The Multilevel Model Framework

Jeff Gill Andrew Womack
Washington University, St. Louis

1 Overview

Multilevel models account for different levels of aggregation that may be present in data. Sometimes researchers are confronted with data that are collected at different levels such that attributes about individual cases are provided as well as the attributes of groupings of these individual cases. In addition, these groupings can also have higher groupings with associated data characteristics. This *hierarchical structure* is common in social science data and is commonly ignored by social science researchers. Unfortunately, neglecting hierarchies in data can have damaging consequences to subsequent statistical inferences.

The frequency of nested data structures in the data-analytic sciences is startling. In the United States and elsewhere, individual voters are nested in precincts which are, in turn, nested in districts, which are nested in states, which are nested in the nation. In healthcare, patients are nested in wards, which are then nested in clinics or hospitals, which are then nested in healthcare management systems, which are nested in states, and so on. In the classic example, students are nested in classrooms, which are nested in schools, which are nested in districts, which are then nested in states, which again are nested in the nation. In another familiar context, it is often the case that survey respondents are nested in areas such as rural versus urban, then these areas are nested by nation, and the nations in regions. Famous studies such as the American National Election, Latinobarometer, Eurobarameter, and Afrobarometer, are obvious cases. Furthermore, the frequency of data at different levels of aggregation is increasing as more data are generated from: geocoding, biometric monitoring, Internet traffic, social networks, an amplification of government and corporate reporting, and high-resolution imaging.

Multilevel models are a powerful and flexible extension to conventional regression frameworks. They extend the linear model and the generalized linear model by incorporating levels directly into the model statement, thus accounting for aggregation present in the data. Therefore all of the familiar model forms for linear, dichotomous, count, restricted range, ordered categorical, and unordered categorical outcomes are supplemented by adding a structural component. This structure classifies cases into known groups, which may have their own set of explanatory variables at the group level. So a hierarchy is established such that some explanatory variables are assigned to explain differences at the individual level and some explanatory variables are assigned to explain differences at the group level. This is powerful because it takes into account correlations between subjects within the same group as distinct from correlations between groups. Thus with nested data structures the multilevel approach immediately provides a set of critical advantages over conventional, flat modeling where these structures emerge as unaccounted-for heterogeneity and correlation.

What does a multilevel model look like? At the core, we have a regression equation that relates an outcome variable on the left-hand-side to a set of explanatory variables on the right-hand-side. This is the basic individual-level specification, and looks immediately like a linear model or generalized linear model in the case of nonlinear outcomes. The departure comes from the treatment of some of the coefficients assigned to the explanatory variables. Suppose that we felt strongly that there was extra heterogeneity of some type introduced by a measured variable that fixed-point estimates would not adequately describe. It might then make sense to introduce variation in the treatment of that variable at this individual-level by allowing it to be described *distributionally* rather than as a fixed point to be estimated in the context of the bigger model. An obvious way to do this is to treat this coefficient as an outcome variable and model it with an additional regression equation. In this manner we can introduce another set of covariates that explain variation in this effect. In this way we remove the restriction that the estimated coefficients are constant across individual cases by specifying levels of additional effects. So any right-hand-side effect can get its own regression expression with its own assumptions about functional form, linearity, independence, variance, distribution of errors, and so on. Such models are usually "mixed," meaning some of these coefficients are *modeled* with this new level and some are *unmodeled* and therefore estimated in the conventional manner as point effects.

What this strategy produces is a method of accounting for structured data through utilizing regression equations at different hierarchical levels in the data. The key linkage is that these higher level models are describing *distributions* at the level just beneath them for the coefficient that they model as if it were itself an outcome variable. This means that multilevel models are highly symbiotic with Bayesian specifications because the focus in both cases is on making supportable distributional assumptions.

Allowing multiple levels in the same model actually provides an immense amount flexibility. First, the researcher is not restricted to a particular number of levels. The coefficients at the second grouping level can also be assigned a regression equation, thus adding another level to the hierarchy. Although it has been shown that there is diminishing return as the number of levels goes up and it is rarely efficient to go past three levels from the individual-level. Second, as stated, any coefficient at these levels can be chosen to be modeled or unmodeled and in this way the mixture of these decisions at any level gives a combinatorically large set of choices. Third, the form of the link function can differ for any level of the model. In this way the researcher may mix linear, logit/probit, count, constrained, and other forms throughout the total specification.

2 Background

It is often the case that fundamental ideas in statistics hide for a while in some applied area before scholars realize that these are generalizeable and broadly applicable principles. For instance, the well-known EM algorithm of Dempster, Laird and Rubin (1977) was predated in less fully articulated forms by Newcomb (1886), McKendrick (1926), Healy and Westmacott (1956), Hartley (1958), Baum and Petrie (1966), Baum and Eagon (1967), and Zangwill (1969) who gives the critical conditions for monotonic convergence. In another famous example, the core Markov chain Monte Carlo (MCMC) algorithm (Metropolis, *et al.* 1953) slept quietly in the *Journal of Chemical Physics* before emerging in the 1990s to revolutionize the entire discipline. It turns out that hierarchical modeling follows this same story-line, roughly originating with the statistical analysis of agricultural data around the 1950s (Eisenhart 1947, Henderson 1950, Scheffé 1956, Henderson *et al.* 1959). A big step forward came in the 1980s when education researchers realized that their data fit this structure perfectly (students nested in classes, classes nested in schools, schools nested in districts, districts nested in states), and that important explanatory variables could be found at all of these levels. This flurry of work focused on the *hierarchical linear model* (HLM) and was developed in detail in works such as: Burstein (1980), Mason *et al.* (1983), Aitkin and Longford (1986), Bryk and Raudenbush (1987), Bryk *et al.* (1988), De Leeuw and Kreft (1986), Raudenbush and Bryk (1986), Goldstein (1987), Longford (1987), Raudenbush (1988), and Lee and Bryk (1989). These applications continue today as education policy remains an important empirical challenge Work in this literature was accelerated by the development of the stand-alone software packages HLM, ML2, VARCL, as well as incorporation into the SAS procedure Genmod. Additional work by Goldstein (notably 1995) took the two level model and extended it to situations with further nested groupings, non-nested groupings, time series cross-sectional data, and more.

Beginning in approximately in the 1990s hierarchical modeling took-on a much more Bayesian complexion now that stochastic simulation tools (eg. MCMC) had arrived to solve the resulting estimation challenges. Since the Bayesian paradigm and the hierarchical reliance on distributional relationships between levels have a natural affinity, many papers were produced and continue to be produced in the intersection of the two. Computational advances during this period centered around customizing MCMC solutions for particular problems (Carlin *et al.* 1992, Albert and Chib 1993, Hobert and Casella 1996, Liu 1994, Jones and Hobert 2004, Cowles 2002). Other works focused on solving specific applied problems with Bayesian models, for instance: Steffey (1992) incorporates expert information into the model, Stangl (1995) develops prediction and decision rules, Cohen *et al.* (1998) model arrest rates, Christiansen and Morris (1997) build on count models hierarchically, Hodges and Sargent (2001) refine inference procedures, Skates *et al.* model cancer risk, and Pettitt *et al.* (2006) model survey data from immigrants. Recently Bayesian prior specifications in hierarchical models have received attention (Hadjicostas, P. and Berry 1999, Daniels and Gatsonis 1999, Gelman 2006, Booth *et al.* 2008). Finally, the text by Gelman and Hill (2007) has been enormously influential.

A primary reason for the large increase in interest in the use of multilevel models in recent years is due to the ready availability of sophisticated general software solutions for estimating more complex specifications. For basic models the lme4 package in R works quite well and preserves R's intuitive model language, and Stata provides some support through the XTMIXED routine. However, researchers now routinely specify generalized linear multilevel models with categorical, count, or truncated outcome variables. It is also now common to see non-nested hierarchies expressing cross-classification, mixtures of nonlinear relationships within hierarchical groupings, and longitudinal considerations such as panel specifications and standard time-series relations. All of this provides a rich, but sometimes complicated, set of variable relationships. Since most applied users are unwilling to spend the time to derive their own likelihood functions or posterior distributions and maximize or explore these forms, software like WinBUGS and its cousin JAGS are popular (Bayesian) solutions (Mplus is also a helpful choice).

