Published in final edited form as:
Stat Med. 2014 October 15; 33(23): 3961–3972. doi:10.1002/sim.6216.

Semiparametric modeling of grouped current duration data with preferential reporting

Alexander C. McLain, **Rajeshwari Sundaram**, **Marie Thoma**, and **Germaine M. Buck Louis**

Abstract

Current duration data arise in cross-sectional studies from questions on the length of time from an initiating event to the time of interview. For example in the National Survey on Family Growth, women who were considered at risk for pregnancy were asked (a) “Are you currently attempting pregnancy?” and (b) “If yes, how many months have you been attempting to get pregnant?” The responses to (b), referred to as the current durations, are length-biased because women with longer durations are more likely to answer yes to question (a) and therefore be included in the sample. Previous methods to analyze such data include continuous time nonparametric and parametric approaches. In this article, we propose a semiparametric Cox model and a piecewise constant baseline model (used to account for digit preference) to analyze grouped current duration data. We discuss and investigate through simulation studies, the robustness properties of the proposed methods when digit preference is present. Lastly, we present an analysis of the current duration data resulting from the 2002 National Survey on Family Growth.

Keywords

Backwards recurrence times; Current duration; Digit preference; Grouped survival data; Proportional hazards model

1. Introduction

Infertility is defined as the absence of pregnancy despite 12+ months of regular unprotected intercourse [1], and is estimated to affect 6% of married women in the United States [2] when using the National Survey on Family Growth (NSFG). However, recent evidence suggests that this construct-derived figure underestimates infertility prevalence by more than half. Specifically, the prevalence of infertility in the United States using the NSFG data was estimated to be 16% when querying women and 12% when querying men [3, 4]. These figures were estimated using the current duration based method proposed by Keiding *et al.* [5], and are more consistent with estimates based on incident data obtained from the few prospective cohort studies followed through 12 months of trying [6, 7, 8]. Underestimation of the prevalence of infertility lessons the perceived impacts of the condition, which includes social stigma in some societies [9], relationship stressors [10], and financial costs associated with treatments or adoption that are often not covered by health insurance [11].

Correspondence to: Alexander C. McLain.

The data we consider in this paper arise from two questions in the NSFG (a) "Are you currently attempting pregnancy?" and (b) "If yes, how many months have you been attempting to get pregnant?" The current duration approach [5, 12, 13] uses responses to (b), denoted by Y, to make inference on the unobserved total duration of pregnancy attempt, denoted by T. The Y values are an entirely right-censored sample from T, commonly referred to as backwards recurrence times [14]. Further, Y is only observed for those who are currently attempting pregnancy, which creates a length-bias because those with long pregnancy attempts are more likely to be in an attempt when surveyed. Statistical methods for analyzing current duration data have focused on continuous time parametric and nonparametric approaches. For example, Keiding *et al.* [5] developed nonparametric and parametric methods to estimate the survivor function of T from continuous Y values (see also [15]), and Keiding *et al.* [16] proposed the use of continuous time accelerated failure time (AFT) models to estimate exposures' effect on the distribution of T.

Methodological research on statistical methods to analyze TTP data have mostly focused on prospective [17, 18, 19], or retrospective [12, 20, 21] study designs. Unlike the retrospective and prospective study designs, a benefit of the current duration method is the ability to use a cross-sectional sample of reproductive aged women that are representative of the population. This sample can include 'non-planners' that can be missed in prospective studies, and couples that will never get pregnant who are missed in retrospectively reported TTP (see [13, 22] for a thorough review of the statistical issues in various TTP study designs).

In this article, we propose a semiparametric proportional hazards model for the total duration of pregnancy attempt. Specifically, we propose a semiparametric grouped backward recurrence Cox model to analyze current duration data. The model results in estimates of exposures effect on the distribution of T, and an estimate of the survivor function of T conditional on covariate values. Some AFT models have no straightforward method of assuring that survivor function of T can be estimated (discussed further in Section 2). This is troublesome since estimating the survivor function of T is a key goal in current duration analyses. Furthermore, parametric inference based on the AFT models may be restrictive, especially when little is known about the true distribution of T. By using a semiparametric framework, the proposed methods are flexible to the distribution of T and result in a proper estimate of the survivor function of T.

An additional challenge with survey-based responses is the presence of digit preference, as reflected in Figure 1 for the 2002 NSFG data. Digit preference is especially detrimental when interest lies in estimating the survivor function at a point of digit preference (i.e., at 6, 12, or 24 months). For example, in the NSFG we are interested in estimating the survivor function at 12 months (i.e., estimating the proportion of infertile couples). When digit preference is present, estimates of the survivor function at 12 months exhibit considerable bias (> 40% in our simulation studies). As a result, we propose a piecewise constant specification for the grouped baseline hazard to control for digit preference. We discuss knot selection of the piecewise model, and apply it to data with and without covariates.

Our methods are based on observed current durations that are integer valued, and allow for $Y = 0$ observations which are problematic with continuous time methods. For example, some parametric models are not defined at zero (e.g., the generalized gamma distribution), a constant must be added to Y to implement AFT models, and methods to estimate continuous nonparametric approaches cannot incorporate zeros or ties. In the NSFG data, Y represents the number of completed months of a pregnancy attempt. Thus, $Y = 0$ represents a woman who was surveyed before the completion of the first month of her pregnancy attempt. Approximately 30% of women are reported to conceive in the first menstrual cycle of trying [7], and the only way of including women who would get pregnant in their first menstrual cycle (which is approximately one month) is by allowing for $Y = 0$ observations. As a result, including such observations is critical to insure the sample is representative of the population.

The paper is organized as follows. In Section 2, we present preliminary theoretical results on backwards recurrence times, and discuss their implications on the current modeling approaches. In Sections 2.1 and 2.2, we propose a semiparametric backward recurrence proportional hazards model when the distribution of T is discrete and continuous, respectively. In Section 2.3, we discuss how to account for digit preference by assuming the baseline hazard is piecewise constant. The estimation of model parameters is discussed in Section 3, and we investigate the properties of our model through simulation studies in Section 4. In Section 5, we illustrate our methods using data from the 2002 NSFG. In the supporting information, we present a discrete nonparametric method for estimating the survivor function of the total duration of pregnancy attempt, additional simulation studies, and the `R` code to implement the proposed methods.

2. Methods

Current duration data consist of observations, Y, that indicate the length of the current pregnancy attempt. Our interest lies in estimating the distribution of the total duration of pregnancy attempt T. Here, $T = \min(X, U)$ where X denotes the couples' time-to-pregnancy (TTP) and U denotes the end of a pregnancy attempt without becoming pregnant (note that $E(T) < \infty$ since $U < \infty$). The unobserved T could arise from a continuous or discrete distribution. The issue of whether T is best regarded as continuous or discrete was discussed in depth by Keiding *et al.* [13]. As will be demonstrated in Section 2.2, our estimation methods are robust for discrete or continuous distributions on T.

