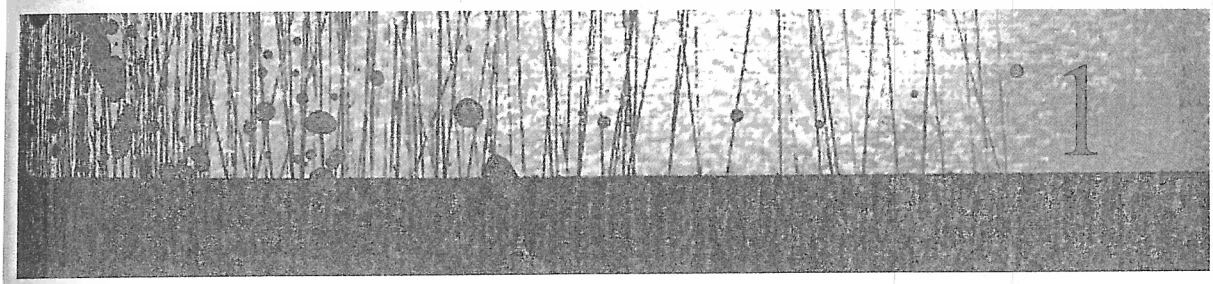

The Multilevel Model Framework

Jeff Gill
Washington University, USA
Andrew J. Womack
University of Florida, USA

1.1 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 data across the sciences, ranging from the social, behavioral, health, and economic sciences to the biological, engineering, and physical sciences, yet is commonly ignored by researchers performing statistical analyses. 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 Studies, Latinobarometer, Eurobarameter, and Afrobarometer are obvious cases. Often in population biology a hierarchy is built using ancestral information, and phenotypic variation is used to estimate the heritability of certain traits, in what is commonly referred to as the "animal model." In image processing, spatial relationships emerge between the intensity and hue of pixels. There are many hierarchies that emerge in language processing, such as topic of discussion, document type, region of origin, or intended audience. In longitudinal studies, more complex hierarchies emerge. Units or groups of units are repeatedly observed over

a period of time. In addition to group hierarchies, observations are also grouped by the unit being measured. These models are extensively used in the medical/health sciences to model the effect of a stimulus or treatment regime conditional on measures of interest, such as socioeconomic status, disease prevalence in the environment, drug use, or other demographic information. 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. As a result, 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, there is 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. The departure comes from the treatment of some of the coefficients assigned to the explanatory variables. What can be done to modify a model when a point estimate is inadequate to describe the variation due to a measured variable? An obvious modification is to treat this coefficient as having a distribution as opposed to being a fixed point. A regression equation can be introduced to model the coefficient itself, using information at the group level to describe the heterogeneity in the coefficient. This is the heart of the multilevel model. 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 often referred to as "mixed," meaning some of the coefficients are *modeled* while others are *unmodeled*.

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 of 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 (Goel and DeGroot 1981, Goel 1983). This is because the effects of the parameterizations at these super-high levels gets washed out as it comes down the hierarchy. 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 combinatorially 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.

1.2 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 generalizable and broadly applicable principles. For instance, the well-known EM algorithm of Dempster, Laird, and Rubin (1977) was pre-dated 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 of statistics. It turns out that hierarchical modeling follows this same storyline, 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 standalone software packages `HLM`, `ML2`, `VARCL`, as well as incorporation into the `SAS` procedure `MIXED`, and others. Additional work by Goldstein (notably 1985) took the two-level model and extended it to situations with further nested groupings, non-nested groupings, time series cross-sectional data, and more. At roughly the same time, a series of influential papers and applications grew out of Laird and Ware (1982), where a random effects model for Gaussian longitudinal data is established. This Laird–Ware model was extended to binary outcomes by Stiratelli, Laird, and Ware (1984) and GEE estimation was established by Zeger and Liang (1986). An important extension to non-linear mixed effects models is presented in Lindstrom and Bates (1988). In addition, Breslow and Clayton (1993) developed quasi-likelihood methods to analyze generalized linear mixed models (GLMMs).

Beginning around the 1990s, hierarchical modeling took on a much more Bayesian complexion now that stochastic simulation tools (e.g. 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, Liu 1994, Hobert and Casella 1996, Jones and Hobert 2001, Cowles 2002). Other works focused on solving specific applied problems with Bayesian models: Steffey (1992) incorporates expert information into the model, Stangl (1995) develops prediction and decision rules, Cohen et al. (1998) model arrest rates, Zeger and Karim (1991) use GLMMs to study infectious disease, Christiansen and Morris (1997) build on count models hierarchically, Hodges and Sargent (2001) refine

inference procedures, Pauler et al. (2001) model cancer risk, and Pettitt et al. (2006) model survey data from immigrants. Recently, Bayesian prior specifications in hierarchical models have received attention (Hadjicostas 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. A review of software available for fitting these models is presented in Chapter 26 of this volume. For basic models, the `lme4` package in `R` works quite well and preserves `R`'s intuitive model language. Also, `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).

