Assessing Sexual Attitudes and Behaviors of Young Women: A Joint Model with Nonlinear Time Effects, Time Varying Covariates, and Dropouts

Pulak Ghosh and **Wanzhu Tu**

Pulak Ghosh is Assistant Professor, Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303 (Email: pghosh@mathstat.gsu.edu).Wanzhu Tu is Associate Professor, Division of Biostatistics, Indiana University School of Medicine, 410 West 10th Street, Suite 3000, Indianapolis IN 46202-3002; He is also Research Scientist at Regenstrief Institute, Inc., (Email: wtu1@iupui.edu)

Abstract

Understanding human sexual behaviors is essential for the effective prevention of sexually transmitted infections. Analysis of longitudinally measured sexual behavioral data, however, is often complicated by zero-inflation of event counts, nonlinear time trend, time-varying covariates, and informative dropouts. Ignoring these complicating factors could undermine the validity of the study findings. In this paper, we put forth a unified joint modeling structure that accommodates these features of the data. Specifically, we propose a pair of simultaneous models for the zero-inflated event counts: Each of these models contains an auto-regressive structure for the accommodation of the effect of recent event history, and a nonparametric component for the modeling of nonlinear time effect. Informative dropout and time varying covariates are modeled explicitly in the process. Model fitting and parameter estimation are carried out in a Bayesian paradigm by the use of a Markov Chain Monte Carlo (MCMC) method. Analytical results showed that adolescent sexual behaviors tended to evolve nonlinearly over time and they were strongly influenced by the day-to-day variations in mood and sexual interests. These findings suggest that adolescent sex is to a large extent driven by intrinsic factors rather than being compelled by circumstances, thus highlighting the need of education on self protective measures against infection risks.

Keywords

Joint modeling; Markov Chain Monte Carlo; Mood; Sexually transmitted infections; Zero-inflated Poisson

1 Introduction

Human sexual contacts are the primary pathway for sexually transmitted pathogens such as *Chlamydia trachomatis* (CT), *Neisseria gonorrhoeae* (NG), and *Trichomonas vaginalis* (TV). Despite the existence of efficacious antimicrobial agents against the organisms, these diseases remain prevalent in the US population. The burden of the diseases is disproportional on adolescents and young adults. For example, although young people aged 15 to 24 account for only a quarter of the sexually active population, they represent nearly half of new infections (Weinstock, Berman, and Cates 2004). Since the diseases are transmitted through behavior, the development of effective prevention strategies requires an improved understanding of human sexual behaviors, particular those of the young. Because adolescent sexual behaviors tend to change with time and experience, and are likely influenced by proximal phenomena such as mood and sexual interest, it is essential to model behavioral events longitudinally with

full consideration of the contextual information. But the analysis of longitudinal behavioral data collected from observational studies is often complicated by potentially nonlinear time effects, large between-subject variability, time-varying covariates, and informative dropouts. This paper presents a unified analytical framework for the modeling of longitudinally collected counts of human sexual events, with explicit accommodation of these various complications.

1.1 An Epidemiological Study of Sexual Behaviors of Young Women

Young women were recruited for participation in a behavioral epidemiological study from three urban primary care clinics. The overall objective of the study is to examine the behavioral factors related to the sexually transmitted infections (STI). Eligibility criteria included that the young woman be between the ages 14 and 17 years, be able to understand English, not have any serious psychiatric disturbances or mental handicaps, and attend one of the three recruiting clinics. These clinics serve a predominantly urban and lower income population. Individuals who did not plan to continue residence in the area for the next three months or who were pregnant were excluded from the study. At the participating sites, all women who met the enrollment criteria, regardless of prior sexual experience, were identified by clinical schedule and those who agreed to participate were enrolled at the current or subsequent clinical visit. Informed consent and parental permission were obtained at the time of enrollment.

All subjects had quarterly clinic visits for the duration of the study period. In addition to the quarterly clinic visits, the study subjects also completed daily behavioral diaries, which provided detailed records of the subject's sexual behaviors in their original time sequence. Specifically, the diary was a structured mini-survey in which the subject reported sexual intercourse, condom protection, STI symptoms, and daily mood and sexual interest. Since coitus is relatively infrequent in adolescents and may exhibit certain day-of-the-week patterns, we summarized the daily events into weekly event counts and focused on the description of weekly rate of sexual intercourse.

The original study is designed to have a total length of followup of 27 months, and it is currently ongoing. In this analysis we used a subset of 282 subjects who had been enrolled into the study for at least six months (24 weeks), including those who had dropped out of the study before the completion of the six-month interview; recent enrollees who had entered the study in the last five weeks were not considered in the current analysis. The subject characteristics that we considered in this analysis included age, lifetime number of partners, and history of STI, all measured at enrollment. STI history is thought to be a marker of the more risky sexual behaviors in young women, and the lifetime number of partners to be markers of subject's sexual experience and partner availability.

The focus of this analysis was to examine whether positive mood and sexual interest were associated with the level of sexual activity in adolescent women. In this study, positive mood was assessed via diary by asking the subject to indicate percent of time in the day that she felt *"happy"*, *"cheerful"*, and *"friendly"*. The responses were on a Likert scale ranging from *"not at all"* (1 point), *"some of it"* (2 points), *"about half"* (3 points), *"most of it"* (4 points), or *"all day"* (5 points). The responses to these three items produced a mood score between 3 and 15 points. The purpose of having these correlated items in the scale is to achieve a better representation of the unobserved underlying construct of positive mood. Similarly, daily *"sexual interest"* was measured by one item in the diary on the same five-point Likert scale. In the present research, we calculated and used the average weekly mood and sexual interest scores in the analysis.

1.2 Analytical Issues of Longitudinally Measured Behavioral Data

The primary response variable of interest of this analysis was the weekly number of coital events. The weekly coital frequency counts of the study cohort are depicted in Figure 1. These counts ranged from 0 to 8 with more than half of the observations being zero. When the number of zeros in the data set exceeds the probability mass that the Poisson distribution allocates to the point of zero, the data are said to be zero-inflated. In the presence of zero inflation, modeling the event count via Poisson regression is no longer appropriate; instead, zero-inflated Poisson (ZIP) regression models are often used in place of the classical Poisson regression models (Lambert 1992). Since Lambert's seminal work on ZIP regression models, a variety of applied ZIP regression models have been successfully used in several important clinical applications (Böhning, Dietz, Schlattmann, Mendonca, and Kirchner, 1999;Yau and Lee, 2001;Cheung, 2002;Lu, Lin, and Shih, 2004,Ghosh, Mukhopadhyay, and Liu, 2006). However, a number of methodological issues have complicated the analysis of the study data.

1. Repeatedly measured event counts contributed by the same subject tend to be correlated, and current event count is likely to be associated with past event counts. Hall and Zhang (2004) discussed model-fitting procedures for marginal ZIP regression models for clustered count data. Min and Agresti (2005) and Hall and Wang (2005) considered mixed-effects ZIP regression models for repeatedly measured or cluster correlated data. But in this application, since the levels of sexual activities vary significantly from person to person, and the event count of the current week is likely to depend on those of the previous weeks, a more natural approach is to develop an autoregressive model where the outcome at current time point depends on its value at previous time points (Diggle *et al*, 2002, Chapter 10). Previous investigation has indicated the validity of this structure in the context of sexual behavioral research (Fortenberry *et al*, 2005).

2. Human behaviors often change gradually over time. Such time effects are typically unobservable but ignoring them could have serious consequences. Additionally, repeated behavioral assessments themselves could have a subtle but non-ignorable impact on the behavior being studied. One of the concerns here is that repeated questioning about sexual activities could "activate" the subject, thus gradually influence her behavior. Methodologically, it is often impossible to independently verify the existence of such activation effects because of the lack of appropriate control subjects who are not subjected to these assessments. The dilemma is that, had there been such a control group, it would produce no behavioral data for the comparison due to the lack of assessments! But, if the data collection instruments have an activating effect, the effect is likely to express itself over time. For this reason, it becomes critically important for us to account for the time effect explicitly in the model to ensure the validity of inference on the important independent variables. This said, we note two difficulties with the modeling of the time effect: First, there is no guarantee that the time effect is linear; indeed, since the true functional form of time covariate is unknown, the assumption of linear time effect may not be always justifiable. Second, the time effect could differ by individual subjects. Examining the event profiles of 20 randomly chosen subjects (Figure 2), it becomes clear that they do not correspond to a particular parametric form, and between-subject variation is quite evident. These considerations have led us to consider a semiparametric approach for the cohort time effect using spline models. We use this approach to explore the features of the population and individual curves within the mixed model framework.

3. The observation of longitudinal cohorts are often accompanied by dropouts. The probability of a subject's dropping out of a study may be related to the subject's self-reported event rate. In the context of adolescent sexual behaviors, some have suspected that dropout is often a marker of higher risk behavior. A missing data

mechanism where a subject's probability of dropping out depends on the rate of the Poisson process was referred to as "informative censoring" by Wu and Carroll (1988) and "informative missingness" by Follman and Wu (1995). We use the term "informative dropout" to describe this situation. In this application, we use a logistic model to depict the drop-out probability as a function of the subject's baseline characteristics, her observed event counts before the dropout time, and a random subject effect. In this example, we had about 20% of subjects that had dropped out of the study during the course of follow-up.

4. Because mood and sexual interest are collected over time together with coital counts, they can be viewed not only as time varying covariates, but also as realizations of some underlying psychological processes (Fortenberry et al., 2006). Therefore, directly modeling them may enhance our understanding of these effects. For example, an explicit covariate model will allow us to assess the strength of correlations of the mood and sexual interest measures within the subject over time. In addition, these time-varying covariates are not observed at the time of dropout leading to missingness in the covariates. Roy and Lin (2005) has shown that ignoring this missingness in the covariates will yield inconsistent estimates of the model parameters. Thus we directly model the time-varying covariates using a mixed model framework.

It should also be noted that in behavioral data, these complications rarely appear in isolation. Therefore in the present paper, we propose a joint modeling approach for the sexual activity data. Specifically, the behavioral events are modeled by an autoregressive ZIP regression structure with a semiparametric component for the accommodation of a potentially nonlinear time effect. The ZIP regression models share the subject-specific random effect with the logistic model for dropout. Time varying covariates such as mood and sexual interest are modeled via linear mixed model with autoregressive structures. The model is thus semiparametric in nature as the time effect is modeled nonlinearly. Within this unified modeling framework, a Bayesian approach was developed for parameter estimation. As an applied statistical method, this work is in contrast to the previous approaches by providing a joint modeling structure for the depiction of behavioral events in a longitudinal study. This research has incorporated some of the more recent modeling techniques in a ZIP regression set up: (1) it simultaneously models the probability weights of the mixture distributions; (2) it incorporates semiparametric functions for the time effects; (3) it explicitly accounts for informative dropouts; and (4) it accommodates the missing covariates. Since these characteristics are not uncommon in longitudinal studies of behavioral outcomes, the method is potentially useful for a wider class of applications.

