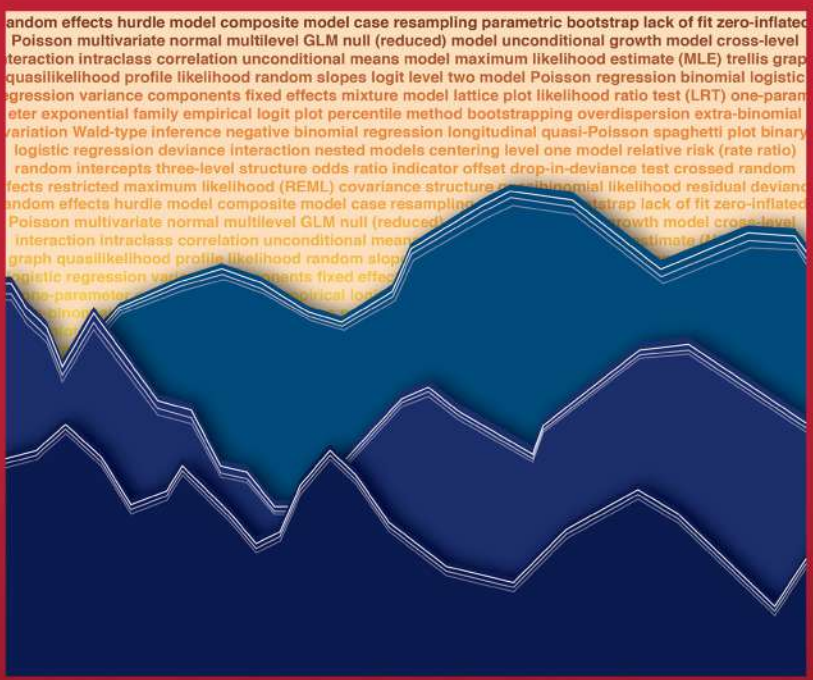
Texts in Statistical Science

Beyond Multiple Linear Regression

Applied Generalized Linear Models and Multilevel Models in R



Paul Roback
Julie Legler

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

8

Introduction to Multilevel Models

8.1 Learning Objectives

After finishing this chapter, you should be able to:

- Recognize when response variables and covariates have been collected at multiple (nested) levels.
- Apply exploratory data analysis techniques to multilevel data.
- Write out a multilevel statistical model, including assumptions about variance components, in both by-level and composite forms.
- Interpret model parameters (including fixed effects and variance components) from a multilevel model, including cases in which covariates are continuous, categorical, or centered.
- Understand the taxonomy of models, including why we start with an unconditional means model.
- Select a final model, using criteria such as AIC, BIC, and deviance.

```
# Packages required for Chapter 8
library(MASS)
library(gridExtra)
library(mnormt)
library(lme4)
library(knitr)
library(kableExtra)
library(tidyverse)
```

8.2 Case Study: Music Performance Anxiety

Stage fright can be a serious problem for performers, and understanding the personality underpinnings of performance anxiety is an important step in determining how to minimize its impact. Sadler and Miller [2010] studied the emotional state of musicians before performances and factors which may affect their emotional state. Data was collected by having 37 undergraduate music majors from a competitive undergraduate music program fill out diaries prior to performances over the course of an academic year. In particular, study participants completed a Positive Affect Negative Affect Schedule (PANAS) before each performance. The PANAS instrument provided two key outcome measures: negative affect (a state measure of anxiety) and positive affect (a state measure of happiness). We will focus on negative affect as our primary response measuring performance anxiety.

Factors which were examined for their potential relationships with performance anxiety included: performance type (solo, large ensemble, or small ensemble); audience (instructor, public, students, or juried); if the piece was played from memory; age; gender; instrument (voice, orchestral, or keyboard); and, years studying the instrument. In addition, the personalities of study participants were assessed at baseline through the Multidimensional Personality Questionnaire (MPQ). The MPQ provided scores for one lower-order factor (absorption) and three higher-order factors: positive emotionality (PEM—a composite of well-being, social potency, achievement, and social closeness); negative emotionality (NEM—a composite of stress reaction, alienation, and aggression); and, constraint (a composite of control, harm avoidance, and traditionalism).

Primary scientific hypotheses of the researchers included:

- Lower music performance anxiety will be associated with lower levels of a subject's negative emotionality.
- Lower music performance anxiety will be associated with lower levels of a subject's stress reaction.
- Lower music performance anxiety will be associated with greater number of years of study.

TABLE 8.1: A snapshot of selected variables from the first three and the last three observations in the Music Performance Anxiety case study.

Obs	id	diary	perf_type	memory	na	gender	instrument	mpqab	mpqpem	mpqnem
1	1	1	Solo	Unspecified	11	Female	voice	16	52	16
2	1	2	Large Ensemble	Memory	19	Female	voice	16	52	16
3	1	3	Large Ensemble	Memory	14	Female	voice	16	52	16
495	43	2	Solo	Score	13	Female	voice	31	64	17
496	43	3	Small Ensemble	Memory	19	Female	voice	31	64	17
497	43	4	Solo	Score	11	Female	voice	31	64	17

8.3 Initial Exploratory Analyses

8.3.1 Data Organization

Our examination of the data from Sadler and Miller [2010] in `musicdata.csv` will focus on the following key variables:

- `id` = unique musician identification number
- `diary` = cumulative total of diaries filled out by musician
- `perf_type` = type of performance (Solo, Large Ensemble, or Small Ensemble)
- `audience` = who attended (Instructor, Public, Students, or Juried)
- `memory` = performed from Memory, using Score, or Unspecified
- `na` = negative affect score from PANAS
- `gender` = musician gender
- `instrument` = Voice, Orchestral, or Piano
- `mpqab` = absorption subscale from MPQ
- `mpqpem` = positive emotionality (PEM) composite scale from MPQ
- `mpqnem` = negative emotionality (NEM) composite scale from MPQ

Sample rows containing selected variables from our data set are illustrated in Table 8.1; note that each subject (id) has one row for each unique diary entry.

As with any statistical analysis, our first task is to explore the data, examining distributions of individual responses and predictors using graphical and numerical summaries, and beginning to discover relationships between variables. With multilevel models, exploratory analyses must eventually account for the level at which each variable is measured. In a two-level study such as this one, **Level One** will refer to variables measured at the most frequently occurring observational unit, while **Level Two** will refer to variables measured on larger observational units. For example, in our study on music performance anxiety, many variables are measured at every performance. These "Level One" variables include:

- negative affect (our response variable)
- performance characteristics (type, audience, if music was performed from memory)
- number of previous performances with a diary entry

However, other variables measure characteristics of study participants that remain constant over all performances for a particular musician; these are considered "Level Two" variables and include:

- demographics (age and gender of musician)
- instrument used and number of previous years spent studying that instrument
- baseline personality assessment (MPQ measures of positive emotionality, negative emotionality, constraint, stress reaction, and absorption)

8.3.2 Exploratory Analyses: Univariate Summaries

Because of this data structure—the assessment of some variables on a performance-by-performance basis and others on a subject-by-subject basis—we cannot treat our data set as consisting of 497 independent observations. Although negative affect measures from different subjects can reasonably be assumed to be independent (unless, perhaps, the subjects frequently perform in the same ensemble group), negative affect measures from different performances by the same subject are not likely to be independent. For example, some subjects tend to have relatively high performance anxiety across all performances, so that knowing their score for Performance 3 was 20 makes it more likely that their score for Performance 5 is somewhere near 20 as well. Thus, we must carefully consider our exploratory data analysis, recognizing that certain plots and summary statistics may be useful but imperfect in light of the correlated observations.

First, we will examine each response variable and potential covariate individually. Continuous variables can be summarized using histograms and summaries of center and spread; categorical variables can be summarized with tables and possibly bar charts. When examining Level One covariates and responses, we will begin by considering all 497 observations, essentially treating each performance by each subject as independent even though we expect observations from the same musician to be correlated. Although these plots will contain dependent points, since each musician provides data for up to 15 performances, general patterns exhibited in these plots tend to be real. Alternatively, we can calculate mean scores across all performances for each of the 37 musicians so that we can more easily consider each plotted point to be independent. The disadvantage of this approach would be lost information which, in a study such as this with a relatively small number of musicians each being observed over many performances, could be considerable. In addition, if the sample sizes varied greatly by subject, a mean based on 1 observation would be given

equal weight to a mean based on 15 observations. Nevertheless, both types of exploratory plots typically illustrate similar relationships.

In Figure 8.1 we see histograms for the primary response (negative affect); plot (a) shows all 497 (dependent) observations, while plot (b) shows the mean negative affect for each of the 37 musicians across all their performances. Through plot (a), we see that performance anxiety (negative affect) across all performances follows a right-skewed distribution with a lower bound of 10 (achieved when all 10 questions are answered with a 1). Plot (b) shows that mean negative affect is also right-skewed (although not as smoothly decreasing in frequency), with range 12 to 23.

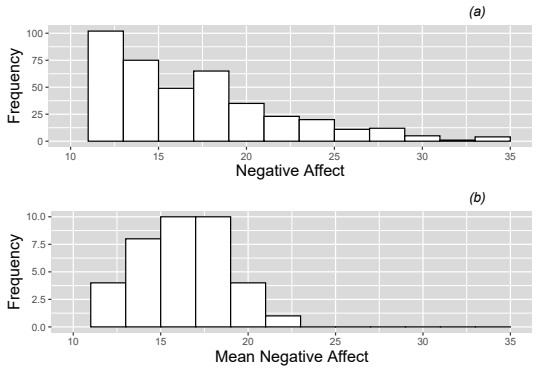


FIGURE 8.1: Histogram of the continuous Level One response (negative effect). Plot (a) contains all 497 performances across the 37 musicians, while plot (b) contains one observation per musician (the mean negative affect across all performances).

We can also summarize categorical Level One covariates across all (possibly correlated) observations to get a rough relative comparison of trends. A total of 56.1% of the 497 performances in our data set were solos, while 27.3% were large ensembles and 16.5% were small ensembles. The most common audience type was a public performance (41.0%), followed by instructors (30.0%), students (20.1%), and finally juried recitals (8.9%). In 30.0% of performances, the musician played by memory, while 55.1% used the score and 14.9% of performances were unspecified.

To generate an initial examination of Level Two covariates, we consider a data set with just one observation per subject, since Level Two variables are constant over all performances from the same subject. Then, we can proceed as we did with Level One covariates—using histograms to illustrate the distributions of continuous covariates (see Figure 8.2) and tables to summarize categorical covariates. For example, we learn that the majority of subjects have positive emotionality scores between 50 and 60, but that several subjects fall into a

lengthy lower tail with scores between 20 and 50. A summary of categorical Level Two covariates reveals that among the 37 subjects (26 female and 11 male), 17 play an orchestral instrument, 15 are vocal performers, and 5 play a keyboard instrument.

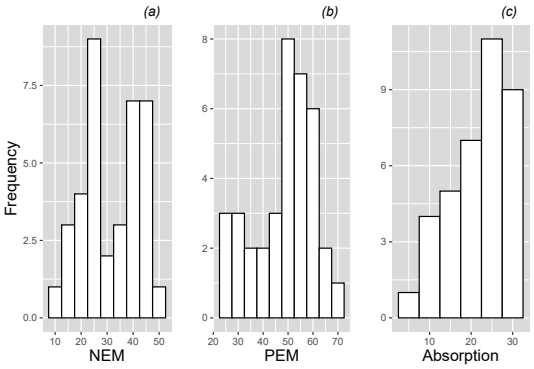


FIGURE 8.2: Histograms of the 3 continuous Level Two covariates (negative emotionality (NEM), positive emotionality (PEM), and absorption). Each plot contains one observation per musician.

8.3.3 Exploratory Analyses: Bivariate Summaries

The next step in an initial exploratory analysis is the examination of numerical and graphical summaries of relationships between model covariates and responses. In examining these bivariate relationships, we hope to learn: (1) if there is a general trend suggesting that as the covariate increases the response either increases or decreases, (2) if subjects at certain levels of the covariate tend to have similar mean responses (low variability), and (3) if the variation in the response differs at different levels of the covariate (unequal variability).

As with individual variables, we will begin by treating all 497 performances recorded as independent observations, even though blocks of 15 or so performances were performed by the same musician. For categorical Level One covariates, we can generate boxplots against negative affect as in Figure 8.3, plots (a) and (b). From these boxplots, we see that lower levels of performance anxiety seem to be associated with playing in large ensembles and playing in front of an instructor. For our lone continuous Level One covariate (number of previous performances), we can generate a scatterplot against negative affect as in plot (c) from Figure 8.3, adding a fitted line to illustrate general trends upward or downward. From this scatterplot, we see that negative affect seems to decrease slightly as a subject has more experience.

To avoid the issue of dependent observations in our three plots from Figure 8.3, we could generate separate plots for each subject and examine trends within and across subjects. These "lattice plots" are illustrated in Figures 8.4, 8.5, and 8.6; we discuss such plots more thoroughly in Chapter 9. While general trends are difficult to discern from these lattice plots, we can see the variety in subjects in sample size distributions and overall level of performance anxiety. In particular, in Figure 8.6, we notice that linear fits for many subjects illustrate the same slight downward trend displayed in the overall scatterplot in Figure 8.3, although some subjects experience increasing anxiety and others exhibit non-linear trends. Having an idea of the range of individual trends will be important when we begin to draw overall conclusions from this study.

