1 Introduction

Political scientists regularly work with hierarchical or multilevel data and political methodologists have spent a considerable amount of time addressing issues inherent in these kinds of data (c.f. Achen and Shively 1995, Boyd and Iverson 1979, Bartels 1996, Jackson 1992, Sprague 1982, Western 1998). In this chapter, I consider broadly a class of “multilevel models.” The historical lineage of these models go back more than fifty years and emerge from disparate perspectives. In econometrics, models for random coefficients using classical methods were considered in the 1960s and brought into prominence with Swamy’s (1970) pioneering work. From the Bayesian perspective, initial forays into hierarchical models can be found as early as James and Stein (1961). Treating regression coefficients explicitly as dependent variables has a history in econometrics dating back to Saxonhouse (1976, 1977) and in political science to Boyd and Iverson (1979). The emergence of the “multilevel model” (in education research) extends back to the 1980s (c.f. Goldstein 1986) and was brought to prominence with Bryk and Raudenbush (1992). Nevertheless, despite the terminology and techniques, all of these approaches have clear commonalities.
2 Multilevel Data

Multilevel data structures consist of data measured at multiple levels. For example, one may have survey data collected on individuals, indexed as $i$. The sample elements, in turn, may reside in distinct geographical units, indexed as $j$. The $j$ units could be census tracts, counties, states, countries. Data structured in this way are implicitly “hierarchical” insofar as there is a clear nesting of “lower level” units ($i$) within “higher” level units ($j$). The natural hierarchy of this kind of data structure is why models considered in this chapter are sometimes called hierarchical (non)linear models. Multilevel data structures need not be hierarchical (Gelman and Hill 2007), however. Some kinds of cross-classification designs or temporal designs may not be obviously hierarchical, yet models considered herein have been usefully applied to these kinds of data structures. Strict maintenance to the concept of hierarchy is unnecessary, though in the discussion that follows, we will proceed as if data are hierarchically structured.

Commonly, researchers are interested in modeling some outcome variable, $y_i$, as a function of lower and higher level factors. To fix terminology we refer to units of analysis as comprising either “level-1” units or “level-2” units. Covariates measured at level-1 will be notated by $x_i$ and covariates measured at level-2 will be notated by $z_j$. The focus in this chapter will primarily be on two-level models. The number of levels in a hierarchy may, of course, extend beyond 2 levels, though the addition of multiple levels may exacerbate the statistical and interpretative complications of multilevel data structures.

3 Pooling

Imagine an analyst has survey data on an outcome variable of interest as well as data on covariates. Further, suppose these data are collected across countries. In this setting, the country would denote the “$j$ unit” and the respondent would denote the “$i$ unit.” If the
researcher is interested in modeling the relationship between a response variable \( y_i \) and some individual-level covariate, \( x_i \), a garden-variety regression model could be estimated,

\[
y_i = \beta_0 + \beta_1 x_i + \epsilon_i.
\] (1)

If estimated using all \( n \) cases, the researcher is explicitly pooling the data. If the data were cross-sectional, as in the case of examining a single administration of a survey, this modeling strategy is equivalent to stacking each survey from country \( j \). To the extent this modeling strategy is problematic squarely hinges on heterogeneity associated with the \( j \) units (Bartels 1996). Optimistically, this approach proceeds as if cross-unit heterogeneity is nonexistent; each of the \( j \) units are homogenous. Frequently, this assumption will be wrong. In the context of (1), unit heterogeneity may induce nonspherical disturbances. Heteroskedasticity may arise because the \( i \) observations in the \( j \) units are subject to different political conditions or simply because measurement error in \( x_i \) varies across \( j \). In either case, the model is agnostic with respect to heterogeneity and so \( \text{var}(\epsilon) \) is no longer a constant, \( \sigma^2 \), but a variable, \( \sigma^2_i \). Autocorrelation may arise in this kind of model because the \( i \) respondents in unit \( j \) may be more alike—hence correlated—with each other than, say, the \( i \) respondents in some other unit, \( k \). In the multilevel modeling literature, this sort of correlation structure is sometimes referred to as the intraclass correlation.

Therefore, the usual assumption of zero covariance between disturbances, conditional on the covariates, may not hold. Since the assumption of spherical disturbances is a conditional one, these problems can be mitigated by better specification of (1), for example, inclusion of covariates thought to “explain” or account for unit-wise heterogeneity. Absent this, blissful estimation of (1) in the presence of the kinds of problems commonly yields inefficient and inconsistent standard errors. This, in turn, renders the usual kinds of hypothesis tests invalid. Moreover, because the intraclass correlation will usually be positive (Hox 2002, Kreft and
standard errors will usually be attenuated, thus increasing the chances for a Type-I error (Barcikowski 1981).

Of course recognition of the problems posed by data clustering is nothing new. With respect to the problems posed by heteroskedastic errors, a wide variety of solutions have been proposed to “fix-up” the covariance matrix produced by (1). White’s heteroskedastic consistent standard errors (White 1980) is a common solution. White’s work extended Huber’s (1967) results and as such, the variance estimator given by White (1980) is sometimes referred to as the “Huber-White” or “Huber-White-sandwich” estimator. A variety of simulation evidence has shown White’s estimator to be robust to the presence of heteroskedasticity; further, the availability of this correction is now hard coded into virtually all statistical software packages. White’s solution does not account for clustering, however.

