Published in final edited form as:

Stat Med. 2017 August 30; 36(19): 3110–3120. doi:10.1002/sim.7325.

Autoregressive and Cross-lagged Model for Bivariate Non-commensurate Outcomes

Fei He^a, Armando Teixeira-Pinto^b, and Jaroslaw Harezlak^{c,*}

^aInstitute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA

^bSchool of Public Health, The University of Sydney, Sydney, NSW, AUS

^cDepartment of Epidemiology and Biostatistics, School of Public Health, Indiana University, Bloomington, IN, USA

Abstract

Autoregressive and cross-lagged models have been widely used to understand the relationship between bivariate commensurate outcomes in social and behavioral sciences, but not much work has been done in modeling bivariate non-commensurate (e.g. mixed binary and continuous) outcomes simultaneously. We develop a likelihood-based methodology combining ordinary autoregressive and cross-lagged models with a shared subject-specific random effect in the mixed model framework to model two correlated longitudinal non-commensurate outcomes. The estimates of the cross-lagged and the autoregressive effects from our model are shown to be consistent with smaller mean squared error than the estimates from the univariate generalized linear models. Inclusion of the subject-specific random effects in the proposed model accounts for between-subject variability arising from the omitted and/or unobservable, but possibly explanatory, subject-level predictors. Our model is not restricted to the case with equal number of events per subject and it can be extended to different types of bivariate outcomes. We apply our model to an ecological momentary assessment study with complex dependence and sampling data structures. Specifically, we study the dependence between the condom use and sexual satisfaction based on the data reported in a longitudinal study of sexually transmitted infections. We find negative cross-lagged effect between these two outcomes and positive autoregressive effect within each outcome.

Keywords

autoregressive and cross-lagged model; non-commensurate; bivariate; mixed model; ecological momentary assessment

1. Introduction

In multivariate longitudinal data analysis, there are multiple questions of interest: longitudinal change of each outcome, auto-dependence on the past observations of each

*Correspondence to: Jaroslaw Harezlak, Department of Epidemiology and Biostatistics, School of Public Health, Indiana University, Bloomington, IN, USA, Phone:(812)855-6719. Fax:(812)855-9107. harezlak@iu.edu.

outcome, correlation between two or more outcomes and cross-dependence between the longitudinal changes of two outcomes. Many research papers have addressed these questions separately, but there is a scarcity of the models developed to deal with the cases that involve all four questions simultaneously. In this paper, our interest is to establish a method to jointly model correlated bivariate non-commensurate longitudinal outcomes and to estimate their auto- and cross-dependence.

In the 1960s and 1970s, Campbell [1], Bohrnstedt [2], Duncan [3], Heise [4], and Joreskog [5] discussed some early social science examples about the autoregressive and cross-lagged models for two or more outcome variables in panel data. Kessler and Greenberg [6] further developed the methodology to support those examples. These have been and continued to be popular modeling approaches for longitudinal data. Bollen [7] also developed the model combining both autoregressive model and cross-lagged model under structural equation model framework with latent variables introduced. Several examples with application of those models can be found in the psychology and behavior literature. Majority of published research on the autoregressive and cross-lagged models considered the outcomes of the same type, i.e. commensurate outcomes. We propose an extension that applies to the mixed (non-commensurate) outcomes, e.g. mixed binary and continuous outcomes. In order to estimate the dependence among different types of outcomes, we borrow the idea of joint modeling of non-commensurate outcomes from multivariate methods and incorporate it in the autoregressive and cross-lagged models.

The challenge of dealing non-commensurate outcomes with multivariate methods is the nonexistence of obvious multivariate distributions. To address this problem, two general likelihood-based approaches have been proposed to avoid the direct specification of the joint distribution of the outcomes: 1) factorization of the joint distribution of the outcomes and 2) introduction of random effects to model the correlation among the multiple outcomes. Conditional models allow the joint distribution to be factorized into a marginal component and a conditional component. Such approaches have been discussed in Tate [8], Olkin and Tate [9], Little and Schluchter [10], Krzanowski [11], and Cox and Wermuth [12][13]. On the other hand, Teixeira-Pinto and Normand [14] proposed a multivariate method to analyze correlated binary and continuous outcomes using probit-type model, which was established by Catalano and Ryan [15], with a shared random variable introduced in cross-sectional data. They have shown that this model is equivalent to the factorization model presented by Catalano and Ryan [15] and the estimates could be obtained via mixed models equivalence.