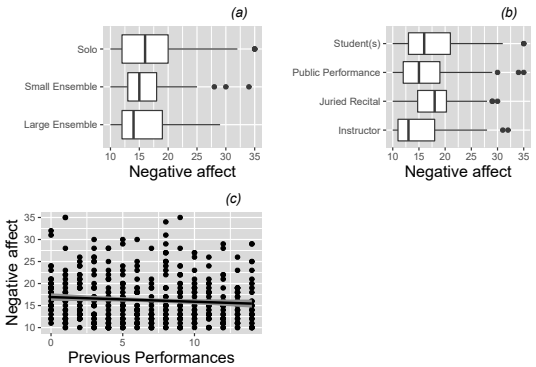


FIGURE 8.3: Boxplots of two categorical Level One covariates (performance type (a) and audience type (b)) vs. model response, and scatterplot of one continuous Level One covariate (number of previous diary entries (c)) vs. model response (negative affect). Each plot contains one observation for each of the 497 performances.

In Figure 8.7, we use boxplots to examine the relationship between our primary categorical Level Two covariate (instrument) and our continuous model response. Plot (a) uses all 497 performances, while plot (b) uses one observation per subject (the mean performance anxiety across all performances) regardless of how many performances that subject had. Naturally, plot (b) has a more condensed range of values, but both plots seem to support the notion that performance anxiety is slightly lower for vocalists and maybe a bit higher for keyboardists.

In Figure 8.8, we use scatterplots to examine the relationships between continuous Level Two covariates and our model response. Performance anxiety appears to vary little with a subject's positive emotionality, but there is some evidence to suggest that performance anxiety increases with increasing negative emotionality and absorption level. Plots based on mean negative affect, with one observation per subject, support conclusions based on plots with all

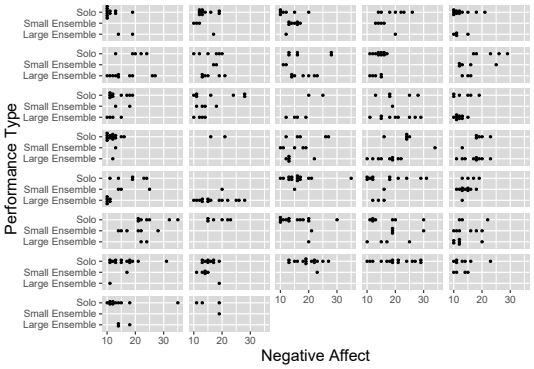


FIGURE 8.4: Lattice plot of performance type vs. negative affect, with separate dotplots by subject.

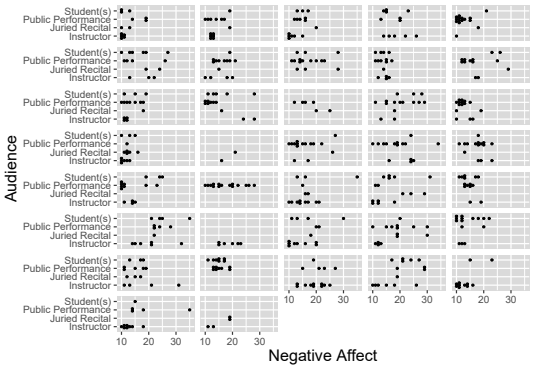


FIGURE 8.5: Lattice plot of audience type vs. negative affect, with separate dotplots by subject.

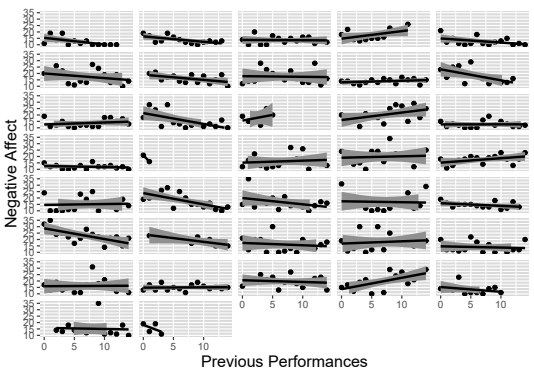


FIGURE 8.6: Lattice plot of previous performances vs. negative affect, with separate scatterplots with fitted lines by subject.

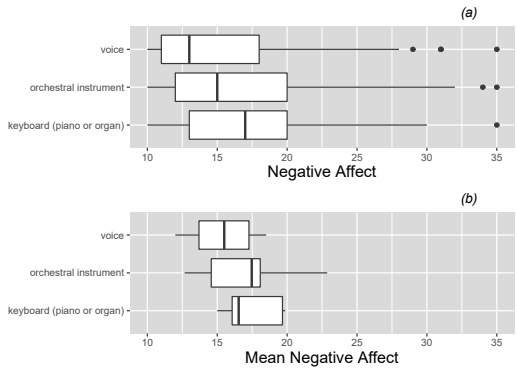


FIGURE 8.7: Boxplots of the categorical Level Two covariate (instrument) vs. model response (negative affect). Plot (a) is based on all 497 observations from all 37 subjects, while plot (b) uses only one observation per subject.

observations from all subjects; indeed the overall relationships are in the same direction and of the same magnitude.

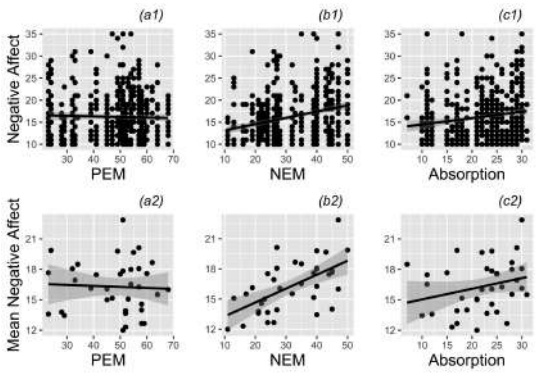


FIGURE 8.8: Scatterplots of continuous Level Two covariates (positive emotionality (PEM), negative emotionality (NEM), and absorption) vs. model response (negative affect). The top plots (a1, b1, c1) are based on all 497 observations from all 37 subjects, while the bottom plots (a2, b2, c2) use only one observation per subject.

Of course, any graphical analysis is exploratory, and any notable trends at this stage should be checked through formal modeling. At this point, a statistician begins to ask familiar questions such as:

- which characteristics of individual performances are most associated with performance anxiety?

- which characteristics of study participants are most associated with performance anxiety?
- are any of these associations statistically significant?
- does the significance remain after controlling for other covariates?
- how do we account for the lack of independence in performances by the same musician?

As you might expect, answers to these questions will arise from proper consideration of variability and properly identified statistical models.

8.4 Two-Level Modeling: Preliminary Considerations

8.4.1 Ignoring the Two-Level Structure (not recommended)

Armed with any statistical software package, it would be relatively simple to take our complete data set of 497 observations and run a multiple linear least squares regression model seeking to explain variability in negative affect with a number of performance-level or musician-level covariates. As an example, output from a model with two binary covariates (Does the subject play an orchestral instrument? Was the performance a large ensemble?) is presented below. Do you see any problems with this approach?

```
# Linear least square regression model with LINE conditions
modelc0 <- lm(na ~ orch + large + orch:large, data = music)
```

```
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  15.7212     0.3591 43.7785 5.548e-172
## orch          1.7887     0.5516  3.2426  1.265e-03
## large        -0.2767     0.7910 -0.3498  7.266e-01
## orch:large   -1.7087     1.0621 -1.6088  1.083e-01

##  R squared =  0.02782
##  Residual standard error =  5.179
```

Other than somewhat skewed residuals, residual plots (not shown) do not indicate any major problems with the LLSR model. However, another key assumption in these models is the independence of all observations. While we might reasonably conclude that responses from different study participants are

independent (although possibly not if they are members of the same ensemble group), it is not likely that the 15 or so observations taken over multiple performances from a single subject are similarly independent. If a subject begins with a relatively high level of anxiety (compared to other subjects) before their first performance, chances are good that they will have relatively high anxiety levels before subsequent performances. Thus, multiple linear least squares regression using all 497 observations is not advisable for this study (or multilevel data sets in general).

8.4.2 A Two-Stage Modeling Approach (better but imperfect)

If we assume that the 37 study participants can reasonably be considered to be independent, we could use traditional linear least squares regression techniques to analyze data from this study if we could condense each subject's set of responses to a single meaningful outcome. Candidates for this meaningful outcome include a subject's last performance anxiety measurement, average performance anxiety, minimum anxiety level, etc. For example, in clinical trials, data is often collected over many weekly or monthly visits for each patient, except that many patients will drop out early for many reasons (e.g., lack of efficacy, side effects, personal reasons). In these cases, treatments are frequently compared using "last-value-carried-forward" methods—the final visit of each patient is used as the primary outcome measure, regardless of how long they remained in the study. However, "last-value-carried-forward" and other summary measures feel inadequate, since we end up ignoring much of the information contained in the multiple measures for each individual. A more powerful solution is to model performance anxiety at multiple levels.

We will begin by considering all performances by a single individual. For instance, consider the 15 performances for which Musician #22 recorded a diary, illustrated in Table 8.2.

Does this musician tend to have higher anxiety levels when he is playing in a large ensemble or playing in front of fellow students? Which factor is the biggest determinant of anxiety for a performance by Musician #22? We can address these questions through multiple LLSR applied to only Musician #22's data, using appropriate indicator variables for factors of interest.

Let Y_{22j} be the performance anxiety score of Musician #22 before performance j. Consider the observed performances for Musician #22 to be a random sample of all conceivable performances by that subject. If we are initially interested in examining the effect of playing in a large ensemble, we can model the performance anxiety for Musician #22 according to the model:

$$Y_{22j} = a_{22} + b_{22}\text{LargeEns}_{22j} + \epsilon_{22j} \text{ where } \epsilon_{22j} \sim N(0, \sigma^2) \text{ and} \qquad (8.1)$$

TABLE 8.2: Data from the 15 performances of Musician #22.

	id	diary	perform_type	audience	na	instrument
240	22	1	Solo	Instructor	24	orchestral instrument
241	22	2	Large Ensemble	Public Performance	21	orchestral instrument
242	22	3	Large Ensemble	Public Performance	14	orchestral instrument
243	22	4	Large Ensemble	Public Performance	15	orchestral instrument
244	22	5	Large Ensemble	Public Performance	10	orchestral instrument
245	22	6	Solo	Instructor	24	orchestral instrument
246	22	7	Solo	Student(s)	24	orchestral instrument
247	22	8	Solo	Instructor	16	orchestral instrument
248	22	9	Small Ensemble	Public Performance	34	orchestral instrument
249	22	10	Large Ensemble	Public Performance	22	orchestral instrument
250	22	11	Large Ensemble	Public Performance	19	orchestral instrument
251	22	12	Large Ensemble	Public Performance	18	orchestral instrument
252	22	13	Large Ensemble	Public Performance	12	orchestral instrument
253	22	14	Large Ensemble	Public Performance	19	orchestral instrument
254	22	15	Solo	Instructor	25	orchestral instrument

$$\text{LargeEns}_j = \begin{cases} 1 & \text{if perf-type = Large Ensemble} \\ 0 & \text{if perf-type = Solo or Small Ensemble} \end{cases}$$

The parameters in this model (a_{22}, b_{22}, and σ^2) can be estimated through least squares methods. a_{22} represents the true intercept for Musician #22—the expected anxiety score for Musician #22 when performance type is a Solo or Small Ensemble (LargeEns = 0), or the true average anxiety for Musician #22 over all Solo or Small Ensemble performances he may conceivably give. b_{22} represents the true slope for Musician #22—the expected increase in performance anxiety for Musician #22 when performing as part of a Large Ensemble rather than in a Small Ensemble or as a Solo, or the true average difference in anxiety scores for Musician #22 between Large Ensemble performances and other types. Finally, the ϵ_{22j} terms represent the deviations of Musician #22's actual performance anxiety scores from the expected scores under this model—the part of Musician #22's anxiety before performance j that is not explained by performance type. The variability in these deviations from the regression model is denoted by σ^2.

For Subject 22, we estimate $\hat{a}_{22} = 24.5$, $\hat{b}_{22} = -7.8$, and $\hat{\sigma} = 4.8$. Thus, according to our simple linear regression model, Subject 22 had an estimated anxiety score of 24.5 before Solo and Small Ensemble performances, and 16.7 (7.8 points lower) before Large Ensemble performances. With an R^2 of 0.425, the regression model explains a moderate amount (42.5%) of the performance-to-performance variability in anxiety scores for Subject 22, and the trend toward lower scores for large ensemble performances is statistically significant at the 0.05 level (t(13)=-3.10, p=.009).

```
regr.id22 = lm(na ~ large, data = id22)
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   24.500       1.96  12.503 1.275e-08
## large         -7.833       2.53  -3.097 8.504e-03
```

```
##  R squared =  0.4245
##  Residual standard error =  4.8
```

We could continue trying to build a better model for Subject 22, adding indicators for audience and memory, and even adding a continuous variable representing the number of previous performances where a diary was kept. As our model R-squared value increased, we would be explaining a larger proportion of Subject 22's performance-to-performance variability in anxiety. It would not, however, improve our model to incorporate predictors for age, gender, or even negative emotionality based on the MPQ—why is that?

