An Analysis for Menstrual Data with Time-Varying Covariates

SUSAN A. MURPHY[1], GILLIAN R. BENTLEY[2] AND MARY ANN O'HANESIAN[3]

September 20, 1996

Summary

This paper concerns the analysis of menstrual data; in particular, methodology to identify variables that contribute to the variability of menstrual cycles both within and between women. The basis for the proposed methodology is a parameterization of the mean length of a menstrual cycle conditional upon the past cycles and covariates. This approach accommodates the length-bias and censoring commonly found in menstrual data. Data from a longitudinal study of menstrual patterns and other variables among Lese women of the Ituri Forest, Zaire, illustrate the methodology. A small simulation illustrates the bias caused by incorrectly deleting the censored cycles.

Key words: Menstrual data; Length-bias; Generalized estimating equations; Longitudinal data

[1]Department of Statistics, Pennsylvania State Univ., 326 Classroom Building, University Park, PA 16802

[2]Department of Anthropology, Northwestern Univ.

[3]Duke Medical Center, Duke Univ.

1. Introduction

Menstrual patterns have been the focus of much research beginning in the 1920s when scholars first recognized the variability inherent in menstrual cycle lengths and periods of menstrual bleeding. This earlier research on menstrual patterns focused on evaluating the characteristics of menstrual cycle lengths in Western populations, and describing variation across the reproductive lifespan (e.g., Chiazze et al.[1]; Treloar et al.[2]; Vollman[3]). In these papers, the statistical interest was in covariates that, differing by woman or slowly changing with time, contribute to the variation in menstrual cycle lengths (e.g. age).

A normal cycle is defined by three specific phases: a) the follicular phase, lasting approximately two weeks, beginning with the first day of menstrual bleeding and ending with ovulation. During this phase, estrogen levels are high and cause the proliferation of the uterine lining for possible conception and implantation; b) ovulation, lasting approximately two days, during which time an ovum is released from the ovary and travels down the fallopian tubes to the uterus; c) the luteal phase, beginning after ovulation and ending with the appearance of menstrual bleeding. This third phase also lasts approximately two weeks. It is characterized by high levels of progesterone which maintain the thickened uterine lining in the case of conception and pregnancy (see Knobil and Hotchkiss, 1988; Yen, 1991).

Since the development of techniques that allow for the measurement of hormonal levels that characterize the menstrual cycle and for the assessment of whether ovulation has occurred, there has been renewed interest in how menstrual cycle lengths and patterns of menstrual bleeding can, by themselves, provide information about a healthy, normal cycle in different women. In particular, there has been considerable increase in the assessment of variables that affect the normal cycle. These include women's nutritional status and dietary composition, patterns of exercise, general health, and the use of various drugs (e.g., Hill et al.[6]; Schweiger et al.[7]; WHO[8]; Yen[5]). In these cases, interest is in assessing covariates that, in varying by cycle and/or varying by woman, contribute to the intra- and inter-woman variability of menstrual cycles.

Indeed, researchers have documented that the intra-woman variation in cycle length is high (see Harlow and Zeger[9]; Harlow and Matanoski[10]; Münster et al.[11]), giving impetus to the search for covariates, varying by cycle within a woman. The development of steroid contraceptives that inevitably alter menstrual characteristics has also stimulated much study in this area. Observed menstrual patterns can often inform the researcher about potential reproductive problems without the necessity of invasive procedures.

The statistical analysis of menstrual patterns is complex. The complexity begins with the very use of the phrase, menstrual cycle, which has particular connotations of normalcy(i.e., follicular and luteal phases, and a mid-cycle ovulation). Without hormonal or other biological measurements to assess these normal characteristics, it is not possible to distinguish between menstrual and non-menstrual bleeds. For example, the withdrawal bleeding experienced by women who use steroid contraceptives is, by definition, a non-menstrual bleed. Therefore instead of the word, "menstrual cycle," the word, "menstrual segment," is commonly used to denote the interval from the first day of bleeding up to, and including, the day preceding the next onset of bleeding (Belsley and Farley[12]; Harlow and Matanoski[10]). In the example we consider here, measurements of salivary steroids were used to verify menstrual bleeds, and so we retain the phrase menstrual cycle for our terminology.

Some of the many statistical challenges to the researcher (Belsley and Farley[12]; Harlow and Zeger[9]) include the following. First, in the analysis of menstrual cycle length, one usually observes women over a fixed length of time (e.g., one year) and records the intervals between menses. This implies that women who have a short habitual cycle contribute more cycles than women with a long habitual cycle. That is, the sampling of cycles is length-biased. Second, since individual women will begin and end their cycles on different days of any given month, the first and last cycle are only partially observed. Even if a planned study is undertaken in which each woman will be followed for a fixed number of cycles, the inevitable dropout of women results in fewer cycles and a partially observed last cycle. Two additional characteristics of menstrual data are the presence of time-varying covariates and the lack of a known parametric distribution for the length of a menstrual cycle. In particular, menstrual cycle data exhibit a long right tail.

Other important statistical challenges in the analysis of menstrual data are reviewed by Burch, Macisco and Parker[13], Belsley and Farley[12] and Harlow and Zeger[9]. We do not address the issue of inter- or pre-menstrual spotting, which is commonly discussed in studies dealing with contraception and which complicates the assessment of cycle lengths (Belsley and Farley[12]; Snowden[14]).

The focus of this paper is methodology for the identification of factors that contribute to menstrual cycle variability. In do this, we first focus attention on the above-mentioned challenges that arise in the statistical analysis of menstrual cycles. In the third section, we propose a methodology based on a sequential model, i.e., we model the mean cycle lengths conditionally on the cycles already observed. This means that we consider covariates for which the variation in covariate level corresponds to variation in mean cycle lengths. In this case changing levels of the covariates can be used to explain variability in cycle length. It could be that a high level of a covariate is associated with higher cycle-to-cycle variability, or the opposite might be the case. For simplicity we do not consider associations of this type.

As we shall see, observation of women over a fixed length of time, requires either implicit or explicit assumptions on the conditional mean cycle length. Additionally, several arguments can be made in favor of a conditional approach. First, this is the way the data actually arise; indeed, both the length-bias and censoring result from the sequential observation of cycle lengths. Second, as mentioned above, researchers have documented that intra-woman variation in cycle length is high, giving impetus to a search for covariates, varying by cycle within a woman. As will be the case in our example, these covariates may depend on the previous cycles. A final reason is the type of inference desired; if one intends to use the data for prediction, then the sequential method is advantageous as it allows one to make predictions without a strong parametric assumption on the distribution of cycle lengths. For example, it is possible to predict that, if a woman loses weight/gains weight/ceases or curtails an exercise regimen, then her cycle lengths will, on average, become shorter/longer by a certain number of days (e.g., Harlow and Matanoski[10]; Jones et al.[15]; Pirke et al.,[16]).

