Data Generating Models of Qian et al. (2020)

Ward B. Eiling

2024-10-17

Table of contents

0.1 Introduction

In this document, I will recreate the data generating models presented by Qian et al. (2020), accompanied by visual representations of these models.

```
# Load packages
library(dagitty)
library(tidyverse)
library(ggdag)
library(lme4)
library(jtools)
library(gee)
library(geex)
library(nlme)
```

0.2 Simple Concrete Example: Without Treatment

0.2.1 Example in section 2.2 of Qian et al. (2020)

As a concrete example, consider the case where each individual is observed for 2 time points ($T_i = 2$), and the covariate at the second time point is the lag-1 outcome: $X_{i2} = Y_{i2}$.

> By lagging the outcome, we essentially have three time points: X_{i1}, $X_{i2} = Y_{i2}$, and Y_{i3}.

Suppose the variables are generated from the following multilevel linear model (MLM) with a random intercept:

$$b_i \sim N(0, \sigma_u^2),$$

$$X_{i1} \sim N(0, \sigma_{X_1}^2) \text{ independently of } b_i,$$

$$Y_{i2} \mid X_{i1}, b_i \sim N(\beta_0 + \beta_1 X_{i1} + b_i, \sigma_\epsilon^2),$$

$$X_{i2} = Y_{i2},$$

$$Y_{i3} \mid X_{i1}, Y_{i2}, X_{i2}, b_i \sim N(\beta_0 + \beta_1 X_{i2} + b_i, \sigma_\epsilon^2).$$

0.2.2 Translating the notation

The notation here is different from the notation used in the psychological sciences. In the psychological sciences, we would typically denote the random intercept as u_{0i} instead of b_i. We would also denote the fixed intercept as γ_{00} instead of β_0 and σ^2_ϵ as σ^2_e, but keep the fixed slope as β_1.

Let's now rewrite the model in this notation

$$u_{0i} \sim N(0, \sigma^2_u),$$

$$X_{i1} \sim N(0, \sigma^2_{X_1}) \text{ independently of } u_{0i},$$

$$Y_{i2} \mid X_{i1}, u_{0i} \sim N(\gamma_{00} + \beta_1 X_{i1} + u_{0i}, \sigma^2_e),$$

$$X_{i2} = Y_{i2},$$

$$Y_{i3} \mid X_{i1}, Y_{i2}, X_{i2}, u_{0i} \sim N(\gamma_{00} + \beta_1 X_{i2} + u_{0i}, \sigma^2_e).$$

0.2.3 Visualizing the Model

We may now draw the DAG for this model

Note that there is an open biasing path from Y_2 to X_2 in the DAG: the exposure/predictor X_2 is *caused by* (in this case equivalent to) the previous outcome Y_2—and thus this time-varying covariate is *endogenous*.

0.2.4 Generating the data

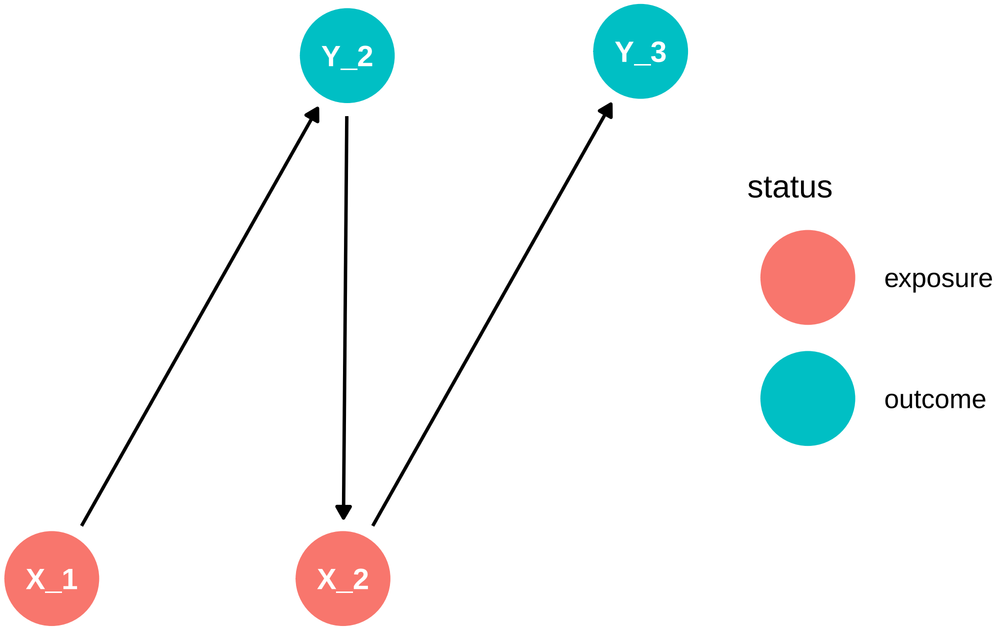
Let's now generate the data according to this model.

```r
set.seed(123)

n_i <- 5000 # number of individuals

sigma_u <- 1 # variance of random intercept
sigma_X1 <- 1 # variance of X1
sigma_e <- 0.1 # residual variance
```

