Hierarchical Linear Models

Jeff Gill

University of Florida

GLOSSARY

[**analysis of covariance model (ANCOVA)**] A varying intercept HLM with the second-level effect fixed across groups.

[**between unit model**] The component of an HLM that describes the variability across the groups.

[**context level variables**] Variables defined at the second or higher level of the HLM.

[**EM algorithm**] An iterative procedure for computing modal quantities when the data are incomplete.

[**empirical Bayes**] Using the observed data to estimate terminal-level hierarchical model parameters.

[**exchangeability**] The joint probability distribution is not changed by re-ordering the data values.

[**fixed effects coefficients**] Model coefficients that are assumed to pertain to the entire population and therefore do not need to be distinguished by subgroups.

[**hierarchy**] The structure of data that identifies units and subunits in the form of nesting.

[**interaction term**] A model specification term that applies to some mathematical composite of explanatory variables, usually a product.

[**random coefficients regression model**] An HLM where the only specified effect from the second-level is seen through error terms.

[**random effects coefficients**] Model coefficients that are specified to differ by subgroups and are treated probabilistically at the next highest level of the model.

[**two-level model**] An HLM specifying a group level and a single contextual level.

[**varying intercept model**] An HLM with only one (non-interactive) effect from the second-level of the model.

[**within unit model**] The component of an HLM that describes variability confined to individual groups.

Hierarchical linear models (HLMs) are statistical specifications that explicitly recognize multiple levels in data. Because explanatory variables can be measured at different points of aggregation, it is often important to structure inferences that specifically identify multilevel relationships. In the classic example, student achievement can be measured at multiple levels: individually, by class, by school, by district, by state, or nationally. This is not just an issue of clarity and organization: if there exist differing effects by level, then the substantive interpretation of the coefficients will be wrong if levels are ignored. Hierarchical linear models take the standard linear model specification and remove the restriction that the estimated coefficients are constant across individual cases by specifying levels of additional effects to be estimated. This approach is also called random effects modeling because the regression coefficients are now presumed to be random quantities according to additionally specified distributions.

I. ESSENTIAL DESCRIPTION OF HIERARCHICAL LINEAR MODELS

The development hierarchical linear model starts with a simple bivariate linear regression specification for individual i:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{1}$$

which relates the outcome variable to the systematic component and the error term. The standard conditions for this model include the Gauss Markov assumptions (linear functional form, independent errors with mean zero and constant variance, and no relationship between regressor and errors), and normality of the errors (provided reasonable sample size): $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Suppose, for example, that we are interested in measuring university student evaluations of faculty teaching through the standard end-of-semester survey. Then the outcome variable, Y_i, is considered the mean score for instructor i for a given class, recorded along with the explanatory variable, X_i, indicating years of teaching experience. In this example setup the intercept, β_0, is the expected score for a new instructor. Now consider that this analysis is temporarily taking place only in department j. This means that (1) becomes:

$$Y_{ij} = \beta_{j0} + \beta_{j1} X_{ij} + \epsilon_{ij}. \tag{2}$$

There is no substantive change here; for the moment the j coefficient is just a placeholder to remind us that we are studying only the jth department so far.

Now consider broadening the analysis to evaluate student evaluation of teaching across the entire university by looking at multiple departments. Although we expect some differences, it would be rare to find that there was no underlying commonality between each of the instructors. A more realistic idea is that although each instructor has idiosyncratic characteristics, because they are teaching at the same university at the same point in time and being evaluated by the same student body, there is a common distribution from which β_0 and β_1 are drawn. Actually it would be unfair and inaccurate to calculate these means across all undergraduate majors at the university. It is well known, for instance, that sociology departments enjoy higher mean student evaluations than chemistry departments. So now add a second level to the model that explicitly nests instructors within departments and index these departments by $j = 1$ to J:

$$\begin{aligned} \beta_{j0} &= \gamma_{00} + \gamma_{10} Z_{j0} + u_{j0} \\ \beta_{j1} &= \gamma_{01} + \gamma_{11} Z_{j1} + u_{j1}, \end{aligned} \tag{3}$$

where all individual level variation is assigned to departments producing department level residuals: u_{j0} and u_{j1}. The variables at this level are called *context level* variables, and the idea of *contextual specificity* is that of the existence of legitimately comparable groups. Here the example explanatory variables Z_{j0} and Z_{j1} are the average class size for department j and the average annual research output per faculty member in department j, respectively. Note that if we were interested in performing two non-hierarchical department-level analyses, this would be straightforward using these two equations provided that the data exists.