2 Model Specification

2.1 Zero-inflated Poisson Model

Let Y_{ij} be the count of behavioral events reported by the ith subject in the jth time unit, $i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$, where m represents the number of subjects in the study, and n is the designed number of time units in the follow-up period. In the context of this research, Y_{ij} is the number of sexual episodes reported by the ith subject in the jth week. Depending on the subject's current state of sexual activeness, a large number of zeros may be observed in Y. Following Lambert (1992), Hall (2000), Dagne (2004) and Ghosh, Mukhopadhyay, and Liu (2006), we further assume that for each observed event count, Y_{ij}, there is an unobserved random variable for the state of sexual activeness, U_{ij}, where $P(U_{ij} = 0) = p_{ij}$ if Y_{ij} comes from the degenerate distribution, and $P(U_{ij} = 1) = 1 - p_{ij}$ if $Y_{ij} \sim$ Poisson λ_{ij}:

$$Y_{ij}= \begin{cases} 0 & \text{with probability } p_{ij} \\ \text{Poisson}(\lambda_{ij}), & \text{with probability } (1-p_{ij}), \end{cases} \tag{1}$$

where Poisson(λ_{ij}) is defined by density function $P(Y_{ij}=y_{ij})=\exp(-\lambda ij)\lambda_{ij}^{y_{ij}}/y_{ij}!$.

It should be noted that both the degenerate distribution and the Poisson process can produce zero observations. Such a formulation is often referred to as the zero-inflated Poisson (ZIP) distribution. It then follows that

$$\Pr(Y_{ij}=0) = p_{ij}+(1-p_{ij})\exp(-\lambda_{ij}) \tag{2}$$

$$\Pr(Y_{ij}=y_{ij}) = (1-p_{ij})\frac{\exp(-\lambda_{ij})\lambda_{ij}^{y_{ij}}}{y_{ij}!}, y_{ij}=1,2,\ldots. \tag{3}$$

In this research, one could conceptualize the degenerate distribution as representing a "sexually inactive" state with probability p_{ij}, while the Poisson process represents a "sexually active" state with λ_{ij} being the mean weekly number of sexual episodes.

Because the weekly event counts are simultaneously influenced by the state that the subject is in during the week and the weekly event rate given she is in an active state, we consider simultaneous modeling of both λ_{ij} and p_{ij}.

2.2 Simultaneous Models of Behavioral Event Counts

We assume the following logistic and log-linear regression models for p_{ij} and λ_{ij}:

$$\begin{aligned} \text{logit}(1-p_{ij})= \ & S_i^T\beta_1^p+T_{ij}^T\beta_2^p+\sum_{q=1}^{Q}\beta_{3q}^p Y_{i,j-q}+Z_{ij2}^T b_{i1} \\ & +f^p(t_{ij})+h_i^p(t_{ij}); \end{aligned} \tag{4}$$

$$\begin{aligned} \log(\lambda_{ij})= \ & S_i^T\beta_1^\lambda+T_{ij}^T\beta_2^\lambda+\sum_{q=1}^{Q}\beta_{3q}^\lambda Y_{i,j-q}+Z_{ij1}^T b_{i2} \\ & + f^\lambda(t_{ij})+h_i^\lambda(t_{ij}). \end{aligned} \tag{5}$$

The logistic model in (4) explicitly depicts the probability that the observation is from the degenerate distribution; and the loglinear model in (5) quantifies the "intensity" of the Poisson process. Herein, S_i denotes the baseline characteristics and T_{ij} is the time-varying covariate vector for the ith subject at time j. Although we assumed the same set of covariates for the p_{ij} and λ_{ij} in the above formulation, the models can easily be modified to accommodate different covariates in the two processes. The parameters, $\beta_1^p, \beta_2^p, \beta_1^\lambda, \beta_2^\lambda$ are vectors of regression coefficients for the fixed effects. Note that in (4) and (5), the subject's response at time j, Y_{ij}, depends on the subject's past events through embedded qth order autoregressive structures. Parameters $\beta_3^p=(\beta_{31}^p,\ldots\beta_{3Q}^p)$ and $\beta_3^\lambda=(\beta_{31}^\lambda,\ldots,\beta_{3Q}^\lambda)$ are associated with the autoregressive

process. In this application, the reported number of sexual episodes may vary from week to week in an unknown fashion. Thus, the time effects on p_{ij} and λ_{ij} are modeled by unspecified nonparametric functions $f^p(t)$ and $f^\lambda(t)$, respectively. These unspecified smooth functions reflect the nonlinear effect of the time. However, these functions represent only the population averages; individual trajectories may still vary from subject to subject and the individual pattern may not follow the pattern of the population curve. These subject effects may also contribute to the correlation of the longitudinal measurements within subjects. Therefore we add a subject specific nonparametric function $h_i(\cdot)$, which represents the subject's deviation from the group curves. The population curves $f(t)$ is important as it describes the overall cohort time effect on the parameter of interest. At the same time, individual curves $h_i(t)$ are introduced to represent subject-specific variations around the population time effect. Together, the population average and individual specific curves serve to improve the fitness of the model to inform the investigators about the nature of the cohort time effect. To accommodate any extra within-subject correlation due to the large within-subject variability in the cohort, we introduce additional random effects $(\mathbf{b}_{i1}, \mathbf{b}_{i2})$ into the models.

He, Fung and Zhu (2005), and Zhao, Staudenmayer, Coull, and Wand (2006) discussed the incorporation of semiparametric population curves in generalized linear models. This research further extends those methods by embedding both population average and subject-specific splines in a ZIP regression model. In doing so, the proposed semiparametric ZIP model offers a greater flexibility in the modeling of zero-inflated event counts. The model reduces to a parametric ZIP model when $f^p(t)$, $f^\lambda(t)$, $h_i^p(t)$, and $h_i^p(t)$ are constants. Following (Ruppert et al. 2003), we assume that the spline functions take the following general forms of a piecewise polynomial of degree τ:

$$f^p(t) = v_1^p t + v_2^p t^2 + \cdots + v_\tau^p t^\tau + \sum_{d=1}^{D_1} u_{d1}^p (t - \kappa_{d1}^p)_+^\tau;$$

$$f^\lambda(t) = v_1^\lambda t + v_2^\lambda t^2 + \cdots + v_\tau^\lambda t^\tau + \sum_{d=1}^{D_1} u_{d1}^\lambda (t - \kappa_{d1}^\lambda)_+^\tau;$$

$$h_i^p(t) = \rho_{1i}^P t + \rho_{2i}^P t^2 + \cdots + \rho_{\tau i}^P t^\tau + \sum_{d=1}^{D_2} u_{id2}^p (t - \kappa_{d2}^P)_+^\tau;$$

$$h_i^\lambda(t) = \rho_{1i}^\lambda t + \rho_{2i}^\lambda t^2 + \cdots + \rho_{\tau i}^\lambda t^\tau + \sum_{d=1}^{D_2} u_{id2}^\lambda (t - \kappa_{d2}^\lambda)_+^\tau;$$

where, $X_+ = x$ if $x > 0$; and 0 otherwise, and k_{d1}^P, k_{d1}^λ, k_{d2}^P, k_{d2}^λs are the known knot points. The choice of the knots will be described in the data analysis section. Note that in the population spline we do not have any intercept to avoid unidentifiability. We assume $u_{d1}^p \sim N(0, \sigma_{up}^2), u_{d1}^\lambda \sim N(0, \sigma_{u\lambda}^2), u_{id2}^p \sim N(0, \sigma_{1p}^2), u_{id2}^\lambda \sim N(0, \sigma_{1\lambda}^2)$. The above spline model of order τ represents adequate fits for most situations. But the number of parameters may not be practical for smaller data sets. In those situations, simpler spline models such as linear splines may be used, or subject specific splines may be dropped. Typically, linear ($\tau = 1$), quadratic ($\tau = 2$) or cubic ($\tau = 3$) splines are common choices in practice as they ensure a certain degree of smoothness in the fitted curve. The above spline models can be embedded in the mixed model framework for a general structure as follows:

Let $\mathbf{X}_{1ij} = (t, \ldots, t^\tau)^T$, $W_{ij1}^p = [(t - k_{11}^P)_+^\tau, \ldots, (t - k_{D_11}^P)_+^\tau]^T$, $\beta_4^p = (v_1^p, \ldots, v_\tau^p)^T$, $u_1^p = (u_{11}^p, \ldots, u_{D_11}^p)^T$, $\rho_i^p = (\rho_{1i}^p, \ldots, \rho_{\tau i}^p)^T$, $u_{i2}^p = (u_{i12}^p, \ldots, u_{iD_22}^p)^T$. In a similar way, we define W_{ij2}^λ, β_4^λ, u_1^λ, ρ_i^λ, u_{i2}^λ. Then,

$$f^p(t) + h_i^p(t) \quad = X_{1ij}^T \beta_4^p + W_{ij1}^{p^T} u_1^p + X_{1ij}^T \rho_i^p + Z_{ij2}^{p^T} u_{i2}^p$$
$$= X_{1ij}^T \beta_4^p + W_{ij1}^{p^T} u_1^p + V_{ij}^{p^T} w_i^p, \tag{6}$$

$$f^\lambda(t) + h_i^\lambda(t) \quad = X_{1ij}^T \beta_4^\lambda + W_{ij2}^{\lambda^T} u_1^\lambda + X_{1ij}^T \rho_i^\lambda + V_{ij2}^{\lambda^T} u_{i2}^\lambda$$
$$= X_{1ij}^{\lambda^T} \beta_4^\lambda + W_{ij2}^{\lambda^T} u_1^\lambda + V_{ij}^{\lambda^T} w_i^\lambda, \tag{7}$$

where $V_{ij}^p = (X_{1ij}^T, Z_{ij2}^{p^T})$, $w_i^p = (\rho_i^p, u_{i2}^p)^T$. Similarly, V_{ij}^λ, w_i^λ is defined. Also, $E(u_1^p) = 0$, $\text{cov}(u_1^p) = \sigma_{up}^2 I_{D_1}$, $E(w_i^p) = 0$, $\text{cov}(w_i^p) = \text{diag}(\sum_\rho^p, \sigma_{1p}^2 I_{D_1})$, $E(u_1^\lambda) = 0$, $\text{cov}(u_1^\lambda) = \sigma_{u\lambda}^2 I_{D_1}$, $E(w_i^\lambda) = 0$, and $\text{cov}(w_i^\lambda) = \text{diag}(\sum_\rho^\lambda, \sigma_{1\lambda}^2 I_{D_1})$.