1.3 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:

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i, \qquad (1.1)$$

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 for making inferences since standard least squares procedures produce an estimate of the standard error (Amemiya 1985, Ravishanker and Dey 2002), but with reasonable sample size and finite variance the central limit theorem applies. For maximum likelihood results, which produce the same estimates as does least squares, the derivation of the estimator begins with the assumption of normality of population residuals. See the discussion on pages 583–6 of Casella and Berger (2001) for a detailed derivation.

1.3.1 Basic Linear Forms, Individual-Level Explanatory Variables

How does one model the heterogeneity that arises because each i case belongs to one of $j = 1, \ldots, J$ groups where $J < n$? Even if there does not exist 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 (1.1) 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 $i = 1, \ldots, n_j$ (where n_j is the size of the jth group):

$$y_{ij} = \beta_{0j} + x_{ij}\beta_1 + \varepsilon_{ij},$$

where the added j subscript indicates that this case belongs to the jth group and gets intercept β_{0j}. The β_{0j} are group-specific intercepts and are usually given a common normal distribution with mean β_0 and standard deviation σ_{u_0}. The overall intercept β_0 is referred to as a *fixed effect* and the difference $u_{0j} = \beta_{0j} - \beta_0$ is a *random effect*. Subscripting the u with 0 denotes that it relates to the intercept term and distinguishes it from other varying quantities to be discussed shortly. Since the β_1

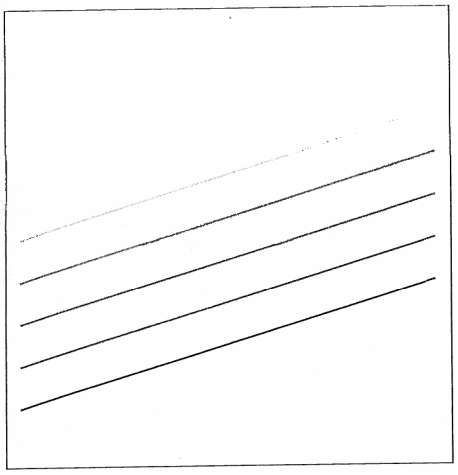

Figure 1.1 Varying Intercepts

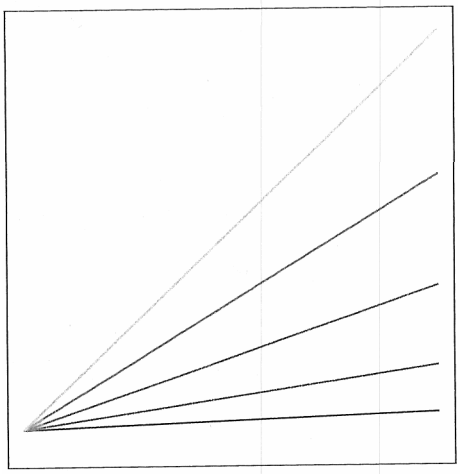

Figure 1.2 Varying Slopes

coefficient is not indexed by the grouping term j, this is still constant across all of the $n = n_1 + \cdots + n_J$ cases and evaluated with a standard point estimate. This model is illustrated in Figure 1.1, which 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-intercept* model.

In a different context, one may want to account for the groupings in the data, but the researcher may feel that the effect is not through the intercept, where the groups start at a zero level of the explanatory variable x, having reason to believe that the grouping affects the slopes instead: as x increases, group membership dictates a different change in y. So now loosen the definition of the single slope, β_1, in (1.1) to account for the groupings according to:

$$y_{ij} = \beta_0 + x_{ij}\beta_{1j} + \varepsilon_{ij},$$

where the added j subscript indicates that the n_j cases in the jth group get slope β_{1j}. The intercept now remains fixed across the cases in the data and the slopes are given a common normal distribution with mean β_1 and standard deviation σ_{u_1}. In this situation, β_1 is a *fixed effect* and the difference $u_{1j} = \beta_{1j} - \beta_1$ is a *random effect*, and subscripting the u with 1 denotes that these random effects relate to the slope as opposed to the intercepts. This is illustrated in Figure 1.2 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.

