Estimation of Effects of Endogenous Time-Varying Covariates: A Comparison Of Multilevel Linear Modeling and Generalized Estimating Equations

Research Report

Ward B. Eiling (9294163)

Supervisors: Ellen Hamaker and Jeroen Mulder

Master's degree in Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Utrecht University

December 22, 2024

Word count: 728

FETC-approved: 24-2003

Candidate journal: Psychological Methods

1 Introduction

Across a wide range of disciplines, researchers analyze clustered longitudinal, observational data to investigate prospective causal relationships between variables. When analyzing such data, the psychological sciences most commonly resort to the multilevel linear model (MLM, McNeish et al., 2017), which—in the context of longitudinal data analysis—separates observed variance into stable between-person differences and within-person fluctuations (Hamaker & Muthén, 2020). Conversely, other fields, such as biostatistics and econometrics often favour generalized estimating equations (GEE) for the analysis of longitudinal data (McNeish et al., 2017). Despite some cross-disciplinary efforts to compare these methods (McNeish et al., 2017; Muth et al., 2016; Yan et al., 2013), their scarcity may leave researchers with limited guidance in choosing the most suitable approach for their application.

Recent evidence has highlighted an issue present in both methods, where controlling for *time-varying endogenous covariates* may lead to biased causal estimates (Pepe & Anderson, 1994; Qian et al., 2020). A time-varying covariate is *endogenous* if it is directly or indirectly influenced by prior treatment or outcome, meaning its value may be determined by earlier stages of the process (Qian et al., 2020). As a result of including these covariates in the mentioned models, ordinary interpretations of the coefficients are no longer valid (Qian et al., 2020, p. 3). According to Diggle (2002), this issue not only pertains GEE and MLM, but *all* longitudinal data analysis methods.

However, due to a divide between the disciplines that employ these methods, such critiques of the MLM appear to have largely failed to reach the applied researcher in psychology. One specific reason might be that the technical jargon in other disciplines makes it difficult for researchers to recognize when and how these issues emerge[1]. As a result, researchers may address related problems in disconnected literatures but fail to understand each other. For instance, while the MLM literature emphasizes on the distinction between different centering methods and the effect of cross-level interactions on parameter interpretations (e.g., Hamaker & Muthén, 2020), the GEE literature appears to focus more on the marginal and conditional interpretations of model parameters (e.g., Pepe & Anderson, 1994).

Through a cross-fertilization of these literatures, this project aims to (1) explain the issue of including endogenous covariates in analyses involving GEE, MLM and DSEM (a widely used framework in the social sciences based on MLM) in a psychological context and (2) establish guidelines on how researchers can prevent this issue in their longitudinal data analysis. Accordingly, the following research questions will be addressed: *In which cases do the inclusion of endogenous variables in multilevel linear models and generalized estimating equations result in a discrepancy between marginal and conditional estimates?* In line with the literature (Diggle, 2002; Pepe & Anderson, 1994; Qian et al., 2020), we expect that the inclusion of endogenous time-varying covariates in longitudinal data analyses may result in bias that—depending on the circumstances—can promote the potential for faulty inferences. To isolate the issue described in Qian et al. (2020), we will focus on the following sub-questions: (1) When removing the interaction β_1 from generative model 3, is there a difference between the marginal and conditional estimates of the treatment effect? (2) When removing the random slope b_{i2} from generative model 3, is there a difference between the marginal and conditional estimates of the treatment effect?

[1] For instance, the term ‘endogeneity’ in econometrics, while related, has a distinct meaning from that of an endogenous variable, which can lead to confusion.

