Broadening Your Statistical Horizons:
Generalized Linear Models and Multilevel Models

J. Legler and P. Roback
6 CONTENTS

4.9.4.1 Skin Cancer: an example. 90
4.9.4.2 Campus Crime 91
4.9.4.3 Dating Online 91
4.9.4.4 Bus Driver Complaints 92
4.9.5 A 3D Graphic of Poisson Regression 92
4.9.6 Guided Exercises 92
4.9.6.1 Overdispersed Poisson 93
4.9.7 Open Ended 95
4.9.8 Other exercises 96

5 Generalized Linear Models (GLMs): A Unifying Theory 97
5.1 One parameter exponential families 97
5.1.1 One Parameter Exponential Family: Possion 97
5.1.2 One parameter exponential family: Normal 98
5.2 Generalized Linear Modeling 99
5.3 Exercises 100
5.4 References cited 101

6 Logistic Regression 103
6.1 Learning Objectives 103
6.2 Introduction 103
6.2.1 Binary Responses 103
6.2.2 Binomial Responses: sums of binary responses 104
6.2.3 Bernoulli Process 104
6.2.4 An Example: Binge Drinking 104
6.2.5 A Graphical Look at Binomial Regression 105
6.3 A Modeling Framework 105
6.4 Case Studies Overview 106
6.5 GLM Theory for Binomial Outcomes 107
6.5.1 Case Study: Soccer Goalie Saves 107
6.5.2 Research Question 107
6.5.3 Data Organization 107
6.6 Case Study: Reconstructing Alabama 109
6.6.1 Research Question 110
6.6.2 Data Collection 110
6.6.3 Data Organization 110
6.6.4 Exploratory Data Analysis 110
6.6.5 Modeling Approach 111
6.6.6 Results 112
6.6.6.1 Models 112
6.6.6.2 Residuals for Binomial Regression 113
6.6.6.3 Overdispersion 114
6.6.7 Least Squares Regression vs. Logistic Regression 116
6.7 Case Study: Who wants to lose weight? Sex, Media, Sports, and BMI. 116
6.7.1 Background 116
6.7.2 Research Questions 117
6.7.3 Data Collection 117
6.7.4 Data Organization 117
6.7.5 Exploratory Data Analysis 118
6.7.6 Modeling 120
6.7.6.1 Model 1: Sex 120
6.7.6.2 Model 2: sex + bmipct 121
8 CONTENTS

8.5.4 Technical issues when estimating and testing parameters (Optional) 159
8.5.5 An initial model with parameter interpretations 161
8.6 Building a multilevel model 163
8.6.1 Model building strategy 163
8.6.2 An initial model: unconditional means or random intercepts 163
8.7 Binary covariates at Level One and Level Two 164
8.7.1 Random slopes and intercepts model 164
8.7.2 Pseudo $R^2$ values 167
8.7.3 Adding a covariate at Level Two 168
8.8 Additional covariates: model comparison and interpretability 169
8.8.1 Interpretation of parameter estimates 170
8.8.2 Model comparisons 172
8.9 Center covariates 172
8.10 A potential final model for music performance anxiety 173
8.11 Modeling the multilevel structure: is it really necessary? 176
8.12 Notes on Using R (Optional) 178
8.13 Exercises 180
8.13.1 Conceptual Exercises 180
8.13.2 Guided Exercise 181
8.13.3 Open-ended Exercises 182

9 Two Level Longitudinal Data 185
9.1 Learning objectives 185
9.2 Case study: Charter schools 185
9.3 Initial Exploratory Analyses 186
9.3.1 Data organization 186
9.3.2 Missing data 186
9.3.3 Exploratory analyses for general multilevel models 188
9.3.4 Exploratory analyses for longitudinal data 189
9.4 Preliminary two-stage modeling 193
9.4.1 Linear trends within schools 193
9.4.2 Effects of level two covariates on linear time trends 194
9.4.3 Error structure within schools 198
9.5 Initial models 199
9.5.1 Unconditional means model 199
9.5.2 Unconditional growth model 200
9.5.3 Modeling other trends over time 202
9.6 Building to a final model 204
9.6.1 Uncontrolled effects of school type 204
9.6.2 Add percent free and reduced lunch as a covariate 206
9.6.3 A potential final model with three Level Two covariates 207
9.7 Covariance structure among observations 209
9.7.1 Standard covariance structure 210
9.7.2 Alternative covariance structures 212
9.7.3 Covariance structure in non-longitudinal multilevel models 213
9.7.4 Final thoughts regarding covariance structures 213
9.7.5 Details of covariance structures (Optional) 213
9.8 Notes on Using R (Optional) 214
9.9 Exercises 216
9.9.1 Conceptual Exercises 216
9.9.2 Guided Exercise 218
9.9.3 Open-ended Exercises 219
## CONTENTS

### 10 Multilevel Data With More Than Two Levels

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.1 Learning Objectives</td>
<td>221</td>
</tr>
<tr>
<td>10.2 Case Studies: Seed Germination</td>
<td>221</td>
</tr>
<tr>
<td>10.3 Initial Exploratory Analyses</td>
<td>222</td>
</tr>
<tr>
<td>10.3.1 Data Organization</td>
<td>222</td>
</tr>
<tr>
<td>10.3.2 Exploratory Analyses</td>
<td>224</td>
</tr>
<tr>
<td>10.4 Initial models: unconditional means and unconditional growth</td>
<td>230</td>
</tr>
<tr>
<td>10.5 Encountering boundary constraints</td>
<td>234</td>
</tr>
<tr>
<td>10.6 Parametric bootstrap testing</td>
<td>239</td>
</tr>
<tr>
<td>10.7 Exploding variance components</td>
<td>243</td>
</tr>
<tr>
<td>10.8 Building to a final model</td>
<td>245</td>
</tr>
<tr>
<td>10.9 Covariance structure (Optional)</td>
<td>249</td>
</tr>
<tr>
<td>10.9.1 Details of covariance structures</td>
<td>252</td>
</tr>
<tr>
<td>10.10 Notes on Using R (Optional)</td>
<td>253</td>
</tr>
<tr>
<td>10.11 Exercises</td>
<td>255</td>
</tr>
<tr>
<td>10.11.1 Conceptual Exercises</td>
<td>255</td>
</tr>
<tr>
<td>10.11.2 Guided Exercise</td>
<td>256</td>
</tr>
<tr>
<td>10.11.3 Open-ended Exercises</td>
<td>258</td>
</tr>
</tbody>
</table>

### 11 Generalized Linear Multilevel Models

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.1 Learning Objectives</td>
<td>261</td>
</tr>
<tr>
<td>11.2 Case Study: College Basketball Referees</td>
<td>261</td>
</tr>
<tr>
<td>11.3 Initial Exploratory Analyses</td>
<td>262</td>
</tr>
<tr>
<td>11.3.1 Data organization</td>
<td>262</td>
</tr>
<tr>
<td>11.3.2 Exploratory analyses</td>
<td>263</td>
</tr>
<tr>
<td>11.4 Two level Modeling with a Generalized Response</td>
<td>266</td>
</tr>
<tr>
<td>11.4.1 A multiple generalized linear model approach (correlation not accounted for)</td>
<td>266</td>
</tr>
<tr>
<td>11.4.2 A two-stage modeling approach (provides the basic idea for multilevel modeling)</td>
<td>267</td>
</tr>
<tr>
<td>11.4.3 A unified multilevel approach (the framework we’ll use)</td>
<td>269</td>
</tr>
<tr>
<td>11.5 Crossed Random Effects</td>
<td>271</td>
</tr>
<tr>
<td>11.6 Model Comparisons Using the Parametric Bootstrap</td>
<td>274</td>
</tr>
<tr>
<td>11.7 A Potential Final Model for Examining Referee Bias</td>
<td>277</td>
</tr>
<tr>
<td>11.8 Estimated Random Effects</td>
<td>280</td>
</tr>
<tr>
<td>11.9 Notes on Using R (Optional)</td>
<td>280</td>
</tr>
<tr>
<td>11.10 Exercises</td>
<td>282</td>
</tr>
<tr>
<td>11.10.1 Conceptual Exercises</td>
<td>282</td>
</tr>
<tr>
<td>11.10.2 Guided Exercise</td>
<td>285</td>
</tr>
<tr>
<td>11.10.3 Open-ended Exercises</td>
<td>285</td>
</tr>
</tbody>
</table>

**Index** 287
Notes for the Instructor

This is a preliminary version of *Broadening Your Statistical Horizons: Generalized Linear Models and Correlated Data (BYSH)*. It is intended to be accessible to undergraduate students who have been successful with a regression course, for example, a course that uses *Stat2*. We started teaching this course in 2003 in response so students would be able to deal with the non-normal, correlated world we live in. It has been offered at St. Olaf ever year since. In fact, it is required for all statistics concentrators. Initially, course materials came from a variety of sources resulting in some dissatisfaction on the part of the students (and the professors as well). After a couple of years of this and no success at finding a suitable text we started writing our own. This text extends statistical concepts to encompass realistic settings beyond iidNormal data. Unlike similar books on the market, this material is presented for an undergraduate audience.

We have selected methods to include that we feel have many applications in the real world. There is a logic to the order of the topics, but professors can choose which chapters and which order as they like. The original intent was to create a course which is an Assumptions Buster. The first chapter (which is under construction) lays out the assumptions most students have encountered when modeling previously. Having a solid understanding of these assumptions will enhance the students overall understanding of the book. It emphasizes the concepts associated with the assumptions. This start reflects an important principle for the course —— the focus is on understanding and interpreting assumptions, parameters, and models in general.

Likelihood is fundamental to fitting most of the models found in BYSH. This can be a challenging chapter for students, but we use a simple application where the construction of the likelihoods is straightforward. We do not attempt to construct likelihoods for complex situations, we stress the ideas of MLE using the simpler example. For example, the section showing that least squares estimators are MLEs is optional. Although the book is not heavy with theory, we have found that the concepts can be understood without an excessive amount of it.

The next chapters of the book provide a mild introduction to generalized linear models. Here students are able to see that it is possible to model responses which are not normally distributed. Not only do the methods have applications in realistic settings, but they also aim to broaden the students’ notion of modeling. With this kind of approach, students will begin to see the possibilities for other approaches to modeling.

The next transition is to correlated data. We have a chapter aimed at understanding just what is meant by correlated data. There are simulations and simulated applications based on real life projects to help fix ideas. We introduce ways correlation can be incorporated into modeling which leads to the remainder of the text which concerns multilevel modeling.

We have used the software package R to carry out computations and construct graphs. Sometimes courses using R can tend to become programming classes where the statistics content is tangential. To avoid that circumstance, we have opted to not include the code within the chapters, but rather include important portions of the code at the end of each chapter. Currently this practice is a little uneven and it is one of the features we will be working on to make more consistent and streamline.

Many portions of the book are under construction and it is possible there are errors that will not appear in the final edition (we hope!) and missing sections that will appear. Even these *Instructors Notes* are preliminary and likely to change.

We’re convinced that this textbook in its entirety is doable in a semester long course. We have
used various versions of it for over 10 years and found that when students successfully complete a regression course, you can move through the text at a reasonable pace and finish it. We have marked some of the more advanced topics optional and you can eliminate them without disturbing the flow of the ideas. You may also be interested in only the GLM material or only the Multilevel Models. These sections can stand on their own. GLM might make for a section in an advanced regression course or Multilevel, being more dense and lengthy may be the basis for an entire semester long course. Because we believe it is possible to cover the entire text, we encourage you to study the chapter on correlation to understand why we find the combination of these sections useful.

Rather than being a black box, we have worked to create a text based on ideas rather than computation or recipes. This should allow students to take what they learn and apply it in other settings with different software or even create their own approaches to modeling. The text as it is will provide students a cornucopia of approaches and ways of thinking about statistics. Your students are likely to find the approaches here if they venture into data science. We hope you and your students broaden your statistical horizons and enjoy your journey through the text!

Julie and Paul
Chapter 1

Introduction

Ecologists count species, criminologists count arrests, and cancer specialists count cases. Political scientists seek to explain who is a Democrat, pre-med students are curious about who gets in to medical school, and sociologists study which people get tattoos. In the first case, ecologists, criminologists and cancer specialists are concerned about outcomes which are counts. The political scientists', pre-med students’ and sociologists’ interest centers on binary responses; Democrat or not, accepted or not, and tattooed or not. We can model these non-Gaussian, non-normal responses in a more natural way by fitting generalized linear models (GLMs) as opposed to using to ordinary least squares (OLS) models.

When models are fit to data using OLS, inferences are possible if we can assume that there is a linear relationship between the response (Y) and an explanatory variable (X), the observations are independent of one another, the responses are approximately normal for each level of the X, and the variation in the responses is the same for each level of X. If we intend to make inferences using GLMs, these assumptions are different. First, we will not be constrained by the normality assumption. When conditions are met, GLMs can accommodate non-normal responses such as the counts and binary data in our preceding examples. While the observations must still be independent of one another, the variance in Y at each level of X need not be equal nor does the assumption of linearity between Y and X need to be plausible.

However GLMs cannot be used for models in the following circumstances. Medical researchers collect data on patients in clinical trials weekly for 6 months, rat dams are injected with teratogenic substances and their offspring are monitored for defects, and musicians’ performance anxiety is recorded for several performances. Each of these examples involves correlated data; in contrast to different patients, the same patient’s outcomes are more likely to be similar from week-to-week, litter mates are more likely to suffer defects at similar rates in contrast to unrelated rat pups, and a musician’s anxiety is more similar from performance to performance than it is with other musicians. Each of these examples violate the independence assumption of simpler linear models for OLS or GLM inference.

The Generalized Linear Models in the book’s title extends models to handle responses that are non-normal in addition to normal outcomes. The Multilevel Methods will allow us to create models for situations where the observations are not independent of one another. Overall, these approaches will permit us to get much more out of data and may be more faithful to the actual data structure than models based on ordinary least squares (OLS).

In order to understand the motivation for handling violations of assumptions, it is helpful to be able to recognize the model assumptions for inference with OLS in the context of different studies. While linearity is sufficient for fitting an OLS model, in order to make inferences and predictions the observations must also be independent, the responses should be approximately normal at each level of the predictor, and the standard deviation of the responses at each level of the predictors
should be approximately equal. After examining circumstances where inference with OLS modeling is appropriate, we will look for violations of these assumptions in another set of circumstances. These are settings where we may be able to use the methods of this text.

We’ve kept the examples in the exposition simple to fix ideas. There are exercises which describe more realistic and complex studies.

1.1 Ordinary Least Squares (OLS) Assumptions

Recall that making inferences or predictions with models fit using ordinary least squares (OLS) require that the following assumptions be tenable. The acronym LINE can be used to recall the assumptions required for making inferences with models based on OLS.

(L) there is a linear relationship between the mean response (Y) and the explanatory variable (X),

(I) the observations are independent,

(N) the responses are normally distributed at each level of X, and

(E) the variance or equivalently the standard deviations of the response are equal for all levels of X.

These assumptions are depicted in Figure 1.1. The mean value for Y at each level of X falls on the regression line. We’ll need to check the design of the study to determine if the observed values for Y and the corresponding values of X are independent of one another. At each level of X, the values for Y are normally distributed with the mean value of Y for X=1 at 4, the mean value of Y at X=2 is 5 and so on. The spread in the Y’s for each level of X is the same satisfying the equal standard deviation assumption.

1.1.1 Cases that do not violate the OLS assumptions for inference

It can be argued that the following studies do not violate the OLS assumptions for inference. We begin by identifying the response and the explanatory variables followed by describing each of the LINE assumptions in the context of the study. We comment on possible problems with the assumptions.
Ordinary Least Squares (OLS) Assumptions

1. **Reaction times and car radios** A researcher suspects that loud music can affect how quickly drivers react. She randomly selects drivers to drive the same stretch of road with varying levels of music volume. Stopping distances for each driver are measured along with the decibel level of the music on their car radio.

   **Response:** Reaction time  
   **Explanatory variable:** Decibel level of music

   The OLS assumptions for inference would apply if:
   
   **L:** The mean reaction time is linearly related to decibel level of the music.  
   **I:** Stopping distances are independent. The random selection of drivers should assure independence.  
   **N:** The stopping distances for a given decibel level of music vary and are normally distributed.  
   **E:** The variation in stopping distances should be approximately the same for each decibel level of music.

   There are potential problems with the linearity and equal standard deviation assumptions. For example, if there is a threshold for the volume of music where the effect on reaction times remains the same, mean reaction times would not be a linear function of music. Another problem may occur if initially the variation in reaction times is large but as the volume of the music increases the variation in reaction times is smaller. In this case, the equal standard deviation assumption would be violated. Often we can think of circumstances where the OLS assumptions may be suspect. Here a plot of the reaction times by music volumes would help in assessing these two assumptions. In general it is a good idea to plot y by x to check assumptions. This is one of the first things we would do in when exploring our data.

2. **Crop yield and rainfall** The yield of wheat per acre for the month of July is thought to be related to the rainfall. A researcher randomly selects acres of wheat and records the rainfall and bushels of wheat per acre.

   **Response:** Yield of a field measured in bushels per acre for July  
   **Explanatory variable:** Rainfall measured in inches for July

   **L:** The mean yield per acre is linearly related to rainfall.  
   **I:** Fields’ yields are independent. The random selection of fields should assure independence if fields are not close to one another.  
   **N:** The yields for a given amount of rainfall are normally distributed.  
   **E:** The standard deviation of yields is approximately the same for each rainfall level.

   Again we may encounter problems with the linearity assumption if mean yields increase initially as the amount of rainfall increases after which excess rainfall begins to ruin crop yield.

3. **Heights of sons and fathers** Galton suspected that a son’s height could be predicted using the father’s height. He collected observations on heights of fathers and their firstborn sons.

   **Response:** Height of the firstborn son  
   **Explanatory variable:** Height of the father

   **L:** The mean height of firstborn sons for fathers of a given height is linearly related to heights of fathers.  
   **I:** The height of one firstborn son is independent of the height of other firstborn sons in the study. This would be the case if firstborn sons were randomly selected.  
   **N:** The heights of firstborn sons for a given fathers’ height are normally distributed.  
   **E:** The standard deviation of firstborn sons’ heights at a given father’s height are the same.
Heights and other similar measurements are often normally distributed. There would be a problem with the independence assumption if all sons not just firstborn sons were selected.

1.1.2 Cases where the OLS assumptions for inference are violated

1. **Grades and studying** Is the time spent studying predictive of success on an exam? The time spent studying for an exam, in hours, and success, measured as Pass or Fail, are recorded for randomly selected students.

   **Response:** Exam grade (Pass or Fail)
   **Explanatory variable:** Time spent studying (in hours)

   Here the response is a binary outcome which violates the OLS inference of a normally distributed response at each level of X. Later in the course we see logistic regression which is more suitable for models with binary responses.

2. **Income and family size** Do wealthy families tend to have fewer children compared to lower income families? Annual income and family size are recorded for a random sample of families.

   **Response:** Family size, number of children
   **Explanatory variable:** Annual income, in dollars

   Family size is a count taking on integer values from 0 to technically no upper bound. The normality assumption may be problematic again because the distribution of family size is likely to be skewed, with more families having one or two children and only a few with a much larger number of children. Both of these concerns lead us to question the validity of the normality assumption. Study design should also specify that families are done adding children to their family.

3. **Exercise, weight and gender** Investigators collected the weight, sex, and amount of exercise for a random sample of college students.