In this paper, we presents autoregressive and cross-lagged model for correlated bivariate non-commensurate longitudinal outcomes under complex dependence and sampling data structures. The path graph of autoregressive and cross-lagged model for one continuous outcome and one binary outcome is shown in Figure 1. Our interest is estimating the cross-lagged effect $\beta^{[c]}$ and $\beta^{[b]}$ as well as the autoregressive effects $\gamma^{[c]}$ and $\gamma^{[b]}$. This work is motivated by a sexually transmitted infection (STI) ecological momentary assessment (EMA) study (IU Phone study). Participants in this study have undergone STI testing and completed EMA forms according to the predetermined schedules. We are interested in analyzing several variables related to the risk of infection and sexual or non-sexual behaviors. In particular, we are interested in the association between the longitudinal

processes of two correlated variables (e.g. condom use behavior and sexual satisfaction) in the study. Sexual behaviors change dynamically with time and have been shown to have influence on the STI transmission. This dynamic aspect makes the understanding of the interrelationships between the longitudinal trajectories crucial in the STI prevention. For example, in the IU Phone study, we have found that the likelihood of condom use declines sharply for both men and women after the sexual relationship initiation with a new partner. Relationship and sexual satisfaction also influence declines in condom use, especially among women [16]. In this paper, we further explore the dynamics of the association between the condom use and sexual as well as relationship satisfaction to study their potential effects on the STI transmission. It will further help us understand the STI transmission and prevention at the event-related level. The design of the IU Phone study and the data structure are introduced in Section 2. In Section 3, the construction of the proposed model and the algorithm of estimation are described. We first develop a model for two continuous outcomes, then extend it to one continuous outcome and one binary outcome. The simulation study under different data generating mechanisms are implemented to compare the proposed model with univariate generalized linear models (GLMs) in Section 4. We apply the proposed model and univariate GLMs to condom use (binary variable) and sexual satisfaction (continuous variable) from IU Phone study in Section 5. The conclusions and limitations are discussed in Section 6.

2. Motivating data set: the IU Phone study

Data were obtained from a prospective 84-days (12-weeks) study designed to examine sexual behaviors and incident STI. Participants were recruited from the patient population of a county sexually transmitted diseases clinic but were not necessarily clinic patients at the time of enrollment. Eligibility criteria were ages 18 to 29 years (inclusive), English speaking, and planning to reside in the area for the subsequent 84 days. The Institutional Review Board of Indiana University Purdue University Indianapolis approved this study. All participants have provided informed consent.

Ecological momentary assessment (EMA) diary technology was used in the study. The primary mode of data collection was via three-times daily self-reports of coital and non-coital sexual behaviors, condom use, and relationship assessments, recorded with project-furnished cellular telephones and service. The expected number of entries was thus 252 entries per participant. Daily diary completion rate of this EMA diary data was 87.7%. Other methodological details were previously published in Hensel et al. [17]. At pre-selected 8-hour intervals, participants responded to a series of questions to identify sexual and non-sexual interactions with specific partners. In each eight-hour reporting period, participants identified any partner, time of each coital event (up to four events within the same eight-hour reporting period), condom use for each coital event, as well as relationship satisfaction and sexual satisfaction. Here, condom use was coded as a binary variable (0 for no condom use; 1 for condom use), and relationship and sexual satisfaction were measured by single item Likert scale from 1 ('very low') to 10 ('very high').

There were 254 participants who have more than one coital events with partner(s) during the reporting period included in our analysis. All the included individuals completed the whole

study and were not jailed during the study period, 58% (147/254) of the participants were women and 90% (229/254) were African American. To study the autoregressive and cross-lagged effects of the outcomes, we included participants who had two or more coital events reported in the study. In this group, the mean number of measurements per subject was 13.6, the median was 8.5 and first and third quartiles were 4.0 and 18.0, respectively.

3. Models and methodology

In our proposed autoregressive and cross-lagged model with two outcomes, we include each outcome's observations at time $(t - 1)$ as predictors of the other outcome's observations at time t . The coefficients of autoregressive effect and cross-lagged effect are assumed to be constant between every two adjacent observation time. We start by establishing the model for two continuous outcomes case and then move to the non-commensurate setting with one continuous and one binary outcome case.

3.1. Two continuous outcomes

Let $y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ be two normally distributed correlated continuous outcomes. The structural regression models of $y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ are

$$\begin{aligned} y_{i,t}^{[c1]} | y_{i,t-1}^{[c1]}, y_{i,t-1}^{[c2]}, u_i &= \alpha^{[c1]} + \beta^{[c1]} y_{i,t-1}^{[c2]} + \gamma^{[c1]} y_{i,t-1}^{[c1]} + \sigma_{\varepsilon 1} u_i + \varepsilon_{i,t}^{[c1]} \\ y_{i,t}^{[c2]} | y_{i,t-1}^{[c1]}, y_{i,t-1}^{[c2]}, u_i &= \alpha^{[c2]} + \beta^{[c2]} y_{i,t-1}^{[c1]} + \gamma^{[c2]} y_{i,t-1}^{[c2]} + \sigma_{\varepsilon 2} u_i + \varepsilon_{i,t}^{[c2]} \end{aligned} \quad (1)$$

where $\varepsilon_{i,t}^{[c1]} \sim N(0, \sigma_{\varepsilon 1}^2)$, $\varepsilon_{i,t}^{[c2]} \sim N(0, \sigma_{\varepsilon 2}^2)$ and $u_i \sim N(0, \sigma_u^2)$. The random variable u_i is introduced into both equations to estimate the correlation between the two outcomes; $\sigma_{\varepsilon 1}$ and $\sigma_{\varepsilon 2}$ are scaling parameters used to standardize the residuals. The correlation between

$y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ given $y_{i,t-1}^{[c1]}$ and $y_{i,t-1}^{[c2]}$ is $\frac{\sigma_u^2}{1 + \sigma_u^2}$ and it is assumed that given u_i , $y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ are independent.