Of course interest here is not in developing separate single-level models, and the two-level model is produced by inserting the department-level specifications, (3), into the original expression for instructor evaluations, (2). Performing this substitution and rearranging produces:

$$\begin{aligned} Y_{ij} &= (\gamma_{00} + \gamma_{10}Z_{j0} + u_{j0}) + (\gamma_{01} + \gamma_{11}Z_{j1} + u_{j1})X_{ij} + \epsilon_{ij} \\ &= \gamma_{00} + \gamma_{01}X_{ij} + \gamma_{10}Z_{j0} + \gamma_{11}X_{ij}Z_{j1} + u_{j1}X_{ij} + u_{j0} + \epsilon_{ij}. \qquad (4) \end{aligned}$$

This equation also shows that the composite error structure, $u_{j1}X_{ij} + u_{j0} + \epsilon_{ij}$, is now clearly heteroscedastic since it is conditioned on levels of the explanatory variable. Unless the corresponding variance-covariance matrix is known, and therefore incorporated as weights in the general linear model, then it must also be estimated. Ignoring this effect and calculating with OLS produces consistent estimators, but incorrect standard errors since it is equivalent to assuming zero intra-class correlation.

Often the first-level expression describing the performance of instructor i in a given department as specified by (2) is labeled the *within unit model* because its effects are confined to the single department. Conversely, the second-level expressions describing the performance of department j as a whole as specified (3) are labeled the *between unit model* which describes the variability across the departments. Looking closely at (4) reveals that there are three different implications of the effects of the coefficients for the explanatory variables:

$\boldsymbol{\gamma_{01}}$ gives the slope coefficient for one-unit the effect of teacher experience in department j.

This slope varies by department.

$\boldsymbol{\gamma_{10}}$ gives the slope coefficient for a one-unit change in department class size in department j completely independent of individual teacher effects in that department.

$\boldsymbol{\gamma_{11}}$ gives the slope coefficient for the product of individual teacher experience by department and mean annual research output by department.

Because this set of variables contains both fixed and random effects, (4) is called a *mixed model.*

The fundamental characteristic of multi-level the data discussed here is that some variables are measured at an individual level and others are measured at differing levels of aggregation. This drives the need for a model such as (4) that classify variables and coefficients by level of hierarchy they affect. Interestingly, a large proportion of HLM models in published work come from education policy studies. This is due to natural nesting of education data through the bureaucratic structure of these institutions. Other applications include studies of voting, bureaucracy, medical trials, and crime rates.

II. SPECIAL CASES OF THE HLM

There are several interesting ramifications from fixing various quantities in the basic HLM model. The most basic is produced by setting the full β_{j1} component and the γ_{10} term in (4) equal to zero. The result is the standard ANOVA model with random effects:

$$Y_{ij} = \gamma_{00} + u_{j0} + \epsilon_{ij}. \tag{5}$$

In another basic case, if the second-level defines a fixed effect rather than random effect model, $(u_{j1}, u_{j0} = 0)$, then the resulting specification is just simple linear regression model with an interaction term between the instructor level explanatory variable and the department level explanatory variable:

$$Y_{ij} = \gamma_{00} + \gamma_{01}X_{ij} + \gamma_{10}Z_{j0} + \gamma_{11}X_{ij}Z_{j1} + \epsilon_{ij}. \tag{6}$$

This is one of the most studied enhancements of the basic linear form in the social sciences.