The splines above are partitioned into fixed linear component plus a random component, with zero expectation, representing smooth deviations about the linear trend.

Letting $X_{ij} = (S_i^T, T_{ij}^T, Y_{i,j-1}, \ldots, Y_{i,j-Q}, X_{1ij}^T)^T$, $\beta^p = (\beta_1^p, \beta_2^p, \beta_3^p, \beta_4^p)^T$, ($\beta^\lambda$ defined similarly), and plugging expressions (6-7) into equations (4-5), we get:

$$\text{logit}(1 - p_{ij}) \quad = S_i^T \beta_1^p + T_{ij}^T \beta_2^p + \sum_{q=1}^Q \beta_{3q}^p Y_{i,j-q} + X_{1ij}^{p^T} \beta_4^p + W_{ij1}^{p^T} u_1^p + V_{ij}^{p^T} w_i^p + Z_{ij1}^T b_{i1};$$
$$= X_{ij}^T \beta^p + W_{ij1}^{p^T} u_1^p + V_{ij}^{p^T} w_i^p + Z_{ij1}^T b_{i1}, \tag{8}$$

$$\log(\lambda_{ij}) \quad = S_i^T \beta_1^\lambda + T_{ij}^T \beta_2^\lambda + \sum_{q=1}^Q \beta_{3q}^\lambda Y_{i,j-q} + X_{1ij}^{\lambda^T} \beta_4^\lambda + W_{ij2}^{\lambda^T} u_1^\lambda + V_{ij}^{\lambda^T} w_i^\lambda + Z_{ij2}^T b_{i2}$$
$$= X_{ij}^T \beta^\lambda + W_{ij2}^{\lambda^T} u_1^\lambda + V_{ij}^{\lambda^T} w_i^\lambda + Z_{ij2}^T b_{i2}. \tag{9}$$

2.3 Informative dropout in Cohort Studies

Dropouts are not uncommon in observational studies of large cohorts. Here we define the dropout as someone who did not come to a scheduled visit and had not come back by the end of the study. Since the measurements are missing after the last kept visit, analysis of the incomplete data poses additional challenges. If the dropouts are due to a mechanism that is unrelated to the investigation, i.e., the unobserved behaviors are missing completely at random, these dropouts can be ignored. However, this is unlikely to be the case for most of the longitudinal studies of human behavior. In adolescent health studies, there are suspicions that the dropouts may be associated with certain traits that can be characterized as lack of discipline. These traits not only influence the dropout process, but also correlate with the sexual behaviors themselves, thus giving us an incentive for a joint modeling of the outcomes and dropout process.

Specifically, for each Y_{ij} we define a missing indicator variable R_{ij}, such that $R_{ij} = 1$ if Y_{ij} was missing, and 0 otherwise. Thus, $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in})^T$ is a vector of missing response indicators

for individual i. Then a simple model could be constructed to describe the nonignorable missing response:

$$R_{ij} \sim \text{Bernoulli}\,(\eta_{ij}) \text{ where } \eta_{ij}=\Pr(R_{ij}=1|Y_{ij},Y_{i(j-1)}) \tag{10}$$

$$g(\eta_{ij})=\mathrm{L}_{ij}^T\xi+\psi_1 Y_{ij}+\sum_{s=2}^{Q_1}\psi_s Y_{i,j+1-s}+\sum_{k=1}^{K}\zeta_k T_{ijk}+Z_{ij3}^T\mathbf{b}_{i3} \tag{11}$$

where $\mathbf{Y}_{i,\,(j-1)}$ denotes a subset of the history of the data, e.g., it can be the previous responses ($y_{i,j-1}$) and/or previous time-varying covariates $T_{i,j-1,k}$. Note that $\psi_1 \neq 0$ gives nonignorable missingness. Here $g(\cdot)$ is a link function; we let $g(x) = \text{logit}(x)$. In the above model, $\mathbf{L}_{ij}$ is the vector of baseline covariates and $\mathbf{b}_{i3}$ is the vector of random subject effects corresponding to the dropout model. The T_{ijk} is the kth time-varying covariate. The unknown parameters are $(\xi, \psi_1, \ldots, \psi_{Q1}, \zeta_1, \ldots, \zeta_K)$. The baseline covariates and the time-varying covariates may be same as in the response model. The non-ignorable dropout mechanism is modeled by the dependence of the dropout probability on the unobserved outcome y_{ij} at the time of dropout and on the outcome before dropping out. As for the random subject effects vector $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i3})^T$ we assume $\boldsymbol{b}_i \sim N(\mathbf{0}, \Delta_b)$. Thus, the correlated random effect allows for the association between dropout and the outcome.

2.4 Modeling Time-varying Covariates

In the above model we have covariates which are also measured over time along with the response variable. It is usual that some of the covariates will be unobserved because of dropout in the data. Due to the presence of this missingness in the time-varying covariates, we need to model the covariate process (Roy and Lin, 2006). We develop a multivariate linear mixed model (Shah et al., 1997) to describe the covariate process.

Let T_{ijk} be the kth covariate for the ith subject measured at time j. We assume the following linear mixed model for the different time-varying covariates:

$$T_{ijk}=\mathrm{A}_{ijk}^T\gamma_{0k}+\gamma_{1k}T_{i,j-1,k}+\mathrm{B}_{ijk}^T\delta_{ik}+e_{ijk}, \tag{12}$$

where $\mathbf{A}_{ijk}$ is the design matrix for the fixed effects. The model assumes that the kth time-varying covariate at the current time depends on its value at the previous time point, $\boldsymbol{\delta}_{ik}$ is the random subject effect for the ith subject in the kth marker, and e_{ijk} is the measurement error.

Let $\boldsymbol{T}_{ij} = (T_{ij1}, \ldots, T_{ijK})^T$, $\boldsymbol{e}_{ij} = (e_{ij1}, \ldots, e_{ijk})$, $\boldsymbol{\gamma}_0 = (\gamma_{01}, \ldots, \gamma_{0K})$, $\boldsymbol{\gamma}_1 = (\gamma_{11}, \ldots, \gamma_{1K})$, $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{iK})^T$. Then in matrix notation we have,

$$\mathrm{T}_{ij}=\mathrm{A}_{ij}^T\gamma_0+\gamma_1\mathrm{T}_{i,j-1}+\mathrm{B}_{ij}^T\delta_i+\mathrm{e}_{ij} \tag{13}$$

where $\boldsymbol{e}_{ij} \sim N(\mathbf{0}, \Sigma = \text{diag}\left\{\sigma_k^2\right\})$, $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \Sigma_\delta)$, Σ_δ being the variance-covariance matrix for the random effects $\boldsymbol{\delta}_i$. We assume independence between the random effects and error distribution.

The Markovian structure of the model allows for a longitudinal correlation structure for the same covariates over time.

3 Bayesian Inference: Likelihood, Priors and Posterior

3.1 The Likelihood Function

Let $\mathbf{Y}_{obs,i} = (Y_{i1}, \ldots, Y_{in_i})^T$, and $\mathbf{T}_{obs,i} = (T_{i1}, \ldots, T_{in_i})^T$ denote the observed values of $\mathbf{Y}_i$ and $\mathbf{T}_i$, respectively. We also assume that for subjects who dropped out from the study, $Y_{drop,i} = (Y_{i,n_i+1}, \ldots, Y_{in})^T$ and $T_{drop,i} = (T_{i,n_i+1}, \ldots, T_{in})^T$ represent the missing response and covariates respectively. Then $\mathbf{Y}_i = (Y_{obs,i}^T, Y_{drop,i}^T)^T$, and $\mathbf{T}_i = (T_{obs,i}^T, T_{drop,i}^T)^T$. We define $\mathbf{S}_i$ and $\mathbf{Z}_{ij}$ similarly. Further, we write $\beta = (\beta_1^\lambda, \beta_2^\lambda, \beta_3^\lambda, \beta_1^p, \beta_2^p, \beta_3^p)^T$, $(\psi_1, \ldots, \psi_{Q1})^T$, and $\zeta = (\zeta_1, \ldots, \zeta_K)^T$.

Let $\Omega = (\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_5)$ be the parameter space. Here $\Omega_1 = (\beta, \sigma_{up}^2, \sigma_{u\lambda}^2, \sigma_{1p}^2, \sigma_{1\lambda}^2)$ is the parameter vector for the joint model, $\Omega_2 = (\xi, \psi, \zeta)^T$ is the parameter vector for the dropout model, $\Omega_3 = (\gamma = (\gamma_0, \gamma_1)^T, \sigma_1^2, \ldots, \sigma_K^2)$ is the parameter vector for the time-varying covariate model, $\Omega_4 (= \Delta_b)$ is the parameter of the random effect b_i, and $\Omega_5 (= \Sigma_\delta)$ is the parameter of the random subject effects $\mathbf{W}_i$.

Then under the assumption of nonignorable dropout ($\psi_1 \neq 0$), the joint likelihood can be written as

$$
\begin{aligned}
&L(Y_{obs,i}, T_{obs,i}, R_i \mid S_i, T_{i1}, b_i, \delta_i; \Omega) \propto \\
&L(Y_i \mid Y_{i1}, S_i, T_i, b_i; \Omega_1) L(T_i \mid T_{i1}, S_i, \delta_i; \Omega_2) \\
&\times L(R_i \mid Y_i, T_{obs,i}, S_{obs,i}, b_i; \Omega_3) L(b_i; \Omega_4) L(\delta_i; \Omega_5),
\end{aligned}
\tag{14}
$$

where

$$
\begin{aligned}
L(Y_i \mid Y_{i1}, S_i, T_i, b_i; \Omega_1) = \prod_{j=2}^{n_i} & \left[p_{ij} + (1 - p_{ij})e^{-\lambda_{ij}} \right]^{I_{[Y_{ij}=0]}} \\
& \times \left[\frac{(1-p_{ij})e^{-\lambda_{ij}} \lambda_{ij}^{Y_{ij}}}{Y_{ij}} \right]^{1-I_{[Y_{ij}=0]}},
\end{aligned}
\tag{15}
$$

with P_{ij} and λ_{ij} given in equation (4-5), and

$$
L(T_i \mid T_{i1}, S_i, \delta_i; \Omega_2) \propto \frac{1}{|\Sigma|^{n_i/2}} \exp \left\{ -\frac{1}{2} \sum_{j=2}^{n_i} (T_{ij} - \mu_{T_{ij}})^T \sum{}^{-1} (T_{ij} - \mu_{T_{ij}}) \right\},
\tag{16}
$$

where $\mu_{T_{ij}} = A_{ij}^T \gamma_0 + \gamma_1 T_{i,j-1} + B_{ij}^T \delta_i$; and

$$
\begin{aligned}
L(R_i \mid Y_i, T_{obs,i}, S_{obs,i}; \Omega_3, b_i) = \prod_{j=2}^{n_i} & \{\Pr(r_{ij}=1 \mid Y_i, T_{obs,i}, S_{obs,i}; \Omega_3, b_i)\}^{r_{ij}} \\
& \times \{1 - \Pr(r_{ij}=1 \mid Y_i, T_{obs,i}, S_{obs,i}; \Omega_3, b_i)\}^{1-r_{ij}},
\end{aligned}
\tag{17}
$$

and $L(\boldsymbol{b}_i; \Omega_4)$ and $L(\boldsymbol{\delta}_i; \Omega_5)$ denote the multivariate normal distributions with zero mean vector and variance-covariance matrices Δ_b and Σ_δ, respectively.

