Chapter 5. Multilevel Models

Abstract. This chapter describes a conditional modeling framework that takes into account hierarchical and clustered data structures. The data and models, known as multilevel, are used extensively in educational science and related disciplines in the social and behavioral sciences. We show that a multilevel model can be viewed as a linear mixed effects model and hence, the statistical inference techniques introduced in Chapter 3 are readily applicable. By considering multilevel data and models as a separate unit, we expand the breadth of applications that linear mixed effects models enjoy.

5.1 Cross-sectional multilevel models

Educational systems are often described by structures in which the units of observation at one level are grouped within units at a higher level of structure. To illustrate, suppose that we are interested in assessing student performance. Students are grouped in classes, classes are grouped in schools and schools are grouped into districts. At each level, there are variables that may affect responses from a student. For example, at the class level, education of the teacher may be important, at the school level, the school size may be important, and at the district level, funding may be important. Further, each level of grouping may be of scientific interest. Finally, there may be not only relationships among variables within each group but also across groups that should be considered.

The term multilevel is used for this nested data structure. In the above situation, we consider students to be the basic unit of observation; they are known as the “level-1” units of observation. The next level up is called “level-2” (classes in this example), and so forth.

We can imagine multilevel data being collected by a cluster sampling scheme. A random sample of districts is identified. For each district selected, a random sample of schools is chosen. From each school, a random sample of classes is taken and from each class selected, a random sample of students. Mechanisms other than random sampling may be used, and this will influence the model selected to represent the data. Multilevel models are specified through conditional relationships, where the relationships described at one level are conditional on (generally unobserved) random coefficients of lower levels. Because of this conditional modeling framework, multilevel data and models are also known as hierarchical.

5.1.1 Two-level models

To illustrate the important features of the model, initially consider only two levels. Suppose that we have a sample of \( n \) schools and, for the \( i \)th school, we randomly select \( n_i \) students (omitting class for the moment). For the \( j \)th student in the \( i \)th school, we assess the student’s performance on an achievement test, \( y_{ij} \), and information on the student’s socio-economic status, \( z_{ij} \), for example, the total family income. To assess achievement in terms of socio-economic status, we could begin with a simple model of the form

\[
y_{ij} = \beta_{0i} + \beta_{1i} z_{ij} + \epsilon_{ij}. \tag{5.1}
\]

Equation (5.1) describes a linear relation between socio-economic status and expected performance, although we allow the linear relationship to vary by school through the notation \( \beta_{0i} \).
and $\beta_{1i}$ for school-specific intercepts and slopes. Equation (5.1) summarizes the “level-1” model that concerns student performance as the unit of observation.

If we have identified a set of schools that are of interest, then we may simply think of the quantities $\{\beta_{0i}, \beta_{1i}\}$ as fixed parameters of interest. However, in educational research, it is customary to consider these schools to be a sample from a larger population; the interest is in making statements about this larger population. Thinking of the schools as a random sample, we model $\{\beta_{0i}, \beta_{1i}\}$ as random quantities. A simple representation for these quantities is:

$$\beta_{0i} = \beta_0 + a_{0i} \quad \text{and} \quad \beta_{1i} = \beta_1 + a_{1i}, \quad (5.2)$$

where $a_{0i}, a_{1i}$ are mean zero random variables. Display (5.2) represents a relationship about the schools and summarizes the “level-2” model.

Displays (5.1) and (5.2) describe models at two levels. For estimation, we combine (5.1) and (5.2) to yield

$$y_{ij} = (\beta_0 + a_{0i}) + (\beta_1 + a_{1i}) z_{ij} + \epsilon_{ij} = a_{0i} + a_{1i} z_{ij} + \beta_0 + \beta_1 z_{ij} + \epsilon_{ij}. \quad (5.3)$$

Equation (5.3) shows that the two-level model may be written as a single linear mixed effects model. Specifically, we define $a_i = (a_{0i}, a_{1i})'$, $z_{ij} = (1, z_{ij})'$, $\beta = (\beta_0, \beta_1)'$ and $x_{ij} = z_{ij}$, to write

$$y_{ij} = z_{ij}' a_i + x_{ij}' \beta + \epsilon_{ij},$$

similar to equation (3.5). Because we can write the combined multilevel model as a linear mixed effects model, we can use the Chapter 3 techniques to estimate the model parameters. Note that we are now using the subscript “*” to denote replications within a subject. This is because we interpret the replication to have no time ordering; generally we will assume no correlation among replications (conditional on the subject). Section 5.2 will re-introduce the “*” subscript when we consider time-ordered repeated measurements.

One desirable aspect of the multilevel model formulation is that we may modify conditional relationships at each level of the model, depending on the research interests of the study. To illustrate, we may wish to understand how characteristics of the school affect student performance. For example, Raudenbush and Bryk (2002EP) discussed an example where $x_i$ indicates whether the school was a Catholic based or a public school. A simple way to introduce this information is to modify the level-2 model in display (5.2) to

$$\beta_{0i} = \beta_0 + \beta_{01} x_i + a_{0i} \quad \text{and} \quad \beta_{1i} = \beta_1 + \beta_{11} x_i + a_{1i}. \quad (5.2a)$$

There are two level-2 regression models in display (5.2a); analysts find it intuitively appealing to specify regression relationships that capture additional model variability. Note, however, that for each model, the left-hand side quantities are not observed. To emphasize this, Raudenbush and Bryk (2002EP) call these models “intercepts-as-outcomes” and “slopes-as-outcomes.” In Section 5.3, we will learn how to predict these quantities.

Combining display (5.2a) with the level-1 model in equation (5.1), we have

$$y_{ij} = (\beta_0 + \beta_{01} x_i + a_{0i}) + (\beta_1 + \beta_{11} x_i + a_{1i}) z_{ij} + \epsilon_{ij} = a_{0i} + a_{1i} z_{ij} + \beta_0 + \beta_{01} x_i + \beta_1 z_{ij} + \beta_{11} x_i z_{ij} + \epsilon_{ij}. \quad (5.4)$$

By defining $a_i = (a_{0i}, a_{1i})'$, $z_{ij} = (1, z_{ij})'$, $\beta = (\beta_0, \beta_{01}, \beta_1, \beta_{11})'$ and $x_{ij} = (1, x_i, z_{ij}, x_i z_{ij})'$, we may again express this multilevel model as a single linear mixed effects model.

The term $\beta_{11} x_i z_{ij}$, interacting between the level-1 variable $z_{ij}$ and the level-2 variable $x_i$, is known as a cross-level interaction. For this example, suppose that we use $x = 1$ for Catholic schools and $x = 0$ for public schools. Then, $\beta_{11}$ represents the difference between the marginal change in achievement scores, per unit of family income, between Catholic and public schools.
Many researchers (see, for example, Raudenbusch and Bryk, 2002EP) argue that understanding cross-level interactions is a major motivation for analyzing multilevel data.

**Centering of variables**

It is customary in educational science to “center” explanatory variables in order to enhance the interpretability of model coefficients. To illustrate, consider the hierarchical models in (5.1), (5.2a) and (5.4). Using the “natural” metric for $z_{ij}$, we interpret $\beta_0$ to be the mean (conditional on the $i$th subject) response when $z = 0$. In many applications such as where $z$ represents total income or test scores, a value of zero falls outside a meaningful range of values.

One possibility is to center level-1 explanatory variables about their overall mean and use $z_{ij} - \bar{z}_i$ as an explanatory variable in equation (5.1). In this case, we may interpret the intercept $\beta_{0,i}$ to be the expected response for an individual with a score equal to the grand mean. This can be interpreted as an adjusted mean for the $i$th group.

Another possibility is to center each level-1 explanatory variable about its level-2 mean and use $z_{ij} - \bar{z}_i$ as an explanatory variable in equation (5.1). In this case, we may interpret the intercept $\beta_{0,i}$ to be the expected response for an individual with a score equal to the mean of the $i$th group.

For longitudinal applications, you may wish to center the level-1 explanatory variables so that the intercept equals the expected random coefficient at a specific point in time, for example, at the start of a training program.

For level-2 explanatory variables, one generally centers on grand means, if centering is done at all (see, for example, van der Leeden, 1998EP).

**Extended two-level models**

To consider many explanatory variables, we extend equations (5.1) and (5.2). Consider a level-1 model of the form

$$y_{ij} = z_{1,ij} \beta_i + x_{1,ij} \beta_1 + \epsilon_{ij}. \quad (5.5)$$

Here, $z_{1,ij}$ and $x_{1,ij}$ represent the set of level-1 variables associated with varying (over level-1) and fixed coefficients, respectively. The level-2 model is of the form:

$$\beta_i = X_{2,i} \beta_2 + \alpha_i, \quad (5.6)$$

where $E \alpha_i = 0$. With this notation, the term $X_{2,i} \beta_2$ forms another set of effects with parameters to be estimated. Alternatively, we could write equation (5.5) without explicitly recognizing the fixed coefficients $\beta_1$ by including them in the random coefficients equation (5.6) but with zero variance. However, we prefer to recognize their presence explicitly because this helps in translating equations (5.5) and (5.6) into computer statistical routines for implementation. Combining equations (5.5) and (5.6) yields

$$y_{ij} = z_{1,ij} (X_{2,i} \beta_2 + \alpha_i) + x_{1,ij} \beta_1 + \epsilon_{ij}$$

$$= z_{ij} \alpha_i + x_{ij} \beta + \epsilon_{ij}, \quad (5.7)$$

with the notation $x_{ij} = (x_{1,ij}', z_{1,ij}'X_{2,i})$, $z_{ij} = z_{1,ij}$, and $\beta = (\beta_1, \beta_2)'$. Again, equation (5.7) expresses this multilevel model in our usual linear mixed effects model form.