3 Foundational Models

The development of multilevel models starts with the simple linear model specification for individual i that relates the outcome variable, Y_i, to the systematic component, $X_i\beta_1$, with unexplained variance falling to the error term, ϵ_i, giving:

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

which is assumed to meet the standard Gauss Markov assumptions (linear functional form, independent errors with mean zero and constant variance, no relationship between X_i and errors). The normality of the errors is not a necessary assumption but comes with reasonable sample size: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

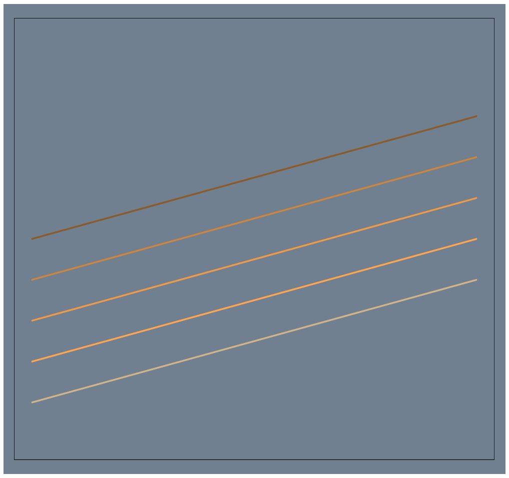
3.1 Basic Linear Forms, Individual-Level Explanatory Variables

Suppose now that we believed that there was heterogeneity because each i case belongs to one of $j = 1, \dots, J$ groups where $J < n$. Even if we do not have explanatory variable information about these j assignments, model fit may be improved by binning each i case into its respective group. This can be done by loosening the definition of the single intercept, β_0, in (1) to j distinct intercepts, β_{0j}, which then groups the n cases giving them a common intercept with other cases if they land in the same group. Formally for $i = 1, \dots, n$:

$$y_i = \beta_{0j} + X_i\beta_1 + \epsilon_i, \tag{2}$$

where the added j subscript indicates that the ith case gets intercept j for the jth group. The β_{0j} are given a common normal distribution with mean β_0 and standard deviation σ_{u_0}. Since the β_1 coefficient is not indexed by the grouping term j, we know that this is still constant across the n cases and evaluated with a standard point estimate. This is illustrated at right and shows that while different groups start at different intercepts, they progress at the same rate (slope). This model is sufficiently fundamental that it has its own name, the *varying-intercept* or *random-intercepts* model.

Figure 1: Varying Intercepts

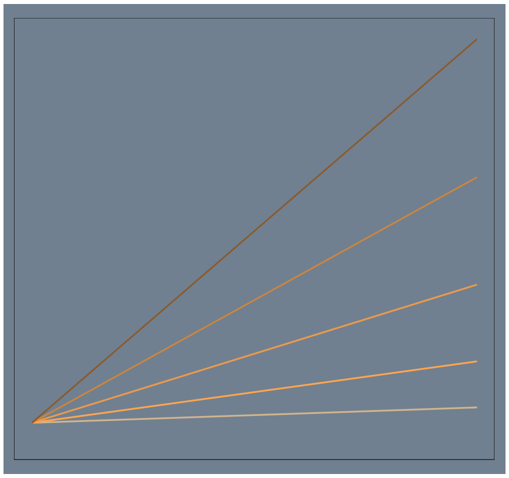


In a different context, we may want to bin the i cases into j groups, but we feel that the effect is not through the intercept where they groups start at a zero level of the explanatory variable X. Now we have reason to believe that the grouping affects the slopes instead: as X increases group membership dictates a different change in Y. So now we loosen the definition of the single slope, β_1, in (1) to allow binning of the i cases by j groups according to:

$$y_i = \beta_0 + X_i\beta_{1j} + \epsilon_i, \tag{3}$$

where the added j subscript indicates that the ith case gets slope j for the jth group. The intercept now remains fixed across the cases in the data and the slopes are given a common normal distribution. This is illustrated in the figure at right showing divergence from the same starting point for the groups as X increases. This model is also fundamental enough that it gets its own name, the *varying-slope* or *random-slope* model.

Figure 2: Varying Slopes

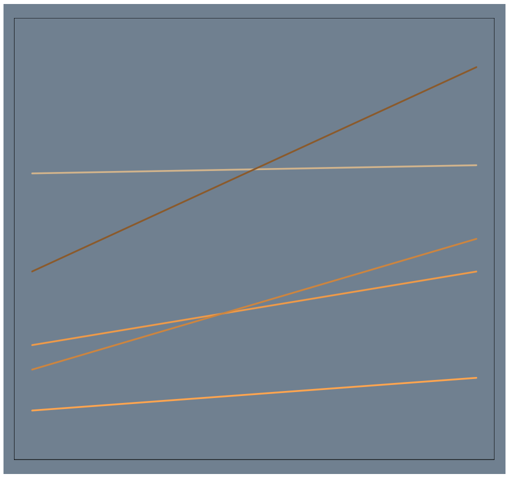


Suppose the researcher suspects that the heterogeneity in the sample across the i cases is sufficiently complex that it needs to be modeled with both a varying-intercept and a varying-slope. This is a simple combination of the previous two models and takes the form:

$$y_i = \beta_{0j} + X_i\beta_{1j} + \epsilon_{ij},$$

where membership in group j for case i has two effects, one that is constant and one that differs from others

Figure 3: Varying Intercepts and Slopes



with increasing X. The vectors $(\boldsymbol{\beta}_{0J}, \boldsymbol{\beta}_{1J})$ are given a common multivariate normal distribution. A synthetic, possibly exaggerated, model result is given at right. Not surprisingly, this is called the *varying-intercept, varying-slope* or *random-intercept, random-slope* model. Notice from the simple artificial example in the figure that we can already model quite intricate differences in the groups for this basic linear form.

So far we have not worried about additional explanatory variables. Obviously we can extend the bivariate linear form additively on the right-hand-side to include more covariates, which may or may not receive the grouping treatment just described. A canonical *mixed* form is one where the intercept and the first $q-1$ explanatory variables have coefficients that vary by the $j = 1, \ldots, J$ grouping as well as the varying-intercept (for a total of q modeled coefficients), but the next p coefficients, $q, q+1, \ldots, k$ coefficients are fixed at the individual level. This is given by the specification:

$$y_i = \beta_{0j} + X_{1i}\beta_{1j} + \ldots + X_{(q-1)i}\beta_{(q-1)j} + X_{qi}\beta_q + \ldots + X_{ki}\beta_k + \epsilon_i, \tag{4}$$

where membership in group j for case i has q effects. This is already getting quite cumbersome from a notational perspective and, like other instances, it is stated more cleanly in matrix terms. First collect the outcome variable by groups such that y_{ij} is now the observed outcome of case i, $i = 1, \ldots n_j$, belonging to group j, $j = 1, \ldots, J$, and these are ordered in the n-length vector:

$$\mathbf{y} = (y_{11}, y_{21}, \ldots, y_{n_1 1}, y_{12}, y_{22}, \ldots, y_{n_2 2}, \ldots y_{1J}, y_{2J}, \ldots, y_{n_J J}), \tag{5}$$

where the first index gives the case number within the group and the second index gives the group number. The same indexing and ordering can be applied to $\mathbf{X}$, so that the ℓth column of $\mathbf{X}$ is the n-length vector:

$$\mathbf{X}_\ell = (X_{\ell 11}, X_{\ell 21}, \ldots, X_{\ell n_1 1}, X_{\ell 12}, X_{\ell 22}, \ldots, X_{\ell n_2 2}, \ldots X_{\ell 1J}, X_{\ell 2J}, \ldots, X_{\ell n_J J}). \tag{6}$$

Allowing the n_j to vary is necessary since we do not want to restrict our data analysis to the perfectly balanced case where all groups have exactly the same number of cases. Since most of the social sciences deals with observational rather than experimental data, this is an important feature. It is also a result of the subscript notation in (5) that $n = \sum_{j=1}^{J} n_j$. So in this way our nesting of cases into groups is explicitly expressed in the notation.