3.2 Prior Distribution

To complete Bayesian specification of the model, we must assign priors to the unknown parameters. Since we have no prior information from historical data or from experiment, we take the usual route and assign conjugate priors to the parameters. We assume elements of the $\Omega=(\beta,\xi,\psi,\zeta,\gamma,\Delta_b,\sum_\delta,\sigma_{up}^2,\sigma_{u\lambda}^2,\sigma_{1p}^2,\sigma_{2p}^2,\sigma_{1\lambda}^2,\sigma_{2\lambda}^2,\sigma_1^2,\ldots,\sigma_K^2)$ are independently distributed. For each fixed effect we assume a normal density prior and for the variance parameter we assume an inverse-Gamma (IG) prior, while for the variance-covariance matrix we assume an inverse Wishart prior. An inverse-Gamma prior with shape parameter a and scale parameter b is denoted by $x \sim IG(a, b)$ and is given by $f(x) \propto x^{-a}\exp(\dfrac{-b}{2x^2})$. Additionally, we assume a Wishart distribution for the inverse of a variance-covariance matrix, where a $W_q(\varrho, S)$ is a q dimensional Wishart distribution with ϱ degrees of freedom and mean ϱS^{-1}. For our analysis, diffuse priors can be chosen so that the analysis is dominated by the data likelihood. Specifically, to represent the vague prior knowledge, we propose to set the degrees of freedom for the Wishart distribution to be the minimum possible, viz., the rank of the variance-covariance matrix.

We specify the following priors on the model parameters for the fixed effects, $\pi(\beta) \sim N(\boldsymbol{\mu}_\beta, \Sigma_\beta)$, $\pi(\xi) \sim N(\mu_\xi, \Sigma_\xi)$, $\pi(\psi) \sim N(\boldsymbol{\mu}_\psi, \Sigma_\psi)$, $\pi(\zeta) \sim N(\boldsymbol{\mu}_\zeta, \Sigma_\zeta)$, $\pi(\gamma) \sim N(\boldsymbol{\mu}_\gamma, \Sigma_\gamma)$.

For the variance parameter we assume an inverse-gamma prior as follows: $\pi(\sigma_{up}^2)\sim IG(a_{up}, b_{up})$, $\pi(\sigma_{u\lambda}^2)\sim IG(a_{u\lambda}, b_{u\lambda})$, $\pi(\sigma_{1p}^2) \sim IG(a_{1p}, b_{1p})$, $\pi(\sigma_{1\lambda}^2) \sim IG(a_{1\lambda}, b_{1\lambda})$, and $\pi(\sigma_k^2) \sim IG(c_k, d_k)$; $k = 1,2, \ldots, K$.

Finally, the variance-covariance parameters of the random subject effect take the following forms: $\pi(\Delta_b^{-1}) \sim \text{Wishart}(\varrho_b, S_\varrho)$, $\pi(\sum_\delta^{-1}) \sim \text{Wishart}(\varrho_\delta, S_\delta)$.

3.3 Posterior Distribution and Inference

The joint posterior distribution of the parameters of the models conditional on the data is obtained by combining the likelihood in (14) and the prior densities using Bayes' theorem:

$$
\begin{aligned}
f(\Omega,b,\delta,u|y) \quad &\propto \quad \prod_{i=1}^{m}\{L(\mathrm{y}_{\mathrm{obs},i},\mathrm{T}_{\mathrm{obs},i},\mathrm{R}_i|S_i,T_{i1},b_i,\delta_i;\Omega)\}\pi(\beta)f(\sigma_{up}^2)\pi(\sigma_{u\lambda}^2)\pi(\sigma_{1p}^2)\\
&\times \quad \pi(\sigma_{1\lambda}^2)\pi(\xi)\pi(\psi)\pi(\zeta)\pi(\gamma)\pi(\Delta_b^{-1})\pi(\textstyle\sum_\delta^{-1})\prod_{k=1}^{K}f(\sigma_k^2).
\end{aligned}
$$

The posterior distributions are analytically intractable. However, models described above can be fitted using the Markov Chain Monte Carlo (MCMC) methods such as the Gibbs sampler (Gelfand and Smith, 1990). Since the full conditional distributions are not standard, a straightforward implementation of the Gibbs sampler using standard sampling techniques may not be possible. However, sampling methods can be performed using adaptive rejection sampling (ARS; Gilks and Wild, 1992). Recently, Ghosh, Mukhopadhyay, and Liu (2006) have advocated the use of ARS for a ZIP model. In this research, we follow their procedure, which first uses a data augmentation step to sample the values of the latent variables (sexual activities) based on the current value of the parameters, and then samples the parameters using the ARS method given the latent variables. Samples were directly obtained from the joint posterior distribution of the parameters as well as the latent variables. Implementation of this method is

relatively easy in publicly available software WinBUGS (2005). The samples from the posterior obtained from the MCMC will allow us to achieve summary measures of the parameter estimates, and to obtain credible intervals of the parameters of interest. See Section 4.2 for more computational details of the data analysis.

4 Data Analysis

4.1 Model Specification

Using the proposed model, we analyzed the data collected from the behavioral epidemiological study. The data set contained weekly coital frequency counts (Y_{ij}) of 282 young women measured over a period of 24 weeks, $i = 1, 2, \ldots, 282; j = 1, 2, \ldots, 24$. The vector of baseline characteristics, $\mathbf{S}_i = (\mathrm{AGE}_i, \mathrm{STD}_i, \mathrm{PTR}_i)^T$, where AGE_i, STD_i, and PTR_i were respectively the ith subject's age, STD history, and lifetime number of sexual partners, is measured at the time of enrollment. The vector of time-varying covariates had two elements, $\mathbf{T}_{ij} = (\mathrm{MOOD}_{ij}, \mathrm{SI}_{ij})^T$, where MOOD_{ij} and SI_{ij} were respectively the weekly average mood and sexual interest scores reported by the ith subject in the jth week. Under a first order auto-regressive structure ($Q = 1$), we had the following semiparametric autoregressive ZIP models:

$$
\begin{aligned}
\mathrm{logit}(1 - p_{ij}) =\ & \beta_{11}^p + \beta_{12}^p \mathrm{AGE}_i + \beta_{13}^p \mathrm{STD}_i + \beta_{14}^p \mathrm{PTR}_i \\
& + \beta_{21}^p \mathrm{MOOD}_{ij} + \beta_{22}^p \mathrm{SI}_{ij} + \beta_{31}^p Y_{i,j-1} \\
& + b_{i1} + f^p(t_{ij}) + h_i^p(t_{ij}), \\
\log(\lambda_{ij}) =\ & \beta_{11}^\lambda + \beta_{12}^\lambda \mathrm{AGE}_i + \beta_{13}^\lambda \mathrm{STD}_i + \beta_{14}^\lambda \mathrm{PTR}_i \\
& + \beta_{21}^\lambda \mathrm{MOOD}_{ij} + \beta_{22}^\lambda \mathrm{SI}_{ij} + \beta_{31}^\lambda Y_{i,j-1} \\
& + b_{i2} + f^\lambda(t_{ij}) + h_i^\lambda(t_{ij}).
\end{aligned}
$$

Please note that for the convenience of model interpretation, we chose to model the probability that the ith subject was in a sexually active state ($1 - p_{ij}$) at jth visit in the logistic model.

For the fitting of the models, there is no clear rule on how many knot points to include or where to locate them in the spline functions. More knots are needed in regions where the function is changing rapidly (Ruppert et al. 2003). Sometimes knowledge of subject matter may be relevant in placing knots where a change in the shape of the curve is expected. Using too few knots or poorly sited knots means the approximation to the curve will be degraded. By contrast, a spline using too many knots will be imprecise. Since the subjects were assessed regularly with equally spaced intervals in this study, we selected the knots from the existing values that were equally spaced within the range [min(x), max(x)]. Thus the six knots were placed at weeks 5, 8, 11, 14, 17, 20.

A model for dropout was assumed in case the dropouts were informative. Preliminary data analysis suggested that the dropout might depend on the current or previous coital frequency counts, as well as on some of the baseline covariates. So we considered the following simple model:

$$
\mathrm{Logit}(\eta_{ij}) = \xi_1 + \xi_2 \mathrm{AGE}_i + \xi_3 \mathrm{STD}_i + \psi_1 Y_{ij} + \psi_2 Y_{i,j-1} + b_{i3}, \tag{18}
$$

where b_{i3} was the random subject effect. As detailed in Section 4.3, although the dropout probability was modeled as a function of subject's enrollment age, STD history, and the weekly coital frequency counts prior to the dropout time, we chose the above model for dropout based on a set of model selection criteria described in section 4.4.

Similarly, the time-varying covariates mood and sexual interest were modeled as follows:

$$\text{MOOD}_{ij} = \gamma_{01} + \gamma_{11}\text{MOOD}_{i,j-1} + W_{i1} + W_{i2}t_{ij} + e_{ij1},$$
$$\text{SI}_{ij} = \gamma_{02} + \gamma_{12}\text{SI}_{i,j-1} + W_{i3} + W_{i4}t_{ij} + e_{ij2}.$$

Again, the autoregressive structures embedded in the time-varying covariate models allowed us to examine the strength of the autocorrelation within the covariates. This was not only of scientific interest to the investigation, but also helpful for the exploration of the modeling structure. For example, a very strong autocorrelation in mood would not only counter the speculation of mood swing in adolescents, but also render it unnecessary to collect mood measure so frequently, or to treat it as a time-varying variable.

Since the study was still ongoing and data were still being collected at the time of this report, the currently available data set was not large enough to be divided for the purpose of prior elicitation. Prior information based on expert opinion, even if available, is nonetheless user specific. Hence, in this analysis we chose our priors to be proper but weakly informative. Specifically, we take a $N(0, 50)$ prior for each of the regression parameters and for each variance parameter (=1/precision) we use an $IG(2.01, 1.01)$ prior, giving rise to a prior mean of 1 and prior variance of 1,00. For the variance-covariance matrix Δ_b^{-1} we assumed Wishart $\left(3, \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}\right)$, and for Δ_w^{-1} we assumed Wishart $\left(2, \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}\right)$.