For the present time, we will model Subject 22's anxiety scores for his 15 performances using the model given by Equation (8.1), with a lone indicator variable for performing in a Large Ensemble. We can then proceed to fit the LLSR model in Equation (8.1) to examine the effect of performing in a Large Ensemble for each of the 37 subjects in this study. These are called **Level One models**. As displayed in Figure 8.9, there is considerable variability in the fitted intercepts and slopes among the 37 subjects. Mean performance anxiety scores for Solos and Small Ensembles range from 11.6 to 24.5, with a median score of 16.7, while mean differences in performance anxiety scores for Large Ensembles compared to Solos and Small Ensembles range from -7.9 to 5.0, with a median difference of -1.7. Can these differences among individual musicians be explained by (performance-invariant) characteristics associated with each individual, such as gender, age, instrument, years studied, or baseline levels of personality measures? Questions like these can be addressed through further statistical modeling.

As an illustration, we can consider whether or not there are significant relationships between individual regression parameters (intercepts and slopes) and instrument played. From a modeling perspective, we would build a system of two **Level Two models** to predict the fitted intercept (a_i) and fitted slopes (b_i) for Subject i:

$$a_i = \alpha_0 + \alpha_1 \text{Orch}_i + u_i \tag{8.2}$$

$$b_i = \beta_0 + \beta_1 \text{Orch}_i + v_i \tag{8.3}$$

where $\text{Orch}_i = 1$ if Subject i plays an orchestral instrument and $\text{Orch}_i = 0$ if Subject i plays keyboard or is a vocalist. Note that, at Level Two, our response variables are not observed measurements such as performance anxiety scores,

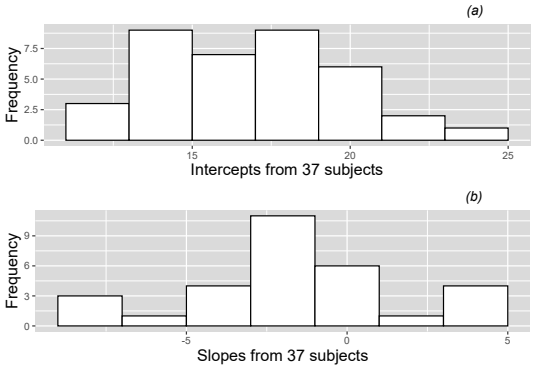


FIGURE 8.9: Histograms of intercepts and slopes from fitting simple regression models by subject, where each model contained a single binary predictor indicating if a performance was part of a large ensemble.

but rather the fitted regression coefficients from the Level One models fit to each subject. (Well, in our theoretical model, the responses are actually the true intercepts and slopes from Level One models for each subject, but in reality, we have to use our estimated slopes and intercepts.)

Exploratory data analysis (see boxplots by instrument in Figure 8.10) suggests that subjects playing orchestral instruments have higher intercepts than vocalists or keyboardists, and that orchestral instruments are associated with slightly lower (more negative) slopes, although with less variability that the slopes of vocalists and keyboardists. These trends are borne out in regression modeling. If we fit Equations (8.2) and (8.3) using fitted intercepts and slopes as our response variables, we obtain the following estimated parameters: $\hat{\alpha}_0 = 16.3$, $\hat{\alpha}_1 = 1.4$, $\hat{\beta}_0 = -0.8$, and $\hat{\beta}_1 = -1.4$. Thus, the intercept (a_i) and slope (b_i) for Subject i can be modeled as:

$$\begin{aligned} \hat{a}_i &= 16.3 + 1.4\text{Orch}_i + u_i \\ \hat{b}_i &= -0.8 - 1.4\text{Orch}_i + v_i \end{aligned} \tag{8.4}$$

where a_i is the true mean negative affect when Subject i is playing solos or small ensembles, and b_i is the true mean difference in negative affect for Subject i between large ensembles and other performance types. Based on these models, average performance anxiety before solos and small ensembles is 16.3 for vocalists and keyboardists, but 17.7 (1.4 points higher) for orchestral instrumentalists. Before playing in large ensembles, vocalists and instrumentalists have performance anxiety (15.5) which is 0.8 points lower, on average, than before solos and small ensembles, while subjects playing orchestral instruments experience an average difference of 2.2 points, producing an average

performance anxiety of 15.5 before playing in large ensembles just like subjects playing other instruments. However, the difference between orchestral instruments and others does not appear to be statistically significant for either intercepts (t=1.424, p=0.163) or slopes (t=-1.168, p=0.253).

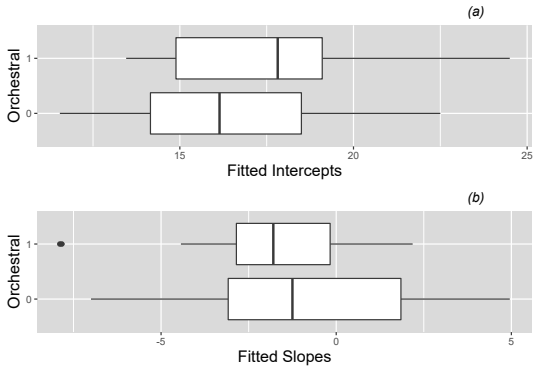


FIGURE 8.10: Boxplots of fitted intercepts, plot (a), and slopes, plot (b), by orchestral instrument (1) vs. keyboard or vocalist (0).

This two-stage modeling process does have some drawbacks. Among other things, (1) it weights every subject the same regardless of the number of diary entries we have, (2) it responds to missing individual slopes (from 7 subjects who never performed in a large ensemble) by simply dropping those subjects, and (3) it does not share strength effectively across individuals. These issues can be better handled through a unified multilevel modeling framework which we will develop in the next section.

8.5 Two-Level Modeling: A Unified Approach

8.5.1 Our Framework

For the unified approach, we will still envision two levels of models as in Section 8.4.2, but we will use likelihood-based methods for parameter estimation rather than ordinary least squares to address the drawbacks associated with the two-stage approach. To illustrate the unified approach, we will first generalize the models presented in Section 8.4.2. Let Y_{ij} be the performance anxiety score of the i^{th} subject before performance j. If we are initially interested in examining the effects of playing in a large ensemble and playing an orchestral instrument, then we can model the performance anxiety for Subject i in performance j with the following system of equations:

- Level One:

$$Y_{ij} = a_i + b_i \text{LargeEns}_{ij} + \epsilon_{ij}$$

- Level Two:

$$a_i = \alpha_0 + \alpha_1 \text{Orch}_i + u_i$$
$$b_i = \beta_0 + \beta_1 \text{Orch}_i + v_i,$$

In this system, there are 4 key **fixed effects** to estimate: α_0, α_1, β_0 and β_1. Fixed effects are the fixed but unknown population effects associated with certain covariates. The intercepts and slopes for each subject from Level One, a_i and b_i, don't need to be formally estimated as we did in Section 8.4.2; they serve to conceptually connect Level One with Level Two. In fact, by substituting the two Level Two equations into the Level One equation, we can view this two-level system of models as a single **Composite Model** without a_i and b_i:

$$\begin{aligned} Y_{ij} = &[\alpha_0 + \alpha_1 \text{Orch}_i + \beta_0 \text{LargeEns}_{ij} + \beta_1 \text{Orch}_i \text{LargeEns}_{ij}] \\ &+ [u_i + v_i \text{LargeEns}_{ij} + \epsilon_{ij}] \end{aligned}$$

From this point forward, when building multilevel models, we will use Greek letters (such as α_0) to denote final fixed effects model parameters to be estimated empirically, and Roman letters (such as a_0) to denote preliminary fixed effects parameters at lower levels. Variance components that will be estimated empirically will be denoted with σ or ρ, while terms such as ϵ and u_i represent error terms. In our framework, we can estimate final parameters directly without first estimating preliminary parameters, which can be seen with the Composite Model formulation (although we can obtain estimates of preliminary parameters in those occasional cases when they are of interest to us). Note that when we model a slope term like b_i from Level One using Level Two covariates like Orch_i, the resulting Composite Model contains a **cross-level interaction term**, denoting that the effect of LargeEns_{ij} depends on the instrument played.

Furthermore, with a binary predictor at Level Two such as instrument, we can write out what our Level Two model looks like for those who play keyboard or are vocalists ($\text{Orch}_i = 0$) and those who play orchestral instruments ($\text{Orch}_i = 1$):

- Keyboardists and Vocalists ($\text{Orch}_i = 0$)

$$a_i = \alpha_0 + u_i$$
$$b_i = \beta_0 + v_i$$

- Orchestral Instrumentalists ($\text{Orch}_i = 1$)

$$a_i = (\alpha_0 + \alpha_1) + u_i$$
$$b_i = (\beta_0 + \beta_1) + v_i$$

Writing the Level Two model in this manner helps us interpret the model parameters from our two-level model. In this case, even the Level One covariate is binary, so that we can write out expressions for mean performance anxiety based on our model for four different combinations of instrument played and performance type:

- Keyboardists or vocalists playing solos or small ensembles: α_0
- Keyboardists or vocalists playing large ensembles: $\alpha_0 + \beta_0$
- Orchestral instrumentalists playing solos or small ensembles: $\alpha_0 + \alpha_1$
- Orchestral instrumentalists playing large ensembles: $\alpha_0 + \alpha_1 + \beta_0 + \beta_1$

8.5.2 Random vs. Fixed Effects

Before we can use likelihood-based methods to estimate our model parameters, we still must define the distributions of our error terms. The error terms ϵ_{ij}, u_i, and v_i represent random effects in our model. In multilevel models, it is important to distinguish between fixed and random effects. Typically, **fixed effects** describe levels of a factor that we are specifically interested in drawing inferences about, and which would not change in replications of the study. For example, in our music performance anxiety case study, the levels of performance type will most likely remain as solos, small ensembles, and large ensembles even in replications of the study, and we wish to draw specific conclusions about differences between these three types of performances. Thus, performance type would be considered a fixed effect. On the other hand, **random effects** describe levels of a factor which can be thought of as a sample from a larger population of factor levels; we are not typically interested in drawing conclusions about specific levels of a random effect, although we are interested in accounting for the influence of the random effect in our model. For example, in our case study, the different musicians included can be thought of as a random sample from a population of performing musicians. Although our goal is not to make specific conclusions about differences between any two musicians, we do want to account for inherent differences among musicians in our model, and by doing so, we will be able to draw more precise conclusions about our fixed effects of interest. Thus, musician would be considered a random effect.

8.5.3 Distribution of Errors: Multivariate Normal

As part of our multilevel model, we must provide probability distributions to describe the behavior of random effects. Typically, we assume that random

effects follow a normal distribution with mean 0 and a variance parameter which must be estimated from the data. For example, at Level One, we will assume that the errors associated with each performance of a particular musician can be described as: $\epsilon_{ij} \sim N(0, \sigma^2)$. At Level Two, we have one error term (u_i) associated with subject-to-subject differences in intercepts, and one error term (v_i) associated with subject-to-subject differences in slopes. That is, u_i represents the deviation of Subject i from the mean performance anxiety before solos and small ensembles after accounting for their instrument, and v_i represents the deviation of Subject i from the mean difference in performance anxiety between large ensembles and other performance types after accounting for their instrument.

In modeling the random behavior of u_i and v_i, we must also account for the possibility that random effects at the same level might be correlated. Subjects with higher baseline performance anxiety have a greater capacity for showing decreased anxiety in large ensembles as compared to solos and small ensembles, so we might expect that subjects with larger intercepts (performance anxiety before solos and small ensembles) would have smaller slopes (indicating greater decreases in anxiety before large ensembles). In fact, our fitted Level One intercepts and slopes in this example actually show evidence of a fairly strong negative correlation ($r = -0.525$, see Figure 8.11).

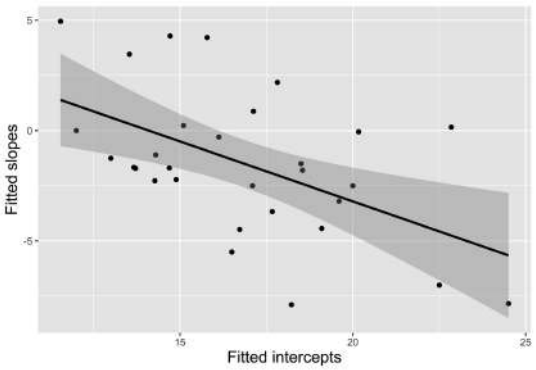


FIGURE 8.11: Scatterplot with fitted regression line for estimated intercepts and slopes (one point per subject).

To allow for this correlation, the error terms at Level Two can be assumed to follow a **multivariate normal distribution** in our unified multilevel model. Mathematically, we can express this as:

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \rho_{uv}\sigma_u\sigma_v \\ \rho_{uv}\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix} \right)$$

where σ_u^2 is the variance of the u_i terms, σ_v^2 is the variance of the v_i terms,

and $\sigma_{uv} = \rho_{uv}\sigma_u\sigma_v$ is the covariance between the u_i and the v_i terms (i.e., how those two terms vary together).

Note that the correlation ρ_{uv} between the error terms is simply the covariance $\sigma_{uv} = \rho_{uv}\sigma_u\sigma_v$ converted to a $[-1, 1]$ scale through the relationship:

$$\rho_{uv} = \frac{\sigma_{uv}}{\sigma_u\sigma_v}$$

With this expression, we are allowing each error term to have its own variance (around a mean of 0) and each pair of error terms to have its own covariance (or correlation). Thus, if there are n equations at Level Two, we can have n variance terms and $n(n-1)/2$ covariance terms for a total of $n + n(n-1)/2$ variance components. These variance components are organized in matrix form, with variance terms along the diagonal and covariance terms in the off-diagonal. In our small example, we have $n = 2$ equations at Level Two, so we have 3 variance components to estimate—2 variance terms (σ_u^2 and σ_v^2) and 1 correlation (ρ_{uv}).