   **Response:** Weight
   **Explanatory variable:** Sex and hours spent exercising in a typical week

   With two predictors, the assumptions now apply to the combination of sex and exercise. For example the linearity assumption implies that there is a linear relationship in mean weight and amount of exercise for males and similarly a linear relationship in mean weight and amount of exercise for females. This data may not be appropriate for OLS modeling because the standard deviation in weight for students who do not exercise for each sex is likely to be considerably more than the standard deviation in weights for students who follow an exercise regime. Thus, equal standard deviation at these different levels of weight may not be equal. Again, we can assess this potential problem by plotting weight by amount of exercise for males and females separately. There may also be a problem with the independence assumption because there is no indication that the subjects were randomly selected. There may be subgroups of subjects likely to be more similar, e.g. selecting students at a gym and others in a TV lounge.

4. **Surgery Outcome and Patient Age** Medical researchers investigated the outcome of a particular surgery for patients with comparable stages of disease but different ages. The ten hospitals in the study had at least two surgeons performing the surgery of interest. Patients were randomly selected for each surgeon at each hospital. The surgery outcome was recorded on a scale of one to ten.

   **Response:** Surgery outcome, scale 1-10
   **Explanatory variable:** Patient age, in years
Outcomes for patients operated on by the same surgeon are more likely to be similar and have similar results. For example, if surgeons’ skills differ or a surgeon’s criteria for electing a patient for surgery varies, patient outcomes will not be independent of one another. Furthermore, outcomes of one hospital may be more similar possibly due to factors associated with different patient populations. The very structure of this data suggests that the independence assumption will be violated. Multilevel models will explicitly take this structure into account for a proper analysis of this study’s results.

While we identified possible violations of OLS assumptions for inference for each of the examples in this section, there may be violations of the other assumptions that we have not pointed out. Prior to taking this course, you have learned some ways to handle these violations such as applying variance stabilizing transformations or logging responses, but you will discover other models in this text that may be more appropriate for the violations we have presented. A fundamental premise of this text is to introduce you to methods beyond OLS and thereby broaden your statistical horizons.

### 1.2 Multiple Regression

Before diving into generalized linear models and multilevel modeling, we review key ideas from multiple regression using an example on car pricing. This data was collected by Shonda Kuiper and first appeared in the *Journal of Statistics Education* entitled “Introduction to Multiple Regression: How Much Is Your Car Worth?” The article provides an excellent review of multiple regression, we highlight only a few of the fundamental concepts here. Many of the concepts we review here carry over to the modeling we do throughout the text. Particularly important are notions of model fitting, checking model assumptions, comparing models and interpreting estimated coefficients.

#### 1.2.1 Car prices

This data set contains Kelly Blue Book prices for 2005 General Motors cars that are in excellent condition and less than one year old. Eleven other related variables including mileage, engine size, number of cylinders and cruise control are included.

##### 1.2.1.1 Exploratory Data Analysis

First we perform an exploratory data analysis to learn about each variable in our dataset, to see how our response, Price, is related to car features and how car features are related to one another. A scatterplot matrix which includes Price and the quantitative covariates appears in Figure 1.2. It is a useful matrix of plots. There is a histogram for each variable, scatterplots of our response Price versus the potential predictors, and plots of how the covariates are related to one.

We examine the first row of the matrix with Price on the vertical axis to get an idea of the relationship between the response and the covariates. The histogram of prices is skewed and may be in need of a transformation. Its form suggests that the log of Price may be a better response to analyze. The plot of Price by Mileage reveals a slight negative correlation with Price indicating as we would expect a depreciation with increasing mileage. A set of outliers appear to be a fixed distance above the linear trend. The points are hardtop Cadillac convertibles which appear to be appreciably more expensive than the other cars. The steeper slope for a line fit to these points suggests that these cars may depreciate at a faster rate than the others. This plot underscores the importance of a graphical inspection. Simply using individual outlier diagnostics might miss the association. It is not advisable to remove outliers without a compelling reason. Those points may provide important information on how Price is related to Mileage and the other car characteristics.
We also note on the scatterplot matrix that **Price** is increasing with the three possible values for cylinders. **Liter** appears to be bimodal with higher values at around 2 and 4. **Price** differs significantly for each of the categorical variables with higher prices for 2-doors, cruise control, leather and better sound systems. There is a large number of observations (over 800) and a multitude of tests so these results are not all that surprising.

### 1.2.2 Simple Linear Regression

We specify a linear model with **Price** as the response and **Mileage** of the form \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \) where \( \epsilon_i \sim N(0, \sigma^2) \). We fit this model using the method of **least squares**, that is, by finding the values for \( \beta_0 \) and \( \beta_1 \) that will minimize the squared vertical distances to the regression line, \( \sum (Y_i - (\beta_0 + \beta_1 X_i))^2 \). For inference, we assume that \( \epsilon_i \sim iidN(0, \sigma^2) \). Output from a model with **Price** as the response and **Mileage** as a covariate appears below.

**Price ~ Mileage**

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 2.476e+04 | 9.044e+02 | 27.383 | < 2e-16 *** |
| Mileage | -1.725e-01 | 4.215e-02 | -4.093 | 4.68e-05 *** |

Residual standard error: 9789 on 802 degrees of freedom
Multiple R-squared: 0.02046, Adjusted R-squared: 0.01924
F-statistic: 16.75 on 1 and 802 DF, p-value: 4.685e-05
The data points along with the estimated regression line, $\hat{\mu} = 24760 - 0.1725$ \textbf{Mileage}, appear in the upper left plot of the scatterplot matrix. An extremely small p-value is associated with the coefficient for \textbf{Mileage} indicating its importance in the model. $R^2$ and the adjusted $R^2$, on the other hand, are extremely small suggesting that the model explains very little of the variation in \textbf{Price}. This is not a contradiction. The test for the coefficient found that the slope of the least squares line differs significantly from 0. The poor $R^2$ implies that there is a lot of variation in \textbf{Price} remaining as evident from the wide band of points about the estimated regression line.

It is good practice to carry out some diagnostics when fitting a regression line to check the assumptions. Linearity, independence, normality and equal variance must hold. We can do this with residual plots. The plot of the residuals versus the fitted values appears in Figure 1.3. It is a little difficult to assess the linearity assumption with the large number of observations, but what we can see looks ok. The independence assumption, as before, should be assessed by the study design and investigating any clumping in the residual plots. The normality assumption can be checked using a qqplot of the residuals. Problems with the common variance assumption appear as fanning on the residual versus fitted values plot.

The plot exhibits some fanning, the variation in the points appears to increase as the fitted values increase suggesting that the constant variance assumption is violated. One way to remedy this is to use the log of \textbf{Price} instead of \textbf{Price}. This is consistent with our observation concerning the distribution of \textbf{Price} in the scatterplot matrix. This plot also reveals a set of outliers that form a line above the general pattern of points and least squares line on the plot. We had spotted these points on the scatterplot matrix. It suggests that there is a missing covariate, possibly \textbf{Make} which we will address later when model building.
10 INTRODUCTION

1.3 Model Building

Model building may be undertaken for a number of reasons. One possible goal is description. We may wish to describe how one or more covariates are related to a response in which case we will be concerned about how the covariates are related to one another, a topic we take up in the next section. A related goal may be to describe the relationship between a covariate and a response controlling for other variables. Another goal of model building may be prediction. Our focus may not center so much on the coefficients but rather how well a model predicts the response. Whatever the goal, a primary aim of model building is parsimony. A null model with no covariates uses the overall mean to predict every observation. On the other hand, a saturated model admits a parameter for every observation, so that the "model" is the data set itself. The principle of parsimony leads to a model somewhere between these two extremes. We want to include all useful covariates and exclude any that are not, we want the model to be parsimonious. This is easier said than done. Whatever is done, subject area expertise should be brought to bear during variable selection and interpretation of the results.

1.3.1 Multicollinearity

The scatterplot matrix can also be used to investigate how the covariates relate to one another. Cylinders and Liter appear to be highly correlated from the plot which is confirmed by a correlation coefficient of 0.95. We want to be aware of correlations among the potential predictors referred to as multicollinearity, because collinear covariates may be providing redundant information when explaining variation in a response. For example, if we fit a model predicting Price as a linear function of Cylinder, another predicting Price as a linear function of Liter and a third model with both Cylinder and Liter we find the following:

Price ~ Cylinder

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -17.06   | 1126.94    | -0.015  | 0.988    |
| Cylinder       | 4054.20  | 206.85     | 19.600  | <2e-16   *** |

Residual standard error: 8133 on 802 degrees of freedom
Multiple R-squared: 0.3239, Adjusted R-squared: 0.323

Price ~ Liter

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 6185.8   | 846.7      | 7.306   | 6.66e-13 *** |
| Liter          | 4990.4   | 262.0      | 19.050  | < 2e-16 *** |

Residual standard error: 8207 on 802 degrees of freedom
Multiple R-squared: 0.3115, Adjusted R-squared: 0.3107

Price ~ Liter + Cylinder

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 1372.4   | 1434.5     | 0.957   | 0.339    |
| Liter          | 1412.2   | 903.4      | 1.563   | 0.118    |
| Cylinder       | 2976.4   | 719.8      | 4.135   | 3.92e-05 *** |

Residual standard error: 8126 on 801 degrees of freedom
Multiple R-squared: 0.3259, Adjusted R-squared: 0.3242
Both Cylinder and Liter coefficients are statistically significant when appearing as the sole covariate, but once in a model together Liter is no longer significant. Once the covariate Cylinders explains as much variation in price that it can, Liter was unable to explain much more of the remaining unexplained variation. The reverse is not true so Cylinder so it is significant and Liter is not.

It should be noted that depending upon the purpose for carrying out regression modeling, multicollinearity may or may not be a concern. If the purpose is prediction multicollinearity may not hurt - it may improve your model fit and improve your prediction. That is not the case here. The adjusted $R^2$ for the model with both covariates is not much more than the model with only Cylinder. On the other hand, if the purpose of the regression is to describe how individual variables are associated with the response, multicollinearity can be a concern. Note how different the coefficient estimates are for Cylinder in the model with only Cylinder and the model with both covariates. Coefficient estimates can be quite unstable in the presence of multicollinearity covariates so it is best to be aware of it when the focus is on individual covariates.

1.3.2 Variable Selection

Kuiper (2008) used Minitab to fit models of all of possible subsets of the covariates from which she selected the “best” model. Her approach included using adjusted $R^2$, Mallows Cp, and the standard deviation of the error term in the model. These measures will be discussed in later sections. While these measures do not consider the significance of individual coefficients and while automated approaches such as stepwise are not generally recommended, with all possible subsets she could consider models important to the analysis such as models with the Mileage variable, which can be forced in the model with some procedures. The balance between too few and too many predictors or parsimony can be kept in mind when selecting a model from all possible subsets. Only quantitative and binary variables with logPrice as the response were used at this stage.

<table>
<thead>
<tr>
<th>Comparing Multiple Regression Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Values favoring the larger model</td>
</tr>
<tr>
<td>• $R^2 = (\text{TOTSS - RESIDUALSS})/(\text{TOTALSS})$ (want large)</td>
</tr>
<tr>
<td>• adjusted $R^2$ (takes into account the number of parameters, want large)</td>
</tr>
<tr>
<td>• F-test ratio (want large)</td>
</tr>
<tr>
<td>• Mallow’s Cp (want to be roughly equal to the number of parameters)</td>
</tr>
<tr>
<td>• BIC (Bayesian Information Criterion, want small)</td>
</tr>
<tr>
<td>• AIC (Akaike Information Criterion, want small)</td>
</tr>
</tbody>
</table>

The model that Kuiper settled on was:

$$ \text{price} = 9.27323 - 0.000007\text{Mileage} + 0.1302\text{Cylinder} - 0.03\text{Doors}$$

$$ + 0.32\text{Cruise} - 0.09\text{Sound} + 0.12\text{Leather} $$

The model was based on comparisons of adjusted $R^2(0.480)$, Mallows Cp(10), and BIC(-485). We force Mileage into the model. Based on our earlier observations, the categorical variable, Make, was added to the model. There are six different makes in the data set. To include them in the model, we create five indicator variables that will allow for a comparison of each make to the baseline make, Buick. The addition of Make considerably improved the model with the adjusted $R^2$ of 0.480.
increasing to 0.90. All of the coefficients are statistically significant except \textit{Sound}. To complete the analysis we add in the categorical variable, \textit{Type} with indicator variables as well. The adjusted $R^2$ is now 0.93 and all of the coefficients are significant including \textit{Sound} (at the 0.057 level). The coefficient for \textit{TypeWagon} is not calculated because it results in a singularity, that is, there is a combination of of the existing covariates that produces the same separation of the data as \textit{TypeWagon} does so it is not needed in the model.

We can compare these models which add multiple predictors such as \textit{Make} and \textit{Type} by performing $F$-tests which use the difference in the residual sums of squares to formally assess the inclusion of these new variables. The output for these comparisons appears below. Note that factor will automatically create the needed indicator levels for us.

\begin{table}[h]
\begin{tabular}{cccccc}
Res. Df & RSS & Df & Sum of Sq & F & Pr(>F) \\
1 & 797 & 69.740 & & & \\
2 & 792 & 13.339 & 5 & 56.401 & 927.53 < 2.2e-16 *** \\
3 & 789 & 9.596 & 3 & 3.743 & 102.60 < 2.2e-16 *** \\
\end{tabular}
\end{table}

The numerator of the $F$-ratio test statistic is the difference in the residual sums of squares (given in the column labelled “Sum of Sq”) divided by the corresponding difference in the residual degrees of freedom. The divisor is the residual sum of squares for the larger model divided by its degrees of freedom. For example, to compare Model 2 to Model 3, we have a difference in the residual sum of squares of 3.743 which we divide by the 3 df difference between the two models. This numerator conveys the improvement using Model 3 compared to Model 2. The denominator is the residual sums of squares for Model.3 (9.596) divided by the number of degrees of freedom associated with the residuals for Model.3.

\[
F - ratio = \frac{(-3.743)/3}{3.743/789} = 102.60
\] (1.1)

If Model.2 is true and there is no need to add the \textit{Type} variable, the $F$-ratio would follow an $F$-distribution with 3 and 789 degrees of freedom. Under Model.2 we expect the $F$-ratio to be close to 1.0. With an $F$-ratio greater than 100, we have evidence that Model.2 is significantly improved with the addition of \textit{Type} to the model. Note that this test requires that the two models being compared are nested.

In place of seeing the data points surrounding an estimated regression line on a graph, now with two or more covariates we have points surrounding planes or hyperplanes. We can check the assumptions by examining the residuals, $Y_i - \hat{Y}_i$. We can assess the linearity assumption in multiple regression by examining a plot of the residuals versus the fitted values, $\hat{Y}_i$s. If the linearity assumption holds the residuals should appear as random scatter with a mean of 0. The residual plots are much improved however there remains some clumping which may indicate the need for other covariates.

In the coming chapters you will find that the assumptions for inference with OLS models will be relaxed. You will perform similar procedures, fitting models, comparing models, and assessing assumptions. You will discover that fitting and comparing models though similar in spirit, will not be based on residuals and sums of squares, but be based on likelihoods (more on this in Chapter ??)
with the OLS models as a special case.

1.4 Preview

This section provides a preview of the type of problems you will encounter in the book. This not only helps you give you the big picture of the text but it will also be useful if you are planning a final project. We introduce problem types along with a description of the type of response and when applicable, the structure of the data. With each description, we are assuming that the conditions for the type of variables considered are met which is not always the case in practice.

1.4.1 Soccer


Abstract

Approach motivation, a focus on achieving positive outcomes, is related to relative left-hemispheric brain activation, which translates to a variety of right-oriented behavioral biases. [...] In our analysis of all Federation Internationale de Football Association (FIFA) World Cup penalty shoot-outs, we found that goalkeepers were two times more likely to dive to the right than to the left when their team was behind, a situation that we conjecture induces approach motivation. Because penalty takers shot toward the two sides of the goal equally often, the goalkeepers right-oriented bias was dysfunctional, allowing more goals to be scored.

The response for this analysis is the direction of the goalie dive. You could define the binary variable, a variable which takes on only one of two possible values. For example, let Y=1 is the dive is to the right and Y=0 if the dive is to the left. This response is clearly not normally distributed. One approach to the analysis is logistic regression as described in Chapter ??.

A binomial random variable could be created for this application by summing the binary variables for each game so that Y= the number of dives right out of the number of dives the goalie makes during a game. These types of responses also can be analyzed using logistic regression.

1.4.2 Elephant Mating

Elephant Mating


Poole and her colleagues recorded for each male elephant, his age, in years, and the number of matings for a given year. The researchers were interested in how age affects the males' mating patterns. Specifically, questions concern whether there is a steady increase in mating success or if there is an optimal age after which the number of matings decline. Because the response of interest are counts (number of matings for each elephant for a given year), we will consider a Poisson regression.

The response for this study is the number of matings per elephant per year. The general form is the count for a specified time, volume or space. The possible values in theory are 0, 1, ... , \infty. This kind of data can be analyzed using Poisson regression when the appropriate conditions are met.
There are a number of modifications of Poisson regression which are useful when the conditions are not met.

1.4.3 Parenting and Gang Activity.

Walker-Barnes and Mason (2001) describe “Ethnic differences in the effect of parenting on gang involvement and gang delinquency: a longitudinal, hierarchical linear modeling perspective.” In this study, 300 ninth graders from one high school in an urban southeastern city were assessed at the beginning of the school year about their gang activity, the gang activity of their peers, behavior of their parents, and their ethnic and cultural heritage. Then, information about their gang activity was collected at 7 additional occasions during the school year.

The response for this study is a gang activity measure which ranges from 1 to 100. While it may be reasonable to assume this measure is approximately normal, the structure of this data implies that it is not a simple regression problem. Individual students have measurements made at 7 different points in time. We cannot assume that we have 2400 independent observations as the same measurements on one individual are more likely to be similar than a measurement of another student. Multilevel modeling as discussed in Chapters ??, ??, and ??.

1.4.4 Crime

An Analysis of the NYPD’s Stop-And-Frisk Policy in the Context of Claims of Racial Bias

Andrew Gelman, Alex Kiss, Jeffrey Fagan

Columbia Public Law Research Paper No. 05-95

Abstract: Recent studies by police departments and researchers confirm that police stop racial and ethnic minority citizens more often than whites, relative to their proportions in the population. However, it has been argued stop rates more accurately reflect rates of crimes committed by each ethnic group, or that stop rates reflect elevated rates in specific social areas such as neighborhoods or precincts. Most of the research on stop rates and police-citizen interactions has focused on traffic stops, and analyses of pedestrian stops are rare. In this paper, we analyze data from 175,000 pedestrian stops by the New York Police Department over a fifteen-month period. We disaggregate stops by police precinct, and compare stop rates by racial and ethnic group controlling for previous race-specific arrest rates. We use hierarchical multilevel models to adjust for precinct-level variability, thus directly addressing the question of geographic heterogeneity that arises in the analysis of pedestrian stops. We find that persons of African and Hispanic descent were stopped more frequently than whites, even after controlling for precinct variability and race-specific estimates of crime participation.

This application involves multilevel data, stops within precincts and a response number of stops. This type of analysis will be the last type you encounter, generalized multilevel modeling as addressed in Chapter ??.

1.5 Exercises

1.5.1 Conceptual

Applications that do not violate the OLS assumptions for inference

Identify the response and explanatory variable(s) for each problem. Write the OLS assumptions for inference in the context of each study.