We also assume the first observations for both outcomes only depend on $\alpha^{[c1]}$ and $\alpha^{[c2]}$, respectively.

$$\begin{aligned} y_{i,1}^{[c1]} | u_i &= \alpha^{[c1]} + \sigma_{\varepsilon 1} u_i + \varepsilon_{i,1}^{[c1]} \\ y_{i,1}^{[c2]} | u_i &= \alpha^{[c2]} + \sigma_{\varepsilon 2} u_i + \varepsilon_{i,1}^{[c2]} \end{aligned} \quad (2)$$

The likelihood based on equation (1) and (2) can be written as the multiplication of conditional probability of the fixed mean part and marginal probability of the random effect u_i .

$$\begin{aligned}
& f(y_{i,1}^{[c1]}, \dots, y_{i,t}^{[c1]}, y_{i,1}^{[c2]}, \dots, y_{i,t}^{[c2]}) \\
&= \prod_{i=1}^N \int f(y_{i,1}^{[c1]} | u_i) f(y_{i,1}^{[c2]} | u_i) \prod_{t=2}^T f(y_{i,t}^{[c1]} | y_{i,t-1}^{[c1]}, y_{i,t-1}^{[c2]}, u_i) f(y_{i,t}^{[c2]} | y_{i,t-1}^{[c1]}, y_{i,t-1}^{[c2]}, u_i) f(u_i) du_i \\
&= \prod_{i=1}^N \int \frac{1}{\sqrt{2\pi\sigma_{\varepsilon 1}^2}} \exp \left[-\frac{(y_{i,1}^{[c1]} - \alpha^{[c1]} - \sigma_{\varepsilon 1} u_i)^2}{2\sigma_{\varepsilon 1}^2} \right] \frac{1}{\sqrt{2\pi\sigma_{\varepsilon 2}^2}} \exp \left[-\frac{(y_{i,1}^{[c2]} - \alpha^{[c2]} - \sigma_{\varepsilon 2} u_i)^2}{2\sigma_{\varepsilon 2}^2} \right]
\end{aligned} \quad (*)$$

$$\prod_{t=2}^T \frac{1}{\sqrt{2\pi\sigma_{\varepsilon 1}^2}} \exp \left[-\frac{(y_{i,t}^{[c1]} - \alpha^{[c1]} - \beta^{[c1]} y_{i,t-1}^{[c2]} - \gamma^{[c1]} y_{i,t-1}^{[c1]} - \sigma_{\varepsilon 1} u_i)^2}{2\sigma_{\varepsilon 1}^2} \right] \quad (**)$$

$$\frac{1}{\sqrt{2\pi\sigma_{\varepsilon 2}^2}} \exp \left[-\frac{(y_{i,t}^{[c2]} - \alpha^{[c2]} - \beta^{[c2]} y_{i,t-1}^{[c1]} - \gamma^{[c2]} y_{i,t-1}^{[c2]} - \sigma_{\varepsilon 2} u_i)^2}{2\sigma_{\varepsilon 2}^2} \right] \quad (***)$$

$$\frac{1}{\sqrt{2\pi\sigma_u^2}} \exp \left[-\frac{u_i^2}{2\sigma_u^2} \right] du_i \quad (****)$$

In the likelihood formula, for the parts after the integral symbol and before the derivative symbol, (*) indicates the multiplication of probability of the first observations of $y_i^{[c1]}$ and $y_i^{[c2]}$ since they have different mean model from other observations of the same subject; (**) and (***) indicates the multiplication of probability of all the observations except the first ones of $y_i^{[c1]}$ and $y_i^{[c2]}$; (****) is the marginal probability of u_i .

In general, there is no simple closed-form solution for the joint likelihood and numerical integration techniques are required. We use Gaussian quadrature to approximate the integral by a weighted sum, where the quadrature points and weights are chosen to provide a good numerical approximation. Maximization of the likelihood is done in an iterative way.

3.2. One continuous outcome and one binary outcome

To extend the case of two continuous outcomes in Section 3.1 to the case of two non-commensurate outcomes, we let $y_{i,t}^{[c]}$ and $y_{i,t}^{*[b]}$ be two normally distributed correlated continuous outcomes. Here, $y_{i,t}^{*[b]}$ is an underlying variable which is used to define the binary variable $y_{i,t}^{[b]}$. If $y_{i,t}^{*[b]} > 0$, then $y_{i,t}^{[b]} = 1$; otherwise $y_{i,t}^{[b]} = 0$. The structural regression models of $y_{i,t}^{[c]}$ and $y_{i,t}^{*[b]}$ is

$$\begin{aligned} y_{i,t}^{[c]} | y_{i,t-1}^{[b]}, y_{i,t-1}^{[c]}, u_i &= \alpha^{[c]} + \beta^{[c]} y_{i,t-1}^{[b]} + \gamma^{[c]} y_{i,t-1}^{[c]} + \sigma_{\varepsilon 1} u_i + \varepsilon_{i,t}^{[c]} \\ y_{i,t}^{*[b]} | y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i &= \alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]} + u_i + \varepsilon_{i,t}^{[b]} \end{aligned} \quad (3)$$

