Published in final edited form as:

J R Stat Soc Ser C Appl Stat. 2016 April ; 65(3): 431–444. doi:10.1111/rssc.12123.

Two-stage model for time varying effects of zero-inflated count longitudinal covariates with applications in health behaviour research

Hanyu Yang,

Pennsylvania State University, University Park, USA

Runze Li,

Pennsylvania State University, University Park, USA

Robert A. Zucker, and

University of Michigan, Ann Arbor, USA

Anne Buu[†]

University of Michigan, Ann Arbor, USA

Summary

This study proposes a two-stage approach to characterize individual developmental trajectories of health risk behaviors and delineate their time-varying effects on short-term or long-term health outcomes. Our model can accommodate longitudinal covariates with zero-inflated counts and discrete outcomes. The longitudinal data of a well-known study of youth at high risk for substance abuse are presented as a motivating example to demonstrate the effectiveness of the model in delineating critical developmental periods of prevention and intervention. Our simulation study shows that the performance of the proposed model improves as the sample size or number of time points increases. When there are excess zeros in the data, the regular Poisson model cannot estimate either the longitudinal covariate process or its time-varying effect well. This result, therefore, emphasizes the important role that the proposed model plays in handling zero-inflation in the data.

Keywords

Functional data; Health risk behavior; Hurdle model; Measurement error; Mixed model; Substance abuse

1. Introduction

Health behavior research often collects longitudinal data on the frequency or quantity of health risk behaviors such as the number of drinking days in a month or the number of drinks consumed on a drinking day (Buu et al., 2012a). Characterizing the *developmental*

[†]Address for correspondence: Anne Buu, School of Nursing, University of Michigan, 400 North Ingalls, Ann Arbor, MI 48109, USA. buu@umich.edu.

Conflict of Interest: None declared.

trajectory of health risk behaviors can facilitate the assessment of disease progression or evaluation of long-term effects of treatment or intervention. It can also provide important information about the timing for prevention and intervention. Furthermore, the developmental trajectory may be predictive of a future health outcome. If we can establish such *predictive validity*, the developmental trajectory can be useful for identifying high risk individuals for prevention and intervention purposes. Our methodological work was motivated by a well-known alcohol study, the Michigan Longitudinal Study (MLS), that has followed children at risk for alcoholism from childhood to adulthood. Our statistical analysis aims to characterize the developmental trajectory of alcohol use behaviors during adolescence and evaluate whether the adolescent alcohol use pattern is predictive of the alcohol dependence diagnosis in adulthood.

In order to achieve these research goals, we have to overcome some methodological challenges. First, the covariate is measured at many time points but the outcome is collected at one future time point so standard longitudinal methods that were designed for longitudinal outcomes are not applicable in this setting. Second, the longitudinal covariates are usually count measures with excess zeros beyond what would be expected from a typical distribution for count data such as the Poisson (Buu et al., 2012a). Third, the effect of longitudinal patterns of health risk behaviors on a short-term or long-term health outcome may be a complex function of time. For example, those periods with frequent occurrence of binge drinking (defined as consuming more than 5 standard alcohol drinks in one episode) tend to have higher negative effects on alcohol related problems. Fourth, health risk behaviors are most of the time self-reported and thus are subject to measurement errors due to recall bias or embarrassment (Tourangeau and Yan, 2007). Fifth, the between-subject and within-subject variability tends to be large in this kind of data, especially for studies on high risk youth who has not yet developed regular patterns of health risk behaviors (Collins et al., 2008).

Zhang et al. (2007) developed a two-stage functional mixed model that was motivated by the need to investigate the effect of the follicle stimulating hormone (FSH) time profile during a menstrual cycle on total hip bone mineral density (BMD) measured at a single time point. The model consists of two stages: the first stage estimates the periodic subject-specific FSH profiles using a nonparametric measurement error model; the second stage plugs the estimated subject-specific profiles in a functional linear model and estimates the functional covariate effect. Another relevant model was proposed by Bhadra et al. (2012) who studied the association between longitudinal prostate-specific antigen exposure history (continuous covariate) and prostate cancer (binary outcome). Bayesian inferential techniques were used to estimate the parameters of the exposure and disease risk models simultaneously by using a joint likelihood.

In the statistics literature, two approaches are commonly used to deal with measurement errors: regression calibration and joint likelihood. The two existing models reviewed above (Zhang et al., 2007; Bhadra et al., 2012) represent these approaches, respectively. Since the longitudinal covariate process in our setting is nonparametric, it requires parameterization with high-dimensional random effects. Thus, the procedure proposed by Bhadra et al. (2012) may encounter computational issues because the reversible jump MCMC algorithm is

expected to converge slowly in the presence of high-dimensional random effects. Furthermore, the procedure of Bhadra et al. (2012) was developed for the setting with continuous longitudinal covariates and a binary response. Extending their Bayesian framework to the setting studied in this paper may be conceptually straightforward and yet challenging for implementation, as our setting involves a zero-inflated Poisson covariate and a response with a conditional distribution in the exponential family. For these reasons, in this study, we adopt the regression calibration approach and extend the model proposed by Zhang et al. (2007), which was originally designed for a continuous longitudinal covariate process and a continuous scalar outcome. The resulting computational complexity is also addressed in our work.