Now consider $\mathbf{X}^*$, the full $n \times k$ positive definite matrix of explanatory variables organized down columns with a leading vector of ones. For the ith case we will denote that row as $\mathbf{x}_{ij}$ where the j index is just a reminder of nesting. The notational problem with (4) is that it makes it difficult to look at $\mathbf{X}^*$ and determine which column variables are modeled with by grouping and which column variables are left at the individual level. So lets partition the $n \times k$ matrix $\mathbf{X}^*$ into two submatrices that will allow us to keep track of this distinction:

$$\underset{n \times k}{\mathbf{X}^*} = \left[\underset{n \times p}{\mathbf{X}} \middle| \underset{n \times q}{\mathbf{Z}} \right], \tag{7}$$

where $\mathbf{X}$ is a matrix of explanatory variables that will have *unmodeled* coefficients, therefore receiving standard point estimates, and $\mathbf{Z}$ is a matrix of explanatory variables that will have *modeled* coefficients and will therefore by described distributionally. Note that $p+q=k$. If we introduce the vector of residuals corresponding to (5), $\boldsymbol{\epsilon}$ containing ϵ_{ij} which are all independent and identically distributed according to $\epsilon_{ij} \sim N_n(0, \sigma)$, then the general model in matrix notation is expressed as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{\beta}$ is a p-length vector of coefficient estimates and $\mathbf{u}$ is a q-length vector of *random effects* with the common assumption of multivariate normality with mean zero and positive definite variance/covariance matrix, such that $\mathbf{u} \sim N_q(\mathbf{0}, \boldsymbol{\Omega})$. So our interest lies in the magnitude and reliability of $\boldsymbol{\beta}$ as well as the magnitude of $\boldsymbol{\Omega}$ relative to $\boldsymbol{\Sigma}$. As we shall see shortly, the larger the variance terms are on the diagonal of $\boldsymbol{\Omega}$ relative to σ, the more variance is being explained by imposed grouping. This result would indicate that our grouping scheme improves the fit of the model since it is taking variation that would normally fall to the residual term and modeling it directly with $\mathbf{Zu}$. We have not yet built an explanatory specification at the q group levels, so for the moment $\mathbf{Z}$ is a vector of ones. This has the effect of adding q group contribution to i's right-hand-side where even though the means are zero, the individual group additions are non-zero. In that way we keep track here of group difference *but not yet why the groups are different.*

We can now write the model in terms of regression equations for each group:

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j, \tag{9}$$

where $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{(q-1)j}, \beta_q, \ldots, \beta_k)$, $\mathbf{y}_j$ is the vector of outcomes for group j, $\mathbf{X}_j$ is the matrix of covariates for group j, and $\boldsymbol{\epsilon}_j$ is the vector of errors for group j. The notational problem with either of these formulations is that it makes it difficult to look at $\mathbf{X}$ and determine which column variables are modeled by grouping and which column variables are left at the individual level. In order to facilitate this, we can look at the mean structure and the error terms in an additive manner. In particular, we can define $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, $\mathbf{u}_j = (\beta_{0j} - \beta_0, \ldots, \beta_{(q-1)j} - \beta_{(q-1)})$, and $\mathbf{Z}_j$ to be a matrix formed by the first q columns of $\mathbf{X}_j$. The model for an individual group can now be expressed as

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\epsilon}_j, \tag{10}$$

where the mean structure is captured by the $\mathbf{X}_j\boldsymbol{\beta}$ term and the error terms are captured by the $\mathbf{Z}_j\mathbf{u}_j + \boldsymbol{\epsilon}_j$. Defining $\mathbf{Z}$ to be a block diagonal matrix with blocks given by the $\mathbf{Z}_j$, the vector $\mathbf{u}$ by stacking the $\mathbf{u}_j$, and the vector $\boldsymbol{\epsilon}$ by stacking the $\boldsymbol{\epsilon}_j$, the general model in matrix notation is expressed as a Laird-Ware model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \tag{11}$$

where $\boldsymbol{\beta}$ is a p-length vector of coefficients and $\mathbf{u}$ is the vector of *random effects* with the assumption of common multivariate normality on the $\mathbf{u}_j$ with mean zero and positive definite variance/covariance matrix, that is $\mathbf{u}_j \sim N_q(\mathbf{0}, \boldsymbol{\Omega})$. So our interest lies in the magnitude and reliability of $\boldsymbol{\beta}$ as well as the magnitude of $\boldsymbol{\Omega}$ relative to $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_n$. As we shall see shortly, the larger the variance terms are on the diagonal of $\boldsymbol{\Omega}$ relative to σ^2, the more variance is being explained by the imposed grouping. This result would indicate that our grouping scheme improves the fit of the model since it is taking variation that would normally fall to the residual term and modeling it directly with $\mathbf{Zu}$. Even with this richer structure, we have yet to incorporate variables that explain the differences between the groups.

3.2 Basic Linear Forms, Adding Group-Level Explanatory Variables

The model given in (11) does not impose explanatory variables at the second level, given by the J groupings, since $\mathbf{u}_j \sim N_q(\mathbf{0}, \boldsymbol{\Omega})$ by assumption. Until we add explanatory variables at this second level, we have not modeled the reason for these differences. Returning to the vary-intercepts, varying-slopes linear model in (3.1), $y_i = \beta_{0j} + X_{ij}\beta_{1j} + \epsilon_{ij}$, model each of the two coefficients with their own regression equation and index these by $j = 1$ to J:

$$\begin{aligned}\beta_{0j} &= \beta_0 + \gamma_{01}X_{j,01} + u_{0j}\\ \beta_{1j} &= \beta_1 + \gamma_{11}X_{j,11} + u_{1j},\end{aligned} \tag{12}$$

where all individual level variation is assigned to groups producing group-level residuals: u_{j0} and u_{j1}. The explanatory variables at the second levels, are called *context level* variables, and the idea of *contextual specificity* is that of the existence of legitimately comparable groups. These context level variables are constant in each group and the additional subscript 01 or 11 denotes the particular coefficient with which they pair.

Substituting the two definitions in (12) into the individual-level model in (3.1) and rearranging produces:

$$\begin{aligned}y_{ij} &= (\beta_0 + \gamma_{01}X_{j,01} + u_{0j}) + (\beta_1 + \gamma_{11}X_{j,11} + u_{1j})X_{ij} + \epsilon_{ij}\\ &= \beta_0 + \beta_1 X_{ij} + \gamma_{01}X_{j,01} + \gamma_{11}X_{j,11}X_{ij} + (u_{1j}X_{ij} + u_{0j} + \epsilon_{ij}),\end{aligned} \tag{13}$$

for the ijth case. The composite fixed effects now have a richer structure. In addition to varying the intercepts and slopes, the variation is now modeled as being due to specific groups level variables. This result also shows that the composite error structure, $(u_{j1}X_{ij} + u_{j0} + \epsilon_{ij})$, is heteroscedastic since it is conditioned on levels of the explanatory variable. This composite error shows that we have increased uncertainty in the multilevel model since we are imposing new structure on the data. We see from the form in (13) that multilevel models can expressed in a single-level expression, although this does not always lead to a more intuitive expression. It is also important to understand the exact role of the new coefficients. First, β_0 is a universally assigned intercept that all i cases share. Second, gives another shared term that is the slope coefficient corresponding to the effect of changes in X_{ij}. These two terms have no effect from the multilevel composition of the model. Third, γ_{01} gives the slope coefficient for one-unit the effect of the variable *for group* j, and applied to all individual cases assigned to this group. It therefore varies by group and not individual. Fourth, and surprisingly, γ_{11} is the coefficient on the interaction term between X_{ij} and But wait, we did not *ask* for an interaction term in the model specification. This illustrates an important point. Any hierarchy that models a slope on the right-hand-side imposes an interaction term if this hierarchy contains group-level covariates. While it is easy to see the multiplicative implications from (13), it is surprising to some that this is an automatic consequence. This deeper structure can also be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$. The matrix $\mathbf{X}$ is expanded to include the group level variables and interaction terms and the vector $\boldsymbol{\beta}$ is expanded to include the appropriate coefficients, in this case γ_{01} and γ_{11}.

