceed as follows: (1) let $\hat{\theta}^{(0)}$ be the MLE using the conditional likelihood in (4.1) so, $\hat{S}^{(0)}(y) = S(y \mid \hat{\theta}^{(0)}) = \int_{y}^{\infty} f(u \mid \hat{\theta}^{(0)}) \, du$; (2) estimate the shape of $\pi(b)$ by $\hat{\pi}^{(1)}(b) \propto \hat{g}_B(b)/S(b \mid \hat{\theta}^{(0)})$; (3) let $\hat{\theta}^{(i)}$ be the MLE using the full likelihood in (4.2) with $\pi(\cdot) = \hat{\pi}^{(i)}(\cdot)$; (4) update $\hat{\pi}^{(i+1)}(b) \propto \hat{g}_B(b)/S(b \mid \hat{\theta}^{(i)})$; and (5) iterate between steps (3) and (4) until convergence criterion is met. Use bootstrap BC$_{a}$ confidence intervals (Efron, 1987) to account for uncertainty in estimating $\pi(\cdot)$.

5. Incorporating a random effect and a covariate

In many applications, it is desirable to account for observed and unobserved residual heterogeneity; therefore, we propose the following model for the observed menstrual cycle length for couple $i$:

$$
\tilde{Y}_i = \tilde{Y}_{0i} W_i e^{\beta z_i},
$$

where $W_i$ is an unobserved couple-specific random effect with support $w_i \in (0, \infty)$ and CDF $H(w_i; \alpha)$, $\alpha \in \mathcal{A}$; $z_i$ is an observed time-independent covariate; $\beta$ is a fixed effect parameter; and $\tilde{Y}_{0i}$ is the baseline cycle length for a couple with $W_i = 1$ and $z_i = 0$ with CDF $F_0(y_{0i}; \theta)$, $\theta \in \Theta$. We assume the distribution of $Z$ does not depend on $\theta$ and that conditional on $Z_i = z_i$ and $W_i = w_i$, the cycle lengths within a woman are independent and identically distributed with CDF $F(y_i \mid w_i, z_i; \theta, \beta) = F_0(y_i w_i^{-\beta} e^{-\beta z_i}; \theta, \beta)$ and PDF $f(y_i \mid w_i, z_i; \theta, \beta) = w_i^{-\beta} f_0(y_i w_i^{-1} e^{-\beta z_i}; \theta, \beta)$.

Our goal is to use the enrollment cycle data to estimate the marginal distribution of menstrual cycle length for the reference population with covariate $Z = z$. For the enrollment cycle, we eliminate the “-” notation and assume as in Section 4 that length-bias operates on $X_i = Y_i \wedge C_i$ with $Y_i$ modeled as in (5.1). Since the random effects are not observed, we assume $\pi(\cdot)$ does not depend on $W_i$, and further we restrict to the case where $\pi(\cdot)$ does not explicitly depend on $z_i$. Incorporating the covariate, the observed enrollment data consist of $(b_i, x_i, \delta_i, z_i); i = 1, \ldots, n$. We assume non-informative and independent censoring conditional on $Z_i = z_i$ and $W_i = w_i$. Under these assumptions, the observed data conditional likelihood of $X$ given $(B, Z)$ is

$$
\mathcal{L}_c(\theta, \alpha, \beta \mid \mathbf{b}, \mathbf{z}, \lambda, \mathbf{x}, \delta) \propto \prod_{i=1}^{n} \frac{f(x_i \mid z_i; \theta, \alpha, \beta)\lambda(x_i)(1/x_i)^{\delta}}{\int_{b_i}^{\infty} f(u \mid z_i; \theta, \alpha, \beta)\lambda(u)(1/u) \, du} \left\{ \int_{x_i}^{\infty} f(u \mid z_i; \theta, \alpha, \beta)\lambda(u)(1/u) \, du \right\}^{(1-\delta)}.
$$
Here, $f(\cdot \mid z_i; \theta, \alpha, \beta) = \int_W f(\cdot \mid w, z_i; \theta, \beta) \, dH(w; \alpha)$, where we use Gaussian quadrature integration to integrate over the unobserved random effect; however, one could alternatively derive the complete data likelihood and use an Expectation Maximization approach. Extending the full likelihood in (4.2) to incorporate the covariate and random effect, we have

