Multilevel (Hierarchical) Modeling: What It Can and Cannot Do

Andrew GELMAN

Department of Statistics and Department of Political Science
Columbia University
New York, NY 10027
(*gelman@stat.columbia.edu*)

Multilevel (hierarchical) modeling is a generalization of linear and generalized linear modeling in which regression coefficients are themselves given a model, whose parameters are also estimated from data. We illustrate the strengths and limitations of multilevel modeling through an example of the prediction of home radon levels in U.S. counties. The multilevel model is highly effective for predictions at both levels of the model, but could easily be misinterpreted for causal inference.

KEY WORDS: Contextual effects; Hierarchical model; Multilevel regression.

1. INTRODUCTION

Multilevel modeling is a generalization of regression methods, and as such can be used for a variety of purposes, including prediction, data reduction, and causal inference from experiments and observational studies (for recent reviews, see Kreft and De Leeuw 1998; Snijders and Bosker 1999; Raudenbush and Bryk 2002; Hox 2002). Compared with classical regression, multilevel modeling is almost always an improvement, but to varying degrees; for prediction multilevel modeling can be essential, for data reduction it can be useful, and for causal inference it can be helpful.

We illustrate the strengths and limitations of multilevel modeling through an example of the prediction of home radon levels in U.S. counties.

2. MULTILEVEL MODELING FOR ESTIMATING HOME RADON LEVELS

2.1 Background and Model

Radon is a carcinogen—a naturally occurring radioactive gas whose decay products are also radioactive—known to cause lung cancer in high concentrations and estimated to cause several thousand lung cancer deaths per year in the United States. The distribution of radon levels in U.S. homes varies greatly, with some houses having dangerously high concentrations. To identify areas of high radon exposure, the Environmental Protection Agency coordinated radon measurements in a random sample of more than 80,000 houses throughout the country.

To simplify the problem somewhat, our goal in analyzing these data was to estimate the distribution of radon levels in each of the approximately 3,000 U.S. counties, so that homeowners could make decisions about measuring or remediating the radon in their houses based on the best available knowledge of local conditions. For the purpose of this analysis, the data were structured hierarchically: houses within counties. (If we were to analyze multiple measurements within houses, then there would be a three-level hierarchy of measurements, houses, and counties.)

In performing the analysis, we had an important predictor—whether the measurement was taken in a basement. (Radon comes from underground and can enter more easily when a house is built into the ground.) We also had an important county-level predictor—a measurement of soil uranium that was available at the county level. We fit a model of the form

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\alpha_j + \beta x_{ij}, \sigma_y^2), \quad \text{for } i = 1, \dots, n_j, j = 1, \dots, J, \\ \alpha_j &\sim \mathrm{N}(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2), \quad \text{for } j = 1, \dots, J, \end{aligned} \tag{1}$$

where y_{ij} is the logarithm of the radon measurement in house i within county j, x_{ij} is an indicator for whether the measurement was taken in a basement, and u_j is the log uranium level in county j. The errors with variance σ_y^2 in the first line of (1) represent "within-county variation," which in this case includes measurement error, natural variation in radon levels within a house over time, and variation between houses (beyond what is explained by the basement indicator). The errors with variance σ_α^2 in the second line represent variation between counties beyond what is explained by the county-level uranium predictor. The hierarchical model allows us to fit a regression model to the individual measurements while accounting for systematic unexplained variation among the 3,000 counties.

Equivalently, the model can be written as a single-level regression with correlated errors

$$\mathbf{y} \sim \mathrm{N}(\gamma_0 \mathbf{1} + \gamma_1 \mathbf{G}\mathbf{u} + \beta \mathbf{x}, \sigma_y^2 \mathbf{I} + \sigma_\alpha^2 \mathbf{G}\mathbf{G}^T),$$

where $\mathbf{G}$ is the $n \times J$ matrix of county indicators.