Figure 1: DAG with fixed effects

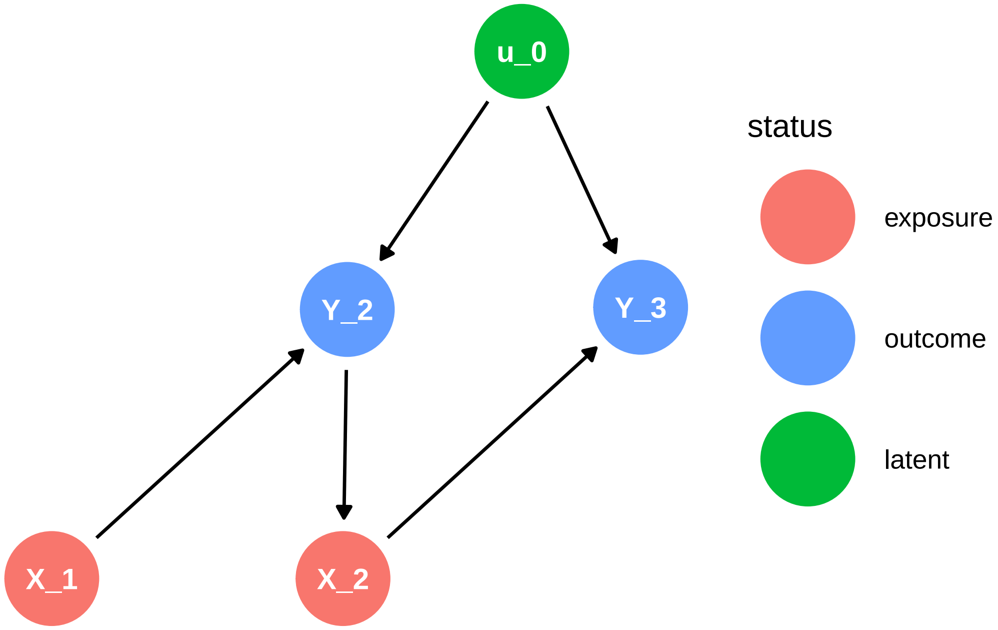


```
beta_1 <- 0.8 # overall slope
gamma_00 <- 2 # overall intercept

# simulate data
u_0i <- rnorm(n_i, 0, sigma_u)
X_i1 <- rnorm(n_i, 0, sigma_X1)
Y_i2 <- rnorm(n_i, gamma_00 + beta_1 * X_i1 + u_0i, sigma_e)
X_i2 <- Y_i2
Y_i3 <- rnorm(n_i, gamma_00 + beta_1 * X_i2 + u_0i, sigma_e)

# # create data frame in wide format
# section2.2.data_wide <- data.frame(id = 1:n_i,
#                                    X_i1 = X_i1,
#                                    Y_i2 = Y_i2,
#                                    X_i2 = X_i2,
#                                    Y_i3 = Y_i3)
#
# head(section2.2.data_wide)
#
# # create data frame in long format
```