The multivariate normal distribution with $n = 2$ is illustrated in Figure 8.12 for two cases: (a) the error terms are uncorrelated ($\sigma_{uv} = \rho_{uv} = 0$), and (b) the error terms are positively correlated ($\sigma_{uv} > 0$ and $\rho_{uv} > 0$). In general, if the errors in intercepts (u_i) are placed on the x-axis and the errors in slopes (v_i) are placed on the y-axis, then σ_u^2 measures spread in the x-direction and σ_v^2 measures spread in the y-direction, while σ_{uv} measures tilt. Positive tilt ($\sigma_{uv} > 0$) indicates a tendency for errors from the same subject to both be positive or both be negative, while negative tilt ($\sigma_{uv} < 0$) indicates a tendency for one error from a subject to be positive and the other to be negative. In Figure 8.12, $\sigma_u^2 = 4$ and $\sigma_v^2 = 1$, so both contour plots show a greater range of errors in the x-direction than the y-direction. Ellipses near the center of the contour plot indicate pairs of u_i and v_i that are more likely. In Figure 8.12 (a) $\sigma_{uv} = \rho_{uv} = 0$, so the axes of the contour plot correspond to the x- and y-axes, but in Figure 8.12 (b) $\sigma_{uv} = 1.5$, so the contour plot tilts up, reflecting a tendency for high values of u_i to be associated with high values of v_i.

8.5.4 Technical Issues when Testing Parameters (optional)

Now, our relatively simple two-level model has 8 parameters that need to be estimated: 4 fixed effects (α_0, α_1, β_0, and β_1), and 4 variance components (σ^2, σ_u^2, σ_v^2, and σ_{uv}). Note that we use the term **variance components** to signify model parameters that describe the behavior of random effects. We can use statistical software, such as the lmer() function from the lme4 package in R, to obtain parameter estimates using our 497 observations. The most common methods for estimating model parameters—both fixed effects and variance components—are maximum likelihood (ML) and **restricted maximum likelihood (REML)**. The method of ML was introduced in

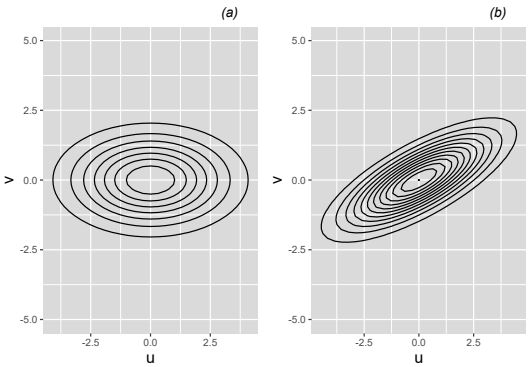


FIGURE 8.12: Contour plots illustrating a multivariate normal density with (a) no correlation between error terms, and (b) positive correlation between error terms.

Chapter 2, where parameter estimates are chosen to maximize the value of the likelihood function based on observed data. REML is conditional on the fixed effects, so that the part of the data used for estimating variance components is separated from that used for estimating fixed effects. Thus REML, by accounting for the loss in degrees of freedom from estimating the fixed effects, provides an unbiased estimate of variance components, while ML estimators for variance components are biased under assumptions of normality, since they use estimated fixed effects rather than the true values. REML is preferable when the number of parameters is large or the primary interest is obtaining estimates of model parameters, either fixed effects or variance components associated with random effects. ML should be used if nested fixed effects models are being compared using a likelihood ratio test, although REML is fine for nested models of random effects (with the same fixed effects model). In this text, we will typically report REML estimates unless we are specifically comparing nested models with the same random effects. In most case studies and most models we consider, there is very little difference between ML and REML parameter estimates. Additional details are beyond the scope of this book [Singer and Willett, 2003].

Note that the multilevel output shown beginning in the next section contains no p-values for performing hypothesis tests. This is primarily because the exact distribution of the test statistics under the null hypothesis (no fixed effect) is unknown, primarily because the exact degrees of freedom is not known [Bates et al., 2015]. Finding good approximate distributions for test statistics (and thus good approximate p-values) in multilevel models is an area of active research. In most cases, we can simply conclude that t-values (ratios of parameter estimates to estimated standard errors) with absolute value above 2 indicate significant

evidence that a particular model parameter is different than 0. Certain software packages will report p-values corresponding to hypothesis tests for parameters of fixed effects; these packages are typically using conservative assumptions, large-sample results, or approximate degrees of freedom for a t-distribution. In Section 1.6.5, we introduced the bootstrap as a non-parametric, computational approach for producing confidence intervals for model parameters. In addition, in Section 9.6.4, we will introduce a method called the parametric bootstrap which is being used more frequently by researchers to better approximate the distribution of the likelihood test statistic and produce more accurate p-values by simulating data under the null hypothesis [Efron, 2012].

8.5.5 An Initial Model with Parameter Interpretations

The output below contains likelihood-based estimates of our 8 parameters from a two-level model applied to the music performance anxiety data:

```
   Linear mixed model fit by REML ['lmerMod']
A)  Formula: na ~ orch + large + orch:large + (large | id)
       Data: music
B)  REML criterion at convergence: 2987

B2)      AIC      BIC   logLik deviance df.resid
        3007     3041    -1496     2991      489

    Random effects:
     Groups   Name        Variance Std.Dev. Corr
C)   id       (Intercept)  5.655   2.378
D)            large        0.452   0.672    -0.63
E)   Residual             21.807   4.670
F)  Number of obs: 497, groups:  id, 37

    Fixed effects:
                Estimate Std. Error t value
G)  (Intercept)   15.930      0.641   24.83
H)  orch           1.693      0.945    1.79
I)  large         -0.911      0.845   -1.08
J)  orch:large    -1.424      1.099   -1.30
```

This output (except for the capital letters along the left column) was specifically generated by the lmer() function in R; multilevel modeling results from other packages will contain similar elements. Because we will use lmer() output to summarize analyses of case studies in this and following sections, we will spend a little time now orienting ourselves to the most important features in this output.

- A: How our multilevel model is written in R, based on the composite model formulation. For more details, see Section 8.12.
- B: Measures of model performance. Since this model was fit using REML, this line only contains the REML criterion.
- B2: If the model is fit with ML instead of REML, the measures of performance will contain AIC, BIC, deviance, and the log-likelihood.
- C: Estimated variance components ($\hat{\sigma}_u^2$ and $\hat{\sigma}_u$) associated with the intercept equation in Level Two.
- D: Estimated variance components ($\hat{\sigma}_v^2$ and $\hat{\sigma}_v$) associated with the large ensemble effect equation in Level Two, along with the estimated correlation ($\hat{\rho}_{uv}$) between the two Level Two error terms.
- E: Estimated variance components ($\hat{\sigma}^2$ and $\hat{\sigma}$) associated with the Level One equation.
- F: Total number of performances where data was collected (Level One observations = 497) and total number of subjects (Level Two observations = 37).
- G: Estimated fixed effect ($\hat{\alpha}_0$) for the intercept term, along with its standard error and t-value (which is the ratio of the estimated coefficient to its standard error). As described in Section 8.5.4, no p-value testing the significance of the coefficient is provided because the exact null distribution of the t-value is unknown.
- H: Estimated fixed effect ($\hat{\alpha}_1$) for the orchestral instrument effect, along with its standard error and t-value.
- I: Estimated fixed effect ($\hat{\beta}_0$) for the large ensemble effect, along with its standard error and t-value.
- J: Estimated fixed effect ($\hat{\beta}_1$) for the interaction between orchestral instruments and large ensembles, along with its standard error and t-value.

Assuming the 37 musicians in this study are representative of a larger population of musicians, parameter interpretations for our 8 model parameters are given below:

- Fixed effects:
  - $\hat{\alpha}_0 = 15.9$. The estimated mean performance anxiety for solos and small ensembles (Large=0) for keyboard players and vocalists (Orch=0) is 15.9.
  - $\hat{\alpha}_1 = 1.7$. Orchestral instrumentalists have an estimated mean performance anxiety for solos and small ensembles which is 1.7 points higher than keyboard players and vocalists.
  - $\hat{\beta}_0 = -0.9$. Keyboard players and vocalists have an estimated mean decrease in performance anxiety of 0.9 points when playing in large ensembles instead of solos or small ensembles.
  - $\hat{\beta}_1 = -1.4$. Orchestral instrumentalists have an estimated mean decrease in performance anxiety of 2.3 points when playing in large ensembles

TABLE 8.3: Comparison of estimated coefficients and standard errors from the approaches mentioned in this section.

Variable	Independence	TwoStage	LVCF	Multilevel
Intercept	15.72(0.36)	16.28(0.67)	15.20(1.25)	15.93(0.64)
Orch	1.79(0.55)	1.41(0.99)	1.45(1.84)	1.69(0.95)
Large	-0.28(0.79)	-0.77(0.85)	-	-0.91(0.85)
Orch*Large	-1.71(1.06)	-1.41(1.20)	-	-1.42(1.10)

instead of solos or small ensembles, 1.4 points greater than the mean decrease among keyboard players and vocalists.

- Variance components
  - $\hat{\sigma}_u = 2.4$. The estimated standard deviation of performance anxiety levels for solos and small ensembles is 2.4 points, after controlling for instrument played.
  - $\hat{\sigma}_v = 0.7$. The estimated standard deviation of differences in performance anxiety levels between large ensembles and other performance types is 0.7 points, after controlling for instrument played.
  - $\hat{\rho}_{uv} = -0.64$. The estimated correlation between performance anxiety scores for solos and small ensembles and increases in performance anxiety for large ensembles is -0.64, after controlling for instrument played. Those subjects with higher performance anxiety scores for solos and small ensembles tend to have greater decreases in performance anxiety for large ensemble performances.
  - $\hat{\sigma} = 4.7$. The estimated standard deviation in residuals for the individual regression models is 4.7 points.

Table 8.3 shows a side-by-side comparison of estimated coefficients from the approaches described to this point. Underlying assumptions, especially regarding the error and correlation structure, differ, and differences in estimated effects are potentially meaningful. Note that some standard errors are greatly *underestimated* under independence, and that no Level One covariates (such as performance type) can be analyzed under a method such as last-visit-carried-forward which uses one observation per subject. Moving forward, we will employ the unified multilevel approach to maximize the information being used to estimate model parameters and to remain faithful to the structure of the data.

Two-level modeling as done with the music performance anxiety data usually involves fitting a number of models. Subsequent sections will describe a process of starting with the simplest two-level models and building toward a final model which addresses the research questions of interest.

8.6 Building a Multilevel Model

8.6.1 Model Building Strategy

Initially, it is advisable to first fit some simple, preliminary models, in part to establish a baseline for evaluating larger models. Then, we can build toward a final model for description and inference by attempting to add important covariates, centering certain variables, and checking model assumptions. In this study, we are particularly interested in Level Two covariates—those subject-specific variables that provide insight into why individuals react differently in anxiety-inducing situations. To get more precise estimates of the effect of Level Two covariates, we also want to control for Level One covariates that describe differences in individual performances.

Our strategy for building multilevel models will begin with extensive exploratory data analysis at each level. Then, after examining models with no predictors to assess variability at each level, we will first focus on creating a Level One model, starting simple and adding terms as necessary. Next, we will move to Level Two models, again starting simple and adding terms as necessary, beginning with the equation for the intercept term. Finally, we will examine the random effects and variance components, beginning with a full set of error terms and then removing covariance terms and variance terms where advisable (for instance, when parameter estimates are failing to converge or producing impossible or unlikely values). This strategy follows closely with that described by Raudenbush and Bryk [2002] and used by Singer and Willett [2003]. Singer and Willett further find that the modeled error structure rarely matters in practical contexts. Other model building approaches are certainly possible. Diggle et al. [2002], for example, begins with a saturated fixed effects model, determines variance components based on that, and then simplifies the fixed part of the model after fixing the random part.

8.6.2 An Initial Model: Random Intercepts

The first model fit in almost any multilevel context should be the **unconditional means model**, also called a **random intercepts model**. In this model, there are no predictors at either level; rather, the purpose of the unconditional means model is to assess the amount of variation at each level—to compare variability within subject to variability between subjects. Expanded models will then attempt to explain sources of between and within subject variability.

The unconditional means (random intercepts) model, which we will denote as Model A, can be specified either using formulations at both levels:

- Level One:
$$Y_{ij} = a_i + \epsilon_{ij} \text{ where } \epsilon_{ij} \sim N(0, \sigma^2)$$

- Level Two:
$$a_i = \alpha_0 + u_i \text{ where } u_i \sim N(0, \sigma_u^2)$$

or as a composite model:
$$Y_{ij} = \alpha_0 + u_i + \epsilon_{ij}$$

In this model, the performance anxiety scores of subject i are not a function of performance type or any other Level One covariate, so that a_i is the true mean response of all observations for subject i. On the other hand, α_0 is the grand mean – the true mean of all observations across the entire population. Our primary interest in the unconditional means model is the variance components – σ^2 is the within-person variability, while σ_u^2 is the between-person variability. The name **random intercepts model** then arises from the Level Two equation for a_i: each subject's intercept is assumed to be a random value from a normal distribution centered at α_0 with variance σ_u^2.