Another very basic model comes from assuming that the second-level introduces no new error terms and there is also no interaction effect. Specifically this means that we can treat the intercept term as a composite of a constant across the sample and a constant across only the j groupings:

$$Y_{ij} = (\gamma_{00} + \gamma_{10}Z_{j0}) + \gamma_{01}X_{ij} + \epsilon_{ij}. \tag{7}$$

This is routinely called a *varying intercept model* because the parenthetical expression is now a group-specific intercept term. If we add the second assumption that there is no articulated structure within the first term, i.e. $(\gamma_{00} + \gamma_{10}Z_{j0})$ is equal to a single context-specific α_j, then this is now the *analysis of covariance model* (ANCOVA) (Kreft and de Leeuw 1998).

Sometimes it is possible to take some specific parameter in the model and fix it at a known level. Thus if substantive information at hand indicates that there is no variability to one of the γ terms, then it is appropriate to fix it in the model. It is also possible design a combination strategy such as to fix the slope coefficient $(\beta_{j1} = \gamma_{01} + \gamma_{11}Z_{j1})$ and let the intercept coefficient remain a random effect, or to fix the intercept coefficient $(\beta_{j0} = \gamma_{00} + \gamma_{10}Z_{j0})$ and let the slope remain a random effect.

Another common variation is to assume that $Z_{j0} = 0$ and $Z_{j1} = 0$, but retain the u_{j0} error term:

$$Y_{ij} = \gamma_{00} + \gamma_{01}X_{ij} + u_{j1}X_{ij} + u_{j0} + \epsilon_{ij}. \tag{8}$$

This model asserts that the j categorization is not important for determining the expected effect on Y_{ij}, but that there is an additional source of error from the categories. Hence specifying the model with only one source of error is to miss a heteroscedastic effect. A specification of this type is typically called a *random coefficients regression model.*

Another related variation is the idea that the effect of the within-unit explanatory variable (years of teaching in the running example) is uniform across departments. This is equivalent to setting $\gamma_{11}X_{ij}Z_{j1} = 0$, producing:

$$Y_{ij} = \gamma_{00} + \gamma_{01}X_{ij} + \gamma_{10}Z_{j0} + +u_{j1}X_{ij} + u_{j0} + \epsilon_{ij}, \tag{9}$$

where sometimes $u_{j1}X_{ij}$ is also set to zero. The common name for this specification is the *additive variance components model.*

III. THE GENERAL STRUCTURE OF THE HLM

The previously developed model is actually a substantial simplification in that typical models in social science research contain many more explanatory variables at both the within-unit level and the between-unit levels. It is therefore necessary to generalize hierarchical linear model to incorporate more specification flexibility. First recast (4) in matrix terms, such that the dimensional assumptions will be generalized to accommodate more useful specifications.

Define a new $\boldsymbol{\beta}_j$ vector according to:

$$\boldsymbol{\beta}_j = \begin{bmatrix} \beta_{j0} \\ \beta_{j1} \end{bmatrix} = \begin{bmatrix} 1 & Z_{j0} & 0 & 0 \\ 0 & 0 & 1 & Z_{j1} \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{01} \\ \gamma_{11} \end{bmatrix} + \begin{bmatrix} u_{j0} \\ u_{j1} \end{bmatrix} \tag{10}$$

which is just a vectorized version of the equations in (3). Therefore it is possible to express (4) in the very concise form:

$$Y_{ij} = \boldsymbol{\beta}'_j [\begin{array}{cc} 1 & X_{ij} \end{array}]' + \epsilon_{ij}. \tag{11}$$

This extra formalism is really not worth the effort for a model of this size, however the real utility is demonstrated when there are more explanatory variables at the contextual level. Define k_0 and k_1 to be the number of explanatory variables defined at the second level for β_{j0} and β_{j1} respectively. Thus far we have had the restrictions: $k_0 = 2$ and $k_1 = 2$, but we can now generalize this dimension:

$$\begin{aligned}\boldsymbol{\beta}_j &= \begin{bmatrix} \beta_{j0} \\ \beta_{j1} \end{bmatrix} \\ &= \begin{bmatrix} 1 & Z_{j01} & Z_{j02} & \dots & Z_{j0(k_0-1)} & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & Z_{j11} & Z_{j12} & \dots & Z_{j1(k_1-1)} \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ \vdots \\ \gamma_{(k_0-1)0} \\ \gamma_{01} \\ \gamma_{11} \\ \vdots \\ \gamma_{(k_1-1)1} \end{bmatrix} + \begin{bmatrix} u_{j0} \\ u_{j1} \end{bmatrix}.\end{aligned} \tag{12}$$