As described previously [5, 13], the observed current duration of pregnancy attempt represents a length-biased sample of backward recurrence times [14]. To discuss the relationship between T and Y, we initially assume that both arise from either continuous or discrete distributions. Let g denote the density (or mass) function of Y. Under regularity conditions discussed below,

$$g(y)=\frac{\bar{F}(y)}{\mu_T} \quad (1)$$

[23] where $\bar{F}(t) = \Pr(T > t)$, $\mu_T = E(T)$ and g is non-increasing. One can use an estimate of g, denoted by $\hat{g}$, to estimate $\bar{F}$ via $\hat{\bar{F}}(t)=\hat{g}(t)/\hat{g}(0)$ where $\hat{\mu}_T = 1/\hat{g}(0)$. Let $\mathscr{P}$ be the space of all probability measures, and $\mathscr{F}$ the space of all survival functions (i.e., the space of non-

increasing positive functions with $\bar{F}(0) = 1$). Viewing the current duration operation as a mapping from $\mathscr{F} \to \mathscr{P}$, non-surjectivity will arise since there exists $g \in \mathscr{P}$ with no corresponding $\bar{F} \in \mathscr{F}$. Letting $\mathscr{P}_0$ denote the space of all non-increasing probability measures with $g(0) < \infty$, all $g \in \mathscr{P}_0$ have a corresponding $\bar{F} \in \mathscr{F}$. For $g \notin \mathscr{P}_0$ we can have either (a) $g(t)/g(0)$ is increasing for some t, or (b) $\bar{F}(0)$ is not defined (when $g(0)$ is not finite). Estimation procedures that restrict $\hat{g} \in \mathscr{P}_0$ (i.e., that map $\mathscr{F} \to \mathscr{P}_0$) are surjective, and will result in an $\hat{\bar{F}}$ which is non-increasing with $\hat{\bar{F}}(0)=1$. Conversely, an unrestricted estimation procedure is non-surjective and can result in $\hat{g} \notin \mathscr{P}_0$. A benefit of the proposed estimation procedure is that it is surjective and results in a valid estimate of $\bar{F}$. The AFT model does not use the $\hat{g} \in \mathscr{P}_0$ restriction, thus it is non-surjective and can result in $\hat{g}$ with no corresponding $\bar{F}$.

The relationship between the distributions of T and Y given in (1) assumes that the renewal process is in equilibrium with the renewal distribution, or that the process is in a ‘steady state’ (cf. [24]). Specifically, the steady state assumption assumes (i) the calendar times at which women are beginning their pregnancy attempts occur at a constant rate, and (ii) the distribution of T is independent of calendar time. Assumption (i) is a stationarity assumption and is satisfied if women are entering pregnancy attempts according to a homogeneous Poisson process. We also assume that (iii) the observations are independent. We discuss the validity of the assumptions in Section 6.

In Section 2.1, we propose a modeling procedure when Y and T are each discrete random variables. In Section 2.2, we consider the situation where Y is a grouped outcome and T is continuous.

2.1. Discrete backwards recurrence Cox model

In this section, we assume a discrete proportional hazards model for T, and we propose a model to estimate regression coefficients and the survivor function $\bar{F}$. Let T have discrete hazard probability $P(T = y|T \geq y, \boldsymbol{Z}) = 1 - \exp\{-\alpha_y \exp(\beta^\top \boldsymbol{Z})\}$, where $\alpha_y \geq 0$. The survivor function of T takes the form

$$\bar{F}(t|\boldsymbol{Z})=\exp\left\{-\exp(\boldsymbol{\beta}^\top\boldsymbol{Z})\sum_{j=0}^{t}\alpha_j\right\},$$

where $\alpha_0 \equiv 0$. This model has been used for the analysis of TTP data by Scheike and Jensen [20] and Sundaram *et al.* [19], and corresponds to the grouped version of the continuous time proportional hazards model [25]. The discrete current durations Y have probability mass function

$$g(y|\boldsymbol{Z})=g(0|\boldsymbol{Z})\exp\left\{-\exp(\boldsymbol{\beta}^\top\boldsymbol{Z})\sum_{j=0}^{y}\alpha_j\right\} \quad (2)$$

where $\alpha_j \geq 0$ for all j with $\alpha_0 \equiv 0$ and

$$g(0|\boldsymbol{Z})=\left[\sum_{y=0}^{\infty}\exp\left\{-\exp(\boldsymbol{\beta}^{\top}\boldsymbol{Z})\sum_{j=0}^{y}\alpha_j\right\}\right]^{-1}. \quad (3)$$

Notice $g(y|\boldsymbol{Z})$ is non-increasing in y with maximum value equal to $g(0|\boldsymbol{Z})$, where $g^{-1}(0|\boldsymbol{Z})=\sum_{y=0}^{\infty}\bar{F}(t|\boldsymbol{Z})=\mu_T$.

2.2. Grouped backwards recurrence Cox model

When T is continuous, the actual current durations, denoted by Y^*, have a continuous density equal to $\bar{F}(y)/\mu_T$. However, we observe the grouped continuous outcomes. We assume the continuous outcomes are grouped by rounding down, rounding upwards or to the closest value can be handled similarly. As a result, $Y = \lfloor Y^* \rfloor = \{y;\ Y^* \in [y, y + 1)\}$ has probability mass function

$$g(y)=\frac{1}{\mu_T}\int_{y}^{y+1}\bar{F}(u)du, \text{for} y=1,2,\ldots \quad (4)$$

We can apply the mean value theorem to (4) to get $g(y) = \bar{F}(y')/\mu_T$ for some $y' \in [y, y + 1]$. Similarly, there exists a $y'' \in [0, 1]$ such that $g(0) = \bar{F}(y'')/\mu_T$. As a result,

$$g(y)=\frac{g(0)\bar{F}(y')}{\bar{F}(y'')}=g(0)\exp[-\{\Lambda(y')-\Lambda(y'')\}] \quad (5)$$

for some $y' \in [y, y + 1]$ and $y'' \in [0, 1]$, where Λ denotes the cumulative hazard function of T.

To incorporate exposures into the model, we assume that T is distributed with proportional hazards form [26] with $\Lambda(t|\boldsymbol{Z}) = \exp(\beta^{\top}\boldsymbol{Z})\Lambda_0(t)$ and $\bar{F}(t|\boldsymbol{Z}) = \exp\{-\Lambda(t|\boldsymbol{Z})\}$, where $\boldsymbol{Z}$ a q-dimensional vector of exposures and β a q-dimensional vector of parameters. Under this model, (5) motivates modeling Y with

$$g(y|\boldsymbol{Z})=g(0|\boldsymbol{Z})\exp\{-H(y)\exp(\boldsymbol{\beta}^{\top}\boldsymbol{Z})\}, \quad (6)$$

where $H(y) = \Lambda_0(y') - \Lambda_0(y'')$ for some $y' \in [y, y + 1]$ and $y'' \in [0, 1]$ with $H(0) \equiv 0$. Notice that if we set $H(y)=\sum_{j=0}^{y}\alpha_j$, (2) and (6) are equivalent. As a result, we can use the method in Section 2.1 regardless of whether $\bar{F}$ is discrete or continuous to estimate of β. The form for g in (6) is an approximation of the true mass function of Y which is $g^*(y|\boldsymbol{Z})=\mu_T^{-1}(\boldsymbol{Z})\int_y^{y+1}\bar{F}(u|\boldsymbol{Z})du$ where $\mu_T^{-1}(\boldsymbol{Z})=g^*(0|\boldsymbol{Z})/\int_0^1\bar{F}(u|\boldsymbol{Z})du$. Directly modeling g^* would require knowledge of $\bar{F}(u|\boldsymbol{Z})$ over $[y, y + 1]$ for all y. When T follows an exponential distribution with $\Lambda_0(t) = \theta t$, we have $g^*(y|\boldsymbol{Z}) = g^*(0|\boldsymbol{Z}) \exp\{-H(y) \exp(\beta^{\top}\boldsymbol{Z})\}$ where $H(y) = \theta t$. As a result, (6) is equal to g^* and the survivor function of T is $\bar{F}(y|\boldsymbol{Z}) = \exp\{-H(y) \exp(\beta^{\top}\boldsymbol{Z})\}$. In the non-exponential setting, $\bar{F}(y|\boldsymbol{Z}) \approx \exp\{-H(y) \exp(\beta^{\top}\boldsymbol{Z})\}$. In our simulation studies, the estimates of $\bar{F}(y|\boldsymbol{Z})$ and β when $\Lambda_0(t) = \theta t^{\gamma}$ were relatively unbiased, suggesting that (6) was a close approximation to g^*.