Using the composite model specification, the unconditional means model can be fit to the music performance anxiety data using statistical software:

```
#Model A (Unconditional means model)
model.a <- lmer(na ~ 1 + (1 | id), REML = T, data = music)
```

```
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept)  4.95    2.22
##  Residual             22.46    4.74

##  Number of Level Two groups =  37

##             Estimate Std. Error t value
## (Intercept)    16.24     0.4279   37.94
```

From this output, we obtain estimates of our three model parameters:

- $\hat{\alpha}_0 = 16.2 =$ the estimated mean performance anxiety score across all performances and all subjects.
- $\hat{\sigma}^2 = 22.5 =$ the estimated variance in within-person deviations.
- $\hat{\sigma}_u^2 = 5.0 =$ the estimated variance in between-person deviations.

The relative levels of between- and within-person variabilities can be compared through the **intraclass correlation coefficient**:

$$\hat{\rho} = \frac{\text{Between-person variability}}{\text{Total variability}} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}^2} = \frac{5.0}{5.0 + 22.5} = .182.$$

Thus, 18.2% of the total variability in performance anxiety scores are attributable to differences among subjects. In this particular model, we can also say that the average correlation for any pair of responses from the same individual is a moderately low .182. As ρ approaches 0, responses from an individual are essentially independent and accounting for the multilevel structure of the data becomes less crucial. However, as ρ approaches 1, repeated observations from the same individual essentially provide no additional information and accounting for the multilevel structure becomes very important. With ρ near 0, the **effective sample size** (the number of independent pieces of information we have for modeling) approaches the total number of observations, while with ρ near 1, the effective sample size approaches the number of subjects in the study.

8.7 Binary Covariates at Level One and Level Two

8.7.1 Random Slopes and Intercepts Model

The next step in model fitting is to build a good model for predicting performance anxiety scores at Level One (within subject). We will add potentially meaningful Level One covariates—those that vary from performance-to-performance for each individual. In this case, mirroring our model from Section 8.4 we will include a binary covariate for performance type:

$$\text{LargeEns}_{ij} = \begin{cases} 1 & \text{if perf-type = Large Ensemble} \\ 0 & \text{if perf-type = Solo or Small Ensemble} \end{cases}$$

and no other Level One covariates (for now). (Note that we may later also want to include an indicator variable for "Small Ensemble" to separate the effects of `Solo` performances and `Small Ensemble` performances.) The resulting model, which we will denote as Model B, can be specified either using formulations at both levels:

- Level One:
$$Y_{ij} = a_i + b_i \text{LargeEns}_{ij} + \epsilon_{ij}$$
- Level Two:
$$a_i = \alpha_0 + u_i$$
$$b_i = \beta_0 + v_i$$

or as a composite model:

$$Y_{ij} = [\alpha_0 + \beta_0 \text{LargeEns}_{ij}] + [u_i + v_i \text{LargeEns}_{ij} + \epsilon_{ij}]$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix} \right).$$

as discussed in Section 8.5.3.

In this model, performance anxiety scores for subject i are assumed to differ (on average) for Large Ensemble performances as compared with Solos and Small Ensemble performances; the ϵ_{ij} terms capture the deviation between the true performance anxiety levels for subjects (based on performance type) and their observed anxiety levels. α_0 is then the true mean performance anxiety level for Solos and Small Ensembles, and β_0 is the true mean difference in performance anxiety for Large Ensembles compared to other performance types. As before, σ^2 quantifies the within-person variability (the scatter of points around individuals' means by performance type), while now the between-person variability is partitioned into variability in Solo and Small Ensemble scores (σ_u^2) and variability in differences with Large Ensembles (σ_v^2).

Using the composite model specification, Model B can be fit to the music performance anxiety data, producing the following output:

```
#Model B (Add large as Level 1 covariate)
model.b <- lmer(na ~ large +  (large | id), data = music)
```

```
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept)  6.333   2.517
##           large        0.743   0.862    -0.76
##  Residual             21.771   4.666

##  Number of Level Two groups =  37

##             Estimate Std. Error t value
## (Intercept)   16.730     0.4908   34.09
## large         -1.676     0.5425   -3.09
```

From this output, we obtain estimates of our six model parameters (2 fixed effects and 4 variance components):

- $\hat{\alpha}_0 = 16.7$ = the mean performance anxiety level before solos and small ensemble performances.

- $\hat{\beta}_0 = -1.7$ = the mean decrease in performance anxiety before large ensemble performances.
- $\hat{\sigma}^2 = 21.8$ = the variance in within-person deviations.
- $\hat{\sigma}_u^2 = 6.3$ = the variance in between-person deviations in performance anxiety scores before solos and small ensembles.
- $\hat{\sigma}_v^2 = 0.7$ = the variance in between-person deviations in increases (or decreases) in performance anxiety scores before large ensembles.
- $\hat{\rho}_{uv} = -0.76$ = the correlation in subjects' anxiety before solos and small ensembles and their differences in anxiety between large ensembles and other performance types.

We see that, on average, subjects had a performance anxiety level of 16.7 before solos and small ensembles, and their anxiety levels were 1.7 points lower, on average, before large ensembles, producing an average performance anxiety level before large ensembles of 15.0. According to the t-value listed in R, the difference between large ensembles and other performance types is statistically significant (t=-3.09).

This random slopes and intercepts model is illustrated in Figure 8.13. The thicker black line shows the overall trends given by our estimated fixed effects: an intercept of 16.7 and a slope of -1.7. Then, each subject is represented by a gray line. Not only do the subjects' intercepts differ (with variance 6.3), but their slopes differ as well (with variance 0.7). Additionally, subjects' slopes and intercepts are negatively associated (with correlation -0.76), so that subjects with greater intercepts tend to have steeper negative slopes. We can compare this model with the random intercepts model from Section 8.6.2, pictured in Figure 8.14. With no effect of large ensembles, each subject is represented by a gray line with identical slopes (0) but varying intercepts (with variance 5.0).

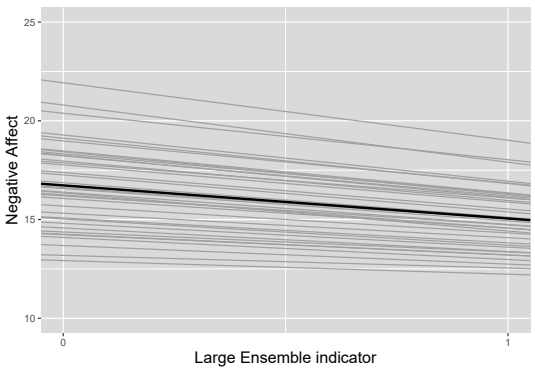


FIGURE 8.13: The random slopes and intercepts model fitted to the music performance anxiety data. Each gray line represents one subject, and the thicker black line represents the trend across all subjects.

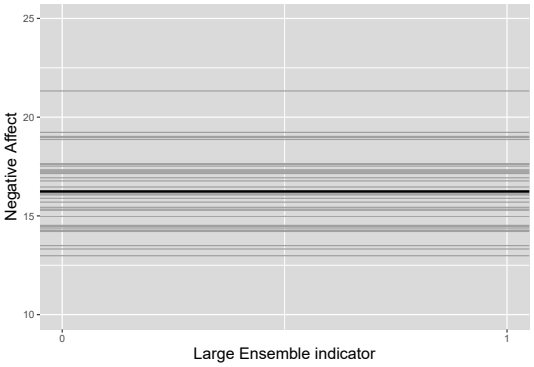


FIGURE 8.14: The random intercepts model fitted to the music performance anxiety data. Each gray line represents one subject, and the thicker black line represents the trend across all subjects.

Figures 8.13 and 8.14 use **empirical Bayes estimates** for the intercepts (a_i) and slopes (b_i) of individual subjects. Empirical Bayes estimates are sometimes called "shrinkage estimates" since they combine individual-specific information with information from all subjects, thus "shrinking" the individual estimates toward the group averages. Empirical Bayes estimates are often used when a term such as a_i involves both fixed and random components; further detail can be found in Raudenbush and Bryk [2002] and Singer and Willett [2003].

8.7.2 Pseudo R-squared Values

The estimated within-person variance $\hat{\sigma}^2$ decreased by 3.1% (from 22.5 to 21.8) from the unconditional means model, implying that only 3.1% of within-person variability in performance anxiety scores can be explained by performance type. This calculation is considered a **pseudo R-squared** value:

$$\text{Pseudo } R^2_{L1} = \frac{\hat{\sigma}^2(\text{Model A}) - \hat{\sigma}^2(\text{Model B})}{\hat{\sigma}^2(\text{Model A})} = \frac{22.5 - 21.8}{22.5} = 0.031$$

Values of $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_v$ from Model B cannot be compared to between-person variability from Model A, since the inclusion of performance type has changed the interpretation of these values, although they can provide important benchmarks for evaluating more complex Level Two predictions. Finally, $\hat{\rho}_{uv} = -0.76$ indicates a strong negative relationship between a subject's performance anxiety before solos and small ensembles and their (typical) decrease in performance anxiety before large ensembles. As might be expected, subjects with higher levels of performance anxiety before solos and small ensembles tend to have smaller increases (or greater decreases) in performance anxiety before large

ensembles; those with higher levels of performance anxiety before solos and small ensembles have more opportunity for decreases before large ensembles.

Pseudo R-squared values are not universally reliable as measures of model performance. Because of the complexity of estimating fixed effects and variance components at various levels of a multilevel model, it is not unusual to encounter situations in which covariates in a Level Two equation for, say, the intercept remain constant (while other aspects of the model change), yet the associated pseudo R-squared values differ or are negative. For this reason, pseudo R-squared values in multilevel models should be interpreted cautiously.

8.7.3 Adding a Covariate at Level Two

The initial two-level model described in Section 8.5.5 essentially expands upon the random slopes and intercepts model by adding a binary covariate for instrument played at Level Two. We will denote this as Model C:

- Level One:
$$Y_{ij} = a_i + b_i \text{LargeEns}_{ij} + \epsilon_{ij}$$

- Level Two:
$$a_i = \alpha_0 + \alpha_1 \text{Orch}_i + u_i$$
$$b_i = \beta_0 + \beta_1 \text{Orch}_i + v_i,$$

  where $\epsilon_{ij} \sim N(0, \sigma^2)$ and

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{bmatrix} \right).$$

We found that there are no highly significant fixed effects in Model C (other than the intercept). In particular, we have no significant evidence that musicians playing orchestral instruments reported different performance anxiety scores, on average, for solos and small ensembles than keyboardists and vocalists, no evidence of a difference in performance anxiety by performance type for keyboard players and vocalists, and no evidence of an instrument effect in difference between large ensembles and other types.

Since no terms were added at Level One when expanding from the random slopes and intercepts model (Model B), no discernible changes should occur in explained within-person variability (although small changes could occur due to numerical estimation procedures used in likelihood-based parameter estimates). However, Model C expanded Model B by using the instrument which a subject plays to model both intercepts and slopes at Level Two. We can use pseudo R-squared values for both intercepts and slopes to evaluate the impact on between-person variability of adding instrument to Model B.

$$\text{Pseudo } R^2_{L2_u} = \frac{\hat{\sigma}^2_u(\text{Model B}) - \hat{\sigma}^2_u(\text{Model C})}{\hat{\sigma}^2_u(\text{Model B})} = \frac{6.33 - 5.66}{6.33} = 0.106$$

$$\text{Pseudo } R^2_{L2_v} = \frac{\hat{\sigma}^2_v(\text{Model B}) - \hat{\sigma}^2_v(\text{Model C})}{\hat{\sigma}^2_v(\text{Model B})} = \frac{0.74 - 0.45}{0.74} = 0.392$$

Pseudo $R^2_{L2_u}$ describes the improvement in Model C over Model B in explaining subject-to-subject variability in intercepts, and Pseudo $R^2_{L2_v}$ describes the improvement in Model C over Model B in explaining subject-to-subject variability in slopes. Thus, the addition of instrument at Level Two has decreased the between-person variability in mean performance anxiety before solos and small ensembles by 10.6%, and it has decreased the between-person variability in the effect of large ensembles on performance anxiety by 39.2%.

We could also run a "random intercepts" version of Model C, with no error term in the equation for the slope at Level Two (and thus no covariance between errors at Level Two as well):

- Level One:
$$Y_{ij} = a_i + b_i \text{LargeEns}_{ij} + \epsilon_{ij}$$

- Level Two:
$$a_i = \alpha_0 + \alpha_1 \text{Orch}_i + u_i$$
$$b_i = \beta_0 + \beta_1 \text{Orch}_i,$$

  where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $u_i \sim N(0, \sigma^2_u)$.