Figure 2: DAG with random intercept



```r
# section2.2data_long <- data.frame(id = rep(1:n_i, each = 3),
#                              time = rep(1:3, n_i),
#                              X = rep(NA, 3 * n_i),
#                              Y = rep(NA, 3 * n_i))
#
# for (i in 1:n_i){
#   section2.2data_long$X[section2.2data_long$id == i & section2.2data_long$time == 1] <-
#   section2.2data_long$Y[section2.2data_long$id == i & section2.2data_long$time == 2] <-
#   section2.2data_long$X[section2.2data_long$id == i & section2.2data_long$time == 2] <-
#   section2.2data_long$Y[section2.2data_long$id == i & section2.2data_long$time == 3] <-
# }
#
#
# head(section2.2data_long)


# create data frame in long format with lagged predictor
section2.2data_long_lagged <- data.frame(id = rep(1:n_i, each = 2),
                             "time" = rep(1:2, n_i),
                             "X_lag1" = rep(NA, 2 * n_i),
                             "Y" = rep(NA, 2 * n_i))

for (i in 1:n_i){
  section2.2data_long_lagged$X_lag1[section2.2data_long_lagged$id == i & section2.2data_lo
  section2.2data_long_lagged$X_lag1[section2.2data_long_lagged$id == i & section2.2data_lo
  section2.2data_long_lagged$Y[section2.2data_long_lagged$id == i & section2.2data_long_la
  section2.2data_long_lagged$Y[section2.2data_long_lagged$id == i & section2.2data_long_la
}

head(section2.2data_long_lagged)
```

```
  id time     X_lag1        Y
1  1    1 -0.4941739 1.281258
2  1    2  1.2812578 2.329146
3  2    1  1.1275935 2.655216
4  2    2  2.6552161 3.836058
5  3    1 -1.1469495 2.733845
6  3    2  2.7338448 5.659680
```

0.2.5 Estimating the model

Let's now estimate the model using a multilevel linear model (MLM) and a generalized estimating equation (GEE) model.

```
# Multilevel Linear Model
section2.2_mlm_reml <- lmer(Y ~ 1 + X_lag1 + (1 | id), data = section2.2data_long_lagged)
section2.2_mlm_mle <- lmer(Y ~ 1 + X_lag1 + (1 | id), data = section2.2data_long_lagged, R
# Generalized Estimating Equations
section2.2_gee_ind <- gee(Y ~ 1 + X_lag1, id = id, data = section2.2data_long_lagged, fami
```

```
(Intercept)      X_lag1
   1.789170    1.010113
```

```
section2.2_gee_exch <- gee(Y ~ 1 + X_lag1, id = id, data = section2.2data_long_lagged, fam
```

```
(Intercept)      X_lag1
   1.789170    1.010113
```

```
section2.2_gee_ar1 <- gee(Y ~ 1 + X_lag1, id = id, data = section2.2data_long_lagged, fami
```

```
(Intercept)      X_lag1
   1.789170    1.010113
```

```
section2.2_gee_unstr <- gee(Y ~ 1 + X_lag1, id = id, data = section2.2data_long_lagged, fa
```

```
(Intercept)      X_lag1
   1.789170    1.010113
```

```
# Generalized linear model (OLS): assumes independence of observations
section2.2_glm <- glm(Y ~ 1 + X_lag1, data = section2.2data_long_lagged, family = gaussian
# Marginal linear model (GLS)
section2.2_gls_symm_mle <- nlme::gls(Y ~ 1 + X_lag1, data = section2.2data_long_lagged, co
section2.2_gls_compsymm_mle <- nlme::gls(Y ~ 1 + X_lag1, data = section2.2data_long_lagged

summary(section2.2_gls_symm_mle)
```

```
Generalized least squares fit by maximum likelihood
  Model: Y ~ 1 + X_lag1
  Data: section2.2data_long_lagged
       AIC      BIC    logLik
  8908.115 8936.956 -4450.057

Correlation Structure: General
 Formula: ~1 | id
 Parameter estimate(s):
 Correlation:
  1
2 0.99

Coefficients:
                Value   Std.Error  t-value p-value
(Intercept) 2.0001338 0.014138919 141.4630       0
X_lag1      0.7983801 0.000897811 889.2519       0

 Correlation:
       (Intr)
X_lag1 -0.063
```

```
Standardized residuals:
         Min           Q1          Med           Q3          Max 
-3.275390043 -0.659680887 -0.001194166  0.672938257  3.513982237 

Residual standard error: 1.000225 
Degrees of freedom: 10000 total; 9998 residual
```

```
section2.2_coefs <- data.frame(
  row.names = c("Intercept", "X_lag1"),
  MLM_reml = fixef(section2.2_mlm_reml),
  MLM_mle = fixef(section2.2_mlm_mle),
  GEE_ind = coef(section2.2_gee_ind),
  GEE_exch = coef(section2.2_gee_exch),
  GEE_ar1 = coef(section2.2_gee_ar1),
  GEE_unstr = coef(section2.2_gee_unstr),
  GLM = coef(section2.2_glm),
  GLS_Symm = coef(section2.2_gls_symm_mle),
  GLS_CompSymm = coef(section2.2_gls_compsymm_mle)
)

knitr::kable(caption = "Section 2.2: Estimated coefficients from the MLM and GEE models",
```

