The patient-as-fixed-effect fallacy: Consequences for statistical power and Type I errors

Examination for the course: 'Open science and reproducible research'

Kristoffer Magnusson

March 13, 2018

1 Introduction

In clinical psychology, and in many other fields, linear mixed-effects models (LMMs) have quickly risen in popularity during the 21st century (Gueorguieva and Krystal 2004). Their usage in clinical psychology is often motivated by LMMs' ability to include participants with missing data (e.g., Kahn 2011). However, LMMs are highly sensitive to some types of model misspecification (Kwok, West, and Green 2007), and investigators are faced with many modelling choices, or researchers degrees of freedom (Wicherts et al. 2016). When analyzing patients that have been repeatedly measured during a treatment, investigators must decide whether to model subjects' change over time as fixed or varying (random). The issues of treating an effect as constant for all individuals (fixed), or as varying between individuals (random), goes back a long time. For instance, in psycholinguistics Clark (1973) coined the term "language-as-fixed-effect fallacy", and Martindale (1978) soon followed by pointing out the unreasonable assumption that therapists have exactly the same success with their patients ("therapist-as-a-fixed-effect fallacy"). In an influential simulation study Barr et al. (2013), recommended to "keep it maximal", i.e. include as many random effects as possible. Others (e.g., Matuschek et al. 2017) have noted that keeping it maximal might be too conservative, and that investigators need to balance the risk of Type I and II errors.

In this paper we will focus on one of the most basic decision an analyst must make, when analyzing longitudinal treatment data—whether subjects' trajectories over time should be seen as an fixed or random effect. Specifically, this paper focuses on the consequence of ignoring subject-specif varying slopes, on both Type I and II errors.

2 Methods

In typical multilevel notation, the simplest case of the two-level model is,

Level 1
$$Y_{ij} = \beta_{0j} + \beta_{1j}t_{ij} + R_{ij} \tag{1}$$
Level 2
$$\beta_{0j} = \gamma_{00} + \gamma_{01}TX_j + U_{0j} \tag{2}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}TX_j + U_{1j} \tag{3}$$
$$\tag{4}$$

$$\text{with } \begin{pmatrix} U_{0j} \\ U_{1j} \end{pmatrix} \sim \mathcal{N} \begin{pmatrix} 0 & , \sigma_{u_0}^2 & \sigma_{u_{01}} \\ 0 & \sigma_{u_{01}} & \sigma_{u_1}^2 \end{pmatrix}, \text{ and } R_{ij} \sim \mathcal{N}(0, \ \sigma^2) \tag{5}$$

The parameter if interest is γ_{11}, which is the mean difference in change between the two groups. The aim of this paper is to investigate if accounting for subject-specific slopes ($\sigma_{u_1}^2 > 0$) is important, both when planning the study, or when analyzing the outcome.

2.1 Simulation

To investigated the impact of wrongly omitting a random slope on the risk of committing a Type I error, a Monte Carlo simulation was performed. A parallel group RCT with 11 weekly time points was assumed, with 50 participants in each treatment group. We also assumed, that at baseline there was an equal amount of variance between and within subjects, which would translate to an intraclass correlation of 0.5, if there was no variation between subjects in change over time, $\sigma_{u_1}^2 = 0$. The simulation consistent of comparing 5 different amounts of random slope variance, $\sigma_{u_1}^2/\sigma^2 = \{0, 0.01, 0.02, 0.03, 0.04\}$. We write the slope variance as a fraction of the error variance, since its the ratio that matters, not the absolute value of, $\sigma_{u_1}^2$.[1]

Power was calculated for the same model, assuming a Cohen's d of 0.5 (standardized using the pretest standard deviation). Simulations and power calculations were done in R (version 3.4.3; R Core Team, 2017), using *powerlmm* (version 0.2.0; Magnusson 2018). LMMs were fit with *lme4* (version 1.1-15; Bates et al. 2015), using restricted maximum likelihood estimation. For each model 5000 data sets were generated, resulting in a 95 % Monte Carlo CI of 0.044–0.056, for a nominal α of 0.05. The simulation code and additional material is available at: https://github.com/rpsychologist/open-science-exam.

3 Results

Figure 1 shows that ignoring even small amounts of random slope variance quickly leads to Type I errors in the range 0.15 to 0.22. Moreover, Figure 2 shows that the power of the test is greatly influenced by the assumed amount of random slope variance. For instance for a

---

[1]Here I use the famous proof: "proof is left as an exercise to the reader", or "It can easily be shown that..." ;)

study with a total of 100 participants, assuming a variance ratio of either 0, 0.01, or 0.02, means power goes from 96 % → 72 % → 54 %.
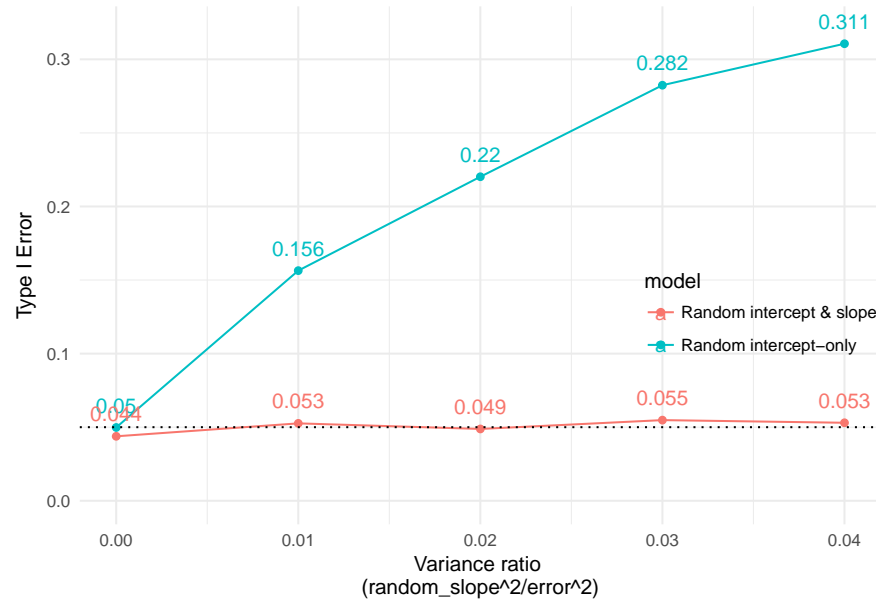

