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 12, 2024

Word count: 2500/2500

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.

A recent study by Qian et al. (2020) highlighted an issue present in both methods—except for GEE with working independence—where controlling for *time-varying endogenous covariates* may lead to biased causal estimates. 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 As a result of including these covariates in these models, ordinary interpretations of the coefficients are no longer valid (Qian et al., 2020, p. 3).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. Therefore, this report aims to understand and explain the issue of including endogenous covariates in analyses involving GEE and MLM in a psychological context. To achieve this aim, the current investigation employs (a) graphical tools such as the directed acyclic graph (DAG) and path diagram to assess potentially relevant assumptions, as well as (b) data simulations with additional scenarios to pinpoint the issue. Accordingly, the following research question will be addressed: *When does the inclusion of endogenous variables in multilevel linear models result in biased estimates of the treatment effect?*

2 Methods

To obtain a better understanding of the issue exposed by Qian et al. (2020), two methods were employed. First, graphical methods were used provide insight into the presence and extent of bias with potential violation of assumptions: (a) path diagrams were used to evaluate the conditional independence assumption and (b) directed acyclic graphs (DAGs) were used to evaluate the backdoor criterion (Pearl, 1988, 2009). Second, a simulation study was performed to reproduce the results for the generative models (GMs) from Qian et al. (2020) and to further isolate the issue using additional GMs. In this simulation, bias in the treatment effect was assessed with analytical multilevel models.

2.1 Data Generation

We consider 2 generative models (GMs) from Qian et al. (2020), where one was a special case of the general model for which bias was found. On top of that, we add 2 GMs, which are also a special case of this general model. We refer to the general model as GM G and the special cases of this model as GM A, B and C. Table 1 summarizes the differences between the generative models. Compared to the general model, GM A is not directly determined by the random intercept b_{i0}; GM B is does not have a random slope b_{i2} for treatment; and GM C does not have a fixed interaction effect β_1 between covariate and treatment.

Table 1: Generative Models: Summary of Differences

Generative Model	Name in Qian et al. (2020)	dependency b_{i0} and X_{it}	random slope treatment b_{i2}	 interaction β_1
G(eneral)	3	✓	✓	✓
A	1	×	✓	✓
B	NA	✓	×	✓
C	NA	✓	✓	×

The details of the generative models are described below. We follow the symbol 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).

2.1.1 Generative Model G

Following the original notation of Qian et al. (2020), the outcome of GM G was generated according to the following model:

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

where Y_{it+1} is the outcome at time $t+1$, X_{it} is the covariate at time t, A_{it} is the treatment at time t, b_{i0} is the random intercept, b_{i2} is the random slope for the treatment, and ϵ_{it+1} is the error term. We may rewrite this model into the repeated-observations or within-person model:

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

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

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

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

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

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

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

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

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

Figure 1a shows the path diagram for the first couple observations of GM G.

2.1.2 Generative Model A

GM A is the same as GM G, except that the covariate X_{it} is not directly determined by the random intercept b_{i0} (see Figure 1b):

$$X_{i1} \sim \mathcal{N}(0,1), \quad X_{it} = Y_{it} + \mathcal{N}(0,1) \text{ for } t \geq 2.$$

2.1.3 Generative Model B

GM B is the same as GM G, except that the random slope b_{i2} for the treatment A_{it} is removed (see Figure 1c). The single equation model then becomes:

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

2.1.4 Generative Model C

GM C is the same as GM G, except that the interaction term $\beta_1 A_{it} X_{it}$ is removed (see Figure 1d). The single equation model then becomes:

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

2.1.5 Parameter Values

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

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

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

2.2 Conditional Independence and Path Diagrams

Qian et al. (2020) proposes the use of the conditional independence assumption to identify whether bias may occur, which is given by:

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

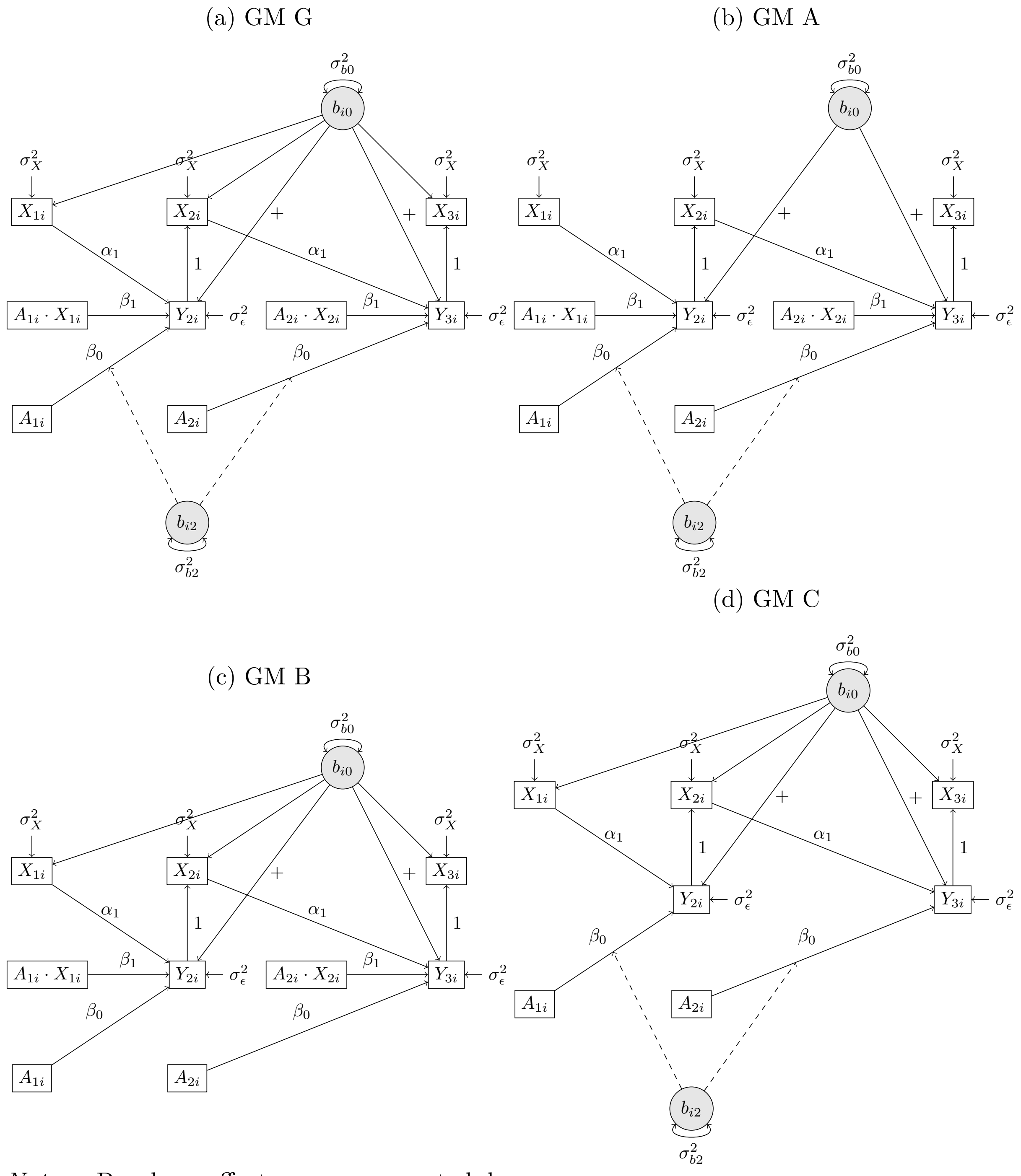
where H_{it-1} refers to the history of the set of covariates, which in this case are all observations of covariate X_{it} prior to the current timepoint t. This allows X_{it} to be endogenous, but the endogenous covariate X_{it} can only depend on the random effects through variables observed prior to X_{it}. If the only endogenous covariates are functions of prior treatments and prior outcomes, then the assumption automatically holds.

To make the application of the assumption more insightful, we accompany the equations of the GMs with path diagrams of the first three observations t for each generative model (see Figure 1). 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.

In GM1, the endogenous covariate X_{it} equals the previous outcome Y_{it} plus some random noiseo isolate the issue. In GM3, the endogenous covariate depends directly on b_{i0}, violating the assumption. We consider two 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. Note that the conditional independence assumption is violated in either of these variations,

When inspecting Figure 1b, we may notice that X_{it} becomes independent of the random effects after conditioning on Y_{it}. On the other hand, we can see that this assumption is violated only in GM G/B/C, as X_{it} depends directly on b_{i0} and can thus not be made independent of the random effects by conditioning on prior variables such as Y_{it} (see Figure 1a, Figure 1c and Figure 1d). Thus, all things considered, we would expect biased estimates of the treatment effect for GM G/B/C but not for GM A.

Figure 1: Path Diagrams for Generative Models G, A, B and C (t = 1, 2, 3)


(a) GM G

(b) GM A

(c) GM B

(d) GM C

Note. Random effects are represented by grey circles, observed variables by squares and relationships across variables by arrows, where dashed lines are reserved for random slopes.