To see this, note White’s heteroskedastic consistent variance estimator is given by

\[
\hat{V}_W = (X'X)^{-1} \left[ \sum_{i=1}^{N} (e_i x_i)' (e_i x_i) \right] (X'X)^{-1}. \tag{2}
\]

Under (2), each observation contributes one variance term and so no adjustment is made to account for the grouping of observations within \( j \) units.\(^1\) Fortunately, the foundational results of Huber and White have been extended to account for group-level clustering and thus contemporaneous correlation among observations within clusters.\(^2\) The general form of the variance estimator for clustered data is given by

\[
\hat{V}_C = (X'X)^{-1} \left[ \sum_{j=1}^{n_c} \left\{ \left( \sum_{i=1}^{n_j} e_i x_i \right)' \left( \sum_{i=1}^{n_j} e_i x_i \right) \right\} \right] (X'X)^{-1}, \tag{3}
\]

\(^1\)In most software packages, equation (2) has a multiplier of \( N/N - k \). This multiplier is a finite-sample correction (corrected for lost degrees of freedom due to inclusion of covariates \( k \); Davidson and MacKinnon (1993); see also Franzese (2005) for a extremely clear discussion of these issues.)

\(^2\)Interestingly, there is no “foundational” cite for these extensions. Cragg (1983) considered the case of clustered observations. Froot (1989) built on Cragg’s work and presented two variance estimators for clustered data. Rogers (1993) derived a variance estimator based on the original Huber (1967) estimator. Finally, Williams (2000) gives a general proof the variance estimator for cluster-correlated data is unbiased.
where \( n_c \) corresponds to the number of clusters and \( n_j \) corresponds to the number of \( i \)-cases within unit \( j \). It is evident from (3) that clustering is explicitly accounted for because the cross-products of the residuals and regressors are computed first within the \( j \)th cluster and then summed over the \( n_j \) clusters (see Franzese 2005 for further discussion of this).\(^3\) In this sense, it need not be assumed that observations \textit{within} clusters are independent, though it is implicitly assumed that clusters are independent. The point of this discussion is naive pooling of multilevel data frequently results in nonspherical disturbances. Well understood solutions to this problem are widely available. Therefore, \textit{if} the sole concern with pooling lies with the inducement of nonspherical errors, then the correctives just considered should go far in remedying problems. Yet these correctives do little to exploit information found in multilevel data: here, the hierarchy is a nuisance. Commonly, researchers will be interested in features of the data not easily modeled through the pooling strategy (even with variance corrections) found in (1). This leads to the consideration of multilevel models.

4 Multilevel Models

Implicit in the discussion of the previous section was concern with cross-unit heterogeneity. This sort of heterogeneity might emerge because of unmeasured factors in \( j \). The model in (1) did not explicitly account for any possible heterogeneity associated with \( j \), but the model can be extended to do this. Inclusion of \( j \) dummy variables into (1) yields what is sometimes called a “fixed effects” model. The model, given by

\[
y_i = \beta_{j[i]} + \beta_1 x_i + \epsilon_i,\tag{4}
\]

\(^3\)The multiplier, which is omitted from (3), for the clustered case is usually taken to be \([N/(N-k)]/[J/(J-1)]\) (Franzese 2005).
gives \( j \) intercepts, one for each unit.\(^4\) The parameter \( \beta_1 \) is the same for all \( j \) and so it is clear that the heterogeneity in (4) is treated as a series of offsets. If \( x_i \) is correlated with the fixed effects, then in general \( \hat{\beta}_1 \) from (4) will differ from \( \beta_1 \) in (1). The fixed effects model has been widely applied but it has some clear drawbacks. First, for small \( n \) and large \( j \), the number of degrees of freedom consumed by the fixed effects will be large. Second, covariates that are constants within \( j \) cannot be estimated. Third, the relationship between \( x_i \) and \( y \) may be conditioned by factors measured at level-\( j \); the model in (4) does not directly permit this kind of relationship to exist.

There are a variety of conventional ways to model unit-level relationships. For example, one could estimate separate models for each \( j \) unit and then compare coefficients, perhaps using graphical means (c.f. Bowers and Drake 2005); however, because the kinds of multilevel data political scientists regularly work with are often unbalanced, some of the \( j \) units may have very few individual-level cases compared to others. Thus, parameter estimates for some covariate of interest across the \( j \) models may exhibit wide variability, but this variability emanates from sample size differences and not from anything substantively interesting. In lieu of fixed effects or explicit non-pooling, researchers interested in modeling both level-1 and level-2 covariate effects may include factors measured at either level and then proceed by estimating

\[
y_i = \beta_0 + \beta_1 x_i + \beta_2 z_j + \epsilon_i. \tag{5}
\]

This model has an amalgamation of level-1 (\( x_i \)) and level-2 (\( z_j \)) factors. Built up in this way, however, the researcher can say little about any conditioning effects level-2 factors have on level-1 factors. Unfortunately, this is often a question of central interest. Nevertheless, this kind of conditioning effect can easily be captured in equation (5) through the inclusion of a

\(^4\)For purposes of estimation, the constant term \( \beta_0 \) would need to be suppressed. Doing so permits estimation of \( j \) intercepts. The notation in (4) and in subsequent equations follows the notational conventions used by Gelman and Hill (2007)
multiplicative term:

\[ y_i = \beta_0 + \beta_1 x_i + \beta_2 z_j + \beta_3 x_i z_j + \epsilon_i. \]  

(6)

The “micro” effect is now conditional on the “macro” effect:

\[ E(Y \mid x_i, z_j) = \beta_0 + \beta_2 z_1 + (\beta_1 x_i + \beta_3 z_1) x_1 + \epsilon_i \]  

(7)

(the conditioning effect of the level-1 variable is also retrievable, although with hierarchical data, interest usually centers on the conditioning effect given in the function above).