Two models have been frequently adopted in applied fields to handle longitudinal zero-inflated count data: the hurdle model and the zero-inflated Poisson model (see Buu et al. (2012a) for a comprehensive review). Both of them are finite mixture models – the hurdle mixes 0 and a truncated Poisson, whereas the zero-inflated Poisson mixes 0 and a regular Poisson. Conceptually, the hurdle model assumes that everybody in the study population is at risk for being involved in health risk behaviors and the level of involvement changes over time, whereas the zero-inflated Poisson model assumes that the population consists of a group of people who are at risk and another group who are not. Thus, the conceptual framework of the hurdle model better fits the common research interest of longitudinal studies that examine health behaviors of the at-risk population developmentally. Furthermore, the hurdle model has greater flexibility (it can handle not only zero-inflated data but also zero-deflated data) and computational advantage. For these reasons, it is adopted in this study to characterize the longitudinal covariate process.

This paper is organized as follows. In Section 2, we present a generalized functional linear model that employs the hurdle model to characterize the longitudinal covariate process and also accommodates any response variable from the exponential family. In Section 3, the longitudinal data of a well-known study of substance abuse are presented as a motivating example. Section 4 delineates the design and results of the simulation study. We present discussion and concluding remarks in Section 5.

2. Statistical Model and Parameter Estimation

Suppose that from the i -th subject ($i = 1, \dots, n$), we collected a response variable y_i , a p -dimensional vector of time-invariant covariates $\mathbf{z}_i$, and an observed longitudinal covariate process w_{ij} measured at time points $t_{ij} \in [T_1, T_2]$, $j = 1, \dots, J_i$. Because w_{ij} is a zero-inflated count variable, we use the hurdle model (Mullahy, 1986) to specify its distribution:

$$\begin{aligned} P(w_{ij}=0) &= 1 - \pi_i, \\ P(w_{ij}=w) &= \pi_i \frac{e^{-\mu_{ij}} \mu_{ij}^w}{w! (1 - e^{-\mu_{ij}})} \text{ for } w > 0, \end{aligned} \quad (1)$$

where $0 < \pi_i < 1$ and $\mu_{ij} > 0$. Let u_i be the latent risk level and $x_i(t)$ be the latent longitudinal covariate process for the i -th subject. The parameters can be modeled as $\text{logit}(\pi_i) = u_i$ and $\log(\mu_{ij}) = x_i(t_{ij})$. The unique contribution of these latent quantities u_i and $x_i(t)$ is, therefore, to

deal with measurement errors in w_{ij} , which is a self-reported health risk behavior in our setting.

Assume that the response y_i comes from an exponential family $f_y(y; \eta_i)$ and $g(\cdot)$ is the corresponding link function. y_i is related to the time-invariant covariates $\mathbf{z}_i$, the latent risk level u_i and the latent longitudinal covariate process $x_i(t)$ through the following generalized functional linear model, which is a generalized version of the model in Müller and Stadtmüller (2005):

$$g(\eta_i) = \mathbf{z}_i^T \boldsymbol{\delta} + u_i \beta + \int_{T_1}^{T_2} x_i(t) \gamma(t) dt, \quad (2)$$

where $\boldsymbol{\delta}$ is a vector of coefficients for the time-invariant covariates $\mathbf{z}_i$; β is a coefficient of u_i indicating the effect of individual risk on health outcome; and $\gamma(t)$ is a coefficient function of $x_i(t)$ that delineates the time-varying effect of the longitudinal covariate process.

Since both $x_i(t)$ and $\gamma(t)$ are infinitely dimensional, we propose to estimate them by the natural cubic smoothing splines. Let $\mathbf{t}^0 = (t_1^0, \dots, t_q^0)^T$ be a q -dimensional vector of ordered distinct values of all time points t_{ij} , and assume that $x_i(t)$ and $\gamma(t)$ belong to the function space S_{ncs} spanned by natural cubic splines in $[T_1, T_2]$ with knots $\mathbf{t}^0$. According to Section 2.2 of Green and Silverman (1994) and Zhang et al. (2007), for any function $s(t) \in S_{ncs}$, there exists a set of q piecewise cubic polynomial basis functions $\mathbf{c}(t) = (c_1(t), \dots, c_q(t))^T$ such that $s(t) = \mathbf{s}^T \mathbf{c}(t)$ where $\mathbf{s} = \mathbf{s}(\mathbf{t}^0)$ is a q -dimensional vector formed by evaluating $s(t)$ at knots $\mathbf{t}^0$, and $\mathbf{s} \rightarrow s(t)$ is a 1-1 mapping from $\mathbb{R}^q$ to S_{ncs} . Under this setting, we have $x_i(t) = \mathbf{x}_i^T \mathbf{c}(t)$ and $\gamma(t) = \boldsymbol{\gamma}^T \mathbf{c}(t)$. Therefore, Equation (2) becomes

$$g(\eta_i) = \mathbf{z}_i^T \boldsymbol{\delta} + u_i \beta + \mathbf{x}_i^T C \boldsymbol{\gamma}, \quad (3)$$

where $C = \int_{T_1}^{T_2} \mathbf{c}(t) \mathbf{c}^T(t) dt$ is a $q \times q$ matrix. Each $c_i(t)$ is also a natural cubic spline function such that $c_i(\mathbf{t}^0) = \mathbf{e}_i$, where $\mathbf{e}_i$ is a q -dimensional vector with 1 for the i -th element and 0 otherwise. Hence, each element in C can be calculated straightforwardly.

We then smooth a nonparametric function $s(t)$ by penalizing the second derivative through $\int_{T_1}^{T_2} (s''(t))^2 dt$. For the particular settings of natural cubic spline, Equation (2.3) in Green and Silverman (1994) specifies a $q \times q$ symmetric matrix K such that $\int_{T_1}^{T_2} (s''(t))^2 dt = \mathbf{s}^T K \mathbf{s}$.

