DGMs of Qian et al. (2020) - Part 2: With Treatment

Ward B. Eiling

November 14, 2024

Table of contents

1 Main Simulation of Qian et al. (2020): With Treatment and with Translated Notation

1.1 Simulation Conditions

In the simulation, we considered three generative models (GMs), all of which have an endogenous covariate. In the first two GMs, the endogenous covariate Z_{ti} equals the previous outcome Y_{ti} plus some random noise, so the conditional independence assumption is valid. In GM 3, the endogenous covariate depends directly on u_{0i}, violating assumption. The details of the generative models are described below. We follow the notation of Schoot et al. (2017), which is largely based on that of Raudenbusch and 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_{ti}. The outcome is generated according to the following repeated-observations or within-person model (level 1)

$$Y_{(t+1)i} = \pi_{0i} + \pi_{1i}Z_{ti} + \pi_{2i}X_{ti} + \pi_{3i}X_{ti}Z_{ti} + e_{(t+1)i}$$

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

$$\pi_{0i} = \beta_{00} + u_{0i} \quad \text{with} \quad u_{0i} \sim \mathcal{N}(0, \sigma^2_{u_0})$$

$$\pi_{1i} = \beta_{10}$$

$$\pi_{2i} = \beta_{20} + u_{2i} \quad \text{with} \quad u_{2i} \sim \mathcal{N}(0, \sigma^2_{u_2})$$

$$\pi_{3i} = \beta_{30}$$

By substitution, we get the single equation model:

$$\begin{aligned} Y_{(t+1)i} &= \pi_{0i} + \pi_{1i}Z_{ti} + \pi_{2i}X_{ti} + \pi_{3i}X_{ti}Z_{ti} + e_{(t+1)i} \\ &= (\beta_{00} + u_{0i}) + \beta_{10}Z_{ti} + (\beta_{20} + u_{2i})X_{ti} + \beta_{30}X_{ti}Z_{ti} + e_{(t+1)i} \\ &= \beta_{00} + \beta_{10}Z_{ti} + u_{0i} + X_{ti}(\beta_{20} + \beta_{30}Z_{ti} + u_{2i}) + e_{(t+1)i} \end{aligned}$$

The random effects $u_{0i} \sim \mathcal{N}(0, \sigma^2_{u0})$ and $u_{2i} \sim \mathcal{N}(0, \sigma^2_{u2})$ are independent of each other. The covariate is generated as $Z_{i1} \sim \mathcal{N}(0, 1)$, and for $t \geq 2$,

$$Z_{ti} = Y_{ti} + \mathcal{N}(0, 1).$$

The randomization probability $p_t = P(X_{ti} = 1 \mid H_{it})$ is constant at 1/2. Thus, $X_{ti} \sim$ Bernoulli(0.5) for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. In other words, each individual

has an independent 50% chance of receiving the treatment at each time point. The exogenous noise is $e_{(t+1)i} \sim \mathcal{N}(0, \sigma_e^2)$.

1.1.2 Generative Model 2

In GM2, we considered the case with a random intercept and a random slope for (1) covariate Z_{ti}, (2) treatment X_{ti}, and (3) the interaction between X_{ti} and Z_{ti}; 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:

$$\pi_{0i} = \beta_{00} + u_{0i} \quad \text{with} \quad u_{0i} \sim \mathcal{N}(0, \sigma_{u_0}^2)$$

$$\pi_{1i} = \beta_{10} + u_{1i} \quad \text{with} \quad u_{1i} \sim \mathcal{N}(0, \sigma_{u_1}^2)$$

$$\pi_{2i} = \beta_{20} + u_{2i} \quad \text{with} \quad u_{2i} \sim \mathcal{N}(0, \sigma_{u_2}^2)$$

$$\pi_{3i} = \beta_{30} + u_{3i} \quad \text{with} \quad u_{3i} \sim \mathcal{N}(0, \sigma_{u_3}^2)$$