where $\varepsilon_{i,t}^{[c]} \sim N(0, \sigma_{\varepsilon 1}^2)$, $\varepsilon_{i,t}^{[b]} \sim N(0, 1)$ and $u_i \sim N(0, \sigma_u^2)$. The random variable u_i again is introduced into both equations to model the correlation between the outcomes. The

correlation between $y_{i,t}^{[c]}$ and $y_{i,t}^{*[b]}$ given $y_{i,t-1}^{[c]}$ and $y_{i,t-1}^{[b]}$ is $\frac{\sigma_u^2}{1 + \sigma_u^2}$. It is assumed that given u_i , $y_{i,t}^{[c]}$ and $y_{i,t}^{*[b]}$ are independent and consequently $y_{i,t}^{[c]}$ and $y_{i,t}^{[b]}$ are also independent given u_i .

We can write the regression equation for the binary outcome in the equation (3) as

$$P(y_{i,t}^{[b]} = 1 | y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}) = P(y_{i,t}^{*[b]} > 0 | y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i) = \Phi(\alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]} + u_i).$$

The final model is

$$\begin{aligned} y_{i,t}^{[c]} | y_{i,t-1}^{[b]}, y_{i,t-1}^{[c]}, u_i &= \alpha^{[c]} + \beta^{[c]} y_{i,t-1}^{[b]} + \sigma_{\varepsilon 1} u_i + \varepsilon_{i,t}^{[c]} \\ \text{Probit} \left[P(y_{i,t}^{[b]} = 1 | y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i) \right] &= \alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + u_i \end{aligned} \quad (4)$$

We also assume the first observations for both outcomes only depend on $\alpha^{[c]}$ and $\alpha^{[b]}$ respectively.

$$\begin{aligned} y_{i,1}^{[c]} | u_i &= \alpha^{[c]} + \sigma_{\varepsilon 1} u_i + \varepsilon_{i,1}^{[c]} \\ \text{Probit} \left[P(y_{i,1}^{[b]} = 1 | u_i) \right] &= \alpha^{*[b]} + u_i \end{aligned} \quad (5)$$

Based on the equation 4 and 5, the likelihood can be written as

$$\begin{aligned} & f(y_{i,1}^{[c]}, \dots, y_{i,t}^{[c]}, y_{i,1}^{[b]}, \dots, y_{i,t}^{[b]}) \\ &= \prod_{i=1}^N \int f(y_{i,1}^{[c]} | u_i) f(y_{i,1}^{[b]} | u_i) \prod_{t=2}^T f(y_{i,t}^{[c]} | y_{i,t-1}^{[b]}, y_{i,t-1}^{[c]}, u_i) f(y_{i,t}^{[b]} | y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i) f(u_i) du_i \\ &= \prod_{i=1}^N \int \frac{1}{\sqrt{2\pi\sigma_{\varepsilon 1}^2}} \exp \left[-\frac{(y_{i,1}^{[c]} - \alpha^{[c]} - \sigma_{\varepsilon 1} u_i)^2}{2\sigma_{\varepsilon 1}^2} \right] \Phi(\alpha^{*[b]} + u_i)^{y_{i,1}^{[b]}} \left[1 - \Phi(\alpha^{*[b]} + u_i)^{1-y_{i,1}^{[b]}} \right] \end{aligned} \quad (*)$$

$$\prod_{t=2}^T \frac{1}{\sqrt{2\pi\sigma_{\varepsilon 1}^2}} \exp \left[-\frac{(y_{i,t}^{[c]} - \alpha^{[c]} - \beta^{[c]} y_{i,t-1}^{[b]} - \gamma^{[c]} y_{i,t-1}^{[c]} - \sigma_{\varepsilon 1} u_i)^2}{2\sigma_{\varepsilon 1}^2} \right] \quad (**)$$

$$\Phi(\alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]} + u_i)^{y_{i,t}^{[b]}} \left[1 - \Phi(\alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]} + u_i) \right]^{1-y_{i,t}^{[b]}}$$

(***)

$$\frac{1}{\sqrt{2\pi\sigma_u^2}} \exp \left[-\frac{u_i^2}{2\sigma_u^2} \right] du_i \quad (****)$$

In the likelihood formula, for the parts after the integral symbol and before the derivative symbol, (*) indicates the multiplication of probability of the first observations of $y_i^{[c]}$ and $y_i^{[b]}$; (**) and (***) indicates the multiplication of probability of all the observations except the first ones of $y_i^{[c]}$ and $y_i^{[b]}$; (****) is the marginal probability of u .

Similar as Section 3.1, there is no simple closed-form solution for the joint likelihood and numerical integration techniques are required. We use Gaussian quadrature to approximate the integral by a weighted sum, where the quadrature points and weights are chosen to provide a good numerical approximation. Maximization of the likelihood is done in an iterative way.