2 Methods

2.1 Data Generation

In the simulation Qian et al. (2020) considered three generative models (GMs), all of which have an endogenous time-varying covariate. In GM1 and GM2, the endogenous covariate X_{it} equals the previous outcome Y_{it} plus some random noise, so the *conditional independence* assumption is valid. In GM3, the endogenous covariate depends directly on b_{i0}, violating the assumption. To isolate the issue in GM3, we consider three variations on this model: GM3A, where the random slope b_{i2} for the treatment A_{it} is removed; GM3B, where the interaction term $\beta_1 A_{it} X_{it}$ is removed. The details of the generative models are described below. We follow the notation of Qian et al. (2020) to allow for direct comparison, but rewrite the equations into within- and between-person models (see Raudenbush & Bryk, 2002). We accompany the equations of the GMs with graphical representations, where random effects are represented by grey circles, observed variables by squares and relationships across variables by arrows. The path diagrams of the three data generating models shows the discrepancies between the different generative models—especially concerning the interaction effects—more clearly than DAGs.

2.1.1 Generative Model 1

In GM1, we considered a simple case with only a random intercept and a random slope for X_{it}. The outcome is generated according to the following repeated-observations or within-person model (level 1):

$$Y_{it+1} = \pi_{0i} + \pi_{1i} X_{it} + \pi_{2i} A_{it} + \pi_{3i} A_{it} X_{it} + \epsilon_{it+1}$$

with the person-level or between-person model (level 2):

$$\pi_{0i} = \alpha_0 + b_{i0}, \quad b_{i0} \sim \mathcal{N}(0, \sigma^2_{b0}),$$

$$\pi_{1i} = \alpha_1,$$

$$\pi_{2i} = \beta_0 + b_{i2}, \quad b_{i2} \sim \mathcal{N}(0, \sigma^2_{b2}),$$

$$\pi_{3i} = \beta_1.$$

By substitution, we get the single equation model:

$$\begin{aligned} Y_{it+1} &= \pi_{0i} + \pi_{1i} X_{it} + \pi_{2i} A_{it} + \pi_{3i} A_{it} X_{it} + \epsilon_{it+1} \\ &= (\alpha_0 + b_{i0}) + \alpha_1 X_{it} + (\beta_0 + b_{i2}) A_{it} + \beta_1 A_{it} X_{it} + \epsilon_{it+1} \\ &= \alpha_0 + \alpha_1 X_{it} + b_{i0} + A_{it}(\beta_0 + \beta_1 X_{it} + b_{i2}) + \epsilon_{it+1}. \end{aligned}$$

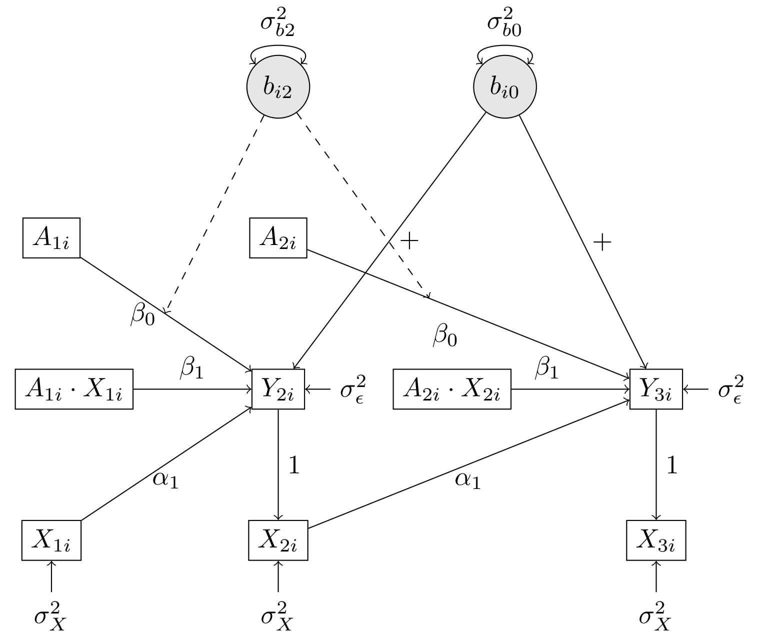
The random effects $b_{i0} \sim \mathcal{N}(0, \sigma^2_{b0})$ and $b_{i2} \sim \mathcal{N}(0, \sigma^2_{b2})$ are independent of each other. The covariate is generated as $X_{i1} \sim \mathcal{N}(0, 1)$, and for $t \geq 2$,