The model in (6) was considered by Boyd and Iverson (1979) as way to explicitly model contextual relationships. Such relationships are hierarchical in that micro-level factors are conditioned by contextual factors (discussion groups, neighborhood factors, etc.). The conditioning effects given by the interactive model are akin to variable coefficients models; the slope for \( x_1 \) is fully conditioned by \( z_j \). As such, there are as many unique conditional slopes as there are \( j \) level-2 units (and equally the same number of standard errors [Friedrich 1982, Franzese and Kam 2007]).

The problem with model (5), however, is its failure to account for the fact that the overall variance is not only a function of variation among the level-1 units, but also (potentially) variation among the level-2 units. Failure to explicitly account for this renders the “variable coefficient” result shown above problematic: while the micro factor varies with respect to the macro factor, the model assumes this variation is deterministic, not stochastic.

### 4.1 Random Coefficients Models

Consideration of multiple levels of variation may leads us to models with random coefficients. I begin in the context of a simple one-way ANOVA. Consider the following model:

\[ y_i = \beta_{j[i]} + \epsilon_i. \]  

(8)
In contrast to the standard regression model, the intercept is subscripted by group \( j \), implying \( \beta_0 \) is variable over the \( j \) groups. In the standard (unsubscripted) model, a “hard constraint” (Gelman and Hill 2007) is imposed on the \( j \)-unit effects in that they all equal a common \( \beta \). The difference between this model and the “fixed effects” model given in equation (4) is the way \( \beta_{j[i]} \) is treated. Under the fixed effects approach, the unit effects are unknown constants that are estimated from the data (Faraway 2006). In lieu of this approach, it is often useful to consider treating \( \beta_{j[i]} \) as a random coefficient (Faraway, 2006; Gelman 2005). From this perspective, the regression coefficient is assumed to follow some probability distribution, typically, the normal distribution (Gelman and Hill 2007):

\[
\beta_j \sim N(\mu_\beta, \sigma_\beta^2), \quad \text{for } j = 1, \ldots, J. \tag{9}
\]

Treating \( \beta_j \) has having some probability distribution is what Gelman and Hill (2007) call a “soft constraint” on \( j \)-unit variability: the mean effect is estimated but is assumed to have some random variability around it. This variability is attributable to unmeasured factors at level-2. With the distributional assumptions given by (9), the model in (8) is equivalent to a one-way random effects ANOVA (Bryk and Raudenbush, 2002, Faraway 2006, Gelman and Hill 2007). To see this explicitly, suppose we treat \( \beta_j \) as a function of a systematic and random component, as in

\[
\beta_j = \gamma_{00} + u_{0j}, \tag{10}
\]

where \( \gamma_{00} \) gives the mean effect across the sample and \( u_{0j} \) gives the residual of group \( j \) (i.e. the level-2 unit) from the mean. In (10), \( u_{0j} \) accounts for random variation in \( y \) due to level-2 factors: if this term is 0 across units, we might conclude unmeasured level-2 factors effectively induce no extra variance in the model. In contrast, departures from 0 imply unmeasured level-2 factors do induce extra variance. In the regression model given by (8), this term constitutes an additional disturbance term. In the parlance of multilevel modeling, equation
(8) gives the level-1 model and equation (10) gives the level-2 model. The multilevel model is obtained when we substitute (10) into (8), yielding

\[ y_i = \gamma_{00} + u_{0j} + \epsilon_{ij}. \]  

While a multilevel model can be expressed as a two-equation system, the model can be summarized as a single reduced-form equation. The interpretation of the model is straightforward: \( \gamma_{00} \) gives the mean; \( \epsilon_{ij} \) gives the level-1 disturbance; and \( u_{0j} \) gives the level-2 disturbance. The additional error term given by \( u_{0j} \) separates this model from the standard regression model. Because there are two sources of residual variation, a ratio of these two variances can be constructed and is commonly referred to as the intraclass correlation,

\[ \rho = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{\epsilon}^2}. \]  

For two-level models, the ratio in (12) measures the proportion of the total variance that is attributable to level-2 units. Hence, the total variance in the model is

\[ \text{var}(y_{ij}) = \sigma_{u0}^2 + \sigma_{\epsilon}^2. \]

With respect to distributional assumptions on the errors, we typically will assume

\[ \epsilon_{ij} \sim N(0, \sigma_{\epsilon}^2) \]
\[ u_{0j} \sim N(0, \sigma_{u0}^2). \]

