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

PROPOSAL

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

September 28, 2024

Word count: 672

FETC-approved: 24-2003

Candidate journal: Psychological Methods

Introduction

Across a wide range of disciplines, researchers analyze longitudinal, observational data to investigate prospective causal relationships between variables. To address such data, the psychological sciences most commonly resort to the multilevel linear model (MLM) (McNeish et al., 2017), which is characterized by its separation of stable between-person differences from possible trends in the dynamics of within-person fluctuations (Hamaker & Muthén, 2020). Conversely, other fields, such as biostatistics and econometrics often favour generalized estimating equations (GEE, McNeish et al., 2017). While some inter-disciplinary research has been carried out on comparing these methods (McNeish et al., 2017; Muth et al., 2016; Yan et al., 2013), the causal structures under which these models may yield meaningful inferences has received scant attention in the research literature.

Recent evidence has shed light on an issue present in both methods, where controlling for moderators of the treatment effect—that depend on prior treatment—may yield biased causal estimands (Pepe & Anderson, 1994; Qian et al., 2020). Such moderators fall under the umbrella of *time-varying endogenous covariates*. As a result of including these covariates in the model, ordinary interpretations of the coefficients are no longer valid (Qian et al., 2020, p. 3).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 segregation of the disciplines that employ these methods, such critiques of the MLM not reach the applied researcher. And when it does, the technical nature and jargon of the literature in other fields may prevent these researchers from figuring out how and when these problems occur.

This indicates a need to understand (1) the consequences of including these covariates in analyses on multilevel longitudinal data, such as GEE, MLM and dynamic structural equation modeling (DSEM; a modeling framework based on MLM) and (2) how this issue may be addressed by researchers that perform longitudinal data analysis. Accordingly, this paper explains how the inclusion of time-varying covariates can yield faulty inferences and intends to establish guidelines on how this may be prevented. More specifically, the current project addresses the following research question: *to what extent does the inclusion of endogenous variables in multilevel linear models and generalized estimating equations result in biased estimands?* 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 that involve a marginal interpretation may result in bias that—depending on the circumstances—can promote the potential for faulty inferences.

Analytic Strategy

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 causal questions. 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 is regressed on one time-varying predictor and stable between person differences are modeled with a random intercept—and increase the complexity until we reach the scenario discussed by Qian et al. (2020) (i.e., MLM that includes a time-varying endogenous covariate). The primary interest of this simulation study is the comparative performance in terms of bias and the secondary interest is the efficiency in mean squared error (MSE). We rely on large sample sizes to avoid potential artifacts and estimation issues that may occur when using small samples (Bou & Satorra, 2018).

Statistical analyses pertaining to the GEE and basic MLM will be performed in `R`, version 4.2.0 (R Core Team, 2022). To fit the GEE, the R-packages `geepack` (Halekoh et al., 2006) and `gee` (Carey et al., 2024) 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 both evaluate restricted maximum likelihood estimation and ordinary maximum likelihood estimation. Extensions of the MLM from the DSEM framework will be fitted using `Mplus`, version 8.10 (Muthén & Muthén, 1998).

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

Bou, J. C., & Satorra, A. (2018). Univariate Versus Multivariate Modeling of Panel Data: Model Specification and Goodness-of-Fit Testing. *Organizational Research Methods*, *21*(1), 150–196. https://doi.org/10.1177/1094428117715509

Carey, V. J., and 4.4), T. S. L. (R. port of versions 3. 13., src/d*), C. M. (LINPACK. routines in, & updates), B. R. (R. port of version 4. 13. and. (2024). *Gee: Generalized estimation equation solver*. https://cran.r-project.org/web/packages/gee/index.html

Curran, P. J., & Bauer, D. J. (2007). Building path diagrams for multilevel models. *Psychological Methods*, *12*(3), 283–297. https://doi.org/10.1037/1082-989X.12.3.283

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

Drikvandi, R., Verbeke, G., & Molenberghs, G. (2024). A framework for analysing longitudinal data involving time-varying covariates. *The Annals of Applied Statistics*, *18*(2), 1618–1641. https://doi.org/10.1214/23-AOAS1851

Erler, N. S., Rizopoulos, D., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2019). Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research*, *28*(2), 555–568. https://doi.org/10.1177/0962280217730851

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, *10*(1), 37–48. https://www.jstor.org/stable/3702180

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

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

Mund, M., Johnson, M. D., & Nestler, S. (2021). Changes in size and interpretation of parameter estimates in within-person models in the presence of time-invariant and time-varying covariates. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.666928

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

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (Eight Edition). Muthén & Muthén.

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

R Core Team. (2022). *R: A language and environment for statistical computing*. https://www.R-project.org/

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

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550. https://journals.lww.com/epidem/fulltext/2000/09000/marginal_structural_models_and_causal_inference_in.11.aspx

Wodtke, G. T. (2020). Regression-based adjustment for time-varying confounders. *Sociological Methods & Research*, *49*(4), 906–946. https://doi.org/10.1177/0049124118769087

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