$$\mathcal{L}(\theta, \alpha, \beta \mid \pi, \lambda, b, x, z, \delta) \propto \prod_{i=1}^{n} \frac{\{ f(x_i | z_i; \theta, \alpha, \beta) \} \{ \int_{y}^{\infty} f(u | z_i; \theta, \alpha, \beta) \lambda(u)(1/u) \, du \}^{1-\delta_i}}{\int_{0}^{\infty} f(u | z_i; \theta, \alpha, \beta) \lambda(u)(1/u) \{ \int_{0}^{u} \pi(t | z_i) \, dt \} \, du}. \quad (5.3)$$

As before, note that $\mathcal{L}_c$ in (5.2) does not depend on $\pi(\cdot)$. Therefore, one could estimate $(\theta, \alpha, \beta)$ using a maximum likelihood approach based on it, and we compare the estimates and standard errors obtained using $\mathcal{L}_c$ to those based on $\mathcal{L}$ in Sections 6 and 7.

5.1 Semi-parametric estimation of $S(y | z; \theta, \alpha, \beta)$

To extend the semi-parametric approach to incorporate a covariate and random effect, we assume a parametric distribution for $\tilde{Y}_0$; but as before, we do not make any parametric assumptions about the form of $\pi(\cdot)$. Starting with an estimate of $g_B(b \mid z)$ (as shown in Section 6, the empirical probability mass function works well), proceed as follows:

1. Let $(\hat{\theta}^{(0)}, \hat{\alpha}^{(0)}, \hat{\beta}^{(0)})$ be the MLE using the conditional likelihood in (5.2) so,

$$S^{(0)}(y | z; \hat{\theta}^{(0)}, \hat{\alpha}^{(0)}, \hat{\beta}^{(0)}) = \int_{y}^{\infty} f(u | z; \hat{\theta}^{(0)}, \hat{\alpha}^{(0)}, \hat{\beta}^{(0)}) \, du.$$

2. 

$$\hat{\pi}^{(1)}(b | z) \propto \frac{\hat{g}_{B \mid Z}(b | z)}{S^{(0)}(y | z; \hat{\theta}^{(0)}, \hat{\alpha}^{(0)}, \hat{\beta}^{(0)})}.$$

3. Let $(\hat{\theta}^{(v)}, \hat{\alpha}^{(v)}, \hat{\beta}^{(v)})$ be the MLE using the full likelihood in (5.3) with $\pi(\cdot \mid z) = \hat{\pi}^{(v)}(\cdot \mid z)$.

4. 

$$\hat{\pi}^{(v+1)}(b | z) \propto \frac{\hat{g}_{B \mid Z}(b | z)}{S^{(v)}(y | z; \hat{\theta}^{(v)}, \hat{\alpha}^{(v)}, \hat{\beta}^{(v)})}.$$

5. Iterate between steps (3) and (4) until convergence criterion is met.

In Sections 6 and 7, we consider the case where $Z$ is a binary covariate, which is useful when comparing the distribution of menstrual cycle length for two populations.

6. Simulation studies

We describe and report a simulation study examining the semi-parametric estimation approach described in Section 5 in the setting in which we model the observed menstrual cycle length dependent on an unobserved random effect $W$ and an observed time-independent binary covariate $z$ (see model (5.1)). Our primary interest is in the estimation of the marginal means and standard deviations: $E(\tilde{Y} \mid z = 0)$, $E(\tilde{Y} \mid z = 1)$, $\text{SD}(\tilde{Y} \mid z = 0)$, and $\text{SD}(\tilde{Y} \mid z = 1)$, which are functions of the parameters $(\theta, \alpha, \beta)$. We evaluate performance of maximum likelihood estimators when proper account is taken of length-bias and selection effects. We focus on censored data using the full likelihood in (5.3).