$$X_{it} = Y_{it} + \mathcal{N}(0, 1).$$

The randomization probability $p_t = P(A_{it} = 1 \mid H_{it})$ is constant at 1/2. Thus, $A_{it} \sim$ Bernoulli(0.5) for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The exogenous noise is $\epsilon_{it+1} \sim \mathcal{N}(0, \sigma^2_\epsilon)$.

Figure 1 shows the path diagram for GM3.

Figure 1: Path diagram for Generative Model 1 ($t = 1, 2, 3$)



2.1.2 Generative Model 2

In GM2, we considered the case with a random intercept and random slopes for (1) covariate X_{it}, (2) treatment A_{it}, and (3) the interaction between A_{it} and X_{it}; and with a time-varying randomization probability for treatment. The outcome is generated according to the same repeated-observations model presented in GM1. However, the person-level model is different:

$$\pi_{0i} = \alpha_0 + b_{i0}, \quad b_{i0} \sim \mathcal{N}(0, \sigma^2_{b0}),$$

$$\pi_{1i} = \alpha_1 + b_{i1}, \quad b_{i1} \sim \mathcal{N}(0, \sigma^2_{b1}),$$

$$\pi_{2i} = \beta_0 + b_{i2}, \quad b_{i2} \sim \mathcal{N}(0, \sigma^2_{b2}),$$

$$\pi_{3i} = \beta_1 + b_{i3}, \quad b_{i3} \sim \mathcal{N}(0, \sigma_{b3}^2).$$

By substitution, we get the single equation model:

$$\begin{aligned} Y_{it+1} &= \pi_{0i} + \pi_{1i} X_{it} + \pi_{2i} A_{it} + \pi_{3i} A_{it} X_{it} + \epsilon_{it+1} \\ &= (\alpha_0 + b_{i0}) + (\alpha_1 + b_{i1}) X_{it} + (\beta_0 + b_{i2}) A_{it} + (\beta_1 + b_{i3}) A_{it} X_{it} + \epsilon_{it+1} \\ &= \alpha_0 + \alpha_1 X_{it} + b_{i0} + b_{i1} X_{it} + A_{it} \left(\beta_0 + \beta_1 X_{it} + b_{i2} + b_{i3} X_{it} \right) + \epsilon_{it+1}. \end{aligned}$$

The random effects $b_{ij} \sim \mathcal{N}(0, \sigma_{bj}^2)$, for $j = 0, 1, 2, 3$, are independent of each other. The covariate is generated as $X_{i1} \sim \mathcal{N}(0, 1)$, and for $t \geq 2$,

$$X_{it} = Y_{it} + \mathcal{N}(0, 1).$$

The randomization probability depends on X_{it}:

$$p_t = P(A_{it} = 1 \mid H_{it}) = \begin{cases} 0.7 & \text{if } X_{it} > -1.27, \\ 0.3 & \text{if } X_{it} \leq -1.27, \end{cases}$$

where the cutoff -1.27 was chosen so that p_t equals 0.7 or 0.3 for about half of the time. In other words, if the value of the covariate for any given person and time point is above the cutoff, the probability of receiving the treatment p_t is 0.7; otherwise, it is 0.3. Accordingly, $A_{it} \sim \text{Bernoulli}(p_t)$ for $i = 1, \dots, N$ and $t = 1, \dots, T$. The exogenous noise is $\epsilon_{it+1} \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