3.3 The Model Spectrum

In language that Gelman and Hill (2007) emphasize, multilevel models can be thought of as sitting between two extremes that are available to the researcher when groupings are known: *fully-pooled* and *fully-unpooled*. The fully-pooled model treats the group-level variables as individual variables, meaning that group-level distinctions are ignored and these effects are treated as if they are case-specific. For a model with one explanatory variable measured at the individual-level (X_1) and one measured at the group-level (X_2) , this specification is:

$$Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + \epsilon_i. \tag{14}$$

In contrast to (13), there is no u_{0j} given explicitly here. This is an assertion that the group distinctions do not matter and the cases should all be treated homogeneously, ignoring the (possibly important) variation between categories. At the other end of the spectrum is a set of models in which we treat each group as a separate dataset and model them completely separately:

$$Y_{ij} = \beta_{0j} + X_{ij}\beta_{1j} + \epsilon_{ij}, \tag{15}$$

for $j = 1, \ldots, J$. Note that the group level predictor X_2 does not enter into this equation because $X_{2i}\beta_2$ is constant within a group and therefore subsumed into the intercept term. Here there is no second level to the hierarchy and the βs are assumed to be fixed parameters, in contrast to the distributional assumptions made in the mixed-model. The fully-unpooled approach is the opposite distinction from the fully-pooled approach and asserts that the groups are so completely different that it does not make sense to associate them in the same model. In particular, the values of slopes and intercept from one group have no relationship to those in other groups. Such separate regression models clearly overstate variation between groups, making them look more different than they really should be.

Between these two polar group distinctions is the multilevel model. The word "between" here means that groups are recognized as different but because there is a single model in they are associated by common individual-level fixed effects as well as distributional assumptions on the random effects. The resulting model therefore compromises between full distinction of groups and the full ignoring of groups. This can be thought of as partial-pooling or semi-pooling in the sense that the groups are collected together in a single model, but their distinctness is preserved.

To illustrate this "betweeness" consider a simple varying-intercepts model with no explanatory variables:

$$y_{ij} = \beta_{0j} + \epsilon_{ij}, \tag{16}$$

which is also called a *mean model* since β_{0j} represents the mean of the jth group. If we assume $\beta_{0j} = \beta_0$ is constant across all cases, then this becomes the fully-pooled model. Conversely, if we create J separate models each with their own β_{0j} which do not derive from a common distribution, then we have the fully-unpooled approach. Estimating (16), as stated, gives group means that are a weighted average of the n_j cases in group j and the overall mean from all cases. Define first:

$\overline{y}_j$	fully-unpooled mean for group j
$\overline{y}$	fully-pooled mean
σ_0^2	within-group variance (assumed equal across groups for now)
σ_1^2	variance among the $\overline{y}_j$ mean estimates
n_j	size of the jth group.

Then an approximation of the multilevel model *estimate* for the group mean is given by:

$$\hat{\beta}_{0j} = \frac{\frac{n_j}{\sigma_0^2}\overline{y}_j + \frac{1}{\sigma_1^2}\overline{y}}{\frac{n_j}{\sigma_0^2} + \frac{1}{\sigma_1^2}}. \tag{17}$$

This is a very revealing expression. The posterior mean for a group is a weighted average of the contribution from the full sample and the contribution from that group, where the weighting depends on relative variances and the size of the group. As the size of arbitrary group j gets small, $\overline{y}_j$ becomes less important and the group estimate borrows more strength from the full sample. A zero size group, perhaps a hypothesized case, relies completely on the full sample size, since (17) reduces to $\hat{\beta}_{0j} = \overline{y}$. On the other hand, as group j gets large, its estimate dominates the contribution from the fully-pooled mean, and is also a big influence on this fully-pooled mean. This is called the *shrinkage* of the mean effects towards the common mean. In addition, as $\sigma_1^2 \to 0$, then $\hat{\beta}_{0j} \to \overline{\mathbf{y}}$, and as $\sigma_1^2 \to \infty$, then $\hat{\beta}_{0j} \to \overline{\mathbf{y}}_j$. Thus the group effect which is at the heart of a multilevel model is a balance between the size of the group relative to the full sample and the standard deviations at the individual and group levels.

4 Extensions Beyond the Two-Level Model

Multilevel models are not restricted to linear forms with interval-measured outcomes over the entire real line, nor are they restricted to hierarchies which contain only one level of grouping or nested levels of grouping. The stochastic assumptions at each level of the hierarchy can be made in any appropriate fashion for the problem being modeled. This added flexibility of the MLM provides a much richer class of models and captures many of the models used in modern social science research.

4.1 Nested Groupings

The generalization of the mixed effects model to nested groupings is straightforward and is most easily understood in the hierarchical framework. Consider the common case of survey respondents nested in regions, which are then nested in states and so on. The individual level comprises the first hierarchy of the model and captures the variation in the data that can be explained by individual level covariates. In this example, the outcome of interest is measured support for a political candidate or party, with covariates that are individualized such as race, gender, income, age, and attentiveness to public affairs. The second level of the model in this example is immediate region of residence, and this comes with its own set of covariates including rural/urban measurement, crime levels, dominant industry, coastal access, and so on. The third level is state, the fourth level is national region, and so on. Each level of the model comes with a regression equation where the variation in intercepts or slopes that are assumed to vary do so with the possible inclusion of group level covariates.

Consider a three level model with individual level covariate X_1, level two group covariate X_2, and level three covariate X_3. The data come as y_{ijk} indicating the ith individual in the jth level two group which is contained in the kth level three group. In the previous example, i represents survey respondents, j represents immediate region, and k represents state. If we allow both varying-intercepts and varying-slopes, then the regression equation at the individual level is:

$$y_{ijk} = \beta_{0jk} + \beta_{1jk} X_{1ijk} + \epsilon_{ijk}, \tag{18}$$

where the ϵ_{ijk} are assumed to be independently and normally distributed. At the second level of the model, we have regression equations for the intercepts and slopes:

$$\begin{aligned}\beta_{0jk} &= \beta_{0k} + \gamma_{0k} X_{2jk} + u_{0jk}\\ \beta_{1jk} &= \beta_{1k} + \gamma_{1k} X_{2jk} + u_{1jk},\end{aligned} \tag{19}$$

where the vectors of (u_{0jk}, u_{1jk}) are assumed to have a common multivariate normal distribution. At the third level of the model, we once again have regression equations for the intercepts and slopes:

$$\begin{aligned}\beta_{0k} &= \beta_0 + \delta_{00}X_{3k} + v_{00k}\\ \gamma_{0k} &= \gamma_0 + \delta_{01}X_{3k} + v_{01k}\\ \beta_{1k} &= \beta_1 + \delta_{10}X_{3k} + v_{10k}\\ \gamma_{1k} &= \gamma_1 + \delta_{11}X_{3k} + v_{11k},\end{aligned} \tag{20}$$

where the vectors of level three residuals $(v_{00k}, v_{01k}, v_{10k}, v_{01k})$ are assumed to have a common multivariate normal distribution. As with any multilevel model, this can be represented by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$, but it is more easily understood in the hierarchical specification. Just from this simple framework, extensions abound. For example, since the level two residuals are indexed by both j and k, a natural relaxation of the model is to let the distribution of $\mathbf{u}_{jk}$ depend on k and then bring these distributions together at the third level. Also, we might have a reason to believe that the level three covariate only effects intercepts and not slopes or that slopes and intercepts vary at level two but only intercepts vary at level three, both of which are easy modifications in this hierarchical specification. In order to see a more clear connection with (11), separate-out the mean structure using the appropriate group level variables and interaction effects and use different design matrices for the random effects at the two levels. Defining $\mathbf{Z}_\ell$ and $\mathbf{u}_\ell$ as the explanatory variable matrix at the group level and random effects vector, respectively, at the ℓth level, we have:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_\ell \mathbf{Z}_\ell \mathbf{u}_\ell + \boldsymbol{\epsilon}, \tag{21}$$

and the model is completed by making distributional assumptions about the $\mathbf{u}_\ell$.