It will be helpful to consider a number of special cases of equations (5.5)-(5.7). To begin, suppose that $\beta_i$ is a scalar and that $z_{1,ij} = 1$. Then, the model in equation (5.7) reduces to the error components model introduced in Section 3.1. Raudenbush and Bryk (2002EP) discuss the further special case, where equation (5.5) does not contain the fixed effects $x_{1,ij}' \beta_1$ portion. In this case, equation (5.7) reduces to

$$y_{ij} = \alpha_i + X_{2,i} \beta_2 + \epsilon_{ij},$$
that Raudenbush and Bryk refer to as the “means-as-outcomes” model. This model, with only level-2 explanatory variables available, can be used to predicts the means, or expected values, of each group \( i \). We will study this prediction problem formally in Section 5.3.

Another special case of equations (5.5)-(5.7) is the random coefficients model. Here, we omit variables the level-1 fixed effects portion \( x_{1,i,j} \) and use the identity matrix for \( X_{2,i} \). Then, equation (5.7) reduces to

\[
y_{ij} = z_{ij}' (\beta_2 + \alpha_i) + \varepsilon_{ij}.
\]

**Example**

As reported in Lee (2000EP), Lee and Smith (1997EP) studied 9,812 Grade 12 students in 1992 who attended 789 public, Catholic, and elite private high schools, drawn from a nationally representative sample from the National Education Longitudinal Study. The responses were achievement gains in reading and mathematics over four years of high school. The main variable of interest was a school level variable, size of the high school. Educational research had emphasized that larger schools enjoy economies of scale and are able to offer a broader curriculum whereas smaller schools offer more positive social environments, as well as a more homogenous curriculum. Lee and Smith sought to investigate the optimal school size. To control for additional student level effects, level-1 explanatory variables included gender, minority status, ability and socio-economic status. To control for additional school level characteristics, level-2 explanatory variables included school average minority concentration, school average socio-economic status and type of school (Catholic, public and elite private). Lee and Smith found that a middle school size, of approximately 600-900 students, produced the best achievement results.

**Motivation for multilevel models**

As we have seen, multilevel models allow analysts to assess the importance of cross-level effects. Specifically, the multilevel approach allows and/or forces researchers to hypothesize relationships at each level of analysis. Many different “units of analysis” within the same problem are possible, thus permitting modeling of complex systems. The ability to estimate cross-level effects is one advantage of multilevel modeling when compared to an alternate research strategy calling for the analysis of each level in isolation of the others.

As described in the introductory Chapter 1, multilevel models allow analysts to address problems of heterogeneity with samples of repeated measurements. Within the educational research literature, not accounting for heterogeneity from individuals is known as aggregation bias; see for example, Raudenbush and Bryk (2002EP). Even if the interest in understanding level-2 relationships, we will get a better picture by incorporating a level-1 model of individual effects. Moreover, multilevel modeling allows us to predict quantities at both level-1 and level-2; Section 5.3 describes this prediction problem.

Second and higher levels of multilevel models also provide us with an opportunity to estimate the variance structure using a parsimonious, parametric structure. Improved estimation of the variance structure provides a better understanding of the entire model and will often result in improved precision of our usual regression coefficient estimators. Moreover, as discussed above, often these relationships at the second and higher levels are of theoretical interest and may represent the main focus of the study. However, technical difficulties arise when testing certain hypotheses about variance components. These difficulties, and solutions, are presented in Section 5.4.
5.1.2 Multiple level models

Extensions to more than two levels follow the same pattern as two level models. To be explicit, we give a three level model based on an example from Raudenbush and Bryk (2002EP). Consider modeling a student’s achievement as the response \( y \). The level-1 model is

\[
y_{i,j,k} = z_{i,j,k} \beta_{i,j} + x_{i,j,k} \beta_1 + \epsilon_{1,i,j,k},
\]

(5.8)

where there are \( i = 1, \ldots, n \) schools, \( j = 1, \ldots, J_i \) classrooms in the \( i \)th school and \( k = 1, \ldots, K_{i,j} \) students in the \( j \)th classroom (within the \( i \)th school). The explanatory variables \( z_{i,j,k} \) and \( x_{i,j,k} \) may depend on the student (gender, family income and so on), classroom (teacher characteristics, classroom facilities and so on) or school (organization, structure, location and so on). The parameters that depend on either school \( i \) or classroom \( j \) appear as part of the \( \beta_j \) vector whereas parameters that are constant appear in the form \( \beta_{1,i,j,k} \). The dependence is made explicit in the higher-level model formulation. Conditional on the classroom and school, the disturbance term \( \epsilon_{1,i,j,k} \) is mean zero and has a variance that is constant over all students, classrooms and schools.

The level-2 model describes the variability at the classroom level. The level-2 model is of the form

\[
\beta_{i,j} = Z_{2,i,j} \gamma_i + X_{2,i,j} \beta_2 + \epsilon_{2,i,j},
\]

(5.9)

Analogous to level-1, the explanatory variables \( Z_{2,i,j} \) and \( X_{2,i,j} \) may depend on the classroom or school but not the student. The parameters associated with the \( Z_{2,i,j} \) explanatory variables, \( \gamma_i \), may depend on school \( i \) whereas the parameters associated with the \( X_{2,i,j} \) explanatory variables are constant. Conditional on the school, the disturbance term \( \epsilon_{2,i,j} \) is mean zero and has a variance that is constant over classrooms and schools. The level-1 parameters \( \beta_{i,j} \) may be (i) varying but nonstochastic or (ii) stochastic. With this notation, we use a zero variance to model parameters that are varying but nonstochastic.

The level-3 model describes the variability at the school level. Again, the level-2 parameters \( \gamma_i \) may be varying but nonstochastic or stochastic. The level-3 model is of the form

\[
\gamma_i = X_{3,i} \beta_3 + \epsilon_{3,i}.
\]

(5.10)

Again, the explanatory variables \( X_{3,i} \) may depend on the school. Conditional on the school, the disturbance term \( \epsilon_{3,i} \) is mean zero and has a variance that is constant over schools.

Putting equations (5.8)-(5.10) together, we have

\[
y_{i,j,k} = z_{i,j,k} \beta_{i,j} + x_{i,j,k} \beta_1 + \epsilon_{1,i,j,k} = x_{i,j,k} \beta_1 + z_{i,j,k} \beta_2 + \epsilon_{1,i,j,k},
\]

(5.11)

where \( x_{i,j,k} = (x_{1,i,j,k} \beta_1 + x_{2,i,j,k} \beta_2 + z_{i,j,k} \epsilon_{1,i,j,k} \beta_1 + \epsilon_{2,i,j,k} \epsilon_{1,i,j,k} \) \) and \( \epsilon_{2,i,j,k} = (z_{i,j,k} \epsilon_{1,i,j,k} \beta_1 + \epsilon_{2,i,j,k} \epsilon_{1,i,j,k} \) \). We have already specified the usual assumption of homoscedasticity for each random quantity \( \epsilon_{1,i,j,k} \), \( \epsilon_{2,i,j,k} \) and \( \epsilon_{3,i} \). Moreover, it is customary to assume that these quantities are uncorrelated with one another. Our main point is that, as with the two-level model, equation (5.11) expresses the three-level model as a linear mixed effects model. (Converting the model in equation (5.11) into the linear mixed effects model in equation (3.5) is a matter of defining vector expressions carefully. Section 5.3 provides further details.) Thus, parameter estimation is a direct consequence of our Chapter 3 results. Many variations of the basic assumptions that we have described are possible. In Section 5.2 on longitudinal multilevel models, we will give a more detailed description of an example of a three-level model. Appendix 5A extends the discussion to higher order multilevel models.
For applications, several statistical software packages exist (such as HLM, MlwiN, and MIXREG) that allow analysts to fit multilevel models without combining the several equations into a single expression such as equation (5.11). However, these specialized packages may not have all of the features that the analyst wishes to display in his or her analysis. As pointed out by Singer (1998EP), an alternative, or supplementary, approach is to use a general purpose mixed linear effects package (such as SAS PROC MIXED) and rely directly on the fundamental mixed linear model theory.

5.1.3 Multilevel modeling in other fields
The field of educational research has been an area of active development of cross-sectional multilevel modeling although it by no means has a corner on the market. This subsection describes examples where these models have been used in other fields of study.

One type of study that is popular in economics is data based on a matched pairs sample. For example, we might select a set of families for level-2 sample and, for each family, observe the behavior of siblings (or twins). The idea underlying this design is that by observing more than one family member we will be able to control for unobserved family characteristics. See Wooldridge (2002E) and Exercise 3.10 for further discussion of this design.

In insurance and actuarial science, it is possible to model claims distributions using a hierarchical framework. Typically, the level-2 unit of analysis is based on an insurance customer, and explanatory variables may include characteristics of the customer. The level-1 model uses claims amounts as the response (typically over time) and typical time-varying explanatory variables include time trends. For example, Klugman (1992O) gives a Bayesian perspective of this problem. For a frequentist perspective, see Frees, Young and Luo (1999O).

5.2 Longitudinal multilevel models
This section shows how to use the conditional modeling framework to represent longitudinal (time-ordered) data. The key change in the modeling set-up is that we now will typically consider the individual as the level-2 unit of analysis and observations at different time points as the level-1 units. The goal is now also substantially different; typically, in longitudinal studies the assessment of change is the key research interest. As with Section 5.1, we begin with the two-level model and then discuss general multilevel extensions.

5.2.1 Two-level models
Following the notation established in Section 5.1, we consider level-1 models of the form

$$y_{it} = z_{1, it} \beta_i + x_{1, it} \beta_1 + \epsilon_{it}.$$  (5.12)

This is a model of \(t = 1, \ldots, T_i\) responses over time for the \(i\)th individual. The unit of analysis for the level-1 model is an observation at a point in time, not the individual as in Section 5.1. Thus, we use the subscript “\(t\)” as an index for time. Most other aspects of the model are as in Section 5.1.1; \(z_{1, it}\) and \(x_{1, it}\) represent sets of level-1 explanatory variables. The associated parameters that may depend on the \(i\)th individual appear as part of the \(\beta_i\) vector whereas parameters that are constant appear in the \(\beta_1\) vector. Conditional on the subject, the disturbance term \(\epsilon_{it}\) is mean zero random variable that is uncorrelated with \(\beta_i\).

