Model Building Strategies in Multilevel Regression

As with ordinary least squares multiple regression and logistic regression, the vast majority of models in psychology and the social sciences are developed based on theoretical considerations (*a priori*). There are usually one or more predictors that are of primary theoretical interest and any number of covariates that the researcher wishes to control for in the analysis. Exploratory model building strategies, in which a "best" model is chosen from a large list of possible predictors also are possible, but less commonly used.

A priori model building

Generally, researchers have a set of a prior notions about which predictors will be of interest and most important in predicting an outcome. Nearly always, they will also be interested in partialing out the effects of potential confounders of that relationship. Multilevel models are typically not limited by the number of predictors (level-1 or level-2) included in the model, so there are few limits on the number of fixed effects that can be estimated. But researchers quite often will find that they cannot estimate all of the random effects that they would like to estimate because models become too complex. For instance, a model with just 4 random slopes will have 15 random effects.[1] More random effects can be estimated when the group sizes (n_j) or time points is larger, when there are more clusters (groups or individuals in the longitudinal case; Buyse, Molenberghs, Burzykowski, Renard, & Geys, 2000; Renard, Geys, Molenberghs, Burzykowski, & Buyse, 2002), and when the cluster sizes are balanced (Van der Elst, Hermans, Verbeke, Kenward, Nassiri, & Molenberghs, 2016). For designs with small groups (e.g., couples, families) or with many longitudinal designs, there will be a theoretical limit to the number of random slopes, and in other cases it will be difficult in practice to estimate too many random slopes. For example, Diallo and colleagues (Diallo, Morin & Parker, 2014) show that models that included random slopes for the quadratic effect had a high rate of nonconvergence. The slope reliability can be a useful diagnosis tool. Slopes with very low reliability (perhaps .1 or .2) will frequently be associated with difficulty in estimating random slopes.

Too many random effects will result in convergence failures (no optimal solution is found even with a large number of iterations allowed) or sometimes error messages indicating difficulties estimating random effect variances or standard errors. When the solution contains a negative variance, an error message may refer to a "boundary" condition, meaning that the estimate exceeds the allowable values for a variance. In other instances, the output may not print any errors but instead will include one or more estimates or a confidence limit for the random slope that is equal to zero. In other instances, standard errors or significance tests may not be printed. For any of these types of results, researchers should not trust the output. Eliminating the random slope from the model for the problematic variable will usually fix the problem, but likely will have some cost in the accuracy of the model.[2]

Two general strategies can be distinguished (Hox, Moerbeek, & Van de Schoot, 2018). Starting with a model with all of the desired fixed and random coefficients and then removing non-significant effects (e.g., West, Welch, & Galecki, 2022) can be called a "top-down approach," whereas starting with and empty model and adding fixed and random effects in some order can be called a "bottom-up" approach. Hox and colleagues (expanding on similar recommendations by Byrk & Raudenbush, 2002) suggest first testing the empty model, then adding the fixed effects from level-1, then adding the level-2 predictors, then adding random slopes for each level-1 predictor, one at a time, and then finally including any cross-level predictors. In this latter approach, when a random effect is not significant, it is not included in the model. Including a random slope even though it is not truly varying, is less problematic, however, assuming the model runs without problems. When testing cross-level interactions, it also is recommended that the random slopes for the level-1 variable involved in the interactions always be estimated, even if the slopes do not vary significantly (Heisig & Schaeffer, 2019; LaHuis & Ferguson, 2009).[3]

[1] To count the random effects, we consider the intercept and the four random slopes, which is five variances. The number of unique covariances is equal to (5 × 4)/2 = 10. So, with the intercept and slope variances and all of the covariances, there are 15 random effects.

[2] It has also been my experience that centering can make an important difference in estimating a multilevel model (this has been demonstrated in the longitudinal case with quadratic effects tests; Diallo et al., 2014). There are more theoretical reasons to center predictors as a general strategy, as I have pointed out in other handouts, but there is likely a (not entirely unrelated) empirical reason for doing so as well.

[3] McNeish, Stapleton, & Silverman (2017) argue that not all nested data circumstances require multilevel models and random effects. General estimating equations (GEE) or complex sampling design adjustments for clustering can address the bias in standard errors. The simulations that suggest biased standard errors for fixed effects tests when omitting random slopes (e.g., LaHuis et al., 2020) seem to contradict this recommendation.