The parameters in the Probit equation of equation (4) and equation (5) are interpreted conditionally on u_i . Given u_i , $\beta^{*[b]}$ and $\gamma^{*[b]}$ are the change on the probit of the expected value of $y_i^{[b]}$ for an increase of one unit in the respective covariates. As the result of that, the parameters of the underlying model cannot be directly compared with the regression parameters of the marginal models. To obtain the marginal effects, we have to average over the distribution of u_i 's.

$$\int P(y_{i,t}^{[b]} | y_{i,t-1}^{[c]}, y_{i,t-1}^{[b]}, u_i) f(u_i) du_i = \Phi \left(\frac{\alpha^{*[b]} + \beta^{*[b]} y_{i,t-1}^{[c]} + \gamma^{*[b]} y_{i,t-1}^{[b]}}{\sqrt{1+\sigma_u^2}} \right)$$

Thus, $\beta^{[b]} = \frac{\beta^{*[b]}}{\sqrt{1+\sigma_u^2}}$ and $\gamma^{[b]} = \frac{\gamma^{*[b]}}{\sqrt{1+\sigma_u^2}}$ are the marginal effects associated with the covariates. In model comparison, the estimates of $\beta^{[b]}$ and $\gamma^{[b]}$ should be used for binary

outcome. Estimates for the marginal effects $\hat{\beta}^{[b]}$ and $\hat{\gamma}^{[b]}$ are obtained using $\frac{\hat{\beta}^{*[b]}}{\sqrt{1+\sigma_u^2}}$ and

$$\frac{\hat{\gamma}^{*[b]}}{\sqrt{1+\sigma_u^2}}.$$

In the following sections, we evaluate our proposed approach through a simulation study in Section 4 and apply the model to the IU Phone study in Section 5.

4. Simulation

A Monte Carlo simulation study is used to investigate the properties of the proposed autoregressive and cross-lagged model. We are especially interested in smaller mean squared error (MSE) gained by modeling the bivariate non-commensurate processes simultaneously. Similar as the Section 3, we discuss the simulation procedure of two cases 1) two continuous outcomes and, 2) one continuous outcome and one binary outcome, respectively.

4.1. Simulation for two continuous outcomes

For the correlated commensurate outcomes case, two continuous outcomes $y_{i,t}^{[c1]}$ and $y_{i,t}^{[c2]}$ are generated with the initial setting

$$\begin{aligned} y_{i,t}^{[c1]} &= -0.5 + 0.4y_{i,t-1}^{[c2]} + 0.6y_{i,t-1}^{[c1]} + \sigma_{\varepsilon 1}u_i + \varepsilon_{i,t}^{[c1]} \\ y_{i,t}^{[c2]} &= -0.5 + 0.6y_{i,t-1}^{[c1]} + 0.3y_{i,t-1}^{[c2]} + \sigma_{\varepsilon 2}u_i + \varepsilon_{i,t}^{[c2]} \end{aligned}$$

where $u_i \sim N(0, \sigma_u^2)$, $(\varepsilon_{i,t}^{[c1]}, \varepsilon_{i,t}^{[c2]}) \sim MNV((0, 0), (\sigma_{\varepsilon 1}^2, 0, \sigma_{\varepsilon 2}^2))$. We assume $\sigma_{\varepsilon 1}^2 = 0.5$ and $\sigma_{\varepsilon 2}^2 = 1$. The correlation between two continuous outcomes at time t given two outcomes'

observations at time $(t-1)$ is defined as $r = \frac{\sigma^2}{1 + \sigma_u^2}$.

For the first observations of the two continuous outcomes, we define the generation mechanism as

$$\begin{aligned} y_{i,1}^{[c1]} &= -0.5 + \sigma_{\varepsilon 1}u_i + \varepsilon_{i,1}^{[c1]} \\ y_{i,1}^{[c2]} &= -0.5 + \sigma_{\varepsilon 2}u_i + \varepsilon_{i,1}^{[c2]} \end{aligned}$$

In our simulation study, we use the univariate GLM to model each outcome separately. Such an assumption is equivalent to modeling the two outcomes jointly without the subject-specific random effect u_i and assuming that the two residual errors are independent. We are interested in comparing the performance of the proposed method with the univariate GLMs under different data generating assumptions. In the model comparison, we test two values of the correlation: $r = 0.05$ ($\sigma_u^2 = 1/19$) as low correlation and $r = 0.03$ ($\sigma_u^2 = 3/7$) as high correlation. For the same correlation assumption, we compare two models under different length of observations per subject (4 time points and 6 time points). For the data set with 6 time points, we test the balanced full data and the data with missing complete at random (MCAR) mechanism. Specifically, in each balanced full data set, we generate 6 sequential time points for both continuous outcomes for all subjects. In data with MCAR, we randomly select subjects from the full data set and restrict their length of time points to 5, 4, 3 and 2 separately. The difference between each adjacent length level is 10 subjects. If we let the n represent the sample size, we end up with n subjects with observations at first two time points; $n - 10$ subjects with observations at first three time points; $n - 20$ subjects with observations at first four time points and so on. For all the scenarios above, we compare models with the sample sizes $n = 100$ and $n = 200$, respectively.

