Modeling menstrual cycle length using a mixture distribution

YING GUO, AMITA K. MANATUNGA*
Department of Biostatistics, Emory University, Atlanta, GA 30322, USA
amanatu@sph.emory.edu

SHANDE CHEN
Department of Biostatistics, University of North Texas Health Science Center at Fort Worth, Fort Worth, TX 76107, USA

MICHELE MARCUS
Department of Epidemiology, Emory University, Atlanta, GA 30322, USA

SUMMARY

In reproductive health studies, epidemiologists are often interested in examining the effects of covariates on menstrual cycle length which is a convenient, noninvasive measure of women’s ovarian and reproductive function. Previous literature (Harlow and Zeger, 1991) suggests that the distribution of cycle length is a mixture of a major symmetric distribution and a component featuring a long right tail. Motivated by the shape of this marginal distribution, we propose a mixture distribution for cycle length, representing standard cycles from a Normal distribution and nonstandard cycles from a shifted Weibull distribution. The parameters are estimated using an estimating equation derived from the score function of an independence working model. The fitted mixture distribution agrees well with the distribution estimated using nonparametric approaches. We propose two measures to help determine whether a cycle is standard or nonstandard, developing tools necessary to identify characteristics of the menstrual cycles that are biologically indicative of ovarian dysfunction. We model the effect of a woman’s age on the mean and variation of both standard and nonstandard cycle lengths using multiple measurements of women.

Keywords: Conditional probability; Estimating equation; Kernel density estimation; Menstrual cycle length; Mixture distribution; Optimum cutoff.

1. INTRODUCTION

Menstrual cycles act as overt indicators of underlying reproductive health. Menstrual dysfunction may both decrease fertility and increase future risk of various chronic diseases such as breast cancer, cardiovascular disease, and diabetes. Menstrual cycle characteristics, including cycle length and bleed length, may serve as sensitive and noninvasive measures of reproductive health. While biological assays have the added advantage of measuring hormone levels and the potential to estimate the day of ovulation,

*To whom correspondence should be addressed.
menstrual cycle characteristics are easy to observe, cost effective, and conveniently monitored by women themselves. Altered patterns of menstruation may indicate subclinical states of reproductive dysfunction and may enable earlier detection of potential menstrual dysfunction. Our understanding of the endocrinology controlling menstrual cycles has advanced in recent years. Yen (1991) describes the endocrinology of the menstrual cycle in terms of three phases: the follicular phase, ovulation, and the luteal phase. However, in terms of observable cycle characteristics, there remain no specific criteria distinguishing normal menstrual function from less-severe forms of dysfunction. Epidemiologists have examined the effects of a range of exposures, from caffeine consumption and smoking to environmental contaminants, on menstrual cycle function. Such studies rely on existing and often limited statistical tools for the analysis of menstrual cycle characteristics.

The statistical analysis of menstrual cycle length is complicated for numerous reasons. First, menstrual cycle lengths are distributed with a long right tail, and the parametric distribution for cycle length has not been described. Second, sampling bias will be present if women are followed for a fixed length of time. Women with generally shorter cycles will be overrepresented as they contribute more cycles to the analysis than women with long cycles. Third, depending on the study design, observed cycles are often censored. For example, if the study ends at a predetermined time, the last observed cycle of each woman is typically right censored. In this paper, we aim to develop statistical tools to address these complexities of menstrual cycle data.

The distribution of cycle length consists of both a symmetric part and a long right tail. However, researchers have frequently ignored this and applied Normal-theory regression models to menstrual cycle length data. Harlow and colleagues recognized this problem and applied a bipartite model to analyze menstrual data (Harlow and Zeger, 1991; Harlow et al., 2000). They classified cycles into two groups: ‘standard’ cycles, those from the symmetric part of the distribution, and ‘nonstandard’ cycles, those from the long right tail. Standard cycles are analyzed using Normal-theory statistical methods such as repeated measure analysis of variance. For example, Harlow and Zeger (1991) and Harlow and Matanoski (1991) defined standard cycles as those less than or equal to 43 days and used linear random-effects models to examine the covariate effects on the mean length of standard cycles. Lin et al. (1997) extended the linear mixed model to account for the heterogeneity of within-woman variance of standard cycles. For nonstandard cycles, Harlow et al. (2000) evaluated the age effect on the probability of having a nonstandard cycle using a generalized estimating equation. A limitation of the bipartite approach is that the analysis of the cycle length pattern typically focuses only on cycles in the symmetric part of the distribution with the analysis of the cycles in the long right tail restricted to modeling the probability of having a nonstandard cycle.

In this paper, our first objective is to develop an appropriate parametric form for the marginal distribution of cycle length that can adequately represent both components of the observed distribution. Motivated by previous literature (Harlow and Zeger, 1991; Harlow et al., 2000), we consider a Normal and shifted Weibull mixture distribution. There are several advantages of specifying a parametric form for the marginal distribution: it provides better understanding of the characteristics of cycle lengths, especially those in the long right tail; it leads to appropriate methods for defining standard and nonstandard cycles, thus facilitating common methods of data analysis used in epidemiology; it is needed for making correct inferences on the dependence structure among cycles within women. The second objective of this paper is to model repeated measures of menstrual cycles of women while maintaining the desired mixture marginal distribution. Menstrual cycle lengths are known to be distributed differentially among various subpopulations. Through the proposed parametric distribution, we are able to examine subject-specific covariate effects on both standard and nonstandard menstrual cycles. Compared to its alternatives, the proposed modeling approach has the following advantages: it does not require the cycles to be categorized by an arbitrary cutoff as in the bipartite models; it allows us to target specific aspects of the cycle length distribution, such as the mean length and variability; it enables us to differentiate covariate effects
on standard and nonstandard menstrual cycles, an appealing feature when modeling covariates such as stress that may have different influences in the two parts of the distribution (Harlow and Zeger, 1991); finally, both complete and censored cycles are taken into account in modeling.