2.3. Piecewise constant baseline model

Previous methods of analyzing data with digit preference include Ridout and Morgan [27], Pickering [28], Price and Seaman [29], and Bar and Lillard [30]. In many circumstances, the parameters governing the digit preference, such as the probability that someone reports $Y = 12$ when actually $Y = 10$, are nuisance parameters. In this case, digit preference can be controlled for by imposing smoothness restrictions on the probability mass function [28], i.e., imposing smoothness restrictions on $g(y|\boldsymbol{Z})$. This type of correction assumes that the rounding in the data is at random, and that people are equally likely to round up as they are to round down. Heitjan and Rubin [31] referred to this assumption as coarsening at random (CAR), for more on CAR see [32, 33].

The models in Sections 2.1 and 2.2 are smoothed by having the α_j's be constant over disjoint intervals. The piecewise constant model is implemented by creating a disjoint partition $(t_0, t_1]$, $(t_1, t_2]$, …, $(t_{L-1}, t_L]$ with $t_0 \equiv 0$ and $t_L \geq \max\{Y\}$. We then assume $\alpha_j = \gamma_l$ for all $j \in (t_{l-1}, t_l]$. Following the results in Section 2.1 and 2.2, the probability mass function of Y is

$$g_P(y|\boldsymbol{Z}){=}g_P(0|\boldsymbol{Z})\exp\left[-\exp(\boldsymbol{\beta}^\top \boldsymbol{Z})\sum_{\{j:t_{j-1}<y\}}\gamma_j\{(y\wedge t_j)-t_{j-i}\}\right], \quad (7)$$

where $x \wedge y = \min(x, y)$ and g_P $(0|\mathbf{Z})$ takes a form similar to (3). The piecewise estimate of survivor function of T is $\bar{F}_P$ $(y|\mathbf{Z}) = g_P$ $(y|\mathbf{Z})/g_P(0|\mathbf{Z})$.

Popular methods of knot selection, such as using the percentiles of the observed Y or information criterion, will not suffice because they will not impose the desired smoothness on g. To specify the t_l's, we consider the shape of a nonparametric estimate of the survivor function of T, given by $\hat{\bar{F}}_{NP}(y){=}\hat{g}_{NP}(y)/\hat{g}_{NP}(0)$. In Appendix A of the supporting information, we discuss the estimation of $\hat{g}_{NP}$ and $\hat{\bar{F}}NP$. In Figure 2, we present a histogram of data with digit preference, the corresponding estimate of $\hat{\bar{F}}_{NP}$ and the true $\bar{F}$.

When digit preference is present, Figure 2 demonstrates that $\hat{\bar{F}}_{NP}(y)$ has large jumps at 7 and 13 months and is flat over the interval [7, 12] months. Heuristically, this results in positive bias over [5, 6] and negative bias over [7, 8] (similarly for [11, 12] and [13, 14], respectively). Let $D_0 = \{d_1, d_2, \ldots\}$ denote the set of points of digit preference, which we assume are known, and let $D_k = \{d_1 + k, d_2 + k, \ldots\}$ be the set of points k units after the points of digit preference. In general, $\hat{\bar{F}}_{NP}(y)$ will have large jumps for $y \in D_1$, and be flat over $[d_{j-1} + 1, d_j]$ for $j > 1$, which results in $\hat{\bar{F}}_{NP}(y)$ having positive bias for $y \in \{D_{-1}, D_0\}$, and negative bias for $y \in \{D_1, D_2\}$.

To smooth the estimate of $\hat{\bar{F}}_{NP}$ we propose choosing the t_l's such that $t_l \notin \{D_{-1}, D_0, D_1, D_2\}$ for $l = 1, 2, \ldots, L$. Under these guidelines note that all $(t_{l-1}, t_l]$ such that $d_j \in (t_{l-1}, t_l]$ contain the set of points $\{d_j - 1, d_j, d_j + 1, d_j + 2\}$. As a result, the knots are chosen such that all intervals that contain a point of digit preference, include points with positive and negative bias, mainly $\{d_j - 1, d_j\}$ and $\{d_j + 1, d_j + 2\}$, respectively. This method acknowledges the difficulty in estimating the grouped baseline hazard around the d_j's, and balances the effect of digit preference. When the maximum observed Y is in D_0 we leave the last interval open ended by setting $t_L = \infty$. Our simulation study found that if $t_l \in \{D_{-1}, D_0, D_1, D_2\}$ for all l, the choice of knots had relatively little effect of the results.

3. Estimation

In this section, we discuss maximum likelihood estimation of the models proposed in Section 2. Let $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_n\}$, and $Y_{(1)}, Y_{(2)}, \ldots, Y_{(m)}$ denote the observed current durations, and the ordered and distinctly observed current durations, respectively. When there is no censoring (2) and (6) can be estimated by setting $\alpha_y = \infty$ for $y > Y_{(m)}$, thus $\sum_{y=0}^{Y_{(m)}} g(y|\boldsymbol{Z})=1$, and $\alpha_y = 0$ for all $y \notin \{Y_{(2)}, Y_{(3)}, \ldots, Y_{(m)}\}$. Notice that by setting $\alpha_y = 0$ for $y < Y_{(2)}$ we have $g(y|\boldsymbol{Z}) = g(0|\boldsymbol{Z})$ for all $0 \le y < Y_{(2)}$, which ensures that (3) is identifiable. We then estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha} = \{\alpha_{Y(2)}, \alpha_{Y(3)}, \ldots, \alpha_{Y(m)}\}$ using maximum likelihood.

It is common to censor all current durations greater than a fixed value, denoted by τ, after which they are not considered to be reliable. Let $\tilde{\boldsymbol{Y}} = \{\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_n\}$ denote the possibly censored current durations, where $\tilde{Y}_i = \min(Y_i, \tau)$ and $\delta_i = I(Y_i \le \tau)$. Let $\tilde{Y}_{(1)}, \tilde{Y}_{(2)}, \ldots, \tilde{Y}_{(m)} \le \tau$ denote the ordered and distinctly observed uncensored current durations, and $\bar{G}(y|\boldsymbol{Z})=1-\sum_{j=0}^{y} g(j|\boldsymbol{Z})$. When censoring is present we cannot set $\alpha_y = \infty$ for $y > \tilde{Y}_{(m)}$ because the likelihood for those censored at τ would be $\bar{G}(\tau|\boldsymbol{Z}) = 0$. To allow for $\bar{G}(\tau|\boldsymbol{Z}) > 0$ we introduce an additional parameter α_τ, and set $\alpha_y = \alpha_\tau$ for all $y > \tilde{Y}_{(m)}$.

To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_\tau = \{\alpha_{Y(2)}, \ldots, \alpha_{Y(m)}, \alpha_\tau\}$, we use maximum likelihood estimation. The likelihood corresponding to the observed $\tilde{\boldsymbol{Y}}$, $\boldsymbol{\delta} = \{\delta_1, \delta_2, \ldots, \delta_n\}$, and $\mathscr{Z} = \{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n\}$ takes the form $L(\boldsymbol{\beta}, \boldsymbol{\alpha}_\tau | \boldsymbol{Y}, \boldsymbol{\delta}, \mathscr{Z})=\prod_{i=1}^{n} g(\tilde{Y}_i|\boldsymbol{Z}_i)^{\delta_i} \bar{G}(\tilde{Y}_i|\boldsymbol{Z}_i)^{1-\delta_i}$ with maxima denoted by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}_\tau$. The survivor function of the unobserved T's at given value of $\boldsymbol{Z}$ is estimated by $\hat{\bar{F}}_{SP}(y|\boldsymbol{Z})=\exp\{-\exp(\hat{\beta}^\top \boldsymbol{Z})\sum_{j=0}^{y}\hat{\alpha}_j\}$ where $\hat{\alpha}_0 \equiv 0$.