We adopt the two-stage approach proposed in Zhang et al. (2007) because it takes into account the measurement error of the observed longitudinal covariate process w_{ij} . This is essential as w_{ij} tends to be self-reported data of health risk behaviors in our setting. In Stage 1, we estimate u_i and $\mathbf{x}_i$; in Stage 2, we estimate $\boldsymbol{\delta}$, β , and $\boldsymbol{\gamma}$ given the estimates obtained in Stage 1. The technical details are described in the following sections.

2.1. The first stage: estimating u_i and x_i

In the estimation procedure, we assume $u_i = u_0 + v_i$ where u_0 is the population risk level and $v_i \sim N(0, \sigma_v^2)$ is the random deviation of an individual's risk from the population risk. Similarly, the individual profile $x_i(t)$ can be decomposed as $x_i(t) = x_0(t) + d_i(t)$ where $x_0(t)$ is the population profile and $d_i(t)$ is the random deviation of an individual profile from the population profile. We assume that both $x_0(t)$ and $d_i(t)$ are natural cubic smoothing spline functions and $d_i(t)$ is a zero-mean Gaussian process, that is, $E(d_i(t)) = 0$ for any t . In addition, v_i and $d_i(t)$ are assumed to be mutually independent. Under this assumption, the likelihood from Model (1) can be factored into two components: one involves a Bernoulli distribution and the other a zero-truncated Poisson distribution. The estimation can be, therefore, conducted in two *independent* steps.

2.1.1. Step 1 estimating u_i —Let $w_{ij}^* = 0$ if $w_{ij} = 0$ and $w_{ij}^* = 1$ if $w_{ij} > 0$. We can estimate u_i using a generalized linear mixed model: $w_{ij}^* \sim \text{Bernoulli}(\pi_i)$, where $\text{logit}(\pi_i) = u_i = u_0 + v_i$. The penalized quasi-likelihood (PQL) method (Breslow and Clayton, 1993) is adopted to carry out the estimation of u_0 and v_i .

2.1.2. Step 2 estimating x_i —We estimate $x_i(t)$ from all $w_{ij} > 0$ (denoted by w_{ij}^+) through the following model

$$\begin{aligned} w_{ij}^+ &\sim \text{ZeroTruncatedPoisson}(\mu_{ij}), \\ \log(\mu_{ij}) &= x_i(t_{ij}) = x_0(t_{ij}) + d_i(t_{ij}). \end{aligned} \quad (4)$$

By stacking over j , $\log(\boldsymbol{\mu}_i) = N_i \mathbf{x}_0 + N_i \mathbf{d}_i$, where $\log(\boldsymbol{\mu}_i) = (\log(\mu_{i1}), \dots, \log(\mu_{iJ_i}))^T$, $\mathbf{x}_0 = x_0(\mathbf{t}^0)$, $\mathbf{d}_i = d_i(\mathbf{t}^0)$, and N_i is a $J_i \times q$ incidence matrix mapping $(t_{i1}, \dots, t_{iJ_i})^T$ to $\mathbf{t}^0$ such that the (j, l) -th element is 1 if $t_{ij} = t_l^0$ and 0 otherwise.

We next transform the random subject-specific deviation as $\mathbf{d}_i = B_* \mathbf{b}_i$. Here, $B_* = (\mathbf{1}_q, \mathbf{t}^0, B)$ is a $q \times q$ matrix with $B = L(L^T L)^{-1}$, where L is a $q \times (q-2)$ full-rank matrix satisfying $L^T (\mathbf{1}_q, \mathbf{t}^0) = \mathbf{0}$ and $LL^T = K$. In addition, $\mathbf{b}_i \sim N(\mathbf{0}, D)$ with a covariance matrix

$D = \text{diag}(\sigma_1^2, \sigma_2^2, \tau_d I_{(q-2) \times (q-2)})$. An advantage of this transformation is that τ_d along with B directly characterize the smoothness of $\mathbf{d}_i$. It can be shown that B_* is a $q \times q$ full-rank matrix, and $\mathbf{d}_i = B_* \mathbf{b}_i$ is a 1-1 transformation. Hence, we should focus on finding the estimation of $\mathbf{b}_i$. Using the following notations: $\log(\boldsymbol{\mu}) = (\log(\boldsymbol{\mu}_1)^T, \dots, \log(\boldsymbol{\mu}_n)^T)^T$,

$N = (N_1^T, \dots, N_n^T)^T$, and $Z = \text{diag}(N_1 B_*, \dots, N_n B_*)$, we may write the model as a generalized additive mixed model (Lin and Zhang, 1999):

$$\log(\boldsymbol{\mu}) = N \mathbf{x}_0 + Z \mathbf{b},$$

where $\mathbf{b} = (\mathbf{b}_1^T, \dots, \mathbf{b}_n^T)^T \sim N(\mathbf{0}, D(\boldsymbol{\theta}))$ with the variance component $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \tau_d)$ and $D = \text{diag}(D, \dots, D)$.

Based on the results in Lin and Zhang (1999), the estimates of $\mathbf{x}_0$ and $\mathbf{b}$ can be derived by maximizing the double penalized quasi-likelihood (DPQL)

$$-\frac{1}{2} \sum_{ij} h_{ij} - \frac{1}{2} \mathbf{b}^T D^{-1} \mathbf{b} - \frac{1}{2} \lambda_x \mathbf{x}_0^T K \mathbf{x}_0 \quad (5)$$

for given λ_x and θ , where

$$h_{ij} = -2 \int_{u_w^{-1}(w_{ij}^+)}^{\mu_{ij}} \frac{m_{ij}(w_{ij}^+ - u_w(s))}{s} ds.$$

In this DPQL, $u_w(\cdot)$ is the mean function of the zero-truncated Poisson distribution; $\lambda_x > 0$ is a smoothing parameter controlling the goodness-of-fit of the model and the smoothness of $x_0(t)$; K has been defined in the previous paragraph; and m_{ij} are the prior weights which are set to be 1 (i.e., equal weights) in most applications, unless we have prior knowledge of the relative contribution of observations. In our study, all m_{ij} are set to be 1.