An important feature of the longitudinal multilevel model that distinguishes it from its cross-sectional counterpart is that time generally enters the level-1 specification. There are a number of ways that this can happen. One way is to let one or more of the explanatory variables be a function of time. This is the approach historically taken in growth curve modeling, described below. Another approach is to let one of the explanatory variables be a lagged response variable.
This approach is particularly prevalent in economics and will be further explored in Chapter 6. Yet another approach is to model the serial correlation through the variance covariance matrix of the vector of disturbance $\mathbf{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$. Specifically, in Sections 2.5.1 and 3.3.1 we developed the notation $\text{Var} \mathbf{\varepsilon}_i = \mathbf{R}_i$ to represent the serial covariance structure. This approach is widely adopted in biostatistics and educational research and will be further developed here.

Like the cross-sectional model, the level-2 model can be represented as $\beta_i = \mathbf{X}_2\beta + \alpha_i$; see equation (5.6). Now, however, we interpret the unobserved $\beta_i$ to be the random coefficients associated with the $i$th individual. Thus, although the mathematical representation is similar to the cross-sectional setting, our interpretations of individual components of the model are quite different. Yet, as with equation (5.7), we may still combine level-1 and level-2 models to get

$$y_{it} = \mathbf{z}_{1, it}'(\mathbf{X}_{2i}\beta + \alpha_i) + \mathbf{x}_{1, it}'\beta_1 + \varepsilon_{it}$$

$$= \mathbf{z}_{1, it}'\alpha_i + \mathbf{x}_{1, it}'\beta + \varepsilon_{it},$$

(5.13)

using the notation $\mathbf{x}_{it} = (\mathbf{x}_{1, it}', \mathbf{z}_{1, it}'\mathbf{X}_{2i})$, $\mathbf{z}_{it} = \mathbf{z}_{1, it}$ and $\beta = (\beta_1', \beta_2')'$. This is the linear mixed effects model introduced in Section 3.3.1.

**Growth curve models**

To develop intuition, we now consider growth curve models, models that have a long history of applications. The idea behind growth curve models is that we seek to monitor the natural development or aging of an individual. This development is typically monitored without intervention and the goal is to assess differences among groups. In growth curve modeling, one uses a polynomial function of age or time to track growth. Because growth curve data may reflect observations from a development process, it is intuitively appealing to think of the expected response as a function of time. Parameters of the function vary by individual, so that one can summarize an individual’s growth through the parameters. To illustrate, we now consider a classic example.

**Example - Dental Data**

This example is originally due to Potthoff and Roy (1964B); see also Rao (1987B). Here, $y$ is the distance, measured in millimeters, from the center of the pituitary to the pteryomaxillary fissure. Measurements were taken on 11 girls and 16 boys at ages 8, 10, 12, and 14. The interest is in the relation between the distance and age, specifically, in how the distance grows with age and whether there is a difference between males and females.

Table 5.1 shows the data and Figure 5.1 gives a graphical impression of the growth over time. From Figure 5.1, we can see that the measurement length grows as each child ages, although it is difficult to detect differences between boys and girls. In Figure 5.1, we use open circular plotting symbols for girls and opaque plotting symbols for boys. Figure 5.1 does show that the ninth boy has an unusual growth pattern; this pattern can also be seen in Table 5.1.
Table 5.1 Dental measurements of 11 girls and 16 boys. Measurements are in millimeters.

<table>
<thead>
<tr>
<th>Number</th>
<th>Age in years</th>
<th></th>
<th></th>
<th></th>
<th>Age in years</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>20</td>
<td>21.5</td>
<td>23</td>
<td>26</td>
<td>25</td>
<td>29</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>21.5</td>
<td>24</td>
<td>25.5</td>
<td>21.5</td>
<td>22.5</td>
<td>23</td>
<td>26.5</td>
</tr>
<tr>
<td>3</td>
<td>20.5</td>
<td>24</td>
<td>24.5</td>
<td>26</td>
<td>23</td>
<td>22.5</td>
<td>24</td>
<td>27.5</td>
</tr>
<tr>
<td>4</td>
<td>23.5</td>
<td>24.5</td>
<td>25</td>
<td>26.5</td>
<td>25.5</td>
<td>27.5</td>
<td>26.5</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>21.5</td>
<td>23</td>
<td>22.5</td>
<td>23.5</td>
<td>20</td>
<td>23.5</td>
<td>22.5</td>
<td>26</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>21</td>
<td>21</td>
<td>22.5</td>
<td>24.5</td>
<td>25.5</td>
<td>27</td>
<td>28.5</td>
</tr>
<tr>
<td>7</td>
<td>21.5</td>
<td>22.5</td>
<td>23</td>
<td>25</td>
<td>22</td>
<td>22</td>
<td>24.5</td>
<td>26.5</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>23</td>
<td>23.5</td>
<td>24</td>
<td>24</td>
<td>21.5</td>
<td>24.5</td>
<td>25.5</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>21.5</td>
<td>23</td>
<td>20.5</td>
<td>31</td>
<td>26</td>
</tr>
<tr>
<td>10</td>
<td>16.5</td>
<td>19</td>
<td>19</td>
<td>19.5</td>
<td>27.5</td>
<td>28</td>
<td>31</td>
<td>31.5</td>
</tr>
<tr>
<td>11</td>
<td>24.5</td>
<td>25</td>
<td>28</td>
<td>28</td>
<td>23</td>
<td>23</td>
<td>23.5</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21.5</td>
<td>23.5</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td>24.5</td>
<td>26</td>
<td>29.5</td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22.5</td>
<td>25.5</td>
<td>25.5</td>
<td>26</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23</td>
<td>24.5</td>
<td>26</td>
<td>30</td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
<td>21.5</td>
<td>23.5</td>
<td>25</td>
</tr>
</tbody>
</table>


Figure 5.1 Multiple Time Series Plot of Dental Measurements. Open circles represent girls; opaque circles represent boys.
A level-1 model is

\[ y_{it} = \beta_{0i} + \beta_{1i} z_{1it} + \epsilon_{it}, \]

where \( z_{1it} \) is the age of the child \( i \) on occasion \( t \). This model relates the dental measurement to the age of the child, with parameters that are specific to the child. Thus, we may interpret the quantity \( \beta_{1i} \) to be the growth rate for the \( i \)th child. A level-2 model is

\[ \beta_{0i} = \beta_{00} + \beta_{01} \text{GENDER}_i + \alpha_{0i} \quad \text{and} \quad \beta_{1i} = \beta_{10} + \beta_{11} \text{GENDER}_i + \alpha_{1i}. \]

Here, \( \beta_{00}, \beta_{01}, \beta_{10} \) and \( \beta_{11} \) are fixed parameters to be estimated. Suppose that we use a binary variable for gender, say, coding the GENDER variable 1 for females and 0 for males. Then, we may interpret \( \beta_{10} \) to the expected male growth rate and \( \beta_{11} \) to be the difference in growth rates between females and males.

Table 5.2 shows the parameter estimates for this model. Here, we see that the coefficient associated with linear growth is statistically significant, over all models. Moreover, the rate of increase for girls is lower than boys. The estimated covariance between \( \alpha_{0i} \) and \( \alpha_{1i} \), which is also the estimated covariance between \( \beta_{0i} \) and \( \beta_{1i} \), turns out to be negative. One interpretation of the negative covariance between initial status and growth rate is that subjects who start at a low level tend to grow more quickly than those who start at higher levels, and vice versa.

**Table 5.2. Dental data growth curve model parameter estimates**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Error Components Model</th>
<th>Growth Curve Model</th>
<th>Growth Curve Model deleting the 9th boy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter Estimates</td>
<td>( t ) statistic</td>
<td>Parameter Estimates</td>
</tr>
<tr>
<td>( \beta_{0i} )</td>
<td>16.341</td>
<td>16.65</td>
<td>16.341</td>
</tr>
<tr>
<td>Age (( \beta_{10} ))</td>
<td>0.784</td>
<td>10.12</td>
<td>0.784</td>
</tr>
<tr>
<td>GENDER (( \beta_{01} ))</td>
<td>1.032</td>
<td>0.67</td>
<td>1.032</td>
</tr>
<tr>
<td>AGE*GENDER (( \beta_{11} ))</td>
<td>-0.305</td>
<td>-2.51</td>
<td>-0.305</td>
</tr>
<tr>
<td>Var ( \epsilon_{it} )</td>
<td>1.922</td>
<td></td>
<td>1.716</td>
</tr>
<tr>
<td>Var ( \alpha_{0i} )</td>
<td>3.299</td>
<td></td>
<td>5.786</td>
</tr>
<tr>
<td>Var ( \alpha_{1i} )</td>
<td>0.033</td>
<td></td>
<td>0.073</td>
</tr>
<tr>
<td>Cov (( \alpha_{0i}, \alpha_{1i} ))</td>
<td>-0.290</td>
<td></td>
<td>-0.734</td>
</tr>
<tr>
<td>-2 Log Likelihood</td>
<td>433.8</td>
<td>432.6</td>
<td>388.5</td>
</tr>
<tr>
<td>AIC</td>
<td>437.8</td>
<td>440.6</td>
<td>396.5</td>
</tr>
</tbody>
</table>

For comparison purposes, Table 5.2 shows the parameter estimates with the 9th boy deleted. The effects of this subject deletion on the parameter estimates are small. Table 5.2 also shows parameter estimates of the error components model. This model employs the same level-1 model but with level-2 models

\[ \beta_{0i} = \beta_{00} + \beta_{01} \text{GENDER}_i + \alpha_{0i} \quad \text{and} \quad \beta_{1i} = \beta_{10} + \beta_{11} \text{GENDER}_i. \]

With parameter estimates calculated using the full data set, there again is little change in the parameter estimates. Because the results appear to be robust to both unusual subjects and model selection, we have greater confidence in our interpretations.
5.2.2 Multiple level models

Longitudinal versions of multiple level models follow the same notation as the cross-sectional models in Section 5.1.2 except that the level-1 replications are over time. To illustrate, we consider a 3-level model in the context of a social work application by Guo and Hussey (1999EP).