Table 1: Section 2.2: Estimated coefficients from the MLM and GEE models

	MLM_reml	MLM_mle	GEE_ind	GEE_exch	GEE_ar1	GEE_unstr	GLM	GLS_Symm	GLS_CompSymm
Intercept	2.000	2.000	1.789	2.000	2.000	2.000	1.789	2.000	2.000
X_lag1	0.798	0.798	1.010	0.798	0.798	0.798	1.010	0.798	0.798

We can clearly see that the MLM and GEE models provide exactly the same estimates for the fixed intercept and fixed regression coefficient, with the exception of the GEE with independence working correlation structure.

> According to Pepe and Anderson (1994), this is the only structure that can avoid bias in the estimation of the fixed effects (i.e., that has a valid marginal interpretation).

As a reminder, the fixed effects were specified as $\gamma_{00} = 2$ and $\beta_1 = 0.8$. Thus, we can see that all models except the GEE with independence working correlation structure returns estimates that are very close to the true values—which represented the conditional mean of Y given X and u_{0i} rather than the marginal mean of Y given X.

To see why this makes sense, it is important to realize that the parameter estimates represent the parsimonious conditional relationship

$$E[Y_{it+1} \mid X_{it}, u_{0i}] = \gamma_{00} + \beta_1 X_{it} + u_{0i}$$

And not the marginal relationship, which is given by:

$$E[Y_{i2} \mid X_{i1}] = \gamma_{00} + \beta_1 X_{i1}$$

$$E[Y_{i3} \mid X_{i2}] = (1 - \rho\zeta - \rho)\gamma_{00} + [(1 - \rho\zeta)\beta_1 + \rho]X_{i2}$$

Let's confirm this by calculating the true marginal effect

```
sigma2_u0 = sigma_u^2
sigma2_e = sigma_e^2

rho = sigma2_u0 / (sigma2_u0 + sigma2_e)

sigma2_X1 = sigma_X1^2
zeta = beta_1 * sigma2_X1 / (beta_1 * sigma2_X1 + sigma2_u0 + sigma2_e)

# Now let's compute the marginal effect of X1 on Y2
marginal_effect_X1_Y2 <- gamma_00 + beta_1
marginaleffect_X2_Y3 <- (1 - rho * zeta - rho) * gamma_00 + ((1 - rho * zeta) * beta_1 + r
```

> This is not correct, but how do we calculate the true marginal effects for intercept and slope?

0.2.6 Intermezzo: What are marginal effects/models?

Marginal models are a class of models that are used to estimate the population average effect of a covariate on an outcome. This may be useful, for instance, when prediction or indeed complete modelling of the data are not the main goal of an analysis (Pepe and Anderson, 1994).

> "Consider, for example, the future practice of screening for risk of respiratory disease, where one might simply ascertain Vitamin A deficiency, weight, height and other covariates at a single time point and make a determination of the child's risk based on these measurements." (Pepe and Anderson, 1994)

Here, the cross-sectional model is of primary interest for use in future screening practices and an in-depth model of longitudinal data is of secondary interest (Pepe and Anderson, 1994).

This contrasts with psychological research, where the cross-sectional model is often deemed problematic in the context of longitudinal data analysis, because it conflates (rather than separates) within-subject and between-subject effects. Instead, we tend to be much more interested in (1) the model that best describes the data (i.e., has the best model fit) and (2) the "why" question: complete model of the effects (including within- and between-person effects). What differs here is the aim of the study.