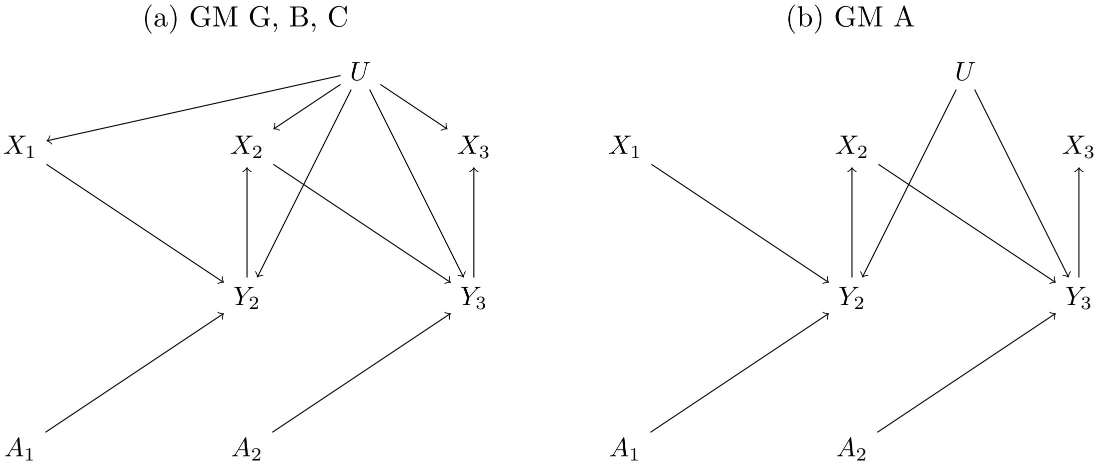
2.3 Backdoor Criterion and DAGs

According to the backdoor criterion (Pearl, 1988, 2009), a requirement for causal identification, causal effects can be identified by blocking non-causal paths through conditioning on intermediate variables (e.g., controlling or matching). If any non-causal paths cannot be blocked due to omitted variables or measurement error, treatment and outcome remain linked via backdoor paths, leading to biased estimates of the treatment effect (Kim & Steiner, 2021).

DAGs are a useful tool for representing causal relationships between variables and to evaluate the assumptions needed for causal identification (see Elwert & Winship, 2014 for a psychological example). We formulated the DAGs for the first three observations of each generative model, where the random disturbance b_{0i} was represented by the node U (e.g., Kim & Steiner, 2021, see Figure 2).

When applying Pearl's backdoor criterion to the GMs, it may be observed that there exists no backdoor path in the treatment effect $A_{it} \rightarrow Y_{it+1}$, as A_{it} does not have any parents. While we need not control for covariate X_{it} to obtain an unbiased total effect, doing so should not introduce bias. All things considered, according to the backdoor criterion, controlling for the covariate X_{it} should not result in biased estimates of the treatment effect for any of the generative models.

Figure 2: DAGs for Generative Models G, A, B and C (t = 1, 2, 3)


(a) GM G, B, C

(b) GM A

Note. The red arrows show the biased back-door path(s) in the treatment efffect (before controlling for X_{it}).

2.4 Data Analysis

We evaluated the performance of the models across a total of 24 different settings, each replicated 1,000 times, by systematically varying the following factors:

- **Generative Models (GM):** G, A, B, C
- **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). After the generation of data generation for any given setting, analytical multilevel linear models were fit that are equivalent to each of the respective data-generating models. To fit the standard MLM, the `lmer` function from the R-package `lme4` (Bates et al., 2015) was employed with restricted maximum likelihood estimation.

3 Results

Table 2 presents the simulation results for each of the generative and analytical models. The estimates for the analytical MLM may be interpreted in terms of bias, where given the value of the treatment effect $\beta_0 = 1$, absolute bias of .05 would imply 5% relative bias. Here we find the greatest absolute bias of $.02 - .06$ for GM3, $\leq .015$ for GM1/2, $\leq .010$ for GM3B and , $\leq .005$ for GM3A. While the bias found for the original GMs 1, 2 and 3 was slightly larger here compared to Qian et al. (2020), the overall pattern remained the same. To conclude, once we remove either the dependency of the random intercept with the covariate (GM1), the random slope b_{i2} (GM3A) or the interaction β_1 (GM3B) from GM3, the bias dissapears or becomes very small. The bias in GM3 decreases as the number of timepoints T increases from 10 to 30. Note that the MLM model fitting success rates are particularly poor for GM2, where in the worst case, only 87 of the 1000 models were fitted.

4 Discussion

This report employed both graphical methods and data simulations to understand and explain the issue of endogenous covariates. Now we first discuss the expected results based on the backdoor criterion Pearl (2009) and the conditional independence assumption (Qian et al., 2020), whereafter we discuss the findings relating to the two research questions.