Using the data generation scheme described in Section 2 of supplementary material available at Biostatistics online, we generate the simulated data consisting of 2000 samples, each with 500 participants enrolled using quota sampling, a sampling scheme in which participants are enrolled until a quota is
Table 1. Performance of semi-parametric algorithm for estimating \( S(y \mid z; \theta, \alpha, \beta) \) using the full likelihood under the following model assumptions: \( \tilde{Y} = \tilde{Y}_0 W e^{b z}, z \) an observed binary covariate, \( W \sim \text{Gamma}(\alpha, 1/\alpha), \tilde{Y}_0 \sim q f_{\text{normal}}(\mu_N, \sigma^2_N) + (1 - q) f_{\text{Gumbel}}(\mu_G, \sigma^2_G), \) and \( \lambda(y) \propto y \)

<table>
<thead>
<tr>
<th>Setting</th>
<th>Parameter</th>
<th>True</th>
<th>Bias</th>
<th>Sim SE</th>
<th>MSE</th>
<th>BC&lt;sub&gt;a&lt;/sub&gt;</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Const ( \pi(b) ) (( n = 250 ))</td>
<td>( E(Y \mid z = 0) )</td>
<td>31.95</td>
<td>0.03</td>
<td>0.81</td>
<td>0.66</td>
<td>92.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( E(Y \mid z = 1) )</td>
<td>30.39</td>
<td>-0.04</td>
<td>0.83</td>
<td>0.70</td>
<td>93.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(Y \mid z = 0) )</td>
<td>9.70</td>
<td>0.02</td>
<td>1.08</td>
<td>1.17</td>
<td>94.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(Y \mid z = 1) )</td>
<td>9.23</td>
<td>0.00</td>
<td>1.04</td>
<td>1.08</td>
<td>94.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
<td>93.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(W) )</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>78.70</td>
<td></td>
</tr>
<tr>
<td>Const ( \pi(b) ) (( n = 500 ))</td>
<td>( E(Y \mid z = 0) )</td>
<td>31.95</td>
<td>0.01</td>
<td>0.55</td>
<td>0.31</td>
<td>94.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( E(Y \mid z = 1) )</td>
<td>30.39</td>
<td>-0.04</td>
<td>0.56</td>
<td>0.31</td>
<td>91.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(Y \mid z = 0) )</td>
<td>9.70</td>
<td>0.02</td>
<td>0.72</td>
<td>0.52</td>
<td>94.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(Y \mid z = 1) )</td>
<td>9.23</td>
<td>0.00</td>
<td>0.69</td>
<td>0.48</td>
<td>94.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>90.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(W) )</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>75.35</td>
<td></td>
</tr>
<tr>
<td>Decr ( \pi(b) ) (( n = 250 ))</td>
<td>( E(Y \mid z = 0) )</td>
<td>31.95</td>
<td>0.04</td>
<td>0.76</td>
<td>0.58</td>
<td>94.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( E(Y \mid z = 1) )</td>
<td>30.39</td>
<td>-0.03</td>
<td>0.76</td>
<td>0.58</td>
<td>94.90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(Y \mid z = 0) )</td>
<td>9.70</td>
<td>0.02</td>
<td>1.10</td>
<td>1.21</td>
<td>95.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(Y \mid z = 1) )</td>
<td>9.23</td>
<td>0.00</td>
<td>1.05</td>
<td>1.11</td>
<td>95.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>93.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(W) )</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>81.15</td>
<td></td>
</tr>
<tr>
<td>Decr ( \pi(b) ) (( n = 500 ))</td>
<td>( E(Y \mid z = 0) )</td>
<td>31.95</td>
<td>0.01</td>
<td>0.53</td>
<td>0.29</td>
<td>95.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( E(Y \mid z = 1) )</td>
<td>30.39</td>
<td>-0.03</td>
<td>0.52</td>
<td>0.27</td>
<td>92.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(Y \mid z = 0) )</td>
<td>9.70</td>
<td>0.01</td>
<td>0.77</td>
<td>0.59</td>
<td>95.40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(Y \mid z = 1) )</td>
<td>9.23</td>
<td>0.00</td>
<td>0.73</td>
<td>0.54</td>
<td>95.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>92.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \text{SD}(W) )</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>73.45</td>
<td></td>
</tr>
</tbody>
</table>

Data generated using algorithm enumerated in Section 2 of supplementary material available at Biostatistics online with two cases for \( \pi(b) \): constant (\( \xi = 0 \)) and decreasing (\( \xi = 5.5 \)) with \( b \). Number of simulations = 2000. Standard errors based on simulation estimates (Sim SE). Bootstrap BC<sub>a</sub> CI constructed using 1000 bootstrap replicates. Two sample sizes: 250 and 500.

reached. The first 250 participants enrolled of each sample compose a small sample size case. In the data generation scheme, the choice of parametric family and the values of the parameters mimic the LIFE data analysis in Section 7.