Guo and Hussey examined subjective assessments of children’s behavior made by multiple raters at two or more time points. That is, the level-1 repeated measurements are over time $t$, where the assessment was made by rater $j$ on child $i$. Raters assessed $n = 144$ seriously emotionally disturbed children receiving services through a large child mental health treatment agency located in Cleveland, Ohio. For this study, the assessment is the response of interest $y$; this response is the Deveroux Scale of Mental Disorders, a score made up of 111 items. Ratings were taken over a two-year period by parents and teachers; at each time point, assessments may be made either by the parent, teacher or both. The time of the assessment was made was recorded as $\text{TIME}_{i,j,t}$, measured in days since the inception of the study. The variable $\text{PROGRAM}_{i,j,t}$ was recorded as a 1 if the child was in program residence at the time of the assessment and 0 if the child was in day treatment or day treatment combined with treatment foster care. The variable $\text{RATER}_{i,j}$ was recorded as a 1 if rater was a teacher and 0 if the rater was a caretaker.

Analogous to equation (5.8), the level-1 model is

$$y_{i,j,t} = z_{1,i,j,t}^{\prime} \beta_{i,j} + x_{1,i,j,t}^{\prime} \beta_{1} + \epsilon_{1,i,j,t},$$ (5.14)

where there are $i = 1, \ldots, n$ children, $j = 1, \ldots, J_i$ raters and $t = 1, \ldots, T_{i,j}$ evaluations. Specifically, Guo and Hussey (1999EP) used $x_{1,i,j,t} = \text{PROGRAM}_{i,j,t}$ and $z_{1,i,j,t} = (1 \times \text{TIME}_{i,j,t})^{\prime}$. Thus, their level-1 model can be written as

$$y_{i,j,t} = \beta_{0,i,j} + \beta_{1,i,j} \times \text{TIME}_{i,j,t} + \beta_{2,i,j} \times \text{PROGRAM}_{i,j,t} + \epsilon_{1,i,j,t}.$$

The variables associated with the intercept and the coefficient for time may vary over child and rater whereas the program coefficient is constant over all observations.

The level-2 model is the same as equation (5.9)

$$\beta_{i,j} = Z_{2,i,j} \gamma_{i} + X_{2,i,j} \beta_{2} + \epsilon_{2,i,j},$$

where there are $i = 1, \ldots, n$ children and $j = 1, \ldots, J_i$ raters. The level-2 model of Guo and Hussey can be written as

$$\beta_{0,i,j} = \beta_{0,0} + \beta_{0,1} \times \text{RATER}_{i,j} + \epsilon_{2,i,j}$$

and

$$\beta_{1,i,j} = \beta_{2,0} + \beta_{2,1} \times \text{RATER}_{i,j}.$$

Again, we leave it as an exercise for the reader to show how this formulation is a special case of equation (5.9).

The level-3 model is the same as equation (5.10)

$$\gamma_{i} = X_{3,i} \beta_{3} + \epsilon_{3,i}.$$

To illustrate, the level-3 model of Guo and Hussey can be written as

$$\beta_{0,0} = \beta_{0,0,0} + \beta_{0,1,0} \times \text{GENDER}_{i} + \epsilon_{3,i}.$$

where $\text{GENDER}_{i}$ is a binary variable indicating the gender of the child.

As with the cross-sectional models in Section 5.1.2, one combines the three levels to form a single equation representation, as in equation (5.14). The hierarchical framework allows analysts to develop hypotheses that are interesting to test. The combined model allows for simultaneous, over all levels, estimation of parameters that is more efficient than estimating each level in isolation of the others.
5.3 Prediction

In Chapter 4, we distinguished between the concepts of estimating model parameters as compared to predicting random variables. In multilevel models, the dependent variables at second and higher levels are unobserved random coefficients. Because it is often desirable to understand their behavior, we wish to predict these random coefficients. To illustrate, if the unit of analysis at the second level is a school, we may wish to use predictions of second level coefficients to rank schools. It may also be of interest to use predictions of second (or higher) level coefficients for prediction in a first level model. To illustrate, if we are studying a children’s development over time, we may wish to make predictions about the future status of a child’s development.

This subsection shows how to use the best linear unbiased predictors (BLUPs) developed in Chapter 4 for these prediction problems. Best linear unbiased predictors, by definition, have the smallest variance among all unbiased predictors. In Chapter 4, we showed that these predictors can also be interpreted as shrinkage “estimators.” Because we have expressed multilevel models in terms of linear mixed effects models, we will not need to develop new theory but will be able to rely directly on the Chapter 4 results.

Two-level models

We begin our prediction discussion with the two-level model, introduced in equations (5.5)-(5.7). To make the multilevel model notation consistent with Chapters 3 and 4, use \( \mathbf{D} = \text{Var} \alpha_i \) = \( \text{Var} \beta_i \) and \( \mathbf{R}_i = \text{Var} \varepsilon_i \), where \( \varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT_i})' \). Suppose that we wish to predict \( \beta_i \). Using the results in Section 4.3.2, it is easy to check that the best linear unbiased predictor (BLUP) of \( \beta_i \) is

\[
\hat{\beta}_{i,\text{BLUP}} = a_{i,\text{BLUP}} + \mathbf{X}_{2,i} b_{2,\text{GLS}},
\]

where \( b_{2,\text{GLS}} \) is the generalized least squares estimator of \( \beta_2 \) and, from equation (4.11),

\[
a_{i,\text{BLUP}} = \mathbf{D} \mathbf{Z}_i \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i b_{2,\text{GLS}}),
\]

Recall that \( \mathbf{z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, ..., \mathbf{z}_{iT_i})' \). Further, \( b_{2,\text{GLS}} = (b_{1,\text{GLS}}', b_{2,\text{GLS}}')' \), \( \mathbf{V}_i = \mathbf{R}_i + \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' \) and \( X_i = (x_{i1}, x_{i2}, ..., x_{iT_i})' \), where \( x_{i'} = (x_{i1}, x_{i2}, x_{iT_i})' \). Thus, it is easy to compute these predictors.

Chapter 4 discussed interpretation in some special cases, the error components and the random coefficients models. Suppose that we have the error components model, so that \( z_{ij} = z_{i1} = 1 \) and \( \mathbf{R}_i \) is a scalar times the identity matrix. Further suppose that there are no level-1 explanatory variables. Then, one can check that the BLUP of the conditional mean of the level-1 response, \( \mathbb{E}(y_{i1} | \alpha_i) = \alpha_i + \mathbf{X}_{2,i} \beta_2 \), is

\[
a_{i,\text{BLUP}} + \mathbf{X}_{2,i} b_{2,\text{GLS}} = \zeta_i (\bar{\mathbf{y}}_{i} - \mathbf{X}_{2,i} b_{2,\text{GLS}}) + \mathbf{X}_{2,i} b_{2,\text{GLS}} = \zeta_i \bar{\mathbf{y}}_{i} + (1 - \zeta_i) \mathbf{X}_{2,i} b_{2,\text{GLS}},
\]

where \( \zeta_i = \frac{T_i}{T_i + (\text{Var} \varepsilon)/(\text{Var} \alpha)} \). Thus, the predictor is a weighted average of the level-2 ith unit’s average, \( \bar{\mathbf{y}}_{i} \), and the regression estimator which is an estimator derived from all level-2 units. As noted in Section 5.1.1, Raudenbush and Bryk (2002EP) refer to this as the “means-as-outcomes” model.

As described in Section 4.3.4, one can also use the BLUP technology to predict the future development of a level-1 response. From equation (4.14), we have that the forecast \( L \) lead times in the future of \( y_{i,T_i} \) is

\[
\hat{y}_{i,T_i+L} = \mathbf{z}'_{i1,T_i+L} \hat{\beta}_{i,\text{BLUP}} + \mathbf{x}'_{i1,T_i+L} \hat{\beta}_{1,\text{GLS}} + \text{Cov}(\varepsilon_{i1,T_i+L}, \varepsilon_i) \mathbf{R}_i^{-1} e_{i,\text{BLUP}},
\]
where is the $e_i, BLUP$ is the vector of BLUP residuals, given in equation (4.13a). As we saw in Section 4.3.4, in the case where the disturbances follow an autoregressive model of order 1 (AR(1)) with parameter $\rho$, we have

$$\hat{y}_{i,t+L} = Z_{i,t+L}' \beta_{BLUP} + X_{i,t+L}' \beta_{GLS} + \rho^{t} \epsilon_{i,t, BLUP}.$$ 

To illustrate, consider the Section 5.1.2 Dental example. Here, there is no serial correlation (so that $R$ is a scalar times the identity matrix), no level-1 fixed parameters and $T_i = 4$ observations for all children. Thus, the $L$ step forecast for the $i$th child is

$$\hat{y}_{i,4+L} = b_{0,i, BLUP} + b_{1,i, BLUP} z_{i,4+L},$$

where $z_{i,4+L}$ is the age of the child at time $4+L$.

**Multiple level models**

For three and higher level models, the approach is the same as with two-level models although it becomes more difficult to interpret the results. Nonetheless, for applied work, the idea is straightforward.

**Procedure for forecasting future level-1 responses**

1. Hypothesize a model at each level.
2. Combine all level models into a single model.
3. Estimate the parameters of the single model, using generalized least squares and variance components estimators, as described in Sections 3.4 and 3.5, respectively.
4. Determine best linear unbiased predictors of each unobserved random coefficient for levels two and higher, as described in Section 4.3.
5. Use the parameter estimators and random coefficient predictors to form forecasts of future level-1 responses.

To illustrate, let’s see how this procedure works for the three-level longitudinal data model.

**Step 1.** We will use the level-1 model described in equation (5.14), together with the level-2 and level-3 models in equations (5.9) and (5.10), respectively. For the level-1 model, let $R_{ij} = \text{Var } e_{i,ij}$, where $e_{i,ij} = (e_{1,ij,1}, \ldots, e_{1,ij,T_{ij}})'$.