4.2 Random Slopes and Group-Level Predictors

Typically, applied workers are interested in the relationship between the response variable and covariates. It is straightforward to extend the basic random ANOVA model. Suppose
there is one level-1 factor and one level-2 factor. The unconditional model would be given by

\[ y_i = \beta_{j[i]} + \beta_{1j[i]}x_{i1} + \epsilon_{ij}, \] (13)

where \( \beta_{1j[i]} \) is the slope coefficient for variable \( x_{i1} \), a level-1 covariate. Further, suppose the constant term, \( \beta_{j[i]} \), randomly varies across units as a function of some level-2 factor \( z_j \). Rewriting the unconditional model to account for this, we obtain

\[
\begin{align*}
\beta_{j[i]} &= \gamma_{00} + \gamma_{01}z_j + u_{0j} \\
\beta_{1j[i]} &= \gamma_{10} + u_{1j}.
\end{align*}
\] (14)

The reduced-form model is given by

\[ y_i = (\gamma_{00} + \gamma_{01}z_j + u_{0j}) + (\gamma_{10} + u_{1j})x_i + \epsilon_i, \]

and rearranging terms gives the multilevel model:

\[ y_i = \gamma_{00} + \gamma_{10}x_i + \gamma_{01}z_j + u_{1j}x_i + u_{0j} + \epsilon_i. \] (15)

In this model, \( \gamma_{00} \) corresponds to the intercept estimate; \( \gamma_{10} \) corresponds to the slope coefficient for the relationship between \( x_i \) and \( y_i \); \( \gamma_{01} \) corresponds to the slope coefficient for the relationship between \( z_j \) and \( y_i \); \( u_{1j} \) corresponds to the disturbance term for the randomly varying slope coefficient \( \gamma_{10} \); \( u_{0j} \) corresponds to the disturbance term for the random intercept term; and \( \epsilon_i \) corresponds to the level-1 disturbance term.

It is useful to contrast this model with a garden variety OLS. Under this model, the \( u_{..} \) are assumed to be zero and so the model reduces to a simple linear model with a mixture of
level-1 and level-2 factors. In standard regression notation, we would posit

\[ y_i = \beta_0 + \beta_1 x_i + \beta_2 z_j + \epsilon_i. \tag{16} \]

It is clear the model in equation (14) is more general than the model in equation (16). Under the multilevel model, the regression parameters are assumed to vary across level-2 units. The variance terms (i.e. the \( \sigma^2_{u} \)) give information on the degree to which between-unit variance exists. With standard regression, these terms are unestimated and are assumed to be exactly 0.

### 4.3 Cross-Level Interactions

The model in (15) assumes the group-level predictor, \( z_j \), has a direct (unconditional) effect on the response variable; however, it is frequently posited that these kinds of factors “moderate” or conditionalize the relationship between an individual-level factor (i.e. \( x_i \)) and \( y_i \). We considered this possibility in the model given by (6). To extend the multilevel model in this way, consider again model (13),

\[ y_i = \beta_{j[i]} + \beta_{1j[i]} x_i + \epsilon_{ij}. \]

Now suppose we hypothesize the bivariate relationship between some covariate and the dependent variable is conditioned by another covariate. This would entail specifying the slope from (13) to be a function of a group-level factor, \( z_j \).

\[
\begin{align*}
\beta_{j[i]} &= \gamma_{00} + \gamma_{01} z_j + u_{0j} \\
\beta_{1j[i]} &= \gamma_{10} + \gamma_{11} z_j + u_{1j}.
\end{align*}
\]  

(17)
The second equation from (17) gives us the change in $\beta_{1j|i}$ as a function of covariate $z_j$. We are explicitly treating the regression slope as conditional on the level-2 factor. The reduced-form model is given by

$$y_i = (\gamma_{00} + \gamma_{01}z_j + u_{0j}) + (\gamma_{10} + \gamma_{11}z_j + u_{1j})x_i + \epsilon_i,$$

and rearranging terms give

$$y_i = \gamma_{00} + \gamma_{10}x_i + \gamma_{01}z_j + \gamma_{11}z_jx_i + u_{1j}x_i + u_{0j} + \epsilon_i.$$  \hfill (18)

In this model, $\gamma_{00}$ corresponds to the intercept estimate; $\gamma_{10}$ corresponds to the slope coefficient for the relationship between $x_i$ and $y_i$ when $z_j = 0$; $\gamma_{01}$ corresponds to the slope coefficient for the relationship between $z_j$ and $y_i$ when $x_i = 0$; $\gamma_{11}$ corresponds to the interaction between $x_i$ and $z_j$; $u_{1j}$ corresponds to the disturbance term for the randomly varying slope coefficient $\gamma_{10}$; $u_{0j}$ corresponds to the disturbance term for the random intercept term; and $\epsilon_i$ corresponds to the level-1 disturbance term. With respect the interaction between $x_i$ and $z_j$, because the two covariates are measured at different levels, this kind of term is sometimes called a “cross-level” interaction. It is important to note that when the covariates included in the cross-level interaction are continuous, it is almost always advisable to mean-center the covariate. This yields an approximate zero point and makes the parameter more naturally interpretable (Raudenbush and Bryk 2002, Faraway 2006, Gelman and Hill 2007, Hox 2002). To complete the model, we will typically assume the following (Goldstein 1995, p. 17):

$$\epsilon_{ij} \sim N(0, \sigma_{\epsilon}^2)$$

$$u_{0j} \sim N(0, \sigma_{u0}^2)$$
\[ u_{1j} \sim N(0, \sigma^2_{u_1}), \]

with \( \text{cov}(u_{0j}, u_{1j}) = \sigma_{u01} \).

