Introduction to multilevel modeling

Instructor: Li An

The term "multilevel" means different levels or units of analysis, which are typically (not always) hierarchically nested: level 1 (lowest; e.g., pupils), then level 2 that each contains level 1 individuals (e.g., schools), and so on.

Groups and their members can exert influences on each other, suggesting that the variability in the outcome of interest may have contribution from both individual members and groups. Ignoring grouping effect may give rise to incorrect conclusion, such as finding differences and relationships that do not exist (the "formal" style example on p.508). The reason may be that individuals within the same group are correlated or clustered.

Multilevel data structures (membership): 1) hierarchical structure: individuals→ year/time→ neighborhood; or response or case→person→neighborhood (the classic design for data collection); 2) non-hierarchical structure: cross-classified structure: individuals at level 1 and workplace and residential neighborhood at level 2; or multiple membership structure: individuals at level 1 and neighborhoods that overlap (i.e., some individuals belong to more than one neighborhood).

Motivations of MLM: 1) Evaluating sources of variations in the outcome: from compositional or from contextual differences? 2) Understanding varying roles of contextual differences on different individual groups. 3) Ascertaining the relative importance of individual and neighborhood covariates. [Note: the term ecology or ecological is kind of used differently, which means context or contextual (related to Fig.C.7.1).].

MLM formulation: Suppose that a number of individual people, belonging to 50 neighborhoods, are sampled for a health study. The dependent variable is a certain health score y_{ij}, and two exemplified independent variables are individual level measure poverty (x_{1ij}; 1 for poor and 0 for non-poor) and a neighborhood-level socioeconomic deprivation index w_{1j} (note i for individual people and j for neighborhoods). Then the model at level-1 is:

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + e_{0ij} \qquad (1)$$ (C.7.1 in the paper—see footnote #1)

This equation means: the health score of an individual i who is in neighborhood j (y_{ij}) can be expressed as the sum of the intercept for neighborhood j (in this example numerically equal to the mean health score of all non-poor people in neighborhood j) in which individual i is located (β_{0j}), a portion explained by individual i's poverty level (x_{1ij}) with a global coefficient (β_1), and an individual level error term (e_{0ij})[2].

[1] The book chapter by Subramanian (2010) with full information attached at the end.

[2] Pay attention to the subscript use: i is for individuals and j for neighborhoods or higher level units; 0 for intercept or error on the y axis, and 1 for a predictor variable (we could have variables x_2, x_3, and so on). Also pay attention to variable or parameter names: English letters for the dependent or response variable (y) or independent variables (x) and errors (e at individual level and u at neighborhood level), Greek letters (α, β) for coefficients or parameters that need to be estimated from data.

We can further break the mean health score of neighborhood j into two parts: a global average for all the neighborhoods (β_0) and a departure of neighborhood j from this global average (u_{0j}):

$$\beta_{0j} = \beta_0 + u_{0j} \qquad (2)^{\text{(C.7.2 in the paper)}}$$

Substituting Equation (2) into (1), we get the basic multilevel model Equation (3):

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (u_{0j} + e_{0ij}) \qquad (3)^{\text{(C.7.3 in the paper)}}$$

which is comprised of a fixed part ($\beta_0 + \beta_1 x_{1ij}$) and a random part ($u_{0j} + e_{0ij}$). The two error terms are assumed to be uncorrelated and follow two normal distributions with zero mean. Of particular interest are the neighborhood level errors u_{0j}, which can be calculated using Equation C.7.7.

Contrary to the random effects model as shown in Equation (3), it is possible to create a set of neighborhood dummy variables (49 if we have 50 neighborhoods) and put them into a regular OLS model (termed fixed effects model; Equation C.7.8, detail skipped). It is pointed out that compared to the fixed effects model, the random effects model (related to the above u_{0j}) is superior for reasons like use of information for all neighborhoods and keeping the neighborhood coefficients from shrinking to an overall coefficient (i, ii, and iii on p.517). Although these two types of models are related (from Equation C.7.9 to C.7.12), it is shown that the fixed effects model is unsuitable for questions that the random effects model can answer.

The above model (Equation (3)) is a random intercept model, i.e., the intercept β_{0j} has a random part u_{0j} that varies from neighborhood to neighborhood. However, we can also relax the "fixed" coefficient β_1 and let it change from neighborhood to neighborhood. Following similar procedure as how Equation (3) is derived from (1) and (2), we can get a random coefficient (or random slope) model as below:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (u_{0j} + u_{1j} x_{1ij} + e_{0ij}) \qquad (4)^{\text{(C.7.16 in the paper)}}$$

Where the poverty differential (or the slope of variable x_{1ij}) is no longer constant across neighborhoods, but varies at a random amount u_{1ij}. There are some normal distribution based assumptions about variance and covariance for the three random terms u_{0j}, u_{1ij}, and e_{0ij}. The model expressed in Equation (4), often called *random intercepts and slopes* (name not appropriate though) model, models variance of the above error terms as function of predictors. The above *random intercepts and slopes* model assumes a homoscedastic variance for the level-1 error term e_{0ij}, which does not have to be so and can be further modeled as a function of predictor variables, e.g., we can use $e_{1ij} x_{1ij} + e_{2ij} x_{2ij}$ to replace e_{0ij} in C.7.20 (x_{1ij} and x_{2ij} are dummy variables representing the non-poor and poor people; detail skipped).

The above models (Equation (3) and (4)) have not used the level-two variable neighborhood-level socioeconomic deprivation index w_{1j}. We can assume that this variable can be used to predict the neighborhood-level intercept (β_{0j}) and slope (β_{1j})—with some mathematical formulation work, the multilevel model can be written as:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \alpha_1 w_{1j} + \alpha_2 w_{1j} x_{1ij} + (u_{0j} + u_{1j} x_{1ij} + e_{1ij} x_{1ij} + e_{2ij} x_{2ij}) \qquad (5)^{\text{(C.7.23 in the paper)}}$$

Note that α_1 and α_2 are coefficients of the neighborhood-level index w_{1j} and an interaction between the two level-1 and level-2 variables. This model allows for not only individual-(by β_1) and neighborhood-level (by α_1) contribution to the dependent variable, but also the cross-level effect (by α_2).

All the above models have a common feature: they model the average (via intercept) of the dependent variable and variation (via slope) around the average at different levels. MLM model can be used in instances with 3 or more levels, with different types of response variables (e.g., binary, proportions, counts, multinomial), with time series data (e.g., measurements at discrete times as level-1 data), with multiple membership (e.g., 20% for neighborhood 1 and 80% for neighborhood 22), and with consideration for spatial autocorrelation (e.g., near neighborhoods exert higher influences than remote ones). In real world, relationships exist at multiple levels, where each is important in its own right. Therefore MLM helps to address the ecological fallacy (higher level observations imposed on lower level entities), individual fallacy (ignoring higher level contexts), and atomistic fallacy (imposing lower-level relationships to higher levels). Technically, MLM may provide more accurate standard errors and thus robust significance tests, and also better estimate influences on the response variable from different levels.

When applying MLM, it is essential to take into consideration a few issues, such as choice of higher levels, representativeness of sampled neighborhoods, sample size (more neighborhoods if possible), and limitation of MLM in making causal inferences.

Subramanian, S.V. (2010). Multilevel modeling. In Fischer M.M. and A. Getis (eds.), Handbook of Applied Spatial Analysis: Software Tools, Methods, and Applications, pp.507-525 (C.7). Springer: New York.