Hierarchical Modeling: Into Statistical Practice

Alan M. Zaslavsky

Harvard Medical School, Boston, Massachusetts, USA

Pardoe, Weidner, and Friese (henceforth PWF) present a nice application of hierarchical modeling, representative of current practice of this methodology. Hierarchical regression modeling now occupies a methodological middle ground. The main principles have been established, as have a fairly general set of computational algorithms for a reasonably general set of models, implemented (in a more or less general manner) in software in some of the major statistical packages. The use of these models is unevenly diffusing across fields of application, scientific journals, and individual researchers and statistical practitioners. Thus an approach that appears novel and challenging in one context might appear routine in another. I am a firm believer in hierarchical modeling as a uniquely effective framework for coherently describing complex social phenomena and for building complex models from simple pieces; on the other hand, as a working applied statistician I have learned to be cautious about treating them as a panacea for dealing with the complexities of data. In this spirit I offer a miscellany of remarks suggested by PWF's article.

Hierarchical modeling and Bayesian modeling

PWF note (section 3) that they will "take a Bayesian approach" and proceed to describe a hierarchical model, but it should be noted that these are not the same thing. In fact, there is a long tradition of hierarchical modeling that is not philosophically Bayesian. The two are often conflated, however, because hierarchical modeling can

involve use of Bayesian inference in at least two ways. First, the inference for the hyperparameters (the unknown parameters at the highest level of the hierarchical structure, typically the variance components and the non-unit-specific coefficients) can be Bayesian, that is, based on posterior inference from a prior distribution and likelihood, or it can use some other inferential approach such as maximum likelihood. This is the sense in which PWF mean that software such as MLwiN and HLM are non-Bayesian (although they are still hierarchical), although it should be noted that MLwiN now includes options for hierarchical modeling. Specification of a prior distribution is the "deal-breaker" for non-Bayesians, but the remainder of the model is the same for the Bayesian and the maximum-likelihoodist.

The other place where Bayes's theorem is used is in inference for the random effects (or functions of them, such as small-area estimates). Use of Bayes's theorem at this level should not be distasteful to anyone on philosophical grounds, since the prior distributions used there are all functions of parameters estimated from the data. Estimators of coefficients and random effects in hierarchical models can also be derived from non-Bayesian principles (minimization of mean squared error under a posited model), leading to similar or identical results (with normal linear models) to those obtained by applying Bayes's theorem; hence we have the BLUP (Best Linear Unbiased Predictor) or EBLUP, parallel to Bayes or Empirical Bayes inference. (See the review by Robinson, 1991.) Use of Bayesian computational tools such as Markov chain Monte Carlo samplers at the lower levels might encourage adoption of a more completely Bayesian approach, however, especially in complex models for which analytic approximations to the likelihood are difficult to implement.

Interpretation of variance components and explained variance

There are at least three classes of targets of inferences from hierarchical models (specifically, multilevel models with GLM specifications at the bottom level and normal linear specifications above it): (1) general (non-unit-specific) regression coefficients, the focus of PWF; (2) predictions (functions of coefficients and unit-specific random effects), the objective of small-area estimation or profiling, and (3) variance components. Of these, the variance components are least likely to be reported, although reporting the fraction of variance explained (R^2) is routine for simple regression models. The concept of explained variance is more complex in a multilevel model (see discussion in Snijders and Bosker 1999, ch. 7), but still meaningful. Furthermore, the unexplained variance of higher-level effects scientifically interesting as an indication of the potential explanatory power of characteristics of the corresponding units that are omitted from the model or as yet unobserved.

Such measures also can have substantial policy relevance. In PWF's analysis, unexplained county-level variance measures the residual unevenness in sentencing practices, a serious equity concern for criminal justice. Similarly, residual variation of educational or health care processes or outcomes reflects the effects of unmeasured factors associated with a type of unit (school, teacher, health care provider). If these factors represent practices of that unit rather than unmeasured variation in the populations (of students or patients), that variation might be interpreted as a measure of the potential for improvement if the practice of all units at that level could be brought up the level of the average unit, or of one moderately above average (say 1 SD higher). While a variance component might be a difficult abstraction for the statistically unsophisticated

reader, it is not hard to describe and comprehend the difference between predictions for a moderately high and a moderately low unit (say, ±1 SD from average), assuming other variables are fixed at their mean values, and to compare it to typical effects of other variables operating at that level.