The model can be expanded in many ways, most naturally by adding more predictors at the individual and county levels and by allowing the slope β and the intercept α to vary by county. For the purposes of this article, however, model (1) is general enough. We further simplify by focusing on a subset of our data—the 919 houses from the state radon survey of the 85 counties of Minnesota (Price, Nero, and Gelman 1996). We fit the model using hierarchical Bayes methods (e.g., Gelman, Carlin, Stern, and Rubin 2003). The posterior density

TECHNOMETRICS, AUGUST 2006, VOL. 48, NO. 3
DOI 10.1198/004017005000000661

is simply,

$$p(\alpha, \beta, \gamma, \sigma_y, \sigma_\alpha|\mathbf{y}, \mathbf{x}, \mathbf{u})$$

$$\propto \prod_{j=1}^{J}\prod_{i=1}^{n_j} \mathrm{N}(y_{ij}|\alpha_j + \beta x_{ij}, \sigma_j^2) \prod_{j=1}^{J} \mathrm{N}(\alpha_j|\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2), \quad (2)$$

where $\mathrm{N}(\cdot|M, S^2)$ represents the normal density function with mean M and standard deviation S and assuming a uniform prior distribution on γ, σ_y, and σ_α, which is reasonable given that the number of counties, J, is large (Gelman 2006).

2.2 Data Reduction: Estimating Associations

Figure 1 displays the estimated multilevel model for a selection of 8 of the 85 counties in Minnesota, along with the completely pooled and unpooled regression line for each county. (The completely pooled line is $y = \alpha + \beta x$, with a common line for all counties, and the unpooled lines are $y = \alpha_j + \beta x$, with the 85 α_j's estimated by least squares.)

Compared with the two classical estimates (no pooling and complete pooling), the inferences from the multilevel models are more reasonable. At one extreme, the complete-pooling method gives identical estimates for all counties, which is particularly inappropriate for this application, whose goal is to identify the locations in which residents are at high risk of radon. At the other extreme, the no-pooling model overfits the data; for example, it gave an implausibly high estimate of the average radon levels in Lac Qui Parle County, in which only two observations were available.

Although the specific assumptions of model (1) could be questioned or improved, it would be difficult to argue against the use of multilevel modeling for the purpose of estimating radon levels within counties.

Another advantage of multilevel modeling for this application is that it allows us to study the relation of the county parameters to county-level predictors—in this case, the uranium measurement, as displayed in Figure 2. It would be possible to estimate this second-level relation using classical regression, first fitting the no-pooling model to estimate the α_j's and then fitting county-level regression to the $\hat{\alpha}_j$'s. The multilevel model has the appeal of fitting the two levels together and actually can be implemented using a Gibbs sampler alternating between the data-level and county-level regression steps. So the point here is not whether the estimates are identified as "multilevel," but whether they take into account the estimation uncertainty of the α_j's, as is done in Figure 1 by shrinking toward the complete-pooling estimate.

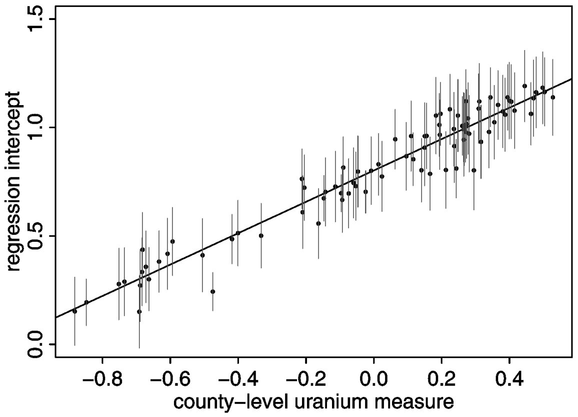

Figure 2. Estimated County Coefficients α_j (±1 standard error) Plotted versus County-Level Log Uranium Measurement u_j, Along With the Estimated Multilevel Regression Line $\alpha = \gamma_0 + \gamma_1 u$. The county coefficients roughly follow the line but not exactly; the deviation of the coefficients from the line is captured in σ_α, the standard deviation of the errors in the county-level regression.

2.3 Prediction

Perhaps the clearest advantage of multilevel models comes in prediction. In our example we can predict the radon levels for new houses in an existing county or a new county. (Because we actually have data on all 85 counties in Minnesota, that would be a new county in a neighboring state.)