Figure 1: Type I errors for models with different amounts of true slope variance. Nominal levels is 0.05.
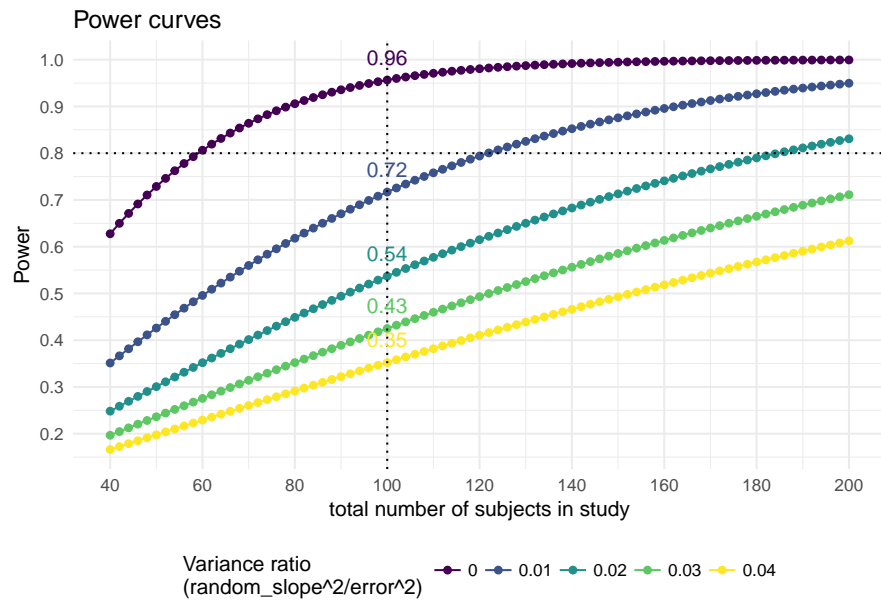

Figure 2: Power curves. Cohen's d = 0.5, 11 time points, and ICC at pretest = 0.5.

4   Conclusion

Investigators planning to estimate a random slope, but who ignore its relative magnitude in the power analysis, will end up with a test much less sensitive to detecting relevant effects than what was planned for. Moreover, and perhaps even worse, investigators who ignore a true random slope in the analysis, will substantially increase their risk of committing a Type I error. It is possible that a data driven approach to model selection could retain Type I errors at nominal levels. However, investigators must still make an *a priori* decision whether the test should be sensitive to detect relevant differences even if random slopes are included. Our recommendation is that investigators should assume that participants will change differently during a treatment trial, and include enough participants so that the statistical tests will not miss clinically relevant effects even under moderate to large amounts of heterogeneity in change. Our simulation showed that ignoring even small amounts of slope variance can substantially inflate Type I errors. Hence, we think it is appropriate that investigators who consider a random intercept-only model a reasonable choice, also provide sufficient evidence and sound arguments for this assumption. To facilitate better reasoning and more transparent reporting, we encourage authors to actually report the model's variance components.

References

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal." *Journal of Memory and Language* 68 (3): 255–78. doi:10.1016/j.jml.2012.11.001.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statistical Software* 67 (1): 1–48. doi:10.18637/jss.v067.i01.

Clark, Herbert H. 1973. "The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research." *Journal of Verbal Learning and Verbal Behavior* 12 (4): 335–59. doi:10.1016/S0022-5371(73)80014-3.

Gueorguieva, Ralitza, and John H. Krystal. 2004. "Move over ANOVA: Progress in Analyzing Repeated-Measures Data and Its Reflection in Papers Published in the Archives of General Psychiatry." *Archives of General Psychiatry* 61 (3): 310–17. doi:10.1001/archpsyc.61.3.310.

Kahn, Jeffrey H. 2011. "Multilevel Modeling: Overview and Applications to Research in Counseling Psychology." *Journal of Counseling Psychology* 58 (2): 257–71. doi:10.1037/a0022680.

Kwok, Oi-man, Stephen G. West, and Samuel B. Green. 2007. "The Impact of Misspecifying the Within-Subject Covariance Structure in Multiwave Longitudinal Multilevel Models: A Monte Carlo Study." *Multivariate Behavioral Research* 42 (3): 557–92. doi:10.1080/00273170701540537.

Magnusson, Kristoffer. 2018. "Powerlmm: Power Analysis for Longitudinal Multilevel

Models.” https://CRAN.R-project.org/package=powerlmm.

Martindale, C. 1978. “The Therapist-as-Fixed-Effect Fallacy in Psychotherapy Research.” *Journal of Consulting and Clinical Psychology* 46 (6): 1526–30. doi:10.1037/0022-006X.46.6.1526.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. “Balancing Type I Error and Power in Linear Mixed Models.” *Journal of Memory and Language* 94 (June): 305–15. doi:10.1016/j.jml.2017.01.001.

R Core Team, 2017. “R: A Language and Environment for Statistical Computing.” Vienna, Austria: R Foundation for statistical computing.

Wicherts, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. “Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid P-Hacking.” *Frontiers in Psychology* 7 (November). doi:10.3389/fpsyg.2016.01832.