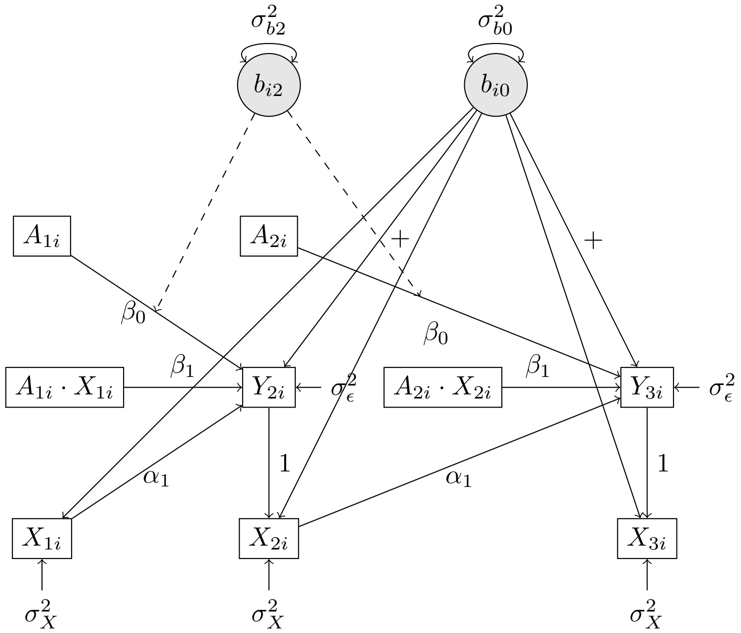
2.1.3 Generative Model 3

GM3 is the same as GM1, except that the covariate X_{it} depends directly on b_{i0}:

$$X_{i1} \sim \mathcal{N}(b_{i0}, 1), \quad X_{it} = Y_{it} + \mathcal{N}(b_{i0}, 1) \text{ for } t \geq 2.$$

Figure 2 shows the path diagram for GM3.

Figure 2: Path diagram for Generative Model 3 ($t = 1, 2, 3$)

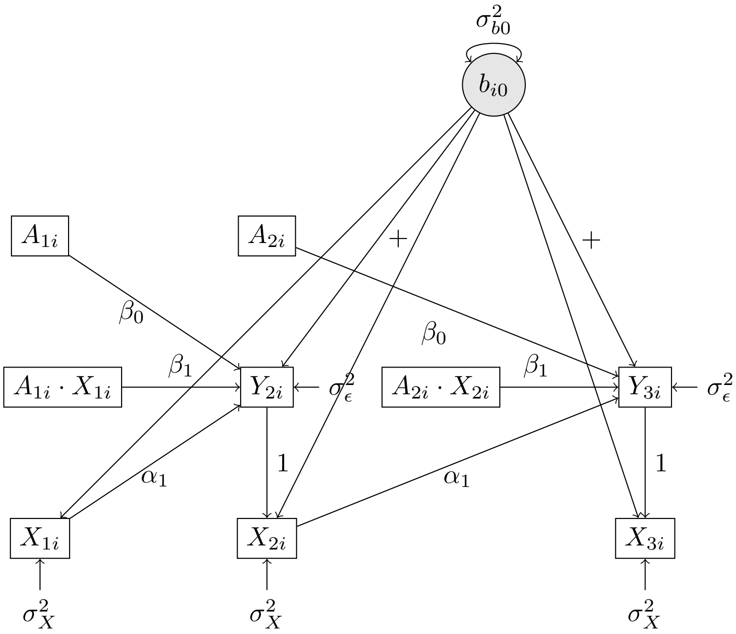


2.1.4 Generative Model 3A

GM3A is the same as GM3, except that the random slope b_{i2} for the treatment A_{it} is removed. The single equation model then becomes:

$$Y_{it+1} = \alpha_0 + \alpha_1 X_{it} + b_{i0} + A_{it}(\beta_0 + \beta_1 X_{it}) + \epsilon_{it+1}.$$

Figure 3: Path diagram for Generative Model 3A ($t = 1, 2, 3$)