We study the performance of the proposed recursive semi-parametric algorithm for empirical estimation of \( \pi(b \mid z) \) and parametric estimation of the distribution of \( \tilde{Y} \), under model (5.1) in which \( \tilde{Y}_0 \) is assumed to be a normal-Gumbel mixture and \( W \sim \text{Gamma}(\alpha, 1/\alpha) \). In particular, we are interested in two data generating models for \( \pi \), constant as found in the data analysis, and decreasing with \( b \). The bias, standard errors, mean squared error, and coverage probabilities for the estimators of \( E(\tilde{Y} \mid z = 0) \), \( E(\tilde{Y} \mid z = 1) \), \( \text{SD}(\tilde{Y} \mid z = 0) \), \( \text{SD}(\tilde{Y} \mid z = 1) \), \( \beta \), and \( \text{SD}(W) \) are summarized in Table 1. For both cases of \( \pi \), the bias of each of the marginal parameters is approximately 0. To account for the uncertainty in estimating both \( \pi \) and the model parameters, we calculated bootstrap BC<sub>a</sub> intervals with 1000 bootstrap replicates. The coverage probability of the nuisance parameter, \( \text{SD}(W) \), is less than the nominal 95%; however, the coverage probability for each of the marginal parameters of interest is very good even for small samples. Using the same simulated data, we also studied the performance of maximum likelihood estimators based on the conditional likelihood, \( \mathcal{L}_c \) in (5.2) which does not require estimation of \( \pi \) (see Table 2). Under both
Table 2. Performance of estimators of $S(y \mid z; \theta, \alpha, \beta)$ using the conditional likelihood under the following model assumptions: \( \tilde{Y} = \tilde{Y}_0 W e^{\beta z} \), \( z \) an observed binary covariate, \( W \sim \text{Gamma}(\alpha, 1/\alpha) \), \( \tilde{Y}_0 \sim qf_{\text{normal}}(\mu_N, \sigma^2_N) + (1 - q) f_{\text{Gumbel}}(\mu_G, \sigma^2_G) \), and \( \lambda(y) \propto y \)

<table>
<thead>
<tr>
<th>Setting</th>
<th>Parameter</th>
<th>True</th>
<th>Bias</th>
<th>Sim SE</th>
<th>MSE</th>
<th>BCa</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Const ( \pi(b) ) (( n = 250 ))</td>
<td>( E(Y \mid z = 0) )</td>
<td>31.95</td>
<td>0.04</td>
<td>0.81</td>
<td>0.66</td>
<td>92.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( E(Y \mid z = 1) )</td>
<td>30.39</td>
<td>0.01</td>
<td>0.81</td>
<td>0.66</td>
<td>93.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (Y \mid z = 0) )</td>
<td>9.70</td>
<td>0.02</td>
<td>1.08</td>
<td>1.17</td>
<td>94.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (Y \mid z = 1) )</td>
<td>9.23</td>
<td>0.02</td>
<td>1.04</td>
<td>1.08</td>
<td>94.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
<td>95.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (W) )</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>78.95</td>
<td></td>
</tr>
<tr>
<td>Const ( \pi(b) ) (( n = 500 ))</td>
<td>( E(Y \mid z = 0) )</td>
<td>31.95</td>
<td>0.01</td>
<td>0.55</td>
<td>0.31</td>
<td>95.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( E(Y \mid z = 1) )</td>
<td>30.39</td>
<td>-0.01</td>
<td>0.54</td>
<td>0.29</td>
<td>95.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (Y \mid z = 0) )</td>
<td>9.70</td>
<td>0.02</td>
<td>0.72</td>
<td>0.52</td>
<td>94.70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (Y \mid z = 1) )</td>
<td>9.23</td>
<td>0.01</td>
<td>0.69</td>
<td>0.47</td>
<td>95.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>96.20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (W) )</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>72.25</td>
<td></td>
</tr>
<tr>
<td>Decr ( \pi(b) ) (( n = 250 ))</td>
<td>( E(Y \mid z = 0) )</td>
<td>31.95</td>
<td>0.04</td>
<td>0.76</td>
<td>0.58</td>
<td>94.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( E(Y \mid z = 1) )</td>
<td>30.39</td>
<td>-0.01</td>
<td>0.76</td>
<td>0.58</td>
<td>95.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (Y \mid z = 0) )</td>
<td>9.70</td>
<td>0.02</td>
<td>1.10</td>
<td>1.21</td>
<td>95.80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (Y \mid z = 1) )</td>
<td>9.23</td>
<td>0.00</td>
<td>1.06</td>
<td>1.12</td>
<td>95.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>95.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (W) )</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>81.05</td>
<td></td>
</tr>
<tr>
<td>Decr ( \pi(b) ) (( n = 500 ))</td>
<td>( E(Y \mid z = 0) )</td>
<td>31.95</td>
<td>0.01</td>
<td>0.53</td>
<td>0.29</td>
<td>95.65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( E(Y \mid z = 1) )</td>
<td>30.39</td>
<td>-0.01</td>
<td>0.52</td>
<td>0.27</td>
<td>95.85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (Y \mid z = 0) )</td>
<td>9.70</td>
<td>0.01</td>
<td>0.77</td>
<td>0.59</td>
<td>95.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (Y \mid z = 1) )</td>
<td>9.23</td>
<td>0.00</td>
<td>0.73</td>
<td>0.54</td>
<td>95.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta )</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>96.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD( (W) )</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>72.25</td>
<td></td>
</tr>
</tbody>
</table>