Both $\mathbf{x}_0$ and $\mathbf{b}$ can be estimated by the method in Section 3 of Lin and Zhang (1999); and the variance component θ can be estimated by the marginal quasi-likelihood method and then bias-corrected following similar procedures described in Sections 4 and 5 of that paper. The tuning parameter λ_x is selected by the generalized cross validation (GCV). It is important to note that $\log(\cdot)$ in Equation (4) is actually a pseudo-link function because it models μ_{ij} rather than the mean of w_{ij}^+ from the zero-truncated Poisson distribution, of which the mean and

variance functions are $u_w(\mu) = \frac{\mu}{1 - e^{-\mu}}$ and $v_w(\mu) = \frac{\mu(1 - e^{-\mu} - \mu e^{-\mu})}{(1 - e^{-\mu})^2}$, respectively. Thus, all the terms corresponding to these functions in the original estimation procedures of Lin and Zhang (1999) should be modified accordingly.

2.2. The second stage: estimating δ , β and γ

We plug the estimates $\hat{u}_i$ (Step 1) and $\hat{\mathbf{x}}_i = \mathbf{x}_0 + B \mathbf{b}_i$ (Step 2) obtained from the first stage into Equation (3) and re-write the model as

$$g(\eta_i) = \mathbf{z}_i^T \delta + \hat{u}_i \beta + \hat{\mathbf{x}}_i^T C \gamma,$$

For a given λ_γ controlling the smoothness of γ , we can estimate δ , β and γ by maximizing the penalized pseudo-quasi-likelihood

$$-\frac{1}{2} \sum_i h_i - \frac{1}{2} \lambda_\gamma \gamma^T K \gamma,$$

where

$$h_i = -2 \int_{y_i}^{\eta_i} \frac{m_i(y_i - u)}{v_y(u)} du.$$

Here, $v_y(\cdot)$ is a specified variance function determined by the distribution f_y ; and m_i are the prior weights which are set to be 1 (i.e., equal weights) in most cases, unless we have prior knowledge of the relative contribution of observations. In our study, all m_i are set to be 1.

This penalized pseudo-quasi-likelihood from the semiparametric model has essentially the same form as DPQL (Equation 5) except for a penalty term from the random effects, thus similar strategies as those in Lin and Zhang (1999) can be applied to estimate δ^* and γ . The tuning parameter λ_γ is also selected by GCV.

It is possible in practical settings that $w_{ij} = 0$ is observed at all time points for some subject i . In this case, u_i can still be estimated (in Step 1 of Stage 1), but the subject-specific deviation $\mathbf{d}_i = B * \mathbf{b}_i$ cannot be estimated (in Step 2 of Stage 1). For such i , we propose to replace $\hat{\mathbf{x}}_i$ with the population profile $\mathbf{x}_0$ as the best estimate.

3. The Michigan Longitudinal Study (MLS)

In this section, we describe a well-known alcohol study, the Michigan Longitudinal Study (MLS), as a motivating example and present the model fitted on the empirical data. The MLS is an ongoing prospective study of people at high risk for alcoholism (Zucker et al., 1996). It is the developmentally earliest study currently extant (from early childhood to adulthood) and is also one of the longest running (funded by the National Institutes of Health for 27 years). The study is highly influential in the alcohol and substance abuse field. We apply our method to data from the MLS and adopt the fitted model as the true model in the simulation, because the features of the data are typical in the field. Thus, our methodological work may have high applicability to the field. More importantly, these rich longitudinal data provide a rare opportunity for us to examine the effect of adolescent risk behaviors on adult health outcomes. Such investigation could shed some light on future prevention and intervention work.

The MLS recruited participant families using fathers' drunk driving conviction records and door-to-door community canvassing in a four-county area in mid-Michigan. All participants received extensive in-home assessments of their substance use and related risk factors and consequences at baseline, and thereafter at 3-year intervals. The children of participant families were followed from early childhood to adulthood. During the critical developmental period of alcohol use: ages 13-20 (Wagner and Anthony, 2002), these children were assessed annually in order to measure drinking onset and patterns more accurately. In this study, we use longitudinal data from a sample of 508 children (71% males) for analysis. The maximum number of time points available is 8, although some participants may skip certain time points. Because the proposed model employing latent covariates does not require every participant to have the same number of time points or a fixed assessment schedule, the proposed procedure is quite flexible in a variety of longitudinal data settings as long as the observed time points across all subjects are dense.

Our investigation aims to (1) characterize alcohol use behavior developmentally during adolescence, and (2) delineate the time-varying effect of adolescent alcohol use patterns on alcohol dependence diagnosis in adulthood. In our analysis, the time-varying covariate at each annual assessment is the number of drinking days in past month, which is a commonly

adopted time frame for retrospective reports of drinking behaviors. Because it is a count measure with excess zeros that can barely be handled by the classical Poisson regression model (67% across waves), statistical models for zero-inflated count data such as the hurdle model are needed. The outcome variable is the DSM IV alcohol dependence diagnosis (APA, 1994) during early adulthood, which is a binary variable with the value 1 for positive diagnosis and 0 otherwise. Because the alcohol research literature shows that positive family history of alcoholism and male gender both contribute to higher risk for meeting alcohol dependence diagnosis (see Buu et al. (2012b) for a comprehensive review), we include two binary time-invariant covariates in the model: parental lifetime alcohol use disorder diagnosis (positive diagnosis from either biological parent was coded as 1) and gender (male was coded as 1).