1. Researchers record the number of cricket chirps per minute and temperature during that time to investigate whether the number of chirps can be varies with the temperature.
Table 1.1: Data for a clinical trial by dose, number at each dose level, and proportion getting relief.

<table>
<thead>
<tr>
<th>dose</th>
<th>n</th>
<th>y</th>
<th>proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.72</td>
<td>60</td>
<td>13</td>
<td>.22</td>
</tr>
<tr>
<td>1.76</td>
<td>62</td>
<td>18</td>
<td>.29</td>
</tr>
<tr>
<td>1.78</td>
<td>56</td>
<td>28</td>
<td>.50</td>
</tr>
<tr>
<td>1.81</td>
<td>63</td>
<td>52</td>
<td>.83</td>
</tr>
<tr>
<td>1.84</td>
<td>59</td>
<td>53</td>
<td>.90</td>
</tr>
<tr>
<td>1.86</td>
<td>62</td>
<td>61</td>
<td>.98</td>
</tr>
<tr>
<td>1.88</td>
<td>60</td>
<td>60</td>
<td>1.00</td>
</tr>
</tbody>
</table>

2. A random selection of women aged 20-24 years are selected and their shoe size is used to predict their height.

1.5.2 Guided Exercises

Applications that do violate the OLS assumptions for inference

All of the examples in this section have at least one violation of the OLS assumptions for inference. Begin by identifying the response and explanatory variables. Write the assumptions in the context of each study as if there were no violations. Then identify which assumptions are likely to be invalid. Explain your answer in the context of the study.

1. **Fragile Family** Researchers using the Fragile Family data are attempting to see if socioeconomic status and parental stability are predictive of low birthweight. Note that a newborn is classified as having a low birthweight with potentially hazardous consequences if their birthweight is less than 2500 gm, hence our response is binary 1 for low birthweight 0 when the birthweight is not low.

2. **Computer Mating** eHarmony.com claims that it is responsible for 2% of the marriages in the U.S. eHarmony.com researchers want to determine which factors predict a successful match.

3. **Clinical Trial I** A Phase I clinical trial is designed to compare the number of patients getting relief at different dose levels.

4. **Clinical Trial II** A randomized clinical trial investigated postnatal depression and the use of an estrogen patch (Gregoire et al. 1996). Patients were randomly assigned to either use the patch or not. Depression scores were recorded on 6 different visits.

5. **Schizophrenia** Thara et al. (1994) investigated the question “Does the course of schizophrenia differ for patients with early and late onset?” Doctors assess patients’ mood using a 0-100 scale at each of 10 visits.

6. **Pollution and traffic** Minnesota Pollution Control Agency is interested in using traffic volume data to generate predictions of particulate distributions as measured in counts per cubic feet.

7. **Canoes and zip codes** For each of over 27,000 overnight permits for the Boundary Water Canoe area, the zip code for the group leader has been translated to the distance traveled and socioeconomic data. This data is used to create a model for the number trips made per zip codes.

8. **Factors affecting beginning salary.** As part of a study investigating possible gender discrimination in beginning salaries at a particular company, researchers study the relationship between years of education and beginning salary among company employees.
9. **Elephant mating.** Researchers are interested in how elephant age affects mating patterns among males. In particular, do older elephants have greater mating success, and is there an optimal age for mating among males? Data collected includes, for each elephant, age and number of matings in a given year.

10. **Mandatory holds.** A mandatory hold is a revocation of a patients right to leave a psychiatric facility without consent of the monitoring physician. These holds have been classified as good or bad, and we want to know if certain groups (Native Americans, poor, uninsured) have rates of good holds that differ from others, after controlling for important factors in the hold decision (reason for hold, alcohol/drug use, physical threats, etc.).

11. **Beating the blues.** Can an interactive multimedia program of behavioral therapy help depression patients more than standard care? 167 subjects were randomized into one of two treatment groups, and their depression levels were assessed monthly for 8 months (2 active treatment and 6 follow-up). We want to compare treatments while controlling for baseline depression, concomitant drugs therapy, etc.

12. **Basketball referee bias.** Do college basketball referees tend to even out the foul calls on the two teams over the course of a game? For example, if several more fouls have been called on the visitors at a certain point in the game, does it become more likely that the next foul will be called on the home team? And do these chances depend on the score of the game, the size of the crowd, or the referees working the game?

13. **No Child Left Behind.** The Minnesota Department of Education makes all test scores publically available at the classroom level for each test type (math, science, reading), scores are available by grade within school within district. School boards are interested in mining this data to learn about how their district is performing, relative to other districts, after adjusting for demographic information at the school and district levels (e.g., percent free and reduced lunch, percent non-white, percent special education). They are also interested in quantifying the impact of potential policy changes, such as increased funding, decreased class size, additional charter schools, etc.

1.6 **Reference**


Chapter 2

Beyond Most Least Squares:
Using Likelihoods to Fit and Compare Models

2.1 Learning Objectives

After finishing this chapter, you should be able to:

- Describe the concept of a likelihood, in words.
- Know and apply the Principle of Maximum Likelihood.
- Identify three ways in which you can obtain or approximate an MLE.
- Use likelihoods to compare models.
- Construct a likelihood for a simple model.

This text encourages you to broaden your statistical horizons and to move beyond independent, identically distributed, Normal (iidN) data. The likelihood chapter focuses on ways to fit models, determine estimates, and compare models for most kinds of data, not just iidN data. In your earlier study of statistics, you often fit simple linear models using ordinary least squares (OLS). Fitting those models assumes linearity and independent observations. In this text, models are not always so simple. Responses may not be normally distributed. The observations may not be independent. For example, a study in education may involve 10 teachers, each having scores for multiple students. Scores for students of the same teacher are more likely to be correlated, not independent. Likelihood methods are often useful when the responses are counts or binary or when the observations are correlated. You might find likelihood methods a bit more complicated, but conceptually the approach is straightforward. As you go through the material here, worry less about calculus and computational details and focus on the concepts. You will have software to help you with computation, but interpretation will be up to you.

2.2 Case Study: Does sex run in families?

Doesn’t it seem that some families tend to have lots of boys while others have more than their share of girls? Is it really the case that each child human couples produce are equally likely to be a male or female? Or does sex run in families? These kinds of questions have implications for population demographics and sibling harmony. A 2009 study at the University of Ulster found that growing up with sisters, as compared to brothers, can enhance the quality of life of an adult. (cite cassidy http://news.bbc.co.uk/2/hi/7977454.stm).

To investigate this question and others, we look at the gender composition of 5,626 families collected by the National Longitudinal Survey of Youth (NLSY, 1994, citation inurl:bls.gov/nls). We fit models to explore whether there is evidence that the gender composition of families is affected by genetics (sex runs in families), by behavior (couples’ stopping behavior because of ‘waiting for a boy’ or ideal family sizes) or neither (random). (Rodgers and Doughty, 2001, CHANCE).
2.2.1 Research Questions

Three specific models we examine are listed below. While we are primarily interested in whether sex runs in families, it may be that the expected patterns are perturbed because of couples’ behavior such as desiring a specific family size or waiting for a boy. Any one of these models may exhibit gender imbalance. We may find that the proportion of boys is larger than 50% with model 1. With model 2, if sex runs in families we might find that sequences of boys in families which tend to produce boys may lead to an imbalance. With respect to the stopping rule models, it is evident that waiting for a boy may lead to an imbalance with fewer boys than girls, for example. It is not so clear the extent to which the ideal family size stopping rule will influence gender parity.

1. **Sex Unconditional Model (Random)** Is having children like flipping a coin with each child’s gender within families and between families independent of one another? What is the sex ratio (males to females at birth) or equivalently the probability of a boy birth for the US population?

2. **Sex Conditional Model (Genetic tendency)**: Do boys or girls run in families? That is, is there a tendency for families with more boys than girls to be more likely to produce another boy? Is the case the same for girls?

3. **Stopping Rule Models (Behavior effects)**:
   
   (a) Do couples have an ideal family size and will they stick with it regardless of the gender composition?
   
   (b) Do couples wait for boys?

Like any study using statistics, we need to consider how gender imbalance is measured. There are a number of ways this is typically done and you could come up with more measures on your own. Here are four intuitive measures which for our purposes, are defined as measured at birth. Obviously you would only need one of these four measures to gauge gender imbalance in a population.

- ratio of males to females
- ratio of females to males
- proportion of males \((p_B)\)
- proportion of females

Sibling harmony aside, why do people care about gender imbalance? Some comparisons of sex ratios between countries illustrate some compelling reasons. Some think that genetic or biological influences within families, such as “sex running in families,” can affect sex ratios. Mating behavior such as waiting until the family includes both sexes affects sex ratios. Before we get to our data, let’s consider the notion of gender imbalance in a more general setting.

2.2.2 Sex ratio for countries and sub-populations

The proportion of males and females born in different countries has become a highly contentious issue. There are a number of reasons why demographers and others feel it is important to measure and compare gender imbalance between countries and between subgroups within countries. Some believe that sex ratios point to the practice of sex selection in a country accomplished through abortion or infanticide. Furthermore, there is speculation that an excess of men could lead to unrest among young males unable to find marriage partners or start families.
It is curious to note that sex ratios were of interest to some early prominent statisticians. In 1930, the statistician Sir Ronald Fisher proposed an explanation for why the sex ratio should tend toward a 50:50 equilibrium for species that reproduce sexually. R. A. Fisher was a statistician renowned for his work in Analysis of Variance (ANOVA) and in the method of maximum likelihood, the subject of this chapter that is used throughout the text. Besides remarkable contributions to statistics, Fisher was also one of the founders of population genetics. Fisher posited his 50:50 equilibrium theory regarding sex ratios in terms of parental expenditure. He predicted that parental expenditure on both sexes should be equal. Sex ratios that are 1:1 are hence known as Fisherian, and those that are not 1:1 are non-Fisherian or extraordinary and occur because they break the assumptions made in Fisher’s model. (Roger and Doughty, 2001)

Most often in practice sex ratios differ from what Fisher predicted. From 1970 to 2002, the sex ratio at birth in the US among white non-Hispanics was 105 boys to 100 girls, but only 103 boys to 100 girls among African Americans and Indians (cite Matthews TJ, et al. (June 2005). "Trend Analysis of the Sex Ratio at Birth in the United States". National Vital Statistics Reports 53 (20.).). A 1997 study in Nature by Komdeur reports evidence which suggests that the human sex ratio may be currently shifting in the United States toward more female babies, closer to Fisher’s prediction! Sex ratio comparisons between countries are also intriguing. For example, Switzerland has a sex ratio of 106 boys to 100 girls whereas there are 112 boys to every 100 girls in China (cite the CIA, The World Fact Book, June 23, 2013, (cite https://www.cia.gov/library/publications/the-world-factbook/fields/2018.html). In the next section, we bring the notion of gender imbalance closer to home by focusing on families instead of countries or sub-populations.

2.2.3 Datasets

Before analyzing the NLSY data, we look at a few smaller datasets to gain an understanding of some features of likelihoods. The first four datasets are smaller and hypothetical, the last is an actual data set from the NLSY.

1. One family: the smallest dataset is a single family consisting of a boy, boy and then girl (BBG in that order)
2. Two families: BBG and GGG
3. n=50 children from 30 families (Table 2.4)
4. n=1,000 children: A proportion of boys of 0.60, the entire dataset is not shown.
5. Family Composition of Boys and Girls, NLSY: A subset of the National Longitudinal Survey of Youth (NLSY, 1994, citation inurl:bls.gov/nls) consisting of 1-, 2-, and 3-child families. (Table 2.5)
The NLSY dataset is large with family compositions for 5,626 respondents. We will perform a thorough analysis of this data with our research questions in mind later in the chapter. You will find the model fitting we do will be similar in flavor to what you have done in the past, only now we will be using likelihoods to obtain estimates and compare models.

2.3 Model 1: Sex Unconditional Model (Random)

This model likens having a child to flipping a loaded coin. Our goal is to incorporate the coin flipping into a likelihood and use it to find the “best” estimate for the probability of having a boy, \( p_B \). We use maximum likelihood criteria to define “best.” When coin flipping, we know that the probability of observing “heads” on the second flip is not affected by what happened on the first flip or future flips. Each flip is independent of other flips. With the Sex Unconditional Model, the chance the second baby is a boy in a family is not affected by the sex of the first or third child. If you believe the Sex Unconditional Model (or if the data leads you to believe it), you do not believe that ‘sex runs in families.’ We will be using likelihoods to fit and compare this model and others. We first explore some basic likelihood ideas using the smaller hypothetical datasets.

2.3.1 What is a likelihood?

As is often the case in statistics, our objective is to find an estimate for a model parameter using our data. Here, we’d like an estimate of the probability of a boy, \( p_B \), given our data. Recall, our data is the number of families for each family gender composition. On a smaller scale, if we observed a family of two boys followed by a girl (BBG), this data would provide information to help us estimate \( p_B \).

With likelihood methods, conceptually we consider different possible values for \( p_B \) and determine how likely we are to see our data in each case. We'll select as our estimate the value of \( p_B \) where our data is most likely. A likelihood is a function that tells us how likely we are to observe our data for each possible parameter value. Some examples of parameter values, \( p_B \), and the corresponding likelihoods appear in Table 2.2 for the data BBG. We see that if the true probability of a boy, \( p_B \), is 0.30 we only see our data (BBG) about 6.3% of the time whereas when \( p_B \) is 0.80, BBG is seen about 12.8% of the time. That is, we are more than twice as likely to see a family in our dataset whose first child is a boy, second is a boy and third is a girl if the true probability of having a boy is 0.80 rather than 0.30. We see that when the probability of a boy is 0.67, the likelihood is largest among the likelihood values displayed in the Table 2.2. When a likelihood is large, the value for the parameter is a plausible parameter estimate. If a likelihood is relatively small for a given parameter value, that parameter value would probably not make a good parameter estimate. The adjective relatively is important because in many settings likelihoods are quite small. It is not the magnitude of a likelihood per se it is how the magnitude of a likelihood compares to the other likelihoods in the same setting. Here \( p_B = 0.67 \) would make a better estimate than any of the other parameter values we consider in Table 2.2 because when the probability of a boy is 0.67, our data (BBG) occurs nearly 15% of the time, the maximum likelihood for any other parameter value from 0 \( \leq p_B \leq 1 \). This illustrates a fundamental idea used with likelihoods, the principle of maximum likelihood.

The Principle of Maximum Likelihood

Assess how likely the data is under each possible parameter value for a given model. The maximum likelihood estimate (MLE) is the value of the parameter that yields the largest likelihood.

In the sections that follow, we look at how likelihoods are constructed, we find better ways to identify the best estimate than making tables like Table 2.2, and we eventually apply this notion of likelihood estimation to our three models using the entire NLSY data set. First let’s start with a
smaller data set to illustrate some ideas about constructing likelihoods.

### 2.3.2 Constructing Likelihoods and finding MLEs.

**Children within families are independent. Data: One family with BBG.**

Here we provide instruction for how to calculate the likelihood values in Table 2.2. To do so, we need some probability basics including the notion of independence. For the Sex Unconditional Model, the probability of having a boy is the same for any child and it does not depend upon the sex of previous or future children. Under this model, the gender of each child is independent of the others in the family just like the result of one coin flip does not affect other flips. When observations are independent, we can multiply probabilities. For example, suppose the true probability of a male child is 0.51 and a 2-child family has two boys (BB). When the sex of each child is independent, 51% of the time the first child will be a boy and of that time 51% of the time the next will be a boy, so 0.51*0.51 or approximately 26% of the time we observe BB when the probability of a boy is 0.51. These ideas extend naturally. We can also multiply probabilities for a 3-child family of BBG so that the probability of BBG is (0.51)*(0.51)*(0.49) or \( 0.51^2 * 0.49 \), or 12.7%. These calculations assume that we know the true probability of a boy, 0.51, but in reality the probability of a boy is not known and so it is denoted by \( p_B \). A likelihood for BBG can be written as:

\[
\text{Lik}(p_B) = p_B^2 (1 - p_B). \tag{2.1}
\]

The data (BBG) appear in the likelihood. The values of parameters \( p_B \) are input into the likelihood which returns how likely the data is for a given \( p_B \). In fact, we can use this likelihood function to create Table 2.2. A graph helps illustrate the idea of a likelihood function. In Figure 2.3.2, you can see that our data appears to be most likely when \( p_B = 2/3 \) depicted by the solid line in Figure 2.3.2. So the Principle of Maximum Likelihood suggests that 2/3 would be our best estimate of \( p_B \). This is intuitively pleasing given that our data contains 2 boys out of 3 children. In contrast on the graph, we see that when \( p_B = 0.3 \) the likelihood is much smaller, close to 6.3% on Table 2.2, suggesting that our data is not as likely to occur when \( p_B \) is as small as 0.3 and values this small would not make good estimates for the probability of a boy. Note that \( p_G = 1 - p_B \).

**Families are independent. Data: 2 families with BBG and GGG.**

In the previous section we exploited the idea that children within families are independent and calculated probabilities for families by multiplying. This model specifies that children within a family are independent of one another, but it also implies that families are independent of one another. For example, if the first family we observe consists of BBG and the next family consists of GGG, then we can determine the likelihood of observing these two families for different values of \( p_B \) by multiplying and writing it as:

\[
\text{Lik}(p_B) = p_B^2 (1 - p_B) \times (1 - p_B)^3 = p_B^2 (1 - p_B)^4
\]
When observations are **independent**, probabilities may be multiplied to construct likelihoods.

2.3.3 MLEs and graphs of likelihoods

The random model implies that the results for each family are independent of one another and therefore not affected by other families’ compositions. As we had done in the previous example, we can graph this likelihood based on the data (BBG, GGG) and approximate what parameter estimates seem plausible for our data. See the lower panel in Figure 2.3.2. Also note that we have 2 boys out of 6 children total in the dataset and from the graph you can see that it appears that the likelihood has its maximum at 1/3 or 0.333. The two likelihoods look quite different because of the difference in the proportion of boys in the two samples.

2.3.4 MLEs using calculus

An alternative approach is to use calculus. We can find the value of $p_B$ where the likelihood is a maximum by finding the first derivative using the Product Rule, setting it to 0 and solving for $p_B$.

\[
\frac{dp_B^2(1-p_B)^4}{dp_B} = 2p_B(1-p_B)^4 - 4p_B^2(1-p_B)^3 \\
= 2p_B(1-p_B)^3(1-3p_B)
\]
FEATURES OF LIKELIHOODS

This result implies that the maximum likelihood may occur at \( p_B = 0 \) or \( p_B = 1 \) or \( p_B = 0.3333 \). Further investigation reveals that the MLE is 0.333333.

This is the same estimate we obtained when we used our intuition and the graph above.

2.3.5 MLEs using software

Yet another way of obtaining a maximum likelihood estimate is to use software to search numerically for the optimal parameter value. There are a large number of statistical programs that produce MLEs. Examples follow. These examples use R to find maximum likelihoods using grid searches. The first example is similar to what we had done with the table of some \( p_B \) and corresponding likelihood values in Table 2.2 and data for a single family. A relatively coarse grid is used to identify values for \( p_B \). The second example uses a finer grid and you can see this in the graph. The next set of data is for two families with compositions: BBG and GGG. This code runs on R. Give it a try.

```r
# Finding MLEs and graphs using software
# Data for two families: BGG, GGG
pb=seq(0,1,length=1001)  # possible values for prob a boy is born
lik=pb^2 * (1-pb)^4  # likelihood of getting observed data
plot(pb,lik,xlab="possible values of pb",ylab="Likelihood",type="l",
     main="Two Families: BBG, GGG")
max(lik)  # maximum likelihood over 1001 values of pb
pb[lik==max(lik)]  # value of pb where likelihood maximized
abline(v=pb[lik==max(lik)],col="red")
```

Using likelihoods to find estimates of parameters is conceptually intuitive - select the estimate for the parameter value where your data is most likely. Often MLEs make a lot of sense in the context of a problem as well - here an estimate of the probability of a boy with the proportion of boys in the data. We have presented three approaches:

**Approaches to Finding MLEs**

1. Graph the likelihood by possible parameter values. Approximate the MLE as the parameter value with the maximum likelihood.
2. Use calculus. Find the first derivative of the log-likelihood, set it equal to 0 and solve for the MLE.
3. Use software to find where the likelihood is a maximum.