By substitution, we get the single equation model:

$$\begin{aligned} Y_{(t+1)i} &= \pi_{0i} + \pi_{1i} Z_{ti} + \pi_{2i} X_{ti} + \pi_{3i} X_{ti} Z_{ti} + e_{(t+1)i} \\ &= (\beta_{00} + u_{0i}) + (\beta_{10} + u_{1i}) Z_{ti} + (\beta_{20} + u_{2i}) X_{ti} + (\beta_{30} + u_{3i}) X_{ti} Z_{ti} + e_{(t+1)i} \\ &= \beta_{00} + \beta_{10} Z_{ti} + u_{0i} + u_{1i} Z_{ti} + X_{ti} \left(\beta_{20} + \beta_{30} Z_{ti} + u_{2i} + u_{3i} Z_{ti}\right) + e_{(t+1)i} \end{aligned}$$

The random effects $u_{ji} \sim \mathcal{N}(0, \sigma_{u_j}^2)$, for $0 \leq j \leq 3$, are independent of each other. The covariate is generated as $Z_{i1} \sim \mathcal{N}(0, 1)$, and for $t \geq 2$,

$$Z_{ti} = Y_{ti} + \mathcal{N}(0, 1).$$

The randomization probability depends on Z_{ti}:

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

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 for the covariate for any given person and timepoint is above the cutoff, the probability of receiving the treatment p_t is 0.7; otherwise, it is 0.3. Accordingly, $X_{ti} \sim \text{Bernoulli}(p_t)$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The exogenous noise is $e_{(t+1)i} \sim \mathcal{N}(0, \sigma_e^2)$.

1.1.3 Generative Model 3

GM3 is the same as GM1, except that the covariate Z_{ti} depends directly on u_{0i}:

$$Z_{i1} \sim \mathcal{N}(u_{0i}, 1), \quad Z_{ti} = Y_{ti} + \mathcal{N}(u_{0i}, 1) \text{ for } t \geq 2.$$

1.1.4 Parameter Values

The following parameter values were chosen:

$$\beta_{00} = -2, \quad \beta_{10} = -0.3, \quad \beta_{20} = 1, \quad \beta_{30} = 0.3,$$

$$\sigma^2_{u0} = 4, \quad \sigma^2_{u1} = \frac{1}{4}, \quad \sigma^2_{u2} = 1, \quad \sigma^2_{u3} = \frac{1}{4}, \quad \sigma^2_e = 1.$$

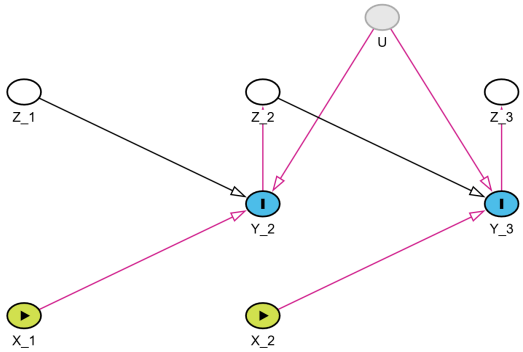
1.2 Graphical representations of Data Generating Models

1.2.1 Directed Acyclic Graphs (DAGs)

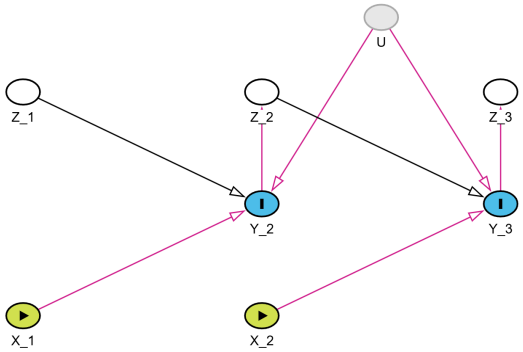
The DAGs for the first three observations of the three data generating models are presented in Figure 1 (GM1 in Figure 1a, GM2 in Figure 1b, and GM3 in Figure 1c). The red arrows show the biased paths after controlling for the covariate Z_{it}.