**Step 2.** The combined model is equation (5.11), except that now we use a “$t$” subscript for time in lieu of the “$k$” subscript. Assuming the level-1, 2 and 3 random quantities are uncorrelated with one another, we define

$$\text{Var } a_{i,j} = \text{Var } \begin{pmatrix} e_{2,i,j} \\ e_{3,i,j} \end{pmatrix} = \begin{pmatrix} \text{Var } e_{2,i,j} & 0 \\ 0 & \text{Var } e_{3,i,j} \end{pmatrix} = \begin{pmatrix} D_2 & 0 \\ 0 & D_3 \end{pmatrix} = D_v$$

and

$$\text{Cov } (a_{i,j}, a_{i,k}) = \begin{pmatrix} \text{Cov } (e_{2,i,j}, e_{2,i,k}) & \text{Cov } (e_{2,i,j}, e_{3,i,j}) \\ \text{Cov } (e_{3,i,j}, e_{2,i,k}) & \text{Var } e_{3,i,j} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & D_3 \end{pmatrix} = D_c.$$
Stacking vectors, we write $y_{ij} = (y_{i1}, \ldots, y_{iT_{ij}})'$, $y_i = (y_{i1}', \ldots, y_{iT_{ij}}')'$, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ_i})'$ and $a_i = (a_{i1}' \ldots a_{iJ_i})'$. Stacking matrices, we have $X_{ij} = (x_{ij1}, \ldots, x_{iT_{ij}})'$, $Z_{ij} = (z_{ij1}, \ldots, z_{iT_{ij}})'$, $X_i = (X_{i1}, \ldots, X_{iJ_i})'$ and

$$Z_i = \begin{pmatrix} Z_{i1} & & 0 \\ & & \vdots \\ 0 & Z_{iJ_i} & \end{pmatrix}.$$

With this notation, we may write equation (5.11) in a linear mixed effects model form as $y_i = Z_i a_i + X_i \beta + \varepsilon_i$. Note the form of $R_i = \text{Var} \varepsilon_i = \text{blockdiagonal}(R_{i1}, \ldots, R_{iJ_i})$ and

Step 3. Having coded the explanatory variables and the form of the variance matrices $D$ and $R_i$, parameter estimates follow directly from the Sections 3.4 and 3.5 results.

Step 4. The BLUP predictors are formed beginning with predictors for $a_i$ of the form $a_{i,BLUP} = D Z_i V_i^{-1} (y_i - X_i \beta + \varepsilon_i)$. This yields the BLUPs for $a_{ij} = (a_{i1}', a_{iJ_i}')'$, say $a_{ij,BLUP} = (e_{2,ij,BLUP}', e_{3,ij,BLUP}')'$. These BLUPs allow us to predict the second and higher level random coefficients through the relations $g_{ij,BLUP} = X_{3,ij} b_{3,GLS} + e_{3,ij,BLUP}$ and $b_{ij,BLUP} = Z_{2,ij} g_{ij,BLUP} + X_{2,ij} b_{2,GLS} + e_{2,ij,BLUP}$, corresponding to equations (5.10) and (5.9), respectively.

Step 5. If desired, we may forecast future level-1 responses. From equation (4.14), for an $L$-step forecast, we have

$$\hat{y}_{i1,1:T_i+L} = Z_{i1,i,T_i+L} b_{1,ij,BLUP} + X_{i1,i,T_i+L} b_{1,GLS} + \text{Cov}(\varepsilon_{i1,i,T_i+L}, e_{i1,i}) R_{i1}^{-1} e_{i1,ij,BLUP},$$

For AR(1) level-1 disturbances, this simplifies to

$$\hat{y}_{i1,i,T_i+L} = Z_{i1,i,T_i+L} b_{1,ij,BLUP} + X_{i1,i,T_i+L} b_{1,GLS} + \rho \varepsilon_{i1,i} e_{i1,ij,BLUP}.$$

### 5.4 Testing variance components

Multilevel models implicitly provide a representation for the variance as a function of explanatory variables. To illustrate, consider the cross-sectional two-level model summarized in equations (5.5)-(5.7). With equation (5.7), we have

$$\text{Var} y_{ij} = z_{ij}' (\text{Var} a_i) z_{ij} + \text{Var} \varepsilon_{ij},$$

and

$$\text{Cov}(y_{ij}, y_{ik}) = z_{ij}' (\text{Var} a_i) z_{ik}.$$

Thus, even if the random quantities $a_i$ and $\varepsilon_{ij}$ are homoscedastic, the variance is a function of the explanatory variables $z_{ij}$. Particularly in education and psychology, researchers wish to test theories by examining hypotheses concerning these variance functions.
Unfortunately, the usual likelihood ratio testing procedure is not valid for testing many variance components of interest. In particular, the concern is for testing parameters where the null hypothesis is on the boundary of possible values. As a general rule, the standard hypothesis testing procedures favors the simpler null hypothesis more often than it should.

To illustrate the difficulties with boundary problems, let’s consider the classic example of i.i.d. random variables $y_1, \ldots, y_n$ where each random variable is distributed normally with known mean zero and variance $\sigma^2$. Suppose that we wish to test the null hypothesis $H_0: \sigma^2 = \sigma_0^2$, where $\sigma_0^2$ is a known positive constant. It is easy to check that the maximum likelihood estimator of $\sigma^2$ is $\sum_{i=1}^{n} y_i^2$. As we have seen, a standard method of testing hypotheses is the likelihood ratio test procedure (described in more detail in Appendix A.7). Here, one computes the likelihood ratio test statistic, which is twice the difference between the unconstrained maximum log-likelihood and the maximum log-likelihood under the null hypothesis, and compares this statistic to a chi-square distribution with one degree of freedom. Unfortunately, this procedure is not available when $\sigma_0^2 = 0$ because the log-likelihoods are not well defined. Because $\sigma_0^2 = 0$ is on the boundary of the parameter space $[0, \infty)$, the regularity conditions of our usual test procedures are not valid.

However, $H_0: \sigma^2 = 0$ is still a testable hypothesis; a simple test is to reject $H_0$ if the maximum likelihood estimator exceeds zero. This procedure will always reject the null hypothesis when $\sigma^2 > 0$ and accept when $\sigma^2 = 0$. Thus, this test procedure has power 1 versus all alternatives and a significance level of zero, a very good test!

For an example closer to longitudinal data models, consider the Section 3.1 error components model with variance parameters $\sigma^2$ and $\sigma_\alpha^2$. In the Exercise 5.4, we outline the proof to establish that the likelihood ratio test statistic for assessing $H_0: \sigma_\alpha^2 = 0$ is $\frac{1}{2}\chi^2_{(1)}$, where $\chi^2_{(1)}$ is a chi-square random variable with 1 degree of freedom. In the usual likelihood ratio procedure for testing one variable, the likelihood ratio test statistic has a $\chi^2_{(1)}$ distribution under the null hypothesis. This means that using nominal values, we will accept the null hypothesis more often than we should; thus, we will sometimes use a simpler model than suggested by the data.

The critical point of this exercise is that we define maximum likelihood estimators to be non-negative, arguing that a negative estimator of variance components is not valid. Thus, the difficulty is that the usual regularity conditions (see, for example, Serfling, 1980G) require that the hypotheses that we test lie on the interior of a parameter space. For most variances, the parameter space is $[0, \infty)$. By testing that the variance equals zero, we are on the boundary and the usual asymptotic results are not valid. This does not mean that tests for all variance components are not valid. For example, for testing most correlations and autocorrelations, the parameter space is $[-1,1]$. Thus, for testing correlations (and covariances) equal to zero, we are in the interior of the parameter space and so the usual test procedures are valid.

In contrast, in Exercise 5.3, we allow negative variance estimators. In this case, by following the outline of the proof, you will see that the usual likelihood ratio test statistic for assessing $H_0: \sigma_\alpha^2 = 0$ is $\chi^2_{(1)}$, the customary distribution. Thus, it is important to know the constraints underlying the software package that you are using.

A complete theory for testing variance components has yet to be developed. When only one variance parameter needs to be assessed for equality to zero, results similar to the error components model discussed above have been worked out. For example, Balagi and Li (1990E) developed a test for a second (independent) error component representing time; this model will be described in Chapter 8. More generally, checking for the presence of an additional random effect in the model implicitly means checking that not only the variance, but also the covariances, are
equal to zero. For example, for the linear mixed effects model with a $q \times 1$ vector of variance components $\alpha_i$, we might wish to assess the null hypothesis

$$H_0 : \mathbf{D} = \text{Var} \alpha_i = \begin{pmatrix} \text{Var}(\alpha_{i,1}, \ldots, \alpha_{i,q-1}) & 0 \\ 0' & 0 \end{pmatrix}.$$ 

In this case, based on the work of Self and Liang (1987S), Stram and Lee (1994S) showed that the usual likelihood ratio test statistic has an asymptotic distribution is

$$\frac{1}{2} \chi^2_{(q-1)} + \frac{1}{2} \chi^2_{(q)},$$

where $\chi^2_{(q-1)}$ and $\chi^2_{(q)}$ are independent chi-square random variables with $q-1$ and $q$ degrees of freedom, respectively. The usual procedure for testing means comparing the likelihood ratio test statistic to $\chi^2_{(q)}$ because we are testing a variance parameter and $q-1$ covariance parameters. Thus, if one rejects using the usual procedure, one will reject using the mixture distribution corresponding to $\frac{1}{2} \chi^2_{(q-1)} + \frac{1}{2} \chi^2_{(q)}$. Put another way, the actual $p$-value (computed using the mixture distribution) is less than the nominal $p$-value (computed using the standard distribution). Thus, we see that the standard hypothesis testing procedures favors the simpler null hypothesis more often than it should.

No general rules for checking for the presence of several additional random effects are available although simulation methods are always possible. The important point is that analysts should not quickly quote $p$-values associated with testing variance components without carefully considering the model and estimator.

**Further reading**

Appendix 5A – High Order Multilevel Models

Despite their widespread application, standard treatments that introduce the multilevel model use at most at only three levels, anticipating that users will be able to intuit patterns (and hopefully equation structures) to higher levels. In contrast, this appendix describes a high order multilevel model using “k” levels.