4.2 Computational Details

We ran two chains of the Gibbs sampler with widely dispersed initial values. The initial values for the fixed parameters were selected by starting with the prior mean and covering ±3 standard deviations. The initial values for the precision were arbitrarily selected. Initially, some evidence of poor mixing was found regarding the standard deviation of the random effect slope in the spline model. Following Zhao, Staudenmayer, Coull, and Wand (2006), we then used several other choices of the inverse gamma and the *folded-t* prior distributions for the standard deviations. The *folded-t* class of prior densities have been recommended by Gelman (2005) in a hierarchical model over the commonly used inverse-gamma distribution. The *folded-t* prior has the advantage of improving computational efficiency by reducing dependence among parameters (Liu, rubin, and Wu, 1998; Liu and Wu, 1999) and yields a Gibbs sampler that is less prone to slow mixing when the standard deviations are near zero. We found that use of a moderate to highly dispersed inverse gamma prior behaved erratically. However, the use of folded-t prior on standard deviations dramatically improved the mixing and fits were stable. Thus, we resort to the folded-t class of prior for our results. See Zhao, Staudenmayer, Coull, and Wand (2006) and Gelman (2005) for details of this prior. We also centered the covariates about mean to have better convergence. In our simulation, 25,000 samples were discarded as burn-in, and of the next 75,000 samples, we used every third value to construct the posterior estimate. Convergence was assessed visually by monitoring the dynamic traces of Gibbs iterations and by computing the Gelman-Rubin convergence statistic (Gelman and Rubin, 1992). In order to check for the sensitivity we ran the proposed model with different sets of priors and found little evidence of any prior sensitivity, although slow mixing was evident in analysis using a highly diffuse prior.

4.3 Analytical Results

Of the 282 subjects enrolled into the study 91% were African American. Enrollment age ranged from 14 to 17 with a mean of 15 years and a standard deviation of 1.1 years. Lifetime number of partners reported at the time of enrollment ranged from zero to 28 with a mean of 2.85

(median 2) and a standard deviation of 3.8. Forty four of the study subjects (15.6%) had a history of STD infection.

Table 1 reports the posterior mean, median, standard deviation (SD), and 95% credible interval (CI) for the parameters in the ZIP regression model. Similarly, the parameter estimates for the dropout and time-varying covariate models are reported in Table 2.

The ZIP regression analysis yielded a number of observations: (1) Older age was associated with increased probability of the subject being in the sexually active state (Odds ratio(OR)=exp $(\widehat{\beta^p_{12}})$= 1.99, 95% CI= [exp(0.0864), exp(1.180)]= [1.09, 3.25]), although an increase in age did not necessarily increase the weekly rate of coital frequency given the subject was in a sexually active state. (2) Baseline STD history was a strong indicator for the subject's state of sexual activity (OR=exp $(\widehat{\beta^p_{13}})$= exp(0.2583)=1.29, 95% CI=[1.10, 6.84]). A young woman that had a positive STD history at enrollment was more likely to be in the sexually active state during the study period. However, STD history did not appear to affect the rate of coital events. (3) Lifetime number of partners that the subject reported at baseline was also positively associated with the probability of being in a sexually active state (OR=exp $(\widehat{\beta^p_{14}})$=1.45, 95% CI=[1.21, 1.59]). Again, no similar effect was observed for the rate of coital frequency. (4) Lower positive mood was associated with an increased probability of being in a sexually active state (OR=exp $(\widehat{\beta^p_{21}})$=0.79, 95% CI=[0.68, 0.91]); but for a subject that was in the sexually active state, higher positive mood was associated with increased coital frequency (incident rate ratio or IRR=exp $(\widehat{\beta^\lambda_{21}})$=1.11, 95% CI=[1.06, 1.16]). (5) Higher sexual interest was associated with increased coital frequency (IRR=exp $(\widehat{\beta^\lambda_{22}})$=1.52, 95% CI=[1.13, 2.47]). (6) Coital frequency in the prior week was associated with both increased probability of being in a sexually active state (OR=exp $(\widehat{\beta^p_{31}})$=3.69, 95% CI=[2.80, 5.55]) and increased coital frequency in the current week (IRR=exp $(\widehat{\beta^\lambda_{31}})$=1.04, 95% CI=[1.02, 1.06]). (7) From Figures 3 and 4, it is evident that both the probability of being in a sexually active state and the rate of coital frequency given the subject was in an active state exhibit nonlinear time effects, and the effects vary from subject to subject. In particular, the rate parameter was monotonically increasing over time, suggesting either a developmental effect of sexuality in adolescents or a possible activation effect of repeated questionnaires. For example, Figure 3 shows that in the study cohort, the probability of a subject being in a sexually active state is not entirely monotone, but the intensity of sexual activities is steadily increasing over time. (8) Finally, we noted that the correlation was modest between random subject effects in the logistic and loglinear models (0.29), suggesting a relatively weak positive link between the subject's current state of sexual activity and her intensity or the level of activity of sexual behaviors given she was in an active state. This last observation demonstrates the usefulness of the proposed joint modeling structure in assessing the interrelationship among latent states, which may be particularly useful in the analysis of behavioral data from a variety of fields.

Similarly, from the estimates of parameters in the dropout and time-varying covariate models (Table 2), we had the following observations: (1) The dropout probability appeared to be related to the baseline covariates AGE, STD, and current and previous coital frequency values. (2) The estimates of the parameters ψ_1, ψ_2 of the dropout models were 1.977 and -0.4828 respectively, suggesting that dropout might be informative and the missing probability of y_{ij} might depend more on the current values of the coital frequency and less on the previous value. Thus, any statistical analysis that ignores the dropout may be biased. (3) Older subjects and those with an STD history were more likely to drop out, perhaps due to competing demands for time in older teens. (4) Both mood and sexual interest measures of the current week were

correlated with their corresponding values of the previous week, suggesting continuity in the adolescent mood and sexual interest.

4.4 Model Comparison

To compare candidate models, we computed $p(Y_{obs,i}, T_{obs,i}, R_i | Y_{obs,-i}, T_{obs,-i}, R_{-i})$ (Geisser and Eddy 1979), which is the posterior predictive density of $(Y_{obs,i}, T_{obs,i}, R_i)$ for subject i conditional on the observed data with a single data point deleted. This value is known as the conditional predictive ordinate (CPO; Gelfand, Dey and Chang (1992); Chen *et al*, (2000)) and has been widely used for model diagnostic and assessment.

For the ith subject, the CPO statistic under model M_l; $1 \leq l \leq L$ is defined as:

$$
\begin{aligned}
\text{CPO}_i &= p(Y_{obs,i}, T_{obs,i}, R_i | Y_{obs,-i}, T_{obs,-i}, R_{-i}) \\
&= E_{\theta_l}[p(Y_{obs,i}, T_{obs,i}, R_i | \theta_l) | Y_{obs,-i}, T_{obs,-i}, R_{-i}]
\end{aligned}
$$

where $-i$ denotes the exclusion of the data from subject i. The θ_l is the set of parameters of model M_l and $p(Y_{obs,i}, T_{obs,i}, R_i / \theta_l)$ is the sampling density of the model evaluated at the ith observation. The expectation above is taken with respect to the posterior distribution of the model parameters θ_l given the cross-validated data $(Y_{obs,-i}, T_{obs,-i}, R_{-i})$. For subject i, the CPO_i can be obtained from the MCMC samples by computing the following weighted average:

$$
\widehat{\text{CPO}}_i = \left(\frac{1}{M} \sum_{m=1}^{M} \frac{1}{f(Y_{obs,i}, T_{obs,i}, R_i | \theta_l^{(m)})} \right)^{-1}
$$

where M is the number of simulations. The θ_l^m denotes the parameter samples at the mth iteration. A large CPO value indicates a better fit. A useful summary statistic of the CPO_i is the logarithm of the pseudomarginal likelihood (LPML), defined as $\text{LPML} = \sum_{i=1}^{n} \log(\widehat{\text{CPO}}_i)$. Models with greater LPML values represent a better fit. LPML is well defined under the posterior predictive density and it is computationally stable. LPML has been used extensively in Bayesian analysis for model selection in situations of simpler and more complicated models and has a long history in the statistics literature (see Chen *et al*, (2000), Chapter 10; Brown and Ibrahim, (2003); Brown, Ibrahim, and DeGruttola (2005)).

We compared the following models using LPML:

Model 1: This is the model that we used in the analysis.

Model 2: Model 1 without the spline components in ZIP model, i.e., the splines are replaced by a linear time effect (t_{ij}).

Model 3: Independent model, i.e., the ZIP model is independent of the dropout process.

We then considered several dropout models, keeping other parts of the model unchanged:

Model 4: $\text{Logit}(\eta_{ij}) = \xi_1 + \xi_2 \text{Age}_i + \xi_3 \text{STD}_i + \psi_1 y_{ij}$

Model 5: $\text{Logit}(\eta_{ij}) = \xi_1 + \psi_1 y_{ij} + \psi_2 y_{i,j-1}$

Model6: $\text{Logit}(\eta_{ij}) = \xi_1 + \psi_1 y_{ij}$

The LPML values for Models 1-6 were -10405.7, -12198.4, -11201.8, -11086.1, -11066.9, and -11132.5, respectively. The proposed model had the highest LPML values, suggesting

that it had the best fit among the six candidate models. The large difference between the LPML values of Models 1 and 2 indicated the presence of a nonlinear time effect, and justified the use of the spline-based model for time effects in the analysis.

4.5 Simulation

In this section, we present a small simulation study to justify the relative complexity of the proposed model and to verify the performance of the model fitting procedure. We first note that the complexity of the model arises primarily from four aspects: (1) explicit modeling of autoregressive effect of the main outcome variable; (2) explicit inclusion of time-varying covariates; (3) spline-based modeling of non-linear time effects; and (4) accommodation of nonignorable dropout. While it is well known that failure to accommodate informative dropouts may lead to questionable inference (Wu and Carroll, 1988; Schluchter, 1992; Little, 1995; Roy and Lin, 2005; Wu, 2007), the impact of inattention to the first three complicating factors has not been well studied. We therefore focus on these three issues in the simulation study. Additionally, the simulation study has also given us a chance to verify the performance of our model fitting procedure.