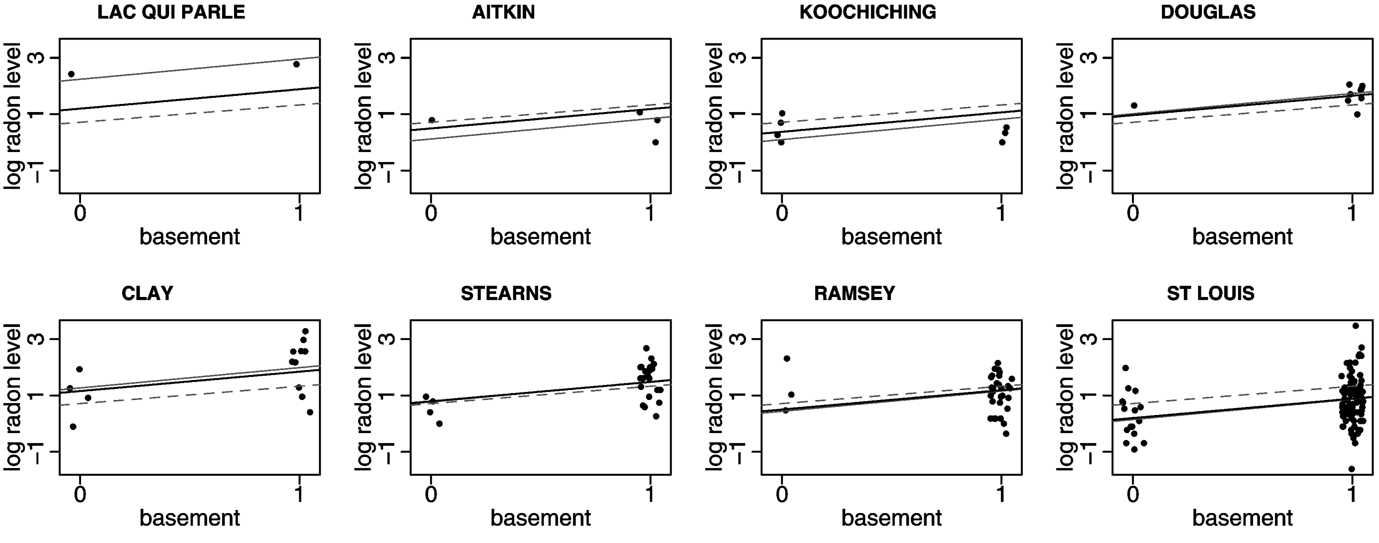

Figure 1. Multilevel (partial pooling) Regression Lines $y = a_j + \beta x$ Fit to Radon Data From Minnesota, Displayed for Eight Counties j With a Range of Sample Sizes. Light-colored dotted and solid lines show the complete-pooling and no-pooling estimates. The x-positions of the points are jittered slightly to improve visibility.

We can use cross-validation to formally demonstrate the benefits of multilevel modeling. We perform two cross-validation tests: first removing single data points and checking the prediction from the model fit to the rest of the data, then removing single counties and performing the same procedure. For each cross-validation step, we compare complete-pooling, no-pooling, and multilevel estimates. Other cross-validation tests for this example were performed by Price et al. (1996).

When removing individual data points and refitting each model, the root mean squared cross-validation prediction errors are .84 for complete pooling, .86 for no pooling, and .79 for multilevel modeling. (In making this comparison, we exclude measurements that, when removed, make the no-pooling model impossible to fit. See Fig. 1, for example; if either of the houses in Lac Qui Parle County or the no-basement house in Aitkin County were removed, then it would not be possible to estimate the regression slope from that county's data alone.)

When removing entire counties one at a time, we summarize by the errors of the predicted county mean responses (given the county-level uranium and the basement information for the houses in the excluded county). The root mean squared predictive errors at the county level are .50 for complete pooling and .40 for multilevel modeling. (Cross-validation cannot be performed at the county level for the no-pooling model because it is does not allow estimation of a county's radon level using data from other counties.)

The multilevel model gives more accurate predictions than the no-pooling and complete-pooling regressions, especially when predicting group averages.

2.4 Causal Inference

We now consider our model as an observational study of the effect of basements on home radon levels. The study includes houses with and without basements throughout Minnesota. The proportion of homes with basements varies by county (see Fig. 1), but a regression model should address that lack of balance by estimating county and basement effects separately. (As noted earlier, we set aside the possibility that basement effects might vary by county.) The estimated coefficient β in model (1) is .67 (with a standard error of .06), implying that within any given county, houses with basements have typical radon levels $\exp(.67) = 2.0$ times higher than houses without basements. (Measurements are made in the lowest living area of the house. The "basement effect" on living-area radon levels thus includes differences between houses explainable by having a basement, as well as differing radon concentrations among levels of a particular house. For our purposes here, we combine these effects.)