To motivate the extensions, begin with the three-level model in Section 5.1.2. Extending equation (5.8), the level-1 model is expressed as

\[ y_{i_1,i_2,...,i_k} = Z_{i_1,i_2,...,i_k}^{(1)} \beta_1^{(1)} + X_{i_1,i_2,...,i_k}^{(1)} \beta_1 + \epsilon_{i_1,i_2,...,i_k}^{(1)}. \]

Here, we might use \( i_k \) as a time index, \( i_{k-1} \) is a student index, \( i_{k-2} \) is classroom index, and so on. We denote the observation set by \( \{ i(k) \} \) as \( y_{i_1,i_2,...,i_k} \) is observed for some \( j_{k+1},...,j_k \). More generally, define

\[ i(k-s) = \{ i_1, ..., i_{k-s} \}: y_{i_1,i_2,...,i_k} \text{ is observed for some } j_{k+1},...,j_k, \]

for \( s = 0, 1, ..., k-1 \). We will let \( i(k) = \{ i_1, i_2, ..., i_k \} \) be a typical element of \( \bar{i}(k) \) and use \( i(k-s) = \{ i_1, ..., i_{k-s} \} \) for the corresponding element of \( \bar{i}(k-s) \).

With this additional notation, we are now in a position to provide a recursive specification of high order multilevel models.

**Recursive specification of high order multilevel models**

1. The level-1 model is

\[ y_{i(k)} = Z_{i(k)}^{(1)} \beta_1^{(1)} + X_{i(k)}^{(1)} \beta_1 + \epsilon_{i(k)}^{(1)}, \quad i(k) \in \bar{i}(k). \]  
   \[ (5A.1) \]

   The level-1 fixed parameter vector \( \beta_1 \) has dimension \( K_1 \times 1 \) and the level-1 vector of parameters that may vary over higher levels, \( \beta_1^{(1)} \), has dimension \( q_1 \times 1 \).

2. For \( g = 2, ..., k-1 \), the level-g model is

\[ \beta_{i(k+1-g)}^{(g)} = Z_{i(k+1-g)}^{(g)} \beta_{i(k-g)}^{(g)} + X_{i(k+1-g)}^{(g)} \beta_{i(k-g)} + \epsilon_{i(k-g)}^{(g)}, \quad \text{for } g = 2, ..., k-1. \]  
   \[ (5A.2) \]

   Here, the level-2 fixed parameter vector \( \beta_2 \) has dimension \( K_2 \times 1 \) and the level-g varying parameter vector \( \beta_{i(k-g)}^{(g)} \) has dimension \( q_{g-1} \times 1 \). Thus, for the covariates, \( Z_{i(k+1-g)}^{(g)} \) has dimension \( q_{g-1} \times q_g \) and \( X_{i(k+1-g)}^{(g)} \) has dimension \( q_{g-1} \times K_g \).

3. The level-k model is

\[ \beta_{i(k)}^{(k-1)} = X_{i(k)}^{(k)} \beta_k + \epsilon_{i(k)}^{(k)}. \]  
   \[ (5A.3) \]

We assume that all disturbance terms \( \epsilon \) are mean zero and are uncorrelated with one another.

Further, define \( D_g = \text{Var}(\epsilon_{i(k+1-g)}^{(g)}) = \sigma_g^2 \mathbb{I}_{q_{g-1}}, \) for \( g \geq 2. \)

We now show how to write the multilevel model as a linear mixed effects model. We do this by recursively inserting the higher level models from equation (5A.2) into the level-1 equation (5A.1). This yields

\[ y_{i(k)} = \epsilon_{i(k)}^{(1)} + X_{i(k)}^{(1)} \beta_1 + Z_{i(k)}^{(2)} \left( Z_{i(k-1)}^{(2)} \beta_{i(k-2)}^{(2)} + X_{i(k-1)}^{(2)} \beta_2 + \epsilon_{i(k-1)}^{(2)} \right). \]
\[ \varepsilon_{i(k)}^{(1)} + X_{i(k)}^{(1)} \beta_1 + Z_{i(k)}^{(1)} \left( \varepsilon_{i(k-1)}^{(2)} + X_{i(k-1)}^{(2)} \beta_2 \right) + Z_{i(k)}^{(2)} \left( Z_{i(k-1)}^{(3)} \beta_3 + X_{i(k-1)}^{(3)} \beta_3 + \varepsilon_{i(k-2)}^{(3)} \right) \]

\[ = \cdots = \varepsilon_{i(k)}^{(1)} + X_{i(k)}^{(1)} \beta_1 + \sum_{s=1}^{k-1} \left( \prod_{j=1}^{s} Z_{i(k-j)}^{(j)} \right) \left( \varepsilon_{i(k-s)}^{(s+1)} + X_{i(k-s)}^{(s+1)} \beta_{s+1} \right). \]

To simplify notation, define the \( 1 \times q_s \) vector

\[ Z_{s,i(k)} = \prod_{j=1}^{s} Z_{i(k-j)}^{(j)}. \]  \hfill (5A.4)

Further, define \( K = K_1 + \ldots + K_k \), the \( K \times 1 \) vector \( \beta = (\beta_1', \beta_2', \ldots, \beta_k')' \) and the \( 1 \times K \) vector

\[ X_{i(k)} = \left( X_{i(k)}^{(1)} Z_{i(i(k))}^{(1)} \right. \ldots \left. Z_{k-1,i(k)} X_{i(k)}^{(k)} \right). \]

This yields

\[ X_{i(k)} \beta = X_{i(k)}^{(1)} \beta_1 + \sum_{s=1}^{k-1} Z_{s,i(k)} X_{i(k-s)}^{(s+1)} \beta_{s+1}. \]

Thus, we may express the multilevel model as

\[ y_{i(k)} = X_{i(k)} \beta + \varepsilon_{i(k)}^{(1)} + \sum_{s=1}^{k-1} Z_{s,i(k)} \varepsilon_{i(k-s)}^{(s+1)}. \]  \hfill (5A.5)

To write equation (5A.5) as a mixed linear model, we require some additional notation. For a fixed set \( \{i_1, \ldots, i_{k-1}\} = i(k-1) \), let \( n(i(k-1)) \) denote the number of observed responses of form \( Y_{i(k),j} \) for some \( j \). Denote the set of observed responses as

\[ y_{i(k-1)} = \begin{pmatrix} y_{i(k-1),1} \\ \vdots \\ y_{i(k-1),n(i(k-1))} \end{pmatrix}. \]

For each \( s=1, \ldots, k-1 \), consider a set \( \{i_1, \ldots, i_{k-s}\} = i(k-s) \) and let \( n(i(k-s)) \) denote the number of observed responses of form \( Y_{i(k-s),j} \) for some \( j \). Thus, we define

\[ y_{i(k-s)} = \begin{pmatrix} y_{i(k-s),1} \\ \vdots \\ y_{i(k-s),n(i(k-s))} \end{pmatrix}. \]

Finally, let \( y = (y_1', \ldots, y_{n(i(k))}'). \) Use a similar stacking scheme for \( X \) and \( \varepsilon^{(s)} \), for \( s = 1, \ldots, k \). We may also use this notation when stacking over the first level of \( Z \). Thus, define

\[ Z_{s,i(k-1)} = \begin{pmatrix} Z_{s,i(k-1),1} \\ \vdots \\ Z_{s,i(k-1),n(i(k-1))} \end{pmatrix}, \text{ for } s = 1, \ldots, k-1. \]

With this notation, when stacking over the first level, we may express equation (5A.5) as

\[ y_{i(k-1)} = X_{i(k-1)} \beta + \varepsilon_{i(k-1)}^{(1)} + \sum_{s=1}^{k-1} Z_{s,i(k-1)} \varepsilon_{i(k-s)}^{(s+1)}. \]
For the next level, define
\[
Z_{s,j(k-2)} = \begin{pmatrix}
Z_{s,j(k-2),1} \\
\vdots \\
Z_{s,j(k-2),n(i(k-2))}
\end{pmatrix}, \text{ for } s = 2, \ldots, k-1.
\]
and
\[
Z_{l,i(k-2)} = \text{blkdiag}(Z_{l,i(k-2),1}, \ldots, Z_{l,i(k-2),n(i(k-2)))}.
\]
With this notation, we have
\[
Z_{l,i(k-2)}e^{(2)}_{i(k-2)} = \begin{pmatrix}
Z_{l,i(k-2),1} & 0 & \cdots & 0 \\
0 & Z_{l,i(k-2),2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & Z_{l,i(k-2),n(i(k-2))}
\end{pmatrix}
\begin{pmatrix}
e^{(2)}_{i(k-2),1} \\
e^{(2)}_{i(k-2),2} \\
\vdots \\
e^{(2)}_{i(k-2),n(i(k-2))}
\end{pmatrix}.
\]
Thus, we have
\[
y_{i(k-2)} = X_{i(k-2)}\beta + e^{(1)}_{i(k-2)} + Z_{l,i(k-2)}e^{(2)}_{i(k-2)} + \sum_{s=2}^{k-1} Z_{s,j(k-2)}e^{(s+1)}_{i(k-s)}.
\]
Continuing, at the gth stage, we have
\[
Z_{s,j(k-g)} = \begin{pmatrix}
Z_{s,j(k-g),1} \\
\vdots \\
Z_{s,j(k-g),n(i(k-g))}
\end{pmatrix}, \text{ for } s \geq g
\]
\[
\text{and } \text{blkdiag}(Z_{s,j(k-g),1}, \ldots, Z_{s,j(k-g),n(i(k-g))}), \text{ for } s < g.
\]
This yields
\[
y_{i(k-g)} = X_{i(k-g)}\beta + e^{(1)}_{i(k-g)} + \sum_{s=1}^{g} Z_{s,j(k-g)}e^{(s+1)}_{i(k-g)} + \sum_{s=g+1}^{k-1} Z_{s,j(k-g)}e^{(s+1)}_{i(k-s)}.
\]
Taking \( g = k-1 \), we have
\[
y_{i(1)} = X_{i(1)}\beta + e^{(1)}_{i(1)} + \sum_{s=1}^{k-1} Z_{s,j(1)}e^{(s+1)}_{i(1)}, \quad (5A.6)
\]
an expression for the usual linear mixed effects model.