For the piecewise constant model the details of the estimation proceed similarly. That is, for a given $\{t_0, t_1, \ldots, t_L\}$ we use g_P given in (7) and $\bar{G}_P(y|\boldsymbol{Z})=1-\sum_{j=0}^{y} g_P(j|\boldsymbol{Z})$ to form the likelihood. If there are censored values at $\tau \ge \tilde{Y}_{(m)}$, an additional parameter γ_τ is used, or set $t_L = \infty$, so that $\bar{G}_p(\tau|\boldsymbol{Z}) > 0$. After $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_L, \gamma_\tau\}$ have been estimated, the piecewise constant estimate of F at given value of $\boldsymbol{Z}$ is given by

$$\hat{\bar{F}}_{PC}(y|\boldsymbol{Z})=\exp[-\exp(\hat{\beta}^\top \boldsymbol{Z})\sum_{\{j:t_{j-1}<y\}}\hat{\gamma}_j\{(y\wedge t_j)-t_{j-1}\}].$$

The R programs [34] to implement the semiparametric and piecewise constant models to simulated data are contained in Appendix B of the supporting information. In our data analysis and simulation study, we estimated the standard error of $\hat{\boldsymbol{\beta}}$ with a numerical approximation to the hessian matrix. If T is continuous, $\exp(\hat{\beta}_j)$ corresponds to the estimated hazard ratio, associated with a one unit increase in Z_j. For discrete T, $\hat{\beta}_j$ can be interpreted as the approximate logarithms of subject-specific risk or odds ratios associated with a one unit increase in Z_j (cf. [35]).

4. Simulation Studies

To test the properties of our models with moderate sample sizes, numerous simulation studies were performed. The current duration for the *i*th subject was simulated by generating the unobserved total durations as $T_{ij} \sim F$ for $j = 1, 2, \ldots, K$, where $K = \min(K; \sum_{j=1}^{k} T_{ij} > M)$ and M is a fixed large integer, replicating a renewal process in equilibrium with renewal distribution (see [36] for details). For the continuous scenario discussed in Section 2.2, the backward recurrence times were grouped with $Y_i = \lfloor M - T_{iK-1} \rfloor$. For the discrete scenario discussed in Section 2.1, the continuous T_{ij} were grouped with $T_{ij}^{*} = \lceil T_{ij} \rceil$ and $Y_i = M - T_{iK^*-1}^{*}$ where $K^{*} = \min(k; \sum_{j=1}^{k} T_{ij}^{*} > M)$. Here, F had hazard function $\lambda(t|\mathbf{Z}_i) = \lambda_0(t) \exp(\beta^{\top} \mathbf{Z}_i)$, and $\mathbf{Z}_i = (Z_{i1}, Z_{i2})$ were independently generated as Bernoulli(0.5) and N(0, 0.5^2), respectively. The baseline hazard was set to $\lambda_0(t) = \theta \gamma t^{\gamma - 1}$ with $\theta = 0.3$ and $\alpha = 0.75$. Note that $\alpha_j \neq \alpha_{j+1}$ for all j, so the piecewise constant model is not correctly specified. In Section 4.1 we test the effect digit preference would have on the results, and in Section 4.2 we test continuous and discrete distributions on T. For brevity, the results for the piecewise model with a continuous F are not presented since they were similar to what was found with discrete F. In Section B of the supporting information, we present expanded simulation studies with piecewise model with a continuous F, an asymmetric rounding mechanism (where CAR is violated), and a comparison to the Weibull AFT model. We found that asymmetric rounding can result in bias in the survival function estimates. Further, the β estimates from the proposed model performed similarly to the Weibull AFT model.

4.1. Digit preference and knot selection

In this section, we explore the effect digit preference has on the proposed methods, and the knot selection for the piecewise model. The discrete scenario was used, and the data were randomly grouped as follows, if $4 \leq Y_i \leq 9$ then Y_i was rounded to 6 with probability 0.4, if $10 \leq Y_i \leq 18$ then Y_i was rounded to 12 with probability 0.6, if $Y_i > 18$ then Y_i was rounded to the nearest multiple of 12 with probability 0.8. The piecewise constant model was fit with four separate knot scenarios, each used 7 knots, the locations for knot scenarios A–C were $\{1, 2, 4, 9, 18, 30, \infty\}$, $\{1, 2, 4, 9, 15, 27, \infty\}$, and $\{1, 2, 4, 10, 17, 29, \infty\}$, respectively, while scenario E used the percentiles of the observed Y_i. For scenarios A–C note that the knots did not coincide with $\{D_{-1}, D_0, D_1, D_2\}$ as recommended in Section 2.3, and the majority of knots were close to 0. The last interval is open ended (i.e., $t_L = \infty$) since it is likely that $Y_{(m)} \in D_0$.

In Table 1, we present the results of the digit preference simulation for β_1, β_2, $\bar{F}(6|\mathbf{0})$, $\bar{F}(12|\mathbf{0})$, $\bar{F}(24|\mathbf{0})$ and the l_2-norm defined as $\| \bar{F} - \tilde{\bar{F}} \|_2 = \sum_{y=1}^{200} \{\bar{F}(y|\mathbf{0}) - \tilde{\bar{F}}(y|\mathbf{0})\}^2$ where $\tilde{\bar{F}}(t|\mathbf{0}) = N^{-1} \sum_{j=1}^{N} \hat{\bar{F}}_j(t|\mathbf{0})$, and $\hat{\bar{F}}_j(t|\mathbf{0})$ is the estimated survivor function for iteration j evaluated at $\mathbf{Z} = \mathbf{0}$. The estimates of β showed little bias. In general, the piecewise estimates of β had more bias than the semiparametric estimates. The average estimate of $\bar{F}$ for knot scenarios A–C were closer to the truth than the semiparametric model, or knot scenario E. This was especially true for the points of digit preference 6, 12, and 24. For these points the semiparametric model showed significant over estimation. Overall, the results from scenarios A–C with the piecewise model were the most robust to digit preference. Further, knot selection did not have a large effect on the estimates under the guidelines discussed in Section 2.3.

4.2. Continuous versus discrete F

In this section, we present the results of simulations designed to test differences when the underlying distribution for T is continuous versus discrete. In Table 2, we present the results from the piecewise constant (discrete F only) and semiparametric models for discrete and continuous F. The piecewise model used 7 knots with locations {1, 2, 5, 8, 11, 18, $Y_{(m)}$}.

The results from the piecewise constant and semiparametric models showed little bias for **β** with either discrete or continuous $\bar{F}$. The bias tended to decrease as the sample size increased. The empirical coverage probabilities for **β** = 0 suggest that hypothesis tests of β_j = 0 will have proper type I error. Further, the empirical coverage probabilities for nonzero β were close to the nominal 0.95 level. For the semiparametric model, properties of the **β** estimates were similar for the discrete and continuous scenario. The values of l_2 norm showed that the estimates of the survivor function accurately estimated the distribution for both scenarios. Notice that the l_2 norm for the continuous setting is similar to the l_2 norm for the discrete setting. This is noteworthy since g is modeled in the continuous setting using (6), which is an approximation to the true mass function of Y, while in the discrete setting g is modeled using the true mass function of Y. These results indicate that (6) is an accurate form for this distribution.

5. Data Analysis

We use a nationally representative cross-sectional sample of 7,643 US women aged 15–44 years from the 2002 NSFG to demonstrate the proposed methods. Details of the study design and survey have been described previously [37]. For this analysis, we included 270 eligible women aged 15–44 years (mean age 31 years) who reported the current duration of their pregnancy attempt (see [3] for further details). Initially we present unadjusted analyses of the data, and then we incorporate covariates and present a comparison to a Weibull AFT model. The validity of the current duration values decrease for longer durations [38], but can be considered reliable over shorter periods of time [39]. As a result, we censor current duration responses at a value after which they are not considered to be reliable. In this analysis, we censored all durations longer than 36 months (21.4% of the data).

To estimate the unadjusted distribution of T, we analyzed the NSFG data with the piecewise constant model without covariates $\hat{\bar{F}}_{PC}$. We compare $\hat{\bar{F}}_{PC}$ to a discrete nonparametric estimate, denoted by $\hat{\bar{F}}_{NP}$, which can handle zeros and ties. The computation and asymptotic properties of $\hat{\bar{F}}_{NP}$ are discussed in Appendix A of the supporting information. After censoring, the points of digit preference for this data are 6, 12, 24, and 36 months. For the piecewise constant model we tested various locations and sizes for the vector of knots and did not see any marked difference in the results. The analyses presented here used seven knots with locations {1, 2, 4, 9, 18, 30, ∞}. Confidence intervals were calculated using the percentiles of 270 bootstrap survivor function estimates.