Suppose the researcher suspects that the heterogeneity in the sample 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:

$$y_{ij} = \beta_{0j} + x_{ij}\beta_{1j} + \varepsilon_{ij},$$

where membership in group j for case ij has two effects, one that is constant and one that differs from others with increasing x. The vectors (β_{0j}, β_{1j}) are given a common multivariate normal distribution with mean vector (β_0, β_1) and covariance matrix Ω_u. The vector of means (β_0, β_1) is the *fixed effect* and the vectors of differences $(u_{0j}, u_{1j}) = (\beta_{0j} - \beta_0, \beta_{1j} - \beta_1)$ are the *random effects*. A synthetic, possibly exaggerated, model result is given in Figure 1.3. Not surprisingly, this is called the *varying-intercept, varying-slope* or *random-intercept, random-slope* model. Notice from the simple artificial example in

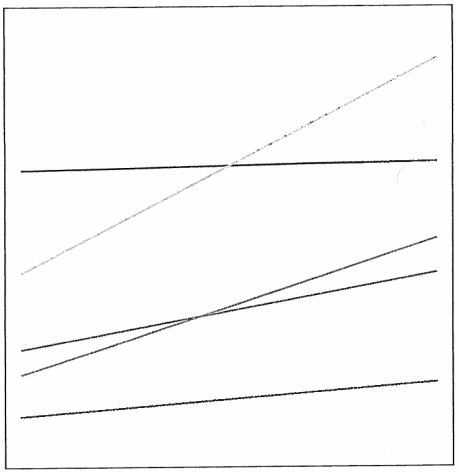

Figure 1.3 Varying Intercepts and Slopes

the figure that it is already possible to model quite intricate differences in the groups for this basic linear form.

Before moving on to more complicated models, a clear explanation of what is meant by *fixed effects* and *random effects* is necessary. *Fixed effects* are coefficients in the mean structure that are assumed to have point values, whereas *random effects* are coefficients that are assumed to have distributions. For example, in the models thus far considered the *random effects* have been assumed to have a common Gaussian distribution. A distinction must be made, then, for when group-level effects are assumed to be *fixed effects* or *random effects*. In the random-intercepts model, for example, the β_{0j} can be assumed to be point values instead of being modeled as coming from a distribution, as *fixed effects* as opposed to *random effects*. This distinction is quite important and modifies both the assumptions of the models as well as the estimation strategy employed in analyzing them.

1.3.2 Basic Linear Forms, Adding Group-Level Explanatory Variables

The bivariate linear form can be additively extended on the right-hand side to include more covariates, which may or may not receive the grouping treatment. A canonical *mixed* form is one where the intercept and the first q explanatory variables have coefficients that vary by the $j = 1, \ldots, J$ groupings (for a total of $q + 1$ modeled coefficients), but the next p coefficients, $q + 1, q + 2, \ldots, q + p$, are fixed at the individual level. This is given by the specification:

$$\begin{aligned} y_{ij} = \; & \beta_{0j} + x_{1i}\beta_{1j} + \cdots + x_{qi}\beta_{qj} \\ & + x_{(q+1)i}\beta_{q+1} \\ & + \cdots + x_{(q+p)i}\beta_{q+p} + \varepsilon_{ij}, \end{aligned}$$

where membership in group j for case ij has $q+1$ effects. The vectors of group-level coefficients $(\beta_{0j}, \ldots, \beta_{qj})$ are given a common distribution, which for now will be assumed to be Gaussian with mean vector $(\beta_0, \ldots, \beta_q)$ (the *fixed effects*) and covariance matrix $\mathbf{\Omega}_u$. As before, the vectors of differences $\mathbf{u}_j = (\beta_{0j} - \beta_0, \ldots, \beta_{qj} - \beta_q)$ are the *random effects* for this model.

An important aspect of these models, which greatly facilitates their use in generalized linear models and nonlinear models, is writing them in a hierarchical fashion. The model is written as a specification of a regression equation at each level of the hierarchy. For example, consider the model where

$$y_{ij} = \beta_{0j} + x_{1ij}\beta_{1j} + x_{2ij}\beta_2 + \varepsilon_{ij}, \quad (1.2)$$

where there are two random coefficients and one fixed coefficient. The group-level coefficients can be written as

$$\beta_{0j} = \beta_0 + u_{0j} \qquad \beta_{1j} = \beta_1 + u_{1j} \quad (1.3)$$

and the u_{0j}, u_{1j} appear as errors at the second level of the hierarchy. Assumptions are then made about the distributions of the individual-level errors (ε_{ij}) and group-level errors (u_{0j}, u_{1j}). For example, a common set of assumptions in linear regression is ε_{ij} are independent and identically distributed (iid) with common variance, and (u_{0j}, u_{1j}) is bivariate normal with mean 0 and covariance matrix $\mathbf{\Omega}$.

The model given in (1.2), (1.3) 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 explanatory variables at this second level are added, there is not a modeled reason for the differences in the group-level coefficients. The fact that one can model the variation at the group level is a key feature of treating them as *random effects* as opposed to *fixed effects*. Returning to the linear model with two group-level coefficients and one individual-level coefficient in (1.2), $y_{ij} = \beta_{0j} + x_{1ij}\beta_{1j} + x_{2ij}\beta_2 + \varepsilon_{ij}$, model each of the two group-level coefficients with their own regression equation and index these by $j = 1$ to J:

$$\begin{aligned} \beta_{0j} &= \beta_0 + \beta_3 x_{3j} + u_{0j} \\ \beta_{1j} &= \beta_1 + \beta_4 x_{4j} + u_{1j}, \end{aligned} \qquad (1.4)$$

and the group-level variation is modeled as depending on covariates. The explanatory variables at the second level 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 are subscripted by j instead of ij to identify that they only depend on group identification.