So far, so good. However, a complication arises if we consider the possibility of correlation between the individual-level predictor, x, and the county-level error, $\alpha_j - \gamma_0 - \gamma_1 u_j$ (see, e.g., Woolridge 2001 for a discussion of this sort of correlation in multilevel models). By simply multiplying the likelihood and prior densities in (1), the posterior density (2) implicitly assumes that the county errors are independent of x. We can allow for possible correlation by including $\bar{x}_j$, the average of x within county j (i.e., the proportion of basements in the houses in county j in the dataset) into the group-level regression,

$$\alpha_j \sim \mathrm{N}(\gamma_0 + \gamma_1 u_j + \gamma_2 \bar{x}_j, \sigma_\alpha^2), \quad \text{for } j = 1, \ldots, J.$$

The new group-level coefficient γ_2 is estimated at $-.39$ (with standard error .20), implying that, all other things being equal, counties with more basements tend to have lower baseline radon levels. For the radon problem, the county-level basement proportion is difficult to interpret directly as a predictor, and we consider it a proxy for underlying variables (e.g., the type of soil prevalent in the county).

In other settings, especially in social science, individual averages used as group-level predictors are often interpreted as "contextual effects." For example, the presence of more basements in a county would somehow have a radon-lowering effect. This makes no sense here, but it serves as a warning that, with identical data of a social nature (e.g., consider substituting "income" for "radon level" and "ethnic minority" for "basement" in our study), it would be easy to leap to a misleading conclusion and find contextual effects where none necessarily exist.

This is related to the "ecological fallacy" studied in geography (see Wakefield 2003 for a recent review with many references), in which group-level correlations can be mistakenly attributed to individual-level causes, but our setting is slightly different in that both individual- and group-level data are available. The available data are modeled correctly but the group-level coefficient γ_2 can be misinterpreted causally. This is related to the problem in meta-analysis that between-study variation is typically observational even if individual studies are randomized experiments (see Rubin 1989; Gelman, Stevens, and Chan 2003).

3. DISCUSSION

Multilevel modeling is an increasingly popular approach to modeling hierarchically structured data, outperforming classical regression in predictive accuracy. This is no surprise, given that multilevel modeling includes least squares regression as a special case. One intriguing feature of multilevel models is their ability to separately estimate the predictive effects of an individual predictor and its group-level mean, which are sometimes interpreted as "direct" and "contextual" effects of the predictor. As we have illustrated in this article, these effects cannot necessarily be interpreted causally for observational data, even if these data are a random sample from the population of interest. Our analysis arose in a real research problem (Price et al. 1996) and is not a "trick" example. The houses in the study were sampled at random from Minnesota counties, and there were no problems of selection bias.

ACKNOWLEDGMENTS

The authors thank Phillip Price for collaboration with the radon example, Jan de Leeuw and a reviewer for helpful comments, and the National Science Foundation for financial support.

[Received June 2005. Revised September 2005.]

REFERENCES

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, to appear.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.), London: CRC Press.

Gelman, A., Stevens, M., and Chan, V. (2003), "Regression Modeling and Meta-Analysis for Decision Making: A Cost-Benefit Analysis of a Incentives in Telephone Surveys," *Journal of Business & Economic Statistics*, 21, 213–225.

Hox, J. (2002), *Multilevel Analysis: Techniques and Applications*, Mahwah, NJ: Lawrence Erlbaum Associates.

Kreft, I., and De Leeuw, J. (1998), *Introducing Multilevel Modeling*, London: Sage.

Price, P. N., Nero, A. V., and Gelman, A. (1996), "Bayesian Prediction of Mean Indoor Radon Concentrations for Minnesota Counties," *Health Physics*, 71, 922–936.

Raudenbush, S. W., and Bryk, A. S. (2002), *Hierarchical Linear Models* (2nd ed.), Thousand Oaks, CA: Sage.

Rubin, D. B. (1989), "A New Perspective on Meta-Analysis," in *The Future of Meta-Analysis*, eds. K. W. Wachter and M. L. Straf, New York: Russell Sage Foundation, pp. 155–166.

Snijders, T. A. B., and Bosker, R. J. (1999), *Multilevel Analysis*, London: Sage.

Wakefield, J. (2003), "Sensitivity Analyses for Ecological Regression," *Biometrics*, 59, 9–17.

Woolridge, J. M. (2001), *Econometric Analysis of Cross-Sectional and Panel Data*, Cambridge, MA: MIT Press.