We illustrate the methodology in the fourth section with data from a longitudinal study of menstrual patterns and other variables among Lese women of the Ituri Forest, Zaire (Bentley et al.,[17], Bentley et al.,[18]). In this last section we also present a small simulation that evaluates the proposed methodology and illustrates the bias resulting from a deletion of the censored cycles.

2. A Discussion Of Some Issues

1. *The observations:* To our knowledge, there has been no precise, quantitative, specification of observed cycle lengths among women. This can be done in the following way: we follow a woman for C days after entering the study and, for simplicity, we assume that she enters the study on the first day of her cycle. Say the length of this cycle is Y_1. We observe the length, Y_1 fully if $Y_1 < C$ otherwise we only observe that $Y_1 > C$. If $Y_1 < C$, then we begin

observation of a second cycle, with length Y_2 days. Continuing in this fashion, we begin observation of the third cycle if $Y_2 + Y_1 < C$, etc. In general, if $\sum_{i=1}^{k-1} Y_i < C$, we observe the minimum of the kth cycle, Y_k and $C - \sum_{i=1}^{k-1} Y_i$; otherwise we do not make an observation. In prospective or longitudinal studies, some individuals inevitably leave the study early. As long as the time of departure of the woman depends on outside influences, or on past cycles and covariates, replacement of C above by the time of departure does not invalidate the statistical analysis below. Since the presence/absence of a cycle is determined by the total length of the previous cycles, we must specify either implicitly or explicitly, a model for the conditional distribution of a cycle length given the past cycle lengths.

2. *Length-biased Sampling:* To focus on length-bias, we simplify the above description by assuming that we follow a woman after time C until she completes her present cycle, thereby temporarily ridding the data of censoring. So we completely observe Y_k if $\sum_{i=1}^{k-1} Y_i < C$. Note that $\sum_{i=1}^{k-1} Y_i$ will be smaller, on average, for a woman with a short habitual cycle length, than for a woman with a long habitual cycle length, thus resulting in an oversampling of cycles from the former woman. We may view the data from two viewpoints. First, the unit of observation is a cycle, and we have a length-biased sampling of cycles. Or secondly, the unit of observation is a woman, and we observe the sequence, Y_k times an indicator of whether $\sum_{i=1}^{k-1} Y_i < C$ for $k \geq 1$ for each woman. Adding in the censoring, we observe on each woman, the minimum of $(Y_k, C - \sum_{i=1}^{k-1} Y_i)$ times an indicator of whether $\sum_{i=1}^{k-1} Y_i < C$ for $k \geq 1$.

3. *Incomplete observation of the first and last cycles:* It is common in the analysis of menstrual cycles to delete each woman's first and last incomplete cycle lengths (for example, in Belsey and Carlson[19] and Harlow and Matanoski[10]). If the length of observation interval, C days, is shorter than the longer cycles, this approach will cause the deletion of all women whose data consists of one long cycle that began prior to the observation interval and ending after the

observation interval. As a result one undersamples women with a long habitual cycle and in addition, the women remaining in the sample will contribute only short cycles.

We assume that C days is long compared to cycle lengths in the population of interest. As a result, even if the first incomplete cycle is deleted, the sample will be composed of a mixture of women with long and short habitual cycles. This combination will, on average, mirror the population mixture of women with long and short habitual cycles. Deleting the first (incomplete) cycle will not introduce bias in the estimation because we make this decision independently of the length of this cycle. However, bias is caused by deleting the last cycle. This occurs because, in order to know whether a cycle is last and thus will be deleted, we must know that the cycle length is greater than the observation time minus (the total length of the previous complete cycles plus the partial length of the first incomplete cycle). Therefore deletion of a woman's last cycle is equivalent to artificial production of informative dropout; in particular, the dropout depends on the length of the deleted cycle and, thus, is not even missing at random as defined by Little and Rubin[20].

Although deletion of the first incomplete cycle does not cause bias, we lose information. The inclusion of the first incomplete cycle is complicated. This is because this cycle is length-biased in the direction of being longer than a woman's habitual cycle. That is, if we choose an arbitrary point in time at which to begin observation of a woman, then a longer cycle (as compared to her habitual cycle length) more likely contains this point in time compared to a shorter cycle. For more on this topic, see Ross[21] (pgs.328,334). For simplicity, we choose in the following to delete the first incomplete cycle; we replace C by C minus the partial length of the first incomplete cycle.

If we observe all of the women long enough for each to contribute more than a few cycles, then the bias caused by deleting the last cycle may not be large (i.e., observe Y_i if

$\sum_{j=1}^{i} Y_j < C$). This is because $\sum_{j=1}^{i} Y_j$ is highly likely to be smaller than C for a high percentage of the observations. Thus, pretending that the ith observation on a woman is Y_i rather than (Y_i times an indicator of $\sum_{j=1}^{i} Y_j < C$) should not produce much bias in estimation. However, when we include time-varying covariates in the model, the last cycles may be very informative. This occurs in the example we discuss below.

4. *Time-varying covariates:* We can partition the variability in cycle lengths into inter- and intra-woman variation. Time-varying covariates may explain some of the observed intra-woman variability through their association with the mean cycle length. Yet, in spite of this, there has been little research using time-varying covariates to explain the intra-woman variation. Some notable exceptions are Harlow and Matanoski[10] and, in the field of sports medicine, Bullen et al.[22]. We consider here the inclusion of time-varying covariates that remain relatively constant over the duration of one cycle, but change from cycle to cycle. We shall make the simplifying assumption that if the covariate changes within a cycle, this change does not affect that cycle's length. If one wishes to assess the impact on cycle length of a covariate (e.g., hormonal measurements) that changes within a cycle, then an event history analysis model such as the one proposed by Aalen and Husebye[23] or Wei et al.[24] is more applicable.

The value of a time-varying covariate for the present cycle may depend in some known or unknown way on the previous cycle lengths. For example, suppose we determine body mass index (BMI) for a cycle by its value on the first day of that cycle. But, the calendar date for the first day of that cycle is given by the total length of the previous cycles as well as the date of woman's entry into the study. So, the particular value of BMI, we use for a cycle depends on the total length of the previous cycles. Another example of a time-varying covariate that depends on the previous cycles is the previous cycle length. This covariate would be useful if one believed that the length of the immediately preceding cycle has an effect on (after allowance for the covariates) the present cycle length.