Most importantly, work to understand the idea of a likelihood. Likelihoods are the foundation upon which estimates are obtained and models compared for most of the models in this course. Do not be overly concerned with calculus and computation at this point.

2.4 Features of Likelihoods

Before we get to the NLSY data, we consider several additional features of likelihoods including its definition, comparing likelihood and log-likelihood graphs for smaller (n=50) and larger (N=1000) problems and applying likelihood ideas when there is more than one parameter.
24 BEYOND MOST LEAST SQUARES: USING LIKELIHOODS TO FIT AND COMPARE MODELS

2.4.1 How does sample size affect the likelihood?

We consider two hypothetical cases both under the Sex Unconditional Model to compare the effect of the size of a sample.

**Case 1** Suppose our data consists of 50 children altogether, 30 of whom are boys. Under the independence model, our likelihood has the form given in Equation 2.2. The data (30 boys and 20 girls) appears in the likelihood, remaining fixed for each possible value of $p_b$.

\[
\text{Lik}(p_b) = p_b^{30} (1 - p_b)^{20} \tag{2.2}
\]

We’ll use the likelihood to determine how likely we are to observe our data (30 boys, 20 girls) for different possible values of $p_b$, $0 \leq p_b \leq 1$. If $p_b=0.1$, the likelihood for our data can be found by replacing $p_b$ in equation 2.2 with 0.1:

\[
\text{Lik}(0.1) = 0.1^{30} * (0.9)^{20} = 1.215767e - 31
\]

\[
\text{Lik}(0.3) = 0.3^{30} * (0.7)^{20} = 1.642852e - 19
\]

This is a very small likelihood, but it is larger than the likelihood for $p_b=0.1$. If we had to choose between $p_b=0.1$ and $p_b=0.3$ it seems our best bet would be $p_b=0.3$ because our data is more likely to be produced when $p_b=0.3$. But ultimately according to the **principle of maximum likelihood**, we’d like to identify the $p_b$ where our data is most likely to be produced to obtain the **maximum likelihood estimate (MLE)** of $p_b$. It isn’t going to be efficient to hunt and peck around to find this best value of $p_b$. As we have seen, we can use calculus or software to determine the MLE. But a graph provides some insights.

2.4.2 Constructing the Likelihood and its Graph

Let’s take a look at the graph of our likelihood in the left panel of Figure 2.2.

\[
\text{Lik}(p_b) = p_b^{30} (1 - p_b)^{20} \tag{2.3}
\]

A computer program can be used to sketch the graph of the likelihood for the 30 boys and 20 girls displayed here in Figure 2.2.

```r
# Data Set: 50 children from 30 families
# Evaluate finer grid of possible values of pb
pb=seq(0,1,length=1001)  # possible values for prob a boy is born
lik=pb^{30} * (1-pb)^{20}  # likelihood of getting observed data
plot(pb,lik,xlab="possible values of pb",ylab="Likelihood",type="l",
     main="50 children from 30 families")
max(lik)  # maximum likelihood over 1001 values of pb
pb[lik==max(lik)]  # value of pb where likelihood maximized
abline(v=pb[lik==max(lik)],col="red")
```

The graph of this likelihood is displayed in the left panel of Figure 2.2. The height of each point is the likelihood and the possible values for $p_b$ appear across the horizontal axis. Where does the likelihood appear the largest to you? A cursory inspection of the plot reveals what you may have suspected all along. In the next section, we determine the MLE more precisely using calculus.
2.4.3 MLEs and Calculus: Likelihoods vs log-likelihoods

Thus far, we have only approximated the MLE for the n=50 children dataset using graphs. The maximum of a likelihood can be determined more precisely using by taking the first derivative, setting it equal to 0 and solving for the MLE as we had done in Section 2.3.4. In this section, the data set was very small and it was not difficult to find the first derivative and solve for the MLE. It looks a little more challenging here and in future applications it could be considerably more difficult.

Here we use calculus to find the MLE for \( p_B \) by must differentiating Equation 2.4. Like this likelihood, most likelihoods are products. Differentiating the likelihood with respect to \( p_B \) can be onerous when working with larger data sets but taking the log of a likelihood can make our work much easier. An attractive benefit of using log-likelihoods is that instead of taking the (very tedious) derivative of a product (as with likelihoods) by using log-likelihoods you can take the (much easier) derivative of a sum. From the left panel of Figure 2.2 you can see that the likelihood appears to attain a maximum when \( p_B = 0.60 \), and from the right panel you can see the maximum of the log-likelihood is also at \( p_B = 0.6 \). Figure 2.2 illustrates how the graphs of the likelihood and log-likelihood compare. While the curves appear quite different, they both attain a maximum at \( p_B = 0.6 \).

This derivation of the MLE using calculus demonstrates how much easier it is to work with derivatives of log-likelihoods as opposed to derivatives of likelihoods. Log-likelihoods can also make computation easier when finding MLEs and comparing models because likelihoods can be very small and subject to error. An example of finding an MLE and log-likelihood using a software package follows.

With calculus we can show our data appears most likely when \( p_B = 30/50 = 0.6 \). This seems like a lot of work for such an obvious result but MLEs have some nice properties that lend some theoretical support for \( p_B = 0.6 \) and we’ll see that much more complex models can be fit using the principle of maximum likelihood. The theoretical properties of MLEs are spelled out in a later...
26 BEYOND MOST LEAST SQUARES: USING LIKELIHOODS TO FIT AND COMPARE MODELS

Figure 2.3 Likelihoods and LogLikelihoods for samples of n=50 and n=1000 from populations where \( p_B = 0.60 \).

Here we demonstrate finding the MLE using calculus.

\[
\text{Lik}(p_B) = p_B^{30}(1-p_B)^{20}.
\]
Logging and differentiating, we have
\[ \log\-lik(p_B) = 30 \log(p_B) + 20 \log(1 - p_B) \]
\[ \frac{d}{dp} \log\-lik(p) = \frac{30}{p} - \frac{20}{1 - p} \]

Setting this expression equal to 0 and solving for \( p_B \), we obtain the MLE, \( \hat{p}_B = 0.6 \).

In summary, we construct a likelihood that reflects features of our Sex Unconditional Model, we then approximate the parameter value for which our data is most likely using a graph or more precisely using calculus, or conveniently using software.

You may not be familiar with calculus, however the primary message is in the graphs. Find the value of \( p_B \) where the likelihood is a maximum. Our “best” estimate for \( p_B \), the MLE, is where our data is most likely to be observed.

**Case 2** Case 1 included 50 children. Case 2 is a much larger study with 600 boys and 400 girls. Our goal is still to find an MLE for \( p_B \) for the independence model, so we proceed as we had done in the smaller study in Case 1. Using the methods in Case 1, we can show that the MLE for \( p_B \) in Case 2 is also 0.60. You should confirm that using a graph, calculus, or software to solve for the MLE.

Graphs of likelihoods and log-likelihoods with different sample sizes (\( n = 50 \) and \( n = 1000 \)) appear in Figure 2.3. These hypothetical datasets have been constructed in order to compare the effect of sample size on the process of finding an MLE. The example with \( n = 50 \) and the example with \( n = 1000 \) each have an MLE of 0.60. Anything from a graphing calculator to a sophisticated computer program can be used to construct Figure 2.3. First the likelihoods for Case 1 (\( n = 50 \)) and Case 2 (\( n = 1000 \)) are compared and contrasted. The likelihoods have similar forms, however the graph for Case 2 is much narrower reflecting how much more certainty we have in our estimate given the large amount of data. Plausible estimates will not stray far from the true value of \( p_B = 0.6 \). Comparing the likelihoods to the log-likelihoods we find that in both cases (\( n = 50 \) and \( n = 1000 \)) the likelihoods and log-likelihoods do not look much alike however they are both concave down and they both have a maximum at the same value of \( p_B \). When comparing the likelihoods for the smaller and larger samples, the graph for the larger sample is narrower. The same is true when comparing the log-likelihoods.

### 2.4.4 More than one parameter

In the previous examples, we estimated only one parameter, the probability a couple has a a boy, \( p_B \). This estimate helps us learn about the sex ratio in this population. But there are many instances when we may want to estimate more than one quantity. Couples behavior is speculated to affect family composition. As a very simple example, we may want to estimate \( p_B \) and the probability that couples stop childbearing when they have a boy, \( p_S \). We can look at the same data we used previously and also note whether the family composition ends with the first boy, that is, identify family compositions that could be construed as waiting for a boy although we have no proof of that. A couple of questions to ponder; if we make the assumption that every family which stops after their first boy was, in fact, waiting for a boy would we be under- or overestimating \( p_S \)? In the other direction, one of the authors has four girls followed by a boy. Surprisingly another girl arrived the following year! Ignoring the fact that this family of six would not be in our data set, our algorithm would not characterize this family as waiting for a boy and maybe that is OK. Another consideration, if we are looking for ‘runs’ of girls or ‘runs’ of boys how would this stopping behavior affect the distribution of family composition? A dataset in Table 2.4 has the identical data for the number of boys and girls (30 and 20, respectively) but the presentation also allows us to assess couples’ stopping behavior. Let \( p_B \) be the probability of a boy as before and \( p_S \) be the probability of stopping
on an only boy. Note we assess every child as to whether they are a last, only boy or not and denote the probability of stopping on a last child who is a boy as \( p_S \). Every child is assessed as to whether they are a last child, only boy because ostensibly the decision to stop childbearing occurs with each child born. Some are assigned \( 1 - p_S \) because they are a girl, some because they are a boy but not the last, only boy.

The likelihood involves the two parameters \( p_B \) and \( p_S \)

\[
p_B^{30}(1 - p_B)^{20} p_S^{12}(1 - p_S)^{38}
\]

With two parameters, the likelihood appears as a surface.

When you are fitting models with more than one parameter, which will be often, keep in mind that you are, in essence, hiking around a surface or hypersurface trying to locate the highest point. In some of the more complex models, the surfaces will have multiple local maxima or very flat terrain so that it may be difficult to locate the maximum. This may be what is happening when you are using software to fit models and get the frustrating message of no convergence.

### 2.4.5 Is a likelihood a probability function? Optional

No. Even though we use probabilities to construct likelihoods, a likelihood is not a probability function. A probability function takes outcomes as input and outputs a probability of a particular outcomes. For example, you flip a loaded coin which comes up heads 25% of the time. After 5 flips, you observe the outcome of three heads and two tails. A probability function provides the probability of observing \((3H,2T)\) when \( p_B = .025 \). If you flip this same coin another 5 times and observe all tails \((5T)\), the probability function provides the probability of \((5T)\).

In contrast, a likelihood is constructed using the data, say \((3H,2T)\). It takes as input possible parameter values and returns the probability of seeing that data for the given parameter value. For example, the likelihood will provide the chance of seeing the data \((3H,2T)\) if \( p_B = .6 \), the likelihood
Our first research question involves determining whether sex runs in the family. Do families with boys tend to have more boys than expected by chance and do families with girls tend to have more girls than expected by chance? What do you think? And how could we use a statistical model to investigate this phenomenon? There are a number of different ways to construct a model for this question. Here’s one possibility.

Unlike the previous model, the $p_B$ in a Sex Conditional Model depends on existing family compositions. We introduce conditional notation to make the dependence explicit. The notation $P(A|B)$ represents the “probability of A given B has occurred.” Another way to read this notation is the “probability of A conditional on B.” Here, let $p_{B|N}$ represent the probability the next child is a boy given that there are equal numbers of boys and girls (sex-neutral) in the existing family. Let $p_{B|B_{bias}}$ represent the probability the next child is a boy if the family is boy-biased, i.e. there are more boys than girls prior to this child. Similarly, let $p_{B|G_{bias}}$ represent the probability the next child is a boy if the family is girl-biased, i.e. there are more girls than boys prior to this child.

Before we are mired in notation and calculus, let’s think about how these conditional probabilities can be used to describe sex running in families. While we only had one parameter, $p_B$, to estimate in the Sex Unconditional Model, here we have three parameters $p_{B|N}$, $p_{B|B_{bias}}$, and $p_{B|G_{bias}}$. Clearly if all three of these probabilities are equal, the probability a child is a boy does not depend upon the gender composition of the family and there is no evidence of sex running in families. A conditional probability $p_{B|B_{bias}}$ that is larger than $p_{B|N}$ suggests families with more boys are more likely to produce more boys in contrast to families with equal boys and girls. This finding would support the theory of “boys run in families.” An analogous argument holds for girls. In addition, comparisons of $p_{B|B_{bias}}$ to the parameter estimate $p_B$ from the Sex Unconditional Model may be interesting and can be performed by comparing models.

While it may seem that including families with a single child (singleton families) would not be helpful for assessing whether there is a preponderance of one sex or another in families, in fact we’ve already noted that comparisons will be made to $p_{B|N}$ and singleton families would be helpful in estimating $p_{B|N}$ because singletons join “neutral families.”

Using the family composition data for 50 children in the 30 families that appears in Table 2.3,
we construct a likelihood. The six singleton families (B) contribute $p_{Bj}^6$ to the likelihood and the seven families with only girls contribute $p_{Gj}^7$ or $(1 - p_{Bj})^7$. There are five families with two boys (BB) each with probability $p_{Bj}^5 * p_{Bj}^{Bbias}$ contributing:

$$[p_{Bj}^5 p_{Bj}^{Bbias}]^5.$$  

We construct the likelihood using data from all 30 families assuming families are independent to get:

$$lik(p_{Bj|N}, p_{Bj|Bbias}, p_{Bj|Gbias}) = p_{Bj|N}^{15} * (1 - p_{Bj|N})^{15} * p_{Bj|Bbias}^5 * (1 - p_{Bj|Bbias})^4 * p_{Bj|Gbias}^8 * (1 - p_{Bj|Gbias})^3$$  

(2.4)  

A couple of points are worth noting. First, there are 50 factors in the likelihood corresponding to the 50 children in these 30 families. Second, in the Sex Unconditional example, we only had one parameter, $p_B$, here we have three parameters. This likelihood does not simplify like the Sex Unconditional Model where we had only two powers in the likelihood; one of $p_B$ and the other of 1-$p_B$. The basic idea we discussed regarding using a likelihood to find a parameter estimate are the same. To obtain the MLEs, we need to find the combination of values for our three parameters where the data is most likely to be observed. Conceptually, we are trying different combinations of possible values for these three parameters, one after another, until we find the combination where the likelihood is a maximum. It will not be as easy to graph this likelihood and we will need multivariable calculus to locate that combination of parameters where the likelihood is a maximum. In this text we do not assume you know multivariable calculus, but we do want you to retain the concepts associated with maximum likelihood estimates. In practice, we use software to obtain values MLEs.

From Table 2.3 note that every first child enters a neutral family with respect to sex composition. It is also clear from this table that there is a probability associated with every child which can be multiplied to obtain likelihood contributions for each sex composition. As we noted with the Sex Unconditional Model, we can write $p_{Gj|Bbias} = 1 - p_{Bj|Bbias}$ and likewise for families with a girl bias. It is unnecessary to have both of the boy and girl parameters. Rewriting the girl parameters in terms of the boy parameters reduces the number of parameters in a model or corresponding likelihood. Table 2.3 should help to give you an idea of how the notation is defined for the Sex Conditional Model. Again there are other ways to do this as is often the case in statistical modeling.

With calculus, we can take partial derivatives of the likelihood with respect to each parameter assuming the other parameters are fixed. As we saw in the first example, differentiating the log of the likelihood often makes things easier. This same approach is recommended here. Set each partial derivative to 0 and solve for the parameter of interest (the one you differentiated with respect to). Repeat for the other two parameters.

Knowing that it is easier to work with log-likelihoods, let’s take the log of the likelihood we constructed Equation 2.5.

$$log\cdot lik(p_{Bj|N}, p_{Bj|Bbias}, p_{Bj|Gbias}) = 15log(p_{Bj|N}) + 15log(1 - p_{Bj|N}) + 5log(p_{Bj|Bbias}) + 4log(1 - p_{Bj|Bbias}) + 8log(p_{Bj|Gbias}) + 3log(1 - p_{Bj|Gbias})$$

Taking a partial derivative with respect to $p_{Bj|N}$

$$\frac{15}{p_{Bj|N}} - \frac{15}{1 - p_{Bj|N}} = 0$$
Table 2.4  Family contributions to the likelihood used to estimate the probability of a \( p_B \) and the probability of stopping on the first boy, \( p_S \), for a data set of \( n=50 \) children.

\[
\hat{p}_{B|N} = \frac{15}{30} = 0.5
\]

Partial differentiation of the log of the likelihood with respect to \( p_{B|B_{bias}} \) and with respect to \( B|G_{bias} \) leads to analogous expressions and MLEs for all three parameters can be recovered. The MLEs are:

\[
\hat{p}_{B|N} = 0.50 \\
\hat{p}_{B|B_{bias}} = 0.56 \\
\hat{p}_{B|G_{bias}} = 0.73 
\]

If we did not anticipate any “sex running in families” effect, we would expect \( \hat{p}_{B|B_{bias}} \) to be larger than the the probability of a boy in the neutral setting, \( p_{B|N} \). In our small hypothetical example, \( \hat{p}_{B|B_{bias}} \) is greater than 0.5 providing support for the “sex runs in families” theory when it comes to boys. What about girls? Do families with more girls than boys tend to have a greater probability of having a girl? We found that the MLE for the probability of a girl in a girl bias setting is 1-0.73=0.27. (A nice property of MLEs is demonstrated here. We have the MLE for \( p_{B|G_{bias}} \), we want the MLE of \( p_{G|G_{bias}} = 1 - p_{B|G_{bias}} \). We can get it by replacing \( p_{B|G_{bias}} \) with its MLE. Often you can get the MLE of a function by applying the function to the original MLE.) This data does not provide evidence that girls run in families; \( \hat{p}_{G|G_{bias}} = 0.27 \). There is a markedly lower probability of a girl if the family is already girl based. This data is hypothetical. Let’s take a look at some real data and see what we find.

2.6 Case Study: Analysis of the NLSY data

2.6.1 Model Building Plan

Finally we put all that you have learned together to find out if sex runs in families or couples wait for a boy or have an ideal family size that will lead them to stop having children. There are many different ways in which this analysis can be done to address the research questions. Our approach begins with the random model. If the random model does fit well, it may be that there is a genetic tendency affecting family composition. We can explore this possibility by comparing the Sex Conditional Model to the Random model. We would probably want to do this in any case because after all, this is our primary question of interest. If we achieve a significant improvement in the model with the ‘sex runs in families’ model, we will have evidence of a genetic tendency as a factor in gender composition. We may also want to investigate some behavioral factors, of which there
Family Composition of Boys and Girls, NLSY: Exploratory Data Analysis

We begin by performing an exploratory data analysis aimed at shedding some light on our research questions. We are looking for clues as to which of our three models is most plausible.

