Multilevel (hierarchical) modeling: what it can and can't do*

Andrew Gelman†

July 29, 2005

Abstract

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.

Keywords: 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 (see Kreft and De Leeuw, 1998, Snijders and Bosker, 1999, Raudenbush and Bryk, 2002, and Hox, 2002, for recent reviews). Compared to classical regression, multilevel modeling is almost always an improvement, but to different 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

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 concentration, 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. In order to identify the areas

*We 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.

†Department of Statistics and Department of Political Science, Columbia University, New York, USA, gelman@stat.columbia.edu, http://www.stat.columbia.edu/∼gelman/

with high radon exposures, the Environmental Protection Agency coordinated radon measurements in a random sample of over 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 3000 counties in the U.S., 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, 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,

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

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 3000 counties.

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

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

where 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 as well as the intercept α to vary by county. For the purposes of this paper, 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, Gelman, and Nero, 1996). We fit the model using hierarchical Bayes methods

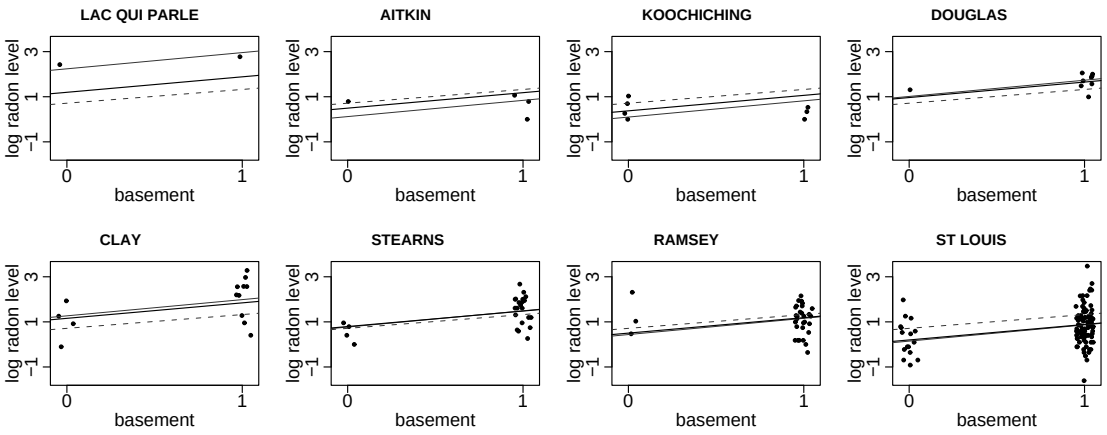

Figure 1: Multilevel (partial pooling) regression lines $y = a_j + \beta x$ fit to radon data from Minnesota, displayed for eight counties j with a range of sample sizes. Light-colored dotted and solid lines show the complete-pooling and no-pooling estimates. The x-positions of the points are slightly jittered to improve visibility.

(e.g., Gelman et al., 2003). The posterior density is simply,

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

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 $\gamma, \sigma_y, \sigma_\alpha$, which is reasonable given that the number of counties J is large (Gelman, 2005).

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 to 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 giving 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

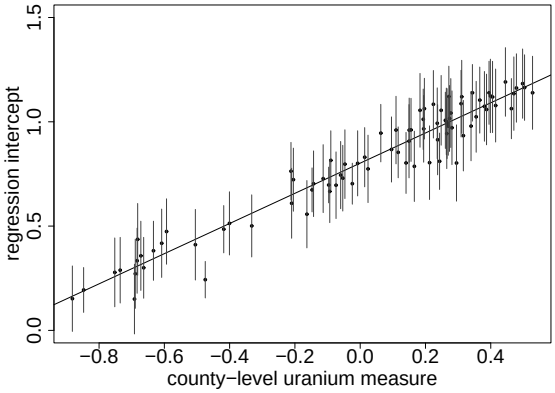

Figure 2: Estimated county coefficients α_j (± 1 standard error) plotted vs. county-level log uranium measurement u_j, along with the estimated multilevel regression line $\alpha = \gamma_0 + \gamma_1 u$. The county coefficients roughly follow the line but not exactly; the deviation of the coefficients from the line is captured in σ_α, the standard deviation of the errors in the county-level regression.

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 can actually 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.

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 for a new county. (Since we actually have data on all 85 counties in Minnesota, that would be a new county in a neighboring state.)

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 in Price, Nero, and Gelman

(1996).

When removing individual data points and re-fitting each model, the root-mean-squared cross-validation prediction errors are 0.84, 0.86, and 0.79 for complete pooling, no pooling, and multilevel modeling. (In making this comparison, we exclude measurements which, when removed, make the no-pooling model impossible to fit. For example, see Figure 1: if either of the houses in Lac Qui Parle County or the no-basement house in Aitkin County is 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 0.50 and 0.40 for complete pooling and multilevel modeling, respectively. (Cross-validation cannot be performed at the county level for the no-pooling model since it is does not allow a county's radon level to be estimated 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.

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 Figure 1), but a regression model should address that lack of balance by separately estimating county and basement effects. (As noted earlier, we set aside the possibility that basement effects might vary by county.) The estimated coefficient β in model (1) is 0.67 (with a standard error of 0.06), implying that, within any given county, houses with basements have typical radon levels $\exp(0.67) = 2.0$ times higher than houses without. (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 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, for example, 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 the county errors are independent of x. We can allow for possible dependence by including $\bar{x}_j$, the average of x within county j (that is, the proportion of basements in the houses in county j in the

dataset), into the group-level regression:

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

The new group-level coefficient γ_2 is estimated at -0.39 (with a standard error of 0.20), implying that, all other things equal, counties with more basements tend to have lower baseline radon levels. For the radon problem, the county-level basement proportion is difficult directly to interpret as a predictor, and we consider it a proxy for underlying variables (for example, the type of soil that is prevalent in the county).

In other settings, especially in social science, individual averages that are 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 (for example, 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, and 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 paper, 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, Nero, and Gelman, 1996) and is not a "trick" example. The houses in the study were sampled at random from the counties of Minnesota, and there were no problems of selection bias.

References

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. 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 and Economic Statistics*.

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. Mahwah, N.J.: 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*, second edition. Thousand Oaks, Calif.: Sage.

Rubin, D. B. (1989). A new perspective on meta-analysis. In *The Future of Meta-Analysis*, ed. K. W. Wachter and M. L. Straf. New York: Russell Sage Foundation.

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 Section and Panel Data*. Cambridge, Mass.: MIT Press.