In what situations may psychological researchers be primarily interested in marginal effects over finding the best description of the data?

1. A *clinical psychologist* may be interested in screening for depression among a large sample of patients in a primary care setting.
2. A *school psychologist* may focus on understanding how classroom behaviors (e.g., attention problems, peer conflicts) relate to academic achievement across a large student population.
3. A *social psychologist* may study the impact of discrimination on mental health outcomes across different demographic groups.
4. A *developmental psychologist* may study how early childhood factors (e.g., parental education, socioeconomic status, early trauma) influence cognitive development in children.

Whether marginal or conditional models are preferred depends simply upon the research question and aim of the study:

> "We do not suggest that marginal models are preferable in general to conditional models. In Section 1 we provided one example where the marginal model is, in fact, preferable but in many cases it will not be. Indeed, which model should be used depends entirely on the questions to be addressed with the data. If a good description of the process generating the data is required then fully conditional or random effects models might be pursued." (Pepe and Anderson, 1994)

0.2.7 Intermezzo: What is the difference between REML and MLE?

When fitting a multilevel linear model (MLM) we can choose between restricted maximum likelihood (REML) and maximum likelihood estimation (MLE). In REML, σ^2 and ρ (intra class correlation) are essentially considered nuisance parameters, which makes sure that small sample bias is reduced. However, since we do not obtain the complete log likelihood, we cannot compare models using the likelihood ratio test. When sample sizes are sufficiently large, the two methods are asymptotically equivalent.

the variance components σ_u^2 are estimated by maximizing the likelihood of the residuals, conditional on the fixed effects. In MLE, the residual variance σ^2 and the variance components σ_u^2 are estimated by maximizing the likelihood of the residuals, conditional on the fixed effects and the random effects.

The difference between the two methods lies in the way they estimate the variance components of the model.

source: STAT 437: 007. Linear Marginal Models: Likelihood, Inference, and Asymptotics (Theory)

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

0.3.1 Original Section: "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:

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

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

0.4 Generative Model 1

0.4.1 Translation of Notation

In the table below, we will provide the translation of original notation in Qian et al. (2020) to notation more common in psychological research

Parameter	Original	New
Fixed intercept	α_0	γ_{00}
Fixed slope for X_{it}	α_1	γ_{01}
Random intercept	b_{i0}	u_{0i}
Random slope for A_{it}	b_{i2}	u_{1i}
Error term	ϵ_{it+1}	e_{it+1}
Fixed effect of A_{it}	β_0	γ_{10}
Interaction effect of A_{it} and X_{it}	β_1	γ_{11}
Covariate	X_{it}	Z_{it}
Treatment	A_{it}	X_{it}

Let's first state the original model:

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

Using this new notation, we may thus rewrite GM1 as a within model:

$$Y_{it+1} = \beta_{0i} + \beta_{1i} X_{it} + e_{it+1},$$

where:

$$\beta_{0i} = \gamma_{00} + \gamma_{01} Z_{it} + u_{0i} \quad \text{with} \quad u_{0i} \sim \mathcal{N}(0, \sigma_u^2),$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11} Z_{it} + u_{1i} \quad \text{with} \quad u_{1i} \sim \mathcal{N}(0, \sigma_u^2).$$

Combining these two equations, the model can be expressed as:

$$Y_{it+1} = \gamma_{00} + \gamma_{01} Z_{it} + u_{0i} + X_{it}(\gamma_{10} + \gamma_{11} Z_{it} + u_{1i}) + e_{it+1}.$$

0.4.2 Visualzing the Model

As mentioned by Ellen in the last meeting (17-10):

> "Conventional DAGs do not only represent main effects but rather the combination of main effects and interactions. Once you have drawn your DAG, you already assume that any variables pointing to the same outcome can modify the effect of the others pointing to the same outcome." (stackexchange)

For $t = 1$, the DAG and path diagram are as follows:

Note that the interaction between X_{it} and Z_{it} is not explicitly shown in the DAG, but is explicit in the path diagram. This is because the interaction is a model assumption, which is not explicitly represented in the non-parametric DAG.