## 5 Estimation Issues

In terms of model matrices, the general two-level model (with cross-level interaction) is given by (following Hox’s (2002) presentation)

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
\end{bmatrix} =
\begin{bmatrix}
  1 & x_1 & z_j & z_j x_1 \\
  1 & x_2 & z_j & z_j x_2 \\
  \vdots & \vdots & \vdots & \vdots \\
  1 & x_n & z_j & z_j x_n
\end{bmatrix}
\begin{bmatrix}
  \gamma_{00} \\
  \gamma_{10} \\
  \gamma_{00} \\
  \gamma_{11}
\end{bmatrix} +
\begin{bmatrix}
  1 & x_1 \\
  1 & x_2 \\
  \vdots \\
  1 & x_n
\end{bmatrix}
\begin{bmatrix}
  u_{0j} \\
  u_{1j} \\
  \vdots \\
  \epsilon_n
\end{bmatrix} +
\begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \vdots \\
  \epsilon_n
\end{bmatrix},
\]

and in more succinct form, this yields

\[ y = x\gamma + Wu + \epsilon. \tag{19} \]

In the general case, \( y \) is an \( n \times 1 \) vector for the response variable; \( x \) is an \( n \times k \) design matrix for the model parameters \( \gamma \); \( \gamma \) is an \( k \times 1 \) vector of model parameters; \( W \) is an \( n \times r \) design matrix for the random effects \( u_\cdot \); \( u \) is an \( r \times 1 \) vector of random level-2 disturbances; and \( \epsilon \) is an \( n \times 1 \) vector of level-1 disturbances. The departure of model (19) from (6) is clear: the presence of the random effects complicates estimation since a mixture of parameters, random and fixed, must be estimated. Estimation of the model, then, entails consideration of \( \text{var}(y) = \text{var}(z\gamma) + \text{var}(\epsilon) = \sigma^2(zDz') + \sigma^2(I) \), where \( D \) is a matrix of variance components (Faraway 2006, p. 155). The unconditional distribution of \( y \) in the normal multilevel model
is (equation adapted from Faraway (2006, p. 155) given by

\[ y \sim N(\mathbf{x}\beta, \sigma^2(I + \mathbf{zDz}')) \]

or

\[ y \sim N(\mathbf{x}\beta, \sigma^2V) \]

where \( V = \sigma^2(I + \mathbf{zDz}') \). Since \( V \) will almost always be unknown, it must be estimated. Once \( V \) is estimated, finding estimates of the model parameters, \( \beta_k \) is straightforward using generalized least squares (or some similar kind of function; Goldstein 1995, Longford 1993, Faraway 2006).

Thus estimation of the variance components is obviously an important issue. In the case of continuous \( y \), most statistical software proceed in the classical manner using either full maximum likelihood (FMLE) or, more commonly, restricted maximum likelihood (RMLE) methods for the normal model. The major distinction between RMLE and FMLE is that the former accounts for consumed degrees of freedom while the latter does not. FMLE, as is well known (Patterson and Thompson 1971; Longford 1993; Faraway 2006), gives biased estimates of the variances, though this bias decreases with increased sizes of level-2 units (Raudenbush and Bryk, 2002, Longford 1993). A variety of algorithms have been proposed for FMLE/RMLE. Among them include the EM algorithm (Dempster, Rubin, and Tsutakawa, 1981; see also Raudenbush and Bryk 2002; Longford 1993; Mason, Wong, and Entwistle 1983), iterative generalized least squares for FMLE (Goldstein (1986, 1995), restricted iterative generalized least squares (for RMLE), the Fisher scoring algorithm (proposed by Longford (1987, 1993) for random coefficients models), and Newton-Raphson. As far as choices among these alternatives go, there is no agreed-upon “optimal” way to proceed. In the case of FMLE, Goldstein (1995) and Goldstein and Rasbash (1992) show that Fisher scoring and IGLS are equivalent (for the normal model; see also Kreft and De Leeuw 1998).
In contrast to classical estimation methods, Bayesian methods for multilevel modeling have been widely proposed. Indeed, the literature on Bayesian hierarchical models goes back more than 40 years (c.f. Hartigan 1969, Hill 1965, Lindley and Smith 1972, Tiao and Zellner 1964). In political science, consideration of Bayesian methods generally (c.f. Jackman and Western 1994, Jackman 2000, Quinn et al 1999,) and for hierarchical data in particular, has been widespread (c.f. Gelman and Hill 2007, Gill 2002, Martin 2001). In the Bayesian approach, the unknown parameters, $\beta, \gamma, \sigma$, are assigned a prior distribution. The prior may be informative or uninformative, for example, a uniform prior (much more detailed discussion of priors for hierarchical models can be found in Gelman (2006)). Following Bayes’ theorem, the likelihood is multiplied by the prior distribution to derive estimates of the posterior distribution of the model parameters. In the multilevel models discussed above, level-1 parameters are conditioned by level-2 parameters. To see the Bayesian approach conceptually, Gill (2002, p. 352), succinctly summarizes it as

$$\text{Pr}(\beta, \gamma \mid D) \propto L(\beta \mid D)p(\beta \mid \gamma)p(\gamma).$$