In the next section, we introduce the reproductive study that motivated this paper. Some practical issues are discussed regarding menstrual cycle length data. In Section 3, a parametric mixture distribution is proposed. An independence working model (IWM, Huster et al., 1989) is used to obtain parameter estimates. These estimates are shown to be consistent regardless of the true dependence structure among within-woman cycle lengths. We also estimate the marginal distribution nonparametrically and compare it to its parametric counterpart. In Section 4, two methods are developed for distinguishing standard and nonstandard cycles. In Section 5, a marginal modeling approach is developed based on the proposed parametric mixture distribution. An illustration is provided using the reproductive study. We conclude with discussion in Section 6.

2. THE MOUNT SINAI STUDY OF WOMEN OFFICE WORKERS DATA

2.1 Study population

The Mount Sinai Study of Women Office Workers (MSSWOW) was a prospective cohort study to explore the effects of video display terminal (VDT) use on rates of spontaneous abortion (Marcus et al., 2000). The participants for the study were recruited between 1991 and 1994 from 14 companies or government agencies in New York, New Jersey, and Massachusetts. A total of 4640 women office workers completed a cross-sectional questionnaire. Women between the ages of 18 and 40 were eligible for the prospective study if they were at risk for pregnancy. This included women who had sexual intercourse at least once in the past month without using contraception or were planning to discontinue regular contraceptive use in the next 6 months. A total of 25% of the qualified women indicated that they were trying to become pregnant. A woman was excluded if she had already been attempting to conceive a child unsuccessfully for 12 months or longer, if she had a hysterectomy, or if her partner had a vasectomy. Women were not excluded if they experienced a year or more of attempted pregnancy sometime in the past. A total of 524 women were finally enrolled in the study. Participants were observed for 1 year, or until a clinical pregnancy.

2.2 Definition of menstrual cycles and cycle length

According to the World Health Organization standard, a menstrual cycle is defined as the interval from the first day of one bleeding episode up to and including the day before the next bleeding episode. During the study, each participant maintained a daily diary recording of whether menstrual bleeding occurred, whether they had intercourse, and, if so, whether birth control was used. They also recorded information on specific exposures (e.g. hours of VDT use) on a daily basis. We excluded those cycles whose starting date or the date when the bleeding episode began was unrecorded.

As with most other reproductive studies, the MSSWOW data contained incomplete menstrual cycles during which pregnancy occurred. When a woman becomes pregnant, her reproductive endocrinology changes and the estrogen and progesterone levels do not decline as in usual nonpregnant cycles. Consequently, the thickened lining of the uterus is not shed during the pregnancy period and the woman will not experience menstrual bleeding until delivery or other kinds of pregnancy termination. Therefore, a woman is temporarily no longer at risk for menstrual bleeding after conception occurs. In other words, the cycle lengths for pregnancy cycles are inherently missing. For this reason and because the exact conception date of pregnancy cycles was not available in the MSSWOW data, clinical or subclinical pregnancy cycles were excluded from our analysis. We also removed cycles during which hormonal medications were used because these medications have well-known effects on menstrual cycle lengths.
Modeling menstrual cycle length using a mixture distribution

Table 1. Observed cycle length distributions by age groups†

<table>
<thead>
<tr>
<th>Age</th>
<th>Women</th>
<th>Cycle</th>
<th>Mean (SD)</th>
<th>Minimum 1%</th>
<th>Median 99%</th>
<th>Maximum</th>
<th>Interquartile range</th>
<th>Cycles &gt; 40 N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19–25</td>
<td>58</td>
<td>381</td>
<td>31.7 (8.3)</td>
<td>9</td>
<td>15.4</td>
<td>81</td>
<td>6</td>
<td>33 (8.7%)</td>
</tr>
<tr>
<td>26–30</td>
<td>134</td>
<td>967</td>
<td>30.4 (8.1)</td>
<td>13</td>
<td>19.7</td>
<td>104</td>
<td>6</td>
<td>67 (6.9%)</td>
</tr>
<tr>
<td>31–35</td>
<td>144</td>
<td>1076</td>
<td>30.0 (11.2)</td>
<td>9</td>
<td>18.6</td>
<td>189</td>
<td>5</td>
<td>47 (4.4%)</td>
</tr>
<tr>
<td>35–41</td>
<td>108</td>
<td>826</td>
<td>27.9 (5.9)</td>
<td>5</td>
<td>15.3</td>
<td>131</td>
<td>5</td>
<td>21 (2.5%)</td>
</tr>
</tbody>
</table>

†Mean (SD), minimum, quantiles, maximum, and interquartile range are calculated based on complete cycles. Number of cycles > 40 includes both complete and censored cycles.
‡QT: quartile.