For $t \geq 2$, the DAG and path diagram are as follows:

So the DAG for the first two observations looks like

0.5 Generative Model 2

0.5.1 Translation of Notation

Now we need to translate more parameters:

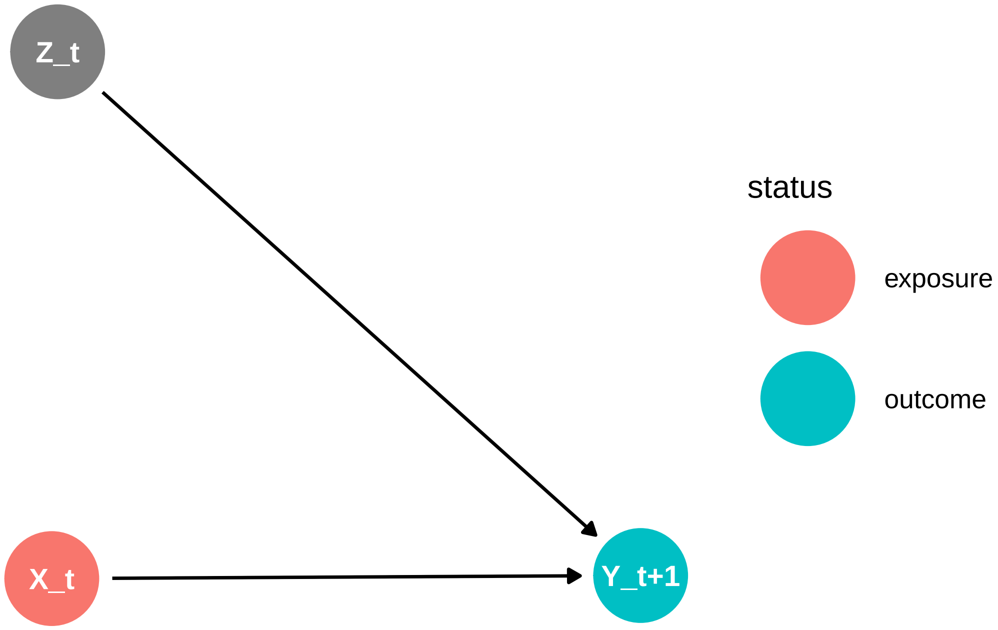


Figure 3: DAG

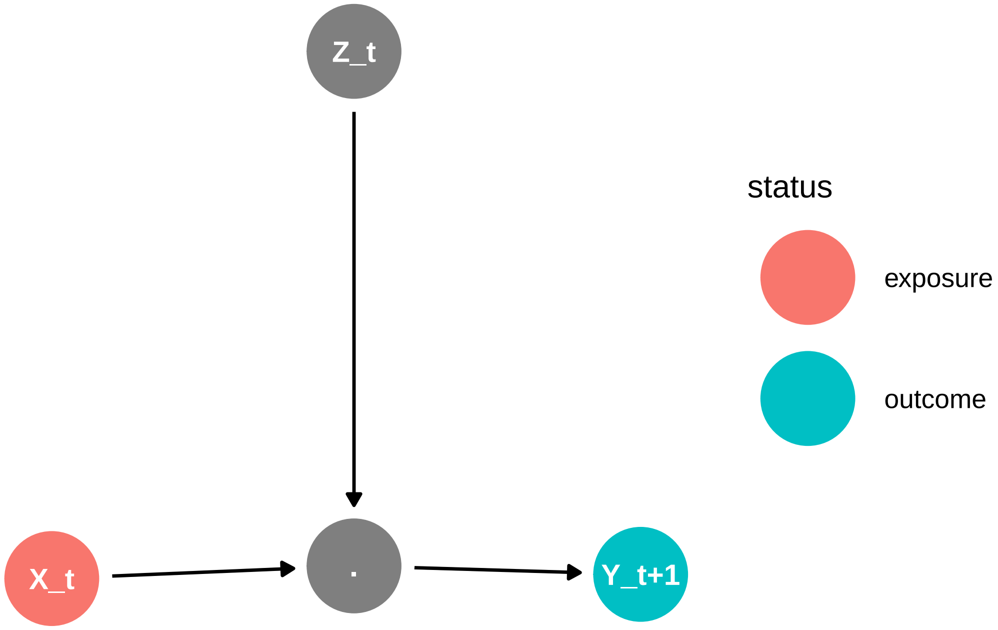


Figure 4: Path Diagam

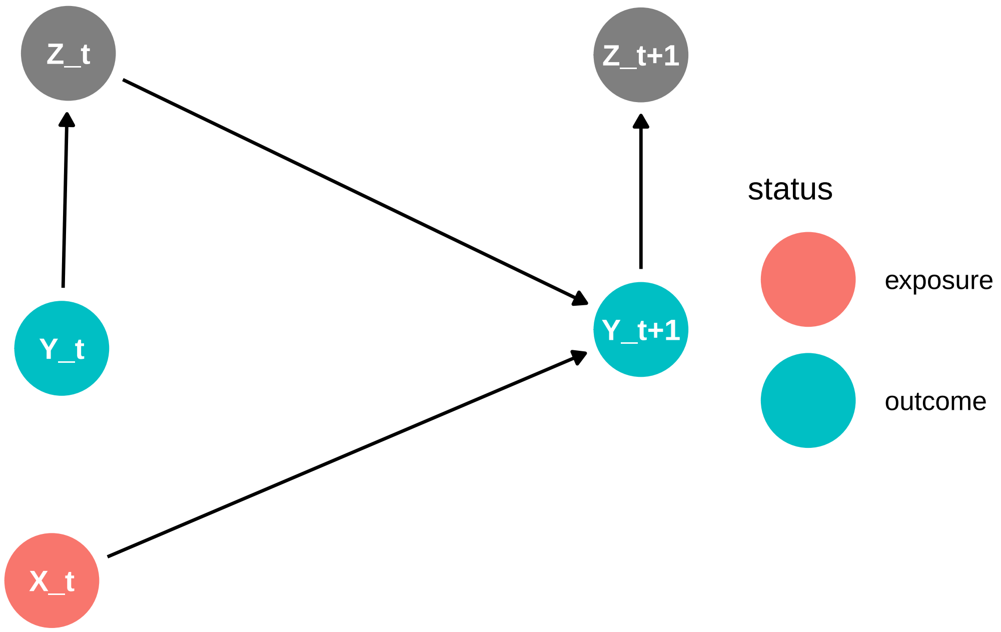


Figure 5: DAG

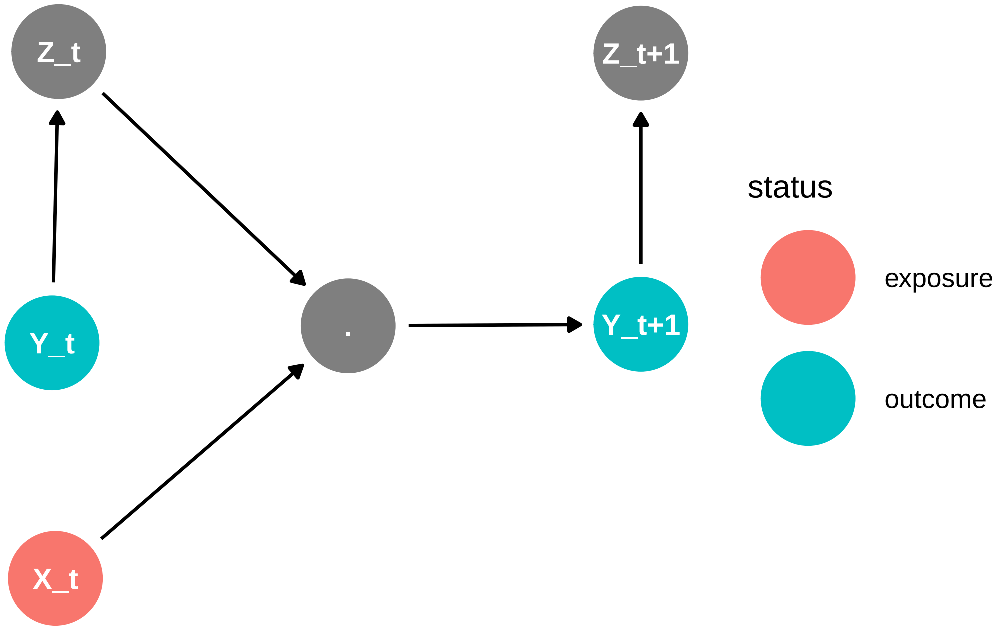
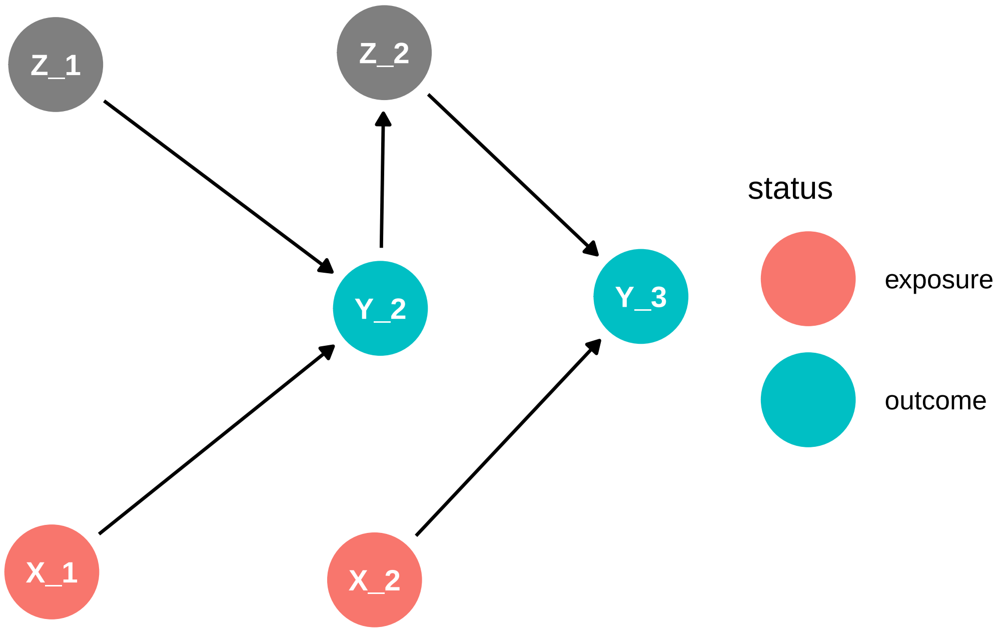


Figure 6: Path Diagam



Parameter	Original	New
Fixed intercept	α_0	γ_{00}
Fixed slope for X_{it}	α_1	γ_{10}
Random intercept	b_{i0}	u_{0i}
Random slope for X_{it}	b_{i1}	u_{1i}