(20)

Gill (2002) notes that the expression in (20) can be expanded to account for more levels in the hierarchy, for example treating $\gamma$ as conditioned by level-3 factors. Adding hierarchies adds substantial complication to the model, however (in either the Bayesian or classical case). Goel and DeGroot (1981; cited in Gill 2002, p. 352) show that the quality of the posterior parameters “decrease at each progressive level of the hierarchical model” (Gill 2002, p. 352); this assertion is formally proved by Goel (1983; cited in Gill 2002, p. 352). Computationally, deriving Bayesian estimates of the model parameters entails specifying conditional distributions for the model parameters, specifying starting values for the parameters, and iterating over values for the model parameters. The iterations produces a chain of simulation draws from the posterior. If several chains based on different starting values are jointly run
(i.e. multiple simulations), the Bayesian model is said to “converge” when the chains have mixed (Gelman and Hill 2007). Simulating chains and mixing, or evaluating convergence, is computationally demanding. Gibbs samplers, or more generally Markov Chain Monte Carlo (MCMC) techniques are usually necessary to derive final parameter estimates (Gelman and Hill 2007, Gill 2002, Jackman 2000). Fortunately, the computing environment BUGS (Spiegelhalter et al 2002) has made MCMC estimation computationally feasible; see also Martin and Quinn’s MCMCpack (2007), a package for the R environment.

A third approach to estimation of the multilevel model differs from the above in that an explicit “two-step” strategy is taken to account for data hierarchy. While this two-step approach was considered by Saxonhouse (1976, 1977), political scientists have recently revisited this strategy (c.f. Duch and Stevenson 2005, Huber et al 2005, Jusko and Shively (2005), Kedar (2005), and Lewis and Linzer (2005)). While the details of estimation vary across the authors just cited, the basic thrust of the two-step strategy is the same. In the first step, \(j\) regressions are estimated as a function of level-1 covariates:

\[
y_{ij} = \beta_{0j} + \beta_{1j}x_i + \epsilon_{ij}.
\]

From the model (21), the regression parameters, \(\hat{\beta}\) are extracted from each model, stacked, and treated as a dependent variable in a model where the covariates are level-2 factors:

\[
\hat{\beta}_j = \gamma_0 + \gamma_1z_j + v_j + u_j.
\]

The dual sources of error in (22) come from that fact that the country-specific parameter will be estimated with uncertainty from the first stage as well as in the second stage (this presentation follows Jusko and Shively 2005; Lewis and Linzer (2005) proceed in a different manner). Jusko and Shively (2005) show this strategy is consistent and efficient (for the normal model) and Achen (2005) notes the two-step approach should be asymptotically
equivalent to the alternative approaches discussed above (see also Saxonhouse 1977). Beck (2005) argues against the two-step approach noting that “one-step” approaches (like the FMLE/RMLE or Bayesian) will never be worse than the two-step strategy. I discuss these issues in the last section.

6 Extensions

The discussion of the multilevel model to this point has centered on models where the response variable, \( y_i \), is continuous; however, consideration of these kinds of models is not limited to the continuous case. Indeed, most software designed for hierarchical data include options for estimating binary or multicategorical response models. In the R environment, Bates and Sarkar’s (2007) \texttt{lme4} package utilizes the general linear modeling (GLM) framework. As such, a wide variety of link functions can be considered for multilevel data. Further, the multilevel framework has been extended to survival models as well as count models (c.f. Barber et al 2000, 2004, Grilli 2005). There is not sufficient space to discuss these developments in detail, though I will discuss the multilevel model in the context of a binary response variable, since such response variables are commonly used in political analysis.

Consider a simple two-level model with binary response variable \( y \) and one level-1 covariate, \( x_i \). The structural model is given by

\[
y_i = \beta_{j[i]} + \beta_{1j[i]} x_{i1}.
\]

If the slope coefficient randomly varies across \( j \) and the intercept is treated as a function of a level-2 factor, the level-2 model is

\[
\beta_{j[i]} = \gamma_{00} + \gamma_{01} z_j + u_{0j} \\
\beta_{1j[i]} = \gamma_{10} + u_{1j}.
\]
The full multilevel model is therefore

$$y_i = \gamma_{00} + \gamma_{10}x_i + \gamma_{01}z_j + u_{1j}x_i + u_{0j}.$$  (23)