In the MSSWOW data, most of the censored cycles occurred at the end of the study. However, a few censored cycles also happened during the study when a subject was too busy or traveling and did not maintain her diary for a period of time.

2.3 Description of the observed menstrual cycle length data

After the exclusions, 3241 menstrual cycles contributed by 444 participants were included in this analysis. Each woman contributed from 1 to 19 cycles with a median of 10 cycles. Among the 3241 cycles, 2901 were complete and 340 were censored. Women’s age ranged from 19 to 41 with a median of 31. The observed complete cycle lengths ranged from 5 to 189 days with a mean of 29.8 days and median of 28 days. Table 1 presents a summary of the observed cycle length distribution characteristics by age groups. Based on the Tremin Trust data, Harlow et al. (2000) suggested 40 days as an appropriate cutoff for standard and nonstandard cycles for women across the reproductive life span. Therefore, we present the observed number of cycles that are longer than 40 days in our study.

3. Estimation of the cycle length distribution

As stated in Section 2, we are interested in characterizing the distribution for menstrual cycles during which no pregnancy occurs and no hormonal medications are used. The cycle length distribution we consider is defined as the distribution of a randomly selected cycle from a randomly chosen woman. To fix notation, let $m$ be the total number of women and $N$ be the total number of menstrual cycles in the data. Let $Y_{ij}$ be the length for the $j$th cycle of the $i$th woman, $j = 1, \ldots, n_i$, where $n_i$ is the number of cycles she contributed. Denote $C_{ij}$ as the censoring time. Let $T_{ij} = \min(Y_{ij}, C_{ij})$ be the observed cycle length and $\delta_{ij} = I(Y_{ij} \leq C_{ij})$ be the censoring indicator. The censoring time $C$ is assumed to be independent of the cycle length $Y$.

Harlow and Zeger (1991) suggested that the distribution of cycle lengths is a mixture of a major symmetric population and a minor long-tailed population. The mixture distribution we propose here is of this type. The Normal distribution is an obvious choice for representing the symmetric dominant part and has been used in most menstrual studies for analyzing standard cycles. There are several possible candidates of parametric distributions that may represent nonstandard cycles. We choose the Weibull distribution because it possesses the desired long right tail shape and favorable mathematical properties. For example, the Weibull distribution has an explicit analytic form for the survival function which is convenient for handling censored cycles. Additionally, the Weibull distribution satisfies the accelerated failure time model, whereby an appealing interpretation of nonstandard menstrual cycles can be obtained by modeling
the logarithm of cycle lengths assuming linear covariate effects. Furthermore, an empirical analysis of our data which will be discussed later supports the Weibull as an appropriate choice for nonstandard cycles.

We write the density of the cycle length \( Y \) as

\[
f(y) = pf_1(y) + (1 - p)f_2(y),
\]

(3.1)

where \( p \) and \( 1 - p \) are the weights for the Normal and Weibull distributions, respectively.

A shifted Weibull distribution is used because nonstandard cycles are defined as cycles with lengths in the long right tail. The shift parameter \( s \) represents the starting point for nonstandard cycles and is estimated from the data.

The model in (3.1) is equivalent to a mixed linear model of the following form:

if a cycle is standard: \[ Y_{ij} = \mu + e^{(1)}_{ij}, \]

(3.2)

if a cycle is nonstandard: \[ \log(Y_{ij} - s) = -\log(1 - \rho) + \kappa e^{(2)}_{ij}, \]

(3.3)

where \( e^{(1)}_{ij} \) follows a Normal distribution with mean zero and variance \( \sigma^2 \), and \( e^{(2)}_{ij} \) follows a log-unit standard exponential distribution. The linear representations of (3.2) and (3.3) allow us to interpret covariate effects on both standard and nonstandard cycle lengths.

Suppose a data set consists of \( (t_{ij}, \delta_{ij}), j = 1, \ldots, n_i, i = 1, \ldots, m \), where \( t_{ij} \) is the observed cycle length and \( \delta_{ij} \) is the censoring indicator. The likelihood for the \( j \)th cycle of the \( i \)th woman is

\[
\{(f(t_{ij}))^{\delta_{ij}} \times (1 - F(t_{ij}))^{1-\delta_{ij}}\}
\]

\[
= \begin{cases} 
\left[p \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t_{ij} - \mu)^2}{2\sigma^2}} + (1 - p)\kappa\rho^\kappa(t_{ij} - s)^{\kappa - 1} e^{-\rho^\kappa(t_{ij} - s)^{\kappa}}\right]^{\delta_{ij}} \\
\times \left[p\Phi\left(\frac{\mu - t_{ij}}{\sigma}\right) + (1 - p)e^{-\rho^\kappa(t_{ij} - s)^{\kappa}}\right]^{1-\delta_{ij}}, & \text{if } t_{ij} > s, \\
\left[p \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t_{ij} - \mu)^2}{2\sigma^2}}\right]^{\delta_{ij}} \times \left[p\Phi\left(\frac{\mu - t_{ij}}{\sigma}\right) + (1 - p)e^{-\rho^\kappa(t_{ij} - s)^{\kappa}}\right]^{1-\delta_{ij}}, & \text{if } t_{ij} \leq s,
\end{cases}
\]

where \( F(\cdot) \) is the cumulative distribution function for the density function \( f(\cdot) \).