The dimension of the matrix of Z variables is now $(2 \times k_0 + k_1)$ and the length of the $\boldsymbol{\gamma}$ vector is $k_0 + k_1$, for any specified values of k_i which are allowed to differ. It is common, and computationally convenient to assume that the error vector, $\mathbf{u}_j$, is multivariate normal distributed around zero with a given or estimated variance-covariance matrix. Note that the row specifications in the $\mathbf{Z}$ matrix always begin with a one for the constant which specifies a level-one constant in the first row and a level-one restricted explanatory variable in the second row.

It is important to observe that since the constant in this model is part of the specification, the indices run to $k_0 - 1$ and $k_1 - 1$ to obtain dimensions k_0 and k_1. Also when there was only one Z variable specified in the second level of the model it was sufficient to subscript simply by j and either 0 or 1 (for the first or second equation of (3). However, now that there are an arbitrary number for each second-level equation they must be further indexed by a third value: 1 to $k_0 - 1$ or 1 to $k_1 - 1$ here. Note also that each group is no longer required to contain the same mix of second-level explanatory variables. This turns out to be useful in specifying many different varying model specifications.

It is possible that there are also more first-level variables in the model (likely in fact). To accommodate this it is required to further generalize the defined matrix structures. Define the $\mathbf{Z}_\ell$ vector as:

$$\mathbf{Z}_\ell = \left[\begin{array}{ccccc} 1 & Z_{j\ell 1} & Z_{j\ell 2} & \dots & Z_{j\ell(k_\ell - 1)} \end{array} \right] \tag{13}$$

for $\ell = 1$ to L coefficients in the first level model, including the constant. Therefore the $\mathbf{Z}$ matrix is now a $(L \times L)$ diagonal matrix according to:

$$\mathbf{Z} = \left[\begin{array}{ccccc} \mathbf{Z}_1 & 0 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & 0 & \dots & 0 \\ 0 & 0 & \mathbf{Z}_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{Z}_L \end{array} \right]. \tag{14}$$

where each diagonal value is a $\mathbf{Z}_\ell$ vector. This can also be fully written out as an irregular

ℓ-diagonalized matrix, but it would be more cumbersome than the given form. Given the new form of the $\mathbf{Z}$ matrix it is necessary to respecify the $\boldsymbol{\gamma}$ vector as follows:

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_0 \\ \boldsymbol{\gamma}_1 \\ \vdots \\ \boldsymbol{\gamma}_L \end{bmatrix} \tag{15}$$

where each $\boldsymbol{\gamma}_\ell$ is a column vector whose length is determined by the k_ℓ dimension specification. Putting these new structures together gives:

$$\boldsymbol{\beta}_j = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}, \tag{16}$$

and:

$$Y_{ij} = \boldsymbol{\beta}_j' \begin{bmatrix} 1 & X_{ij1} & X_{ij2} & \dots & X_{ijL} \end{bmatrix}' + \epsilon_{ij}. \tag{17}$$

Thus the HLM in this form allows any number of first and second level explanatory variables as well as differing combinations across contextual levels. Note also that there is no restriction for number of individual units, n_j, be equal across the contexts (although this can make the estimation process more involved).

The final basic way that the HLM can be made more general is to add further levels of hierarchy with respect to levels. That is, it is possible to specify a third-level in exactly the way that the second level was added by parameterizing the $\boldsymbol{\gamma}$ terms according to:

$$\gamma_{pq} = \delta_{0q} + \delta_{1q} W_{pq} + v_{pq},$$

where the p subscript indicates a second level of contexts ($p = 1, \dots, P$), and the q subscript indexes the number of equations ($q = 1, \dots, Q$) specified at this level (analogous to k at the lower level). In this specification W_{pq} is a third-level measured explanatory variable and v_{pq} is the level-associated error term. Obviously the specifications can be made very complex at this level.