Specifically, we consider the following model:

$$
\begin{aligned}
\operatorname{logit}(1 - p_{ij}) &= \beta_{11}^p + \beta_{13}^p X_i + \beta_{21}^p Z_{ij} + \beta_{31}^p Y_{i,j-1} + b_{i1} + f^p(t_{ij}) \\
\log(\lambda_{ij}) &= \beta_{11}^\lambda + \beta_{13}^\lambda X_i + \beta_{21}^\lambda Z_{ij} + \beta_{31}^\lambda Y_{i,j-1} + b_{i2} + f^\lambda(t_{ij}),
\end{aligned}
\tag{19}
$$

where we use $f^p(t) = \frac{1}{2}\cos^2(\frac{t+12}{12})$ and $f^\lambda(t) = 0.6\sin^2(\frac{t-3}{12})$, for $t = 1, 2, \ldots, 24$ to depict the nonlinear time effects (see Figure 5). Also in this model we consider a subject-specific covariate X_i, random intercepts $\mathbf{b}_i = (b_{i1}, b_{i2})^t$, as well as a time-varying covariate Z_{ij}, where $i = 1, 2, \ldots, 50, j = 1, 2, \ldots, 24$.

Data were generated from (19) to mimic the real data presented in the paper. Specifically, for the ith subject, we first generated X_i from a Bernoulli distribution with probability $p_0^{(x)}$. For the same subject, we then generated a 24-dimensional vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{i24})^t \sim \text{MVN}(\mu_Z, \Sigma_Z)$ to represent the values of the time-varying covariate Z_{ij} for the 24 time points. We gave Σ_z, an AR(1) variance-covariance structure to maintain a correlation between the Z values in adjacent weeks within the subject. Similarly, random intercepts $\mathbf{b}_i = (b_{i1}, b_{i2})^t$ were generated from a bivariate normal distribution. We then calculated the values of $f^p(t_{ij}) = \frac{1}{2}\cos^2(\frac{t_{ij}+12}{12})$, and $f^\lambda(t_{ij}) = 0.6\sin^2(\frac{t_{ij}-3}{12})$ at each time points. Finally, we generate the baseline value for the Y_{i1} from ZIP (p_1, λ_1). Models in (19) were then used to calculate p_{i2} and λ_{i2}. From p_{i2} and λ_{i2}, we generated $Y_{i2} \sim \text{ZIP}(p_{i2}, \lambda_{i2})$. We then repeated this last step to generate the rest of the Y values. Parameter values used in the simulation were chosen to produce data that are similar to the real data. In particular, we take $\beta_{11}^p = 0.45$, $\beta_{13}^p = 0.25$, $\beta_{21}^p = -0.21$, $\beta_{31}^p = 1.3$ and $\beta_{11}^\lambda = 0.1$, $\beta_{13}^\lambda = 0.25$, $\beta_{21}^\lambda = 0.2$, $\beta_{31}^\lambda = 0.04$. One hundred simulated data sets were used in the simulation study.

Using generated data, we fitted our proposed semiparametric ZIP regression model as well as the ZIP regression model with a linear time effect (i.e., without splines for time effect). Results are presented in Table 3. We computed the "relative bias" (RB) which is defined as the ratio of bias and the absolute value of the true parameter, mean square error (MSE) and coverage probability (CP). The numbers in parenthesis in Table 4 are the true value of the parameters.

A number of observations can be made from the simulation results: (1) that the proposed method is able to produce accurate estimates of the model parameters with minimal bias, MSE, and which have good coverage probabilities. (2) In the presence of nonlinear time effects, traditional ZIP regression models with linear time effect often produce biased estimates, larger MSEs, and substantially lower coverage probability in the time-varying covariates, although other covariates appear to be spared from such effects of the model mis-specification. Figure 5 clearly shows that proposed model is able to recover the unobserved nonlinear time effects reasonably well. (3) Finally we observed from the simulation that the proposed semiparametric regression model had larger LPML values then their parametric counterparts, suggesting improved fit of the new model. Based these observations, we contend that the models used in the analysis have good performance in the modeling of zero-inflated behavioral counts. Despite the increased complexity, the new analysis provides a safe-guard against potential effects of mis-specification of the time effects, thus preventing the occurrence of large biases in the estimation of time-varying effects.

5 Discussion

Sexually transmitted infections (STI) are spread primarily through sexual intercourse. A young woman is at risk of STI once she becomes sexually active. Yet little is known about the contextual factors that are associated with the occurrence of coitus, and the temporal patterns in which adolescent sexual behaviors evolve. This study is perhaps the first longitudinal examination of these issues based on a sizable cohort. A major strength of this investigation is the inclusion of young teens that were still in their early years of sexual experience. Other strengths of the study include the longitudinal follow-up and the extensive behavioral information that were collected in the process.

Contrary to the alarming anecdotes reported by the lay press in recent years, the findings of this paper reveal a more complicated picture: Sexual behaviors in adolescents are influenced strongly by intrinsic factors such as mood and sexual interest, rather than being driven completely by circumstances over which the teens have little control. This gives us reason to believe that more effective education and promotion of self-protective behaviors might help to reduce the risk of disease transmission. It also suggests that future prevention strategies should take into account of the emotional needs of the teens. The increasing levels of sexual activity over time are not surprising, but their individual-specific patterns seem to suggest that there are no uniformly followed patterns in terms of sexual behavioral development. Considering the age range of our study participants (14 to 17 years), we believe that intervention measures must start early to be effective.

Methodologically, the most challenging aspect in the modeling of human behavior is perhaps the incorporation of relevant contextual information. In studies of STD epidemiology and human sexuality, this contextual information often includes the concurrent mood, sexual interest, prior behavior, and subtle time effects that cannot be dismissed. Along the same line, issues such as dropout and autocorrelations existing among the time-varying covariates also complicate the analysis. These various factors form an interactive system in which the behaviors of interest are influenced by the other factors, which in turn are influenced by the observed behaviors. Therefore, a unidimensional modeling approach with a narrower focus often fails to capture the full complexity of the situation, and may produce an overly simplistic depiction of the behavior.

To address these shortcomings, we propose a new analytical framework that takes into account most of the major components in the modeling of human behaviors. Our joint model is a complicated system, but it is also necessary in order to place the behavioral event in its original context. Such an approach is likely to help investigators achieve a more comprehensive

understanding of the studied behavior. As an applied statistical tool, this method is motivated by a real epidemiological investigation. Although the data analysis that we presented in this paper is preliminary in nature due to the fact that the full data are still being collected, the initial results are promising and they have revealed some previously under-appreciated characteristics of adolescent behavior. As a result, we feel that the basic construction of the model might be appropriate for other longitudinal studies as well. Although we recognize that this is an initial step in seeking a more comprehensive solution, our effort has demonstrated the feasibility of this general strategy. Preliminary results from the simulation study have provided assurance of the modeling procedure. Additionally, it has also highlighted the potential pitfalls of using mis-specified parametric ZIP regression models.

Technically, the model employs some of the more recent developments in statistical methodology. The proposed joint model is flexible and new in several aspects: (1) it represents a semiparametric development of the ZIP model. The semiparametric approach is useful, particularly when the linearity of time effect is in question. (2) it incorporates the dropout process in the ZIP model. Without the accommodation of dropout, models may produce biased results. (3) It take into account of time-varying covariates. (4) the autocorrelation structure among behaviors and time-varying covariates. Our joint modeling approach deals with both missing response and missing covariate simultaneously and is built to borrow strength from each of the modeling components. Through a real application, we have demonstrated that the joint modeling increased the LPML values and resulted in a better fit of the data.

A few limitations of our methods must be underlined. One of the major issues is the robustness of the distributional assumption. In our application, we use a parametric normal distribution for the random effects. A broader class of distributions such as Dirichlet processes, may be a viable alternative. Another issue is the fixed knot points. Random knot points would be more flexible; however, such models will be numerically challenging in this set up. Third, the parametric dropout model may be overly simplistic. Given a sufficient number of dropouts, one can build a more complex parametric or semiparametric structure as suggested by Chen and Ibrahim (2006). It will be worthwhile to see if the complex modeling of the dropout and the use of robust random effects would improve the goodness-of-fit and change the results of the analysis. We are currently exploring these aspects of the modeling. Notwithstanding these limitations, this research has pointed to a new road map for the analysis of longitudinally measured behavioral data.

Acknowledgements

References

Böhning D, Dietz E, Schlattmann S, Mendonca L, Kirchner U. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of Royal Statistical Society: Ser A 1999;162:95209.

Brown ER, Ibrahim JG. Bayesian Approaches to Joint cure-Rate and Longitudinal Models with Applications to Cancer Vaccine Trials. Biometrics 2003;59:686–693. [PubMed: 14601770]

Brown ER, Ibrahim JG, DeGruttola V. A flexible B-spline Mdoel for multiple Longitudinal Biomarkers and Survival. Biometrics 2005;61:64–73. [PubMed: 15737079]

Chen Q, Ibrahim JG. Semiparametric Models for Missing Covariate and Response Data in Regression Models. Biometrics 2006;62:177–184. [PubMed: 16542244]

Chen, Ming-Hui; Shao, Qi-man; Ibrahim, Joseph George. Monte Carlo methods in Bayesian computation. Springer-Verlag Inc; Berlin; New York: 2000.

Cheung YB. Zero-inflated models for regression analysis of count data: a study of growth and development. Statistics in Medicine 2002;21:1461–1469. [PubMed: 12185896]

Dagne GA. Hierarchical Bayesian Analysis of Correlated Zero-inflated Count Data. Biometrical Journal 2004;6:653–663.

Diggle, P.; Heagerty, P.; Liang, KY.; Zeger, S. Analysis of Longitudinal Data. Oxford University Press; 2002.

Follmann D, Wu M. An approximate generalized linear model with random effects for informative missing data. Biometrics 1995;51:151168.

Fortenberry J, Temkit M, Tu W, Graham C, Katz B, Orr D. Daily mood, partner support, sexual interest, and sexual activity among adolescent women. Health Psychology 2005;24:252257.

Fortenberry J, Katz B, Llythe M, Juliar B, Tu W, Orr D. Factors associated with time of day of sexual activity among adolescent women. Journal of Adolescent Health 2006;38:275281.

Geisser I, Eddy W. A Predictive Approach to Model Selection. Journal of the American Statistical Association 1979;74:153–160.

Gelfand, AE.; Dey, DK.; Chang, H. Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. Bayesian Statistics. Vol. 4. Oxford: Oxford University Press; 1992. p. 147-159.

Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 1990;85:398–409.

Gelman A, Rubin D. Inference from alternative simulation using multiple sequences. Statistical Science 1992:457–472.

Ghosh SK, Mukhopadhyay P, Lu JC. Bayesian analysis of zero-inflated regression models. Journal of Statistical Planning and Inference 2006;136:1360–1375.

Hall DB. Zero-inflated Poisson and binomial regression with random effects: A case study. Biometrics 2000;56:1030–1039. [PubMed: 11129458]

Hall DB, Zhang Z. Marginal models for zero inflated clustered data. Statistical Modelling 2004;4:161–180.