The system of notation takes us directly from the multilevel model in equations (5A.1)-(5A.3) to the linear mixed effects model in equation (5A.6). Properties of parameter estimates for linear mixed effects model are well established. Thus, parameter estimators of the multilevel model also enjoy these properties. Moreover, by showing how to write multilevel models as linear mixed effects model, no special statistical software is required. One may simply use software written for linear mixed effects models for multilevel modeling.
5. Exercises and Extensions

Section 5.3

5.1. Two-level model

Consider the two-level model described in Section 5.1.1 and suppose that we have the error components model, so that \( z_{ij} = z_{1ij}, z_{2ij} = 1 \) and \( R_i \) is a scalar times the identity matrix. Further suppose that there are no level-1 explanatory variables. Show that the BLUP of the conditional mean of the level-1 response, \( E(y_{it} | \alpha_i) = \alpha_i + X_{2ij} \beta_2, \) is \( \zeta_i y_{ij} + (1 - \zeta_i) X_{2ij} b_{2, GLS}, \) where \( \zeta_i = \frac{T_i}{T_i + (\text{Var} \varepsilon) / (\text{Var} \alpha)} \).

5.2. Three-level model

Assume that we observe \( i = 1, \ldots, n \) school districts. Within each school district, we observe \( j = 1, \ldots, J \) students. For each student, we have \( t = 1, \ldots, T_{ij} \) observations. Assume that the model is given by

\[
y_{ijt} = x_{ijt}' \beta + \alpha_i + \nu_{ij} + \varepsilon_{ijt}.
\]

Here, assume that each of \( \{ \alpha_i \}, \{ \nu_{i1} \}, \ldots, \{ \nu_{ij} \} \) and \( \{ \varepsilon_{ijt} \} \) are independently and identically distributed as well as independent of one another. Also assume that \( \{ \alpha_i \}, \{ \nu_{i1} \}, \ldots, \{ \nu_{ij} \} \) and \( \{ \varepsilon_{ijt} \} \) are mean zero with variances \( \sigma^2, \sigma^2 \nu_1, \ldots, \sigma^2 \nu_J \) and \( \sigma^2 \varepsilon \), respectively.

Define \( z_{ij} \) to be a \( T_{ij} \times (J+1) \) matrix with ones in the first and \( j+1 \)st columns and zeroes elsewhere. For example, we have

\[
\begin{pmatrix}
1 & 0 & 1 & \cdots & 0 \\
1 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & \cdots & 0
\end{pmatrix}
\]

Further define \( Z_i = (z_{i1}' \ z_{i2}' \ \cdots \ z_{ij}' \ \cdots)' \), \( \alpha_i = (\alpha_i \ \nu_{i1} \ \cdots \ \nu_{ij} \ \cdots)' \) and

\[
D = \text{Var} \alpha_i = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 \nu_1 & \cdots & \sigma^2 \nu_J \end{pmatrix}, \quad \text{where} \quad D_v = \text{diag}\left( \sigma^2 \nu_1, \sigma^2 \nu_2, \ldots, \sigma^2 \nu_J \right).
\]

a. Define \( y_i, x_i \) and \( \varepsilon_i \) in terms of \( \{ y_{ijt} \}, \{ x_{ijt} \} \) and \( \{ \varepsilon_{ijt} \} \), so that we may write \( y_i = Z_i \alpha_i + X_i \beta + e_i \), using the usual notation.

b. For the appropriate choice of \( R_i \), show that

\[
Z_i' R_i^{-1} (y_i - X_i b_{GLS}) = \frac{1}{\sigma^2 \varepsilon} \left( T_i e_{i1}' \ T_i e_{i1}' \ \cdots \ T_i e_{iJ}' \right),
\]

where \( e_{ijt} = y_{ijt} - x_{ijt} b_{GLS} \), \( e_{ij} = T_i^{-1} \sum_{t=1}^{T_{ij}} e_{ijt} \) and \( T_i \bar{e}_i = \frac{1}{J} \sum_{j=1}^{J} \sum_{t=1}^{T_{ij}} e_{ijt} \).

c. Show that

\[
\left( D^{-1} + Z_i' R_i^{-1} Z_i \right)^{-1} = \begin{pmatrix}
C_{11}^{-1} & -C_{11}^{-1} \zeta_{10}' \\
-C_{11}^{-1} \zeta_{0}' & C_{22}^{-1}
\end{pmatrix},
\]
where $\zeta_i = \frac{\sigma_i^2 T_i}{\sigma_e + \sigma_i^2 T_i}$, $\zeta_{ij} = \frac{\sigma_{ij}^2 T_{ij}}{\sigma_e + \sigma_{ij}^2 T_{ij}}$, $\zeta = (\zeta_{11}, \zeta_{12}, \ldots, \zeta_{1L})'$.

$T_{i,2} = \text{diag}(T_{i1}, T_{i2}, \ldots, T_{iL})$, $C_{11}^{-1} = \sum_{j=1}^J T_{ij} (1 - \zeta_i \zeta_{ij})$ and

$C_{22}^{-1} = (D_{ij}^{-1} + \sigma_e^{-2} T_{i,2})^{-1} + C_{11}^{-1} \Sigma \zeta_{ij}$.

d. With the notation $a_{i,\text{BLUP}} = (a_{i,\text{BLUP}} \, v_{i,\text{BLUP}} \, \ldots \, v_{i,\text{BLUP}})'$, show that

$$a_{i,\text{BLUP}} = \frac{\sum_{j=1}^J T_{ij} (1 - \zeta_i \zeta_{ij}) \overline{y}_j}{\sum_{j=1}^J T_{ij} (1 - \zeta_i \zeta_{ij})}$$

and

$$v_{ij,\text{BLUP}} = \zeta_{ij} \left( \overline{y}_{ij} - a_{i,\text{BLUP}} \right).$$

Section 5.4

5.3. MLE variance estimators without boundary conditions

Consider the basic random effects model and suppose that $T_i = T_i$, $K = 1$ and that $x_{it} = 1$. Parts (a) and (b) are the same as Exercise 3.10 (a) and (b). As there, for this problem, we ignore boundary conditions so that the estimator may become negative with positive probability.

a. Show that the maximum likelihood estimator of $\sigma_e^2$ may be expressed as:

$$\hat{\sigma}_e^2,\text{ML} = \frac{1}{n(T-1)} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \overline{y}_i)^2.$$  

b. Show that the maximum likelihood estimator of $\sigma^2$ may be expressed as:

$$\hat{\sigma}_e^2,\text{ML} = \frac{1}{n(T-1)} \sum_{i=1}^n (\overline{y}_i - \overline{y})^2 - \frac{1}{T} \hat{\sigma}_e^2,\text{ML}.$$  

c. Show that the maximum likelihood may be expressed as:

$$L(\hat{\sigma}_{a,\text{ML}}, \hat{\sigma}_{e,\text{ML}}) = -\frac{n}{2} \left[ T \ln(2\pi) + T + (T-1) \ln \hat{\sigma}_e^2,\text{ML} + \ln(T \hat{\sigma}_{a,\text{ML}}^2 + \hat{\sigma}_e^2,\text{ML}) \right].$$

d. Consider the null hypothesis $H_0$: $\sigma_a^2 = 0$. Under this null hypothesis, show that the maximum likelihood estimator of $\sigma_e^2$ may be expressed as:

$$\hat{\sigma}_e^2,\text{Reduced} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \overline{y})^2.$$  

e. Under the null hypothesis $H_0$: $\sigma_a^2 = 0$, show that the maximum likelihood may be expressed as:

$$L(0, \hat{\sigma}_{e,\text{Reduced}}^2) = -\frac{n}{2} \left[ T \ln(2\pi) + T + T \ln \left( \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \overline{y})^2 \right) \right].$$

f. Use a second order approximation of the logarithm function to show that twice the difference of log-likelihoods may be expressed as:

$$2 \left[ L(\hat{\sigma}_{a,\text{ML}}, \hat{\sigma}_{e,\text{ML}}^2) - L(0, \hat{\sigma}_{e,\text{Reduced}}^2) \right] = \frac{1}{2nT(T-1)\sigma_e^2} \{SSW - (T-1)SSB \}^2,$$
where \( SSW = \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \bar{y})^2 \) and \( SSB = T \sum_{i} (\bar{y}_i - \bar{y})^2 \).

f. Assuming normality of the responses and the null hypothesis \( H_0: \sigma_a^2 = 0 \), show that
\[
2(L(\hat{\sigma}_{a,ML}^2, \hat{\sigma}_{e,ML}^2) - L(0, \hat{\sigma}_{e,Reduced}^2)) \xrightarrow{D} \chi^2(1),
\]
as \( n \to \infty \).