The other principle approach to modeling a third level is to specify a Bayesian prior distribution for the $\boldsymbol{\gamma}$ coefficients. These priors are typically assigned normal distributions, but this is not a restriction and many others have been used. As a consequence of linearity, the normal property then ripples down the hierarchy, making estimation relatively easy. This then model specifies hierarchies of linear hyperpriors each of which has its own prior plus an associated matrix of explanatory variables, and only nodes at the highest level of the hierarchy have fixed hyperprior values.

IV. ESTIMATION OF THE HLM

No matter how complex the right-hand-side of the HLM equation becomes, the left-hand-side always consists of Y_{ij} which is assumed to be normal with mean equal to the systematic component of the model and variance from the collected error terms. If was known that the error structure in the model was uncorrelated to explanatory variables, then it would easy to estimate the coefficients with standard maximum likelihood or least squares approaches. Actually we know that in general the form of the errors *is* conditional on levels of the explanatory variables since in (4) there is the term: $u_{j1}X_{ij}$. In addition, there are increasing numbers of dependencies as the model becomes progressively more complex and realistic.

If we knew for certain the form of the relationship between regressors and errors, then it could be expressed through a weighting matrix and general least squares would provide consistent estimates of the coefficients and their standard errors. Unfortunately this information is rarely available. The classic alternative is to specify a likelihood function and employ maximum likelihood estimation of the full set of unknown parameters, including variances using Fisher scoring. This is often a cumbersome process so many software implementations work with the profile likelihood: first estimating the higher order variance terms only then fixing them in the likelihood function equation for the lower level parameters. This tends to underestimate the magnitude of the higher order variance terms since uncertainty is ignored in the first step, leading to overconfident model results. An improved process is employing restricted maximum likelihood (REML) by integrating out the fixed effects terms in the calculation of the profile likelihood, and after obtaining the lower level parameter estimates, recalculating the higher order variance terms conditional on these. However, the best method is the quasi-Bayesian procedure, Empirical Bayes/Maximum Likelihood (EB/ML). A fundamental principle of Bayesianism is that unknown parameters are treated as random variables possessing their own distributions to be estimated as a consequence of applying Bayes Law. By analogy we can consider the unknown HLM estimates as having their own distributions conditioned on unknown quantities from the higher level of the model. Rather than stipulating explicit priors for the parameters, as a Bayesian would do, it is possible to use a prior suggested by the data: empirical Bayes.

The EM algorithm is essential to this estimation process and therefore warrants some description. Expectation-Maximization (EM) is a flexible and often-used method for incomplete data problems: i.e. to “fill-in” missing information given a specified model. The notion of what is “missing” is general here: it can be unknown parameters, missing data, or both. There are two basic steps. First, temporary data that represent a reasonable guess are assigned to the missing data (expectation). Second, proceed with maximum likelihood estimation of the parameters as if there now exists a complete-data problem (maximization). The algorithm is iterative in the sense that it is now possible to use these parameter estimates to update the assignment of the temporary data values with better guesses, and repeat the process. It can be shown that the EM algorithm gives a series of parameter estimates that are monotonically increasing on the likelihood metric and are guaranteed to converge to a unique maximum point under very general and non-restrictive regularity conditions. The utility here is that the HLM model with linear specifications and normal assumptions is a particularly well-behaved application of EM.

Detailed summaries of the EB/ML computational procedure for obtaining coefficient estimates and measures of reliability are found in Chapter 10 of Bryk and Raudenbush, the Appendix of Wong and Mason, and from an exclusively Bayesian perspective in Lindley and Smith, along with Smith. The basic strategy is to obtain estimates of the variance terms using the EM algorithm and the joint likelihood function for the coefficients and the variances, plug these estimates into the top hierarchy of the model, perform maximum likelihood calculations as if these were the correct weightings, update the estimate of the coefficients by using the mean of the subsequent posterior. This is a very general description of the procedure and there are many nuances which depend on the particular form of the model and configuration of the data.