To write the full likelihood for the observed data, one needs to specify the multivariate distribution function for the repeated menstrual cycle lengths from the same woman. Additional assumptions are needed to describe the dependence structure of within-woman cycles. Since the goal of this paper is to estimate and model the marginal distribution of cycle length with minimal assumptions, we use an estimating equation based on the IWM. The associated working likelihood is the product of marginal likelihoods over all cycles presuming independence among within-woman observations. Parameters in the marginal models are then estimated by solving the score equation of the working likelihood. Specifically, we write the IWM for woman \( i \) as,

\[
\prod_{j=1}^{n_i} \{([f(t_{ij}; \theta)]^{\delta_{ij}} \times [(1 - F(t_{ij}; \theta))^{1-\delta_{ij}}\}]^{w_{ij}},
\]

(3.4)

where \( \theta = (p, \mu, \sigma, \kappa, \rho)' \). Harlow and Zeger (1991) proposed to use a ‘By-women’ weight function to avoid overrepresenting women with shorter cycles. The weight assigned to the cycle lengths of each woman is inversely proportional to the number of cycles contributed by her, that is \( w_{ij} = \frac{N}{n_i m} \). The ‘By-Women’ weight \( w_{ij} \) is adopted here to balance likelihood contributions across women by downweighting.
subjects with more cycles and upweighting those with fewer ones. The overall IWM likelihood is the product of the likelihood (3.4) over all women.

The corresponding estimating function for woman \( i \) is derived from the score function of the IWM,

\[
U_i(\theta) = \sum_{j=1}^{n_i} \omega_{ij} \left[ \delta_{ij} \frac{\partial \log f(t_{ij}; \theta)}{\partial \theta} + (1 - \delta_{ij}) \frac{\partial \log(1 - F(t_{ij}; \theta))}{\partial \theta} \right].
\]

Under the assumptions that the cycle length \( Y \) is missing completely at random and that the marginal distribution of \( Y \) is correctly specified, a standard argument shows that the estimating function \( \sum_{i=1}^{m} U_i(\theta) \) is unbiased. Hence, the solutions of this estimating equation, \( \hat{\theta} \), are consistent regardless of the nature of dependence among within-woman cycle lengths.

Assuming that cycles from different women are independent and \( \sup\{n_i, i = 1, \ldots, m\} = O(m) \), \( \hat{\theta} \) asymptotically follows a Normal distribution under mild regularity conditions,

\[
\sqrt{m}(\hat{\theta} - \theta) \xrightarrow{d} N(0, D^{-1}(\theta) C(\theta) D^{-T}(\theta)),
\]

where

\[
D(\theta) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} E \left( -\frac{\partial U_i}{\partial \theta} \right), \quad C(\theta) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} E(U_i U_i^T).
\]

One challenging task in fitting this mixture distribution is the choice of the shift parameter \( s \) for the shifted Weibull distribution. The usual likelihood defined as the product of densities evaluated at each observation is in fact a first-order approximation of the true likelihood—the product of probability increments at each observation. When the usual regularity conditions hold, this approximation works well. However, with the shifted Weibull distribution, the shift parameter represents the lower limit of the Weibull distribution and the usual likelihood may go to infinity as the shift parameter approaches the smallest observation, leading to inconsistent estimates of the other parameters. The corrected likelihood proposed by Cheng and Iles (1987) solved this problem by using the proper probability increment, instead of the marginal density, to calculate the likelihood for the smallest observation. A profile likelihood approach, by Cheng and Iles (1987) solved this problem by using the proper probability increment, instead of the marginal density, to calculate the likelihood for the smallest observation. A profile likelihood approach, based on the corrected IWM likelihood, is applied to estimate the shift parameter \( s \). For each value of \( s \) in the grid, we maximize the IWM likelihood over the other parameters. The estimate of \( s \) is then chosen as the value corresponding to the maximum profile likelihood; we find \( \hat{s} = 36 \). The estimated shift parameter \( \hat{s} \) is then plugged into the estimating equation \( \sum_{i=1}^{m} U_i(\theta) = 0 \) to obtain the estimates for the other parameters. With the proposed mixture distribution, the estimating equation does not have an explicit solution and needs to be solved iteratively.

If the shift parameter \( s \) were viewed as fixed, the variance matrix for the other parameters could be readily obtained using the standard sandwich variance estimator. However, we need to take into account the uncertainty in estimating \( s \) when making inferences on the other parameters. In this paper, we use a bootstrap approach for this purpose. Because each woman contributed multiple cycles to the data, a two-step bootstrapping strategy is applied. We first randomly select \( m \) women with replacement from the data, i.e. \( i_1, i_2, \ldots, i_m \) are chosen such that \( i_k \in \{1, \ldots, m\} \) for \( k = 1, \ldots, m \). For each of the selected woman \( i_k \), we then draw \( n_{i_k} \) observations with replacement from her observed data: \( (t_{ikj}, \delta_{ikj}), j = 1, \ldots, n_{i_k} \). Using this strategy, 100 bootstrap samples are selected. For each of these samples, the shift parameter \( s \) and the other parameters in the mixture distribution are estimated using the procedure described earlier. Bootstrap variance estimators are obtained from bootstrap parameter estimates. A Normal approximation is used for hypothesis testing and the validity of the approximation is confirmed by Q–Q plots.