Substituting the two definitions in (1.4) into the individual-level model in (1.2) and rearranging produces:

$$\begin{aligned} y_{ij} &= (\beta_0 + \beta_3 x_{3j} + u_{0j}) \\ &\quad + (\beta_1 + \beta_4 x_{4j} + u_{1j}) x_{1ij} \\ &\quad + \beta_2 x_{2ij} + \varepsilon_{ij} \\ &= (\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} \\ &\quad + \beta_4 x_{4j} x_{1ij}) \\ &\quad + (u_{0j} + u_{1j} x_{1ij} + \varepsilon_{ij}) \end{aligned} \qquad (1.5)$$

for the ijth case. The composite fixed effects now have a richer structure, accounting for variation at both the individual and group levels. In addition, the error structure is now modeled as being due to specific group-level variables. In this example, the composite error structure, $(u_{1j}x_{1ij} + u_{0j} + \varepsilon_{ij})$, is heteroscedastic since it is conditioned on levels of the explanatory variable x_{1ij}. This composite error shows that uncertainty is modified in a standard linear model (ε_{ij}) by introducing terms that are correlated for observations in the same group. Though it seems that this has increased uncertainty in the data, it is just modeling the data in a fuller fashion. This richer model accounts for the hierarchical structure in the data and can provide a significantly better fit to observed data than standard linear regression.

It is important to understand the exact role of the new coefficients. First, β_0 is a universally assigned intercept that all i cases share. Second, β_1 gives another shared term that is the slope coefficient corresponding to the effect of changes in x_{1ij}, as does β_2 for x_{2ij}. These three terms have no effect from the multilevel composition of the model. Third, β_3 gives the slope coefficient for the effect of changes in the variable x_{3j} for group j, and are applied to all individual cases assigned to this group. It therefore varies by group and not individual. Fourth, and surprisingly, β_4 is the coefficient on the interaction term between x_{1ij} and x_{4j}. But though no interaction term was specified in the hierarchical form of the model 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 (1.5), it is surprising to some that this is an automatic consequence.

In the Laird–Ware form of the model, the fixed effects are separated from the random effects as in (1.5). In this formulation, the covariates on the random effects are represented by zs rather than xs in order to distinguish the two. This structure can be written (in matrix form) for group j as $\mathbf{y}_j = \boldsymbol{X}_j\beta + \boldsymbol{Z}_j\mathbf{u}_j + \varepsilon_j$. Here, $\mathbf{y}_j$ is the vector of observations, $\boldsymbol{X}_j$ is the fixed effects design matrix, and $\boldsymbol{Z}_j$ is a the random effects design matrix for group j. The vector $\boldsymbol{\beta}$ is the vector of fixed effects, which are assumed to have point values, whereas $\mathbf{u}_j$ is the vector of random effects for group j which are modeled

by a distributional assumption. Finally, ε_j is a vector of individual-level error terms for group j. From this formulation, it is apparent that multilevel models can be expressed in a single-level expression, although this does not always lead to a more intuitive expression.

1.3.3 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:

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i.$$

In contrast to (1.5), there is no u_{0j} modeled 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 each group can be treated as a separate dataset and modeled completely separately:

$$y_{ij} = \beta_{0j} + x_{ij}\beta_{1j} + \varepsilon_{ij},$$

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 lies the multilevel model. The word "between" here means that groups are recognized as different, but because there is a single model in which 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:

$$y_{ij} = \beta_{0j} + \varepsilon_{ij}, \tag{1.6}$$

which is also called a *mean model* since β_{0j} represents the mean of the jth group. If there is an assumption that $\beta_{0j} = \beta_0$ is constant across all cases, then this becomes the fully pooled model. Conversely, if there are J separate models each with their own β_{0j} which do not derive from a common distribution, then it is the fully unpooled approach. Estimating (1.6) as a partial pooling model (with Gaussian distributional assumptions) 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 (variance of the ε_{ij})
σ_1^2	between-group variance (variance of the $\overline{y}_j$)
n_j	size of the jth group.

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

$$\hat{\beta}_{0j} = \frac{\frac{n_j}{\sigma_0^2}\overline{y}_j + \frac{1}{\sigma_1^2}\overline{y}}{\frac{n_j}{\sigma_0^2} + \frac{1}{\sigma_1^2}}. \tag{1.7}$$

This is a very revealing expression. The estimate of the 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 (1.7) 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 $\hat{\beta}_{0j} \to \overline{y}$, and as $\sigma_1^2 \to \infty$, then $\hat{\beta}_{0j} \to \overline{y}_j$. Thus, the group effect which is at the heart of a multilevel model, is a balance between the size of the group and the standard deviations at the individual and group levels.