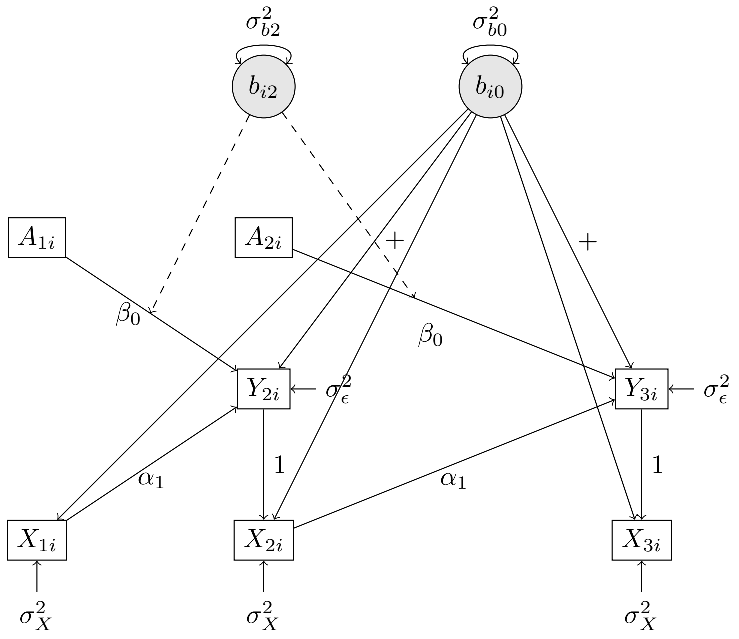


2.1.5 Generative Model 3B

GM3B is the same as GM3, except that the interaction term $\beta_1 A_{it} X_{it}$ is removed. The single equation model then becomes:

$$Y_{it+1} = \alpha_0 + \alpha_1 X_{it} + b_{i0} + A_{it}(\beta_0 + b_{i2}) + \epsilon_{it+1}.$$

Figure 4: Path diagram for Generative Model 3B ($t = 1, 2, 3$)



2.1.6 Parameter Values

The following parameter values were adapted from Qian et al. (2020):

$$\alpha_0 = -2, \quad \alpha_1 = -0.3, \quad \beta_0 = 1, \quad \beta_1 = 0.3,$$