In each data set, we use the proposed method to model the two outcomes simultaneously and used univariate GLMs to model each outcome separately. PROC NLMIXED in SAS version 9.3 is used to execute both models to assure that same numerical algorithms are used in likelihood maximization. Each simulation procedure is repeated 500 times. We compare the percentage of the bias and MSE of estimates from the proposed model and the univariate GLMs based on the 500 simulation results under different data settings.

The percentage of bias and MSE of estimates of two commensurate outcomes (two continuous outcomes) under different correlation ($r=0.05$ and $r=0.3$), missing mechanism (full data and MCAR), sample size ($N=100$ and $N=200$) and sequence length ($p=6$ and $p=4$) for the proposed model (red triangle) and the univariate GLMs (blue circle) are displayed in Figure 2. The summary of the estimation results under the low correlation assumption ($r=0.05$) are displayed in the top row of Figure 2 and the results under high correlation assumption ($r=0.30$) are displayed in the bottom row of Figure 2. Based on the simulation results, the proposed model provide estimates with smaller percentage of bias and smaller MSE comparing to univariate GLMs. Higher correlation between two outcomes increases the bias and MSE of coefficient estimates under univariate GLMs, especially for the intercepts in the model; but have little influence on the proposed model. Data missingness slightly decreases the bias and MSE of intercept estimates and slightly increases the bias and MSE of autoregressive and cross-lagged estimates for univariate GLMs when there is a higher correlation between two outcomes, but have no influence on the proposed model. Larger sample size decreases the MSE for both models but have very small influence on bias. Shorter length of the observed sequence for each subject decreases the bias and MSE of intercept estimates and increases the bias and MSE of autoregressive and cross-lagged effect estimates for univariate GLMs, but have little influence on the proposed model.

4.2. Simulation for one continuous outcome and one binary outcome

To deal with the one continuous outcome and one binary outcome case, two continuous variables $y_{i,t}^{[c]}$ and $y_{i,t}^{*[b]}$ are generated with the initial setting as below. Here, $y_{i,t}^{*[b]}$ is a latent variable of $y_{i,t}^{[b]}$.

$$\begin{aligned} y_{i,1}^{[c]} &= -0.5 + 0.4y_{i,t-1}^{[b]} + 0.6y_{i,t-1}^{[c]} + \sigma_{\varepsilon 1}u_i + \varepsilon_{i,t}^{[c]} \\ y_{i,1}^{*[b]} &= -0.5 + 0.6y_{i,t-1}^{[c]} + 0.3y_{i,t-1}^{[b]} + \sigma_{\varepsilon 2}u_i + \varepsilon_{i,t}^{[b]} \end{aligned}$$

where $u_i \sim N(0, \sigma_u^2)$, $(\varepsilon_{i,t}^{[c]}, \varepsilon_{i,t}^{[b]}) \sim MNV((0, 0), (\sigma_{\varepsilon 1}^2, 0, 0, \sigma_{\varepsilon 2}^2))$. We assume $\sigma_{\varepsilon 1}^2 = 0.5$ and $\sigma_{\varepsilon 2}^2 = 1$.

For the first observations, we define the generation mechanism as

$$\begin{aligned} y_{i,1}^{[c]} &= -0.5 + \sigma_{\varepsilon 1}u_i + \varepsilon_{i,t}^{[c]} \\ y_{i,1}^{*[b]} &= -0.5 + \sigma_{\varepsilon 2}u_i + \varepsilon_{i,t}^{[b]} \end{aligned}$$

In each simulation, we keep the $y_{i,t}^{[c]}$ as the continuous outcome and then create the binary outcome $y_{i,t}^{[b]}$ based on the rule that if the $P(Z < y_{i,t}^{*[b]}) > 0.5$, then $y_{i,t}^{[b]} = 1$; otherwise $y_{i,t}^{[b]} = 0$, where Z is the quantile of the standard normal distribution. The first observation of the binary outcome is generated by following the same transformation procedure. The correlation between the continuous outcome and the underlying continuous outcome of the binary outcome at time t given two outcomes' observations at time $(t-1)$ equals to $\frac{\sigma_u^2}{1+\sigma_u^2}$.

We follow the same model performance comparison procedure in Section 4.1. The percentage of bias and MSE of estimates of two non-commensurate outcomes (one continuous outcome and one binary outcome) for the proposed model and the univariate GLMs are presented in Figure 3. The proposed model provides estimates with smaller percentage of bias and smaller MSE comparing to univariate GLMs. The estimates of intercepts from two models of one continuous and one binary outcome are closer than the intercept estimates of two models of two continuous outcomes. Higher correlation between two outcomes increases the bias and MSE of coefficient estimates under univariate GLMs, but nearly no influence on proposed model. Data missingness slightly decreases the bias of intercepts estimates and slightly increases the bias and MSE of the autoregressive and cross-lagged estimates for univariate GLMs when there is a higher correlation between two outcomes, but have no influence on the proposed model. Larger sample size decreases the MSE for both models but have very small influence on bias. Shorter length of observed sequence for each subject decreases the bias and MSE of intercepts estimates and increases the bias and MSE of autoregressive and cross-lagged effect estimates for univariate GLMs, but have no much influence on the proposed model.