1.4 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 scientific research.

1.4.1 Nested Groupings

The generalization of the mixed effects model to nested groupings is straightforward and is most easily understood in the hierarchical framework of (1.2), (1.3) as opposed to the single equation of (1.5).

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. Allowing both varying-intercepts and varying-slopes in the regression equation at the individual level, gives:

$$y_{ijk} = \beta_{0jk} + \beta_{1jk}x_{1ijk} + \varepsilon_{ijk},$$

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

$$\begin{aligned}\beta_{0jk} &= \beta_{0k} + \beta_{2k}x_{2jk} + u_{0jk}\\ \beta_{1jk} &= \beta_{1k} + \beta_{3k}x_{2jk} + u_{1jk},\end{aligned}$$

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

$$\begin{aligned}\beta_{0k} &= \beta_0 + \beta_4 x_{3k} + u_{3k}\\ \beta_{2k} &= \beta_2 + \beta_5 x_{3k} + u_{4k}\end{aligned}$$

$$\beta_{1k} = \beta_1 + \beta_6 x_{3k} + u_{5k}$$
$$\beta_{3k} = \beta_3 + \beta_7 x_{3k} + u_{6k},$$

where the vectors of level three residuals $(u_{3k}, u_{4k}, u_{5k}, u_{6k})$ are assumed to have a common multivariate normal distribution. In analogy to (1.5), this model includes eight fixed effects parameters capturing the intercept (β_0), the three main effects ($\beta_1, \beta_2, \beta_4$), the three two-way interactions ($\beta_3, \beta_5, \beta_6$), and the three-way interaction of the covariates (β_7) as well as a rich error structure capturing the nested groupings within the data. 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. Alternatively, one can specify the model such that the level three covariate only affects 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.

1.4.2 Non-Nested Groupings

In order to generalize to the case of non-nested groupings, consider data with two different groupings at the second level. In an economics example, imagine modeling the income of individuals in a state who have both an immediate region of residence and an occupation. These workers 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 one 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. Represent region of residence with the index r and occupation with the index o, letting iro refer to the ith individual who is in both the rth region class and the oth occupation class. Of course, such an individual does not necessarily exist—there may be no ranchers in New York City. A regression equation with individual-level covariate x and intercepts which vary with both groupings is given by:

$$y_{iro} = \beta_0 + x_{iro}\beta_1 + u_r + u_o + \varepsilon_{iro}, \quad (1.8)$$

where the random effects u_r have one common normal distribution and the random effects u_o have a different common normal distribution. To add varying slopes to (1.8), simply modify the equation to be

$$y_{iro} = \beta_0 + x_{iro}\beta_1 + u_{0r} + u_{0o} + u_{1r}x_{iro} + u_{1o}x_{iro} + \varepsilon_{iro},$$

and make appropriate distributional assumptions about the random effects.

The addition of a random effect u_{ro} to (1.8) that depends on both region and occupation would give this model 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 affect the outcome variable and then include these in appropriately defined regression equations.

1.4.3 Generalized Linear Forms

The extension to generalized linear models with non-Gaussian distributional assumptions at the individual level 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. For a

two-level model, this means modeling the linked variable as:

$$\eta_{ij} = [\beta_{0j} + x_{1ij}\beta_{1j}] + \cdots + x_{qij}\beta_{qj} + x_{(q+1)ij}\beta_{q+1} + \cdots + x_{(q+p)ij}\beta_{q+p},$$

where this is then related to the conditional mean of the outcome, $\eta_{ij} = g(E[y_{ij}|\boldsymbol{\beta}_j])$, where there is conditioning on the vector of group-level coefficients $\boldsymbol{\beta}_j$. This two-level model is completed by making an assumption about the distribution of $\boldsymbol{\beta}_j$. In contrast to (1.5), the stochastic components of the model at the individual and group levels cannot simply be added together because the link function g is only linear in the case of normally distributed data. Thus, the random effects are difficult or impossible to "integrate out" of the likelihood, and the marginal likelihood of the data cannot be obtained in closed form.