The output below contains REML estimates of our 6 parameters from this simplified version of Model C (which we'll call Model C2):

```
#Model C2 (Run as random intercepts model)
model.c2 <- lmer(na ~ orch + large + orch:large +
  (1|id), data = music)
```

```
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept)  5.13    2.27
##  Residual             21.88    4.68
```

```
##  Number of Level Two groups =  37
```

```
##             Estimate Std. Error t value
## (Intercept)  15.9026     0.6187  25.703
## orch          1.7100     0.9131   1.873
## large        -0.8918     0.8415  -1.060
## orch:large   -1.4650     1.0880  -1.347
```

```
         df  AIC
model.c   8 3003
model.c2  6 2999

         df  BIC
model.c   8 3037
model.c2  6 3025
```

Note that parameter estimates for the remaining 6 fixed effects and variance components closely mirror the corresponding parameter estimates from Model C. In fact, removing the error term on the slope has improved (reduced) both the AIC and BIC measures of overall model performance. Instead of assuming that the large ensemble effects, after accounting for instrument played, vary by individual, we are assuming that large ensemble effect is fixed across subjects. It is not unusual to run a two-level model like this, with an error term on the intercept equation to account for subject-to-subject differences, but with no error terms on other Level Two equations unless there is an *a priori* reason to allow effects to vary by subject or if the model performs better after building in those additional error terms.

8.8 Adding Further Covariates

Recall that we are particularly interested in this study in Level Two covariates—those subject-specific variables that provide insight into why individuals react differently in anxiety-inducing situations. In Section 8.3, we saw evidence that subjects with higher baseline levels of negative emotionality tend to have higher performance anxiety levels prior to performances. Thus, in our next step in model building, we will add negative emotionality as a Level Two predictor to Model C. With this addition, our new model can be expressed as a system of Level One and Level Two models:

- Level One:

$$Y_{ij} = a_i + b_i \text{LargeEns}_{ij} + \epsilon_{ij}$$

- Level Two:

$$a_i = \alpha_0 + \alpha_1 \text{Orch}_i + \alpha_2 \text{MPQnem}_i + u_i$$
$$b_i = \beta_0 + \beta_1 \text{Orch}_i + \beta_2 \text{MPQnem}_i + v_i,$$

or as a composite model:

$$
\begin{aligned}
Y_{ij} = &[\alpha_0 + \alpha_1 \text{Orch}_i + \alpha_2 \text{MPQnem}_i + \beta_0 \text{LargeEns}_{ij} \\
&+ \beta_1 \text{Orch}_i \text{LargeEns}_{ij} + \beta_2 \text{MPQnem}_i \text{LargeEns}_{ij}] \\
&+ [u_i + v_i \text{LargeEns}_{ij} + \epsilon_{ij}]
\end{aligned}
$$

where error terms are defined as in Model C.

From the R output below, we see that, as our exploratory analyses suggested, subjects with higher baseline levels of stress reaction, alienation, and aggression (as measured by the MPQ negative emotionality scale) had significantly higher levels of performance anxiety before solos and small ensembles (t=3.893). They also had somewhat greater differences between large ensembles and other performance types, controlling for instrument (t=-0.575), although this interaction was not statistically significant.

```
# Model D (Add negative emotionality as second L2 covariate)
model.d <- lmer(na ~ orch + mpqnem + large + orch:large +
  mpqnem:large + (large | id), data = music)
```

```
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept)  3.286   1.813
##           large        0.557   0.746    -0.38
##  Residual             21.811   4.670
```

```
##  Number of Level Two groups =  37
```

```
##              Estimate Std. Error t value
## (Intercept)  11.56801    1.22057  9.4775
## orch          1.00069    0.81713  1.2246
## mpqnem        0.14823    0.03808  3.8925
## large        -0.28019    1.83412 -0.1528
## orch:large   -0.94927    1.10620 -0.8581
## mpqnem:large -0.03018    0.05246 -0.5753
```

8.8.1 Interpretation of Parameter Estimates

Compared to Model C, the directions of the effects of instrument and performance type are consistent, but the effect sizes and levels of significance are reduced because of the relative importance of the negative emotionality term. Interpretations will also change slightly to acknowledge that we have controlled for a covariate. In addition, interpretations of fixed effects involving negative emotionality must acknowledge that this covariate is a continuous measure and not binary like instrument and performance type:

- $\hat{\alpha}_0 = 11.57$. The estimated mean performance anxiety for solos and small ensembles (`large=0`) is 11.57 for keyboard players and vocalists (`orch=0`) with negative emotionality of 0 at baseline (`mpqnem=0`). Since the minimum negative emotionality score in this study was 11, this interpretation, while technically correct, is not practically meaningful.
- $\hat{\alpha}_1 = 1.00$. Orchestral instrument players have an estimated mean anxiety level before solos and small ensembles which is 1.00 point higher than keyboardists and vocalists, controlling for the effects of baseline negative emotionality.
- $\hat{\alpha}_2 = 0.15$. A one point increase in baseline negative emotionality is associated with an estimated 0.15 mean increase in anxiety levels before solos and small ensembles, after controlling for instrument.
- $\hat{\beta}_0 = -0.28$. Keyboard players and vocalists (`orch=0`) with baseline negative emotionality levels of 0 (`mpqnem=0`) have an estimated mean decrease in anxiety level of 0.28 points before large ensemble performances compared to other performance types.
- $\hat{\beta}_1 = -0.95$. After accounting for baseline negative emotionality, orchestral instrument players have an estimated mean anxiety level before solos and small ensembles which is 1.00 point higher than keyboardists and vocalists, while the mean anxiety of orchestral players is only .05 points higher before large ensembles (a difference of .95 points).
- $\hat{\beta}_2 = -0.03$. After accounting for instrument, a one-point increase in baseline negative emotionality is associated with an estimated 0.15 mean increase in anxiety levels before solos and small ensembles, but only an estimated 0.12 increase before large ensembles (a difference of .03 points).

Some of the detail in these parameter interpretations can be tricky—describing interaction terms, deciding if a covariate must be fixed at 0 or merely held constant, etc. Often it helps to write out models for special cases to isolate the effects of specific fixed effects. We will consider a few parameter estimates from above and see why the interpretations are written as they are.

- $\hat{\alpha}_1$. For solos and small ensembles (`LargeEns=0`), the following equations describe the fixed effects portion of the composite model for negative affect score for vocalists and keyboardists (`Orch=0`) and orchestral instrumentalists (`Orch=1`):

$$\begin{aligned} &\text{Orch} = 0: \\ &\qquad Y_{ij} = \alpha_0 + \alpha_2 \text{MPQnem}_i \\ &\text{Orch} = 1: \\ &\qquad Y_{ij} = (\alpha_0 + \alpha_1) + \alpha_2 \text{MPQnem}_i \end{aligned}$$

Regardless of the subjects' baseline negative emotionality (`MPQnem`), $\hat{\alpha}_1$ represents the estimated difference in performance anxiety between those playing orchestral instruments and others. This interpretation, however, only holds for

solos and small ensembles. For large ensembles, the difference between those playing orchestral instruments and others is actually given by $\hat{\alpha}_1 + \hat{\beta}_1$, holding `MPQnem` constant (Show!).

- $\hat{\beta}_0$. Because `LargeEns` interacts with both `Orch` and `MPQnem` in Model C, $\hat{\beta}_0$ only describes the estimated difference between large ensembles and other performance types when both `Orch=0` and `MPQnem=0`, thus removing the effects of the interaction terms. If, for instance, `Orch=1` and `MPQnem=20`, then the difference between large ensembles and other performance types is given by $\hat{\beta}_0 + \hat{\beta}_1 + 20\hat{\beta}_2$.
- $\hat{\beta}_1$. As with $\hat{\alpha}_1$, we consider equations describing the fixed effects portion of the composite model for negative affect score for vocalists and keyboardists (`Orch=0`) and orchestral instrumentalists (`Orch=1`), except here we leave *LargeEns* as an unknown rather than restricting the model to solos and small ensembles:

$$\begin{aligned}
&\text{Orch} = 0: \\
&\qquad Y_{ij} = \alpha_0 + \alpha_2 \text{MPQnem}_i + \beta_0 \text{LargeEns}_{ij} \\
&\qquad\quad + \beta_2 \text{MPQnem}_i \text{LargeEns}_{ij} \\
&\text{Orch} = 1: \\
&\qquad Y_{ij} = (\alpha_0 + \alpha_1) + \alpha_2 \text{MPQnem}_i + (\beta_0 + \beta_1) \text{LargeEns}_{ij} \\
&\qquad\quad + \beta_2 \text{MPQnem}_i \text{LargeEns}_{ij}
\end{aligned}$$

As long as baseline negative emotionality is held constant (at any level, not just 0), then $\hat{\beta}_1$ represents the estimated difference in the large ensemble effect between those playing orchestral instruments and others.

8.8.2 Model Comparisons

At this point, we might ask: do the two extra fixed effects terms in Model D provide a significant improvement over Model C? Nested models such as these can be tested using a **likelihood ratio test** (drop in deviance test), as we've used in Sections 4.4.4 and 6.5.4 with certain generalized linear models. Since we are comparing models nested in their fixed effects, we use full maximum likelihood methods to estimate model parameters, as discussed in Section 8.5.4. As expected, the likelihood is larger (and the log-likelihood is less negative) under the larger model (Model D); our test statistic (14.734) is then -2 times the difference in log-likelihood between Models C and D. Comparing the test statistic to a chi-square distribution with 2 degrees of freedom (signifying the number of additional terms in Model D), we obtain a p-value of .0006. Thus, Model D significantly outperforms Model C.

```
# anova() automatically uses ML for LRT tests
drop_in_dev <- anova(model.d, model.c, test = "Chisq")
```

```
        npar  AIC  BIC logLik  dev Chisq Df      pval
model.c    8 3007 3041  -1496 2991    NA NA        NA
model.d   10 2996 3039  -1488 2976 14.73  2 0.0006319
```

Two models, whether they are nested or not, can be compared using AIC and BIC measures, which were first seen in Chapter 1 and later used in evaluating generalized linear models. In this case, the AIC clearly favors Model D (2996.7) over Model C (3007.3), whereas the BIC favors Model D (3038.8) only slightly over Model C (3041.0) since the BIC imposes a stiffer penalty on additional terms and additional model complexity. However, the likelihood ratio test is a more reliable method for comparing nested models.

Finally, we note that Model D could be further improved by dropping the negative emotionality by large ensemble interaction term. Not only is the t-value (-0.575) associated with this term of low magnitude, but a likelihood ratio test comparing Model D to a model without `mpqnem:large` produces an insignificant p-value of 0.5534.

```
drop_in_dev <- anova(model.d, model.d1, test = "Chisq")
```

```
         npar  AIC  BIC logLik  dev  Chisq Df   pval
model.d1    9 2995 3033  -1488 2977     NA NA     NA
model.d    10 2996 3039  -1488 2976 0.3513  1 0.5534
```

8.9 Centering Covariates

As we observed above, the addition of baseline negative emotionality in Model D did not always produce sensible interpretations of fixed effects. It makes no sense to draw conclusions about performance anxiety levels for subjects with MPQNEM scores of 0 at baseline (as in $\hat{\beta}_0$), since the minimum NEM composite score among subjects in this study was 11. In order to produce more meaningful interpretations of parameter estimates and often more stable parameter estimates, it is often wise to **center** explanatory variables. Centering involves subtracting a fixed value from each observation, where the fixed value represents a meaningful anchor value (e.g., last grade completed is 12; GPA

is 3.0). Often, when there's no pre-defined anchor value, the mean is used to represent a typical case. With this in mind, we can create a new variable

$$\begin{aligned} \text{centeredbaselineNEM} &= \text{cmpqnem} \\ &= \text{mpqnem - mean(mpqnem)} \\ &= \text{mpqnem} - 31.63 \end{aligned}$$

and replace baseline NEM in Model D with its centered version to create Model E:

```
# Model E (Center baseline NEM in Model D)
model.e <- lmer(na ~ orch + cmpqnem + large + orch:large +
  cmpqnem:large + (large | id), REML = T, data = music)
```

```
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept)  3.286   1.813
##           large        0.557   0.746    -0.38
##  Residual             21.811   4.670

##  Number of Level Two groups =  37

##                Estimate Std. Error t value
## (Intercept)   16.25679    0.54756 29.6893
## orch           1.00069    0.81713  1.2246
## cmpqnem        0.14823    0.03808  3.8925
## large         -1.23484    0.84320 -1.4645
## orch:large    -0.94927    1.10620 -0.8581
## cmpqnem:large -0.03018    0.05246 -0.5753
```

As you compare Model D to Model E, you should notice that only two things change – $\hat{\alpha}_0$ and $\hat{\beta}_0$. All other parameter estimates—both fixed effects and variance components—remain identical; the basic model is essentially unchanged as well as the amount of variability in anxiety levels explained by the model. $\hat{\alpha}_0$ and $\hat{\beta}_0$ are the only two parameter estimates whose interpretations in Model D refer to a specific level of baseline NEM. In fact, the interpretations that held true where `NEM=0` (which isn't possible) now hold true for `cmpqnem=0` or when NEM is at its average value of 31.63, which is possible and quite meaningful. Now, parameter estimates using centered baseline NEM in Model E change in value from Model D and produce more useful interpretations:

- $\hat{\alpha}_0 = 16.26$. The estimated mean performance anxiety for solos and small ensembles (`large=0`) is 16.26 for keyboard players and vocalists (`orch=0`) with an average level of negative emotionality at baseline (`mpqnem=31.63`).

- $\hat{\beta}_0 = -1.23$. Keyboard players and vocalists (`orch=0`) with an average level of baseline negative emotionality levels (`mpqnem=31.63`) have an estimated mean decrease in anxiety level of 1.23 points before large ensemble performances compared to other performance types.