5. *Unknown cycle length distribution:* Many authors have documented, this important problem (Harlow and Zeger[9]; Chiazze et al.[1]; Treloar et al.[2]). In particular, these authors document the presence of a long right tail in the distribution of cycle lengths. One possible distribution is a mixture of a normal distribution and an exponential as mentioned by Harlow and Zeger. One can take various approaches with long cycles. It may be possible to explain long cycles by either covariates varying with a cycle, or covariates varying within a cycle. Or, one may analyze separately cycles over a determined length (arbitrarily in the interval 40 to 50 days) as is done in Harlow and Zeger[9], and Harlow and Matanoski[10]. This is equivalent to giving long cycles zero weight in a joint analysis with shorter cycles. Another possibility is to progressively discount the information in the very long cycles by using a robust form of estimation. Then it is possible to conduct a joint analysis of the effect of a covariate on the length of the cycle (whether long or short). We discuss this further below.

3. A Methodology

Our basic approach is to parameterize the conditional mean of a cycle length given the previous cycle lengths and the present and past covariate values. We then consider a variant of the generalized estimating equation, given in Murphy and Li[25], to estimate the parameters in the conditional mean. Define Y_{ij} as the jth cycle length of the ith woman and Z_{ij} as the corresponding covariate value. Denote the length of time from the first complete cycle to the last partially observed cycle as C_i (C_i may denote the woman's departure time or termination of the study). Suppose we have already parameterized the conditional mean of the cycle length, denoted by $\mu_{ij}(\beta)$ and the conditional variance, denoted by $V_{ij}(\beta)$; note both the conditional mean and variance can be functions of past cycle lengths and past and present covariate values. We do not require that $V_{ij}(\beta)$ be a function of β only through $\mu_{ij}(\beta)$ as is the case in generalized estimating equations. We discuss some possibilities for parametrization below.

If we could observe the last cycle completely, then we would proceed as follows. To estimate β we solve the generalized estimating equation,

$$S_C(\beta) = \sum_{i=1}^{n} \sum_{j \geq 1} \dot{\mu}_{ij}(\beta) V_{ij}(\beta)^{-1} (Y_{ij} - \mu_{ij}(\beta)) \delta_{ij} = 0,$$

where δ_{ij} is one if $\sum_{l=1}^{j-1} Y_{il} < C_i$ and zero otherwise and $\dot{\mu}_{ij}(\beta)$ is the derivative of $\mu_{ij}(\beta)$ with respect to β. As in generalized estimating equations, approximately correct specification of the conditional covariance of a cycle length given the previous cycle lengths improves the efficiency of the estimator (Zeger and Liang[26]; Liang and Zeger[27]). As long as we specify the conditional mean correctly and the C_i's depend on only the past observations, the above estimating equation has expectation zero even if we misspecify V_{ij}. Note that deleting the last cycle is equivalent to setting δ_{ij} to one if $\sum_{l=1}^{j} Y_{il} < C_i$ and zero otherwise. Since δ_{ij} depends on Y_{ij}, and therefore no longer depends only on the past, the estimating equation will, in general, have a nonzero mean and will yield inconsistent estimators of β.

If we desire to discount the long cycles progressively then we could replace the term $V_{ij}(\beta)^{-1/2}(Y_{ij} - \mu_{ij}(\beta))$ by $\psi(V_{ij}(\beta)^{-1/2}(Y_{ij} - \mu_{ij}(\beta)))$ where ψ is Huber's ψ function as is done in M-estimation (see Ruppert[28]). We would also have to generalize the steps in the following paragraphs. The disadvantage of this approach is that μ_{ij} is no longer a conditional mean but is similar to a conditional trimmed mean. We do not discuss this approach further.

The additional complication, that the last cycle length is not fully observed but is known only to be larger than C_i minus the sum of the previous cycles, suggests the following approach. As an estimating equation, take the conditional expectation of the above estimating equation given the observed data. This yields,

$$S_I(\beta) = \sum_{i=1}^{n} \sum_{j \geq 1} \dot{\mu}_{ij}(\beta) V_{ij}^{-1}(\beta) \Big(E_\beta[Y_{ij}|obs.] - \mu_{ij}(\beta) \Big) \delta_{ij},$$

where $E_\beta[Y_{ij}|obs.]$ is equal to Y_{ij} if $\sum_{l=1}^{j} Y_{il} \leq C_i$ and is equal to $E_\beta[Y_{ij}|Y_{ij} > C_i - \sum_{l=1}^{j-1} Y_{il}]$ otherwise.

To solve $S_I(\hat{\beta}) = 0$ for $\hat{\beta}$, we borrow an idea from the E-M algorithm of Dempster, Laird and Rubin[29]. That is, we iterate between imputing the conditional expectation, $E_\beta[Y_{ij}|Y_{ij} > C_i - \sum_{l=1}^{j-1} Y_{il}]$, using the current value of the parameters and and solving the equation with fixed conditional expectations. Unlike Dempster, Laird and Rubin we do not maximize a likelihood function; instead we solve an estimating equation set to zero.

Note that $E_\beta[Y_{ij}|Y_{ij} > C_i - \sum_{l=1}^{j-1} Y_{il}]$ is a function of β and it involves greater knowledge of the conditional distribution of a cycle length than we have assumed. In our analysis we consider three working conditional distributions, a normal, and a sum of an exponential and a normal, and an exponential in order to impute the conditional expectation, $E_\beta[Y_{ij}|Y_{ij} > C_i - \sum_{l=1}^{j-1} Y_{il}]$. Harlow and Zeger[9] use a mixture of a normal and an exponential for the distribution of cycle lengths, to account for the long right tail present in most menstrual data. One could use other distributions such as a log normal or distributions with longer right tails. Note that to preserve the robustness of our model we use the parametric form of the cycle length distribution only in the imputation of the missing cycle length. In the simulation to follow we compare the effects of different imputation distributions. We describe a computational method for solving $S_I(\hat{\beta}) = 0$ in the appendix.

An estimator of the asymptotic variance of the parameters can be formed in a similar fashion to the method used in generalized estimating equations (see Liang and Zeger[27] and the appendix). The estimator is, $\dot{S}_I^{-1}(\hat{\beta})\hat{\Sigma}_I(\hat{\beta})\dot{S}_I^{-1}(\hat{\beta})^T$ for

$$\dot{S}_I(\beta) = -\sum_{i=1}^{n}\sum_{j\geq 1}\dot{\mu}_{ij}(\beta)V_{ij}^{-1}(\beta)\Big(\frac{dE_\beta[Y_{ij}|obs.]}{d\beta} - \dot{\mu}_{ij}(\beta)\Big)^T\delta_{ij},$$

and

$$\Sigma_I(\beta) = \sum_{i=1}^{n}\sum_{j\geq 1}\dot{\mu}_{ij}(\beta)V_{ij}^{-2}(\beta)\Big(E_\beta[Y_{ij}|obs.] - \mu_{ij}(\beta)\Big)^2\dot{\mu}_{ij}^T(\beta)\delta_{ij}.$$

Note that for all but the last cycle, $E_\beta[Y_{ij}|obs.] = Y_{ij}$ so that the above terms simplify greatly; the term, $\frac{dE_\beta[Y_{ij}|obs.]}{d\beta}$, is the necessary adjustment for the imputation.