We fitted the proposed model on the MLS data. Figure 1 shows the overall developmental trajectory of alcohol use behavior during adolescence $x_0(t)$ and the time-varying effect of adolescent alcohol use on adult alcohol dependence diagnosis $\gamma(t)$. As demonstrated by the left panel, in general, the frequency of alcohol use increased during adolescence, and the curve became flat when development approached early adulthood. The right panel shows that drinking behavior in early adolescence (before age 15) had the highest effect on adult alcohol dependence. In addition, around the time when most of the youth entered college (age 18), the time-varying effect had a small peak. Moreover, the confidence interval at ages 15 and 18 tended to be wider because of the drastic change in curvature.

In terms of those time-invariant effects, our model indicates that males were at higher risk for meeting alcohol dependence criteria ($\hat{\delta}_1 = 1.4353$ with 95%CI=[0.8586, 2.0120]); parental alcohol use disorder did not have a significant effect ($\hat{\delta}_2 = 0.3256$ with 95%CI=[-0.2194, 0.8706]); individual risk for alcohol use contributed to greater likelihood of adult alcohol dependence ($\hat{\beta} = 1.2449$ with 95%CI=[0.8973, 1.5924]). Our analysis also estimated the population risk level u_0 to be -0.7310 (with 95%CI=[-0.8637, -0.5983]) that corresponds to 65% zeros. Moreover, the other variance parameters in the model were estimated as $\hat{\sigma}_v^2 = 1.0630^2$, $\hat{\sigma}_1^2 = 0.6198^2$, $\hat{\sigma}_2^2 = 0.3057^2$, $\hat{\tau}_d = 1.7353^2$, and the smoothing parameters were selected as $\lambda_x = 1.3$, $\lambda_\gamma = 0.003$.

4. Simulation

We conducted two simulation experiments in this study. The first experiment was designed to evaluate the performance of the proposed model with different sample sizes, numbers of time points, and proportions of zeros. The fitted model in Section 3 was adopted as the true model to generate simulated data that closely represent the structure of real data so that the results can be more generalizable to the substance abuse field (Burton et al., 2006). The other experiment was motivated by our previous work finding that zero-inflation was frequently ignored in practice and yet could result in poor estimation, if the classical Poisson model was employed in the settings involving longitudinal outcomes (Buu et al., 2012a) or variable selection (Buu et al., 2011). Thus, our second simulation experiment evaluated the potential negative consequence of ignoring zero-inflation in the setting involving longitudinal covariates with zero-inflated counts and a discrete outcome. The detailed designs and results of the two experiments are described in the following sections.

4.1. Experiment 1: evaluating the performance of the proposed model

In this experiment, we manipulated three factors: (1) the sample size with the small size $n = 250$, medium size $n = 500$, or large size $n = 750$; (2) the number of time points with $q = 8$, 16, or 32; and (3) the population risk level with $u_0 = -0.75$ or $u_0 = 0.75$. When $u_0 = -0.75$, the data contained approximately 65% zeros and this setting was similar to that of the real data. When $u_0 = 0.75$, there were around 35% zeros in the data and we used this setting to simulate the situation to which a regular Poisson may apply.

We first set $\mathbf{t}^0$ to be q equally spaced time points in $[-1.5, 1.5]$, where q is 8, 16, or 32. In order to simulate unbalanced observation time points t_{ij} among n subjects, we allowed the longitudinal covariate process w_{ij} for subject i to be observed at t_{ij} for $j = 1, \dots, J_i$, where J_i was an integer randomly chosen from $\{\lceil \frac{3}{4}q \rceil, \dots, q\}$, and t_{ij} were J_i distinct points among $\mathbf{t}^0$.

The longitudinal covariate data w_{ij} were zero-inflated counts generated by Model (1). Here, we set u_0 to be -0.75 or 0.75 , $\sigma_v^2 = 1^2$, $x_0(t) = 0.5 + 0.8 \arctan(2t)$, and $d_i(t)$ was determined by $\mathbf{d}_i = B * \mathbf{b}_i$ with $\mathbf{b}_i$ being a random sample from $N(\mathbf{0}, \text{diag}(\sigma_1^2, \sigma_2^2, \tau_d I_{(q-2) \times (q-2)}))$, where $\sigma_1^2 = 0.6^2$, $\sigma_2^2 = 0.3^2$, and $\tau_d = 1^2$. The response data y_i were then generated from

$$y_i \sim \text{Bernoulli}(\eta_i), \\ \text{logit}(\eta_i) = z_i \delta + u_i \beta + \int_{-1.5}^{1.5} x_i(t) \gamma(t) dt,$$

where $\beta = 1.25$, $\delta = 1.5$, and $\gamma(t) = -1.5t^3 + t^2 + 1.5t - 0.5$. In addition, z_i were randomly drawn from $\{0, 1\}$ with $P(z_i = 1) = \frac{2}{3}$ for each subject i .

In summary, we manipulated three factors, the sample size (n), number of time points (q), and the population risk level (u_0), and considered $3 \times 3 \times 2 = 18$ situations in total. For each combination, we generated $N_S = 100$ data sets and applied the proposed two-stage method to estimate the parameters. For each parameter, the mean with standard error (se) and the mean squared error (MSE) were calculated from N_S replications. For the nonparametric coefficient functions $x_0(t)$ and $\gamma(t)$, the mean integrated squared error (MISE) was used to summarize the results.