4.2 Non-Nested Groupings

The extension to non-nested (crossed) groupings is straightforward from (21) focusing on the $\mathbf{u}$ subscripts that represent levels in a hierarchical model. In order to generalize to the non-nested case, let those subscripts represent any particular grouping of interest. Consider data where we have two different groupings at the second level. Returning to the survey example, imagine respondents in a state who have both an immediate region of residence and an occupation. These respondents are then naturally grouped by the multiple regions and the jobs, where these groups obviously are not required to have the same number of individuals: there are more residents in a large urban region than a nearby rural county, and we would expect more clerical office workers than clergymen, for example. This is non-nested in the sense that there are multiple people in the same region with the same *and* different jobs. Notate region of residence with the index r and the occupations with the index o, letting iro refer to the ith respondent who is in both the rth region class and the oth occupation class. A regression equation with individual level covariate X and intercepts which vary with both groupings is given by:

$$y_{iro} = \beta_0 + X_{iro}\beta_1 + u_r + u_o + \epsilon_{iro}, \tag{22}$$

where the random effects u_r have one common normal distribution and the random effects u_o have a different common normal distribution. To write this in the form of (21), we can use directly reference the different design matrices for history and composition classes.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_R\mathbf{u}_R + \mathbf{Z}_O\mathbf{u}_O + \boldsymbol{\epsilon}. \tag{23}$$

The addition of a random effect u_{ro} that depends on both region and occupation means that this model has three levels: the individual level, the level of intersections of region and occupation, and the level of region or occupation. The second level of the hierarchy would naturally nest in both of the level three groupings. There are many ways to extend the MLM with crossed groupings to take into account complicated structures that could generate observed data. The key to effectively using these models in practice is to consider the possible ways in which different groupings can effect the outcome variable and then including these in appropriately defined regression equations.

4.3 Generalized Linear Forms

The extension to generalized linear models with qualitative outcomes is also straightforward. This involves inserting a link function between the outcome variable and the additive-linear form based on the right-hand-side with the explanatory variables (see Gill [1999] for details on the over-arching theory behind generalized linear models from a social science perspective). For a two level model, this means modeling the linked variable as:

$$\eta_{ij} = \beta_{0j} + X_{1ij}\beta_{1j} + \ldots + X_{(q-1)ij}\beta_{(q-1)j} + X_{qij}\beta_q + \ldots + X_{(p-1)ij}\beta_{(p-1)} + \epsilon_{ij}, \tag{24}$$

where this is then related to the mean of the outcome, $\eta_{ij} = g(\mu_{ij}) = g(E[Y_{ij}])$. In general, since the matrices $\mathbf{X}$ and $\mathbf{Z}$ can contain the appropriate group level variables, we can express (24) as:

$$g(E[Y_i]) = \eta_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u},$$

and complete the model by making appropriate distributional assumptions about $\mathbf{u}$.

In a simplified logistic case, suppose we have outcomes from the same binary choice at the individual-level, a single individual level covariate, and some possible level two covariates. Then the regression equation for varying-intercepts and varying-slopes with a level 2 grouping is given by:

$$p(Y_{ij} = 1) = \text{logit}^{-1}(\beta_{0j} + X_{ij}\beta_{1j})$$

Make the following assumption for the $j = 1, \ldots, J$ intercepts and slopes:

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} = N_2\left(\begin{bmatrix} m_{0j} \\ m_{1j} \end{bmatrix}, \Omega\right)$$

where m_{0j} and m_{1j} are the group-level mean structures for the intercept and slope of group j. The parameters to be estimated are the coefficients in the mean structure (the fixed effects), and the elements of the covariance matrix Ω. Notice that the distribution at the second level is given explicitly as a normal distribution whereas the distribution at the first level is implied by the assumption of having Bernoulli trials. It is common to stipulate normal forms at higher levels in the model since these are both convenient and supported by asymptotic theory. Other distributions are possible, and may be reasonable under specific circumstances, but tend to heavily tax conventional statistical software solutions. Other standard forms for the link function in $g(\mu)$ include probit, Poisson (log-linear), gamma, multinomial, ordered categorical forms, and more. Many statistical packages are available for fitting a variety of these models, but as assumptions are relaxed about distributional forms or more exotic generalized linear models are used, one must resort to more flexible estimation strategies, such as using `WinBUGS` or `JAGS` to sample from the appropriate posterior distributions.

5 Vocabulary

An unfortunate consequence of the development of multilevel models in disparate literatures is that the vocabulary describing these models differs, even for the exact same specification. The primary confusion is between the synonymous terms of *multilevel model* or *hierarchical model* and the descriptions that use *effects*. Varying-coefficients models, either intercepts or slopes, are often called *random effects models* since they are associated with distributional statements like $\beta_{0j} = N(\gamma_{00} + \gamma_{01} Z_{0j}, u_{0j})$ above. Regretfully, *fixed effects* is more nebulous with different meanings from different settings. Sometimes this is applied to unmodeled coefficients that are constant across individuals, or "nuisance" coefficients that are uninteresting but included by necessity in the form of controls, or even in the case where the data represent a population instead of a sample. A related term, as noted before, is *mixed models*, meaning that the specification has both modeled and unmodeled coefficients.

It is annoying that "fixed" and "random" can also differ in definition by literature (Kreft and De Leeuw 1988, Section 1.3.3, Gelman 2005), The obvious solution to this confusion is to not worry about labels but to pay attention to the implications of *subscripting* in the described model. We conceptualize these specifications as members of a larger multilevel family where indices are purposely *turned-on* to create a level, or *turned-off* to create a point estimate.

6 Case Study: Party Identification in Western Europe

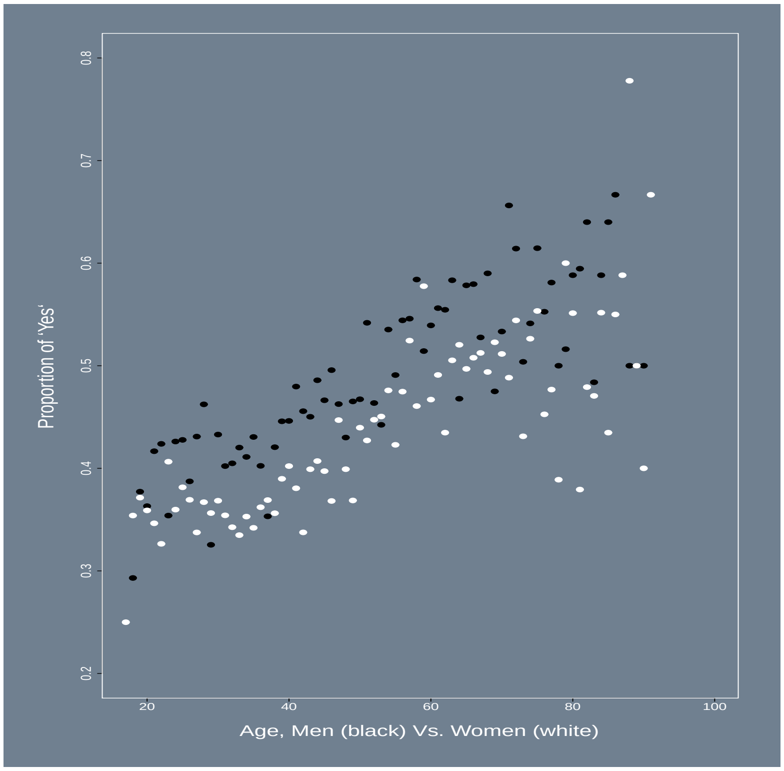
As an illustration, consider 23,355 citizen's feeling of alignment with a political party in ten Western European countries[1] taken from the Comparative Study of Electoral Systems (CSES) for 16 elections from 2001 to 2007. The natural hierarchy for these eligible voters is: district, election, and country (some countries held more than one parliamentary election during this time period). The percentage of those surveyed who felt close of one party varies from 0.29 (Ireland 2002) to 0.65 (Spain 2004).