He X, Fung WK, Zhu Z. Robust Estimation in Generalized Partial Linear Models for Clustered Data. Journal of the American Statistical Association 2005;100:1176–1184.

Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 1992;34:114.

Lu SE, Lin Y, Shih WJ. Analyzing excessive no changes in clinical trials with clustered data. Biometrics 2004;60:257267.

Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. Statistical Modelling 2005;5:119.

Ruppert D, Carroll RJ, Wand MP. Semiparametric Regression. 2003Cambridge Series in Statistical and Probabilistic Mathematics (No. 12

Roy J, Lin X. Missing Covariates in Longitudinal Data With Informative Dropouts: Bias Analysis and Inference. Biometrics 2005;61:837–846. [PubMed: 16135036]

Schluchter MD. Methods for the analysis of informatively censored longitudinal data. Statistics in Medicine 1992;11:1861–1870. [PubMed: 1480878]

Shah A, Laird N, Schoenfeld D. A random-effects model for multiple characteristics with possibly missing data. Journal of the American Statistical Association 1997;92:775779.

Spiegelhalter, D.; Thomas, A.; Best, N.; Lunn, D. MRC Biostatistics Unit, Institute of Public Health and Department of Epidemiology & Public Health, Imperial College School of Medicine; 2005. WinBUGS User Manual, Version 1.4. available at http://www.mrc-bsu.cam.ac.uk/bugs

Weinstock H, Berman S, Cates W Jr. Sexually transmitted diseases among American youth: Incidence and prevalence estimates, 2000. Perspectives on Sexual and Reproductive Health 2004;36:6–10. [PubMed: 14982671]

Wu L. HIV viral dynamic models with dropouts and missing covariates. Statistics in Medicine 2007;26:3342–3357. [PubMed: 17221835]

Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics 1988;44:175188.

Yau KKW, Lee AH. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. Statistics in Medicine 2001;20:2907–2920. [PubMed: 11568948]

Zhao Y, Staudenmayer J, Coull BA, Wand MP. General Design Bayesian Generalized Linear Mixed Models. Statistical Science 2006;21:35–51.
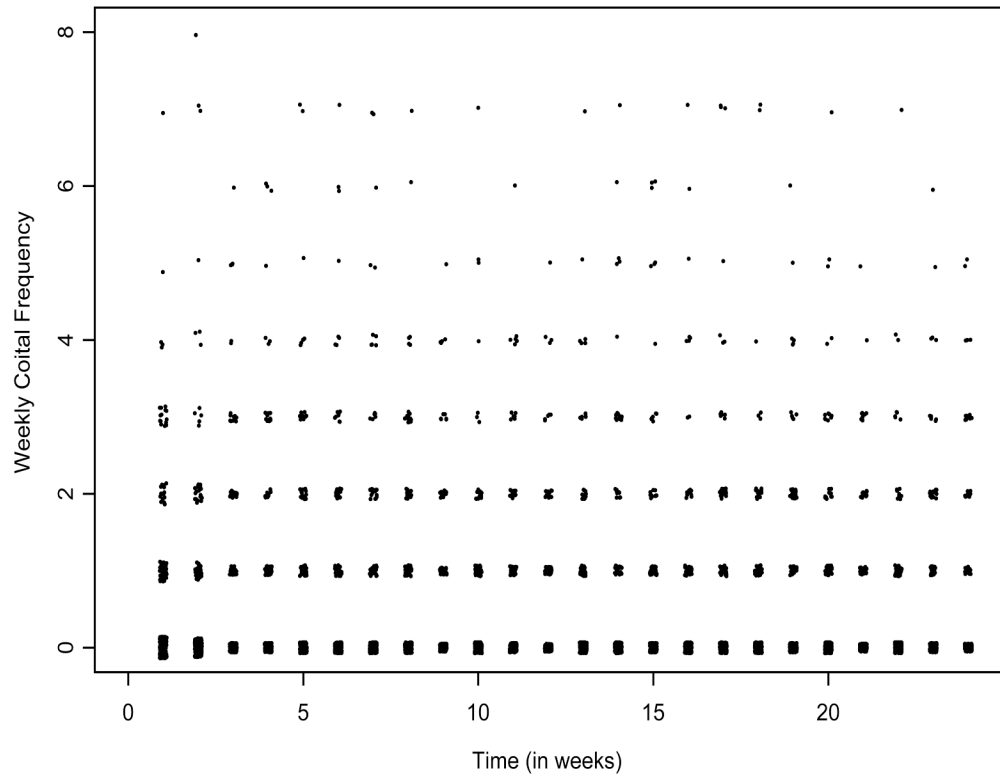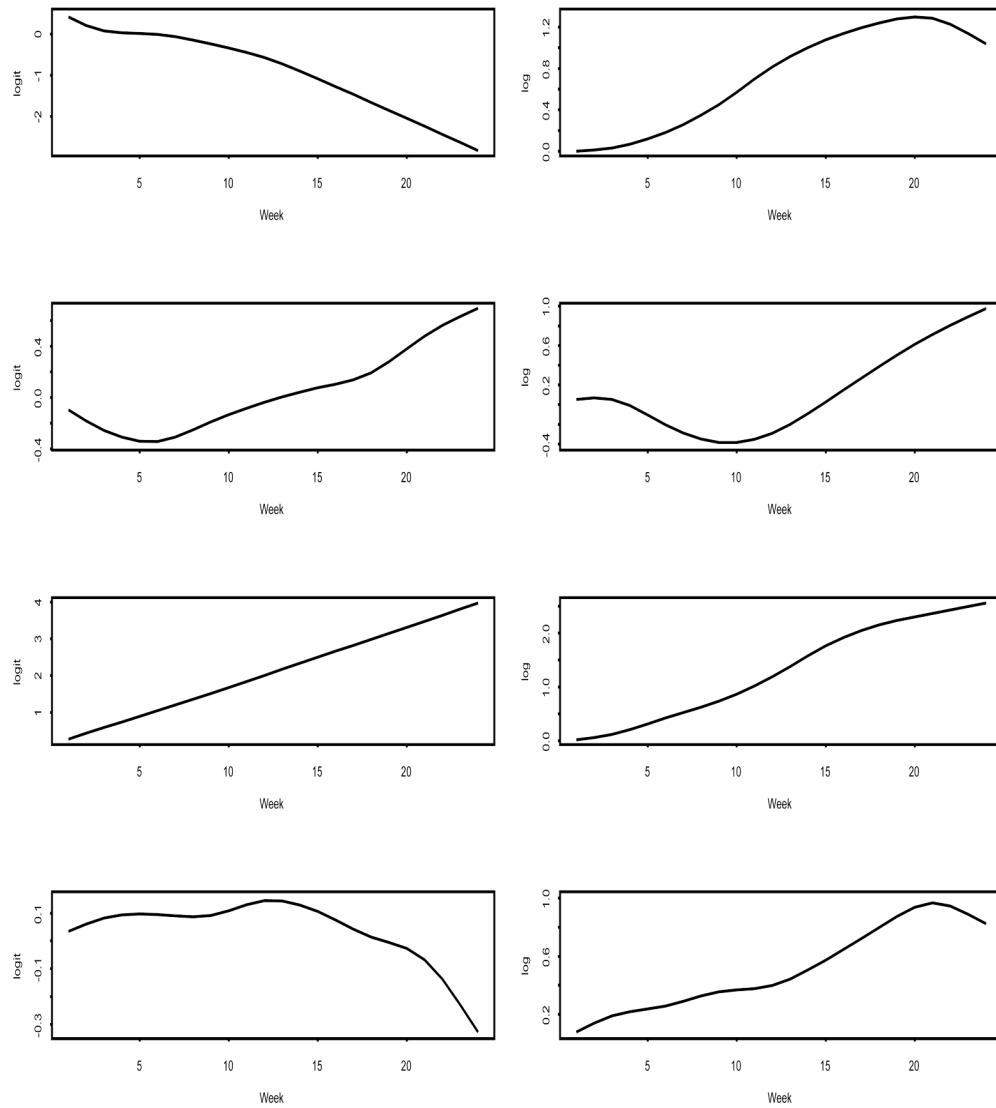
Figure 1.
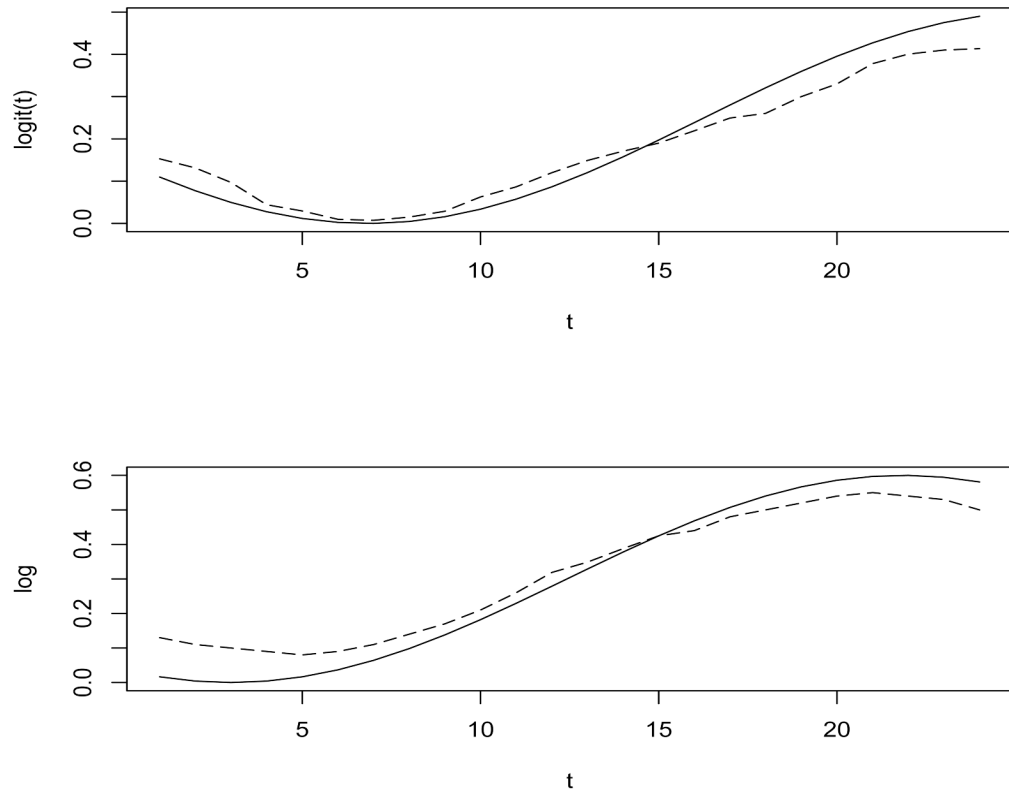Weekly coital frequency counts of 282 young women over a 24 week period.
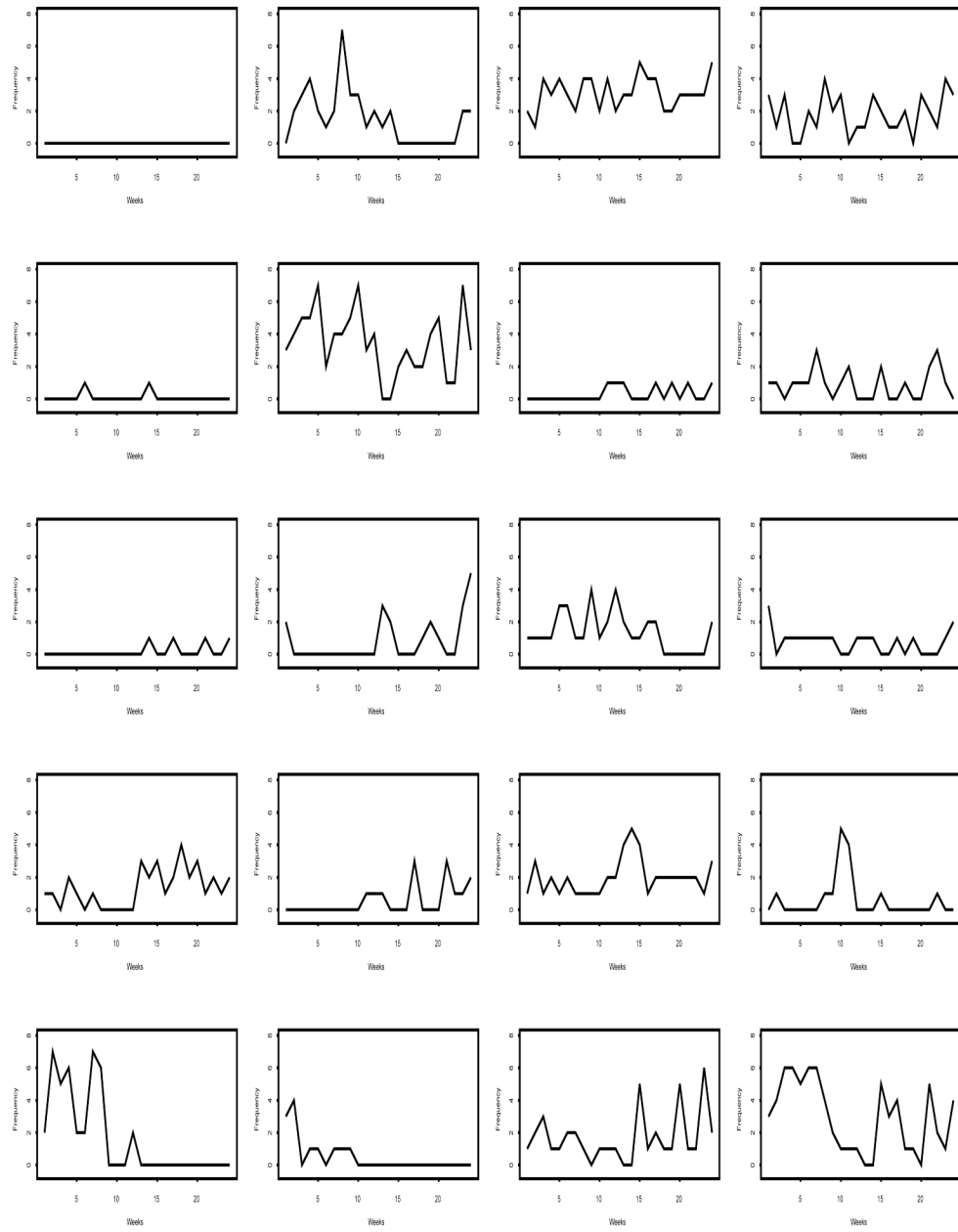
Figure 2.
Sample Longitudinal Profiles of Weekly Coital Frequency for 20 Subjects.
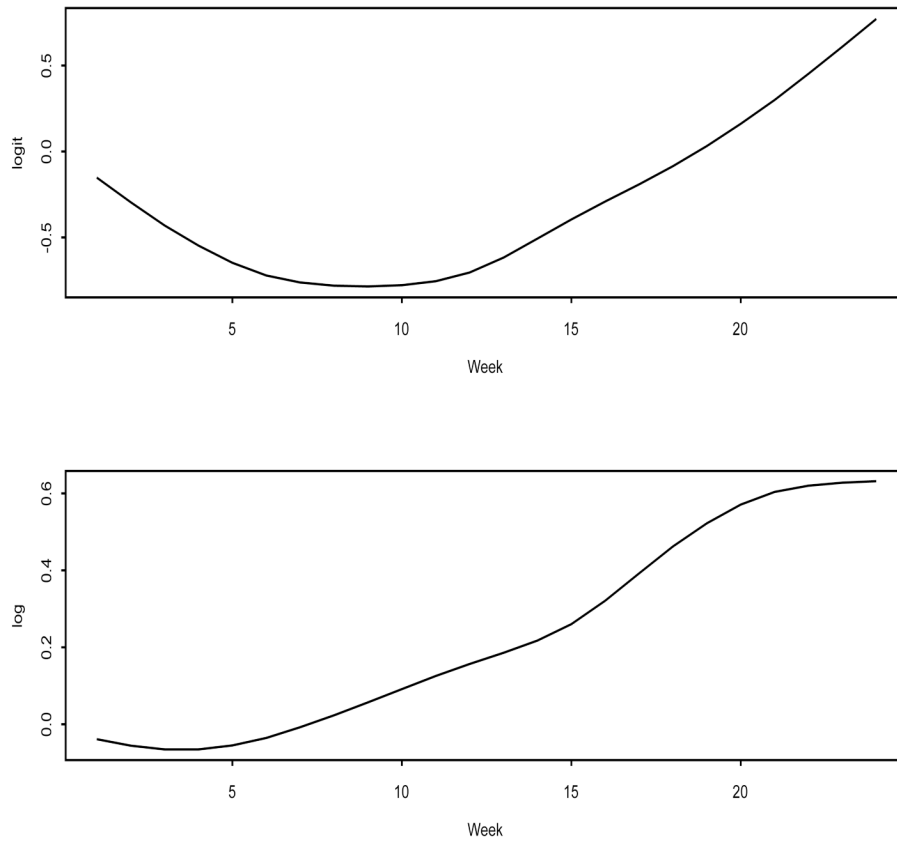
Figure 3.
Spline Estimates of the Average Time Effects on logit$p$ and logλ.