Figure 1: Directed Acylcic Graphs (DAGs) for the three Generative Models

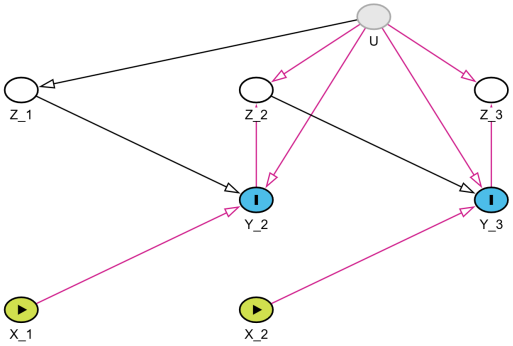
(a) Generative Model 1



(b) Generative Model 2



(c) Generative Model 3



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 Z_{ti} on the random intercept u_{0i}.

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

$$Z_{ti} \perp (u_{0i}, u_{1i}) \mid H_{(t-1)i}, X_{(t-1)i}, Y_{ti}.$$

This allows Z_{ti} to be endogenous, but the endogenous covariate Z_{ti} can only depend on the random effects through variables observed prior to Z_{ti}: $H_{(t-1)i}$, $X_{(t-1)i}$, 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 Z_{ti} depends directly on u_{0i}; and is thus not independent of the random effects u_{0i} and u_{1i}. Notice that GM1 and GM2 are also not marginally independent of u_{0i} and u_{1i}, but they are conditionally independent given $H_{(t-1)i}$, $X_{(t-1)i}$, and Y_{ti}.

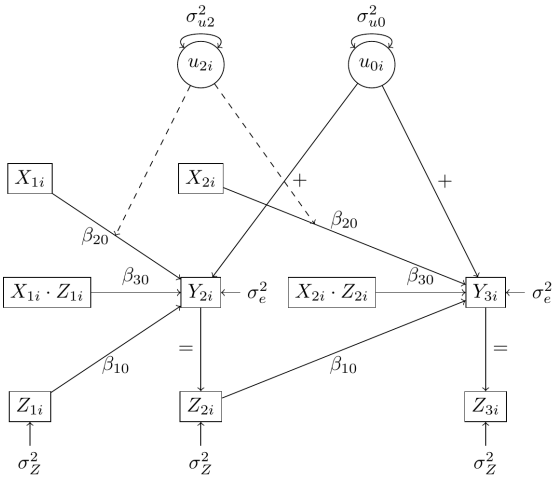
1.2.2 Path Diagrams

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

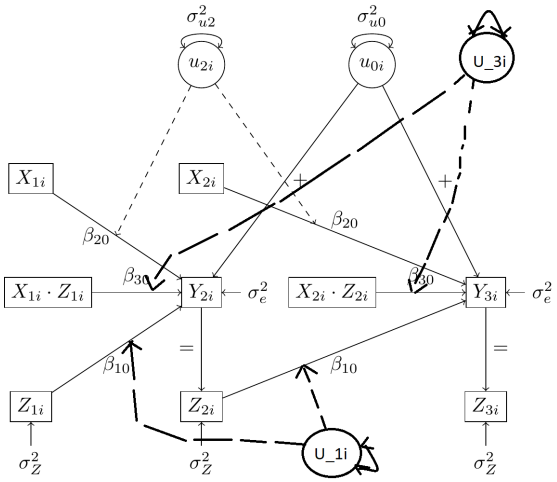
in Figure 2 (GM1 in Figure 2a, GM2 in Figure 2b, and GM3 in Figure 2c), which shows the discrepancies between the different generative models more clearly than the DAGs.