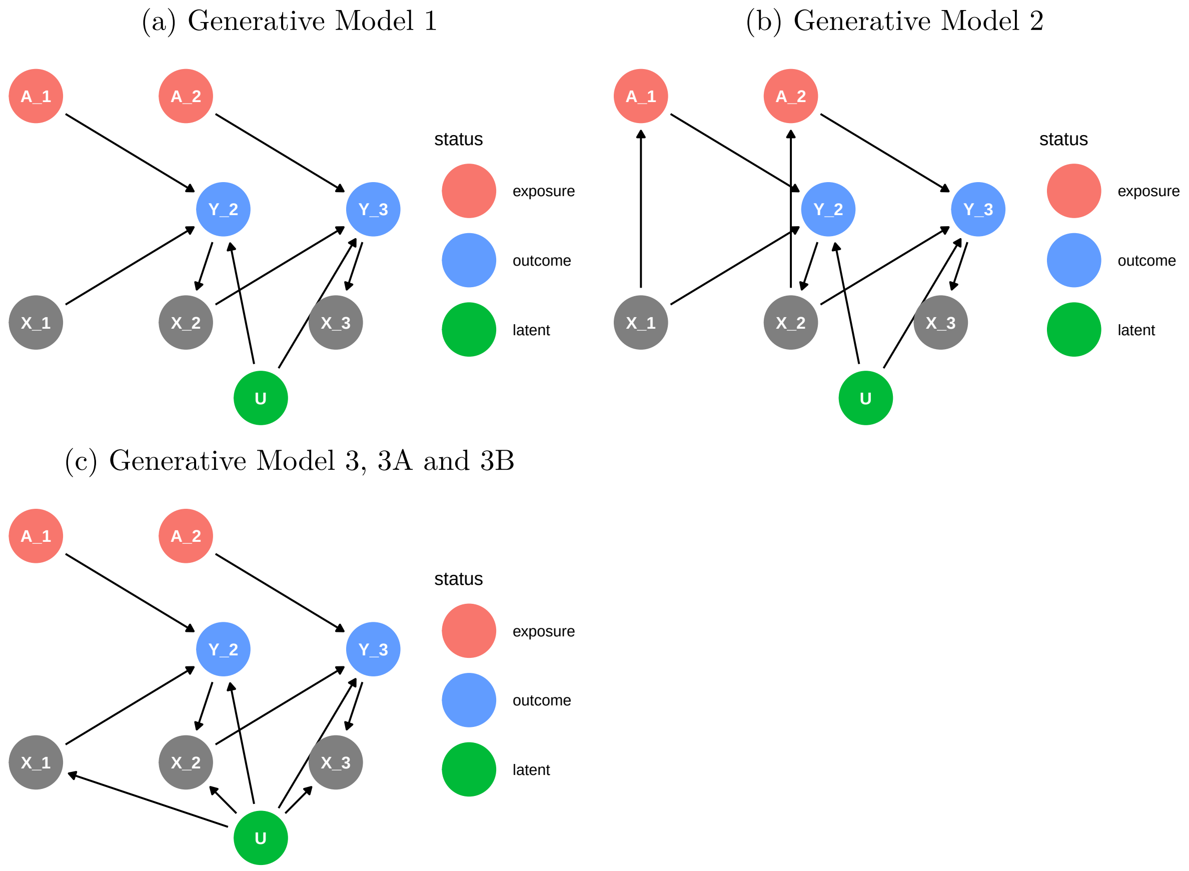
$$\sigma^2_{b0} = 4, \quad \sigma^2_{b1} = \frac{1}{4}, \quad \sigma^2_{b2} = 1, \quad \sigma^2_{b3} = \frac{1}{4}, \quad \sigma^2_\epsilon = 1.$$

2.2 Graphical representations of Data Generating Models: DECIDE WHAT TO DO WITH THIS!!!

2.2.1 Directed Acyclic Graphs (DAGs)

The DAGs for the first three observations of the three data generating models are presented in Figure 5. The red arrows show the biased paths after controlling for the covariate X_{it}.

Figure 5: DAG for Generative Model 1



(a) Generative Model 1

(b) Generative Model 2

(c) Generative Model 3, 3A and 3B

We may notice that the DAGs for GM1 and GM2 are identical (there are only differences in random effects and randomization probabilities), while GM3 has a different structure due to the dependency of the covariate X_{it} on the random intercept b_{i0}.

Paraphrasing Qian et al. (2020), the conditional independence assumption is:

$$X_{it} \perp (b_{i0}, b_{i1}) \mid H_{it-1}, A_{it-1}, Y_{it}.$$

This allows X_{it} to be endogenous, but the endogenous covariate X_{it} can only depend on the random effects through variables observed prior to X_{it}: H_{it-1}, A_{it-1}, and Y_{it}. If the only endogenous covariates are functions of prior treatments and prior outcomes, then the assumption automatically holds.

When inspecting Figure 5, we can see that this assumption is violated in GM3, as

X_{it} depends directly on b_{i0} and is thus not independent of the random effects b_{i0} and b_{i1}. Notice that GM1 and GM2 are also not marginally independent of b_{i0} and b_{i1}, but they are conditionally independent given H_{it-1}, A_{it-1}, and Y_{it}.