There is a wide variety of choices to parameterize the conditional mean of a cycle length given the previous cycles. One possibility is to use an autoregressive model of order one in which $\mu_{ij}(\beta) = Z_{ij}\gamma + \eta(Y_{i(j-1)} - Z_{i(j-1)}\gamma)$. In this case the vector β is composed of the vector, γ and the scalar, η. Another possibility is to include the previous cycle length in the covariate vector, Z_{ij} for the present (jth) cycle length.

In the example below, we construct a very simple model that conforms to a classical analysis. If equal numbers of cycles had been collected for each woman, a plausible working model would have been a mixed effects model with a random block effect for woman and fixed covariate effects. Harlow and Mantanoski[10] use a model of this type to explore the relationship between weight, other variables and menstrual cycle length. We use this working model to form the conditional means. That is, we parameterize the mean of a ith woman's jth cycle given the previous cycles by,

$$\mu_{ij}(\beta) = Z_{ij}\gamma + \frac{\rho}{\rho(j-1)+1-\rho}\left[\sum_{l=1}^{j-1}Y_{il} - \sum_{l=1}^{j-1}Z_{il}\gamma\right], \quad (1)$$

where $\beta = (\gamma, \rho)$. In the mixed effects model, ρ is the intra-woman correlation coefficient. Note that we have not made a distributional assumption on the cycle length; we have only parameterized the conditional mean. In this case, ρ functions as an intra-woman correlation coefficient in that if ρ is zero, the covariance between $Y_{ij} - Z_{ij}\gamma$ and $Y_{il} - Z_{il}\gamma$ is zero for $j \neq l$ and the past is no longer helpful in estimating the mean of the next cycle length.

We also use the working model to specify the conditional variance as, $\sigma^2 V_{ij}(\beta) = \sigma^2\Big(1+\frac{\rho}{\rho(j-1)+1-\rho}\Big)$. In the imputation step, we need an estimator of σ^2; analogous to generalized estimating equations, we estimate σ^2 by solving

$$\sum_{i=1}^{n}\sum_{j\geq 1}\left(\frac{(Y_{ij}-\mu_{ij}(\hat{\beta}))^2}{V_{ij}(\beta)} - \sigma^2\right)\delta_{ij} = 0. \tag{2}$$

As before, we do not know the value of $(Y_{ij} - \mu_{ij}(\beta))^2$ for the last cycles so we replace this term by the conditional variance of Y_{ij} given $Y_{ij} > C_i - \sum_{l=1}^{j-1} Y_{il}$. The presence of the additional parameter, σ^2, requires an adjustment to the variance estimator given in the previous section. We give this adjustment in the appendix.

4. An Example And Simulation Results

The data considered here are a subset of information collected in a longitudinal study of menstrual patterns and other variables among Lese women of the Ituri Forest, Zaire (Bentley et al.[17], Bentley et al.[18]). Menstrual cycle lengths were verified using salivary progesterone profiles obtained for most cycles and analyzed using radioimmunoassay. Details of the methods used for the analysis of salivary progesterone can be found in Ellison[30], 1988, and Ellison, Peacock and Lager[31,32], 1986, 1989). A discussion of the biological implications of this data set with a complete analysis will also appear in Bentley et al.[33]. The study followed women through a period of eight months (January through August, 1989) at a time when Lese women experienced varying nutritional conditions and fluctuations in body mass. This time period consists of a three-month prehunger season (January through March), a hunger season (April through June), and a posthunger season (July and August). The Lese suffer from an annual hunger season generally occurring from March through June that varies in severity from year to year. It is ultimately tied to the pattern of rainfall in the Ituri Forest that determines how large a garden can be cleared, burned and planted (see Bailey et al.[34] for further details). In years of greatest shortage, the Lese have been known to lose up to eight percent of their total body weight. Since previous studies on other populations have found that menstrual cycles lengthen as women lose weight (Bullen et al[22]; Schweiger[35]), we expect cycles to lengthen with weight loss during the hunger season.

Other covariates are age group (age1 is 16-22 years, age2 is 23-34 years, age3 is 35-45 years and age4 is 45+ years), parity (an indicator of numbers of live children ever-born), pregnancy (an indicator of ever being pregnant), and locale (an indicator of one of two village sites where the field study occurred, one at a greater distance from the researchers' field station than the other). These covariates do not vary with time. The covariates that do vary with time are the BMI (body mass index = weight in kilograms divided by the square of height in meters) and lactation (an indicator of whether the woman is presently lactating). The BMI and lactation for a cycle were set to their values on the first day of the cycle.

As in the analysis of Harlow and Zeger[9], there is no evidence to support a serial correlation, beyond that due to a woman effect ($p > .2$, Cliff and Ord[36]). Hence, we adopt the parameterization (1) for the conditional mean of a cycle length given the previous cycles and covariate values. Along with this, we specify $\sigma^2 V_{ij} = \sigma^2 \left(1 + \frac{\rho}{\rho(j-1)+1-\rho}\right)$ as the conditional variance of the jth cycle of the ith woman conditional on her past cycle lengths and covariate values. To solve $S_I(\gamma, \rho) = 0$ for (γ, ρ), we use three working distributions to impute the censored cycle lengths: a normal distribution (skewness = 0), a normal plus an exponential (skewness = 1) and an exponential (skewness = 2).

