Methods

Ward B. Eiling

November 25, 2024

Table of contents

1 Main Simulation of Qian et al. (2020): With Treatment

1.1 Simulation Conditions

In the simulation, we considered nine generative models (GMs), all of which have an endogenous covariate. In the first 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 described by Qian et al. (2020), we consider 6 models on top of the three models described in the original paper. More specifically, we considered models with an "a" suffix, where all random slopes were removed, and models with a "b" suffix, were on top of the removal of random slopes, the interaction terms were 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 Raudenbusch & Bryk, 2002).

1.1.1 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 $\epsilon_{it+1} \sim \mathcal{N}(0, \sigma^2_\epsilon)$.

1.1.2 Generative Model 1A

GM1A is the same as GM1, except that the random slope b_{i2} for the treatment A_{it} is removed. The single equation model thus becomes:

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

1.1.3 Generative Model 1B

GM1B is the same as GM1A, except that the interaction term $\beta_1 A_{it} X_{it}$ is removed. The single equation model thus becomes:

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

1.1.4 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:

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

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

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

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

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 $b_{ij} \sim \mathcal{N}(0, \sigma^2_{bj})$, 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 $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The exogenous noise is $\epsilon_{it+1} \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

1.1.5 Generative Model 2A

GM2A is the same as GM2, except that the random slopes b_{i1}, b_{i2} and b_{i3} are removed. The single equation model then becomes the same as GM1A, but with the time-varying randomization probabilities of GM2.

1.1.6 Generative Model 2B

GM2B is the same as GM2A, except that the interaction term $\beta_1 A_{it} X_{it}$ is removed. The single equation model then becomes the same as GM1B, but with the time-varying randomization probabilities of GM2.

1.1.7 Generative Model 3

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

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

1.1.8 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 the same as GM1A, but including the dependency of the covariate X_{it} on the random intercept b_{i0} present in GM3.

1.1.9 Generative Model 3B

GM3B is the same as GM3A, except that the interaction term $\beta_1 A_{it} X_{it}$ is removed. The single equation model then becomes the same as GM1B, but including the dependency of the covariate X_{it} on the random intercept b_{i0} present in GM3.

1.1.10 Parameter Values

The following parameter values were chosen:

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

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

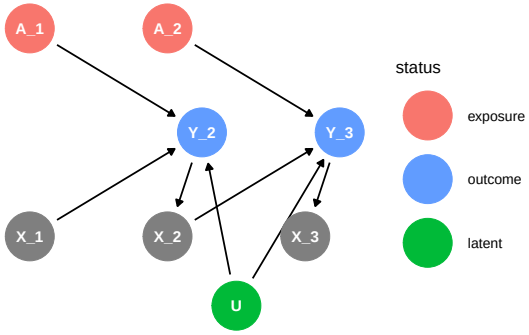
1.2 Graphical representations of Data Generating Models

1.2.1 Directed Acyclic Graphs (DAGs)

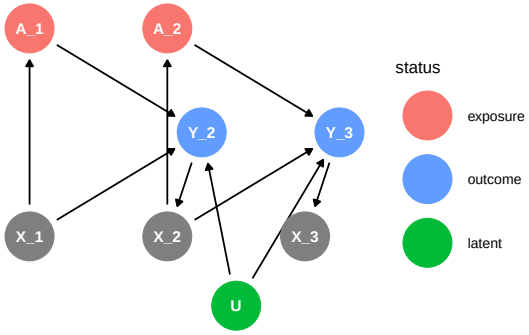
The DAGs for the first three observations of the three data generating models are presented in Figure 1. The red arrows show the biased paths after controlling for the covariate X_{it}.

Figure 1: DAG for Generative Model 1

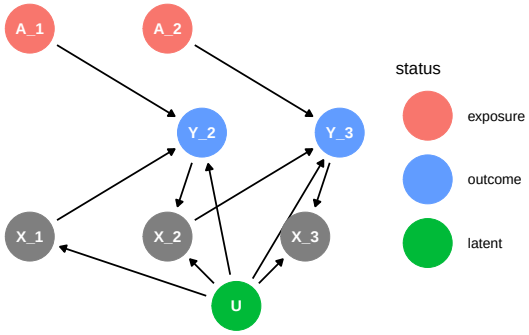
(a) Generative Model 1, 1A and 1B



(b) Generative Model 2, 2A and 2B



(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 1, 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}.

