Sampling Size Planning for Intensive Longitudinal Investigations

Ziting Zhang, Xinyi Zhang, Madeline Craft, and Shelley Blozis
University of California, Davis

Abstract

Mixed-effects models are used to analyze longitudinal data where the focus is on the individual. A mixed-effects location scale (MELS) model (Hedeker, Mermelstein, & Demirtas, 2008) in particular was developed for intensive longitudinal data.

A MELS model allows researchers to study within- and between-person variation of a behavioral measure observed over time.

One question for users of a MELS model is, what are optimal combinations of sample sizes (i.e., number of repeated measures and number of subjects)) to yield unbiased estimates of the different sources of variation?

Using a Monte Carlo data simulation, this study investigates the impact of sample size combinations on the estimated parameters of a MELS model.

Introduction

With the advent of technology, such as smartphones and physiological sensors, data collection in behavioral research has expanded to include real-time data measured intensively over time. These data are often referred to as intensive longitudinal data.

Intensive longitudinal data allow researchers to study how individuals differ in the degree to which a behavioral measure varies over time, such as in the degree of within-person variation in daily positive affect (see Figure 1).

Given the increasing popularity of using a MELS model to analyze such data, this research sought to evaluate sample size requirements to obtain unbiased estimates of a MELS model by simulating data (e.g., Figure 2) that differed according to combinations of sample sizes – that is, the number of repeated measures and the number of participants.

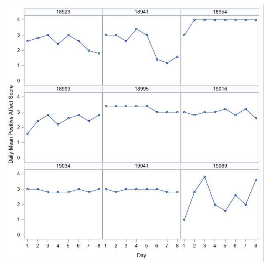

Figure 1. Daily positive affect for 9 individuals

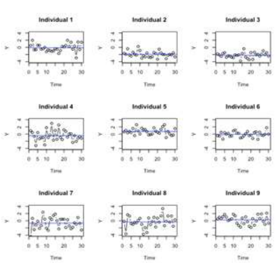

Figure 2. Simulated data for 9 individuals

Methods

Data were simulated using R according to a MELS model. 1000 data sets were simulated per experimental condition.

$$\text{Model: } Y_{ti} = \beta_{0i} + \beta_1 X_{cti} + \beta_2 X_{bti} + \gamma_1 W_{ci} + \gamma_2 W_{bi} + \varepsilon_{ti}$$

where

$$var(\beta_{0i}) = exp\{\alpha_0 + \alpha_1 W_{ci} + \alpha_2 W_{bi}\}$$

$$\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Theta})$$

where

$$\boldsymbol{\Theta} = \sigma_{ti}^2 \mathbf{I}_i$$
$$\sigma_{ti}^2 = exp\{\tau_0 + \tau_1 X_{cti} + \tau_2 X_{bti} + \tau_3 W_{ci} + \tau_4 W_{bi} + w_i\}$$
$$w_i \sim N(0, \sigma_w^2)$$

X_{cti}, X_{bti} are level-1 continuous and binary covariates, respectively
W_{ci}, W_{bi} are level-2 continuous and binary covariates, respectively

Conditions: All combinations of the number of repeated measures (J = 10, 30, 50, 75) and subject count (N = 50, 100, 200, 500, 2000), and different parameter values to model the within- and between-subject variance models. Data were analyzed using SAS PROC NLMIXED using Gaussian quadrature.

Results

We evaluated performance of the model in terms of bias and coverage of parameter estimates. Results for which <50% of samples did not converge (most often problematic for the smallest sample combination (J=10, N=50)), are excluded from the summaries reported here.

Bias = difference between the average parameter estimate across all replications and the true population value.

Coverage = number of replications for which the population value falls within the estimated confidence interval divided by the total number of replications and multiplied by 100.

Mean structure:
Estimated effects of level-1 and level-2 continuous and binary covariates on Y showed low bias (~< 5%), good coverage (~ 95%).

Covariance structure:
Within-subject: Estimated effects of level-1 and level-2 continuous and binary covariates showed low bias (~< 5%), good coverage (~ 95%).

Between-subject:
- The estimated random location variance (e^{α_0}) and scale variance σ_w^2 tended to be underestimated in general and especially so for J=10.
- Estimated effects of level-1 (α_1) and level-2 (α_2) continuous and binary covariates on the intercept variance showed low bias (~< 5%), good coverage (~ 95%) except for the level-2 binary predictor at N<200.

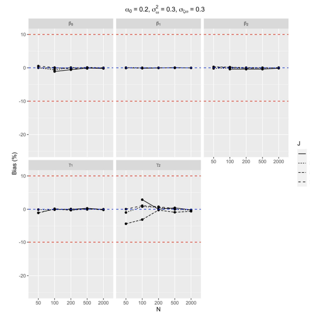

Figure 3. Bias plot for $\beta_0, \beta_1, \beta_2, \gamma_1, \gamma_2$

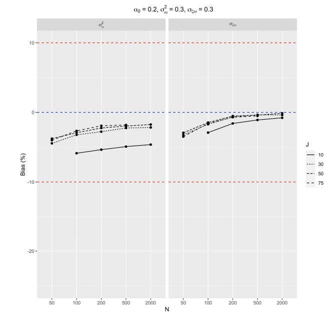

Figure 4. Bias plot for $\sigma_w^2, \sigma_{\zeta w}$

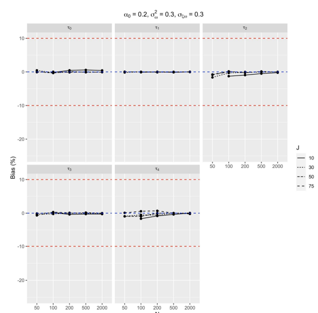

Figure 5. Bias plot for $\tau_0, \tau_1, \tau_2, \tau_3, \tau_4$

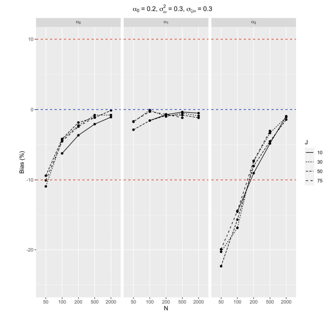

Figure 6. Bias plot for $\alpha_0, \alpha_1, \alpha_2$

Discussion

A low number of repeated measures, coupled with a small number of subjects, tends to result in non-converged solutions (especially for J=10/N=50). Otherwise, bias and coverage for fixed effects and the within-person variance model were generally acceptable for other sample size combinations. Between-person variances tended to be underestimated. The binary predictor at level 2 was problematic at N<200, but we used a binary variable with highly unbalanced counts.

Conclusions

When planning intensive longitudinal data collection and use of a MELS model to study fixed effects and within-person variation, researchers anticipating parameter estimates similar to those studied here may consider data collection for which the number of repeated measures is as small as 10 if the number of participants is about 200 or greater. With careful sample size planning, applications of a MELS model offer a rich understanding of both within- and between-person variation in longitudinal data.

Contact

Prof. Shelley Blozis
Department of Psychology, University of California
Email: sablozis@ucdavis.edu
Website: https://ilvrem.faculty.ucdavis.edu/
Phone: 530.754.9457

References

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. *Biometrics*, *64*(2), 627-634. https://doi.org/10.1111/j.1541-0420.2007.00924.x
Also see: Hedeker, D., & Nordgren, R. (2013). MIXREGLS: A Program for Mixed-Effects Location Scale Analysis. *Journal of Statistical Software*, *52*(12), 1-38. https://doi.org/10.18637/jss.v052.i12