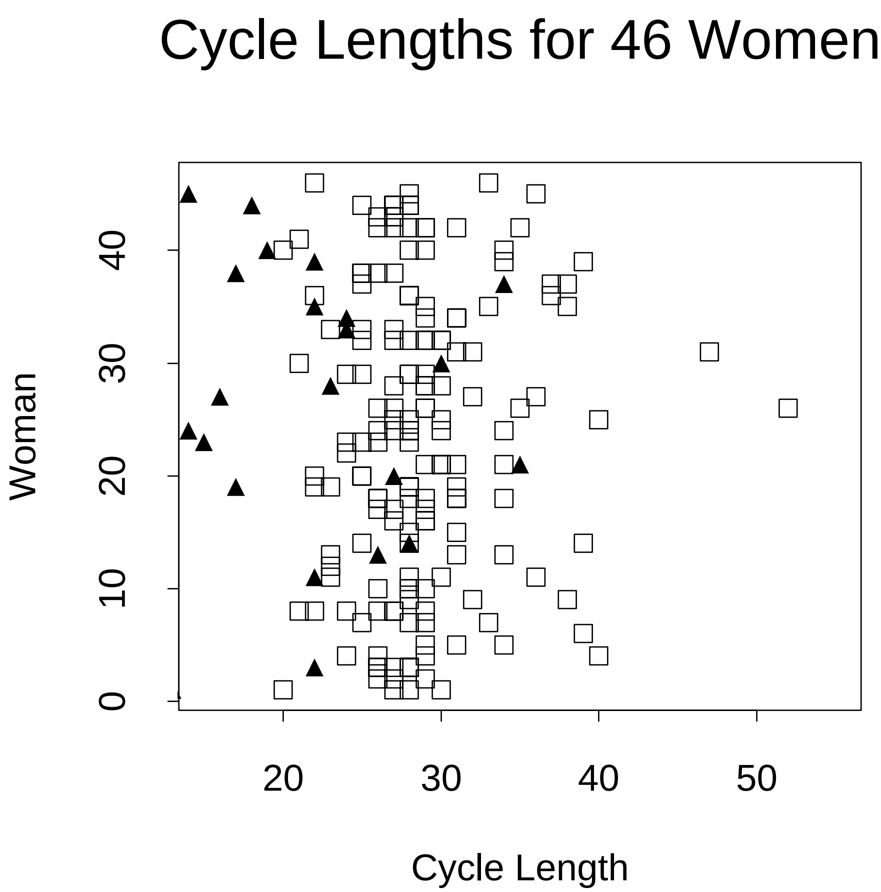

Figure 1: Cycle Length versus Woman Labeled by Censoring, (uncensored:squares) (censored:triangles).

We consider 176 cycle lengths for 46 women. In this example, we begin observation with the first complete cycle during which a researcher was present and for which there was a measurement of BMI. Each horizontal row on Figure 1 gives the cycle lengths for a woman. Open squares depict the complete cycles and dark triangles depict censored cycles of length greater than 14 days.

The results of the statistical analysis appear in Table 1. Pregnancy is insignificant ($p > .5$) and is missing for two women, so we have omitted this variable from Table 1. Cycle length is centered at 28 days and BMI at 21 kilograms/(meters)^2. We used the SAS package, Interative Matrix Language, and in all three cases, we found three iterations sufficient for convergence of the parameters. The reported standard errors are calculated via the formula given in the appendix. These estimated standard errors are virtually identical to the standard errors based on (2); recall that (2) is not adjusted for the estimation of σ^2.

All three imputation distributions yield similar conclusions. This is to be expected since we only use the imputation distribution to garner information from the censored cycles. This analysis lends credence to the conjecture that as the BMI of women in the nearer locale decreases, their cycle lengths increase. Two possible explanations for the lack of a negative trend between cycle length and the BMI for women in the the further locale is that, first, they had access to more food and had better nutritional status. Alternatively, the data might not be as reliable for these women, since the researchers could not question them as frequently about their menstrual status. An incorrect analysis of the data based on the same parameterization above but with the last cycle deleted does not provide evidence for an inverse relationship between BMI and cycle length. This is not surprising since, by deleting the last censored cycle, we differentially exclude long cycles that start near the end of the observation interval and include only short cycles that start near the end of the observation interval.

Table 1: Model Results

	Normal (α_3[1]=0.0)	Sum (α_3=1.0)	Exponential (α_3=2.0)
Covariate	*Estimate(Stderr)*	*Estimate(Stderr)*	*Estimate(Stderr)*
Intercept	1.53(0.83)	1.55(0.83)	1.55(0.83)
age = 1	0.76(1.21)	0.74(1.21)	0.74(1.22)
age = 2	-0.65(0.72)	-0.64(0.73)	-0.64(0.73)
age = 4	2.09(1.20)	2.09(1.20)	2.08(1.20)
lactating	2.11(1.16)	2.13(1.16)	2.17(1.16)
parity = yes	-1.91(0.83)	-1.93(0.84)	-1.95(0.85)
locale = far	-0.48(0.95)	-0.50(0.95)	-0.52(0.96)
BMI*(local = far)	0.22(0.54)	0.24(0.54)	0.26(0.55)
BMI*(local = near)	-0.40(0.19)	-0.41(0.19)	-0.42(0.19)
ρ	0.03(0.08)	0.03(0.08)	0.03(0.08)
σ^2	18.9	19.1	19.2

[1] α_3=skewness

The following small simulation illustrates both the bias that results from a deletion of the censored cycles and the robustness of the model to the assumption of a distribution for cycle length in the imputation. In this simulation, we follow each of 50 women for 225 days after the onset of the first cycle. Therefore, if a woman has cycles of around 28 days, we will include about 7 complete cycles and one censored cycle in the data set. There is one covariate, BMI, that decreases linearly from 22 kg/m^2 on day 1 to 20 kg/m^2 on day 195 and then increases linearly back to 21 kg/m^2 on day 225. We set the (intercept-28) to .6 days, the coefficient of (BMI-21) to -.4, σ^2 to 11.0 and ρ to .03.

Table 2: Estimators with an Empirical Standard Error

	Distribution			
Truth:	Normal	Normal	Exponential	Exponential
Imputed:	Normal	Exponential	Normal	Exponential
Intercept				
Without Censored Cycles	.57(.21)	.57(.20)	.57(.21)	.57(.21)
With Censored Cycles	.59(.20)	.59(.20)	.60(.21)	.60(.21)
BMI				
Without Censored Cycles	-.34(.19)	-.35(.18)	-.32(.19)	-.32(.18)
With Censored Cycles	-.41(.19)	-.42(.19)	-.40(.19)	-.40(.19)
ρ				
Without Censored Cycles	.03(.03)	.03(.03)	.03(.03)	.03(.03)
With Censored Cycles	.03(.03)	.03(.03)	.03(.03)	.03(.04)

This simulation consists of 1000 samples. In general, the analysis that deletes the censored observations produces a positively biased regression coefficient. The direction of the bias is as expected, since by deleting the censored cycle we differentially include short cycles that begin near the end of the observation interval and exclude long cycles that begin near the end of the observation interval.

As an indication of the robustness of this methodology to the assumption of the form of the conditional variance, V_{ij}, consider Table 3. As before the intercept and BMI have been centered so that the true values are .6, and -.4, respectively, and $\rho = .03$ and $\sigma^2 = 11.0$.