Table 1 presents the results of Experiment 1 with varied sample sizes n and fixed number of time points ($q = 16$) and population risk level ($u_0 = -0.75$). When the sample size increased, the performance of the proposed model improved (i.e. smaller MSE or MISE). In Table 2, the number of time points q was varied while holding the other two factors constant ($n = 500$, $u_0 = -0.75$). The performance of the proposed model improved with the number of time points. The effect of the population risk level u_0 is summarized in Table 3 with the other two factors fixed at their medium values ($n = 500$, $q = 16$). By varying the value of u_0 from -0.75 to 0.75 , we manipulated the proportion of zeros in the data from a zero-inflated situation ($\approx 65\%$) to a situation similar to a regular Poisson ($\approx 35\%$). The result shows that the performance of the proposed model was only slightly better in the regular Poisson situation. This means that our model can handle the zero-inflation pretty well.

According to the values of mean(se) reported in Tables 1-3, some biases exist in the estimation of parameters. This may result from the Laplace approximation employed in the high-dimensional integration for quasi-likelihood calculation, or the two-stage approach. Nevertheless, given that the true values are covered by the 95% confidence intervals, these biases may still be at an acceptable level.

Since the MISE of estimates for $\gamma(\cdot)$ is noticeably larger than that for $x_0(\cdot)$, we further examine the estimates of $\gamma(\cdot)$. Figure 2 depicts the true curve of $\gamma(t)$ used in this simulation (the solid curve), the average of estimated curves over 100 replications for $n = 500$, $q = 16$, and $u_0 = -0.75$ (the dashed line), and its pointwise 95% confidence intervals calculated as the average of estimated curves ± 1.96 multiplying the standard error of the 100 estimated curves (the dotted curves). It can be seen from Figure 2 that the large MISE of $\hat{\gamma}(\cdot)$ is mainly due to the large variance of $\hat{\gamma}(\cdot)$ because the average of the estimated curves is very close to the true one.

4.2. Experiment 2: evaluating the consequence of ignoring the zero-inflation

In the second experiment, we fitted the regular Poisson model (instead of the hurdle model) on the simulated data generated according to the true model described in Experiment 1. Here, we only manipulated the proportion of zeros in the data (through the two different values of u_0) while holding the sample size and number of time points at their medium values ($n = 500$, $q = 16$).

Table 4 summarizes the results of the simulation. As expected, when the data contained excess zeros (i.e. $u_0 = -0.75$ corresponding to 65% zeros), the regular Poisson did not perform as well as in the situation without zero-inflation (i.e. $u_0 = 0.75$ corresponding to 35% zeros). In Figure 3, we present the fitted longitudinal covariate process $x_0(t)$ and its time-varying effect $\gamma(t)$ under the two situations. The estimated curves were derived from evaluating the fitted functions of the 100 replications at a set of grid points and connecting the means at all grid points. As demonstrated in the left panel, the regular Poisson model estimated $x_0(t)$ well when there was no zero-inflation (the bottom graph) whereas the model did very poorly in the situation of zero-inflation (the top graph). On the other hand, the right panel of Figure 3 shows that the performance of the regular Poisson did not vary much with the existence of zero-inflation in the data. Specifically, the regular Poisson model estimated the curve of $\gamma(t)$ to be pretty flat in both situations, whereas the true effect changed developmentally across time.

5. Discussion

This study proposes a generalized time-varying effect model that employs the hurdle model to characterize the longitudinal covariate process with zero-inflated counts and also accommodates any response variable from the exponential family. Our model delineates both the longitudinal covariate process and its time-varying effect on a scalar outcome using natural cubic smoothing splines. Our two-stage approach also takes into account measurement errors in the observed longitudinal covariate process. Further, the proposed model involves random effects to deal with individual differences in risk levels and developmental trajectories.

The longitudinal data of the MLS, a well-known study of youth at high risk for substance abuse, are presented as a motivating example to demonstrate that the proposed model is very useful for informing critical developmental periods of prevention or intervention. For example, Figure 1 shows that alcohol use in early adolescence or the freshman year of college could have profound impact on adult alcohol dependence. Further, our empirical analysis indicates that family history of alcoholism has a distal and nonsignificant effect on adult alcohol dependence in comparison to adolescent alcohol use. This indicates that prevention and intervention efforts during adolescence can still counteract genetic vulnerability. Moreover, the proposed model can be applied to not only multi-wave longitudinal studies like the MLS, but also to short-term studies that involve intensive data collection such as the ecological momentary assessment or the daily process data, which are collected through modern health communication technology (Cranford et al., 2010; Liu et al., 2013; Selya et al., 2013; Vasilenko et al., 2014).

The design of our simulation studies is unique because it simulates the features of the empirical data so that the results can be generalizable to the substance abuse field. The performance of the proposed model improves as the sample size or the number of time points increases. Our simulation study also shows that when there are excess zeros in the data, the two-stage model involving the regular Poisson cannot estimate either the longitudinal covariate process or its time-varying effect well. This emphasizes the important role that the hurdle model plays in handling zero-inflation in the data. Furthermore, for most settings in health behavior research, it is practical to assume that each individual carries a different level of risk u_i for engaging in health risk behaviors. Our simulation experiment shows that when data are generated under this conceptual framework, the regular Poisson model which constraints the value of u_i to be 0 for each individual does not estimate the time-varying effect $\gamma(t)$ well, even when the degree of zero-inflation is minimal.

In this study, we only consider the Poisson link function for the hurdle model because it fits well with the real data example. In some applications where overdispersion is common, the negative binomial link should be considered. Furthermore, the two-stage approach ignores the uncertainty inherited from the first stage in the estimation of the second stage. There is, therefore, an issue related to uncertainty propagation in the practical implementation of the two-stage method. The Bayesian frame-work for joint models proposed by Bhadra et al. (2012) has been shown to be effective for dealing with this kind of issue in the setting with continuous longitudinal covariates and a binary response. Thus, future research may extend the Bayesian framework to the setting studied in this paper that involves zero-inflated count longitudinal covariates and a discrete response.