2.3 Data Analysis

We evaluated the performance of the models across a total of 30 different settings, each replicated 1,000 times, by systematically varying the following factors:

- **Generative Models (GM):** 1, 2, 3, 3A, 3B
- **Number of timepoints (T):** 10, 30
- **Sample size (N):** 30, 100, 200

All data generation and estimation was performed in `R`, version 4.4.2 (Team, 2024). To fit the standard MLM, the `lmer` function from the R-package `lme4` (Bates et al., 2015) was employed with restricted maximum likelihood estimation. For the MLM, the analytical models were equivalent to each of the respective data-generating models. To fit the GEE with the "exchangeable", "independent" and "AR(1)" working correlation structures, the `geeglm` function from the R-package `geepack` (Halekoh et al., 2006) was employed with the identity link function. Unsurprisingly, since the GEE does not explicitly model random effects, the specification of the analytical GEE models is different the analytical MLM models for all GMs. Compared to their MLM counterparts, the analytical GEE models simply excluded the random effects. Since the fixed effects modeled in GM1, GM2, GM3, GM3a are the same (the only differences pertain to the modeling of random effects), the analytical *GEE model* is identical across these conditions:

$$Y_{it+1} = \alpha_0 + \alpha_1 X_{it} + \beta_0 A_{it} + \beta_1 A_{it} X_{it} + \epsilon_{it+1}.$$

In GM3b, the fixed interaction effect is removed, so the analytical *GEE model* is given by:

$$Y_{it+1} = \alpha_0 + \alpha_1 X_{it} + \beta_0 A_{it} + \epsilon_{it+1}.$$

3 Results

As shown in Table X, GM3 results in bias

4 Discussion

5 References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 148. https://doi.org/10.18637/jss.v067.i01

Diggle, P. (2002). *Analysis of Longitudinal Data*. OUP Oxford.

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The r package geepack for generalized estimating equations. *Journal of Statistical Software*, *15/2*, 111.

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140. https://doi.org/10.1037/met0000078

Muth, C., Bales, K. L., Hinde, K., Maninger, N., Mendoza, S. P., & Ferrer, E. (2016). Alternative Models for Small Samples in Psychological Research: Applying Linear Mixed Effects Models and Generalized Estimating Equations to Repeated Measures Data. *Educational and Psychological Measurement*, *76*(1), 64–87. https://doi.org/10.1177/0013164415580432

Pepe, M. S., & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation*, *23*(4), 939–951. https://doi.org/10.1080/03610919408813210

Qian, T., Klasnja, P., & Murphy, S. A. (2020). Linear mixed models with endogenous covariates: Modeling sequential treatment effects with application to a mobile health study. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, *35*(3), 375–390. https://doi.org/10.1214/19-sts720

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). SAGE.

Team, R. C. (2024). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Yan, J., Aseltine, R. H., & Harel, O. (2013). Comparing Regression Coefficients Between Nested Linear Models for Clustered Data With Generalized Estimating Equations. *Journal of Educational and Behavioral Statistics*, *38*(2), 172–189. https://doi.org/10.3102/1076998611432175

6 Appendix

6.1 Original Section from Qian et al. (2020): "4. Simulation"

In the simulation, we considered three generative models (GMs), all of which have an endogenous covariate. In the first two GMs, the endogenous covariate X_{it} equals the previous outcome Y_{it} plus some random noise, so the conditional independence assumption (10) is valid. In GM 3, the endogenous covariate depends directly on b_i, violating assumption (10). The details of the generative models are described below.

In GM1, we considered a simple case with only a random intercept and a random slope for A_{it}, so that $Z_{i(t_0)} = Z_{i(t_2)} = 1$ in model (7). The outcome is generated as:

$$Y_{it+1} = \alpha_0 + \alpha_1 X_{it} + b_{i0} + A_{it}(\beta_0 + \beta_1 X_{it} + b_{i2}) + \epsilon_{it+1}.$$

The random effects $b_{i0} \sim N(0, \sigma_{b0}^2)$ and $b_{i2} \sim N(0, \sigma_{b2}^2)$ are independent of each other.