In this table we assume exponential errors in the imputation, but the true simulated error has a skewness of one and is a sum of a normal and an exponential. Furthermore, we misspecify the variance by assuming that the conditional variance does not change with the

Table 3: Misspecified Variance and Imputation Distribution

	Deleting Censored Cycles			Including Censored Cycles		
Covariate	*Estimate*	*Stderr*	$\widehat{Stderr}$	*Estimate*	*Stderr*	$\widehat{Stderr}$
Intercept	.57	0.198	0.205	.60	0.197	0.206
BMI	-0.33	0.186	.195	-0.40	0.188	0.201
ρ	0.033	0.033	0.034	0.029	0.034	0.033

cycle. The values under the column *Estimate* are averages of the 1000 simulations, the values under the column *Stderr* are the standard deviations of the estimator as estimated across the 1000 simulations and lastly the values under the column, $\widehat{Stderr}$, are the averages of the 1000 estimated standard deviations. Note that the biases of the estimators that utilize the censored cycles are quite low even though we misspecified both the imputation distribution and variance.

5. Discussion

We have described quantitatively the observations of menstrual cycle length in a study of fixed duration. This fixed duration results in length-biased sampling of cycles and in a censored last cycle length. We note that the length-bias arises because we observe a cycle length only if the total length of the woman's previous cycles is less than the duration of the study. To identify variables that contribute to the variability of menstrual cycles within and between women, and to accommodate the length-bias and censoring, we propose a methodology based on a conditional parameterization of the mean response (conditional on past cycle lengths and covariate values). This methodology allows for estimation of the parameters via an unbiased estimating equation even though short cycle lengths are oversampled.

As illustrated by the example in section four, the censored cycle lengths can be very informative. In addition, the deletion of the censored cycles from the estimating equation S_I, will produce biased estimating equations and subsequently biased estimators. The simulation given in the same section illustrates the bias that results from deletion of a woman's last censored cycle.

An alternate approach to the one given here is to specify a parametric distributional form for cycle length, construct a likelihood for a woman's cycles including the censored cycle, and use maximum likelihood estimators for the regression coefficients. If the assumed distribution is nonnormal then the estimators can be highly dependent upon aspects of the distribution other than the mean and variance. At this time, there does not appear to be a widely accepted distributional form for cycle length. Therefore we are reluctant to use strong parametric assumptions on the cycle length form in order to estimate parameters in the mean. However, to utilize the information in the censored cycles, we need to use a distributional form for cycle length. By using the assumed parametric distributional form only in the imputation, we minimize deleterious effects caused by an incorrect assumption.

An alternative approach is to use a nonparametric estimator of the cycle length in the imputation. This is a subject for further research. If a parametric form is used, it is important to consider several possible imputation distributions as we do here. The estimators of the regression coefficient and the variance are asymptotically consistent if we use the correct imputation distribution. However, the results of the last simulation in Section 4, indicate the robustness of the estimation procedure to an incorrect parametric form for the cycle length distribution. This is not surprising since the imputation distribution is used to garner information from the censored cycles and not from the complete cycles.

An important area for further research is the generalization of the estimating equation to a robust estimating equation. This would allow one to treat very long cycles in a similar way to cycles of average length (e.g., 22-40 days). Often, one treats long cycles as separate entities (Harlow and Zeger[9]). However, to assess those covariates that may be important contributors to greater cycle length, it would be advantageous to treat both types of cycles in a similar manner. In Section 3 we indicated very briefly how one might use Huber's ψ function to do this. Another area for further research is the assessment of covariates that, in varying within a cycle, may influence cycle length. For example, it may be of interest to assess how various hormonal measurements affect a cycle. In this case an event history analysis, as discussed by Aalen and Husebye[23] or Wei et al.[2]4, is a starting point for research.

6. Appendix

Variance Estimator

Although the form of the variance estimator given here is not identical to the form of the variance estimator given by Liang and Zeger[27], the derivation of the estimator is very similar. This derivation is based on a Taylor series in $\hat{\beta}$ about the true β. We have, $0 = S_I(\hat{\beta}) =$

$$S_I(\beta) + \frac{dS_I(\beta)}{d\beta}(\hat{\beta} - \beta)$$

plus lower order terms. Solving for $(\hat{\beta} - \beta)$ we get,

$$\sqrt{n}(\hat{\beta} - \beta) = \left(-1/n \frac{d}{d\beta} S_I(\beta)\right)^{-1} n^{-1/2} S_I(\beta).$$

In forming the derivative of $1/nS_I(\beta)$ with respect to β we use chain rule. So this derivative is a sum of terms; most of these converge to zero since the expectation of $E_\beta[Y_{ij}|obs.] - \mu_{ij}(\beta)$ conditioned on past cycle lengths and present and past covariates is zero. The only term in the derivative of $1/nS_I(\beta)$ that does not converge to zero is $1/n\dot{S}_I(\beta)$ given previously. Hence the asymptotic variance of $\sqrt{n}(\hat{\beta}-\beta)$ is given by the limit of $\left(1/n\dot{S}_I(\beta)\right)^{-1}$ $Var\left(n^{-1/2}S_I(\beta)\right)\left(1/n\dot{S}_I(\beta)\right)^{-1}$. An estimator of the $Var\left(n^{-1/2}S_I(\beta)\right)$ is given by $1/n\Sigma_I(\hat{\beta})$. Estimators of this type are discussed in Murphy and Li[25].

The asymptotic consistency of the variance estimator requires that the distribution used in the imputation is the true distribution of the cycle lengths. In addition, the above argument is valid under general conditions, if either the conditional variance, V_{ij} depends only on parameters present in β or if there is only length-biased sampling but no censoring. The model that we consider in our analysis has the additional parameter, σ^2 in the conditional variance and there is censoring. In the imputation for the censored cycles, the conditional expectations involve an estimator of σ^2. As a result the asymptotic variance of $\hat{\beta}$ will depend on the asymptotic normal distribution of the estimator of σ^2. Denote the estimating equation in (2) by $S_{\sigma^2}(\beta,\sigma^2)$ and concatenate S_{σ^2} to S_I to get the vector $S(\beta,\sigma^2)$. The dimension of $S(\beta,\sigma^2)$ is equal to the number of covariates plus one for the intercept, plus one for ρ and plus one for σ^2. To derive an estimator for the asymptotic variance use a Taylor series for $0=S(\hat{\beta},\hat{\sigma}^2)$, in $(\hat{\beta},\hat{\sigma}^2)$ about the true value, (β,σ^2). As in the above we get

$$\sqrt{n}\begin{pmatrix} \hat{\beta}-\beta \\ \hat{\sigma}^2-\sigma^2 \end{pmatrix} = \left(-1/n\frac{d}{d(\beta,\sigma^2)^T}S(\beta,\sigma^2)\right)^{-1} n^{-1/2}S(\beta,\sigma^2)$$

plus terms of lower order. The asymptotic variance of $\left(\sqrt{n}(\hat{\beta}-\beta),\sqrt{n}(\hat{\sigma}^2-\sigma^2)\right)$ is given by the limit of

$$\left(-1/n\frac{d}{d(\beta,\sigma^2)^T}S(\beta,\sigma^2)\right)^{-1} Var\left(n^{-1/2}S(\beta,\sigma^2)\right)\left(-1/n\frac{d}{d(\beta,\sigma^2)^T}S(\beta,\sigma^2)\right)^{-1}.$$