Our work in this paper focuses on the setting that involves a longitudinal covariate process with zero-inflated counts and a discrete outcome. Future studies may consider modeling *multiple* longitudinal covariate processes that tend to co-occur such as substance use behaviors, violence behaviors, and HIV sexual risk behaviors. Furthermore, although a scalar outcome such as the adult alcohol dependence diagnosis in our motivating example is quite common in practice, there are some settings that may involve multiple outcomes. For example, a health outcome may be recorded at several time points. In this kind of setting, we would add some terms with respect to time on the right hand side of Equation (2) to

delineate the longitudinal change in the outcome. We may also add a random term to account for the dependence among repeated measures within each individual (i.e., a random effect model), or take a marginal (GEE) approach with robust standard errors (see Diggle et al. (2002) for a comprehensive account of these alternative approaches). This is, however, beyond the scope of this paper.

6. Software

Software in the form of R code, together with a sample input data set and complete documentation is available on request from the corresponding author (buu@umich.edu).

Acknowledgments

Funding

This work was supported by the National Institutes of Health (NIH) [P50 DA010075, P50 DA036107 & R01 CA168676 to H.Y. and R.L.; R37 AA007065 to R.A.Z.; and K01 AA016591 & R01 DA035183 to A.B.] and National Science Foundation DMS 1512422 to R.L. The content is solely the responsibility of the authors and does not necessarily represent the official view of the NIH.

References

- APA. Diagnostic and Statistical Manual. 4. Washington DC: American Psychiatric Association; 1994.
- Bhadra D, Daniels MJ, Kim S, Ghosh M, Mukherjee B. A Bayesian semiparametric approach for incorporating longitudinal information on exposure history for inference in case-control studies. *Biometrics*. 2012; 68:361–370. [PubMed: 22313248]
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. 1993; 88:9–25.
- Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine*. 2006; 25:4279–4292. [PubMed: 16947139]
- Buu A, Johnson NJ, Li R, Tan X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*. 2011; 30:2326–2340. [PubMed: 21563207]
- Buu A, Li R, Tan X, Zucker RA. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*. 2012a; 31:4074–4086. [PubMed: 22826194]
- Buu A, Wang W, Schroder SA, Kalaida NL, Puttler LI, Zucker RA. Developmental emergence of alcohol use disorder symptoms and their potential as early indicators for progression to alcohol dependence in a high risk sample: a longitudinal study from childhood to early adulthood. *Journal of Abnormal Psychology*. 2012b; 121:897–908. [PubMed: 21842966]
- Collins LR, Kashdan TB, Koutsky JR, Morsheimer ET, Vetter CJ. A self-administered timeline followback to measure variations in underage drinkers' alcohol intake and binge drinking. *Addictive Behaviors*. 2008; 33:196–200. [PubMed: 17720324]
- Cranford JA, Tennen H, Zucker RA. Feasibility of using interactive voice response to monitor daily drinking, moods, and relationship process on a daily basis in alcoholic couples. *Alcoholism: Clinical and Experimental Research*. 2010; 34:499–508.
- Diggle, PJ.; Heagerty, P.; Liang, K.; Zeger, SL. *Analysis of Longitudinal Data*. New York: Oxford University Press; 2002.
- Green, PJ.; Silverman, BW. *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall; 1994.
- Lin X, Zhang D. Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B*. 1999; 61:381–400.

- Liu X, Li R, Lanza ST, Vasilenko S, Piper M. Understanding the role of cessation fatigue in the smoking cessation process. *Drug and Alcohol Dependence*. 2013; 133:548–555. [PubMed: 23954071]
- Mullahy J. Specification and testing of some modified count data models. *Journal of Econometrics*. 1986; 33:341–365.
- Müller H-G, Stadtmüller U. Generalized functional linear models. *The Annals of Statistics*. 2005; 33:774–805.
- Selya AS, Dierker LC, Rose JS, Hedeker D, Tan X, Li R, Mermelstein RJ. Time-varying effects of smoking quantity and nicotine dependence on adolescent smoking regularity. *Drug and Alcohol Dependence*. 2013; 128:230–237. [PubMed: 22995764]
- Tourangeau R, Yan T. Sensitive questions in surveys. *Psychological Bulletin*. 2007; 133:859–883. [PubMed: 17723033]
- Vasilenko S, Piper M, Lanza ST, Liu X, Yang J, Li R. Time-varying processes involved in smoking lapse in a randomized trial of smoking cessation therapies. *Nicotine & Tobacco Research*. 2014; 16S2:S135–S143. [PubMed: 24711627]
- Wagner FA, Anthony JC. From first drug use to drug dependence: Developmental periods of risk for dependence upon marijuana, cocaine, and alcohol. *Neuropsychopharmacology*. 2002; 26:479–488. [PubMed: 11927172]
- Zhang D, Lin X, Sowers MF. Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome. *Biometrics*. 2007; 63:351–362. [PubMed: 17688488]
- Zucker RA, Ellis DA, Fitzgerald HE, Bingham CR, Sanford K. Other evidence for at least two alcoholisms ii: life course variation in antisociality and heterogeneity of alcoholic outcome. *Development and Psychopathology*. 1996; 8:831–848.

Fig. 1.

The fitted model on the empirical data: the overall developmental trajectory of alcohol use behavior during adolescence $x_0(t)$ and the time-varying effect of adolescent alcohol use on adult alcohol dependence diagnosis $\chi(t)$.

Fig. 2.