Table 2 presents the estimated mean and standard deviation of both standard and nonstandard cycles for all women and for each 5-year age subgroup. The mean and standard deviation for nonstandard cycles...
are presented on the log-scale but can be transformed to the original scale for interpretation. One observation from Table 2 is that the mean of the Normal distribution decreases with increasing women’s age. Women’s age also seems to have a quadratic effect on the variation of standard cycles. For nonstandard cycles, there is no clear trend in mean cycle length but the variation increases linearly with age. According to the parameter estimates in Table 2, the probabilities for a cycle length to be greater than 40 are 8.6%, 7.0%, 4.6%, and 3.0% in ascending order of age; these probabilities are very close to the observed proportions in Table 1.

To examine the validity of the shifted Weibull distribution, diagnostic plots are obtained using cycles with lengths greater than the estimated shift parameter $\hat{s} = 36$. The Kaplan–Meier estimate $\hat{S}$ is obtained and the plot of $\log(-\log(\hat{S}(t - 36)))$ versus $\log(t - 36)$ yields a fairly straight line, indicating that the Weibull distribution is an appropriate choice for nonstandard cycles. Similar diagnostic plots are obtained with cycles greater than 40 days.

To determine the appropriateness of the fitted parametric mixture distribution, we estimate the marginal distribution of menstrual cycle lengths nonparametrically. Harlow and Zeger (1991) proposed a nonparametric method based on kernel density estimation. The density estimator for complete cycles is

$$f(x) = \frac{1}{N^*} \sum_{i=1}^{m} \sum_{j=1}^{n_i^*} w_{ij} K\left(\frac{x - t_{ij}}{h}\right),$$

where $w_{ij}$ is the ‘By-Women’ weight assigned to each cycle, $n_i^*$ is the number of complete cycles contributed by the $i$th woman, and $N^*$ is the total number of complete cycles for all women. The variable $x$ is the cycle length (in days) for which the kernel density is estimated. In this paper, we use a Normal kernel $K$. The bandwidth $h$ is chosen by the maximum likelihood cross-validation method (Hardle, 1990). To adjust for censored cycles, the Monte Carlo Expectation and Maximization (EM) algorithm is used for the kernel density estimator (Harlow and Zeger, 1991).

As an alternative, we propose a nonparametric approach based on Kaplan–Meier estimates. Assuming that within-woman cycle lengths are independent and identically distributed conditional on a given woman, we first obtain the Kaplan–Meier estimates $\hat{S}_i$, $i = 1, \ldots, m$, for each individual woman based on the multiple cycle lengths she contributed. Then, the overall Kaplan–Meier survival function estimate for cycle length is defined as

$$\hat{S}_{KM}(t) = \frac{1}{m} \sum_{i=1}^{m} \hat{S}_i(t).$$
In this way, each woman contributes equally in calculating the overall Kaplan–Meier estimates. Based on \( \hat{S}_{KM} \), kernel density estimation is applied to obtain the smoothed density estimates. Let \( t_g \) (\( g = 1, \ldots, G \)) denote the \( g \)th distinct cycle length when the complete cycles from all the women are sorted. Denote \( \Delta \hat{S}_{KM}(t_g) \) as the amount of jump at time \( t_g \) in \( \hat{S}_{KM} \), i.e.

\[
\Delta \hat{S}_{KM}(t_g) = \hat{S}_{KM}(t_g^-) - \hat{S}_{KM}(t_g).
\]

The Kaplan–Meier density estimate for cycle length is

\[
\hat{f}(x) = \sum_{g=1}^{G} \Delta \hat{S}_{KM}(t_g) \ K\left( \frac{x - t_g}{h} \right).
\]

The proposed Kaplan–Meier method yields very similar density estimates as does the kernel density estimation method of Harlow and Zeger. The advantages of the Kaplan–Meier method are that it does not require iterations in computation to handle censored observations and is much easier to apply with statistical software such as SAS or SPSS.

In Figure 1, the kernel density estimates are overlaid on the estimated parametric mixture distribution. The parametric and nonparametric estimations agree very well except in two areas: the interval between (30, 50) days and the peak area around 28 days. In both locations, the curvature of the density is large. Therefore, these discrepancies may be attributable to the bias of kernel density estimation, which is proportional to the second derivative of the density (Hardle, 1990).

We also calculated the kernel density estimates for 5-year age subgroups to examine the effect of women’s age on the distribution of cycle lengths. Figure 2 shows that the symmetric part of the distribution slightly shifts to the left with the increase in women’s age. The left-shifting trend is observed in previous papers (Harlow et al., 2000; Treloar et al., 1967) and is also consistent with the trend in the estimated mean of standard cycles for the four age subgroups in Table 2.
4. METHODS FOR DISTINGUISHING STANDARD AND NONSTANDARD CYCLES

In this section, we discuss two methods developed from the proposed mixture distribution for identifying nonstandard menstrual cycles. These two methods provide quantitative and qualitative criteria for distinguishing the two kinds of cycles.

4.1 Conditional probability for nonstandard cycles

We first propose the conditional probability for nonstandard cycles as a measure to quantify the evidence for a cycle to be nonstandard. Given a cycle length, it can be expressed as the conditional probability for the cycle to come from the shifted Weibull distribution, that is

\[
\text{prob}(Y \text{ is from the shifted Weibull distribution } | Y = y) = \frac{(1 - p)f_2(y)}{pf_1(y) + (1 - p)f_2(y)}.
\]

Figure 3 shows that the conditional probability for a cycle to be nonstandard starts from a small value at a cycle length close to 36 days and rises to almost 1 when the cycle length reaches 45 days. The probability is 0.5 at the cycle length around 38 days.