To illustrate this model, consider a simplified logistic case. Suppose there are outcomes from the same binary choice at the individual level, a single individual-level covariate (x_1), and a single group-level covariate (x_2). Then the regression equation for varying-intercepts and varying-slopes is given by:

$$\begin{aligned} p(y_{ij} = 1|\beta_{0j}, \beta_{1j}) &= \text{logit}^{-1}(\beta_{0j} + x_{1ij}\beta_{1j}) \\ &= 1 - \left(1 + e^{\beta_{0j} + x_{1ij}\beta_{1j}}\right)^{-1} \\ \beta_{0j} &= \beta_0 + \beta_2 x_{2j} + u_{0j} \\ \beta_{1j} &= \beta_1 + \beta_3 x_{2j} + u_{1j}. \end{aligned}$$

Assume that the random effects $\mathbf{u}_j$ are multivariate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}$. The parameters to be estimated are the coefficients in the mean structure (the fixed effects), and the elements of the covariance matrix $\boldsymbol{\Omega}$. 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 they provide an easy way to consider the possible correlation of the random effects. However, it is important to understand that the assumption of normality at this level is exactly that, *an assumption*, and thus must be investigated. If evidence is found to suggest that the random effects are not normally distributed, this assumption must be relaxed or fixed effects should be used.

Standard forms for the link function in $g(E[y_{ij}|\boldsymbol{\beta}_j])$ include probit, Poisson (log-linear), gamma, multinomial, ordered categorical forms, and more. The theory and estimation of GLMMs is discussed in Chapter 15, and the specific case of qualitative outcomes is discussed in Chapter 16. Many statistical packages (see Chapter 26) 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.

1.5 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 $\beta_{0j} \sim N(\beta_0 + \beta_1 x_{0j}, \sigma_u^2)$. A related term is *mixed models*, meaning that the specification has both modeled and unmodeled coefficients. The term *fixed effects* is unfortunately not used as cleanly as implied above, with different meanings in different particular 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 control variables, or even in the case where the data represent a population instead of a sample.

The meanings of "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. These specifications can be conceptualized 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.

1.6 CASE STUDY: PARTY IDENTIFICATION IN WESTERN EUROPE

As an illustration, consider 23,355 citizens' 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 to 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 to take into account higher-level 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. If a researcher expects heterogeneity to be explained by dynamics within and between particular elections, the developed model will be hierarchical with two levels based on districts and elections. In Figure 1.4 this is shown by plotting the observed fraction of "Yes" answers for each age, separated by gender (gray dots for men, black dots for women), 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.

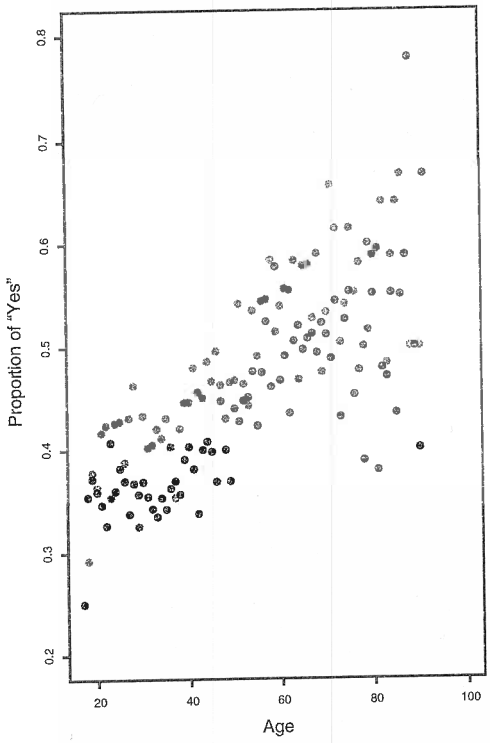

Figure 1.4 Empirical Proportion of "Yes" Versus Age, by Gender

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). Here attention is focused on only a modest subset of the total 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, ignore the nested structure of the data and simply analyze the dataset using a logistic generalized linear model. The outcome variable is modeled as

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

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 are men for `Female`, the first income quantile, and `Rural/Village` for region. Table 1.1 provides this standard logistic model in the first block of results.

For the second model, 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 variable is now modeled as:

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

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, with a difference in AIC (Akaike information criterion) of 525 and a difference in BIC (Bayesian information criterion) of 509 in favor of the multilevel model. This shows that the multilevel model fits the data dramatically better than the GLM. As a description of model fit, also consider 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 1.1 there are essentially no differences in the coefficient estimates between the two models for `Age`, `Female`, `Income Level 4`, `Income Level 5`, and

Table 1.1 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		
				$\sigma_d = 0.2402$, $\sigma_e = 0.32692$		

`Small/Middle Town`. However, notice the differences between the two models for the coefficient estimates of `Income Level 2`, `Income Level 3`, `City Suburb`, `City Metropolitan`, `Parties`, and `Seats`. In all cases where observed differences exist, 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.