Table 2.5 displays family composition data for the 5,626 families with one, two, or three children in the data set. This dataset includes 10,672 children. Because our interest centers on sex ratios and proportions of males, let’s calculate sex ratios and proportions of males for each family size. For one-child families the male to female ratio is less than one (97 males:100 females) whereas the two- and three-child families have ratios of 104 boys to 100 girls, what we may expect in a population which favors males. Similarly the probably of a male birth is less than 0.50 for one-child families and more than 51% for the larger families. While our research questions do not specifically call for these measures stratified by family size, it will still provide us with an idea of gender imbalance in the data.

Table 2.6 provides insight into whether sex runs in families if the probability of a boy is 0.5. Simple probability suggests that the proportion of 2 child families with all the same sex would be 50% (BB, GG) but in our data we see only 45%. For 3-child families, we have 8 possible orderings of boys and girls and so we would expect 2 out of the 8 orderings to be of the same sex (BBB, GGG) or 25% to have all of the same sex, but in fact 27% have the same sex among the 3-children families. These results do not provide overwhelming evidence of sex running in families. There are some potentially complicating factors: the probability of a boy may not be 0.5 or couples may be waiting for a boy or a girl.

Table 2.7 contains the number of families by size and the percentage of those which are families with one boy who is last. So one might surmise that these families ‘waited’ and then quit childbearing after a boy was born. We see the proportion of one child families with a boy is slightly less than
Table 2.6 Proportion of families with all the same sex by number of children in the family. Note that 1-child families are all homogeneous with respect to sex so we look at 2- and 3-child families.

<table>
<thead>
<tr>
<th>Number of children</th>
<th>Number of families</th>
<th>Number with all same sex</th>
<th>Percent with same sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two children</td>
<td>2444</td>
<td>1112</td>
<td>45%</td>
</tr>
<tr>
<td>Three children</td>
<td>1301</td>
<td>345</td>
<td>27%</td>
</tr>
</tbody>
</table>

Table 2.7: Proportion of families with only one boy who is born last.

<table>
<thead>
<tr>
<th>Number of children</th>
<th>Number of families</th>
<th>Number with one boy last</th>
<th>Percent with boy last</th>
</tr>
</thead>
<tbody>
<tr>
<td>One child</td>
<td>1881</td>
<td>930</td>
<td>49.4%</td>
</tr>
<tr>
<td>Two children</td>
<td>2444</td>
<td>666</td>
<td>27.2%</td>
</tr>
<tr>
<td>Three children</td>
<td>1301</td>
<td>125</td>
<td>8.6%</td>
</tr>
</tbody>
</table>

the 50% expected. We’d expect one out of four, or 25%, of 2-child family configurations to have one boy last and there is 27% in our dataset. Only 8.6% of 3-child families has one boy last, but in theory we would expect one out of eight or 12.5% of 3-child families to have one boy last.

2.6.3 Constructing a Likelihood for the Sex Unconditional Model: the NLSY data

We can combine these ideas to construct a likelihood for the NLSY data for one-, two- and three-child families that appear in Table 2.5.

There are 177 families with BBG compositions in our data and each family contributes this same amount to the likelihood. See equation 2.1. Because the families’ compositions are independent, those 177 BBG families contribute to the likelihood:

\[
\text{Lik}(p_B) = [p_B^2(1-p_B)]^{177} = p_B^{354}(1-p_B)^{177}
\]  

(2.5)

The simplicity of the sex unconditional model results in a great simplification of the likelihood. There are 177 families of BBG which yield 2 \times 177 or 354 boys and 177 girls. Note that the contribution to the likelihood for these 177 families in equation 2.5 has 354 factors of \(p_B\) and 177 factors of \((1-p_B)\). Families with different compositions will contribute different factors to the likelihood. For example, here is a sample of contributions from other family compositions. See Table 2.5 to see where these numbers are coming from.

\[
\begin{align*}
G & (1-p_B)^{951} \\
GB & (1-p_B)^{666} p_B^{666} \\
BGB & p_B^{296}(1-p_B)^{148}
\end{align*}
\]

Now we are ready to put together the entire likelihood for our data under the Sex Unconditional Model.

\[
L(p_B) = p_B^{930} p_G^{951} p_{BB}^{582} \cdots p_{BBG}^{177} \cdots p_{GGG}^{159}
\]

\[
= p_B^{930+2+582+666+666+\cdots+125}(1-p_B)^{951+666+666+2+530+\cdots+3+159}
\]

or

\[
L(p_B) = p_B^{5416}(1-p_B)^{5256}.
\]

(2.6)

This very simple likelihood implies that each child contributes a factor of the form \(p_B\) or \(1-p_B\). Equation 2.7 indicates that there are 5416+5256= 10,672 children. What would be your best guess of
Beyond Most Least Squares: Using Likelihoods to Fit and Compare Models

The estimated probability of a boy for this model? That is, for what value of \( p_B \) is \( L(p_B) \) maximized?

Taking the log of the likelihood, then using calculus, differentiating with respect to \( p_B \), setting the derivative to 0, and solving for the MLE of \( p_B \):

\[
\log\text{-lik}(p_B) = 5416\log(p_B) + 5256\log(1 - p_B)
\]

\[
\frac{d}{dp_B} = \frac{5416}{p_B} + \frac{5256}{1 - p_B}
\]

\[
\hat{p}_B = \frac{5416 + 5256}{5416 + 5256} = 0.507
\]

2.6.4 Likelihood for the Sex Conditional Model: NLSY Data

While the EDA (in section 2.6.2) results are useful, it is not possible to answer our first research question straightaway, so we consider another model.

Consider the sex composition of every family with one, two or three children the NLSY data. As in previous applications, the contribution to the Sex Conditional likelihood for each family composition can be written out. (See Table 2.8 for examples.) Hence, we write the likelihood as a function of the three parameters \( p_{B|N} \), \( p_{B|bias} \), and \( p_{B|Gbias} \).

\[
\text{lik}(p_{B|N}, p_{B|bias}, p_{B|Gbias}) = p_{B|N}^{930}(1 - p_{B|N})^{951}(p_{B|N}p_{B|bias})^{582}(p_{B|N}(1 - p_{B|bias}))^{666}
\]

\[
\cdots((1 - p_{B|N})(1 - p_{B|Gbias})(1 - p_{B|Gbias}))^{159}
\]

\[
= p_{B|N}^{3131}(1 - p_{B|N})^{3119}p_{B|bias}^{1131}(1 - p_{B|bias})^{1164}p_{B|Gbias}^{1124}(1 - p_{B|Gbias})^{973}
\]

\[
\log\text{-lik}(p_{B|N}, p_{B|bias}, p_{B|Gbias}) = 3161\log p_{B|N} + 3119\log(1 - p_{B|N}) +
\]

\[
1131\log(p_{B|bias}) + 1164\log(1 - p_{B|bias}) +
\]

\[
1124\log p_{B|Gbias} + 973\log(1 - p_{B|Gbias})
\]

To use this likelihood to estimate the probability of a boy entering a neutral setting, \( p_{B|N} \), we begin with the logarithm of the likelihood in equation 2.7. Differentiating the log-likelihood with respect to \( p_{B|N} \) holding all other parameters constant yields a result much like our first model. That is,

\[
\hat{p}_{B|N} = \frac{3136}{3136 + 3119} = 0.5014
\]

We could think of this as the estimated probability of a boy joining a family where the number of boys and girls are equal including when there are no children in a family. Our estimate is sensible.

Similarly, MLEs for \( p_{B|bias} \) and \( p_{B|Gbias} \) can be obtained:

\[
\hat{p}_{B|bias} = \frac{1131}{1131 + 1164} = 0.4928
\]

\[
\hat{p}_{B|Gbias} = \frac{1124}{1124 + 973} = 0.5360
\]
CASE STUDY: ANALYSIS OF THE NLSY DATA

<table>
<thead>
<tr>
<th>Family Composition</th>
<th>Number of families</th>
<th>Sex Unconditional</th>
<th>Sex Conditional Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>930</td>
<td>$p_B$</td>
<td>$P_B</td>
</tr>
<tr>
<td>G</td>
<td>951</td>
<td>$1 - p_B$</td>
<td>$1 - P_B</td>
</tr>
<tr>
<td>BB</td>
<td>582</td>
<td>$p_Bp_B$</td>
<td>$P_B</td>
</tr>
<tr>
<td>BG</td>
<td>666</td>
<td>$p_B(1 - p_B)$</td>
<td>$P_B</td>
</tr>
<tr>
<td>GB</td>
<td>666</td>
<td>$(1 - p_B)p_B$</td>
<td>$P_B</td>
</tr>
<tr>
<td>GG</td>
<td>530</td>
<td>$(1 - p_B)(1 - p_B)$</td>
<td>$P_B</td>
</tr>
<tr>
<td>BBB</td>
<td>186</td>
<td>$p_Bp_Bp_B$</td>
<td>$P_B</td>
</tr>
<tr>
<td>BBG</td>
<td>177</td>
<td>$p_Bp_B(1 - p_B)$</td>
<td>$P_B</td>
</tr>
<tr>
<td>BGG</td>
<td>173</td>
<td>$p_B(1 - p_B)(1 - p_B)$</td>
<td>$P_B</td>
</tr>
<tr>
<td>BGB</td>
<td>148</td>
<td>$p_Bp_Bp_B$</td>
<td>$P_B</td>
</tr>
<tr>
<td>GBB</td>
<td>151</td>
<td>$1 - p_Bp_Bp_B$</td>
<td>$P_B</td>
</tr>
<tr>
<td>GGB</td>
<td>125</td>
<td>$(1 - p_B)(1 - p_B)p_B$</td>
<td>$P_B</td>
</tr>
<tr>
<td>GBB</td>
<td>182</td>
<td>$(1 - p_B)(1 - p + B)p_B$</td>
<td>$P_B</td>
</tr>
<tr>
<td>GGG</td>
<td>159</td>
<td>$(1 - p_B)(1 - p + B)(1 - p_B)$</td>
<td>$P_B</td>
</tr>
<tr>
<td>log-likelihood</td>
<td></td>
<td>-7396.067</td>
<td>-7374.238</td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td>14794.13</td>
<td>14751.48</td>
</tr>
<tr>
<td>BIC</td>
<td></td>
<td>14810.68</td>
<td>14749.18</td>
</tr>
</tbody>
</table>

Table 2.8: Composition of and number of families, and Stopping Model Parameters.

Are these results consistent with the notion that boys or girls run in families? Our investigation into “sex runs in families” through the Sex Conditional Model hypothesized a higher probability of boys among children born into families with a boy bias. However, we found that if there is a boy bias the probability of a subsequent boy was estimated to be 0.493. Similarly, girls join families with more girls than boys approximately 46.4% of the time so that there is little support for the idea that “girls run in families.” First let’s see if the Sex Conditional Model is significantly better than a simple random one. If it is, our parameter estimates will not be as we expected for sex running in the family. It may be that indeed sex does run in the family but there are other factors at work that alter the parameter estimates. Is the notion of sex running with families fallacious or could there possibly be another explanation? The heart of the matter is either the Sex Conditional Model is wrong and the data confirmed that or there is possibly another model that does not contradict the “sex runs in families” notion but includes other information that may explain this result. What other factors may be relevant? Couples may be stop when they get a boy (or a girl) in which case the proportion of boys may be changed (in what way?).

# Model 2: Sex Conditional Model (small 3 family data set)
# Assumes probability of having a boy depends on whether you’ve had
# boys previously
pbb=1/2    # fix parameters at MLEs
pbg=1/2
pbn=1/4
max.L.m2.small = pbn^1 * (1-pbn)^3 * pbb^1 * (1-pbb)^1 * pbg^1 * (1-pbg)^1
max.logL.m2.small = 1*log(pbn)+3*log(1-pbn) + 1*log(pbb)+1*log(1-pbb) + 1*log(pbg)+1*log(1-pbg)
2.6.5 Comparing the Sex Unconditional to the Sex Conditional Model

Does the Sex Conditional Model provide a statistically significant improvement over the Random Model? The Sex Conditional Model provides support of a genetic explanation of the variation in family composition.

Likelihoods are not only useful for fitting models but they are also useful when comparing models. If the predictors for one model are a subset of predictors for a larger model, we say the models are nested and the difference between their likelihoods can be incorporated into a statistical test to help judge the benefit of including additional predictors. If the predictors are not nested, the likelihood can still be useful but will take a different form. The Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are functions of the likelihood that can be used to compare models particularly when we are comparing different models where predictors are not nested. Either way we see that this notion of likelihood is pretty useful.

2.6.6 Nested Models

There are systematic ways in which likelihoods can be used to compare two models. We’ll begin with the case where the models are nested. When one model is nested within another, it can be thought of as a reduced version of the more complex model. For example, one way to think of this is to consider whether parameters in the more complex model can be equated to obtain the simpler model or whether some parameters in the complex model can be equated to constants. We certainly have that case here. If \( p_{B|bias} = p_{B|N} = p_{B|Gbias} = p_B \) in the Sex Conditional Model, we have the simpler, coin-flipping model.

We start with the idea of comparing the likelihoods of each model at their maxima. To do so, we write out likelihoods, find the MLEs, and replace the parameters in the likelihood with their MLEs thereby finding the maximum value for the likelihood. Here we will refer to the first model, the Sex Unconditional Model, as the reduced model noting that it has only a single parameter, \( p_B \). The more complex model, the Sex Conditional Model, has three parameters and is unsurprisingly referred to as the larger model. Recall the likelihoods and log-likelihoods derived earlier. (Equations 2.7). Next as we did earlier in the chapter, we use the log-likelihoods to determine the MLEs for each model.

**Hypotheses**

\( H_0 : p_{B|N} = p_{B|bias} = p_{B|Gbias} = p_B \)

The probability of a boy does not depend on the prior family composition.

\( H_a : \) At least one parameter from \( p_{B|N}, p_{B|bias}, p_{B|Gbias} \neq 0 \)

The probability of a boy does depend on the prior family composition.

**The reduced model: the Sex Unconditional Model**

\[
L(p_B) = p_B^{5416}(1-p_B)^{5256}
\]

\[
\log L(p_B) = 5416 \log p_B + 5256 \log (1-p_B)
\]

For the reduced model,

\[
\hat{p}_B = \frac{5416}{5416 + 5256} = .5075
\]

The maximum of the log-likelihood for the reduced model can be found by replacing \( p_B \) in the
CASE STUDY: ANALYSIS OF THE NLSY DATA

log-likelihood with the MLE of $p_B$.

\[ \text{log-lik}(\hat{p}_B) = 5416 \times \log(0.5075) + 5256 \times \log(1 - 0.5075) \]
\[ = -7396.067 \]

The larger* model: the Sex Conditional Model

\[ \text{log-lik}(p_{B|N}, p_{B|B\text{bias}}, p_{B|G\text{bias}}) = 3136 \log p_{B|N} + 3119 \log(1 - p_{B|N}) + 1131 p_{B|B} + 1164(1 - p_{B|B}) \]
\[ + 1124 \log p_{B|G} + 973 \log(1 - p_{B|G}) \]

\[ \hat{p}_{B|N} = 0.5013, \quad \hat{p}_{B|B} = 0.4928, \quad \hat{p}_{B|G} = 0.5360 \]

The maximum of the log-likelihood for the larger model can be found by replacing $p_{B|N}, p_{B|B\text{bias}}, p_{B|G\text{bias}}$ in the log-likelihood with 0.501, 0.493, and 0.536, respectively.

\[ \text{log-lik}(0.501, 0.493, 0.536) = -7374.238 \]

*We use the term ‘larger’ for the model with more parameters but in the literature it is more often referred to as the ‘full’ model. We chose to use ‘larger’ here to reduce confusion. Sometimes students misinterpret the ‘full’ model as the model with all of the parameters. In some settings, the ‘full’ model is intended to refer to larger model not the one with all of the parameters in it.

Take a look at the likelihoods - the maximum likelihood for the larger model is indeed larger. The maximum likelihood for the larger model is guaranteed to be at least as large as the maximum likelihood for the reduced model, so we’ll be interested in whether this (always nonnegative) difference in maximum likelihoods is significant.

\[ \max(\text{log-lik}(\text{larger model})) - \max(\text{log-lik}(\text{reduced model})) \]

But what should we compare this difference to?

To judge significance, look at twice the difference of the maximum log-likelihoods. A difference of the maximum log-likelihoods can also be looked at as the log of the ratio of the likelihoods and for that reason the test is referred to as the Likelihood Ratio Test (LRT). Distribution theory says that twice the log of the ratio of the maximum likelihoods will follow a $\chi^2$ distribution with the degrees of freedom equal to the difference in the number of parameters between the two models when the reduced model is the true model.

Our test statistic is

\[ 2[\text{log-lik}(\text{larger model}) - \text{log-lik}(\text{reduced model})] \]
\[ = 2 \log \left( \frac{\text{Lik}(\text{larger model})}{\text{Lik}(\text{reduced model})} \right) \]
\[ = -2 \times (-7374.238 + 7396.073) = 43.67(p < 0.0001) \]

Intuitively, when the likelihood for the larger model is much greater than it is for the reduced
model, we have evidence that the larger model is more closely aligned with the observed data.

This isn’t really a fair comparison on the face of it. We need to account for the fact that more parameters were estimated and used for the larger model. That is accomplished by the degrees of freedom for the $\chi^2$ distribution. The expected value of the $\chi^2$ distribution is its degrees of freedom. Thus when the difference in the number of parameters is large, the test statistic will need to be much larger to convince us that it is not simply chance variation with two identical models.

Note: You may notice that the LRT is similar in spirit to the extra-sum-of-squares F-test used in linear regression. Recall that the extra-sum-of-squares F-test involves comparing two nested models. When the smaller model is true, the F-ratio follows an F-distribution which on average is 1.0. A large, unusual F-ratio provides evidence that the larger model provides a significant improvement.

```
# Model comparisons - Model 1 vs. Model 2
lrt = 2*(max.logL.m2.small-max.logL.m1.small)
lrt # likelihood ratio test statistic
1-pchisq(lrt,df=2) # p-value for testing Ho: no diff between Models 1&2
```

The Sex Conditional Model provides a significant improvement over the Random Model lending credibility to the ‘sex runs in families’ hypothesis. This provides evidence that the gender composition of families may be in part due to genetics. However our point estimates for a probability of a boy were not what we had expected for ‘sex runs in families.’ When the gender composition of a family is biased, say having more boys than girls, we expected the probability the next child is a boy would be greater, but we found that the estimated probability of a boy in this circumstance to be 0.4928. It may be that this discrepancy stems from behavioral aspects of family formation. There are at least two forms of couples’ behavior that may affect the relative proportions of family compositions in the data. They are:

1. Couples may stop when they reach their desired number of children regardless of the family composition.
2. Couples may be waiting for either (1) a boy or (2) a girl or (3) balance, that is a boy and a girl.

2.7 Model 3: Stopping Rule Models

Rodgers and Doughty (2001) offer one reason to explain the contradictory results: waiting for a male child. It has been noted by demographers that some parents are only interested in producing a boy so that the appearance of a boy in the family leads more often to the end of childbearing. This would affect the sex ratios. In what way? Fewer runs of boys? There are several different models one could conceive of (no pun intended) for examining this phenomenon.

2.7.1 Ideal Family Size Stopping Model

Do parents have an ideal family size after which they stop having children? This stopping rule is not looking at particular gender compositions, but simply counting the number of children.