4.2 Optimum cutoff for standard versus nonstandard cycles

In situations such as clinical diagnosis, it is desirable to determine a cutoff in cycle length to define standard and nonstandard cycles. Cycles with length exceeding the cutoff are classified as nonstandard cycles. Previous studies (Harlow and Zeger, 1991; Harlow et al., 2000) have proposed their cutoffs based on the empirical distribution. For example, the cutoff cycle length proposed by Harlow and Zeger (1991) was
chosen as the 99th percentile of the Normal distribution estimated from the observed central tendency and spread of the data. We propose an alternative criterion to decide the optimum cutoff. Two types of classification errors are considered: classify a cycle to be nonstandard when it is actually from the Normal distribution and classify a cycle to be standard when it is from the shifted Weibull distribution. The optimum cutoff is defined to be the cycle length that minimizes the weighted sum of the two misclassification errors. Let $D$ be the optimum cutoff. Obviously, $D \geq s$. We find $D$ to minimize

$$Q = p \times \text{prob}(Y > D|Y \text{ is from the Normal distribution}) + (1 - p) \times \text{prob}(Y \leq D|Y \text{ is from the shifted Weibull distribution})$$

$$= p \left[ \int_{D}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, dy \right] + (1 - p) \left[ \int_{s}^{D} k \rho^k (y-s)^{k-1} e^{-\rho(y-s)} \, dy \right].$$

Using the parameter estimates of model (3.1), we calculate the misclassification probability $Q$ and find the optimum cutoff to be 38 days. In fact, Figure 4 shows that any cutoff between 36 and 45 will result in very small misclassification probabilities.

One interesting observation is that the conditional probability for a cycle to be nonstandard is 0.5 at the cycle length of 38 days. Therefore, cycles that are classified as nonstandard by the optimum cutoff are those that have greater conditional probability to be nonstandard than to be standard. In this respect, the conditional probability method and the optimum cutoff approach agree well.

5. MODELING COVARIATES

One major task in the analysis of menstrual data is to investigate covariate effects on menstrual cycle lengths. In this paper, we focus on modeling the mean and variability of cycle lengths. These two attributes
of the distribution are important indicators of the menstrual function and have been frequently investigated (Treloar et al., 1967; Lin et al., 1997). With the proposed mixture distribution, the mean and variation of standard and nonstandard cycles are summarized by the parameters of the two component distributions. The covariate effects on the two kinds of cycles can then be modeled simultaneously through the corresponding parameters.

Marginally, the cycle length $Y$ is assumed to follow the proposed mixture distribution with the density function defined in (3.1),

$$f(y) = pf_1(y; \mu_i, \sigma_i) + (1-p)f_2(y; \kappa_i, \rho_i),$$

where $f_1$ and $f_2$ are the densities for the Normal and shifted Weibull distributions, respectively. Due to the considerable complexity in estimating the shift parameter $s$, we do not model it in terms of covariates but rather estimate it from all the data. The proportion parameter $p$ in the mixture distribution is not modeled due to our focus on the mean and variation.

Let $x_i$ denote the vector of covariates of interest for the $i$th woman. Here, we choose women’s age as the covariate to illustrate the modeling approach. Based on observations from Table 2, we model the mean and variance of standard cycles using a linear and a quadratic model, respectively. In the quadratic model, ages are centered around the median age in the data set. The models for standard cycles are then

$$\mu_i = \beta_0 + \beta_1 x_i,$$

$$\log(\sigma_i^2) = \alpha_0 + \alpha_1 (x_i - 31) + \alpha_2 (x_i - 31)^2.$$

For nonstandard cycles, let $\eta$ and $\nu$ denote the mean and variance for the logarithm of the shifted Weibull distribution. The mean $\eta$ is modeled with distinct parameters for the four age subgroups in Table 2.
A Wald test is used to test the homogeneity of the four age-specific means. A log-linear model is fitted for the variance $\nu_i$,

$$\eta_i = \eta^{(1)} I_{(x_i \leq 25)} + \eta^{(2)} I_{(26 \leq x_i \leq 30)} + \eta^{(3)} I_{(31 \leq x_i \leq 35)} + \eta^{(4)} I_{(36 \leq x_i \leq 41)},$$

$$\log(\nu_i) = \gamma_0 + \gamma_1 x_i,$$

where $I$ is an indicator variable. The estimating equation based on the IWM is used to obtain parameter estimates and the standard errors are estimated using the bootstrap approach described in Section 3. Table 3 summarizes the results. In comparison, we also report the sandwich variance estimator of the estimating equation where $s$ is assumed to be fixed. As expected, the bootstrap standard errors are slightly larger than the sandwich standard errors due to the variability in estimating $s$. The Q–Q plots confirm that the bootstrap parameter estimates approximately follow Normal distributions. Therefore, the Normal approximation is used for hypothesis testing.