Based on the summary of the simulation results, the correlation between two outcomes becomes the main factor that distinguish the performance of the proposed model and the univariate GLMs. The proposed model provides consistent estimates with smaller bias and MSE under different assumptions of the data structure. We also check the mean standard error (SE) of the estimates in both univariate GLM and the proposed model from the simulation result. The SE of some estimates from the proposed model is slightly larger than the univariate GLM. Similar results have been found in Teixeira-Pinto and Normand [14] that in a multivariate model for non-commensurate multiple outcomes, there is no gains in efficiency for the multivariate model compared to the univariate models.

5. Application to the IU Phone Study

In the motivating data example, we are interested in studying the relationship between condom use behavior and sexual satisfaction, which are reported by participants in each 8-hour reporting period. We model the association between the evaluation of the sexual satisfaction and condom use during the next coital event via cross-lagged effects in our model. In an analogous fashion, the association between the current condom use and the sexual satisfaction evaluation to the current event is captured via a different cross-lagged effect. In addition, we model the within-outcome associations via autoregressive effects.

In this study, we use coital event order as the time scale, since participants might not report sexual events during every reporting period. This assumption also allows us to directly utilize the proposed model, since the event-scale interval is consistent within and across the study participants. Though the information on condom use and sexual satisfaction are collected at the same time, the sexual satisfaction reflects the evaluation of the coital event that happened before the reporting time, but after the condom use. As a result, the sexual satisfaction evaluation has in practice a minor lag from the coital event reported in the same period. However, the event intervals between the adjacent events are the same for sexual satisfaction and condom use, since they are corresponding to the same coital event. The path graph of the relationship between these two outcomes of interest is illustrated in Figure 4. One cross-lagged regression coefficient ($\beta^{[c]}$) estimates the association between the condom use at event t and sexual satisfaction at event t . An analogous cross-lagged regression coefficient ($\beta^{[b]}$) estimates the association between the sexual satisfaction at event t and condom use probability at event $(t + 1)$. An autoregressive coefficient ($\gamma^{[b]}$) estimates the association between the condom use at event t and condom use at event $(t + 1)$. A similar autoregressive coefficient ($\gamma^{[c]}$) estimates the association between sexual satisfaction at event t and sexual satisfaction at event $(t + 1)$. From the scientific point of view, the cross-lagged associations are of primary interest for understanding reciprocal influences between condom use behaviors and sexual satisfaction over time.

All 254 participants included in our analysis have at least two coital events reported during the study period. There are very few subjects who have more than 20 coital reported. Thus, we restrict our analysis to the first 20 participant-specific events. We analyze the data using both the proposed model and the univariate GLMs. P-values are provided to show the significance of the estimates. The results of proposed model show that the cross-lagged effect ($\beta^{[c]} = -0.775$, $p < 0.0001$) of condom use on sexual satisfaction and the cross-lagged effect ($\beta^{[b]} = -0.144$, $p < 0.0001$) of sexual satisfaction on condom use are both negative and significant, which indicate the use of the condom is negatively associated the high sexual satisfaction of the same coital event and higher sexual satisfaction is negatively associated to the probability of condom use of next coital event. Both of the autoregressive effect of sexual satisfaction ($\gamma^{[c]} = 0.091$, $p < 0.0001$) and condom use ($\gamma^{[b]} = 2.066$, $p < 0.0001$) are positive and significant. The evaluation of sexual satisfaction and the behavior of condom use of next coital event have positive association with the sexual evaluation and condom use of previous coital event respectively. On the other side, the estimates from GLMs did not change the conclusion of our findings in the proposed models. However, the univariate GLMs have smaller absolute values of cross-lagged effect estimates ($\beta_{GLM}^{[c]} = -0.462$, $p < 0.0001$ and $\beta_{GLM}^{[b]} = -0.102$, $p < 0.0001$) than the proposed model, while larger absolute values of autoregressive effect estimates ($\gamma_{GLM}^{[c]} = 0.159$, $p < 0.0001$ and $\gamma_{GLM}^{[b]} = 2.109$, $p < 0.0001$) than the proposed model. We also get the correlation between the sexual satisfaction and assumed latent variable of condom use of the same coital event as 0.869 based on the estimate of the variance of the shared random variable in the proposed model. This finding is an additional strong argument for using our proposed model rather than GLMs to model the association between these two correlated outcomes. All the results of two models are presented in the Table 1 for comparison.

6. Discussion

The proposed model provides joint estimates of the relationship between two non-commensurate longitudinal outcomes (one continuous and one binary), which extends the commonly used autoregressive and cross-lagged models. By introducing a common subject-specific random effect to estimate the correlation between two correlated outcomes, we combine the univariate mixed model methodology with the cross-lagged models to model correlated bivariate longitudinal outcomes, which relaxes the independent error assumption in the univariate GLMs. Traditional panel models treat all autoregressive and cross-lagged effects as fixed without considering the variation among subjects. Inclusion of the subject-specific random effects in the proposed model accounts for between-subject variability arising from the omitted subject-level predictors. We include both cross-lagged and autoregressive effects in the model in order to minimize bias in the estimation of cross-lagged effects, which is supported by the compelling arguments made by Gollob and Reichardt [18] and Cole and Maxwell [19]. The estimates obtained from the proposed model in the simulation studies are consistent and have smaller variability than the estimated obtained from the ordinary GLMs. The current model is developed for one continuous outcome and one binary outcome, however, this likelihood-based approach can be applied to outcomes with different measurement types. The proposed model also does not require the data to have a balanced structure and can be used when subjects contribute varying numbers of events. However, we require the inter-observation time intervals to be the same.