An estimator of the $Var\left(n^{-1/2}S(\beta,\sigma^2)\right)$ is,

$$\begin{pmatrix} \Sigma_{\hat{\beta}\hat{\beta}} & \Sigma_{\hat{\beta}\hat{\sigma}} \\ \Sigma_{\hat{\beta}\hat{\sigma}}^T & \Sigma_{\hat{\sigma}\hat{\sigma}} \end{pmatrix}$$

where $\Sigma_{\beta\beta}$ is Σ_I from before,

$$\Sigma_{\beta\sigma} = 1/n \sum_{i=1}^{n} \sum_{j \geq 1} \dot{\mu}_{ij}(\beta) V_{ij}^{-1}(\beta) \Big(E_\beta[Y_{ij}|obs.] - \mu_{ij}(\beta) \Big) \left(\frac{E_\beta \left[(Y_{ij} - \mu_{ij}(\beta))^2 |obs. \right]}{V_{ij}(\beta)} - \sigma^2 \right) \delta_{ij},$$

and

$$\Sigma_{\sigma\sigma} = 1/n \sum_{i=1}^{n} \sum_{j \geq 1} \left(\frac{E_\beta \left[(Y_{ij} - \mu_{ij}(\beta))^2 |obs. \right]}{V_{ij}(\beta)} - \sigma^2 \right)^2 \delta_{ij}.$$

These two equations look more complicated than they are; for example, recall that $E_\beta \left[(Y_{ij} - \mu_{ij}(\beta))^2 |obs. \right] = (Y_{ij} - \mu_{ij}(\beta))^2$ for all but the censored cycles.

Computational Method

We consider the parametrization, (1). To solve for $\hat{\beta}$ in $S_I(\hat{\beta}) = 0$, we use two iteration loops. Iteration loop, M, consists of calculating the estimator $\hat{\beta}$, for a given set of data. Iteration loop, E, revises the data; that is, it carries out the imputation for the censored cycles using the most recent values of $\hat{\beta}$. We cycle between M and E until $S_I(\hat{\beta})$ and the relative change in $\hat{\beta}$ are close to zero.

Before beginning the iteration loops we simulate a very large number (e.g. 30,000) of replicates from the chosen imputation distribution. By appropriate subtraction and division we standardize the replicates so as to have mean zero and variance one. These replicates will be used in loop E.

In iteration loop, M, we have a given set of cycles. If we have not been through E, we use each woman's cycles except for the last censored cycle. After E, we use each woman's cycles plus the imputed values of the last cycles. Note that S_I is a vector with dimension equal to the dimension of γ plus one (for ρ). Partition S_I into $S_{I\gamma}$ and $S_{I\rho}$. We consider a grid of values for ρ, say $\rho_1, \ldots, \rho_k$. For each ρ_j, we solve $S_{I\gamma}(\gamma_j, \rho_j) = 0$ for γ_j. This turns out to be weighted least squares with weight matrix equal to a block diagonal matrix; each block corresponding to a woman. Each block is the inverse of a matrix with one on the diagonal and ρ_j in all of the off-diagonal entries. If $S_{I\rho}(\gamma_j, \rho_j)$ is close to zero, we set $\hat{\beta} = (\gamma_j, \rho_j)$. If necessary we make the grid for ρ finer.

In between loops M and E we estimate σ^2. The values of γ and ρ are fixed at the results of loop M. From (2), $\hat{\sigma}^2$ is given by

$$\sum_{i=1}^{n} \sum_{j \geq 1} \left(\frac{(Y_{ij} - \mu_{ij}(\hat{\beta}))^2}{\sqrt{\left(1 + \frac{\hat{\rho}}{\hat{\rho}(j-1)+1-\hat{\rho}}\right)}} \right) \delta_{ij}$$

divided by the number of cycles in the data. If we have not yet been through loop E then the data consists of each woman's cycles except for the last censored cycle. Otherwise we replace the value of $(Y_{ij} - \mu_{ij}(\hat{\beta}))^2$ for each last censored cycle by the estimated conditional variance calculated in loop E.

In iteration loop, E, we fix the parameter values at their most recent estimated values. For each censored cycle, say Y_{ik} we calculate $\mu_{ik}(\hat{\beta})$ and $V_{ik}(\hat{\sigma}^2, \hat{\rho})$. We multiply each replicate, simulated at the beginning, by the square root of the above estimated conditional variance and to this we add the above estimated conditional mean. We call these the standardize replicates. To estimate the mean and variance of a censored cycle, Y_{ik} given that $Y_{ik} < C_i - \sum_{l=1}^{k-1} Y_{il}$ and the observed past, we calculate the mean and variance of the standardized replicates that are greater than $C_i - \sum_{l=1}^{k-1} Y_{il}$. The calculated mean becomes the imputed value of Y_{ik} and we use the calculated variance in forming $\hat{\sigma}^2$.

7. References

1. Chiazze, L., F.T. Brayer, J.J. Macisco, M.P. Parker and B.J. Duffy. "The Length and Variability of the Human Menstrual Cycle," *Journal of American Medical Association*, **203**, 377-380 (1968).

2. Treloar, A.E., R.E. Boynton, B.G. Behn, and B.W. Brown. "Variation of the Human Menstrual Cycle through Reproductive Life," *International Journal of Fertility*, **12**, 77-126 (1967).

3. Vollman, R. F. "The degree of variability of the length of the menstrual cycle in correlation with age of woman," *Gynaecologia*, **142**, 310-315 1956.

4. Knobil, E. and Hotchkiss, J. "The menstrual cycle and its neuroendocrine control," in **The Physiology of Reproduction**, ed. Knobil E, and Neill JD,s., Raven Press, New York, vol 2, 1971-1994, (1988).

5. Yen, S.S.C. "The Human Menstrual Cycle: Neuroendocrine Regulation," In **Reproductive Endocrinology.** ed. Yen, S.S.C, and Jaffe, R.B., W.B. Saunders Company, Philadelphia, 273-308 (1991).

6. Hill, P., L. Garbaczewski, J. Huskisson and E.L. Wynder. "Diet, Lifestyle, and Menstrual Activity," *American Journal of Clinical Nutrition*, **33**, 1192-1198 (1980).