If logistic distribution function is specified for $y$, then a familiar logit model is obtained (transformed in log-odds)

$$\log \left[ \frac{\Pr(y_i)}{1 - \Pr(y_i)} \right] = \gamma_{00} + \gamma_{10}x_i + \gamma_{01}z_j + u_{1j}x_i + u_{0j},$$  (24)

with the complication of the random disturbance terms given by $u_{..}$. The model in (24) is a linear model in the logits and so the parameters are on the logit scale (i.e. are log-odds ratios). The model in (24) can be extended in ways similar to the normal model discussed previously and made to account for greater levels of hierarchy. As in the case of the normal, it is typically assumed that

$$u_{0j} \sim N(0, \sigma_{u0}^2)$$

$$u_{1j} \sim N(0, \sigma_{u1}^2).$$

A variety of estimation techniques have been proposed for the multilevel binary response model (Longford 1993 and Rodriguez and Goldman 2001 provide an excellent overview; see also Browne and Draper 2006 or Guo and Zhao 2000). The basic issue involves estimating the unconditional distribution of $y$ by integrating out the random effects $u_{..}$. For example, with a simple random-intercepts model (and no group-level predictors), the conditional density is given by (following Guo and Zhao 2000)

$$f(y_i \mid x_i) = \int_{-\infty}^{\infty} f(y_i \mid x_i, u_j) g(u_j) d(u_j).$$  (25)
The unconditional density, $f(y \mid x_i)$, however, is what we are interested in. With many random effects, (25) becomes a high-dimensional integral. Three general techniques for estimation of the parameters of interest have been proposed. Numerical methods (like Gaussian quadrature or adaptive quadrature) have been proposed to approximate the integral in (25); quasi-likelihood methods have been proposed to approximate the integrand in (25); and Bayesian methods. Quadrature methods approximate the integral over a number of integration points within the domain of the function. As the number of approximation points increase, the computational burden substantially increases. The GLLAMM package (Rabe-Hesketh and Skrondal 2005, Rabe-Hesketh et al, 2004), which can be implemented in Stata utilizes the quadrature approach (note also that the Stata module xtlogit uses Gauss-Hermite quadrature [StataCorp 2005]). Hedeker and Gibbons (1994) provide a nice overview of quadrature as applied to an ordinal logit model.

Quasi-likelihood methods are more commonly used in standard multilevel software programs. These methods focus on the integrand in (25). With respect to this estimation technique, there are two general varieties employed, penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL). The difference in the two methods involves how Taylor series expansions are carried out (Taylor expansions are used to linearize the binary response model). There is not sufficient space to discuss the technical details of these two approximation methods in detail; Longford (1993) and Browne and Draper (2006), among others, provide considerable detail. However, some remarks are in order about implementation. Simulation evidence shows considerable differences can occur in the binary response model between PQL and MQL methods (even greater distinctions can be made if one accounts for the order of the Taylor expansion; see Browne and Draper 2006, Guo and Zhao 2000, or Rodriguez and Goldman 2001). Although there is not widespread agreement here, the general advice to offer is that PQL outperforms MQL. This has been found to be particularly true when the within-cluster correlations are very large (Rodriguez and Goldman 2001).
many software packages, for example R’s \texttt{lme4} and SAS’s \texttt{proc mixed}, the user can specify which technique to use.

With respect to Bayesian methods, the parameters (both fixed and random) are assigned priors and then multiplied by the likelihood to obtain a posterior distribution. The posterior distribution is then sampled from and updated using a Gibbs sampler. In principal, the approach taken here is similar to that followed for the normal model; however, as Browne and Draper (2006) note, “Gibbs sampling in (random effects logit) models is not straightforward.” In their study, they utilize a Gibbs sampler for the variance terms and univariate-update random-walk Metropolis sampling for the fixed effects. The details of either the Gibbs sampler or the Metropolis algorithm are far beyond the scope of this chapter. I recommend Brown (1998), Gelman et al (2003), or Gill (2002) for extensive treatment of these issues. Gelman and Hill (2007) also go into great detail with respect to these approaches and implementation of them in \texttt{BUGS}.

7 Level-1 Coefficients

To this point, the concern has been with estimating the fixed and random effects from the multilevel model. These parameters will frequently be of primary interest; however, it is often useful to examine the level-1 coefficients directly. In the “two-step” approach discussed above, the “level-1” coefficients are explicitly derived from the first stage estimation, since these regressions are separate models for each $j$ unit. The separate-models approach has the virtue of giving unbiased estimates of the level-1 coefficients, although the reliability (or precision) of these estimates will decline as the number of level-1 observations within each $j$ unit decrease.

However, in some instances, the number of level-1 cases may in fact be very small. For example, clustering survey respondents within an areal unit (zip code region, county,
congressional district, etc.) often produces a hierarchical data structure with many $j$ but few $i \in j$. In lieu of the separate regression approach, it is useful to consider *shrink estimators* (James and Stein 1961, Efron and Morris 1975). The basic concept of a shrink estimator in the context of multilevel data, is to treat the level-1 coefficients as a weighted average of, essentially, group means and grand means. In the simple case of the random one-way ANOVA, this implies that if the within-group mean is highly reliable, level-1 estimates of $\beta_{j[i]}$ will be weighted more heavily in terms of the group mean than the grand mean. In the general setting with random slopes and intercepts, the degree to which within-unit information gives reliable estimates of the level-1 parameters of interest, these coefficients will be weighted more heavily in terms of the cluster-level data. When cluster-level data yield low-reliability estimates, level-1 estimates “borrow” information from the full sample. Because of this, Morris (1983) referred to these kinds of estimators as “compromise estimator(s).” The shrink estimator used to derive level-1 coefficients in multilevel models is typically an empirical Bayes (EB) estimator (Carlin and Louis 1996; Efron and Morris 1975; Morris 1983).