Figure 4.
Spline Estimates of Individual Time Effects on logit$p$ and logλ of Four Individual Subjects.

Figure 5.
The True and Estimated Average Time Effects on logit$p$ and logλ from the Simulation Study.
Note: Solid line represents the true curves and dashed line represents the estimated curves.

Table 1
Parameter Estimates of the ZIP Regression Models

Parameter	Mean	Median	SD	95% C.I.
Zero-inflated				
Logit				
β_{11}^{p} (Intercept)	0.4463	0.4402	0.102	(0.046, 1.69)
β_{12}^{p} (AGE)	0.6874	0.7155	0.249	(0.0864, 1.18)
β_{13}^{p} (STD)	0.2583	0.2445	0.2892	(0.0949, 1.923)
β_{14}^{p} (PTR)	0.375	0.337	0.118	(0.1902, 0.462)
β_{21}^{p} (MOOD)	-0.2318	-0.2305	0.077	(-0.3904, -0.0908)
β_{22}^{p} (SI)	0.5769	0.5758	0.4489	(-0.295, 1.55)
β_{31}^{p} (AR(1))	1.306	1.288	0.173	(1.029, 1.713)
Log-linear				
β_{11}^{λ} (Intercept)	0.0296	0.0302	0.0035	(-0.0722, 0.1259)
β_{12}^{λ} (AGE)	-0.0251	-0.027	0.14	(-0.031, 0.0306)
β_{13}^{λ} (STD)	0.2307	0.2383	0.258	(-0.3383, 0.6504)
β_{14}^{λ} (PTR)	0.0325	0.0293	0.024	(-0.0101, 0.0802)
β_{21}^{λ} (MOOD)	0.1034	0.105	0.026	(0.0572, 0.1493)
β_{22}^{λ} (SI)	0.4193	0.3945	0.2482	(0.1181, 0.906)
β_{31}^{λ} (AR(1))	0.0355	0.03257	0.00882	(0.0239, 0.0567)

Table 2

Parameter Estimates for the Dropout Model and Time-varying Covariates

Parameter	Mean	Median	SD	95% C.I.
Dropout parameter				
ζ_1 (Intercept)	0.2651	0.2809	0.8061	(-0.414, 1.79)
ζ_2 (AGE)	0.7953	0.7694	0.1051	(0.6348, 1.032)
ζ_3 (STD)	0.0392	0.0296	0.16	(0.013, 1.319)
ψ_1 (Current obs)	1.977	1.97	0.2052	(1.571, 2.391)
ψ_2 (Prev. Obs)	-0.4828	-0.4806	0.1444	(-0.7759, -0.2086)
MOOD (Covariate)				
α_{01} (Intercept)	3.556	3.65	0.182	(2.692, 4.95)
α_{11} (AR(1))	0.0955	0.0954	0.084	(0.0788, 0.1116)
SI (Covariate)				
α_{02} (Intercept)	1.138	1.137	0.04707	(1.047, 1.232)
α_{12} (AR(1))	0.2908	0.291	0.0103	(0.2698, 0.3117)

Table 3

Simulation Results

Parameter	Model with Spline				Linear Model			
	Mean	RB	MSE	CP	Mean	RB	MSE	CP
β_{11}^{P} (0.45)	0.47	-0.02	0.042	.94	0.43	0.03	0.048	.93
β_{12}^{P} (0.23)	0.25	0.02	0.023	.92	0.21	0.02	0.037	.91
β_{13}^{P} (-0.21)	-0.2	-0.03	0.029	.96	-0.13	0.063	1.89	.89
β_{14}^{P} (1.3)	1.37	-0.02	0.047	.95	1.1	0.05	0.1	.92
β_{11}^{λ} (0.1)	0.1	0.01	0.041	.94	0.12	0.01	0.052	.94
β_{12}^{λ} (0.25)	0.23	0.02	0.08	.92	0.2	0.04	0.11	.90
β_{13}^{λ} (0.2)	0.22	0.02	0.06	.97	0.08	0.07	1.46	.88
β_{14}^{λ} (0.004)	0.004	0.01	0.032	.97	0.004	0.03	0.04	.98