Snijders and Bosker (2012; see Section 6.4) note simply that random intercepts suggest predictors are needed to account for between-cluster variation and that random slopes suggest predictors are needed to account for random variance. These notions often drive researchers to include additional predictors in an attempt to account for between-cluster variation. One can conceivably argue that when there is no significant variance remaining in the intercept that the model requires no additional predictors (Hoffman & Walters, 2022).

In practice, it is almost always the case that the multilevel model that can be tested in practice is not the model the researcher ideally would like to test. So, one must be prepared for some level of compromise. And comprise it is, because, assuming the slope truly varies, not estimating random slopes for a variable can bias the standard errors of the fixed effect for that variable, leading to increase type I error (LaHuis, Jenkins, Hartman, Hakoyama, & Clark, 2020). This is all the more reason for study planning for sufficient samples sizes within and between clusters.

Exploratory model building

I have focused primarily on more *a priori*-based modeling approach, partly because this seems to be heavily favored within psychology (as at least evidenced by the fairly rare instances of exploratory model building approaches in published articles) and partly because software for exploratory model building approaches is not very widely available and the relative performance of the multitude of approaches have not been very thoroughly compared in the multilevel modeling context.

Methods of choosing the “best” model based on the variance accounted for can be a challenge to implement, because fit or accounted for variance in multilevel models is not as simply defined as it is with single-level regression models (Wang & Gelman, 2014). There are, however, a number of proposed approaches to extend exploratory model selection used with single-level regression to multilevel models, such as random forests (Shi et al., 2019), decision trees (e.g., Fokkema et al., 2021; Speiser et al., 2020) and boosting model selection (e.g., Griesbach et al., 2021; Sigrist, 2022).

References

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models: applications and data analysis methods, second edition*. Sage.

Buyse M, Molenberghs G, Burzykowski T, Renard D, & Geys H. (2000). The validation of surrogate
endpoints in meta-analyses of randomized experiments. Biostatistics,1,49–67.

Diallo, T. M., Morin, A. J., & Parker, P. D. (2014). Statistical power of latent growth curve models to detect quadratic growth. *Behavior research methods, 46*, 357-371.

Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2021). Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. Psychotherapy Research, 31(3), 329-341.

Griesbach, C., Säfken, B., & Waldmann, E. (2021). Gradient boosting for linear mixed models. *The International Journal of Biostatistics, 17*(2), 317-329.

Heisig, J. P., & Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review, 35*(2), 258-279.

Hoffman, L., & Walters, R. W. (2022). Catching up on multilevel modeling. *Annual Review of Psychology, 73*, 659-689.

Hox, J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications, third edition*. Routledge.

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of statistical tests for variance components
in multilevel random coefficient modeling. *Organizational Research Methods, 12*(3), 418-435.

LaHuis, D. M., Jenkins, D. R., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2020). The effects of misspecifying the random part of multilevel models. *Methodology, 16*(3), 224-240.

McNeish, D., & Bauer, D. J. (2022). Reducing incidence of nonpositive definite covariance matrices in mixed effect models. *Multivariate Behavioral Research, 57*(2-3), 318-340.

Renard D, Geys H, Molenberghs G, Burzykowski T, Buyse M. (2002). Validation of surrogate endpoints in multiple randomized clinical trials with discrete outcomes. *Biometrical Journal, 44*, 921–935

Shi, L., Westerhuis, J. A., Rosén, J., Landberg, R., & Brunius, C. (2019). Variable selection and validation in multivariate modelling. *Bioinformatics, 35*(6), 972-980.

Sigrist, F. (2022). Gaussian process boosting. *Journal of Machine Learning Research, 23*(232), 1-46.

Snijders, T.A.B., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling, second edition.* Sage.

Speiser, J. L., Wolf, B. J., Chung, D., Karvellas, C. J., Koch, D. G., & Durkalski, V. L. (2020). BiMM tree: a decision tree method for modeling clustered and longitudinal binary outcomes. *Communications in Statistics-Simulation and Computation, 49*(4), 1004-1023.

Van der Elst, W., Hermans, L., Verbeke, G., Kenward, M. G., Nassiri, V., & Molenberghs, G. (2016). Unbalanced cluster sizes and rates of convergence in mixed-effects models for clustered data. *Journal of Statistical Computation and Simulation, 86*(11), 2123-2139.

Wang, W., & Gelman, A. (2015). Difficulty of selecting among multilevel models using predictive accuracy. *Statistics and Its Interface 8* (2): 153–160.

West, B. T., Welch, K. B., & Galecki, A. T. (2022). *Linear mixed models: a practical guide using statistical software.* Chapman and Hall/CRC.