Missing data

PWF's *ad hoc* approach to missing data seems discordant with their advanced technology for model specification and estimation. Perhaps the fraction of missing data is small enough that this is acceptable. Be that as it may, PWF's missing data regression models (section 5) ignore much of the available information that could improve imputations of the missing data. For example, the intraclass correlation of offender's race is likely to be substantial, so knowing the racial composition of the offender's county would improve imputations for individuals.

In general, the outcome variable of the complete-data analysis (here, incarceration) should almost always be included in the imputation model, even if that appears to be a counterintuitive "double use" of the outcome data. In essence, imputation is a means for representing (by drawing from) the joint distribution of the data so that the conditional distributions of interest (outcome given predictors) can be inferred from it. Thus, the covariation of the outcome with potential predictors is among the most important aspects of that joint distribution. Indeed, imputing the predictors independent of the outcome will almost always lead to attenuation of the regression coefficients in the final model. This can be corrected most simply by including incarceration as a predictor of race, sex, etc. in the imputation model. PWF's use of generalized Bayesian software also allows

them to incorporate a marginal model for the predictors into the inference, thereby automatically combining the likelihood from the outcome (incarceration) model with the prior from a marginal model (stripped down to the essentials most important to imputation). Such a principled approach confers greater confidence in the validity of the inference, even with substantial rates of missing data.

Contextual variables, the Southeast, mapping, and causality

Policy analyses commonly use variation among political/geographical units (states or counties, in the U.S.) to characterize relationships of outcomes to characteristics that might affect policy in those units. Hierarchical models are natural tools for such investigations because if correctly specified, they can distinguish among relationships operating at different levels of units, for example the association of race with sentencing of individuals versus the association of racial composition with areas with their sentencing patterns. To distinguish between these two effects, the model must include both variables as predictors, as did PWF.

Typically many contextual variables (compositional or other characteristics of jurisdictions or areas) are strongly correlated. The "story" that emerges from an analysis of associations thus can be strongly shaped by the choice of predictors. Given this, it is incumbent on the analyst to report a wide range of possibly related variables and consider which sets of variables cannot be distinguished in the analysis, especially when the number of units is modest (39 in this study).

Policy outcomes also are likely to display strong geographical correlations. In the United States, a major contributor to this geographical correlation is the complex of differences between the southeastern region (with its conservative political tradition, concentration of Black population, and relatively stingy social services) and the remainder of the country. When these differences predominate, a study with an apparent sample size of 50 (states) or over 3000 (counties) might effectively be based on a sample size of two (major regions), a common fallacy in analyses across U.S. geographical areas. PWF have sensibly chosen to include a regional indicator variable in their model to reflect these differences.

More generally, the assumption of conditional independence across units is a strong one and might not be justified, although it is essential to the validity of inferences under the models unless dependencies are explicitly modeled. One simple diagnostic is to map the geographical effects (predictions or residuals from the regression model for geographical units); patterns in these maps could alert the knowledgeable analyst to spatial correlations and/or potential confounding variables.

When does hierarchical modeling make a difference?

While hierarchical models are appealing for their unified expression of models for hierarchically structured phenomena, they also are more complex to estimate and interpret than single-stage models. Thus it is worth considering when full hierarchical modeling is most essential and when simpler strategies will work, such as estimating regression models within individual areas and then regressing the area-specific coefficients on area characteristics. Broadly I would argue that the hierarchical model is

most useful when the interunit reliability of estimates is not very close to 1, so the error of estimation within units cannot be ignored. (Reliability in a univariate variance components problem can be defined as $\tau^2/(\tau^2+V_i)$, where τ^2 is between-unit variance component and V_i is the within-unit estimation variance.) With very large samples in each unit, maximum likelihood estimates can be analyzed like data at the next level and then the analysis might be broken up into two distinct stages, which is not possibly with small sample sizes, especially with unbalanced data structures..

The data set analyzed by PWF is most likely in the range in which this naïve approach would not work very well, given the limited sample in many counties and the dichotomous outcome. This makes this set an appropriate candidate for hierarchical modeling.

PWF have performed a service by demonstrating a nice hierarchical modeling application. Statisticians and other researchers should continue to work to bring hierarchical modeling methods into standard curricula and widespread use.

REFERENCES:

Robinson, G. K. (1991), "That BLUP is a good thing: The estimation of random effects," *Statistical Science*, 6, 15-32.

Snijders, T. A. B. and Bosker, R. J. (1999), *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Newbury Park, CA: Sage Publications.