8.10 A Final Model for Music Performance Anxiety

We now begin iterating toward a "final model" for these data, on which we will base conclusions. Typical features of a "final multilevel model" include:

- fixed effects allow one to address primary research questions
- fixed effects control for important covariates at all levels
- potential interactions have been investigated
- variables are centered where interpretations can be enhanced
- important variance components have been included
- unnecessary terms have been removed
- the model tells a "persuasive story parsimoniously"

Although the process of reporting and writing up research results often demands the selection of a sensible final model, it's important to realize that (a) statisticians typically will examine and consider an entire taxonomy of models when formulating conclusions, and (b) different statisticians sometimes select different models as their "final model" for the same set of data. Choice of a "final model" depends on many factors, such as primary research questions, purpose of modeling, tradeoff between parsimony and quality of fitted model, underlying assumptions, etc. So you should be able to defend any final model you select, but you should not feel pressured to find the one and only "correct model", although most good models will lead to similar conclusions.

As we've done in previous sections, we can use (a) t-statistics for individual fixed effects when considering adding a single term to an existing model, (b) likelihood ratio tests for comparing nested models which differ by more than one parameter, and (c) model performance measures such as AIC and BIC to compare non-nested models. Below we offer one possible final model for this data—Model F:

- Level One:

$$Y_{ij} = a_i + b_i\text{previous}_{ij} + c_i\text{students}_{ij} + d_i\text{juried}_{ij} + e_i\text{public}_{ij} + f_i\text{solo}_{ij} + \epsilon_{ij}$$

- Level Two:

$$
\begin{aligned}
a_i &= \alpha_0 + \alpha_1 \text{mpqpem}_i + \alpha_2 \text{mpqab}_i + \alpha_3 \text{orch}_i + \alpha_4 \text{mpqnem}_i + u_i \\
b_i &= \beta_0 + v_i, \\
c_i &= \gamma_0 + w_i, \\
d_i &= \delta_0 + x_i, \\
e_i &= \varepsilon_0 + y_i, \\
f_i &= \zeta_0 + \zeta_1 \text{mpqnem}_i + z_i,
\end{aligned}
$$

where `previous` is the number of previous diary entries filled out by that individual (`diary-1`); `students`, `juried`, and `public` are indicator variables created from the `audience` categorical variable (so that "Instructor" is the reference level in this model); and, `solo` is 1 if the performance was a solo and 0 if the performance was either a small or large ensemble.

In addition, we assume the following variance-covariance structure at Level Two:

$$
\begin{bmatrix} u_i \\ v_i \\ w_i \\ x_i \\ y_i \\ z_i \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & & & & & \\ \sigma_{uv} & \sigma_v^2 & & & & \\ \sigma_{uw} & \sigma_{vw} & \sigma_w^2 & & & \\ \sigma_{ux} & \sigma_{vx} & \sigma_{wx} & \sigma_x^2 & & \\ \sigma_{uy} & \sigma_{vy} & \sigma_{wy} & \sigma_{xy} & \sigma_y^2 & \\ \sigma_{uz} & \sigma_{vz} & \sigma_{wz} & \sigma_{xz} & \sigma_{yz} & \sigma_z^2 \end{bmatrix} \right).
$$

Being able to write out these mammoth variance-covariance matrices is less important than recognizing the number of variance components that must be estimated by our intended model. In this case, we must use likelihood-based methods to obtain estimates for 6 variance terms and 15 correlation terms at Level Two, along with 1 variance term at Level One. Note that the number of correlation terms is equal to the number of unique pairs among Level Two random effects. In later sections we will consider ways to reduce the number of variance components in cases where the number of terms is exploding, or the statistical software is struggling to simultaneously find estimates for all model parameters to maximize the likelihood function.

From the signs of fixed effects estimates in the R output below, we see that performance anxiety is higher when a musician is performing in front of students, a jury, or the general public rather than their instructor, and it is lower for each additional diary the musician previously filled out. In addition, musicians with lower levels of positive emotionality and higher levels of absorption tend to experience greater performance anxiety, and those who play orchestral instruments experience more performance anxiety than those who play keyboards or sing. Addressing the researchers' primary hypothesis, after controlling for all these factors, we have significant evidence that musicians with higher levels of negative emotionality experience higher levels of performance anxiety, and

that this association is even more pronounced when musicians are performing solos rather than as part of an ensemble group.

Here are how a couple of key fixed effects would be interpreted in this final model:

- $\hat{\alpha}_4 = 0.11$. A one-point increase in baseline level of negative emotionality is associated with an estimated 0.11 mean increase in performance anxiety for musicians performing in an ensemble group (`solo=0`), after controlling for previous diary entries, audience, positive emotionality, absorption, and instrument.
- $\hat{\zeta}_1 = 0.08$. When musicians play solos, a one-point increase in baseline level of negative emotionality is associated with an estimated 0.19 mean increase in performance anxiety, 0.08 points (73%) higher than musicians playing in ensemble groups, controlling for the effects of previous diary entries, audience, positive emotionality, absorption, and instrument.

```
# Model F (One - of many - reasonable final models)
model.f <- lmer(na ~ previous + students + juried +
    public + solo + mpqpem + mpqab + orch + mpqnem +
    mpqnem:solo + (previous + students + juried +
    public + solo | id), REML = T, data = music)
```

```
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 14.4802  3.805
##           previous     0.0707  0.266    -0.65
##           students     8.2151  2.866    -0.63  0.00
##           juried      18.3177  4.280    -0.64 -0.12
##           public      12.8094  3.579    -0.83  0.33
##           solo         0.7665  0.876    -0.67  0.47
##  Residual             15.2844  3.910
##
##
##
##
##   0.84
##   0.66  0.58
##   0.49  0.21  0.90
##
##  Number of Level Two groups =  37
##             Estimate Std. Error t value
## (Intercept)  8.36883    1.91369  4.3731
```

```
## previous     -0.14303    0.06247 -2.2895
## students      3.61115    0.76796  4.7022
## juried        4.07332    1.03130  3.9497
## public        3.06453    0.89274  3.4327
## solo          0.51647    1.39635  0.3699
## mpqpem       -0.08312    0.02408 -3.4524
## mpqab         0.20377    0.04740  4.2986
## orch          1.53138    0.58384  2.6230
## mpqnem        0.11465    0.03591  3.1930
## solo:mpqnem   0.08296    0.04158  1.9951
```

8.11 Modeling Multilevel Structure: Is It Necessary?

Before going too much further, we should really consider if this multilevel structure has gained us anything over linear least squares regression. Sure, multilevel modeling seems more faithful to the inherent structure of the data—performances from the same musician should be more highly correlated than performances from different musicians—but do our conclusions change in a practical sense? Some authors have expressed doubts. For instance Robert Bickel, in his 2007 book, states, "When comparing OLS and multilevel regression results, we may find that differences among coefficient values are inconsequential, and tests of significance may lead to the same decisions. A great deal of effort seems to have yielded precious little gain" [Bickel, 2007]. Others, especially economists, advocate simply accounting for the effect of different Level Two observational units (like `musicians`) with a sequence of indicator variables for those observational units. We contend that (1) fitting multilevel models is a small extension of LLSR regression that is not that difficult to conceptualize and fit, especially with the software available today, and (2) using multilevel models to remain faithful to the data structure *can* lead to different coefficient estimates and *often* leads to different (and larger) standard error estimates and thus smaller test statistics. Hopefully you've seen evidence of (1) in this chapter already; the rest of this section introduces two small examples to help illustrate (2).

Figure 8.15 is based on a synthetic data set containing 10 observations from each of 4 subjects. For each subject, the relationship between previous performances and negative affect is linear and negative, with slope approximately -0.5 but different intercepts. The multilevel model (a random intercepts model as described in Section 8.7) shows an overall relationship (the dashed black line) that's consistent with the individual subjects—slope around -0.5 with an intercept that appears to average the 4 subjects. Fitting a LLSR model,

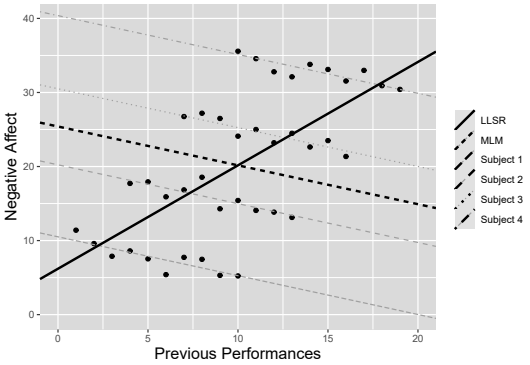


FIGURE 8.15: Hypothetical data from 4 subjects relating number of previous performances to negative affect. The dashed black line depicts the overall relationship between previous performances and negative affect as determined by a multilevel model, while the solid black line depicts the overall relationship as determined by an LLSR regression model.

however, produces an overall relationship (the solid black line) that is strongly positive. In this case, by naively fitting the 40 observations as if they were all independent and ignoring subject effects, the LLSR analysis has gotten the estimated slope of the overall relationship backwards, producing a continuous data version of Simpson's Paradox.

Our second example is based upon Model C from Section 8.7.3, with single binary predictors at both Level One and Level Two. Using the estimated fixed effects coefficients and variance components from random effects produced in Model C, we generated 1000 sets of simulated data. Each set of simulated data contained 497 observations from 37 subjects just like the original data, with relationships between negative affect and large ensembles and orchestral instruments (along with associated variability) governed by the estimated parameters from Model C. Each set of simulated data was used to fit both a multilevel model and a linear least squares regression model, and the estimated fixed effects ($\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$, and $\hat{\beta}_1$) and their standard errors were saved. Figure 8.16 shows density plots comparing the 1000 estimated values for each fixed effect from the two modeling approaches; in general, estimates from multilevel modeling and LLSR tend to agree pretty well, without noticeable bias. Based on coefficient estimates alone, there appears to be no reason to favor multilevel modeling over LLSR in this example, but Figure 8.17 tells a different story. Figure 8.17 shows density plots comparing the 1000 estimated standard errors associated with each fixed effect from the two modeling approaches; in general, standard errors are markedly larger with multilevel modeling than LLSR. This is not unusual, since LLSR assumes all 497 observations are independent, while

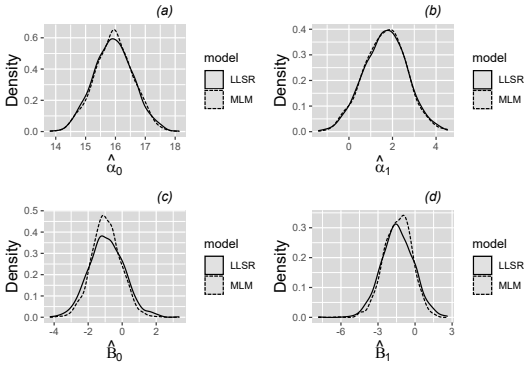


FIGURE 8.16: Density plots of parameter estimates for the four fixed effects of Model C under both a multilevel model and linear least squares regression. 1000 sets of simulated data for the 37 subjects in our study were produced using estimated fixed and random effects from Model C. For each set of simulated data, estimates of (a) α_0, (b) α_1, (c) β_0, and (d) β_1 were obtained using both a multilevel and an LLSR model. Each plot then shows a density plot for the 1000 estimates of the corresponding fixed effect using multilevel modeling vs. a similar density plot for the 1000 estimates using LLSR.

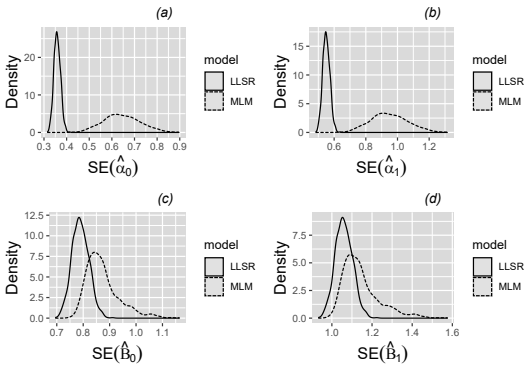


FIGURE 8.17: Density plots of standard errors of parameter estimates for the four fixed effects of Model C under both a multilevel model and linear least squares regression. 1000 sets of simulated data for the 37 subjects in our study were produced using estimated fixed and random effects from Model C. For each set of simulated data, estimates of (a) SE(α_0), (b) SE(α_1), (c) SE(β_0), and (d) SE(β_1) were obtained using both a multilevel and an LLSR model. Each plot then shows a density plot for the 1000 estimates of the corresponding standard error term using multilevel modeling vs. a similar density plot for the 1000 estimates using LLSR.

multilevel modeling acknowledges that, with correlated data within subject, there are fewer than 497 independent pieces of data. Therefore, linear least squares regression can overstate precision, producing t-statistics for each fixed effect that tend to be larger than they should be; the number of significant results in LLSR are then too great and not reflective of the true structure of the data.

8.12 Notes on Using R (optional)

Initial examination of the data for Case Study 8.2 shows a couple of features that must be noted. First, there are 37 unique study participants, but they are not numbered successively from 1 to 43. The majority of participants filled out 15 diaries, but several filled out fewer (with a minimum of 2); as with participant IDs, diary numbers within participant are not always successively numbered. Finally, missing data is not an issue in this data set, since researchers had already removed participants with only 1 diary entry and performances for which the type was not recorded (of which there were 11).