The estimated piecewise constant and nonparametric survivor functions are given in Figure 3, along with 95% pointwise confidence intervals (CI). In Table 3, we present point estimates and confidence intervals for the probability that a pregnancy attempt is longer than 12 or 24 months. The nonparametric method found $\hat{\bar{F}}_{NP}(12)=0.336$ with 95% CI (0.212, 0.481), while the piecewise constant method found $\hat{\bar{F}}_{PC}(12)=0.223$ with 95% CI (0.148, 0.310). Here, $\hat{\bar{F}}_{PC}(12)$ is closer to historical values than $\hat{\bar{F}}_{NP}(12)$. The over estimation of $\hat{\bar{F}}_{NP}$ at the points of digit preference corroborates the results in Section 4.1.

To assess the association between exposures and the distribution of T, we used the proposed backwards recurrence Cox model (piecewise and semiparametric). The covariates used in the model were parity (z_1), the indicator of at least one live birth, and the woman's age minus 31 years (z_2). A woman is said to be parous if she has had a previous live birth ($z_1 = 1$), and nulliparous otherwise ($z_1 = 0$). In Table 3, we present the estimated β coefficients, and the estimated probability that $T > 12$ months for a 31-year old parous woman, denoted by $\bar{F}(12|1, 0)$, and a 31-year old nulliparous woman, denoted by $\bar{F}(12|0, 0)$. In Figure 4, we display the estimated survivor functions $\hat{\bar{F}}(\cdot|1,0)$ and $\hat{\bar{F}}(\cdot|0,0)$.

The point estimates for the effect of parity on T were consistent for both models. These estimates suggest that parous women become pregnant faster than nulliparous women. The effect of age was negative in both models, and achieved significance at the 0.05 level for the semiparametric model. The estimates for the prevalence of infertility corroborate the nonparametric results, showing that at the points of digit preference there appears to be over estimation of the survivor function for the semiparametric model. Further, the width of the 95% confidence intervals indicates greater precision for the piecewise model in the survivor function estimate. These results indicate that with 0.358 probability, a 31-year-old nulliparous women will have 12+ month pregnancy attempt.

We compared the above results to those obtained from the parametric AFT model proposed in [16]. To fit the AFT model we altered the data to $T^* = T + C$ where $C = 0.5$, and fit $\log(T^*) = -(\mu + \gamma\boldsymbol{Z}) + \in/\gamma$ where $\in$has an extreme value distribution. The estimates of the coefficients from the fitted model were in the same direction as those presented in Table 3, where $\hat{\gamma}_1 = 0.76$ and $\hat{\gamma}_2 = -0.06$ for parity and age, respectively, and both were significant at the 0.05 level. Further, the estimated shape was $\hat{\gamma} = 0.84$. Note that $\hat{\gamma} < 1$ and $\hat{g}(0|\boldsymbol{Z})$ is unbounded. Thus, using the notation in Section 2, $\hat{g} \notin \mathscr{P}_0$ and $\bar{F}$ cannot be estimated from this model (see [13] for further discussion on this issue). The results were similar for $C = 0.1$ and 1.

6. Discussion

The current duration approach is an evolving method that is gaining attention in light of global concerns about declining human fecundity accompanied by purported increases in the prevalence of infertility. In this paper, we have proposed semiparametric and piecewise constant backward recurrence Cox models to estimate the distribution of the total length of pregnancy attempt from the observed current length of pregnancy attempt. The use of the semiparametric framework is advantageous over the parametric AFT model, since little is known about the underlying distribution. As discussed in Section 2, applying the AFT model to current duration data is a non-surjective operation since estimates of $\hat{\bar{g}}$ can have no corresponding $\bar{F}$. The proposed methods result in valid estimates of $\bar{F}$, and can incorporate issues that will be encountered in practice, such as current durations equal to zero and digit preference. This gives the proposed methods practical relevance to those analyzing current duration data.

Our findings suggest that inattention to digit preference overestimates the percentage of women not achieving pregnancy at 12 and 24 months. Specifically, when controlling for digit preference we estimated that 21.7% of parous women aged 31 years will require a 12+ month attempt for pregnancy, when ignoring such reporting preferences this value was more than 50% larger at 33.6%. Our analyses also examined the effect that exposures have on the distribution of the total length of pregnancy attempt. Assuming that F is continuous, the semiparametric model found that the rate of the total length of pregnancy attempts was 1.64 times higher for parous women than it was for nulliparous women, and that an additional year of maternal age lowers this rate by 0.97. When the Weibull AFT model was used, the estimated coefficients resulted in a model where $\bar{F}$ could not be estimated. As a result, a different modeling approach would need to be implemented to estimate $\bar{F}$. The proposed modeling approach has assumed a proportional hazards form for T, which can not be verified by the observed data. One could, however, check if (2) shows signs of lack of fit based on the observed Y values. This would not guarantee that the proportional hazards assumptions holds, but it would verify if the data are consistent with the assumed structure. Validity and recall bias of current duration data has not been studied and if present could bias parameter estimates.

Current duration analyses provides inference on the total length of pregnancy attempt (T), which is the minimum of TTP (X) and the length of an unsuccessful pregnancy attempt (U). The fact that the outcome is total length of pregnancy attempt and not TTP needs to be taken into consideration when interpreting model estimates. For example, if we had found that older age was associated with shorter T it could be due to older age being related to shorter U, and thus not related to the outcome of interest X (we found older age being associated with longer T so this was not an issue). Development of statistical methods that can delineate between the competing factors that end pregnancy attempts is an area of future development.

The steady state assumptions given in Section 2 would be violated if there is a change in fecundity over time, or if there was a pattern to the initiations of the women’s pregnancy attempts. With regard to the former the research has been mixed [40, 41], and this assumption should be reasonable over the time period of interest (1999–2002). In a short time frame such as this (censored at 3 years), the variation in attempt times would mainly be due to seasonal differences in pregnancy attempts over a given year. Sensitivity analyses discussed in Slama *et al.* [42] showed some seasonal fluctuation in the distribution of starting dates for planned pregnancies; however, they found little difference in findings when accounting for this variation in their model. An area of future research is the development of methods to address the validity of the steady state assumptions.

The field of fecundity has greater relevancy in light of sociodemographic changes in childbearing for most developed countries, resulting in women or couples' expectation for more immediate pregnancy results. This underscores the relevancy of monitoring couple fecundity as measured by TTP [8] for estimating fecundity related impairments such as conception delay or infertility. The benefit of the current duration approach is that it can incorporate women who are not planning to become pregnant (missed in prospective cohorts), and women that have never been pregnant (missed in most retrospective studies). The difficulty with current duration data is the lack of methods available to estimate the distribution of the total length of pregnancy attempt (previously, only nonparametric and parametric methods are available), and to estimate the association exposures have with the total length of pregnancy attempt (previously, only parametric methods are available). The proposed proportional hazards methods are semiparametric. Further, digit preference, which is common in studies where recall is involved, has been addressed with the piecewise model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by the NICHD Intramural Research Program (Louis and Sundaram).