Using the conditional independence assumption of Qian et al. (2020), we would expect, based on the path diagrams, that the treatment effect would be biased for GM3, 3A and 3B. On the other hand, the backdoor criterion suggested the absence of bias for all generative models. While Qian et al. (2020) show that GM3 is the only model with bias

Table 2: Treatment effect bias for Generative Models G, A, B and C over 1000 replications

			β_0		
GM	T	N	Bias	SD	SR
G	10	30	-0.052	0.245	0.999
		100	-0.064	0.134	1.000
		200	-0.051	0.096	1.000
	30	30	-0.024	0.206	0.997
		100	-0.030	0.108	0.996
		200	-0.023	0.080	0.997
A	10	30	0.000	0.238	0.998
		100	-0.012	0.129	1.000
		200	0.003	0.093	0.999
	30	30	-0.001	0.203	0.998
		100	-0.007	0.107	0.996
		200	0.001	0.079	0.996
B	10	30	0.000	0.126	1.000
		100	0.004	0.073	1.000
		200	0.002	0.048	1.000
	30	30	-0.001	0.071	1.000
		100	0.000	0.040	1.000
		200	0.000	0.028	1.000
C	10	30	0.001	0.217	0.999
		100	-0.008	0.121	1.000
		200	0.005	0.087	1.000
	30	30	0.000	0.193	1.000
		100	-0.008	0.103	0.997
		200	0.001	0.075	0.999

Note. GM: generative model. T: number of timepoints. N: sample size. SD: standard deviation of estimates across replications. SR: model fitting success rate. Bias: $\hat{\beta}_{0,MLM} - \beta_{0,MLM}$.

in the treatment effect, the backdoor criterion failed to identify this bias, as there is no backdoor path in the treatment effect. This may be explained by the fact that the classic DAG does not impose restrictions based on (a) the random slopes and (b) interaction effects.

The first research question—pertaining to the extent of treatment effect bias in MLM estimates of generative model that were nested in GM3—was investigated using the analytical multilevel model. First, we reproduced the findings by Qian et al. (2020) who found consistent estimators for GM1 and and inconsistent ones for GM3. Using additional generative models, we found that bias became indiscernable when removing from GM3 either the dependency between the random intercept and covariate (GM1), the random slope for treatment (GM3A) or the interaction effect (GM3B). This finding is in sharp contrast to the suggestion of the conditional independence assumption that the treatment effect would be biased for GM3, 3A and 3B.

The current research report leaves several avenues unexplored. First, it is unclear whether the simulation findings pertaining the generative models in Qian et al. (2020) and here generalize to other generative models. For instance, we found here that removal of a random slope or interaction from GM3 got rid of most if not all of the treatment effect bias. Thus, it is important to establish how this generalizes, so that practical recommendations can be formulated. This is particularly important, since while violations of model assumptions are never desired, the robustness against and the practical implications of a violation is what matters. Second, it is unclear how exactly the divide between the literatures pertaining to the focus of the MLM on different centering methods and within- and between-person interpretations and the focus of the GEE on marginal and conditional interpretations may be bridged. Consequently, future research could assess the implications of centering methods in MLMs on the extent to which the marginal interpretation of MLM breaks down. Third, we found that the classical DAG may not be sufficient to identify bias in the treatment effect for GM3, especially due to its lack of specification of interaction effects. Concerns regarding the use of Pearl's backdoor criterion in situations with interaction effects have been voiced by several people (see Weinberg (2007); Attia et al. (2022)). Future research could explore to what extent proposed extensions of the DAG may be useful in identifying bias in the treatment effect for GM3. Finally, it may be interesting to investigate the implications of endogenous covariates in MLMs for other types of longitudinal data analysis methods, such as dynamic structural equation modelling (DSEM; a widely used framework in the social sciences based on MLM).

Third, since the issue extends to all longitudinal data analysis methods according to Diggle (2002), in future research it may be interesting to investigate the implications of endogenous covariates in MLMs for other types of longitudinal data analysis methods, such as dynamic structural equation modelling (DSEM; a widely used framework in the social sciences based on MLM).

5 References

Attia, J., Holliday, E., & Oldmeadow, C. (2022). A proposal for capturing interaction and effect modification using DAGs. *International Journal of Epidemiology*, *51*(4), 1047–1053. https://doi.org/10.1093/ije/dyac126

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.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, *40*, 31–53. https://doi.org/10.1146/annurev-soc-071913-043455

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

Kim, Y., & Steiner, P. M. (2021). Causal graphical views of fixed effects and random effects models. *British Journal of Mathematical and Statistical Psychology*, *74*(2), 165–183. https://doi.org/10.1111/bmsp.12217

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

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

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/

Weinberg, C. R. (2007). Commentary: Can DAGs clarify effect modification? *Epidemiology*, *18*(5), 569–572. https://www.jstor.org/stable/20486428

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