Running a logistic regression model on these data using individual, district, and country level covariates as though they are individual specific (fully-pooled) requires dramatically different ranges for the explanatory variables to produce reliable coefficients. Since Western European countries do not show such differences in individual level covariates and the country level covariates do not vary strongly with the outcome variable (correlation of -0.25, the model needs take into account higher levels variations. This is done by specifying a hierarchical model to take into account the natural groupings in the data.

The CSES dataset provides multiple ways to consider hierarchies through region and time. Respondents can be nested in voting districts, elections, and countries. Additionally, one could add a time dynamic taking into account that elections within a single country have a temporal ordering. Twelve of the elections considered belong to groupings of size two (grouped by country) and in four countries there was a single election. Since we expect heterogeneity to be explained by dynamics within an between particular elections, the developed model will be hierarchical with two levels based on districts and elections. In Figure 4 we have plotted the observed fraction of "Yes" answers for each age, separated by gender for these elections. Notice that in the aggregated data women are generally less likely to identify with a party for any given age, even though identification for both men and women increases with age.

[1] These are: Switzerland, Germany, Spain, Finland, Ireland, Iceland, Italy, Norway, Portugal, and Sweden.

Figure 4: Empirical Proportion of “Yes” Versus Age, by Gender



The outcome variable for our model is a dichotomous measure from the question “Do you usually think of yourself as close to any particular political party?” coded zero for “no” and one for “yes” (numbers B3028, C3020_1). We focus here on only a modest subset of the total possible number of possible explanatory variables. The individual-level demographic variables are: `Age` of respondent in years (number 2001), `Female` with men equal to zero and women equal to one (number 2002), the respondent’s income quintile labeled `Income`, and the respondent’s municipality coded `Rural/Village`, `Small/Middle Town`, `City Suburb`, `City Metropolitan`, with the first category used as the reference category in the model specification (numbers B2027, C2027). Subjects are nested within electoral districts with a district level variable describing the number of seats elected by proportional representation in the given district. Additionally, these districts are nested within elections with an election level variable describing the effective number of parties in the election. The variable `Parties` (number 5094) gives the effective number of political parties in each country, and `Seats` (number 4001) indicates the number of seats contested in each district of the first segment of the legislature’s lower house where the respondent resides. Further details can be found at `http://www.cses.org/varlist/varlist_full.htm`.

For an initial analysis, we ignore the nested structure of the data and simply analyze the dataset using a logistic generalized linear model. The outcome variable is modeled as

$$Y_i|p_i \sim \text{Bern}(p_i) \qquad\qquad \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{X}_i\boldsymbol{\beta},$$

where $\mathbf{X}_i$ is the vector of covariates for the ith respondent and $\boldsymbol{\beta}$ is a vector of coefficients to be estimated. The base categories for this model is men for `Female`, the first income quantile, and `Rural/Village` for region. Table 6 provides this standard logistic model in the first block of results.

For the second model we analyze a two level hierarchy: one at district level and one at the election level, represented by random intercept contributions to the individual-level. The outcome

Table 1: Contrasting Specifications, Voting Satisfaction

	Standard Logit GLM			Random Intercepts Version		
	Estimate	Std. Error	z-score	Estimate	Std. Error	z-score
Intercept	-0.3946	0.0417	-9.46	-0.1751	0.0944	-1.855
Age	0.0156	0.0008	19.25	0.0168	0.0008	20.207
Female	-0.2076	0.0267	-7.76	-0.2098	0.0272	-7.709
Income Level 2	0.1178	0.0425	2.77	0.0350	0.0436	0.801
Income Level 3	0.2282	0.0427	5.35	0.1474	0.0439	3.357
Income Level 4	0.2677	0.0443	6.04	0.2468	0.0453	5.454
Income Level 5	0.2388	0.0451	5.30	0.2179	0.0466	4.673
Small/Middle Town	0.0665	0.0392	1.70	-0.0822	0.0429	-1.916
City Suburb	0.1746	0.0431	4.05	-0.0507	0.0500	-1.014
City Metropolitan	0.1212	0.0359	3.38	0.0636	0.0417	1.525
Parties	-0.0408	0.0142	-2.87	-0.1033	0.0846	-1.222
Seats	0.0047	0.0010	4.82	0.0027	0.0019	1.403
Residual Deviance	31608 on 23343 df			31079 on 23341 df		
Null Deviance	32134 on 23354 df			31590 on 23352 df		
				$\sigma_d = 0.2402, \sigma_e = 0.32692$		

variable is now modeled as:

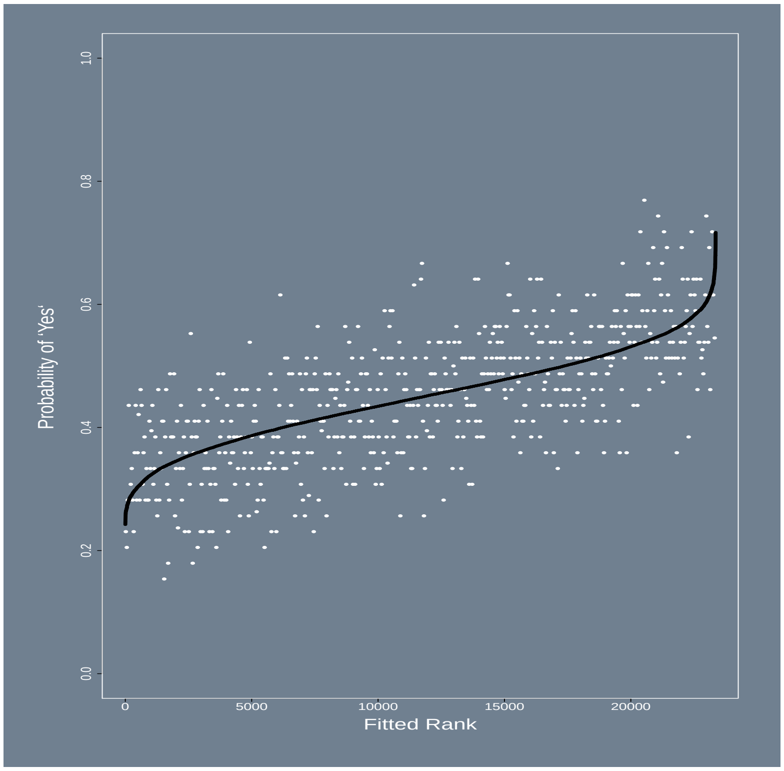
$$\begin{aligned}
Y_{ijk}|p_{ijk} &\sim \text{Bern}(p_{ijk}) \\
\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) &= \beta_{0jk} + \mathbf{X}_{ijk}\beta \\
\beta_{0jk} &= \beta_{0k} + \beta_{\texttt{Seats}} \times \texttt{Seats}_{jk} + u_{0jk} \\
\beta_{0k} &= \beta_0 + \beta_{\texttt{Parties}} \times \texttt{Parties}_k + v_{0k} \\
u_{0jk} &\sim \mathcal{N}(0, \sigma^2_{u_0}) \\
v_{0k} &\sim \mathcal{N}(0, \sigma^2_{v_0}).
\end{aligned}$$

Therefore `Seats` and `Parties` are predictors at different levels of the model. Since `Parties` is constant in an election, it predicts the change in intercept at the election level of the hierarchy. `Seats` is constant in districts but varies within an election, so it predicts the change in intercept at the district level. In addition, all of the district level random effects have a common normal distribution which does not change depending on election and the election level random effects have a different common normal distribution. The three levels of the hierarchy are evident and stochastic processes are introduced at each level: Bernoulli distributions at the data level and normal distributions at the district and election level.

