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

1 Introduction

- Start with a paragraph describing a problem in the real-life world (so that a BoS member not familiar with statistics understands why you are pursuing research in this direction);

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), whereas other fields, such as biostatistics and econometrics often favour generalized estimating equations [GEE; McNeish et al. (2017)]. 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.

- Then, add a paragraph describing what is known in the literature;

Recent evidence has shed light on an issue present in both methods, where controlling for covariates may yield biased causal estimands (Qian et al., 2020). More specificially, 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).

- Also, describe what is not known and which gap you will address in your thesis;

...

- End with a clear research question and, if applicable, your hypothesis (for confirmatory research questions – it should be a testable hypothesis) or expectations (for exploratory research questions – can be much vaguer because of the explorative nature of the question).

...

This project aims to address and better understand the biases in the parameter estimates of a multilevel model, when there are endogenous time-varying covariates (e.g., Micro-randomized trials). The study will investigate the issue introduced in Qian et al. (2020), reframing the problem using visualizations such as Directed Acyclic Graphs (DAGs) and path diagrams. Additionally, it will compare the multilevel models with alternative approaches, including Generalized Estimating Equations (GEE) and Dynamic Structural Equation Modeling (DSEM), through simulations and empirical analysis.

In the research report, the problem will be restated using visualizations to enhance understanding. The biases associated with multilevel models in the presence of endogenous time-varying confounders will be explored, utilizing path diagrams or Directed Acyclic Graphs (DAGs) to illustrate the problem as described by Qian et al. (2020). To further investigate, data will be simulated based on the descriptions provided in the study, replicating the conditions under which time-varying confounders and endogenous covariates affect the analysis. The next step involves reproducing the results of multilevel model and GEE as discussed in Qian et al. (2020). This will include a thorough investigation of the effects of group mean centering (GMC) and centering within cluster (CWC) on the analysis.

In the thesis, the focus will shift to implementing solution proposed by Qian et

al. (2020) in Dynamic Structural Equation Modeling (DSEM) using software like Mplus. Utilizing DSEM estimation, the study will explore (1) the potential bias in parameter estimates in multilevel models when including endogenous covariates and (2) how interpretation of these parameters is affected in the presence of marginal and conditional effects. Following this, a large-scale simulation study will be conducted to compare when GEE delivers better results than DSEM. This study will evaluate the sensitivity of each method to model misspecification, particularly in areas where the model is not directly focused. We will take a multidisciplinary approach, by comparing methods for the analysis of endogenous time-varying covariates from the fields of biostatistics, econometrics, and social sciences. Finally, the thesis will include a comprehensive comparative analysis of multilevel regression, GEE, and DSEM. The advantages and disadvantages of each method will be summarized, as well as their similarities and differences, providing clear guidelines on their appropriate use cases.

In conclusion, the research report will provide a detailed visualization and simulation-based exploration of the problem, while the thesis will delve deeper into methodological comparisons and practical implementations. Together, they will offer a comprehensive evaluation of handling endogenous time-varying confounders, providing guidelines on different estimation methods for researchers across different fields.

2 Analytic strategy

- Describe how you are planning to answer your research question and how to test your hypothesis or explore your question. Be as specific as possible and preferably use an illustration to help the reader understand how all your plans connect.
- Also, include information on what data you are going to use and if you already have

ethical consent. If you do already have consent, putting the FETC case number on your front page also suffices.

- Describe what software/packages you will use.

...

3 References

- Add a reference list for the literature used in the text – you can use the formatting style of the chosen journal, but this is not mandatory.
- Add about 20 extra references that will be used for the thesis (you do not have to have read these already).

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

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