Fixed effect of A_{it}	β_0	γ_{20}
Interaction effect of A_{it} and X_{it}	β_1	γ_{30}
Random slope for A_{it}	b_{i2}	u_{2i}
Random interaction effect for $A_{it} \times X_{it}$	b_{i3}	u_{3i}
Error term	ϵ_{it+1}	e_{it+1}
Covariate	X_{it}	Z_{it}
Treatment	A_{it}	X_{it}

Let's first restate the original model:

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

Using the psychological notation, we rewrite GM2 as a within-person model:

$$Y_{it+1} = \beta_{0i} + \beta_{1i} Z_{it} + \beta_{2i} X_{it} + \beta_{3i} X_{it} Z_{it} + e_{it+1},$$

with:

$$\beta_{0i} = \gamma_{00} + u_{0i} \quad \text{where} \quad u_{0i} \sim \mathcal{N}(0, \sigma_u^2),$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \quad \text{where} \quad u_{1i} \sim \mathcal{N}(0, \sigma_u^2),$$

$$\beta_{2i} = \gamma_{20} + u_{2i} \quad \text{where} \quad u_{2i} \sim \mathcal{N}(0, \sigma_u^2),$$

$$\beta_{3i} = \gamma_{30} + u_{3i} \quad \text{where} \quad u_{3i} \sim \mathcal{N}(0, \sigma_u^2).$$

Combining these, the full model becomes:

$$Y_{it+1} = (\gamma_{00} + u_{0i}) + (\gamma_{10} + u_{1i})Z_{it} + (\gamma_{20} + u_{2i})X_{it} + (\gamma_{30} + u_{3i})X_{it}Z_{it} + e_{it+1}.$$

0.5.2 Visualizing the Model

0.6 Generative Model 3

0.6.1 Translation of Notation

...

0.6.2 Visualizing the Model

...