The covariate is generated as $X_{i1} \sim N(0,1)$, and for $t \geq 2$,

$$X_{it} = Y_{it} + N(0,1).$$

The randomization probability p_t is constant at 1/2. The exogenous noise is $\epsilon_{it+1} \sim N(0, \sigma_\epsilon^2)$.

In GM2, we considered the case where $Z_{i(t_0)} = Z_{i(t_2)} = 1$, with time-varying randomization probability. The outcome is generated as:

$$Y_{it+1} = \alpha_0 + \alpha_1 X_{it} + b_{i0} + b_{i1} X_{it} + A_{it}(\beta_0 + \beta_1 X_{it} + b_{i2} + b_{i3} X_{it}) + \epsilon_{it+1}.$$

The random effects $b_{ij} \sim N(0, \sigma_{b_j}^2)$, for $0 \leq j \leq 3$, are independent of each other. The covariate is generated as $X_{i1} \sim N(0,1)$, and for $t \geq 2$,

$$X_{it} = Y_{it} + N(0,1).$$

The randomization probability depends on X_{it}:

$$p_t = 0.7 \cdot 1(X_{it} > -1.27) + 0.3 \cdot 1(X_{it} \leq -1.27),$$

where $1(\cdot)$ represents the indicator function, and the cutoff -1.27 was chosen so that p_t equals 0.7 or 0.3 for about half of the time. The exogenous noise is $\epsilon_{it+1} \sim N(0, \sigma_\epsilon^2)$.

GM3 is the same as GM 1, except that the covariate X_{it} depends directly on b_i:

$$X_{i1} \sim N(b_{i0}, 1), \quad X_{it} = Y_{it} + N(b_{i0}, 1) \text{ for } t \geq 2.$$

We chose the following parameter values:

$$\alpha_0 = -2, \quad \alpha_1 = -0.3, \quad \beta_0 = 1, \quad \beta_1 = 0.3,$$

$$\sigma_{b0}^2 = 4, \quad \sigma_{b1}^2 = \frac{1}{4}, \quad \sigma_{b2}^2 = 1, \quad \sigma_{b3}^2 = \frac{1}{4}, \quad \sigma_{\epsilon}^2 = 1.$$

6.2 Overview of Variations on Generative Model 3

Table 1: Models with 1 Parameter Less

Generative Model	random slope treatment b_{i2}	interactie β_1	fixed slope covariate α_1	bias
3	yes	yes	yes	yes, negative
3a	no	yes	yes	no
3d	yes	no	yes	no
3h	yes	yes	no	yes, positive

Generative Model	random slope treatment b_{i2}	interactie β_1	fixed slope covariate α_1	bias
3	yes	yes	yes	yes, negative
3b	no	no	yes	no
3i	no	yes	no	
3j	yes	no	no	

6.3 Simulation Plan Proposal

To uncover the undesirable effects of endogenous covariates and investigate robustness against these effects, we will carry out simulations in which data will be generated according to several increasingly complex scenarios. These scenarios will be visually represented using directed acyclic graphs and analyzed using GEE, MLM and DSEM. We will start out with a scenario of the basic MLM—where a time-varying outcome Y is regressed on a single time-varying predictor X and in the presence of stable between person differences in the intercept—and increase the complexity until we reach the scenario that includes a time-varying endogenous covariate. The primary interest of this simulation study is the comparative performance of different specifications of the MLM and GEE in terms of bias in the estimation of the effect of X to Y. The secondary interest is the efficiency in mean squared error (MSE). We consider settings with timepoints $T = 10, 30$ and sample size $N = 30, 100, 200$.

Statistical analyses pertaining to the GEE and basic MLM will be performed in `R`, version 4.4.2 (Team, 2024). To fit the GEE, the R-package `geepack` (Halekoh et al., 2006) will evaluate several different working correlation structures, including independent, exchangeable, AR(1) and unstructured. To fit the basic MLM, the R-package `lme4` (Bates et al., 2015) will be employed, where we will use restricted maximum likelihood estimation.