Data generated using algorithm enumerated in Section 2 of supplementary material available at *Biostatistics* online with two cases for \( \pi(b) \): constant \( (\xi = 0) \) and decreasing \( (\xi = 5.5) \) with \( b \). Number of simulations = 2000. Standard errors based on simulation estimates (Sim SE). Bootstrap BCa CI constructed using 1000 bootstrap replicates. Two sample sizes: 250 and 500.

generating models for \( \pi \) and for each sample size, we did not observe a notable difference between the two approaches in any of the performance measures.

7. Application to the LIFE Study

We apply the proposed semi-parametric approach to estimate the enrollment probability function and the distribution of menstrual cycle length for the population studied in the LIFE Study, focusing on enrollment cycle data with adjustment for length-bias and selection features of the sampling plan. For this prospective pregnancy study, recruitment took place in the U.S. in Michigan and Texas in 2005–2009 with minimal criteria for eligibility listed as follows: female ages 18–44 years and male ages 18+ years; in a committed relationship; ability to communicate in English or Spanish; menstrual cycle length between 21 and 42 days; no hormonal contraception injections during past year; and no sterilization procedures or physician diagnosed infertility. Prior to conception, 501 couples satisfying the eligibility criteria and interested in becoming pregnant were enrolled on a calendar date irrespective of the day of the
woman’s menstrual cycle and followed until pregnancy, exit from the study, or censored at the completion of 12 menstrual cycles at-risk for pregnancy. Under these inclusion criteria and protocol, the sampling plan is independent of the participants’ cycle lengths. For further details, see the complete study protocol (Buck Louis, Schisterman and others, 2011).

The enrollment menstrual cycle length was observed as the sum of two segments: the backward recurrence time, which was recalled on the enrollment date as the time since LMP; and the residual time, which was prospectively observed from enrollment until the next menstrual bleeding event determined from the woman’s daily recorded level of bleeding or spotting (0 = none to 4 = heavy), using an algorithm to distinguish menstrual bleeding from non-menstrual bleeding. The residual time was censored if the couple exited the study for one or more of the following reasons: no longer wishing to participate, relocation, diagnosis of sterility, and non-compliance with study protocol; or if pregnancy was detected using Clearblue® Easy pregnancy tests which detect levels of human chorionic gonadotropin over 25 mIU/mL. For couples becoming pregnant in the enrollment cycle, we would have ideally censored the residual time on the day on which the next menstrual bleeding is precluded by the absence of a decline in estrogen and progesterone. In the absence of knowledge of this date, we censored on the date of the first positive pregnancy date.