In previous works, various approaches have been applied to analyzing the effect of women’s age on the mean of standard cycles. For example, Harlow and Zeger (1991) used the linear mixed model

$$Y^*_{ij} = \beta_0 + \alpha_i + \beta_1 x_i + \epsilon_{ij},$$

where $Y^*_{ij}$ is the $j$th standard cycle for woman $i$, $\beta_0$ is the common intercept, $\alpha_i$ is the deviation from the common intercept for woman $i$ and has an expectation of 0, $\beta_1$ is the fixed-effect coefficient for age, and $\epsilon_{ij}$ is a random error that follows a zero-mean Normal distribution. According to both the linear mixed model and our model parameterization, the expectation of a standard cycle length $Y^*_{ij}$ is $\beta_0 + \beta_1 x_i$. To

Table 3. Age effect on the mean and variance of standard and nonstandard cycle lengths

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>SE$^\dagger$</th>
<th>SE$^\ddagger$</th>
<th>Z</th>
<th>P-value$^\§$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of standard cycle length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>34.240</td>
<td>1.208</td>
<td>1.089</td>
<td>28.34</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.172</td>
<td>0.038</td>
<td>0.034</td>
<td>-4.53</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Variance of standard cycle length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>2.455</td>
<td>0.116</td>
<td>0.106</td>
<td>21.25</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.016</td>
<td>0.020</td>
<td>0.016</td>
<td>-0.78</td>
<td>0.438</td>
</tr>
<tr>
<td>(Age-31)$^2$</td>
<td>0.013</td>
<td>0.004</td>
<td>0.0033</td>
<td>3.22</td>
<td>0.001</td>
</tr>
<tr>
<td>Mean of nonstandard cycle length (log-scale$^\parallel$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age: 19–25</td>
<td>2.857</td>
<td>0.396</td>
<td>0.343</td>
<td>7.22</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Age: 26–30</td>
<td>2.171</td>
<td>0.217</td>
<td>0.172</td>
<td>9.99</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Age: 31–35</td>
<td>2.446</td>
<td>0.300</td>
<td>0.273</td>
<td>8.15</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Age: 36–41</td>
<td>2.824</td>
<td>0.801</td>
<td>0.919</td>
<td>3.53</td>
<td>0.0004</td>
</tr>
<tr>
<td>Test of homogeneity for age-specific means</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.297</td>
</tr>
<tr>
<td>Variance of nonstandard cycle length (log-scale$^\parallel$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.540</td>
<td>1.591</td>
<td>1.288</td>
<td>-2.23</td>
<td>0.026</td>
</tr>
<tr>
<td>Age</td>
<td>0.138</td>
<td>0.051</td>
<td>0.041</td>
<td>2.73</td>
<td>0.006</td>
</tr>
</tbody>
</table>

$^\dagger$Bootstrap standard errors. The Z statistic and P-values are based on bootstrap standard errors.

$^\ddagger$Sandwich standard errors.

$^\§$P-value for the hypothesis test that the null value equals zero.

$^\parallel$Results from the linear mixed model with the cutoff of 38 days.

$^\parallel$The mean and variance of nonstandard cycle lengths are calculated on the scale of log($Y - 36$).
compare the linear mixed model with the estimating equation approach, we use 38 days as the cutoff for standard and nonstandard cycles and apply the above linear mixed model to standard cycles. Results are also presented in Table 3.

With a woman’s age increasing from 19 to 41, both the estimating equation and the linear mixed model reveal a significant decrease in the average length of the standard menstrual cycles, and the results based on the two methods are quite close. The relatively small standard errors based on the linear mixed model may be due to the fact that all standard cycles are cut off at 38 days, therefore, cycles are more homogeneous. We also fit the linear mixed model using the cutoff of 40 days suggested by Harlow et al. (2000) and obtain similar estimates. Additionally, age has a significant quadratic effect on the variance of standard cycle lengths with the variability first decreasing in the early 20s, reaching the nadir at age 32, and rising thereafter. This quadratic pattern is also observed with the estimated variance presented in Figure 3 of Harlow et al. (2000). A plausible biological explanation is that as women enter their 20s, the menstrual function becomes stable, resulting in less variable menstrual cycles. However, when women approach the later part of the reproductive life span, the stability of menstruation declines due to the aging effect. For nonstandard cycles, the estimated means in the four age subgroups appear to be similar and the Wald test does not reject the homogeneity hypothesis ($p = 0.297$). We identify a significant increasing trend in the variability in nonstandard cycle lengths in our data. This result suggests that older women in the MSSWOW study experience more variable nonstandard cycles than younger women.

6. Discussion

Our scientific focus in this paper is to characterize the marginal distribution of menstrual cycle length and to determine its change with respect to subject-specific covariates. We use an IWM for making inference on the parameters in the marginal model. This method provides consistent estimates, accommodates censored cycles, and offers easy implementation and interpretation. However, there may be an efficiency loss in using this approach, particularly with strong dependence among within-woman cycles. In addition, this marginal approach is not applicable if one is interested in cycle-specific prediction or longitudinal effects. In such a case, additional assumptions regarding the dependence structure are required. A possible extension of the proposed approach is to include a random-effect $a_i$ in both the Normal and shifted Weibull components of the mixture distribution to account for the within-woman correlation. More specifically, one can add $a_i$ in (3.2) and a scaled $a_i$ in (3.3). Under the assumptions that within-woman cycle lengths are independent conditional on the random effect and that the random effect follows a Normal distribution, a full likelihood can be constructed. Since the marginal likelihood does not have an explicit form in this case, an EM algorithm or Gibbs sampler is needed for the parameter estimates. It should be noted that the random-effects model permits a mixture distribution conditional on the random effect, but the unconditional or marginal distribution will not then have the same form. Alternatively, one can model the repeated measures of menstrual cycle lengths using a copula model. Let $Y_{i1}, \ldots, Y_{ni}$ represent cycle lengths from woman $i$. The joint survival function can be defined from marginals through a copula. For example, if an Archimedean copula $\phi$ is used, the joint survivor function is

$$S(y_1, \ldots, y_{ni}) = \phi^{-1}\left\{\phi(S_1(y_1)), \ldots, \phi(S_{ni}(y_{ni}))\right\}.$$  