To further evaluate the fit of the regular GLM, consider a plot of the ranked fitted probabilities from the model against binned estimates of the observed proportion of success in the reordered observed data. Figure 1.5 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 it is possible to 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, one could 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 taking the binned empirical values and breaking down their variance in various ways. To see that the random effects have a meaningful impact on data fit, 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, as small group effects relative to the residual effect indicate that the grouping is not effective in the model specification. Use of the binned empirical values mimics this kind of analysis for a generalized linear mixed model. The variance of the binned values minus the fixed effects is 0.0111

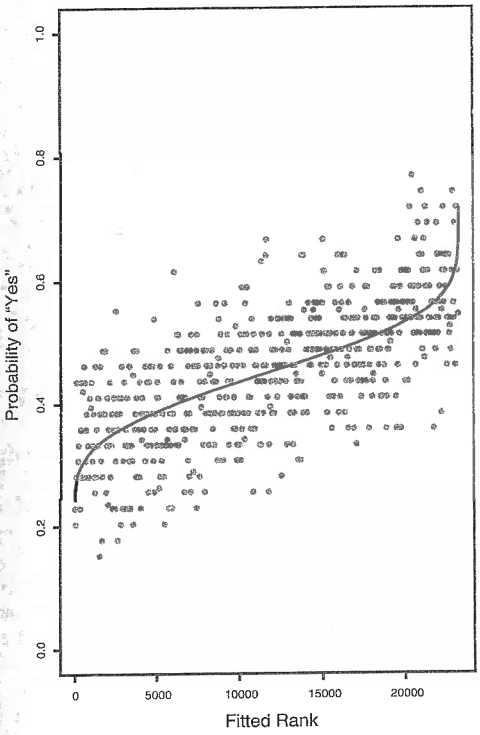

Figure 1.5 Probability of "Yes" Versus Fitted Rank

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 indeed captured by the random effects.

1.6.1 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, it is possible to integrate out all levels of the hierarchy to be left with a likelihood for y which only depends on the fixed effects and the induced covariance structure. Maximum likelihood (or restricted maximum likelihood) estimates for the coefficients on the fixed effects as well as the parameters of the variance components can be computed.

Moving beyond the simple Gaussian–Gaussian model greatly complicates estimation considerations. This includes both relaxations of the assumptions of the linear mixed effects models and complications arising from nonlinear link functions in generalized linear mixed models. A sophisticated set of estimation strategies using approximations to the likelihood have been developed to overcome these difficulties and are discussed in Chapter 3. Alternatively, the Bayesian paradigm offers a suite of MCMC methods for producing samples from the posterior distribution of the model parameters. These alternatives are discussed in Chapter 4.

1.7 SUMMARY

This introduction to multilevel models provides an overview of a class of regression models that account for hierarchical structure in data. Such data occur 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. It provides a general description of the model features that enable multilevel models to account for such structure, demonstrates that ignoring hierarchies produces incorrect inferential statements in model summaries, and has illustrated that point with a simple example using a real dataset.

Aitkin and his co-authors (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 nonlinear 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 taken into account. 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 scientific theories, which routinely consider settings where individual cases are contained in larger groups, which themselves are contained in even larger groups, and so on. Moreover, there are real problems with ignoring hierarchies in data. The resulting models will have the wrong standard errors on group-affected coefficients since fully pooled results assume that the apparent commonalities are results of individual effects. This problem also spills over into covariances between coefficient estimates. Thus, the multilevel modeling framework not only respects the data, but provides better statistical inference which can be used to describe phenomena and inform decisions.

NOTE

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

REFERENCES

Aitkin, M., Anderson, D., and Hinde, J. (1981) 'Statistical Modeling of Data on Teaching Styles', *Journal of the Royal Statistical Society, Series A*, 144: 419–61.

Aitkin, M. and Longford, N. (1986) 'Statistical Modeling Issues in School Effectiveness Studies', *Journal of the Royal Statistical Society, Series A*, 149: 1–43.

Albert, J.H. and Chib, S. (1993) 'Bayesian Analysis of Binary and Polychotomous Response Data', *Journal of the American Statistical Association*, 88: 669–79.

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

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: 360–3.

Baum, L.E. and Petrie, T. (1966) 'Statistical Inference for Probabilistic Functions of Finite Markov Chains', *Annals of Mathematical Statistics*, 37: 1554–63.

Booth, J.G., Casella, G., and Hobert, J.P. (2008) 'Clustering Using Objective Functions and Stochastic Search', *Journal of the Royal Statistical Society, Series B*, 70: 119–39.

Breslow, N.E. and Clayton, D.G. (1993) 'Approximate Inference In Generalized Linear Mixed Models', *Journal of the American Statistical Association*, 88: 9–25.

Bryk, A.S. and Raudenbush, S.W. (1987) 'Applications of Hierarchical Linear Models to Assessing Change', *Psychological Bulletin*, 101: 147–58.

Bryk, A.S., Raudenbush, S.W., Seltzer, M., and Congdon, R. (1988) *An Introduction to HLM: Computer Program and User's Guide*. 2nd edn. 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*, 41: 389–405.

Casella, G. and Berger, R.L. (2001) *Statistical Inference*. 2nd edn. Belmont, CA: Duxbury Advanced Series.

Christiansen, C.L. and Morris, C.N. (1997) 'Hierarchical Poisson Regression Modeling', *Journal of the American Statistical Association*, 92: 618–32.