7. Schweiger, U., R. Laessle, H. Pfister, C. Hoehl,M. Schwingenschloegel, M. Schweiger, and K. Pirke. " Diet-Induced Menstrual Irregularities: Effects of Age and Weight Loss," *Fertility and Sterility*, **48**, 746-751 (1987).

8. World Health Organization. "A Prospective Multicentre Trial of the Ovulation Method of Natural Family Planning. III. Characteristics of the Menstrual Cycle and of the Fertile Phase," *Fertility and Sterility*, **40**, 773-778 (1983).

9. Harlow, S.D., and S.L. Zeger. "An Application of Longitudinal Methods to the Analysis of Menstrual Diary Data," *Journal of Clinical Epidemiology*, **44**, 1015-1025 (1991).

10. Harlow, S.D., and G.M. Matanoski. "The Association between Weight, Physical Activity, and Stress and Variation in the Length of the Menstrual Cycle," *American Journal of Epidemiology*, **133**, 38-49 (1991).

11. Münster K., L. Schmidt and P. Helm. "Length and Variation in the Menstrual Cycle-A Cross-Sectional Study from a Danish County." *British Journal of Obstetrics and Gynaecology*, **99**, 422-429 (1992).

12. Belsley, E.M. and T.M. Farley. "The Analysis of Menstrual Bleeding Patterns: A Review," *Applied Stochastic Models and Data Analysis*, **3**, 125-150 (1987).

13. Burch, T.K., J. Macisco, and M.P. Parker. "Some Methodologic Problems in the Analysis of Menstrual Data," *International Journal of Fertility*, **12**, 67-76 (1967).

14. Snowden, R. "The statistical analysis of menstrual bleeding patterns," *Journal of Biosocial Science*, **9**, 107-120 (1977).

15. Jones, Y., J.T. Judd, W.S. Cambell, P. Nair. "Influence of dietary fat on menstrual cycle and menses length," *Human Nutrition: Clinical Nutrition*, **41**, 341-345 (1987).

16. Pirke, K.M., U. Schweiger, W. Lemmel, J.C. Krieg, and M. Berger. "The influence on the menstrual cycle of healthy young women," *Journal of Clinical Endocrinology and Metabolism*, **60**, 1174-1179 (1985).

17. Bentley, G.R., A. M. Harrigan and P. T. Ellison. "Ovarian Cycle Length and Days of Menstruation of Lese Horticulturalists," *Paper presented at the 59th Annual Meeting of the American Association of Physical Anthropologists, Miami, Florida.* (1990).

18. Bentley, G.R., A. M. Harrigan and P. T. Ellison. "Ecological Influences on Ovarian Patterns among Lese Horticulturalists of the Ituri Forest, Zaire," *Paper presented at the Annual Meeting of the Population Association of America, Denver, Colorado.* (1992).

19. Belsey, E.M. and N. Carlson. "The Description of Menstrual Bleeding Patterns: Towards Fewer Measures," *Statistics in Medicine* **10** 267-284 (1991).

20. Little, R.J. and D.B. Rubin. **Statistical Analysis with Missing Data**, Wiley, New York 1987.

21. Ross S.M. **Introduction to Probability Models**, 5th edition, Academic Press, Inc., San Diego (1993).

22. Bullen, B.B., G.S. Skrinar, I.Z. Beitins, G. von Mering, B.A. Turnbull, and J.W. McArthur. "Induction of Menstrual Disorders by Strenuous Exercise in Untrained Women," *New England Journal of Medicine*, **312**, 1349-1353 (1985).

23. Aalen, O.O., and E. Husebye. "Statistical Analysis of Repeated Events Forming Renewal Processes," *Statistics in Medicine*, **10**, 1227-1240 (1991).

24. Wei, L.J. D.Y. Lin and L. Weissfeld. "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions," *Journal of the American Statistical Association*, **84**, 1065-1073.

25. Murphy, S.A. and B. Li. "Projected Partial Likelihood and its Application to Longitudinal Data," *submitted for publication* (1993).

26. Zeger S.L. and K. Liang. "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, **42**, 121-130 (1986).

27. Liang K. and S.L. Zeger. "Longitudinal Data Analysis using Generalized Linear Models," *Biometrika*, **73**, 13-22 (1986).

28. Ruppert D. "M-Estimators," in **Encyclopedia of Statistical Sciences**. vol. 5, S. Kotz and N. Johnson,(Eds.) Wiley, New York (1985).

29. Dempster, A.P., N.M. Laird, and D.B. Rubin. "Maximum Likelihood Estimation from Incomplete Data via the E-M Algorithm (with discussion)," *Journal of the Royal Statistical Society, Series B*, **39**, 1-38 (1977).

30. Ellison, P.T. "Human salivary steroids: Methodological considerations and applications in physical anthropology," *Yearbook of Physical Anthropology*, **31**, 115-142 (1988).

31. Ellison, P.T., Peacock, N. and Lager, C. "Salivary progesterone and luteal function in two low-fertility populations of northeast Zaire," *Human Biology*, **58**,473-483 (1986).

32. Ellison, P.T., Peacock, N. and Lager, C. "Ecology and ovarian function among Lese women In the Ituri Forest, Zaire," *American Journal of Physical Anthropology* **78**, 519-526 (1989).

33. Bentley, G.R., A.M. Harrigan, S. Murphy, M.A. O'Hanesian, and P.T. Ellison. "Menstrual patterns among Lese women of the Ituri Forest, Zaire," *in progress*.

34. Bailey, R.C., M.R. Jenike, P.T. Ellison, G.R. Bentley, A. M. Harrigan, and N. Peacock. "The Ecology of Birth Seasonality in the Ituri Forest of Zaire," *Journal of Biosocial Science*, **24**, 393-412 (1992).

35. Schweiger U. "Menstrual Function and Luteal-Phase Deficiency in Relation to Weight Changes and Dieting," *Clinical Obstetrics and Gynecology*, **34**, 191-197 (1991).

36. Cliff, A.D. and J.K. Ord. **Spatial Autocorrelation**, Pion Limited, London 1973.

ACKNOWLEDGEMENTS

We would like to thank Alisa Harrigan for her help in collecting the field data on which this study was based, Peter Ellison and his laboratory for the endocrinological analyses verifying the menstrual patterns reported here, as well as Steven Arnold and two anonymous reviewers for their perspicacious comments. This research was supported by an NSF grant DMS-93-07255 to S.A. Murphy, a National Institute of Aging Postdoctoral Fellowship to G.R. Bentley, and an NSF grant BNS-88-10931 to P.T. Ellison.