1.2.2 Path Diagrams

Alternatively, we can display the data generating models as a path diagram, where latent variables are represented by circles, observed variables by squares and relationships across variables by arrows. The path diagrams of the three data generating models is presented in **?@fig-pathdiagrams** (GM1 in **?@fig-GM1_pd**, GM2 in **?@fig-GM2_pd**, and GM3 in **?@fig-GM3_pd**), which shows the discrepancies between the different generative models more clearly than the DAGs.

We can make a couple observations from this path diagram:

- Contrary to the DAG, this path diagram shows the moderation effect (1) of X_{it} on the relationship between X_{ti} and Y_{it+1} and (2) of u_{2i} on the relationship between X_{it} and Y_{it+1}.
- Similar to the example without treatment in section 2.2, the covariate X_{it} is determined by the previous value of the outcome Y_{ti}—which makes it an endogenous time-varying covariate.
- The path diagram does not display the difference in the randomized treatment assignment probabilities between GM1 and GM2.

1.3 Data Estimation/Analysis

For the multilevel linear model, the analytical models are equivalent to each of the respective data-generating models. The multilevel linear models were fitted using the `lmer` function in R. As a reminder, the analytical *multilevel model* for GM1 and GM3 is given by:

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

which is fitted as `lmer(Y ~ X * A + (1 + A | id), data = data)` in R.

The analytical *multilevel model* for GM2 is given by:

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

which is fitted as `lmer(Y ~ X * A + (X * A | id), data = data)` in R.

The analytical *multilevel model* for GM1A, GM2A, and GM3A is given by:

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

which is fitted as `lmer(Y ~ X * A + (1 | id), data = data)` in R.

The analytical *multilevel model* for GM1B, GM2B, and GM3B is given by:

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

which is fitted as `lmer(Y ~ X + A + (1 | id), data = data)` in R.

The specification of the GEE models related to each of the generative models is unsurprisingly different considering that GEE does not explicitly model random effects. For each of the generative models, we will fit three GEE models: one with an exchangeable correlation structure, one with an independent correlation structure, and one with an AR(1) correlation structure. The GEE models were fitted using the `geeglm` function in R. Since the fixed effects modeled in GM1, GM1a, GM2, GM2a, GM3, GM3a are the same (the only differences pertain to the modeling of random effects), the analytical *GEE model* is identical across these three conditions. The analytical *GEE model* is given by:

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

The GEE models were fitted as `geeglm(Y ~ X * A, id = id, data = data, family = gaussian, corstr = "exchangeable")`, `geeglm(Y ~ X * A, id = id, data = data, family = gaussian, corstr = "independence")`, and `geeglm(Y ~ X * A, id = id, data = data, family = gaussian, corstr = "ar1")` in R.

In GM1b, GM2b, GM3b, the fixed interaction effect is removed, so the analytical *GEE model* is given by:

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

The GEE models were fitted as `geeglm(Y ~ X + A, id = id, data = data, family = gaussian, corstr = "exchangeable")`, `geeglm(Y ~ X + A, id = id, data = data, family = gaussian, corstr = "independence")`, and `geeglm(Y ~ X + A, id = id, data = data, family = gaussian, corstr = "ar1")` in R.

2 Appendix

2.1 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 $b_{i0} \sim N(0, \sigma^2_{b0})$ and $b_{i2} \sim N(0, \sigma^2_{b2})$ 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 $\epsilon_{it+1} \sim N(0, \sigma^2_\epsilon)$.

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 $b_{ij} \sim N(0, \sigma^2_{b_j})$, 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 $\epsilon_{it+1} \sim N(0, \sigma^2_\epsilon)$.

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:

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

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