The estimates of $\gamma(t)$ with pointwise 95% confidence intervals for $n = 500$, $q = 16$, and $u_0 = -0.75$ in Experiment 1.

Fig. 3.

The fitted models without accounting for zero-inflation in Experiment 2: the longitudinal covariate process $x_0(t)$ and its time-varying effect $\gamma(t)$ under different levels of population risk: $u_0 = -0.75$ (65% zeros), $u_0 = 0.75$ (35% zeros).

Table 1

Simulation results of Experiment 1 with varied sample sizes n ($q = 16$, $u_0 = -0.75$).

	$n = 250$			$n = 500$			$n = 750$		
	mean(se)	MISE		mean(se)	MISE		mean(se)	MISE	
$u_0 = -0.75$	-0.7033(0.0706)	0.0071		-0.7000(0.0480)	0.0048		-0.7010(0.0396)	0.0039	
$\beta = 1.25$	1.1641(0.1938)	0.0446		1.1700(0.1444)	0.0271		1.1614(0.1296)	0.0245	
$\delta = 1.5$	1.3476(0.3048)	0.1152		1.3417(0.2150)	0.0708		1.3276(0.1565)	0.0540	
$\sigma_v^2 = 1^2$	0.9022(0.1241)	0.0248		0.8845(0.0843)	0.0204		0.8996(0.0704)	0.0150	
$\sigma_1^2 = 0.6^2$	0.2958(0.0446)	0.0061		0.2986(0.0309)	0.0047		0.2996(0.0243)	0.0042	
$\sigma_2^2 = 0.3^2$	0.0591(0.0235)	0.0015		0.0562(0.0171)	0.0014		0.0565(0.0129)	0.0013	
$\tau_d = 1^2$	0.6815(0.2067)	0.1437		0.6882(0.1474)	0.1188		0.7018(0.1202)	0.1032	
		MISE			MISE			MISE	
$x_0(\cdot)$		0.0712			0.0621			0.0594	
$\gamma(\cdot)$		5.0362			3.9944			2.7841	

Table 2

Simulation results of Experiment 1 with varied numbers of time points q ($n = 500$, $u_0 = -0.75$).

	$q = 8$			$q = 16$			$q = 32$		
	mean(se)	MSE		mean(se)	MSE		mean(se)	MSE	
$u_0 = -0.75$	-0.6766(0.0475)	0.0076		-0.7000(0.0480)	0.0048		-0.7115(0.0524)	0.0042	
$\beta = 1.25$	1.1886(0.1528)	0.0269		1.1700(0.1444)	0.0271		1.1648(0.1426)	0.0274	
$\delta = 1.5$	1.2706(0.2068)	0.0950		1.3417(0.2150)	0.0708		1.3528(0.2252)	0.0719	
$\sigma_v^2 = 1^2$	0.8648(0.1049)	0.0292		0.8845(0.0843)	0.0204		0.9417(0.0854)	0.0106	
$\sigma_1^2 = 0.6^2$	0.2813(0.0357)	0.0075		0.2986(0.0309)	0.0047		0.3110(0.0264)	0.0031	
$\sigma_2^2 = 0.3^2$	0.0570(0.0209)	0.0015		0.0562(0.0171)	0.0014		0.0619(0.0112)	0.0009	
$\tau_d = 1^2$	0.6422(0.2016)	0.1683		0.6882(0.1474)	0.1188		0.7348(0.0974)	0.0797	
<hr/>									
	MISE			MISE			MISE		
$x_0(\cdot)$	0.1084			0.0621			0.0405		
$\gamma(\cdot)$	4.1672			3.9944			2.5494		

Table 3Simulation results of Experiment 1 with varied population risk levels u_0 ($n = 500$, $q = 16$).

	$u_0 = -0.75$ (65% zeros)		$u_0 = 0.75$ (35% zeros)	
	mean (se)	MSE	mean (se)	MSE
u_0	-0.7000(0.0480)	0.0048	0.7082(0.0483)	0.0041
$\beta = 1.25$	1.1700(0.1444)	0.0271	1.2044(0.1710)	0.0310
$\delta = 1.5$	1.3417(0.2150)	0.0708	1.3080(0.2206)	0.0851
$\sigma_v^2 = 1^2$	0.8845(0.0843)	0.0204	0.8985(0.0900)	0.0183
$\sigma_1^2 = 0.6^2$	0.2986(0.0309)	0.0047	0.3079(0.0257)	0.0034
$\sigma_2^2 = 0.3^2$	0.0562(0.0171)	0.0014	0.0603(0.0119)	0.0010
$\tau_d = 1^2$	0.6882(0.1474)	0.1188	0.7384(0.0948)	0.0773
		MISE		
$x_0(\cdot)$		0.0621		0.0418
$\gamma(\cdot)$		3.9944		3.3347

Table 4

Simulation results of Experiment 2 with varied population risk levels u_0 ($n = 500$, $q = 16$).

	$u_0 = -0.75$ (65% zeros)		$u_0 = 0.75$ (35% zeros)	
	mean (se)	MSE	mean (se)	MSE
$\delta = 1.5$	1.2418(0.1782)	0.0981	1.4091(0.2058)	0.0502
$\sigma_1^2 = 0.6^2$	0.7579(0.0551)	0.1613	0.3112(0.0279)	0.0031
$\sigma_2^2 = 0.3^2$	0.1812(0.0254)	0.0090	0.0894(0.0091)	0.0001
$\tau_d = 1^2$	19.3488(4.7072)	358.6152	2.4217(0.4721)	2.2420
	MISE		MISE	
$x_0(\cdot)$		3.5962		0.3112
$\gamma(\cdot)$		3.5962		4.3391