The basic concept behind shrink estimators is nicely summarized in Gelman and Hill (2007, p. 476) who show that

$$
\beta_{j[i]} = \omega_j \mu_\beta + (1 - \omega_j) \bar{y}_j,
$$

(26)

where

$$
\omega_j = 1 - \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_y/n_j}.
$$

Gelman and Hill (2007) refer to this expression as a “pooling factor.” It is clear why. When there is “no pooling,” the estimate of $\beta_{j[i]}$ will equal the group mean; when there is complete pooling, the group-level coefficient is indistinguishable from the grand mean. Similarly, the variance of the parameter will be affected by the pooling factor in that (following Gelman
and Hill 2007, p. 477, eq. 21.11)
\[
\text{var}(\beta_j) = (1 - \omega_j)\sigma^2_y/n_j.
\]

This again shows how uncertainty around the level-1 parameter relates to the separation among the \(j\)-level units: “borrowing strength” is directly connected to the degree to which the data exhibit pooling.

8 Recommendations and Conclusion

The enterprise of multilevel modeling is far too vast to give adequate coverage here. However, a few concluding remarks and recommendations are in order. Unlike many of the traditional econometric models applied by political scientists where estimation properties are well understood (and may be robust under adverse conditions), the same cannot be said for many of the estimation choices available here. This is particularly true with respect to hierarchical non-linear models. Colossally different estimates may be obtained dependent merely on the choice of estimation technique. Browne and Draper (2006), Guo and Zhao (2000), and Rodriguez and Goldman (2001) demonstrate this with respect to multilevel binary response models. Because of the increasing ease to which software packages allow us to estimate complicated models, failure to take note of “software defaults” can have deleterious consequences. “Deleterious” in this case means one may get fundamentally different parameter estimates contingent on: a) what the default choice was versus some alternative (this is especially true with PQL and MQL) and b) which software program/computing environment was actually used. It seems clear that “good practice” with respect to these models will require users to have a “better-than-usual” understanding of what is being estimated and how it is being estimated.
Nevertheless, if appropriately applied (or even approximately appropriately applied),
these models yield considerable information about model parameters (fixed effects), level-wise
variance, and micro-level parameters. Unfortunately, much applied work using multilevel
models are often presented as if this extra information was never estimated. Minimally, if
random coefficients are posited, the estimated level-2 variances should be presented along
with model parameters. Frequently, this is not done. In principal, these terms convey
substantive information about variability around the model parameters. “Insignificance” of
a level-2 variance term, on its own, may be of some interest. Further, discussion and/or
interpretation of level-1 coefficients would seem to be of interest. After all, variability across
j is one reason multilevel models are considered. Gelman and Hill (2007) provide outstanding
coverage of ways to use level-1 information.

Indeed, the explicit consideration of micro-level effects makes the “two-step” strategy
discussed above attractive; however, it should be clear that any of the estimation techniques
permit estimation of the micro-level parameters. With respect to the “two-step” approach
versus its alternatives (FMLE/RMLE, MCMC/Bayes, quadrature), then, I agree with Beck’s
(2005) conclusion that there is not much leverage gained from taking the two-step approach
(outside of, perhaps, initial model diagnostics). Achen (2005) makes a compelling case in
favor of the two-step approach, noting that political science hierarchical data frequently do
not mimic the students-in-classrooms data used by education researchers. That is, at least
with comparative politics data, sample sizes are often very large and disparate (see also
Bartels 1996) and so concern with “borrowing strength” (or shrink estimators) will not have
the same import in this context as in, say, the education research context. Nevertheless,
the two-step approach automatically assumes each level-1 coefficient is explicitly non-pooled
(in equation (26), \( \omega = 0 \)), which seems to be a strong assumption, one testable inside
what Gelman (2005) calls a “full multilevel model.” Given the fact that FMLE/RMLE or
MCMC/Bayes techniques are getting easier to implement (easy being a relative term) in
Stata, R or BUGS, it is not clear how useful the two-stage strategy will ultimately become.

To conclude, this chapter has attempted to convey the scope of multilevel models but in a very compact way. Many issues were touched upon but not covered in great detail, for example, extension of the GLM framework to hierarchical nonlinear models is an extremely important growth area but was only briefly referenced in the discussion of the binary response model. Also, I did not provide much detailed discussion about modeling decisions involving centering (Hoffman and Gavin 1998) and conditional versus unconditional shrink estimators (Raudenbush and Bryk 2002), although these are clearly important issues. Nevertheless, carefully applied, multilevel models offer great promise for exploiting information in hierarchical data structures. There are a range of alternatives for such data and it bears repeating, that sometimes, simpler-to-apply correctives like those discussed in Section 3 will suffice. Often they will not, however, and so extensions of the basic forms given in Section 4 and 6 of this chapter will need to be considered.

References


Duch, Raymond M. and Randy Stevenson. “Context and the Economic Vote: A Multilevel


Sprague, John. 1982. “Is there a micro theory consistent with contextual analysis?” In
38: 311–323.
Tiao, George C. and Arnold Zellner. 1964. “Bayes’s Theorem and the Use of Prior Knowledge in
Western, Bruce. 1998. “Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical
White, H. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for
Biometrics. 56: 645-646.