In this case, the marginal survival functions $S_1, \ldots, S_{ni}$ can be specified as mixture distributions. The dependence among within-woman cycle lengths is characterized by the copula $\phi$. This approach will allow modeling the dependence structure while maintaining the mixture form for the marginal distributions.

Harlow and Zeger (1991) and Harlow et al. (2000) proposed 43 and 40 days, respectively, as the cutoffs for standard and nonstandard cycles. They determined cutoffs based on the empirical Normal distribution and focused on reducing one of the misclassification errors, namely classifying a standard cycle as nonstandard. The optimum cutoff that we propose is based on both components of the mixture
distribution and is defined to minimize the sum of both misclassification errors, leading to 38 days as the optimum cutoff. The optimum cutoff may change with age (Harlow et al., 2000). Since two-thirds of the subjects in our study were within the age of 26–35, the 38 days cutoff may not be generalizable to women across the entire reproductive life span, especially to the years close to the menarche and menopause.

The majority of prior studies on menstrual cycles are focused on standard cycles. With the mixture distribution, we are now able to study nonstandard cycles. However, one needs to be cautious of the fact that the number of nonstandard cycles in healthy women is generally much smaller than the number of standard ones. Furthermore, the length of nonstandard cycles has a much wider range and is more variable than that of standard cycles. Consequently, a small data set with few very long cycles may result in unstable estimates of the parameters for the shifted Weibull distribution.

In strict terms, nonstandard cycles include cycles that are either atypically long or atypically short, though nonstandard long cycles are generally much more common than nonstandard short cycles. Harlow et al. (2000) suggested that nonstandard short cycles are most likely to exist in older women resulting from phenomena of intermenstrual bleeding and polymenorrhea. In this paper, we only consider nonstandard cycles in the long right tail because the number of extremely short cycles in our data set is too small to be used for valid analysis.

The attributes of the menstrual cycle length distribution are related to the subject characteristics of the study population. The subject age range of the MSSWOW study is 19–41. Thus, women close to menarche or menopause are not represented. Since it is known that the menstrual cycle is more variable close to menarche and menopause, the results in this paper may not be generalized to women beyond the age range examined here. Another feature of the MSSWOW data is that the population for this study was selected because they were ‘at risk’ of pregnancy and 25% of these women identified themselves as trying to become pregnant. The attributes of the distribution may also be affected by the study design. For example, the cycle length variability observed with the MSSWOW data (Table 1) is greater than that in the Tremin Trust data (Treloar et al., 1967; Harlow et al., 2000) but is not different from other studies (e.g. Chiazze et al., 1968). A plausible explanation is that the Tremin Trust data are based on women who were followed over many years while the subjects in Chiazze et al. (1968) and in the MSSWOW data were followed for only one or two years. Women who participate in many years of diary keeping may be a subpopulation with more regular cycles.

The definitions of standard and nonstandard cycles in this paper are based on the menstrual cycle length and differ from the concepts of ‘normal’ and ‘abnormal’ cycles in the biological sense. To determine whether a cycle is ‘abnormal’ or not, one would need to measure hormone levels or use other techniques that directly monitor ovarian function. It is likely that menstrual cycles, like other physiologic indicators, are best described on a continuous scale rather than a discrete one, i.e. ‘normal’ and ‘abnormal.’ The aim of this paper is to develop statistical tools that will enable researchers to make better use of menstrual cycle data as indicators of underlying biological function.

One of the reviewers pointed out the potential problem of informative censoring on cycle length due to pregnancy. There are several challenging issues in addressing this problem. First, the censoring time, which is the conception time, is very difficult to obtain in practice because the conception date can only be measured with error using techniques such as ultrasound, or be approximated through back-calculation based on a gestational period of 40 weeks. In the MSSWOW data as well as in many other reproductive studies, the conception dates are not available. Second, even when the conception time is measured, there are additional difficulties in adjusting for the potential-dependent censoring due to pregnancy. For example, when conception occurs, a woman’s risk for the event, which is the occurrence of menstrual bleeding, becomes zero due to the change in her reproductive endocrinology. More complex statistical methods with additional assumptions regarding the censoring mechanism are needed to address these issues related to pregnancy cycles.
ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health grants, R01-ES012458-01 and R01-HD24618, and a grant from the University Research Committee of Emory University. We are grateful to the referees and editors for constructive suggestions that have significantly improved this paper. We also thank Chanley Small for her helpful comments in revising the manuscript.

REFERENCES


[Received December 11, 2002; first revision January 26, 2004; second revision December 9, 2004; third revision May 23, 2005; fourth revision June 29, 2005; accepted for publication July 13, 2005]