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: XXX

Candidate journal: Psychological Methods

Introduction

Recent trends in data-collection methods and rises in longitudinal research have led to a proliferation of studies that employ clustered data. To address such data, the psychological sciences most commonly resort to multilevel linear models (MLMs), which allow us insight into between-person heterogeneity of effects. Conversely, other fields, such as biostatistics and econometrics often favour generalized estimating equations (GEE, McNeish et al., 2017), which does not include random effects. However, blind application of either analysis (e.g., not for its advantages over the other in a particular case but because of the frequency of use by fellow researchers or by it being unknown) may cause researchers to obtain biased estimates that do not represent the measures that they intend to report.

Recent evidence has shed light on an issue present in both methods, where controlling for endogenous covariates may yield biased causal estimands (Qian et al., 2020). More specifically, in a standard MLM with fixed covariates, coefficients may be interpreted in the marginal (population-averaged) manner, as well as in the conditional-on-the-random-effects manner (Qian et al., 2020, p. 3). However, once we include time-varying endogenous covariates, the marginal interpretation is no longer appropriate. In a similar manner, once we include endogenous covariates when carrying out GEE, parameter estimates no longer follow the marginal interpretation unless the working correlation matrix is specified as independent (Pepe & Anderson, 1994). According to Diggle (2002), this issue not only pertains GEE and MLM, but *all* longitudinal data analysis methods.

This indicates a need to understand (1) the consequences of including these covariates in analyses on clustered data, including GEE and extensions of the basic MLM such as the (random-intercept) cross-lagged panel model (CLPM) and dynamic structural equation modeling (DSEM) and (2) how this issue may be addressed by researchers

that perform longitudinal data analysis. Accordingly, this paper explores the ways in which 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 this exploratory study, we expect that the inclusion of endogenous time-varying covariates elicits non-trivial bias in longitudinal data analyses that involve a marginal interpretation.

Analytic strategy

To introduce the problem, this paper will visualize the issue in fictional examples using the directed acyclic graph.

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, followed by an analysis of these questions using the GEE, basic MLM and extensions thereof. We consider varying number of timepoints and sample sizes. 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, such as the (RI-)CLPM and the bayesian DSEM 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

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

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

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The r package geepack for generalized estimating equations. *Journal of Statistical Software*, *15/2*, 111.

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

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/