2.7.2 Waiting for a Boy or a Girl or a Balance Preference Model.

Do parents ‘try’ for a boy? Are they more likely to stop childbearing once they produce a boy? Or are some parents waiting for a girl? Others might wish to have at least one boy and girl. The
exploratory data analysis results in Table 2.7 provides some insight, but cannot definitively settle the question about couples’ stopping once they have a boy.

Suppose a population’s reproduction follows the Sex Conditional Model, suppose ‘sex runs in families.’ However if the couples in this population also adhere to a Waiting for a Boy stopping rule, the gender compositions expected under the Sex Conditional Model may be perturbed. For this reason, we consider models with stopping rules in addition to the Sex Unconditional or Sex Conditional models.

Consider a ‘Waiting for a Boy or a Girl’ stopping rule model, for illustration. For each child, ask if there are more children in the family after that child or if the couple has stopped (S) having kids. If the family is stopped at that child, is the child the only boy in the family (OB) or the only girl (OG)? To keep things simple, we do not incorporate balance preference nor preferred family size at this stage, although it is possible to do so.

2.7.3 Parameters for Stopping Rule Models

**Parameters for the First Male or First Female Stopping Model**

- \( p_{S|OB} \) probability of stopping after a child who is the only boy up to that point
- \( 1 - p_{S|OB} \) probability of not stopping after a child who is the only boy up to that point
- \( p_{S|OG} \) probability of stopping after a child who is the only girl up to that point
- \( 1 - p_{S|OG} \) probability of not stopping after a child who is the only girl up to that point
- \( p_{S|NO} \) probability of stopping after a child who is neither an only boy or only girl
- \( 1 - p_{S|NO} \) probability of not stopping after a child who is neither an only boy or only girl

**Parameters for the Ideal Family Stopping Model**

- \( p_{S|C1} \) probability that no more children follow the first child
- \( 1 - p_{S|C1} \) probability that more children follow the first child
- \( p_{S|C2} \) probability that no more children follow the second child
- \( 1 - p_{S|C2} \) probability that more children follow the second child
- \( p_{S|C3} \) probability that no more children follow the third child

The Ideal Family parameterization introduces two more parameters; \( p_{S|C1} \) and \( p_{S|C2} \). Note that for our dataset \( p_{S|C3}=1 \). In addition to the number of parameters for the Sex Unconditional or Conditional Models whereas the First Male Stopping parameterization introduces one more parameter; \( p_{S|OB} \).

2.7.4 Data for Stopping Models

For this analysis, we use the data from Table 2.9 which consists of gender composition information on completed one, two or three child families along with the corresponding model parameters.

2.7.5 Analyzing Models which incorporate Stopping Rules

Let’s construct a model using our First Male or First Female stopping rule model with the Sex Conditional Model to investigate whether the expected “sex runs in families” estimates are altered because of couples “waiting for a boy or girl” stopping rule. Table 2.9 contains the probabilities for a family compositions under the First Male or First Female Stopping Rule Model. For example, the family BGB, the last child is neither an only boy nor only girl. These 148 families did not appear to be motivated by waiting for one gender or another. The 177 families with BBG compositions may be families that were rewarded for waiting for a girl. To include the portions of the likelihood that
involve stopping, see the tallies for each child corresponding to their stopping status in Table 2.9. Note that each child whether they be first, second or last, has a corresponding probability related to their stopping status. For each child, the tally takes into account whether the child is the only boy up to that point. In the exploratory data analysis results in Table 2.10, a couple stops at an only boy 43.2% of the time and at an only girl 45.7% of the time. This difference does not provide support for the notion that couples wait for and stop once a boy arrives.

The likelihood for the First Boy or First Girl Stopping Rule portion of our likelihood looks like:

$$P_{S|OB}^{1721}(1 - P_{S|OB})^{2265} \cdot P_{S|OG}^{1794}(1 - P_{S|OG})^{2134} \cdot P_{S|NO}^{2111}(1 - P_{S|NO})^{647}$$

After finding a derivative, set equal to 0, and solve for the MLEs:

$$\hat{P}_{S|OB} = \frac{1721}{1721 + 2265} = 0.432$$

$$\hat{P}_{S|OG} = \frac{1794}{1794 + 2134} = 0.457$$

<table>
<thead>
<tr>
<th>Family Composition</th>
<th>Number of families</th>
<th>Ideal Family Size</th>
<th>Up to that point is the child the only boy or only girl?</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>930</td>
<td>(P_{S</td>
<td>C1})</td>
</tr>
<tr>
<td>G</td>
<td>951</td>
<td>(P_{S</td>
<td>C1})</td>
</tr>
<tr>
<td>BB</td>
<td>582</td>
<td>(1 - P_{S</td>
<td>C1})P_{S</td>
</tr>
<tr>
<td>BG</td>
<td>666</td>
<td>(1 - P_{S</td>
<td>C1})P_{S</td>
</tr>
<tr>
<td>GB</td>
<td>666</td>
<td>(1 - P_{S</td>
<td>C1})P_{S</td>
</tr>
<tr>
<td>GG</td>
<td>530</td>
<td>(1 - P_{S</td>
<td>C1})P_{S</td>
</tr>
<tr>
<td>BBB</td>
<td>186</td>
<td>(1 - P_{S</td>
<td>C1})(1 - (P_{S</td>
</tr>
<tr>
<td>BBG</td>
<td>177</td>
<td>(1 - P_{S</td>
<td>C1})(1 - (P_{S</td>
</tr>
<tr>
<td>BGG</td>
<td>173</td>
<td>(1 - P_{S</td>
<td>C1})(1 - (P_{S</td>
</tr>
<tr>
<td>BGB</td>
<td>148</td>
<td>(1 - P_{S</td>
<td>C1})(1 - (P_{S</td>
</tr>
<tr>
<td>GBB</td>
<td>151</td>
<td>(1 - P_{S</td>
<td>C1})(1 - (P_{S</td>
</tr>
<tr>
<td>GGB</td>
<td>125</td>
<td>(1 - P_{S</td>
<td>C1})(1 - (P_{S</td>
</tr>
<tr>
<td>GBG</td>
<td>182</td>
<td>(1 - P_{S</td>
<td>C1})(1 - (P_{S</td>
</tr>
<tr>
<td>GGG</td>
<td>159</td>
<td>(1 - P_{S</td>
<td>C1})(1 - (P_{S</td>
</tr>
</tbody>
</table>

Table 2.9: Composition of and number of families and Stopping Rule Parameters.

<table>
<thead>
<tr>
<th>Child is...</th>
<th>total children</th>
<th>prop of all children</th>
<th>n.stops (n.families)</th>
<th>prop stopped after these children</th>
</tr>
</thead>
<tbody>
<tr>
<td>a boy who is the only boy in the family up to that point</td>
<td>3,986</td>
<td>37.4%</td>
<td>1,721</td>
<td>0.432</td>
</tr>
<tr>
<td>a girl who is the only girl in the family up to that point</td>
<td>3,928</td>
<td>36.8%</td>
<td>1,794</td>
<td>0.457</td>
</tr>
<tr>
<td>neither an only boy nor only girl</td>
<td>2,758</td>
<td>25.8%</td>
<td>2,111</td>
<td>0.765</td>
</tr>
<tr>
<td></td>
<td>10,672</td>
<td>100%</td>
<td>5,626</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.10: What percentage of all births are children who are the only boy or only girl up to that point.
\[ \hat{p}_{S|NO} = \frac{2111}{2111 + 647} = 0.765 \]

After replacing \( p_{S|OB}; p_{S|OG}; p_{S|NO} \) with their MLEs and logging the likelihood we obtain the maximum log-likelihood, -7565.558. The estimated probability of stopping in these three different circumstances do not provide any compelling evidence to suggest that the probability of stopping differs markedly for first boys, first girls or neither. The MLEs suggest that of the 3,986 first boys, parents stop at about 43% of the time and of the 3,928 first girls childbearing stopped at about 46% of the time. Stopping occurred for about over 76% of the remaining 2,758 children who are neither first girls nor first boys. The astute reader might notice that the MLEs are equal to the proportion in Table 2.10. Nonetheless, there are some subtleties to these arguments and it might be a good idea to look for evidence of stopping for a balance of boy and girl, i.e. “sex balance” behavior or define alternative models.

We have only worked with the portion of the likelihood involving the waiting stopping parameters \( p_{S|OB}; p_{S|OG}; \) and \( p_{S|NO} \). These results ignore that an increasing number of children may lead to an increased likelihood of stopping regardless of the sex or whether the parents are hoping for a boy and a girl. Models with parameters representing ideal family size, \( p_{SC1}; p_{SC2}; p_{SC3} \), can be considered for inclusion.

\[
\begin{align*}
\hat{p}_{SC1} &\left(1 - p_{SC1}\right)^{3745} \\
\hat{p}_{SC2} &\left(1 - p_{SC2}\right)^{1301} \\
\hat{p}_{SC3} &\left(1 - p_{SC3}\right)^{1881}
\end{align*}
\]

You can verify that the MLEs are \( \hat{p}_{SC1} = 0.334, \hat{p}_{SC2} = 0.65 \) and \( \hat{p}_{SC3} = 1 \). For our dataset the probability of stopping once you reach the third child is one. Therefore, we really only have two parameters to add with this stopping model.

**2.7.6 Non-nested Models**

How does the waiting for a boy or girl model compare to the ideal family model? Thus far we’ve seen how nested models can be compared. But these two models are not nested. Two measures referred to as information criteria are useful when comparing non-nested models. For each, a small value suggests a superior model. You can find the log-likelihood on output from most modeling software packages. A big idea in this chapter is that a (relatively) large likelihood implies your data is likely to be produced. Two parameters, two likelihoods, we select the parameter value associated with the larger likelihood. Log-likelihoods are typically negative, so multiplying by -2 will produce positive quantities where the smaller value is associated with the better model. So both of our information criteria measures have the same first term, -2log-likelihood and for a good model this term would be small.

- **AIC = -2*log-likelihood + 2*variables**
  where variables represents the number of parameters in the fitted model. AIC stands for Akaike Information Criterion. Because smaller AICs imply better models, we can think of the second term as a penalty, the more variables we use the larger the AIC, specifically the penalty is twice the number of variables.

- **BIC= -2*log-likelihood + variables*log(nobs)**
  where variables is the number of parameters and nobs is the number of observations. BIC stands for Bayesian Information Criterion(BIC). It is also known as Schwarz’s Bayesian criterion (SBC). Here we see that the penalty for the BIC differs from the AIC. The log of the number of observations heavily penalizes models built with large datasets. The variables have less influence with BIC than with the AIC.

When assessing model performance, AIC and BIC will penalize you for too many variables whereas BIC also includes the number of observations in the penalty on a log scale. Which explanation of the data seems more plausible - sex running in families or couples stopping more often
following a male birth? The Sex Conditional Model and the Waiting for a Boy or Girl Model could be compared to consider this question, however these models are not nested. They both involve three parameters none of which can be set to constants or equal to one another to obtain the other model. One way models are compared in this setting is to calculate and compare information criteria.

<table>
<thead>
<tr>
<th></th>
<th>Sex Conditional</th>
<th>Waiting for a Boy or Girl</th>
</tr>
</thead>
<tbody>
<tr>
<td>log-likelihood</td>
<td>-7374</td>
<td>-7565</td>
</tr>
<tr>
<td>AIC</td>
<td>14754.47</td>
<td>15137.12</td>
</tr>
<tr>
<td>BIC</td>
<td>14776.3</td>
<td>15158.94</td>
</tr>
</tbody>
</table>

Smaller BIC are preferred so here the Sex Conditional Model is judged superior to Waiting for a Boy or Girl Model suggesting that the Sex Conditional Model is a better explanation of the data than our First Male Stopping Rule model.

With likelihood methods, we are no longer restricted to independent, identically distributed normal responses (iidN). Likelihood methods can accommodate non-normal responses and correlated data. Strictly speaking, we do not need likelihood methods for every model in this text, nevertheless they are needed for enough models here so that it is worth your time and effort to understand them. Models that in the past you would fit using ordinary least squares can also be fit using the principle of maximum likelihood. It is pleasing to discover that under the right assumptions the maximum likelihood estimates (MLEs) for the intercept $a$, and slope $b$ are identical to ordinary least squares estimators (OLS) despite the fact that they are obtained in quite different ways.

Beyond the intuitively appealing aspects of MLEs, they also have some very desirable statistical properties. You learn more about these features in a statistical theory course. Here we briefly summarize the highlights. MLEs are consistent, i.e. MLEs converge in probability to the parameter being estimated as the sample size increases. MLEs are asymptotically normal, as the sample size increases, the distribution of MLEs look more normal. MLEs are efficient because no consistent estimator has a lower mean squared error. Of all the estimators that produce estimates that get closer to the true parameter value as the sample size increases, no estimator will have a smaller mean square errors than the MLE.

Maximum likelihood estimation is not a panacea, however. There are circumstances where MLEs do not exist, where likelihoods cannot be written down, or where MLEs cannot be written explicitly.
2.8 Likelihoods and this Course

Rodgers and Doughty (2001) noted that many factors have been identified that can potentially affect the human sex ratio at birth. A 1972 paper by Michael Teitelbaum accounted for around 30 such influences, including drinking water, coital rates, parental age, parental socioeconomic status, birth order, and even some societal-level influences like wars and environmental pathogens.

This chapter on likelihood ignored these complicating factors and was intentionally kept simple to impress you with the fact that likelihoods are conceptually straightforward. Likelihoods answer the sensible question of how likely you are to see your data in different settings. When the likelihood is simple as in this chapter, you can roughly determine an MLE by looking at a graph or be a little more precise by using a little calculus or most conveniently using software. As we progress throughout the course, the likelihoods will become more complex and may require numerical methods to obtain MLEs, yet the concept of an MLE will remain the same and its use will turn up all through the course. Likelihoods will show up in parameter estimation, model performance assessment and comparisons.

One if the reasons many of the likelihoods will become complex is because of covariates. Here we estimated probabilities of having a boy in different settings, but we did not use any specific information other than sex composition about families. The problems in the remainder of the book will typically employ covariates. For example, suppose we had information of paternal age for each family. Consider the Sex Unconditional Model, let $p_B = \beta_0 + \beta_1 (\text{paternal age})$. Then the likelihood would not simply be a function of $p_B$, rather it would be:

$$\text{Lik} (\beta_0, \beta_1) = (\beta_0 + \beta_1 (\text{paternal age}))^{n_{\text{boys}}} (1 - \beta_0 - \beta_1 (\text{paternal age}))^{n_{\text{girls}}}$$

So our focus might then be on how the covariate, paternal age, is associated with the probability of having a boy. Consequently, we must estimate $\beta_0$ and $\beta_1$. Here we see the benefit of using likelihoods - the linear models you have used previously would not be helpful for many of these more complicated models.

Another compelling reason for likelihoods occurs when we encounter correlated data. For example, models with conditional probabilities do not conform to the independence assumption. The Sex Conditional Model is an example of such a model. We’ll see that likelihoods can be useful when the data has structure such as multilevel that induces a correlation. A good portion of the book addresses this.

When the responses are are not normal such as in generalized linear models where we see binary responses and responses which are counts we’ll find it difficult to use the linear models of the past and we find the flexibility of likelihood methods to be extremely useful. Likelihood methods will significantly broaden your statistical horizons!

2.9 R code notes

```r
# Looking at Likelihoods
#
# Estimating the probability of a boy, p
# using the n.boys and n.girls born
# in a random sample
#
# Assumes the gender for each birth is
# independent of other births
```
# Intuitively $\hat{p} = \frac{n.\text{boys}}{n.\text{boys} + n.\text{girls}}$
# We will provide rationale for that intuition

$p = \text{seq}(0,1,.05)$  # possible values for $p$

par(mfrow=c(2,2))  # looking at graphs for diff sample sizes

# Defining a function that plots log-likelihoods

log-lik.plot = function(n.boys,n.girls){
  Lik = p^n.boys*(1-p)^n.girls
  plot(p,Lik,type="l",main=(paste("Likelihood for n.boys", n.boys)))
  log-lik = log(Lik)
  plot(p,log-lik,type="l",main=(paste("log-lik for n.boys", n.boys)))
  #main=expression(paste("log-likelihood for n.boys" n.boys) )
  
}

postscript("Likelihood.eps")
par(mfrow=c(2,2))  # looking at graphs for diff sample sizes
log-lik.plot(20,30)
log-lik.plot(400,600)
dev.off()

R has a \texttt{bic} function. If you fit a model, \texttt{bic(model.fit)}
will calculated the BIC for you.

# Computing Max Likelihoods for NLSY data

# Model 1: Sex Unconditional
# MLE

pb=0.507

max.Lik.m1=pb^5416*(1-pb)^5256
max.Lik.m1

max.log.Lik.m1=5416*log(pb)+5256*log(1-pb)
max.log.Lik.m1

aic1= -2*max.log.Lik.m1+2*1
aic1
bic1=-2*max.log.Lik.m1 +log(10572)*1
bic1

# Model 2: Sex Conditional
# ML\texttt{Es}

pbn=0.5014
pbb=0.4928
pbg=0.5360

max.Lik.m2=pbn^3136*(1-pbn)^3119*pbb^1131*(1-pbb)^1164*pbg^1124*(1-pbg)^973
max.Lik.m2

max.log.Lik.m2=3136*log(pbn)+3119*log(1-pbn)+1131*log(pbb)+1164*log(1-pbb)+1124*log(pbg)+973*log(1-pbg)
R CODE NOTES

max.log.Lik.m2

-2*(max.log.Lik.m1-max.log.Lik.m2)

aic2= -2*max.log.Lik.m2+2*3
aic2
bic2=-2*max.log.Lik.m2+log(10572)*3
bic2

# Model 1: Sex Unconditional Model
# Assumes the gender for each birth is independent of other births

## Data for one family: BGG
### possible values of p_B and lik(p_B)
### as in Table 2.2 Some parameter values and likelihoods

# Evaluate courser grid of possible values of pb
pb=c(.1,.3,.5,.67,.8,.9) # possible values for prob a boy is born
lik=pb^2 *(1-pb) # likelihood of getting observed data
cbind(pb,lik) # print table of results
max(lik) # maximum likelihood over 6 values of pb
pb[lik==max(lik)] # value of pb where likelihood maximized
plot(pb,lik,xlab="possible values of p_B", ylab="Likelihood",type="l", main="Courser grid for BBG")
abline(v=pb[lik==max(lik)],col="red")

# Evaluate finer grid of possible values of pb
pb=seq(0,1,length=1001) # possible values for prob a boy is born
lik=pb^2 *(1-pb) # likelihood of getting observed data
max(lik) # maximum likelihood over 1001 values of pb
pb[lik==max(lik)] # value of pb where likelihood maximized
plot(pb,lik,xlab="possible values of pb", ylab="Likelihood",type="l", main="Denser grid for BBG")
abline(v=pb[lik==max(lik)],col="red")

## Stopping Rule Models: Ideal Family Size

#MLEs
pc1=1881/(1881+3745) # 0.334
pc2=2444/(2444+1301) # 0.653
pc3=1.0

max.log.Lik.size=1881*log(pc1)+3745*log(1-pc1)+2444*log(pc2)+1301*log(1-pc2)+1301*log(pc3)
max.log.Lik.size

aic.size= -2*max.log.Lik.size+2*2
aic.size
bic.size= -2*max.log.Lik.size+log(10572)*2
bic.size

## Stopping Rule Models: Waiting for a Boy or a Girl

#MLEs
46BEYOND MOST LEAST SQUARES: USING LIKELIHOODS TO FIT AND COMPARE MODELS

pob=0.432
pog=0.457
pno=0.765

max.log.Lik.wait=1721*log(pob)+2265*log(1-pob)+1794*log(pog)+2134*log(1-pog)
+2111*log(pno)+2758*log(1-pno)

max.log.Lik.wait

aic.wait= -2*max.log.Lik.wait+2*3
aic.wait

bic.wait=-2*max.log.Lik.wait +log(10572)*3
bic.wait

> # Comparing Models 2 and 3
> # note that these models are not nested

> bic.2 = -2*logL_2 + 3*log(n.families)
> bic.2
[1] 14992.48
>

# with random model, all children are independent

total.boys = rbinom(1000,10762,0.5)
total.girls = 10672-total.boys
pB=total.boys/10672
loglik=total.boys*log(pB)+total.girls*log(1-pB)
obs.loglik=5416*log(0.507) +5256*log(1-0.507)
postscript("simRandom.eps")
  hist(loglik, main="Simulated Log-likelihoods for the Random Model")
  abline(v=obs.loglik,col='red')
  dev.off()

2.10 Exercises

2.10.1 Conceptual

1. Suppose we plan to use data to estimate one parameter.
   (a) When using a likelihood to obtain an estimate for the parameter, which is preferred a large or a small likelihood value? Why?
   (b) The height of a likelihood curve is the probability of the data for the given parameter.
   (c) The horizontal axis represents different possible parameter values.