We consider the enrollment cycles of 467 couples for which both the backward recurrence time and residual time (or censoring time) could be determined, with total observed length in the range of (8, 90) days and backward recurrence time less than 50 days. Since a menstrual cycle length of 90 days is more than twice the study’s inclusion criteria of 42 days, we have excluded lengths outside this range as they are likely an artifact of a missing cycle stop/start. Furthermore, we have excluded 4 women because the onset date of the enrollment menstrual cycle could not be determined due to breastfeeding, miscarriage, or discontinuation of hormonal birth control. This restriction excludes women who recently were pregnant which may be important if cycle length is associated with pregnancy, but this is a very small number of exclusions. We excluded 6 women who exited the study before recording any bleeding information due to non-compliance or no longer wishing to participate. This exclusion is based on the assumption that these exits were not associated with menstrual cycle length.

We estimate the marginal distribution of menstrual cycle length using model (5.1) for the observed menstrual cycle length, with \( z = 0 \) for the women aged < 30 years and \( z = 1 \) for the women aged \( \geq 30 \) years, where 30 is the median age; and we assume \( W \sim \text{Gamma}(\alpha, 1/\alpha) \) so \( E(W) = 1 \) and \( \text{Var}(W) = 1/\alpha \). Motivated by the statistical model for menstrual cycle length implemented by McLain and others (2012), we assume that \( Y_0 \) has a normal-Gumbel mixture distribution with unknown parameter vector: \( \theta = (q, \mu_N, \sigma_N, \mu_G, \sigma_G) \), where \( q \) is the proportion of normally distributed lengths and \( (\mu_N, \sigma_N) \) and \( (\mu_G, \sigma_G) \) are the mean and standard deviation of the normal and Gumbel distributions, respectively. As described in Section 3 of supplementary material available at Biostatistics online, we find this parametric choice for \( Y_0 \) fits the menstrual cycle length data better than the log-normal model. We assume \( \lambda(y) \propto y \) and provide a check of this assumption as suggested by a reviewer in Section 4 of supplementary material available at Biostatistics online. Rather than assuming a model for \( \pi(\cdot \mid z) \) in the full likelihood (see (5.3)), we use the semi-parametric approach proposed in Section 5.1 to estimate the shape of \( \pi(\cdot \mid z) \) and \( (\theta, \beta, \alpha) \).

In order to illustrate the effects of adjusting for the sampling plan, we show the marginal survival curves (Figure 1(a)) for women aged < 30 years, point estimates, and 95% confidence intervals (Table 3) with and without adjustment for length-bias and selection effects. The estimates of the unadjusted marginal survivor function and marginal means and standard deviations are located above those which have been adjusted, due to the over-representation of long cycles. The marginal survivor function that adjusts for both length-bias and selection effects lies in between the other curves, suggesting some cancelation of the length-bias by the selection effects. The same pattern was observed for the women aged \( \geq 30 \) years (not shown). The point estimate and confidence interval of the standard deviation of the random effect are small indicating that the residual variation is small. Of biological interest is that the estimated marginal survivor curve (see Figure 1(b)) and marginal mean and standard deviation for the women aged \( \geq 30 \) years
is below that of women aged < 30 years which agrees with the findings of Harlow and others (2000). If estimation of $\pi(\cdot \mid z)$ is not of interest, we can alternatively estimate the parameters by maximizing the censored form of the conditional likelihood, $L_c$, in (5.2). For these data, the point estimates and standard errors for the full and conditional approaches are approximately identical (not shown); however, this is not always the case as seen in the simulation results in Table 2. Additionally, in Section 5 of supplementary
Estimation accounting for length-bias and selection effects

Fig. 2. Smoothed shape of empirical estimate of $\pi(b \mid z)$ for LIFE Study for women aged < 30 years (a) and women aged $\geq 30$ years (b), scaled so that maximum value is 100%. Rug shows backward recurrence times. Note that a non-smoothed estimate of the shape of $\pi(b \mid z)$ was used to calculate the estimates shown in column 3 of Table 3. Restricted to $n = 467$ enrollment cycles of known observed length within (8, 90) days and with backward recurrence time less than 50 days.

material available at *Biostatistics* online, we compare the unadjusted and adjusted estimated distribution of $\tilde{Y}$ based on enrollment cycle data to the estimated distribution based on first post-enrollment cycle data.