References

1. Practice Committee of the American Society for Reproductive Medicine. Definitions of infertility and recurrent pregnancy loss: a committee opinion. Fertility and Sterility. 2013; 99:63. [PubMed: 23095139]
2. Chandra A, Copen CE, Stephen EH. Infertility and impaired fecundity in the united states, 1982–2010: Data from the national survey of family growth. National health statistics reports. 2013; (67)
3. Thoma ME, McLain AC, Louis JF, King RB, Trumble AC, Sundaram R, Louis GMB. Prevalence of infertility in the united states as estimated by the current duration approach and a traditional constructed approach. Fertility and Sterility. 2013; 99:1324–1331. e1. [PubMed: 23290741]
4. Louis JF, Thoma ME, Srensen DN, McLain AC, King RB, Sundaram R, Keiding N, Buck Louis GM. The prevalence of couple infertility in the united states from a male perspective: evidence from a nationally representative sample. Andrology. 2013; 1:741–748. [PubMed: 23843214]
5. Keiding N, Kvist K, Hartvig H, Tvede M, Juul S. Estimating time to pregnancy from current durations in a cross-sectional sample. Biostatistics. 2002; 3:565–578. [PubMed: 12933598]
6. Tietze C. Fertility after discontinuation of intrauterine and oral contraception. International journal of fertility. 1968; 13:385. [PubMed: 5700397]
7. Zinaman MJ, Clegg E, Brown CC, O'connor J, Selevan S. Estimates of human fertility and pregnancy loss. Fertility and sterility. 1996; 65:503–509. [PubMed: 8774277]

8. Louis GMB, Sundaram R, Schisterman EF, Sweeney AM, Lynch CD, Gore-Langton RE, Chen Z, Kim S, Caldwell KL, Barr DB. Heavy metals and couple fecundity, the LIFE study. Chemosphere. 2012; 87:1201–1207. [PubMed: 22309709]
9. Bak CW, Seok HH, Song SH, Kim ES, Her YS, Yoon TK. Hormonal imbalances and psychological scars left behind in infertile men. Journal of Andrology. 2012; 33:181–189. [PubMed: 21546616]
10. Schmid J, Kirchengast S, Vytiska-Binstorfer E, Huber J. Infertility caused by pcoshealth-related quality of life among austrian and moslem immigrant women in austria. Human Reproduction. 2004; 19:2251–2257. [PubMed: 15333601]
11. Wu AK, Elliott P, Katz PP, Smith JF. Time costs of fertility care: the hidden hardship of building a family. Fertility and Sterility. 2013; 99:2025–2030. [PubMed: 23454007]
12. Weinberg CR, Gladen BC. The beta–geometric distribution applied to comparative fecundability studies. Biometrics. 1986; 42:547–560. [PubMed: 3567288]
13. Keiding N, Højbjerg Hansen OK, Sørensen DN, Slama R. The current duration approach to estimating time to pregnancy. Scandinavian Journal of Statistics. 2012; 39:185–204.
14. Allison PD. Survival analysis of backward recurrence times. Journal of the American Statistical Association. 1985; 80:315–322.
15. Ali MM, Marshall T, Babiker AG. Analysis of incomplete durations with application to contraceptive use. Journal of the Royal Statistical Society. Series A (Statistics in Society). 2001; 164:549–563.
16. Keiding N, Fine JP, Hansen OH, Slama R. Accelerated failure time regression for backward recurrence times and current durations. Statistics & Probability Letters. 2011; 81:724–729.
17. Dunson DB, Stanford JB. Bayesian inferences on predictors of conception probabilities. Biometrics. 2005; 61:126–133. [PubMed: 15737085]
18. Scarpa B, Dunson DB. Bayesian methods for searching for optimal rules for timing intercourse to achieve pregnancy. Statistics in Medicine. 2007; 26:1920–1936. [PubMed: 17328097]
19. Sundaram R, McLain AC, Buck Louis GM. A survival analysis approach to modeling human fecundity. Biostatistics. 2012; 13:4–17. [PubMed: 21697247]
20. Scheike TH, Jensen TK. A discrete survival model with random effects: an application to time to pregnancy. Biometrics. 1997; 53:318–329. [PubMed: 9147597]
21. Scheike TH, Petersen JH, Martinussen T. Retrospective ascertainment of recurrent events: An application to time to pregnancy. Journal of the American Statistical Association. 1999; 94(447): 713–725.
22. Scheike TH, Keiding N. Design and analysis of time-to-pregnancy. Stat. Methods Med. Res. 2006; 15:127–140. [PubMed: 16615653]
23. Cox, DR. Some sampling problems in technology. In: Johnson, NL.; Smith, H., editors. New Developments in Survey Sampling. New York: Wiley; 1969. p. 506-527.
24. Mandel M. The competing risks illnessdeath model under cross-sectional sampling. Biostatistics. 2010; 11:290–303. [PubMed: 19933879]
25. Fahrmeir, L.; Tutz, G. Multivariate statistical modelling based on generalized linear models. Second edn.. New York: Springer Series in Statistics, Springer-Verlag; 2001.
26. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society, Series B. 1972; 34:187–220.
27. Ridout M, Morgan B. Modelling digit preference in fecundability studies. Biometrics. 1991; 47:1423–1433. [PubMed: 1786326]
28. Pickering R. Digit preference in estimated gestational age. Statistics in medicine. 1992; 11:1225–1238. [PubMed: 1509222]
29. Price KL, Seaman JW. Bayesian modeling of retrospective time-to-pregnancy data with digit preference bias. Mathematical and Computer Modelling. 2006; 43:1424–1433.
30. Bar HY, Lillard DR. Accounting for heaping in retrospectively reported event data a mixture-model approach. Statistics in Medicine. 2012; 31:3347–3365. [PubMed: 22733577]
31. Heitjan DF, Rubin DB. Ignorability and coarse data. The Annals of Statistics. 1991; 19(4):2244–2253. 12.

32. Gill, R.; Laan, M.; Robins, J. Coarsening at random: Characterizations, conjectures, counter-examples. In: Lin, D.; Fleming, T., editors. Proceedings of the First Seattle Symposium in Biostatistics, Lecture Notes in Statistics. Vol. 123. Springer; US: 1997. p. 255-294.

33. Gill, R.; Robins, J. Sequential models for coarsening and missingness. In: Lin, D.; Fleming, T., editors. Proceedings of the First Seattle Symposium in Biostatistics, Lecture Notes in Statistics. Vol. 123. US: Springer; 1997. p. 295-305.

34. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.

35. Ecochard R, Clayton DG. Multivariate parametric random effect regression models for fecundability studies. Biometrics. 2000; 56:1023–1029. [PubMed: 11129457]

36. Feller, W. An introduction to probability theory and its applications. Vol. II. New York: John Wiley & Sons Inc.; 1966.

37. Lepkowski JM, Mosher WD, Davis KE, Groves RM, van Hoewyk J, Willem J. National survey of family growth, cycle 6: Sample design, weighting, imputation, and variance estimation. National health statistics reports. 2006; (142)

38. Cooney MA, Louis GMB, Sundaram R, McGuiness BM, Lynch CD. Validity of self-reported time to pregnancy. Epidemiology. 2009; 20(1):56–59. [PubMed: 19057382]

39. Zielhuis GA, Hulscher MEJL, Florack EIM. Validity and reliability of a questionnaire on fecundability. International Journal of Epidemiology. 1992; 21(6):1151–1156. [PubMed: 1483821]

40. Joffe M. Time trends in biological fertility in britain. The Lancet. 2000; 355(9219):1961–1965.

41. Joffe M, Holmes J, Jensen TK, Keiding N, Best N. Time trends in biological fertility in western europe. American Journal of Epidemiology. 2013; 178(5):722–730. [PubMed: 23887045]

42. Slama R, Ducot B, Carstensen L, Lorente C, Rochebrochard EdL, Leridon H, Keiding N, Bouyer J. Feasibility of the current–duration approach to studying human fecundity. Epidemiology. 2006; 17(4):440–449. [PubMed: 16755258]