The multilevel model fits better by several standard measures. The chi-square comparison of the regular GLM deviance to the null deviance for that model is well into the tail for 11 degrees of freedom (the difference from Table 6), indicating excellent overall fit. The multilevel model chi-square deviance comparison with 11 degrees of freedom produces a similarly small tail value. A more striking comparison comes from considering the deviance of the multilevel model and the GLM. The decrease in deviance is 529 with only two additional degrees of freedom used. This shows that the multilevel model fits the data dramatically better than the GLM. We can also look at the percent correctly predicted with the naïve criterion (splitting predictions at the arbitrary 0.5 threshold). The standard GLM gives 57.367% estimated correctly whereas the multilevel model gives 60.522%.

In Table 6 we see essentially no difference in the coefficient estimates between the two models for `Age`, Female, `Income Level 3`, `Income Level 4`, `Income Level 5`, and `Small/Middle Town`. However, notice the differences between the two models for the coefficient estimates of `Income Level 2`, `City Suburb`, `City Metropolitan`, `Parties`, and `Seats`. In all cases where we observe these differences the standard generalized linear model gives more reliable coefficient estimates (smaller standard errors), but this is misleading. Once the correlation structure of the data is taken into account, there is more uncertainty in the estimated values of these parameters.

Figure 5: Probability of "Yes" Versus Fitted Rank



To further evaluate the fit of the regular GLM we plot the ranked fitted probabilities from the model against binned estimates of the observed proportion of success in the reordered observed data. Figure 5 indicates that although the model does describe the general trend of the data, it misses some features since the point-cloud does not adhere closely to the fitted line underlying the assumption of the model. The curve of fitted probabilities only describes 47% of the variance in these empirical proportions of success. This suggests that there are additional features of the data, and we will capture these with the multilevel specification. The fitted probabilities of the multilevel model describe 71% of the variation in the binned estimates of the observed proportion of success.

After running the initial GLM, we would have improperly concluded that the number of seats in a district, the effective number of political parties, and city type significantly influenced the response. However, these variables in this specification are mimicking the correlation structure of the data, which was more effectively taken into account through the multilevel model framework, as evidenced by the large gains in predictive accuracy. This is also apparent if we take the binned empirical values and break down their variance in various ways. To see that the random effects have a meaningful impact on data, fit, we can compare how well the fixed effects and the full model predict the binned values. Normally a researcher would be happy if the group level standard deviations were of similar size to the residual standard deviation, and small group effects relative to the residual effect are indications that the grouping is not effective in the model specification. We use the binned empirical values to mimic this kind of analysis for a generalized linear mixed model. The variance of the binned values minus the fixed effects is 0.0111 and the variance of the binned values minus the fitted values from the full model is 0.0060, indicating that a significant amount of variation in the data that is captured is indeed captured by the random effects.

6.1 Computational Considerations

Multilevel models are more demanding of statistical software than more standard regressions. In the case of the Gaussian-Gaussian multilevel model such as the one-way random effects model, we can integrate out all levels of the hierarchy and are left with a likelihood for $\boldsymbol{y}$ which only depends on the fixed effects and the induced covariance structure. Maximum likelihood estimates for the coefficients on the fixed effects as well as the parameters of the variance components can be computed. However, with the maximum likelihood estimation of the standard linear model, the estimates of the parameters of the variance component are biased. This has led to the use of **RE**stricted **M**aximum **L**ikelihood (REML) methods for estimating variance components from a standard likelihood point of view.

At the heart of the REML method is a partition of the likelihood function into two pieces, one which is free of the fixed effects and one which depends on the fixed effects. The variance components are then estimated by maximizing the piece of the likelihood which only depends on the variance components and not the fixed effects. In the Gaussian-Gaussian setting, this step is achieved through an appropriate linear transformation of the data. The estimate of the variance component is then used when finding estimates of the fixed effects. This second step is equivalent to maximizing the entire likelihood conditioned upon the assumption that the variance component is equal to the one estimated using the restricted likelihood.

Even using the REML approach, computation can be burdensome and time-consuming and so EM algorithm of Dempster, Laird, and Rubin is often employed when computing estimates. While REML uses a restricted likelihood, the EM algorithm expands the data into a "complete" data vector by appending some parameters (usually location parameters) to the end of the data vector. This expanded data vector now depends only on the remaining parameters (those of the variance component) and one iterates the **E**xpectation and **M**aximization steps. In the M step, an MLE (REMLE) for the variance components is found conditioned on the current value of the fixed effects. In the E step, the expected values of the sufficient statistics for the fixed effects is found conditioned on the current value of the variance components.

Beyond these methods for computing ML and REML estimates, the Bayes paradigm offers powerful Markov chain Monte Carlo (MCMC) simulation tools for computing posterior statistics of interest. It has been shown that an empirical Bayes approach provides estimators that are equivalent to those from the REML method. In the Gaussian-Gaussian multilevel model, the hierarchical specifications and conjugate prior (a mathematically convenient form) for the parameters lead to a Gibbs sampler for the parameters of the model. In Gibbs sampling, the full conditional distributions of the parameters are sampled iteratively, producing a Markov chain that eventually converges such that these conditional draws behave as if they are draws from the marginal posterior distributions for the parameters. When model prior specifications do not lend themselves well to Gibbs sampling, the Metropolis-Hastings algorithm (an iterative modification of rejection sampling) is utilized to produce samples from the appropriate posterior distribution. One possible complication in the utilization of the MCMC techniques is the dimensionality of the random effects. This has become a well-studied problem and there are a wide range of tools for addressing this problem. Many nonlinear multilevel models can be estimated conveniently by moving to MCMC procedures with the easy-to-used (and free) software packages WinBUGS and JAGS already mentioned.

7 Summary

This introduction to multilevel models provides an overview of a class of regression models that account for hierarchical structure in data. Such data occurs when there are natural levels of aggregation whereby individual cases are nested within groups, and those groups may also be nested in higher level groups. We provide a general description of the model features that enable multilevel models to account for such structure, demonstrate that ignoring hierarchies produces incorrect inferential statements in model summaries, and illustrate our points with a simple example using a real dataset.

Aitkin and his coauthors (especially, 1981, 1986) introduced the linear multilevel model in the 1980s, concentrating on applications in education research since the hierarchy in that setting is obvious: students in classrooms, classrooms in schools, schools in districts, and districts in states. These applications were all just linear models and yet they substantially improved fit to the data in educational settings. Since this era more elaborate specifications have been developed for non-linear outcomes, non-nested hierarchies, correlations between hierarchies, and more. This has been a very active area of research both theoretically and in applied settings. These developments are described in detail in subsequent chapters.

Multilevel models are flexible tools because they exist in the spectrum between fully-pooled models, where groupings are ignored, and fully-unpooled models, where each group gets its own regression statement. This means that multilevel models recognize both commonalities within the cases and differences between group effects. The gained efficiency is both notational and substantive. The notational efficiency occurs because there are direct means of expressing hierarchies with subscripts, nested subscripts, and sets of subscripts. This contrasts with messy "dummy" coding of group definitions with large numbers of categories. Multilevel models account for individual versus group-level variation because these two sources of variability are both explicitly accounted-for. Since all non-modeled variation falls to the residuals, multilevel models are guaranteed to capture between-group variability when it exists. These forms are also a convenient way of estimating separately, but concurrently, regression coefficients for groups. The alternative is to construct separate models whereby between-group variability is completely lost. In addition, multilevel models provide more flexibility for expressing social science theories. We routinely consider settings where individual cases are contained in larger groups, which themselves are contained in even larger groups, and so on. This is relatively common in sociology, political science, public administration, anthropology, and other related fields.