Although the estimates shown in Figure 1 and Table 3 are based on an empirical estimate of $\pi(b \mid z)$, for the purpose of visualizing the enrollment probability function we display a smoothed version of the estimate for each age group in Figure 2 which was obtained using a kernel density smoothing approach with oversmoothed bandwidth selection (Wand and Jones, 1994), implemented in R via the function `bkde`. The magnitude of the estimate is scaled by its maximum so that the values on the $y$-axis are the percent of maximum value. For both groups, the shape of the smoothed estimate is approximately uniform for most values of $b$; therefore, we do not see evidence of selection effects with regard to the length of the backward recurrence time.

8. Discussion

We have considered the problem of estimating the distribution of menstrual cycle length from enrollment cycle data for which the distribution is potentially weighted by features of the sampling process, in particular length-bias and selection effects. Addressing these features in the likelihood, we developed a two-stage semi-parametric approach for estimating the population distribution along with the shape of the enrollment probability. Using the empirical probability mass function to estimate the distribution of backward recurrence times, we obtained approximately unbiased estimates of the parameters of the distribution and traded very little variance for estimating the shape of the enrollment probability.

In the analysis of the LIFE Study data, we showed that adjustment for both length-bias and selection effects leads to visible differences in parameter and survivor curves estimates. These differences may be especially important when making risk predictions for TTP or disease; or when studying covariate effects in studies where an effect size on the order of 1–3 days is not uncommon such as those investigating the effects of environmental pollutants on menstrual cycle length (see, e.g. Buck Louis, Rios and others, 2011).
This approach is particularly useful in settings where there is a desire to model the first gap time or an interest in the characteristics of the selection process. For example, one could compare the estimated enrollment probability across multiple studies or recruitment strategies and, if a significant difference is found, then account for selection effects to make inference on a reference population.

As is done in many studies involving a renewal process, the backward recurrence time was ascertained using self-report at enrollment. Since the inclusion criteria restricted to women with regular cycles, we expect the recall bias and uncertainty to be small. As is common with retrospective recalled information, we do observe some digit preference, in particular for an LMP of 7, 14, and 21 days before enrollment. In a more comprehensive analysis of these data, we could conduct a sensitivity analysis on digit preference generating distributions around those digits.

Focusing solely on enrollment cycle data, we were able to obtain relatively small confidence intervals for all parameters of the menstrual cycle length distribution; suggesting potential gain for inclusion of these data even when post-enrollment data are available. Our future work will focus on modeling the full history of menstrual cycle data from enrollment to pregnancy along with the time-to-pregnancy process. In modeling the pregnancy cycles for this work, we used time-to-positive test date as a proxy for the date on which the menstrual cycle is paused; however, the true date is at a time before the positive test date. Improvements can be made using data on intercourse patterns, ovulation, and pregnancy tests and by conditioning on the date for which the menstrual cycle is paused being earlier than the positive pregnancy test.

This work impacts the analysis of existing longitudinal data by providing an approach to estimate \( \pi(\cdot) \) for a particular study and to use this estimate to adjust inferences to a reference population. The estimate of \( \pi(\cdot) \) can also be useful in designing recruitment for a new study. Additionally, it might be of interest to collect data on the number of participants who decline enrollment in order to estimate the magnitude of \( \pi(\cdot) \).

**Supplementary material**

Supplementary material is available online at [http://biostatistics.oxfordjournals.org](http://biostatistics.oxfordjournals.org).

**Acknowledgments**

The authors thank Dr Germaine Buck Louis for providing us access to the LIFE Study. The authors acknowledge that this study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD ([http://biowulf.nih.gov](http://biowulf.nih.gov)). **Conflict of Interest:** None declared.

**Funding**

This work was supported by the Intramural Research Program of the U.S., National Institutes of Health, Eunice Kennedy Shriver National Institute of Child Health and Human Development [Contracts N01-HD-3-3355, N01-HD-3-3356, N01-HD-3-3358] and by the U.S., National Institutes of Health, National Institute of Environmental Health Sciences [Training Grant 2T32ES012871].

**References**

Estimation accounting for length-bias and selection effects


[Received July 30, 2013; revised March 6, 2014; accepted for publication June 10, 2014]