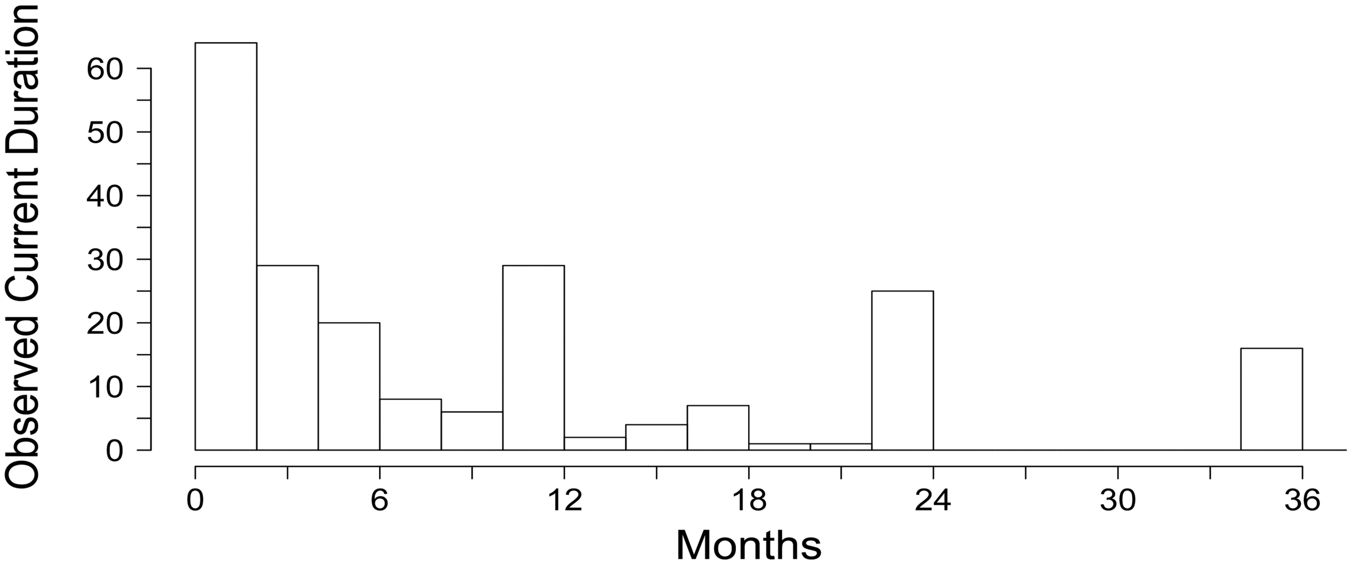

Figure 1.

Histogram of observed current durations from the National Survey on Family Growth. Apparent digit preference is displayed at 12, 24, and 36 months

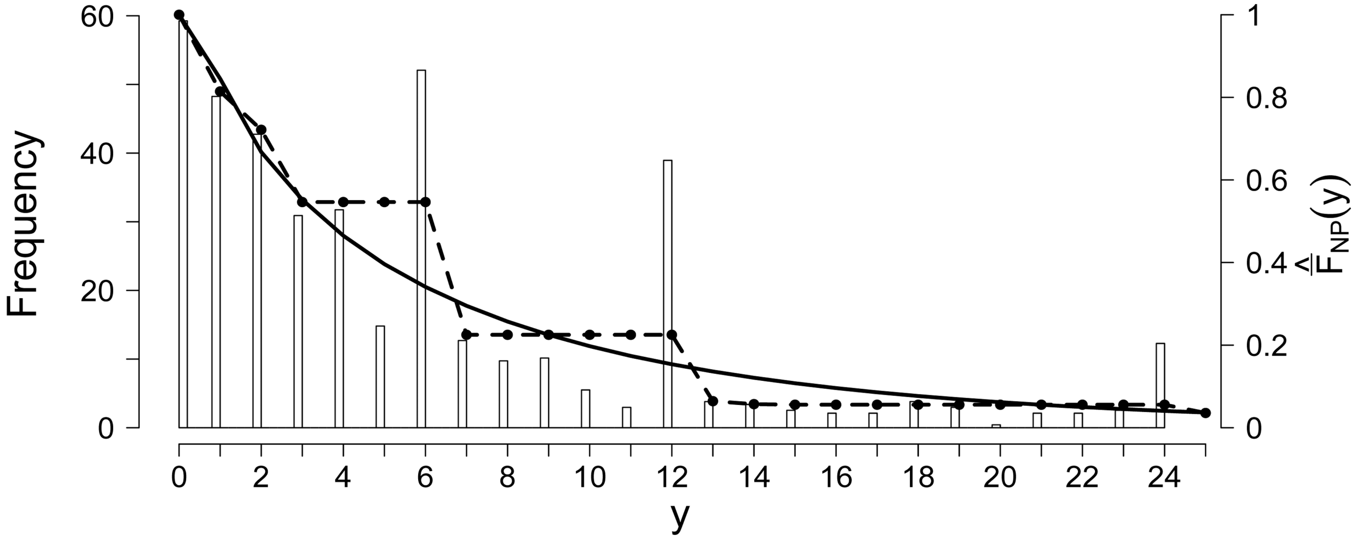

Figure 2.

A histogram of data simulated to have digit preference at y = 6, 12, and 24, overlayed with an unconstrained nonparametric current duration survivor function estimate (dashed line, axis on right) and the true survivor function (solid line).

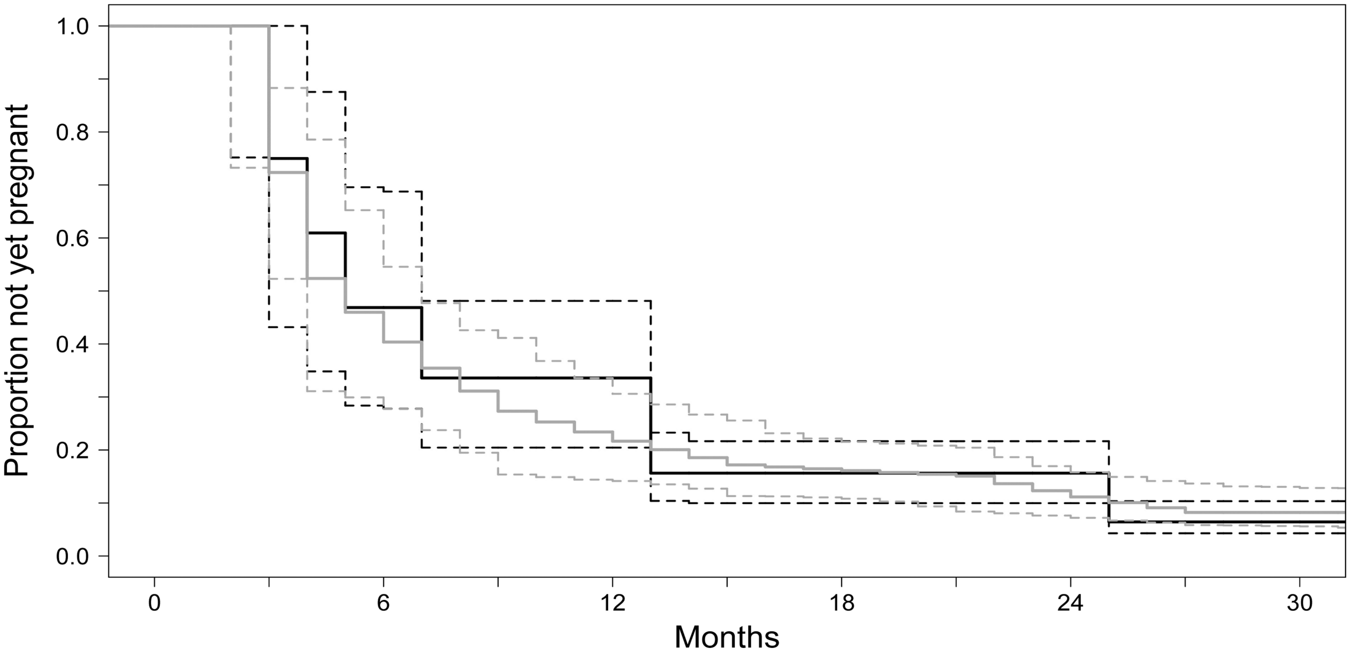

Figure 3.

Current duration estimate of the survivor function of TTP for the NSFG data (solid line), with 95% pointwise bootstrap confidence intervals (dashed line) for the nonparametric method (black) and the unadjusted piecewise constant model (gray).