Cohen, J., Nagin, D., Wallstrom, G., and Wasserman, L. (1998) 'Hierarchical Bayesian Analysis of Arrest Rates', *Journal of the American Statistical Association*, 93: 1260–70.

Cowles, M.K. (2002) 'MCMC Sampler Convergence Rates for Hierarchical Normal Linear Models: A

Simulation Approach', *Statistics and Computing*, 12: 377–89.

Daniels, M.J. and Gatsonis, C. (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 Kreft, I. (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.

Eisenhart, C. (1947) 'The Assumptions Underlying the Analysis of Variance', *Biometrics*, 3: 1–21.

Gelman, A. (2005) 'Analysis of Variance: Why It Is More Important Than Ever', *Annals of Statistics*, 33: 1–53.

Gelman, A. (2006) 'Prior Distributions for Variance Parameters in Hierarchical Models', *Bayesian Analysis*, 1: 515–33.

Gelman, A. and Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Goel, P.K. (1983) 'Information Measures and Bayesian Hierarchical Models', *Journal of the American Statistical Association*, 78: 408–10.

Goel, P.K. and DeGroot, M.H. (1981) 'Information About Hyperparameters in Hierarchical Models', *Journal of the American Statistical Association*, 76: 140–7.

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

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

Hadjicostas, P. and Berry, S.M. (1999) 'Improper and Proper Posteriors with Improper Priors in a Poisson-Gamma Hierarchical Model', *Test*, 8: 147–66.

Hartley, H.O. (1958) 'Maximum Likelihood Estimation From Incomplete Data', *Biometrics*, 14: 174–94.

Healy, M. and Westmacott, M. (1956) 'Missing Values in Experiments Analysed on Automatic Computers', *Journal of the Royal Statistical Society, Series C*, 5: 203–6.

Henderson, C.R. (1950) 'Estimation of Genetic Parameters', *Biometrics*, 6: 186–7.

Henderson, C.R., Kempthorne, O., Searle, S.R., and von Krosigk, C.M. (1959) 'The Estimation of Environmental and Genetic Trends From Records Subject To Culling', *Biometrics*, 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: 367–79.

Jones, G.L. and Hobert, J.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: 127–37.

Laird, N.M. and Ware, J.H. (1982) 'Random-effects Models for Longitudinal Data', *Biometrics*, 38: 963–74.

Lee, V.E. and Bryk, A.S. (1989) 'Multilevel Model of the Social Distribution of High School Achievement', *Sociology of Education*, 62: 172–92.

Lindstrom, M.J. and Bates, D.M. (1988) 'Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data', *Journal of the American Statistical Association*, 83: 1014–22.

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: 958–66.

Longford, N.T. (1987) 'A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models With Nested Random Effects', *Biometrika*, 74: 817–27.

Mason, W.M., Wong, G.Y., and Entwistle, B. (1983) 'Contextual Analysis Through the Multilevel Linear Model', in S. Leinhardt (ed.), *Sociological Methodology 1983–1984*. 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: 1087–91.

Newcomb, S. (1886) 'A Generalized Theory of the Combination of Observations So As to Obtain the Best Results', *American Journal of Mathematics*, 8: 343–66.

Pauler, D.K., Menon, U., McIntosh, M., Symecko, H.L., Skates, S.J., and Jacobs, I.J. (2001) 'Factors Influencing Serum CA125II Levels in Healthy Postmenopausal Women', *Cancer Epidemiology, Biomarkers & Prevention*, 10: 489–93.

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.W. and Bryk, A.S. (1986) 'A Hierarchical Model for Studying School Effects', *Sociology of Education*, 59: 1–17.

Ravishanker, N. and Dey, D.K. (2002) *A First Course In Linear Model Theory*. New York: Chapman & Hall/CRC.

Scheffé, H. (1956) 'Alternative Models for the Analysis of Variance', *Annals of Mathematical Statistics*, 27: 251–71.

Stangl, D.K. (1995) 'Prediction and Decision Making Using Bayesian Hierarchical Models', *Statistics in Medicine*, 14: 2173–90.

Steffey, D. (1992) 'Hierarchical Bayesian Modeling With Elicited Prior Information', *Communications in Statistics–Theory and Methods*, 21: 799–821.

Stiratelli, R., Laird, N., and Ware, J.H. (1984) 'Random-Effects Models for Serial Observations with Binary Response', *Biometrics*, 40: 961–71.

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

Zeger, S.L. and Karim, M.R. (1991) 'Generalized Linear Models With Random Effects: A Gibbs Sampling Approach', *Journal of the American Statistical Association*, 86: 79–86.

Zeger, S.L. and Liang, K.-Y. (1986) 'Longitudinal Data Analysis for Discrete and Continuous Outcomes', *Biometrics*, 42: 121–30.