In the real data application, we employ the proposed model to the EMA longitudinal dataset. We are able to depict the timing and sequencing link of condom use and sexual satisfaction. In our analysis, we assume that the sexual satisfaction follows a normal distribution to simplify the illustration. In further studies, the use of either skewed-normal or a Beta distribution rescaled to the interval 1 to 10 might be a better distributional assumption as the empirical distribution of the sexual satisfaction is heavily left-skewed. In the IU Phone study example, we use coital event order instead of chronological time sequence to define the “time lag”. However, the chronological time between every two adjacent events might be different from event to event. This choice of time lag might influence the estimates. Till now, all the autoregressive and cross-lagged models required the outcomes share the same time lag across subjects. In the data arising in the observational studies this assumption might not be satisfied. In the future model development, we plan to combine the variable-domain functional regression method established by Gellar et al. [20] with the proposed model to deal with the sparse outcomes with different inter-event time intervals.

Acknowledgments

This project was supported by grant R01 HD053231-01A1 to Dr. J. Dennis Fortenberry from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). Dr. Teixeira-Pinto is partially supported by the NHMRC Program Grant BeatCKD (APP1092957). Dr. Harezlak's research was partially supported by the NIMH R01 grant MH108467.

References

1. Campbell DT. From description to experimentation: Interpreting trends as quasiexperiments. *Problems in measuring change*. 1963:212–242.

2. Bohrnstedt GW. Observations on the measurement of change. *Sociological methodology*. 1969; 1:113–133.
3. Duncan OD. Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin*. 1969; 72:177–182.
4. Heise DR. Separating reliability and stability in test-retest correlation. *American Sociological Review*. 1969; 34:93–101.
5. Joreskog, KG., Sorbom, D. *Advances in factor analysis and structural equation models*. Cambridge, Mass: Abt; 1979.
6. Kessler, RC., Greenberg, DF. *Linear Panel Analysis*. New York: Academic Press; 1981.
7. Bollen KA, Curran PJ. Autoregressive Latent Trajectory (ALT) Models A Synthesis of Two Traditions. *Sociological Methods & Research*. 2004; 32(3):336–383.
8. Tate RF. Correlation between a discrete and a continuous variable. *Annals of Mathematical Statistics*. 1954; 25:603–607.
9. Olkin I, Tate RF. Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics*. 1961; 32:448–465.
10. Little RJA, Schluchter MD. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*. 1985; 72:497–512.
11. Krzanowski WJ. *Principles of Multivariate Analysis*. 1988
12. Cox DR, Wermuth N. Response models for mixed binary and quantitative variables. *Biometrika*. 1992; 79:441–461.
13. Cox DR, Wermuth N. A note on the quadratic exponential binary distribution. *Biometrika*. 1994; 81:403–408.
14. Teixeira-Pinto A, Normand SL. Correlated bivariate continuous and binary outcomes: issues and applications. *Statistics in Medicine*. 2009; 28(13):1753–1773. [PubMed: 19358234]
15. Catalano PJ, Ryan LM. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*. 1992; 87:651–658.
16. Fei H, Hensel DJ, Fortenberry JD, Harezlak J. Condom Use as a Function of Number of Coital Events in New Relationships. *Sex Transmitted Disease*. 2016; 43(2):67–70.
17. Hensel DJ, Fortenberry JD, Harezlak J, Craig D. The feasibility of cell phone based electronic diaries for STI/HIV research. *BMC Medical Research Methodology*. 2012; 12(75)
18. Gollob HF, Reichardt CS. Taking account of time lags in causal models. *Child Development*. 1987; 58:80–92. [PubMed: 3816351]
19. Cole DA, Maxwell SE. Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*. 2003; 112:558–577. [PubMed: 14674869]
20. Gellar JE, Elizabeth C, Dale MN, Ciprian MC. Variable-Domain Functional Regression for Modeling ICU Data. *Journal of the American Statistical Association*. 2014; 109(508):1425–1439. [PubMed: 25663725]

Figure 1.

Figure 2.

Figure 3.

Figure 4.

Table 1

Estimates comparison between proposed model and univariate GLMs on IU Phone study

Models	Proposed model		Univariate GLMs	
	Estimates	Standard Error	Estimates	Standard Error
$\alpha^{[c]}$	8.562	0.088	7.956	0.083
$\beta^{[c]}$	-0.775	0.079	-0.462	0.074
$\gamma^{[c]}$	0.091	0.009	0.159	0.009
$\alpha^{[b]}$	-0.356	0.090	-0.587	0.077
$\beta^{[b]}$	-0.144	0.010	-0.102	0.009
$\gamma^{[b]}$	2.066	0.085	2.109	0.075
χ^2_u	0.435	0.052	–	–