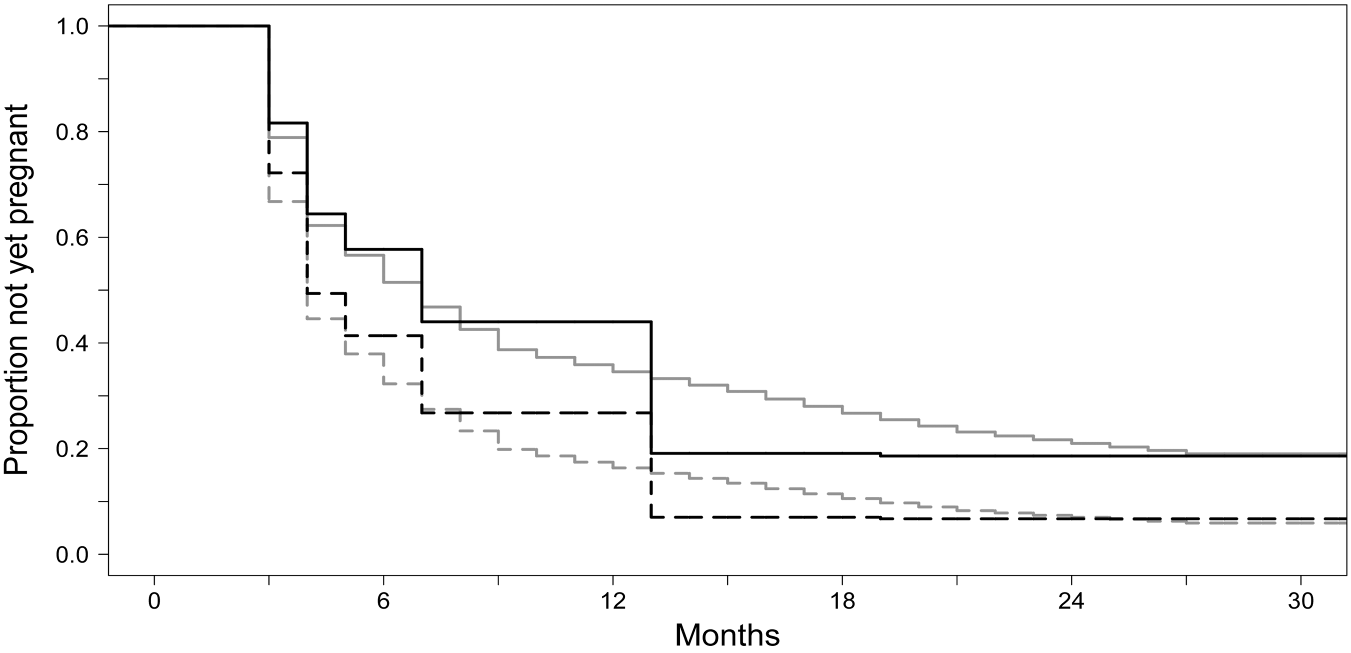

Figure 4.

Semiparametric (black) and piecewise constant (gray) current duration estimates of the survivor function of TTP for parous women (dashed line), nulliparous women (solid line).

Table 1

Summary of 1, 000 simulated samples with n = 250 for the piecewise constant and semiparametric (SP) models when digit preference is present. The piecwise model was fit with knot scenario A {1, 2, 4, 9, 18, 30, $Y_{(m)}$}, scenario B {1, 2, 4, 9, 15, 27, $Y_{(m)}$}, scenario C {1, 2, 5, 8, 11, 18, $Y_{(m)}$}, and scenario E that uses empirical percentiles. Displayed is the empirical bias (BIAS), empirical standard deviation (SD), and $\| \bar{F} - \tilde{\bar{F}} \|_2 (l_2)$.

		Piecewise knot scenario				
		A	**B**	**C**	**E**	**SP**
	TRUE	BIAS (SD)	BIAS (SD)	BIAS (SD)	BIAS (SD)	BIAS (SD)
β_1	−0.5	−0.002 (0.12)	−0.009 (0.12)	−0.016 (0.13)	−0.019 (0.13)	0.004 (0.13)
β_2	−0.5	−0.012 (0.30)	−0.021 (0.30)	−0.027 (0.30)	−0.029 (0.31)	−0.007 (0.30)
$\bar{F}(6)$	0.317	0.026 (0.06)	0.025 (0.06)	0.026 (0.06)	0.109 (0.10)	0.127 (0.08)
$\bar{F}(12)$	0.145	−0.004 (0.03)	−0.007 (0.03)	−0.005 (0.03)	0.024 (0.05)	0.063 (0.05)
$\bar{F}(24)$	0.039	−0.003 (0.01)	−0.001 (0.01)	−0.001 (0.01)	−0.004 (0.01)	0.005 (0.02)
l_2	-	0.077	0.074	0.068	0.155	0.217

Table 2

Summary of 1,000 simulated samples with $n = 250$, and 500 for the piecewise constant and semiparametric models under the discrete (true F is discrete) and continuous (true F is continuous) scenarios. Displayed is the average coefficient (MEAN), empirical standard deviation (SD), empirical coverage probabilities (ECP), and l_2 norm (l_2)

		Discrete *F*				Continuous *F*	
		Piecewise		Semiparametric		Semiparametric	
	TRUE	MEAN(SD)	ECP	MEAN(SD)	ECP	MEAN(SD)	ECP
				n=250			
β_1	−0.5	−0.505 (0.148)	0.964	−0.510 (0.162)	0.950	−0.511 (0.167)	0.949
β_2	−0.5	−0.507 (0.354)	0.951	−0.509 (0.368)	0.945	−0.510 (0.389)	0.952
l_2		0.068		0.053		0.056	
β_1	0.0	−0.005 (0.143)	0.948	−0.002 (0.145)	0.950	−0.003 (0.143)	0.953
β_2	0.0	0.005 (0.349)	0.955	−0.003 (0.356)	0.947	0.008 (0.375)	0.944
l_2		0.074		0.053		0.049	
				n=500			
β_1	−0.5	−0.506 (0.105)	0.952	−0.507 (0.104)	0.956	−0.509 (0.115)	0.944
β_2	−0.5	−0.506 (0.252)	0.949	−0.505 (0.245)	0.965	−0.507 (0.259)	0.954
l_2		0.050		0.029		0.026	
β_1	0.0	0.002 (0.095)	0.948	0.002 (0.097)	0.950	−0.002 (0.100)	0.947
β_2	0.0	0.007 (0.230)	0.965	−0.002 (0.241)	0.950	−0.017 (0.251)	0.957
l_2		0.050		0.034		0.031	

Table 3

(Top) Unadjusted estimates and 95% confidence intervals (95% CI) for the prevalence of total durations longer than 12 and 24 months for the nonparametric and piecewise constant approaches. (Bottom) Regression coefficients and estimates intertility prevalence for 31-year old women that are parous ($\bar{F}(12|1, 0)$) or nulliparious ($\bar{F}(12|0, 0)$) from the semiparametric, and piecewise constant models.

	Unadjusted			
	Nonparametric		Piecewise	
	EST	95% CI	EST	95% CI
$\bar{F}(12)$	0.336	(0.212, 0.481)	0.223	(0.148, 0.310)
$\bar{F}(24)$	0.223	(0.162, 0.290)	0.123	(0.082, 0.167)

	Covariate Adjusted			
	Semiparametric		Piecewise	
	EST	95% CI	EST	95% CI
Parity	0.492	(0.257, 0.728)	0.747	(0.206, 1.289)
Age	−0.035	(−0.054,−0.015)	−0.036	(−0.074,0.003)
$\bar{F}(12|1, 0)$	0.303	(0.138, 0.463)	0.191	(0.110, 0.295)
$\bar{F}(12|0, 0)$	0.475	(0.291, 0.610)	0.358	(0.244, 0.473)