2. Suppose the families with an “only child” were excluded for the Sex Conditional Model. How might the estimates for the three parameters be affected? Would it still be possible to perform a Likelihood Ratio Test to compare the Sex Unconditional and Sex Conditional Models? Why or why not?
   (a) How would couples waiting for a boy change the estimated probability of a boy in the neutral, boy-biased, and girl-biased settings?
   (b) Come up with an alternative models to investigate whether “sex runs in families.”
   (c) What concerns do you have for this formulation of the Sex Conditional Model? Can you think of some alternatives? Would the existence of stopping behavior affect our results? How?
EXERCISES

3. Write out the likelihood for a model which assumes the probability of a girl equals the probability of a boy. Carry out a LRT to determine whether there is evidence that the two probabilities are not equal. Comment on the practical significance of this finding (there is not necessarily one correct answer).

4. **Four child families** Repeat the likelihood analysis including four children families. Show all of your work. How do your results differ from the one to three child families?

5. **Case 2** Case 1 included 50 children. Case 2 is a much larger study with 600 boys and 400 girls. Our goal is to find an MLE for $p_B$ for the independence model so we proceed as we had done in the smaller study in Case 1. Using the methods in Case 1, we can show that the MLE for $p_B$ in Case 2 is 0.60. Show your work when using the first derivative to solve for the MLE.

6. For each value of $p_B$ below,
   
   (a) Write out an expression for the likelihood of seeing our data (5,416 boys and 5,256 girls) if the true probability of a boy is:
      
      i. $p_B = 0.5$.
      ii. $p_B = 0.45$
      iii. $p_B = 0.55$
      iv. $p_B = 0.5075$

   (b) Compute the value of the log-likelihood for each of the values of $p_B$ above.

   (c) Which of these four possibilities, $p_B = 0.45$, $p_B = 0.5$, $p_B = 0.55$, or $p_B = 0.5075$ would be the best estimate of $p_B$ given what we observed (our data)?

7. For the **Sex Conditional Model**
   
   (a) Add three more rows to Table 2.3.

   (b) In our first example, we had one parameter, $p_B$, to estimate. How many parameters need to be estimated for this model?

   (c) For these reasons, Rodgers et al., introduce $t_1$, $t_2$, and $t_3$ to represent the probability of stopping after the first, second, and third child respectively. They also defined gender-specific stopping parameters, $b_1$ and $g_1$, $b_2$ and $g_2$, plus $b_3$ and $g_3$.

   (d) Compute the BIC for models 1 and 2 (the null and sex conditional models) and comment.

   (e) Can you provide any rationale for why Model 2 appears to be better than Model 3?

   (f) Compute the AIC for models 2 and 3. How do the results of the BIC analysis compare to your AIC results. Look at the formulas for the BIC and AIC and comment on what you found.

8. **More Stopping Rule Models**
   
   (a) Do you think that the Stopping Rule portion of the likelihood should be separate from the Random or Sex Conditional Model portion of the likelihood?

   (b) Can you come up with an alternative method for parameterize ‘waiting for a boy’?

   (c) Should waiting for a girl be included in the model?

9. **Still More Stopping Rule Models** Write out the likelihood for a model with stopping parameters $t_1$, $t_2$ and $t_3$ combined with the sex unconditional model. Find the MLEs for the parameters $t_1$, $t_2$ and $t_3$.

10. **Yet Again More Stopping Rule Models** Write out the likelihood for a model with stopping parameters $b_1$, $g_1$, $b_2$, $g_2$ and $b_3$, $g_3$ combined with the sex unconditional model. Find the MLEs for the parameters $b_1$, $g_1$, $b_2$, $g_2$, and $b_3$, $g_3$. 
48BEYOND MOST LEAST SQUARES: USING LIKELIHOODS TO FIT AND COMPARE MODELS

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex Unconditional</td>
<td>(p_B = 0.507)</td>
</tr>
<tr>
<td>Sex Conditional</td>
<td>(p_{B</td>
</tr>
</tbody>
</table>

Table 2.11: MLEs for model parameters.

11. **Balance-preference** Can you construct a model which suggests that childbearing is stopped when couples have a boy *and* a girl? Define the parameter(s) for balance-preference stopping combined with the sex conditional model and write out likelihood contributions for the same family compositions that appear in Table 2.3.

   **Extra Credit** Obtain the MLEs for your balance-preference model parameters and interpret the results.

   **Extra Extra Credit** Use the methods of the next sections to compare models with and without a balance-preference.
Chapter 3

Distribution Theory

To the Instructor

The material in this chapter can be treated as a chapter on its own say right after the likelihood chapter or as a reference to be covered “just in time” while making your way through the text. While it covers topics typically covered in a probability course, it is not at the same depth or mathematical rigor. Knowing about distributions is extremely useful for students of this text for several reasons. We highlight three:

- **With only the Normal distribution, students will not appreciate the fundamental ideas behind Generalized Linear Models and how it can broaden their modeling capabilities.** In general, a primary objective of this text is to introduce students to distributions other than the normal distribution. When students have been introduced to only iidNormal observations, their approaches to modeling will be severely limited given the vast array of data types regularly encountered.

- **Data from models in this text can be simulated to understand underlying theory of correlated and multilevel data.** For example, simulations can be used to illustrate the notion of overdispersion which in turn can be used to introduce correlated data. Simulations can be used to generate very simple mixture distributions. These distributions can be compared to the shape, mean and variances of data that that is not produced by a mixture. This exposes students to the idea of mixing parameters and later generalizes to multilevel models. Software such as R can be used to explore the characteristics of a distribution such as shape, location, and scale by changing parameter values with R.

- **Distribution theory can be used to simulate data from processes, especially complex processes, to anticipate what might be found when performing exploratory data analysis or modeling.** With a larger repertoire of distributions at their disposal, students can creatively build simulations that approximate familiar, complicated processes. Students construct a simulation that makes use of three or more distributions to model processes such as simple traffic flow or basic epidemiological models.

3.1 Learning Objectives

- Write definitions of two or more non-normal random variables in the context of an application.

- Identify possible values for each random variable.

- For a given distribution, identify how changing values for a parameter affects the characteristics of the distribution.

- Recognize a form of the pdf for each distribution.

- Identify the mean and variance for each distribution.

- Match the response for a study to a plausible random variable and provide reasons for ruling out other random variables.
• Match a histogram of sample data to plausible distributions.
• Come up with your own applications for each random variable.
• Create a mixture of distributions and compare the shape, mean, and variance to a similar distribution that is not a mixture.
• Create a simulation using several distributions providing rationale for each part.

3.2 Introduction
What if it is not plausible that a response is normally distributed? You may want to construct a model to predict whether a prospective student will enroll at a school or model the lifetimes of patients following a particular surgery. In the first case you have a binary response (enrolls (1) or does not enroll (0)) and in the second case you are likely to have very skewed data with many similar values and a few hardy souls with extremely long survival. These responses are not expected to be normally distributed, other distributions will be needed to describe and model binary or lifetime data. Non-normal responses responses are encountered in a large number of situations. Luckily, there are quite a few possibilities for models. In this appendix, we begin with some general definitions, terms and notation for different types of distributions with some examples of applications. We then create new random variables using combinations of random variables.

In our discussion of probability, a random variable is a rule that assigns values to outcomes of a chance experiment and a probability mass function or probability density function quantifies the probability or proportion of the time we expect the random variable to assign particular values for discrete and continuous random variables, respectively. For example, we could consider randomly selecting a women from the US population of women to be a chance experiment. There are an infinite number of random variables we could come up with for this “experiment.” An example of a random variable for this experiment is \( Y \) = the height of a randomly selected woman from this population. The distribution of heights provides the basis for the probability distribution of \( Y \). Often physical measurements like height and weight are approximately normally distributed. A normal distribution has two parameters that fully define it, its mean and standard deviation denoted by \( \mu \) and \( \sigma \), respectively. We see a familiar bell curve when examining a graph of the probability density function (pdf) for \( Y \).

\[
f(y) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}
\]

The mean and standard deviation of heights in the population of US women can be thought of as the mean and standard deviation of our random variable \( Y \)

3.2.1 Characteristics of Random Variables
Random variables can be characterized in the following ways.
• The set of values it takes on are Discrete or Continuous.
• The set of values it takes on are Bounded or Unbounded.
• The shape of the distribution is Skewed or Symmetric.

3.2.2 Location, scale and shape parameters
Location parameters provide information on an average of the values a random variable takes on. Scale parameters quantify how spread out the values of a probability distribution. The scale parameter for a normal distribution is denoted by \( \sigma \). Not surprisingly, shape parameters describe the shape of a distribution providing information such as how skewed or symmetric a distribution might
be. Most distributions are defined using a location and possibly a scale parameter. Alternatively some distributions are parameterized in terms of scale and shape parameters.

For our purposes, we will focus on a selection of random variables and their distributions encountered in three different settings.
1. Random Variables used for Modeling Responses
2. Random Variables referenced for Interval Estimation and Testing
3. Random Variables used to create Mixtures

It is worth noting that the few random variables in this appendix appear in many more settings than presented here. Furthermore, there are alternative descriptions and parameterizations of the distributions we consider. The scope of our view of distribution theory is limited. You can explore these ideas in much more detail if you take a course in probability theory.

While we have an extremely useful tool with the Gaussian distribution at our disposal, it is worthwhile filling our modeling tool with a number of other distributions. Throughout our discussion we reference the comprehensive compendium compiled by Van Hauwermeiren and Vose at A Compendium of Distributions at http://www.vosesoftware.com/content/ebook.pdf. You are also encouraged to reference Wikipedia which contains nice graphs and summaries for the distributions we consider.

3.3 Modeling Responses

You are undoubtedly most familiar with normally distributed responses. With the study of generalized linear models we will accommodate different types of responses. For each of the random variables we use to model responses, we explicitly describe the “chance experiment,” a generic definition, in words, of the random variable, the possible values the random variable takes on, its pdf and its parameters. We start out with random variables that assign discrete values. Some variables have a finite number of possible values, 0, 1, 2, 3, 4, 5 others an infinite such as 0, 1, 2, . . . . In some cases, we also state the mean and standard deviation.

3.3.1 Discrete Random Variables

With discrete random variables, the associated probabilities can be calculated for each possible value using a probability mass function.

3.3.1.1 Hypergeometric

**Chance experiment:** Randomly select $n$ items without replacement from $m$ successes and $N - m$ failures.

$Y =$ number of successes out of $n$ selected without replacement.

$$\text{Mean} = n \frac{m}{N}$$

Possible values: $0, 1, 2, \ldots n$ for $n \leq m$.

$$P(Y = y) = \binom{m}{y} \frac{\binom{N - m}{n - y}}{\binom{N}{n}}$$

**Example:** Suppose 12 jurors are randomly selected (without replacement) from a jury pool of 20 men and 15 women.
Let \( Y \) = number of women out of 12 selected.
Parameters: \( m=15 \) women, \( N=35 \), \( n=12 \).
Mean = \( 12 \times (15/35) = 5.14 \)
Possible values: 0, 1, \ldots , 12

What is the probability that exactly 6 women are selected?

\[
P(Y = 6) = \frac{\binom{15}{6} \cdot \binom{20}{6}}{\binom{35}{6}}
\]

which equals 0.16.

### 3.3.2 Bernoulli Process

Many random variables are based on observations from a Bernoulli process. A Bernoulli process consists of a series of independent trials that take on one of two values, say 0 or 1, often referred to as success or failure. The probability of success is identical for each trial.

#### 3.3.2.1 Binary Random Variable

**Chance experiment:** Perform a trial with two possible outcomes; success or failure.
\( Y = 0 \) if failure or \( Y = 1 \) for a success for a random trial.

**Parameter:** probability of success, \( p \)

**Mean** = \( p \)  
**Standard deviation** = \( \sqrt{p/(1 - p)} \)

**Possible values:** 0 or 1

\[
P(Y = y) = p^y (1 - p)^{1-y}
\]

**Example:** Your iPod shuffle randomly selects a song from among 2000 songs, 1500 of which are rock songs. A rock song is a success.
\( Y = 1 \) if a rock song is selected, 0 if the selection is not a rock song.

Parameter: \( p = 1500/2000 \) or 0.75

Mean = 0.75  
Standard deviation = 0.75 * 0.25 = 0.188

Possible values: 0 or 1

What is the probability that a rock song is not selected? 0.25.

#### 3.3.2.2 Binomial

**Chance experiment:** Perform \( n \) independent and identical Bernoulli trials.

\( Y \) = number of successes in \( n \) independent, identical trials.

**Parameter:** probability of success, \( p \) on a single trial

**Mean** = \( np \)  
**Standard deviation** = \( \sqrt{np(1 - p)} \)  
**Possible values:** 0, 1, \ldots , \( n \)

\[
P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}
\]

**Example:** Ten randomly selected emails are classified as spam or not spam. The chance that any randomly selected email is spam is 0.2

\( Y \) = the number of emails that are spam out of the ten emails.

Parameter: \( p = 0.2 \)

Mean = 10 * 0.2 = 2  
Standard deviation = 10 * 0.2 * 0.8 = 1.6
MODELING RESPONSES

Possible values: 0, 1, . . . , 10

What is the probability that exactly 4 of the emails are spam?

\[
P(Y = 4) = \binom{10}{4} 0.2^4 0.8^6
\]

which equals 0.09.

Figure 3.1: Some binomial distributions.

Figure 3.1 displays some binomial distributions.

Note: Observations of binary and binomial responses are based on a Bernoulli process resulting in a success (spam) or failure (not spam). The probability of spam is identical for each email. Emails should be independent of one another when emails coming in do not take previous emails into account. The probability of a spam email should remain the same from email to email.

3.3.2.3 Geometric

**Chance experiment:** Perform independent, identical Bernoulli trials up to and including the first success.

\[Y=\text{number of independent, identical trials up to and including the first success.}\]

**Parameter:** mean or probability of success, \(p\)

\[
\text{Mean} = \frac{1}{p} \quad \text{Standard deviation} = \sqrt{(1-p)/p^2}
\]

**Possible values:** 0, 1, 2, \(\infty\)

Note: An alternative parameterization is \(Y = \text{the number of trials until, but not including the first success.}\) The corresponding random variable has a different probability distribution function.

\[
P(Y = y) = (1-p)^{y-1} p
\]

**Example:** Randomly select emails from a process that produces 20% spam. Randomly select emails until a spam email is found.

\[Y = \text{the number of emails selected up to and including a spam email.}\]

Parameter: \(p = 0.20\)

Mean = 1/0.2 = 5 \quad \text{Standard deviation} = \sqrt{0.8}/0.04 \text{ or } 4.47

Possible values: 0, 1, 2, \ldots, \infty
What is the probability that exactly 7 emails are observed before a spam email appears?

\[ P(Y = 8) = 0.8^7 \cdot 0.2 \]

which equals 0.034.

Figure 3.2: Some geometric distributions.

3.3.2.4 Negative Binomial

**Chance experiment:** Perform independent, identical Bernoulli trials until \( k \) successes are observed. 

\( Y \) = the number of trials up to the \( k \)th success occurs.

**Parameters:** \( p \) = probability of a success, \( k \) the number of successes

**Mean** = \( k/p \) \hspace{1em} **Standard deviation** = \( \sqrt{pk/(1 - p)^2} \)

\[ P(Y = y) = \binom{y-1}{k-1} (1 - p)^{y-k} p^k \]

**Example:** Randomly select emails until six spam emails are selected from a process that produces 20% spam emails. 

\( Y \) = the number of emails selected up to and including the sixth spam mail 

Parameter: \( p = 0.2, \ k = 6 \)

Mean = \( \sqrt{0.2 \times 6 / 0.8} \) Standard deviation = \( \sqrt{0.2 \times 6 / 0.8^2} = 1.37 \)

Possible values: 0, 1, 2, …, \( \infty \)

What is the probability that exactly 10 emails including the 6 spam emails are selected before 6 spam emails are drawn?

\[ P(Y = 10) = \binom{15}{5} 0.8^9 0.2^6 \]

which equals 0.021.

3.3.3 Poisson Process

Many random variables are based on a Poisson process. A time period or space can be broken into very small, equal size units. The probability of an event within a given unit is very small and
independent of other units. For example, the probability of an animal bite during a very small time period is rare. A Poisson model implies that the probability of a bite occurring during another very small, equal length time period is the same and independent of a bite occurring within another time period of equal length.

3.3.3.1 Poisson Random Variable

**Chance experiment:** Randomly select a specific amount of time or space.

**Y =** count of items for a specified time or space.

**Parameter:** mean count per unit time or space, \( \lambda \)

**Mean** = \( \lambda \) \( \text{Standard deviation} = \sqrt{\lambda} \)

\[
P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}
\]

**Example:** Let \( Y \) = the number of fish caught after two hours of fishing at a fishing hole where, on average, 5 fish are caught every hour.

Parameter: \( \lambda = 5 \)

Mean = \( 5 \times 2 = 10 \) fish for 2 hours of fishing \( \text{Standard deviation} = \sqrt{5 \times 10} \)

Possible values: 0, 1, 2, \ldots \infty

What is the probability that no fish are caught?

\[
P(Y = 0) = \frac{e^{-10} 10^0}{0!}
\]

which equals 0.008.

**Another example:** Let \( Y \) = the number of mosquitos in a four-man tent at a particular point in time. The unit of measure for this example is the volume of the tent. On average, 2 mosquitos can be found in the tent for a particular point in time.

Parameter: \( \lambda = \) the average number of mosquitos in a four-man tent at a particular point in time

Possible values: 0, 1, 2, \ldots \infty

Mean = 2 mosquitos per tent \( \text{Standard deviation} = \sqrt{2} \)

What is the probability that exactly one mosquito is in a randomly selected four-man tent?

\[
P(Y = 1) = \frac{e^{-2} 2^1}{1!}
\]

which equals 0.271.

Figure ?? displays some Poisson distributions.

**Good to Know:** The mean and variance of a poisson random variable are equal.

3.3.4 Continuous Random Variables

With continuous random variables, we define probabilities using probability density functions, \( pdf \). Probabilities are calculated by computing the area under the density curve over the interval of interest.