Figure 2: Path Diagrams for the three Generative Models

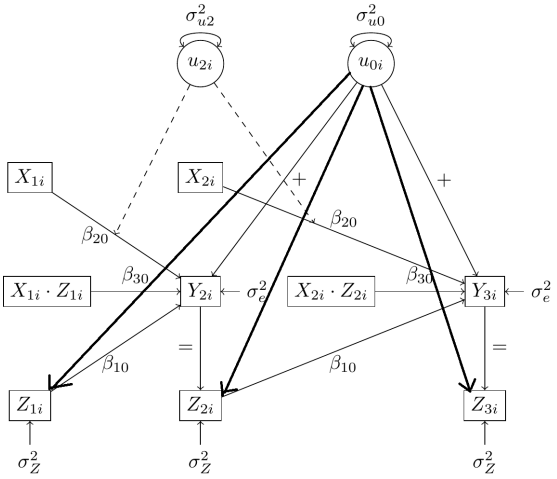
(a) Generative Model 1



(b) Generative Model 2



(c) Generative Model 3



We can make a couple observations from this path diagram:

- Contrary to the DAG, this path diagram shows the moderation effect (1) of Z_{ti} on the relationship between X_{ti} and Y_{ti+1} and (2) of u_{2i} on the relationship between Z_{ti} and $Y_{(t+1)i}$.
- Similar to the example without treatment in section 2.2, the covariate Z_{ti} 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. As a reminder, the analytical *multilevel model* for GM1 and GM3 is given by

$$Y_{(t+1)i} = (\beta_{00} + u_{0i}) + \beta_{10} Z_{ti} + (\beta_{20} + u_{2i}) X_{ti} + \beta_{30} X_{ti} Z_{ti} + e_{(t+1)i}$$

and the analytical *multilevel model* for GM2 is given by

$$Y_{(t+1)i} = (\beta_{00} + u_{0i}) + (\beta_{10} + u_{1i}) Z_{ti} + (\beta_{20} + u_{2i}) X_{ti} + (\beta_{30} + u_{3i}) X_{ti} Z_{ti} + e_{(t+1)i}$$

However, for GEE, the analytical model is different, as they do not explictly model random effects. As the main effects modeled in GM1 through GM3 are the same (the only differences pertains to modelling of random effects), the analytical *GEE models* are identical for these different conditions. The analytical *GEE model* is given by

$$Y_{(t+1)i} = \beta_0 + \beta_1 Z_{ti} + \beta_2 X_{ti} + \beta_3 X_{ti} Z_{ti} + e_{(t+1)i}$$

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:

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

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

2.2 Translation of Notation

In the table below, we will provide the translation of original notation in Qian et al. (2020) to the notation by Schoot et al. (2017). This notation in turn, is very similar to and likely based on the notation by Raudenbush and Bryk (2002). However, rather than representing the random effects with an r, the notation by Schoot et al. (2017) uses u.

Parameter/variable	Qian et al. (2020)	Schoot et al. (2017)
Covariate	X_{it}	Z_{ti}
Randomized Treatment	A_{it}	X_{ti}
Outcome	Y_{it+1}	$Y_{(t+1)i}$
Fixed intercept	α_0	β_{00}
Random intercept	b_{i0}	u_{0i}
Fixed slope for covariate	α_1	β_{10}
Random slope for covariate	b_{i1}	u_{1i}
Fixed slope of treatment	β_0	β_{20}
Random slope for treatment	b_{i2}	u_{2i}
fixed interaction effect of covariate and treatment	β_1	β_{30}
Random interaction effect of covariate and treatment	b_{i3}	u_{3i}
Error term (exogenous noise)	ϵ_{it+1}	$e_{(t+1)i}$
Randomization probability	p_t	p_t
Residual variance	σ_{ϵ}^2	σ_e^2
Random intercept variance	σ_{b0}^2	σ_{u0}^2
random slope variance for covariate	σ_{b1}^2	σ_{u1}^2
Random slope variance for treatment	σ_{b2}^2	σ_{u2}^2
Random slope variance for interaction	σ_{b3}^2	σ_{u3}^2