5.4. MLE variance estimators with boundary conditions

Consider the basic random effects model and suppose that \( T_i = T, K = 1 \) and that \( x_{it} = 1 \). Unlike problem 5.3, we now impose boundary conditions so that variance estimators must be nonnegative.

a. Using the notation of Exercise 5.3, show that the maximum likelihood estimators of \( \sigma_e^2 \) and \( \sigma_a^2 \) may be expressed as:
\[
\hat{\sigma}_{a,CML}^2 = \begin{cases} \hat{\sigma}_{a,ML}^2 & \text{if } \hat{\sigma}_{a,ML}^2 > 0 \\ 0 & \text{if } \hat{\sigma}_{a,ML}^2 \leq 0 \end{cases} \quad \text{and} \quad \hat{\sigma}_{e,CML}^2 = \begin{cases} \hat{\sigma}_{e,ML}^2 & \text{if } \hat{\sigma}_{a,ML}^2 > 0 \\ \hat{\sigma}_{e,Reduced}^2 & \text{if } \hat{\sigma}_{a,ML}^2 \leq 0 \end{cases}.
\]
An early reference for this result is Herbach (1959G).

b. Show that the maximum likelihood may be expressed as:
\[
L(\hat{\sigma}_{a,CML}^2, \hat{\sigma}_{e,CML}^2) = \begin{cases} L(\hat{\sigma}_{a,ML}^2, \hat{\sigma}_{e,ML}^2) & \text{if } \hat{\sigma}_{a,ML}^2 > 0 \\ L(0, \hat{\sigma}_{e,Reduced}^2) & \text{if } \hat{\sigma}_{a,ML}^2 \leq 0 \end{cases}.
\]

c. Define the cut-off \( c_n = (T - 1) \frac{SSB}{SSW} - 1 \). Check that \( c_n > 0 \) if and only if \( \hat{\sigma}_{a,ML}^2 > 0 \). Confirm that we may express the likelihood ratio statistic as
\[
2(L(\hat{\sigma}_{a,CML}^2, \hat{\sigma}_{e,CML}^2) - L(0, \hat{\sigma}_{e,Reduced}^2)) = \begin{cases} n \left[ T \ln \left( 1 + \frac{c_n}{T} \right) - \ln(1 + c_n) \right] & \text{if } c_n > 0 \\ 0 & \text{if } c_n \leq 0 \end{cases}.
\]

d. Assuming normality of the responses and the null hypothesis \( H_0: \sigma_a^2 = 0 \), show that the cut-off \( c_n \to_p 0 \) as \( n \to \infty \).

e. Assuming normality of the responses and the null hypothesis \( H_0: \sigma_a^2 = 0 \), show that
\[
\sqrt{n}c_n \xrightarrow{D} N\left( 0, \frac{2T}{T - 1} \right) \quad \text{as } n \to \infty,
\]
where \( \Phi \) is the standard normal distribution function.

f. Assume normality of the responses and the null hypothesis \( H_0: \sigma_a^2 = 0 \), Show, for \( a > 0 \), that
\[
\text{Prob}\left[ 2(L(\hat{\sigma}_{a,CML}^2, \hat{\sigma}_{e,CML}^2) - L(0, \hat{\sigma}_{e,Reduced}^2)) > a \right] \xrightarrow{D} 1 - \Phi\left( \sqrt{a} \right) \quad \text{as } n \to \infty.
\]

g. Assume normality of the responses and the null hypothesis \( H_0: \sigma_a^2 = 0 \). Summarize the results above to establish that the likelihood ratio test statistic asymptotically has a distribution that is 50% equal to 0 and 50% a chi-square distribution with one degree of freedom.
Empirical Exercise
5.5. Student Achievement

These data were gathered to assess the relationship between student achievement and education initiatives. Moreover, they can also be used to address related interesting questions, such as how one can rank the performance of schools or how one can forecast a child’s future performance on achievement tests based on their early test scores.

Webb et al. (2002) investigated relationships between student achievement and Texas school district participation in the National Science Foundation Statewide Systemic Initiatives program between 1994 and 2000. They focused on the effects of systemic reform on performance on a state mathematics test. We consider here a subset of these data to model trajectories of students’ mathematics achievement over time. This subset consists of a random sample of 20 elementary schools in Dallas, with 20 students randomly selected from each school. All available records for these 400 students during elementary school are included. In Dallas, Grades 3 through 6 correspond to elementary school.

Although there exists a natural hierarchy at each time point (students are nested within schools), this hierarchy was not maintained completely over time. Several students switched schools (see variable SWITCH_SCHOOLS) and many students were not promoted (see variable RETAINED). To maintain the hierarchy of students within schools, a student was associated with a school at the time of selection. To maintain a hierarchy over time, a cohort variable was defined as 1, 2, 3, 4 for those in grades 6, 5, 4 and 3, respectively, in 1994, and a 5 for those in grade 3 in 1995, and so on up to a 10 for those in grade 3 in 2000. The variable FIRST_COHORT attaches a student to a cohort during the first year of observation whereas the variable LAST_COHORT attaches a student to a cohort during the last year of observation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADE</td>
<td>Grade when assessment was made (3-6)</td>
</tr>
<tr>
<td>YEAR</td>
<td>Year of assessment (1994-2000)</td>
</tr>
<tr>
<td>TIME</td>
<td>Observed repeated occasions for each student</td>
</tr>
<tr>
<td>RETAINED</td>
<td>Retained in grade for a particular year (1=yes, 0=no)</td>
</tr>
<tr>
<td>SWITCH_SCHOOLS</td>
<td>Switched schools in a particular year (1=yes, 0=no)</td>
</tr>
<tr>
<td>DISADVANTAGED</td>
<td>Economically disadvantaged (1= free/reduced lunch, 0=no)</td>
</tr>
<tr>
<td>TLI_MATH</td>
<td>Texas Learning Index on mathematics – assessment measure</td>
</tr>
<tr>
<td>CHILDID</td>
<td>Student identification number</td>
</tr>
<tr>
<td>MALE</td>
<td>Gender of students (1=male, 0=female)</td>
</tr>
<tr>
<td>ETHNICITY</td>
<td>White, black, hispanic, other (“other” includes asian as well as mixed races)</td>
</tr>
<tr>
<td>FIRST_COHORT</td>
<td>First observed cohort membership</td>
</tr>
<tr>
<td>LAST_COHORT</td>
<td>Last observed cohort membership</td>
</tr>
<tr>
<td>SCHOOLID</td>
<td>School identification number</td>
</tr>
<tr>
<td>MATH_SESSIONS</td>
<td>Number of teachers attended mathematics sessions</td>
</tr>
<tr>
<td>N_TEACHERS</td>
<td>Total number of teachers in the school</td>
</tr>
</tbody>
</table>

a Basic summary statistics
   i Summarize the school level variables. Produce a table to summarize the frequency of the USI variable. For MATH_SESSIONS and N_TEACHERS, provide the mean, median, standard deviation, minimum and the maximum.
   ii Summarize the child level variables. Produce tables to summarize the frequency of gender, ethnicity and the cohort variables.
   iii Provide basic relationships among level 2 and 3 variables. Summarize the number of teachers by gender. Examine ethnicity by gender.
   iv Summarize the level 1 variables. Produce means for the binary variables RETAINED, SWITCH_SCHOOLS and DISAVANTAGED. For TLI_MATH, provide the mean, median, standard deviation, minimum and the maximum.
   v Summarize numerically some basic relationships between TLI_MATH and the explanatory variables. Produce tables of means of TLI_MATH by GRADE, YEAR, RETAINED, SWITCH_SCHOOLS and DISAVANTAGED.
   vi Summarize graphically some basic relationships between TLI_MATH and the explanatory variables. Produce boxplots of TLI_MATH by GRADE, YEAR, RETAINED, SWITCH_SCHOOLS and DISAVANTAGED. Comment on the trend over time and grade.
   vii Produce a multiple time series plot of TLI_MATH. Comment on the dramatic declines of some students in year-to-year test scores.

b Two level - Error components model
   i Ignoring the school level information, run an error components model using child as the second level unit of analysis. Use the level 1 categorical variables GRADE and YEAR and binary variables RETAINED and SWITCH_SCHOOLS. Use the level 2 categorical variables ETHNICITY and the binary variable MALE.
   ii Repeat your analysis in part b(i) but include the variable DISAVANTAGED. Describe the advantages and disadvantages of including this variable in the model specification.
   iii Repeat your analysis in part b(i) but include an AR(1) specification of the error. Does this improve the model specification?
   iv Repeat your analysis in part b(iii) but include a (fixed) school level categorical variable. Does this improve the model specification?

c Three level model
   i Now incorporate school level information into your model in b(i). At the first level, the random intercept that varies by child and school. Include GRADE, YEAR, RETAINED and SWITCH_SCHOOLS as level 1 explanatory variables. At the second level, the random intercept that varies by school. Include ETHNICITY and MALE as level 2 explanatory variables. At the third level, include USI, MATH_SESSIONS and N_TEACHERS as level 3 explanatory variables. Comment on the appropriate of this fit.
   ii Is the USI categorical variable statistically significant? Re-run the part c(i) model without USI and use a likelihood ratio test statistic is respond to this question.
   iii Repeat your analysis in part c(i) but include an AR(1) specification of the error. Does this improve the model specification?
Appendix 5A

5.6. BLUP predictors for a general multilevel model

Consider the general multilevel model developed in Appendix 5A and the mixed linear model representation in equation (5A.6). Let $V_{i(i)} = \text{Var} \ y_{i(i)}$.

a. Using best linear unbiased prediction introduced in Section 4.2, show that we can express the BLUP predictors of the residuals as

$$e_{i(k+1-g),BLUP}^{(g)} = \text{Cov}(e_{i(i)}^{(g)}, e_{i(k+1-g)}^{(g)})Z'_{g-1,i(i)}V^{-1}_{i(i)}(y_{i(i)} - X_{i(i)}b_{GLS})$$

for $g = 2, \ldots, k$, and, for $g=1$,

$$e_{i(k),BLUP}^{(1)} = \text{Cov}(e_{i(i)}^{(1)}, e_{i(k)}^{(1)})V^{-1}_{i(i)}(y_{i(i)} - X_{i(i)}b_{GLS})$$

b. Show that the BLUP predictor of $b_{i(k+1-g)}^{(g-1)}$ is

$$b_{i(k+1-g),BLUP}^{(g-1)} = X_{i(k+1-g)}^{(g)}b_{GLS} + Z_{i(k+1-g)}^{(g)}b_{GLS,BLUP} + e_{i(k+1-g),BLUP}^{(g)}$$

c. Show that the BLUP forecast of $Y_{i_{1}j_{2},...,j_{L}}$ is

$$\hat{Y}_{i_{1}j_{2},...,j_{L}} = Z_{i_{1}j_{2},...,j_{L}}^{(1)}b_{i_{1}j_{2},...,j_{L}}^{(1),BLUP} + X_{i_{1}j_{2},...,j_{L}}^{(1)}b_{GLS}$$

$$+ \text{Cov}(e_{i_{1}j_{2},...,j_{L}}^{(1)}, e_{i(i)}^{(1)})V^{-1}_{i(i)}(y_{i(i)} - X_{i(i)}b_{GLS})$$