It is important to observe that multilevel models are actually an explicit Bayesian statement. The core tenets of Bayesian inference are: everything unknown is described probabilistically, and probability statements are updated by conditioning on data as it becomes available. Multilevel models build distributional statements around unknown coefficient effects with a probability model at the next higher level. Therefore these are directly Bayesian specifications. Furthermore, all likelihood models are equivalent to Bayesian models with the appropriate uniform prior distribution on the coefficient estimates, and asymptotically there is no difference with any non-degenerate prior. This Bayesian perspective is quite helpful since prior information abounds in the social sciences and it is important and helpful to use it. These issues are explored in detail in Chapter 4 of this volume.

There are also real problems with ignoring hierarchies in data with existing aggregation. Resulting models will have the wrong standard errors on group-affected coefficients since fully-pooled results assume that the apparent commonalities are a results of individual effects and decisions. This problem also spills over into covariances between coefficient estimates. There are also substantive reasons to consider multilevel forms. Many social science theories are actually expressed hierarchically. We believe that individuals behave differently in different political, social, or spatial, groups. Social groups, political parties, nations, regions, institutional settings, all impose effects on those individuals contained within. To hypothesize a complicated theory of human interaction and then ignore these interactions when performing statistical modeling is troublesome to say the least. Therefore multilevel models are an especially important modeling tool for empirical social scientists.

8 References

Aitkin, Murray , Dorothy Anderson and John Hinde. 1981. “Statistical Modelling of Data on Teaching Styles.” *Journal of the Royal Statistical Society, Series A* 144, 419-461.

Aitkin, M. and N. Longford. 1986. “Statistical Modelling Issues in School Effectiveness Studies.” *Journal of the Royal Statistical Society, Series A* 149, pp. 1-43.

Albert, J. H. and Chib, S. 1993. “Bayesian Analysis of Binary and Polychotomous Response Data.” *Journal of the American Statistical Association* 88, 669-679.

Baum, L. E. and Petrie, T. 1966. “Statistical Inference for Probabilistic Functions of Finite Markov Chains.” *Annals of Mathematical Statistics* 37, 1554-1563.

Booth, James G. , George Casella, James P. Hobert. 2008. “Clustering Using Objective Functions and Stochastic Search.” *Journal of the Royal Statistical Society, Series B* 70, 119-139.

Bryk, A. S., and S. W. Raudenbush. 1987. “Applications of Hierarchical Linear Models to Assessing Change.” *Psychological Bulletin* 101, 147-158.

Bryk, A. S., S. W. Raudenbush, M. Seltzer, and R. Congdon. 1988. *An Introduction to HLM: Computer Program and User’s Guide.* Second Edition. Chicago: University of Chicago Department of Education.

Burstein, L. 1980. “The Analysis of Multi-Level Data in Educational Research and Evaluation.” *Review of Research in Education* 8, 158-233.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. 1992. “Hierarchical Bayesian Analysis of Changepoint Problems.” Applied Statistics, Vol. 41, No. 2. (1992), 389-405.

Christiansen, C. L. and Morris, C. N. 1997. “Hierarchical Poisson Regression Modeling.” *Journal of the American Statistical Association* 92, 618-632.

Cohen, J., Nagin, D., Wallstrom, G., and Wasserman, L. 1998. “Hierarchical Bayesian Analysis of Arrest Rates.” *Journal of the American Statistical Association* 93, 1260-1270.

Cowles, M. K. 2002. “MCMC Sampler Convergence Rates for Hierarchical Normal Linear Models: A Simulation Approach.” *Statistics and Computing* 12, 377-389.

Daniels, Michael J. and Constantine Gatsonis. 1999. “Hierarchical Generalized Linear Models in the Analysis of Variations in Health Care Utilization.” *Journal of the American Statistical Association* 94, 29-42.

De Leeuw, J. and I. Kreft. 1986. “Random Coefficient Models for Multilevel Analysis.” *Journal of Educational Statistics* 11, 57-85.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B* 39, 1-38.

Baum, L. E. and Eagon, J. A. 1967. "An Inequality with Applications to Statistical Estimation for Probabilistic Functions of Markov Processes and to a Model for Ecology." *Bulletin of the American Mathematical Society* 73, 360-363.

Eisenhart, C. 1947. "The Assumptions Underlying the Analysis of Variance." *Biometrics* 3, 1-21.

Gelman, Andrew 2005. "Analysis of Variance: Why It Is More Important Than Ever." *Annals of statistics* 33, 1-53.

Gelman, Andrew. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis* 1, 515-533.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52, 647-674.

Goldstein, H. 1987. *Multilevel Models in Education and Social Research.* Oxford: Oxford University Press.

Goldstein, H. 1985. *Multilevel Statistical Models.* New York: Halstead Press.

Hadjicostas, P. and Berry, S. M. 1999. "Improper and Proper Posteriors with Improper Priors in a Poisson-Gamma Hierarchical Model." *Test* 8, 147-166.

Hartley, H. O. 1958. "Maximum Likelihood Estimation From Incomplete Data." *Biometrics* 14, 174-194.

Henderson, C. R. 1950. "Estimation of Genetic Parameters." *Biometrics* 6, 186-187.

Healy, Michael and Michael Westmacott. 1956. "Missing Values in Experiments Analysed on Automatic Computers." *Journal of the Royal Statistical Society, Series C* 5, 203-206.

Henderson C. R. O. Kempthorne, S. R. Searle and von Krosigk, C. M. 1959. "The Estimation of Environmental and Genetic Trends From Records Subject To Culling." *Biometric* 15, 192.

Hobert, J. P. and Casella, G. 1996. "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models." *Journal of the American Statistical Association* 91, 1461-73.

Hodges, J. S. and Sargent, D. J. 2001. "Counting Degrees of Freedom in Hierarchical and Other Richly Parameterized Models." *Biometrika* 88, 367-379.

Jones, Galin L. and Hobert, James P. 2001. "Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo." *Statistical Science* 16, 312-34.

Kreft, I. G. G. and De Leeuw, J. 1988. "The Seesaw Effect: A Multilevel Problem?" *Quality and Quantity* 22, 127-137.

Laird, N.M. and Ware, J.H. 1982. "Random-effects Models for Longitudinal Data." *Biometrics* 38, 963–974.

Lee, V. E. and Bryk, A. S. 1989. "Multilevel Model of the Social Distribution of High School Achievement." *Sociology of Education* 62, 172-192.

Liu, J. S. 1994. "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem." *Journal of the American Statistical Association* 89, 958-966.

Longford, N. T. 1987. "A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models With Nested Random Effects." *Biometrika* 74, 817-827.

Mason, W. M., Wong, G. Y., and Entwistle, B. 1983. "Contextual Analysis Through the Multilevel Linear Model." In *Sociological Methodology 1983-1984*, S. Leinhardt (ed.). Oxford: Blackwell, 72-103.

McKendrick, A. G. 1926. "Applications of Mathematics to Medical Problems." *Proceedings of the Edinburgh Mathematical Society* 44, 98-130.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller E. 1953. "Equation of State Calculations by Fast Computing Machines." *Journal of Chemical Physics* 21, 1087-1091.

Newcomb, S. 1886. "A Generalized Theory of the Combination of Observations So As to Obtain the Best Results." *American Journal of Mathematics* 8, 343-366.

Pettitt, A. N., Tran, T. T., Haynes, M. A., and Hay, J. L. 2006. "A Bayesian Hierarchical Model for Categorical Longitudinal Data From a Social Survey of Immigrants." *Journal of the Royal Statistical Society, Series A* 169, 97-114.

Raudenbush, S. W. 1988. "Education Applications of Hierarchical Linear Models: A Review." *Journal of Educational Statistics* 12,85-116.

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

Scheffé 1956, Henry. 1956. "Alternative Models for the Analysis of Variance." *The Annals of Mathematical Statistics* 27, 251-271.

Stangl, D. K. 1995. "Prediction and Decision Making Using Bayesian Hierarchical Models." *Statistics in Medicine* 14, 2173-2190.

Steffey, D. 1992. "Hierarchical Bayesian Modeling With Elicited Prior Information." *Communications in Statistics* 21, 799-821.

Zangwill, W. I. 1969. *Nonlinear Programming: A Unified Approach.* Englewood Cliffs, NJ: Prentice-Hall.