V. CRITICAL ADVANTAGES OF THE HLM

Hierarchical linear models are a compromise between two opposite approaches to clustering. On one side it is possible to simply pool all of the observations an estimate coefficients of interest as if the between group effects do not matter. Conversely, it is also possible to aggregate the data by the groups and calculate the coefficient estimates on these aggregations as if they are the primary object of interest. The first approach ignores between-group variation and the second approach ignores within-group variation. It is possible that either of these approaches is entirely appropriate and reliable inferences can be obtained. Of course if there actually are important differences by groupings, then neither will be correct. Hierarchical linear models provide a method for producing models that explicitly recognize this distinction by incorporating the nesting of the data into the model specification.

Hierarchical linear models also have several specific methodological advantages over standard linear models:

- Hierarchical models are ideal tools for identifying and measuring structural relationships that fall at different levels of the data generating procedure.
- There is virtually no limit to the dimension of the hierarchy.
- Hierarchical models directly express exchangeability of units.
- Non-hierarchical models applied to multi-level data typically underestimate variance.
- Hierarchical models facilitate the testing of hypotheses across different levels of analysis.
- Non-hierarchical models are nested within hierarchical models allowing a likelihood or Bayes factor test of the validity of the proposed hierarchical structure.

While these reasons are compelling, it is only relatively recently that hierarchical models have been actively pursued in the social sciences. This is parallel (and related) to the attachment social scientists have for the linear model in general. What precipitated the change was dramatic improvements in statistical computing that provided solutions to previously intractable problems. These stochastic simulation tools include the EM algorithm as well as Markov Chain Monte Carlo techniques (MCMC) such as the Metropolis-Hastings algorithm and the Gibbs sampler whereby an iterative chain of consecutive computationally generated values is setup carefully enough and run long enough to produce *empirical* estimates of integral quantities of interest from later chain values. Although these approaches are typically associated with Bayesian modeling such iteration techniques are not *limited* to Bayesian or even hierarchical applications. They do, however, greatly help naturally occurring computational problems in these settings.

BIBLIOGRAPHY

Goldstein, Harvey. (1995). *Multilevel Statistical Models.* Edward Arnold, New York.

Heck, Ronald H., Thomas, Scott Loring Thomas. (2000). *Introduction to Multilevel Modeling Techniques.* Lawrence Erlbaum Associates, Mahwah, NJ.

Kreft, Ita, and de Leeuw, Jan. (1998). *Introducing Multilevel Modeling.* Sage Publications, Newbury Park, CA.

Leyland, A. H., and Goldtsein, H. (2001). *Multilevel Modelling of Health Statistics.* John Wiley & Sons, New York.

Lindley, D. V., and Smith, A. F. M. (1972). "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society, Series B* **34,** 1- 41.

Nelder, J. A. (1977). A Reformulation of Linear Models (with discussion). *Journal of the Royal Statistical Society, Series A* **140,** 48-76.

Raudenbush, Stephen, and Bryk, Anthony S. (1986). A Hierarchical Model for Studying School Effects. *Sociology of Education* **59,** 1-17.

Raudenbush, Stephen, and Bryk, Anthony S. (2002). *Hierarchical Linear Models.* Second Edition. Sage Publications, Newbury Park, CA.

Reise, Steven P., and Duan, Naihua. (2001). *Multilevel Models: A Special Issue of Multivariate Behavioral Research.* Lawrence Erlbaum Associates, Mahwah, NJ.

Smith, A. F. M. (1973). A General Bayesian Linear Model. *Journal of the Royal Statistical Society, Series B* **35**, 61-75.

Wong, George Y., and Mason, William M. (1991). Contextually Specific Effects and Other Generalizations of the Hierarchical Linear Model for Comparative Analysis. *Journal of the American Statistical Association* **86,** 487-503.