The R code below runs the initial multilevel model in Section 8.5.5. Multilevel model notation in R is based on the composite model formulation. Here, the response variable is `na`, while `orch`, `large`, and `orch:large` represent the fixed effects α_1, β_0, and β_1, along with the intercept α_0 which is included automatically. Note that a colon is used to denote an interaction between two variables. Error terms and their associated variance components are specified in `(large|id)`, which is equivalent to `(1+large|id)`. This specifies two error terms at Level Two (the `id` level): one corresponding to the intercept (u_i) and one corresponding to the large ensemble effect (v_i); the multilevel model will then automatically include a variance for each error term in addition to the covariance between the two error terms. A variance associated with a Level One error term is also automatically included in the multilevel model. Note that there are ways to override the automatic inclusion of certain variance components; for example, `(0+large|id)` would not include an error term for the intercept (and therefore no covariance term at Level Two either).

```
model0 <- lmer(na ~ orch + large + orch:large +
  (large | id), REML = T, data = music)
summary(model0)
```

8.13 Exercises

8.13.1 Conceptual Exercises

1. **Housing prices.** Brown and Uyar [2004] describe "A Hierarchical Linear Model Approach for Assessing the Effects of House and Neighborhood Characteristics on Housing Prices". Based on the title of their paper: (a) give the observational units at Level One and Level Two, and (b) list potential explanatory variables at both Level One and Level Two.

2. In the preceding problem, why can't we assume all houses in the data set are independent? What would be the potential implications to our analysis of assuming independence among houses?

3. In the preceding problem, for each of the following sets of predictors: (a) write out the two-level model for predicting housing prices, (b) write out the corresponding composite model, and (c) determine how many model parameters (fixed effects and variance components) must be estimated.

   - Square footage, number of bedrooms
   - Median neighborhood income, rating of neighborhood schools
   - Square footage, number of bedrooms, age of house, median neighborhood housing price
   - Square footage, median neighborhood income, rating of neighborhood schools, median neighborhood housing price

4. **Music performance anxiety.** Describe a situation in which the two plots in Figure 8.7 might tell different stories.

5. Explain the difference between a_i in Equation (8.2) and $\hat{a}_i$ in Equation (8.4).

6. Why is the contour plot for multivariate normal density in Figure 8.12(b) tilted from southwest to northeast, but the contour plot in Figure 8.12(a) is not tilted?

7. In Table 8.3, note that the standard errors associated with estimated coefficients under independence are lower than standard errors under alternative analysis methods. Why is that often the case?

8. Why is Model A (Section 8.6.2) sometimes called the "unconditional means model"? Why is it also sometimes called the "random intercepts model"? Are these two labels consistent with each other?

9. Consider adding an indicator variable in Model B (Section 8.7.1) for Small Ensemble performances.
   - Write out the two-level model for performance anxiety,
   - Write out the corresponding composite model,
   - Determine how many model parameters (fixed effects and variance components) must be estimated, and
   - Explain how the interpretation for the coefficient in front of Large Ensembles would change.
10. Give a short rule in your own words describing when an interpretation of an estimated coefficient should "hold constant" another covariate or "set to 0" that covariate (see Section 8.8.1).
11. The interpretation of $\hat{\alpha}_1$ in Section 8.8.1 claims that, "This interpretation, however, only holds for solos and small ensembles. For large ensembles, the difference between those playing orchestral instruments and others is actually given by $\hat{\alpha}_1 + \hat{\beta}_1$, holding MPQNEM constant." Show that this claim is true.
12. Explain how the interpretations of the following parameter estimates change (or don't change) as we change our model:
   - $\hat{\alpha}_0$ from Model A to B to C to D to E
   - $\hat{\beta}_1$ from Model B to C to D to E
   - $\hat{\alpha}_1$ from Model C to D to E
   - $\hat{\beta}_1$ from Model C to D to E
   - $\hat{\sigma}_u$ from Model A to B to C to D to E
   - $\hat{\sigma}_v$ from Model B to C to D to E
13. When moving from Model B to Model C in Section 8.7.3, $\hat{\sigma}_u^2$ increases slightly. Why might this have occurred?
14. Interpret other estimated parameters from Model F beyond those interpreted in Section 8.10: $\hat{\alpha}_0$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, $\hat{\beta}_0$, $\hat{\gamma}_0$, $\hat{\zeta}_0$, $\hat{\rho}_{wx}$, $\hat{\sigma}^2$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_z^2$.
15. Explain Figure 8.15 in your own words. Why would LLSR produce a misleading analysis in this case, but multilevel models would not?
16. Summarize Figures 8.16 and 8.17 in your own words.

8.13.2 Guided Exercises

1. **Music performance joy.** In this chapter, we studied models for predicting music performance anxiety, as measured by the negative affect scale from the PANAS instrument. Now we will examine models for predicting the happiness of musicians prior to performances, as measured by the positive affect scale from the PANAS instrument.

To begin, run the following models:

- Model A = unconditional means model
- Model B = indicator for instructor audience type and indicator for student audience type at Level One; no Level Two predictors
- Model C = indicator for instructor audience type and indicator for student audience type at Level One; centered MPQ absorption subscale as Level Two predictor for intercept and all slope terms
- Model D = indicator for instructor audience type and indicator for student audience type at Level One; centered MPQ absorption subscale and a male indicator as Level Two predictors for intercept and all slope terms

1. Perform an exploratory data analysis by comparing positive affect (happiness) to Level One and Level Two covariates using appropriate graphs. Comment on interesting trends, supporting your comments with appropriate summary statistics.
2. Report estimated fixed effects and variance components from Model A, using proper notation from this chapter (no interpretations required). Also report and interpret an intraclass correlation coefficient.
3. Report estimated fixed effects and variance components from Model B, using proper notation from this chapter. Interpret your MLE estimates for $\hat{\alpha}_0$ (the intercept), $\hat{\beta}_1$ (the instructor indicator), and $\hat{\sigma}_u$ (the Level Two standard deviation for the intercept). Also report and interpret an appropriate pseudo R-squared value.
4. Write out Model C, using both separate Level One and Level Two models as well as a composite model. Be sure to express distributions for error terms. How many parameters must be estimated in Model C?
5. Report and interpret the following parameter estimates from Model C: $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\gamma}_0$, $\hat{\beta}_1$, $\hat{\sigma}_u$, $\hat{\sigma}_v$, and $\hat{\rho}_{uv}$. Interpretations for variance components should be done in terms of standard deviations and correlation coefficients.
6. Report and interpret the same parameter estimates listed above from Model D. In each case, the new interpretation should involve a small modification of your interpretation from Model C. Use underlines or highlights to denote the part of the Model D interpretation that differs from the Model C interpretation.
7. Also report and interpret the following parameter estimates from Model D: $\hat{\alpha}_2$ and $\hat{\beta}_2$.
8. Use a drop in deviance statistic (likelihood ratio test) to compare Model C vs. Model D. Give a test statistic and p-value, then state a conclusion. Also compare Models C and D with appropriate pseudo R-squared value(s) and with AIC and BIC statistics.

8.13.3 Open-Ended Exercises

1. **Political ambiguity.** Chapp et al. [2018] explored 2014 congressional candidates' ambiguity on political issues in their paper, *Going Vague: Ambiguity and Avoidance in Online Political Messaging.* They hand-coded a random sample of 2012 congressional candidates' websites, assigning an ambiguity score. A total of 870 websites from 2014 were then automatically scored using Wordscores, a program designed for political textual analysis. In their paper, they fit a multilevel model for candidates' ambiguities with predictors at both the candidate and district levels. Some of their hypotheses include that:

   - "when incumbents do hazard issue statements, these statements will be marked by a higher degree of clarity." (Hypothesis 1b)
   - "ideological distance [from district residents] will be associated with greater ambiguity." (Hypothesis 2a)
   - "controlling for ideological distance, ideological extremity [of the candidate] should correspond to less ambiguity." (Hypothesis 2b)
   - "more variance in attitudes [among district residents] will correspond to a higher degree of ambiguity in rhetoric" (Hypothesis 3a)
   - "a more heterogeneous mix of subgroups [among district residents] will also correspond to a higher degree of ambiguity in rhetoric" (Hypothesis 3b)

   Their data can be found in `ambiguity.csv`. Variables of interest include:

   - `ambiguity` = assigned ambiguity score. Higher scores indicate greater clarity (less ambiguity)
   - `democrat` = 1 if a Democrat, 0 otherwise (Republican)
   - `incumbent` = 1 if an incumbent, 0 otherwise
   - `ideology` = a measure of the candidate's left-right orientation. Higher (positive) scores indicate more conservative candidates and lower (negative) scores indicate more liberal candidates.
   - `mismatch` = the distance between the candidate's ideology and the district's ideology (candidate ideology scores were regressed against district ideology scores; mismatch values represent the absolute value of the residual associated with each candidate)
   - `distID` = the congressional district's unique ID
   - `distLean` = the district's political leaning. Higher scores imply more conservative districts.
   - `attHeterogeneity` = a measure of the variability of ideologies within the district. Higher scores imply more attitudinal heterogeneity among voters.

- `demHeterogeneity` = a measure of the demographic variability within the district. Higher scores imply more demographic heterogeneity among voters.

With this in mind, fit your own models to address these hypotheses from Chapp et al. [2018]. Be sure to use a two-level structure to account for variables at both the candidate and district levels.

2. **Airbnb in Chicago.** Trinh and Ameri [2018] collected data on 1561 Airbnb listings in Chicago from August 2016, and then they merged in information from the neighborhood (out of 43 in Chicago) where the listing was located. We can examine traits that are associated with listings that command a higher price. Conduct an EDA, build a multilevel model, and interpret model coefficients to answer questions such as: What are characteristics of a higher priced listing? Are the most influential traits associated with individual listings or entire neighborhoods? Are there intriguing interactions where the effect of one variable depends on levels of another?

   The following variables can be found in `airbnb.csv` or derived from the variables found there:

   - `overall_satisfaction` = rating on a 0-5 scale.
   - `satisfaction` = 1 if `overall_satisfaction` is 5, 0 otherwise
   - `price` = price for one night (in dollars)
   - `reviews` = number of reviews posted
   - `room_type` = Entire home/apt, Private room, or Shared room
   - `accommodates` = number of people the unit can hold
   - `bedrooms` = number of bedrooms
   - `minstay` = minimum length of stay (in days)
   - `neighborhood` = neighborhood where unit is located (1 of 43)
   - `district` = district where unit is located (1 of 9)
   - `WalkScore` = quality of the neighborhood for walking (0-100)
   - `TransitScore` = quality of the neighborhood for public transit (0-100)
   - `BikeScore` = quality of the neighborhood for biking (0-100)
   - `PctBlack` = proportion of black residents in a neighborhood
   - `HighBlack` = 1 if `PctBlack` above .60, 0 otherwise

3. **Project 5183.** The Colorado Rockies, a Major League Baseball team, instigated a radical experiment on June 20th, 2012. Hopelessly out of contention for the playoffs and struggling again with their pitching, the Rockies decided to limit their starting pitchers to 75 pitches from June 20th until the end of the year with the hope of improving a struggling starting rotation, teaching pitchers how to pitch to contact (which results in low pitch counts), and at the same time trying to conserve young arms. Data has shown that, as a game

progresses, fatigue becomes a big factor in a pitcher's performance; if a pitcher has to tweak his mechanics to try to make up for a fatigued body, injuries can often occur. In addition, pitchers often struggle as they begin facing the same batters over again later in games. The Rockies called their experiment "Project 5183" to acknowledge the altitude at Coors Field, their home ballpark, and the havoc that high altitude can wreak on pitchers.

A team of students collected 2012 data on Rockies pitchers from FanGraphs to evaluate Project 5183 [Sturtz et al., 2013]. In a successful experiment, Colorado pitchers on a strict limit of 75 pitches would throw more strikes and yet record fewer strikeouts (pitching to contact rather than throwing more pitches to attempt to strike batters out). Different theories explain whether these pitchers would throw harder (since they don't have to save themselves) or throw slower (in order to throw more strikes). But the end result the Rockies hoped to observe was that their pitchers pitch better (allow fewer runs to the opponent) with a pitch limit.

The data set `FinalRockiesdata.csv` contains information for 7 starting pitchers who started at least one game before June 20th (without a pitch limit) and at least one game after June 20th (with a limit of 75 pitches). Key response variables include:

- `vFA` = average fastball velocity
- `K.9` = strikeouts per nine innings
- `ERA` = earned runs per nine innings
- `Pitpct` = percentage of strikes thrown

The primary explanatory variable of interest is `PCL` (an indicator variable for if a pitch count limit is in effect). Other potential confounding variables that may be important to control for include `Coors` (whether or not the game was played in Coors Field, where more runs are often scored because of the high altitude and thin air) and `Age` of the pitcher.

Write a short report summarizing the results of Project 5183. (You may notice a few variance components with unusual estimates, such as an estimated variance of 0 or an estimated correlation of 1. These estimates have encountered boundary constraints; we will learn how to deal with these situations in Section 10.5. For now ignore these variance components; the fixed effects coefficients are still reliable and their interpretations valid.)

4. **Replicate the Sadler and Miller paper.** Try to replicate Models 1 and 2 presented in Table 2 of Sadler and Miller [2010]. We expect small differences in parameter estimates, since they use SAS (with an unstructured covariance structure) instead of R.

- Do the parameter estimates, SEs, AIC, and variance explained compare well?
- Explain ways in which the model equations from page 284 of Sadler and Miller do not align with Model 2 from Table 2, or with the manner in which we write out two-level models.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com