3.3.4.1 Exponential

**Chance experiment:** Observe a poisson process and wait until the first event occurs.

**Y =** time until the first event occurs.
Mean = \frac{1}{\lambda} \text{ Standard deviation } = \sqrt{\frac{1}{\lambda}} \text{ Parameter: } \lambda \text{ is a rate of an event per time unit.}

Possible values: \((0, \infty)\)

\[ f(y) = \lambda e^{-\lambda y} \]

**Example:** Wait until a hard drive crashes for a randomly selected PC. Most will have similar lifetimes but a few may last a very long time without a crash. The mean time until a crash is 250 days. On average the \(1/\lambda\) or 0.004 crashes occur daily.

\( Y = \text{time, in days, until the first crash} \)

Parameter: \(1/\lambda = 200\) average time, in days, until a hard drive crashes

Possible values: \((0, \infty)\) days What is the probability that a randomly selected hard drive crashes in 200 days?

\[ f(y) = 0.004 e^{-0.004 \times 200} \]

which equals 0.368.

**Good to Know:** Exponentials are often used to model waiting times.

### 3.3.4.2 Gamma

**Chance experiment:** Observe a Poisson process and wait until \(k\) events occur for events that occur at \(\text{lambda}\) per time unit.

\( Y = \text{time until } k \text{ events occur. A gamma random variable is the sum of } k \text{ independent exponentially distributed random variables, each with the same } \lambda. \)

**Parameters:** \(\lambda, k\), the number of events or alternatively a gamma function can be parammeterized in terms of its shape, \(\alpha > 0\) and scale, \(\beta > 0\).

Mean = \( k \lambda \) \text{ Standard deviation } = \sqrt{k \lambda^2}

Mean = \(\alpha/\beta\) \text{ Standard deviation } = \sqrt{(\alpha/\beta^2)}

\[ f(y) = \frac{1}{\Gamma(k)\lambda^k} y^{k-1} e^{-y/\lambda} \]

**Example**
Chance experiment: While fishing, wait until 5 fish are caught at a fishing hole where one fish is caught every hour, on average. Let $Y = \text{time, in hours, until 5 fish are caught while fishing}$. Contrast this example to the Poisson random variable above.

**Parameters:** $1/\lambda = 1$, the average time until a fish is caught

Possible values: $(0, \infty)$ hours

What is the probability that you wait 6 or more hours before you catch a fish?

$$f(y) = \frac{1}{1!} \frac{6^1}{5^1} e^{-6/5}$$

which is 0.616.

**Good to Know:** Its two parameters make gamma distributions a more flexible alternative to the exponential for modeling waiting time.

### 3.3.4.3 Gaussian or Normal

You are already very familiar with this distribution.

**Chance experiment:** Randomly select an item from a population.

$Y = \text{value for a randomly selected observation from a population characteristic which is normally distributed.}$

Parameters: $\mu$, the population mean, and $\sigma$, the population standard error.

$$f(y) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

**Example:** The weight of boxes of Fruity Tootie cereal boxes is approximately normally distributed with an average weight of 15 ounces and a standard deviation of 0.5 ounces.

Let $Y = \text{weight of the contents of a randomly selected box of Fruity Tooties.}$

Parameters: $\mu = 15$ ounces and $\sigma = 0.5$ ounces

Possible values: $(0, \infty)$

What is the probability that the weight is more than 15.5 ounces?
0.0
0.1
0.2
0.3
0.4
−10 0 10 20 30
values
density
Normal Distributions: N(0,1)  N(5,1)  N(10,5)

Figure 3.5: Some Normal distributions.

\[ \int_{15.5}^{\infty} f(y) dy = \frac{1}{\sqrt{2\pi}0.5^2} e^{-\frac{(15.5-15)^2}{20.25}} \]

which is 16%.

You may have noticed that a random variable’s parameters always appear in the probability functions.

3.3.4.4 Beta

3.4 Distributions used in Testing

3.4.1 \( \chi^2 \)

You have probably already encountered \( \chi^2 \) tests before. For example, \( \chi^2 \) tests are used with contingency tables to investigate the association between row and column variables. \( \chi^2 \) tests are also used in goodness-of-fit testing such as comparing counts expected according to Mendelian ratios to observed data. In general, \( \chi^2 \) tests compare observed counts to what would be expected under the null hypotheses and reject the null when these observed discrepancies are too large.

In this course, we encounter \( \chi^2 \) distributions in several testing situations, for example we frequently perform likelihood ratio tests (LRTs) to compare models. When a larger model provides no significant improvement over a reduced model, the LRT statistic (which is twice the difference in the log-likelihoods) follows a \( \chi^2 \) with the degrees of freedom equal to the difference in the number of parameters. The shape of \( \chi^2 \) distributions are right skewed with a mean at its df.

What does a \( \chi^2 \) distribution look like? Values from a \( \chi^2 \) distribution are non-negative. When squared values of standard normal deviates, \( z \)-scores, are added, they follow a \( \chi^2 \)-distribution. Like \( t \)-distributions, \( \chi^2 \)-distributions are indexed by degrees-of-freedom with higher df corresponding to more information. The expected value of a \( \chi^2 \) distribution is equal to its degrees of freedom.
3.4.2 $F$

$F$-distributions are also used when performing statistical tests. Like the $\chi^2$ distribution, the values from an $F$-distribution are non-negative and the distribution is right skewed. RA Fisher (for whom the test is named) devised this test statistic to compare two different estimates of the same variance parameter and it has a prominent role in Analysis of Variance. Model comparisons are often based on the comparison of variance estimates, e.g. the extra sums-of-squares $F$-test. $F$-distributions are indexed by two sets of degrees-of-freedoms, one for the numerator and one for the denominator. The expected value for an $F$-distribution under the null hypothesis is 1.0.

3.5 Mixtures

Random variables can be combined to model more complex processes.

3.5.1 Zero-inflated Poisson

**Example:** Sometimes you encounter a distribution of counts which you would expect to have a Poisson distribution but there seems to be too many zero counts. An example might be responses to a survey asking students at a small college how many drinks they had the previous weekend. If the population of students includes a subset of students who are non-drinkers you will see many more reports of no drinks than you would expect from a population of drinkers with the number of drinks distributed as a Poisson. The Poisson distribution for the drinkers would have some zeros, drinkers who did not happen to drink that weekend. However, some of the zeros would come from non-drinkers, ‘true zeros.’ So non-drinkers are mixed in with the drinkers, the proportion of non-drinkers being $\alpha$. When you look at your observations you only see the reported number of drinks. You don’t know who is a drinker or who is a non-drinker. The proportion of non-drinkers, $\alpha$, is unknown or latent but it is a parameter of interest that we’d like to estimate. The observations can be modeled using a random variable referred to as a Zero-inflated Poisson. $Y$ = either 0 for non-drinkers or a Poisson count (which could be 0) for drinkers.

**Parameters:** $\lambda$, the average number of drinks for the population of drinkers, $\alpha$ a mixing parameter, the proportion of non-drinkers (true 0’s) in the population.

The method for estimating the parameters involves the likelihood.
Example: Suppose the time of onset for a population of schizophrenic patients is recorded and graphed. It may look something like Figure 3.7. Clearly this distribution does not appear to be normal, but rather a combination of two normal distributions, one centered around 20 and another at 40. Furthermore it appears that the standard deviation for the younger population is less than the standard deviation for the older population. Based on this graph, it may be conjectured that there are in fact two different diseases labelled schizophrenia which may have different prognoses or may be associated with different genes. It may of interest to determine what proportion of the patients fall into this younger population of patients and what proportion of the entire population the older group comprises. A mixture of normals can be modeled in the following way.

$$\text{Mixture: } N(22,2) \text{ and } N(40,5)$$

**Parameters:** $\mu_1, \mu_2$, the population means for the first and second distributions, and $\sigma_1, \sigma_2$, the population standard deviations, and $\alpha$ a mixing parameter, the proportion of the population in population 1.

To generate a mixture of two normal distribution of two random variables, randomly select a value from a Binary distribution which takes on the value 1 with probability $\alpha$ and 2 with probability $1 - \alpha$. When the binary observation is 1, it implies that an observation be selected from normal distribution one with mean, $\mu_1$, and standard deviation, $\sigma_1$ otherwise randomly select from another distribution with mean, $\mu_2$ and standard deviation, $\sigma_2$, respectively. The method for estimating the parameters involves the likelihood.

3.5.3 Beta-Binomial

**Beta** random variables take on values between 0 and 1. They are very flexible with the shapes being characterized by two parameters, $\alpha$ and $\beta$. These distributions are useful in generating probabilities and have the following pdf.

$$f(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1 - y)^{\beta-1}$$

**Beta-binomial** random variables are created when a probability, $p$, is drawn from a Beta distribution($\alpha, \beta$) and used as the parameter for a binomial random variable where the number of successes out of $n$ independent, identical trials is generated. Following the first random draw, $p_1$ from the beta distribution, the first observation $Y_1$ is generated using a binomial distribution with
$n$ trials and probability of success, $p_1$. The next observation is generated by selecting a new probability of success, $p_2$, from the same beta distribution. Again a binomial random variable, $Y_2$, is generated this time with $n$ trials and probability of success, $p_2$. This continues with a new probability of success drawn from the same beta distribution for each binomial observation with $n$ trials. Note that all of the observations will be integer values from $0, 1, \ldots, n$. So the observed values look like binomial random variables but each one does not have the same probability of success. We will revisit this idea later in the text when we discuss correlated data.

3.5.4 Negative Binomial

**Gamma** random variables take on continuous non-negative values and are useful for producing non-negative parameters such as rates. This distribution serves a role similar to the beta in the beta-binomial variable. While draws from a beta provided $p_i$ for a binomial, here a gamma provides the rate parameter, $\lambda_i$, which is used for Poisson random variable to produce a count, $Y_i$. Like the beta-binomial, the $Y_i$s are counts and may appear to be Poisson observations. However the rates, $\lambda_i$s differ so the distribution is not Poisson, but rather **Negative Binomial** (which corresponds to the Beta-Binomial in the previous section).

3.6 Notes on Using R

For the discrete distributions provided:

- `droot` returns a probability,
- `proot` returns a cumulative probability,
- `qroot` returns a quantile, and
- `rroot` returns a random number.

For the continuous distributions:

- `droot` returns the density,
- `proot` returns a probability,
- `qroot` returns a quantile, and
- `rroot` returns a random number.

Different roots available include:

<table>
<thead>
<tr>
<th>Hyper</th>
<th>Hypergeometric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binom</td>
<td>Binomial</td>
</tr>
<tr>
<td>Geom</td>
<td>Geometric</td>
</tr>
<tr>
<td>Pois</td>
<td>Poisson</td>
</tr>
<tr>
<td>Exp</td>
<td>Exponential</td>
</tr>
<tr>
<td>Norm</td>
<td>Normal</td>
</tr>
<tr>
<td>T</td>
<td>Student’s-t</td>
</tr>
<tr>
<td>Beta</td>
<td>Beta</td>
</tr>
<tr>
<td>Chisq</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

# Mixture of Normal Plots

```
mu1=20
s1=4
mu2=40
```
3.7 Worked Example

Generate 1000 observations from a Beta-binomial random variable. You can use the R code below to help you.

1. Investigate the properties of a Beta distribution by drawing 1000 observations from a Beta distribution with the arbitrarily selected parameters \( \alpha = 4 \) and \( \beta = 1 \).
   - What is the mean of this distribution? Can you relate it to the parameters? Plot a graph to approximate the probability density function. From the graph, describe what values you expect to observe when drawing from this distribution.

2. Simulate a plain old vanilla binomial random variable with \( n = 10 \) and \( p = 0.8 \) by generating 1000 draws from \( \text{binomial}(10, 0.8) \). Plot the histogram of these binomial observations.

3. Create a distribution of beta-binomial random variables by:
   - (a) Draw \( p_i \) from the beta distribution with \( \alpha = 4 \) and \( \beta = 1 \).
   - (b) Generate a binomial random variable \( Y_i \) with \( n = 10 \) trials and probability of success, \( p_i \).
   - (c) Repeat 1000 times \( (i = 1, \ldots, 1000) \) and plot the histogram of these beta-binomial random variables.
   - (d) Although both sets of random variables are based on \( n = 10 \) trials and an average \( p = 0.8 \), the two distributions differ in important ways. Compare the shapes, means and standard deviations for these two distributions.

```r
# Worked Example
# Distribution Thy Chapter
# Creating a Beta-Binomial RV
#
# First draw 1000 obs from a Beta distribution
#
alpha=4
beta=1

# Mean of Beta is \( \alpha/(\alpha+\beta) \)

mean.beta = alpha/(alpha+beta)

mean.beta

n=10

# The values from the Beta distribution will serve as
# randomly selected \( p=\)probability of success
# for each binomial generated

random.p = rbeta(1000, alpha, beta)

mean(random.p)
```
3.8 In Class Activity: A Poisson Process

Use a Poisson process generator from the internet. Divides the class into three groups. One group will watch until the first event occurs and records the time. Another group waits until 3 events occur and records the time. Group three counts the number of events in 5 minutes. Which random variables apply to each of the three sets of these generated values?

3.9 A Schematic for the Relationship between Distributions

A very cool schedmatic representing how probability distributions are related to one another can be found at

http://www.johndcook.com/distribution_chart.html

3.10 Exercises

3.10.1 Conceptual Exercises

1. Identify two uses for random variables in statistics.
2. Consider a normal random variable. Is it continuous or discrete? Are the values bounded or unbounded? Is its distribution skewed or symmetric?
3. What is the difference between sampling with or without replacement as for the hypergeometric and binomial random variables?
4. What is the difference between a Bernoulli and Poisson process?
5. Do mixture distributions apply only to normal distributions?
6. What does the parameter $\alpha$ represent for a mixture distribution?
<table>
<thead>
<tr>
<th>Random Variable</th>
<th>Continuous or Discrete?</th>
<th>Possible Values</th>
<th>Parameters</th>
<th>pdf</th>
<th>..</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypergeometric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bernoulli</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binomial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geometric</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative Binomial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poisson</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exponential</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gamma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1: Characteristics of Selected Random Variables

3.10.2 Guided Exercises

1. **Random Variable Table** Complete Table 1.

2. **Random Variable Applications** For each random variable in Table 1, describe a setting where the random variable may be used for modeling. Specifically define $Y$ in the context of the setting. For example, customers arriving at a service counter approximate a Poisson process. Let $Y =$ the number of customers arriving in a 5 minute time period.

   Many of the following exercises require you to use R and its built-in distributions. See the section 3.6 as a reference.

3.10.3 Guided Exercises

1. **Graphs of distributions** Create two graphs for each random variable in the Table 1. Use different parameter values for the two graphs and describe how changing these values affects the graph. Note that your choice of possible values will depend upon the parameter values. Be sure to use lines for continuous random variables (type="l") and spikes for discrete random variables (type = "h").

   - Is this random variable discrete or continuous?
   - What are the possible values this random variable can take on?
   - Is the distribution symmetric? Or does it depend on...?

   When creating a graph, for discrete random variables, use something like
   
   ```r
   x=0:10 this gives integer values,
   ```
   
   for continuous random variables, use something like
   
   ```r
   x <- seq(-3,3, by=.1).
   ```

   For discrete random variables, use graph type “h.”
   For continuous random variables, use graph type “l.”

   Here is an example of some code to examine the effect of changing the mean or standard deviation for a normal random variable. Use these ideas to explore other distributions. In particular, determine the following:
EXERCISES 65

# Distribution Thy
# A standard normal distribution
x <- seq(-3,3, by=.1)
y <- dnorm(x)
plot(x, y, type="l")

# Same Mean, Change Sigma
y5 <- dnorm(x,0,5)
lines(x, y5, type="l")

# Same Mean, Change Sigma
y2 <- dnorm(x,0,2)
lines(x,y2,col="red")

# Same Mean, Change Sigma
y_5<dnorm(x,0,0.5)
lines(x,y_5,col="green")

# Change the Mean and Sigma
y3<-dnorm(x,1,1)
lines(x,y3,col="blue")

2. **Negative Binomial activity** Obtain a histogram of 1,000 draws of a plain old vanilla Poisson random variable, y, with $\lambda=1.5$. Next, create 1000 $\lambda$s by selecting from a Gamma distribution with parameters 3 and 2. Produce a Negative Binomial random variable by generating a Poisson count for each of the 1000 $\lambda$s, $y_{bn}$. Plot the histograms for y and $y_{bn}$ and compare. How do the means and standard deviations compare?

3. **Beta-Binomial** Using the approach taken for the generation of Negative Binomial random variable values, generate values for a Beta-Binomial random variable.

4. **Lazy DJ** A lazy DJ doesn’t try to gauge the crowd but simply uses a shuffle to select tunes. Worse still, the shuffle selects tunes with replacement so that the same tune may show up again. If only 1 out of 4 tracks is a dance hit,
   (a) How many tracks are played until a dance hit comes up? Which random variable did you use?
   (b) Out of 10 tracks played, what is the probability that no dance hits are played? What random variable did you use?
   (c) Patrons insist on staying until 5 dance hits are played. How many tracks will the patrons listen to until 5 dance hits are played, on average? What random variable did you use?

5. **Trashball I** Typically, you can make a basket using a trash ball about 35% of the time. To test this claim, use a crumbled piece of paper, aim for the trash basket, and take a shot.
   (a) If you take 5 shots, what is the probability that you make them all?
   (b) What random variable did you use?
   (c) What assumptions did you make when applying this random variable? Do you think they are realistic?

6. **Trashball II** Ten students each take 5 shots. Why would a binomial random variable not be appropriate for the total baskets out of the 50 shots?

7. **Mixtures** Describe how you could use R to generate the following mixtures:
   (a) Zero-inflated poisson random variables
   (b) Mixture of two Normal distributions
   (c) Negative binomial
3.11 References

Chapter 4

Poisson Regression

**Learning Objectives**

- Describe a Poisson process and a corresponding Poisson random variable.
- Describe why simple linear regression is not ideal for Poisson data.
- Write out a Poisson regression model and identify the assumptions for inference.
- Write out the likelihood for a Poisson regression and describe how it could be used to estimate coefficients for a model.
- Interpret an estimated coefficient from a Poisson regression and construct a confidence interval for it.
- Use deviances for Poisson regression models to compare and assess models.
- Use an offset to account for varying effort in data collection.
- Optional: Fit and use a zero-inflated Poisson (ZIP) regression.

4.1 Preface

This chapter introduces themes and approaches to modeling that you will use throughout *Broaden Your Statistical Horizons*. Major ideas you will see again and again include:

1. **Response**: Explicitly define your response. This will be very helpful in choosing a modeling approach.
2. **Research question**: The research question will be your road map as you proceed through your analysis, so do your best to articulate it.
3. **EDA**: Tailor your exploratory data analysis to your research question(s). What kinds of graphs and numerical summaries will help you understand your research question?
4. **Modeling**
   (a) Fitting a model involves estimating coefficients using likelihoods, not least squares.
   (b) Checking assumptions for models using graphs and numerical summaries.
   (c) Comparing models using deviances.

4.2 Introduction to Poisson Regression

4.2.1 Initial Examples

Are the number of motorcycle deaths related to a state’s helmet laws in a given year? Does the number of rapes on a campus during a year differ for public and private colleges? Does the daily number of asthma-related visits to an Emergency Room differ depending on an air pollution index? Has the number of deformed fish in randomly selected Minnesota lakes been affected by changes in trace minerals in the water over the last decade? Like earlier chapters, each example involves predicting a **response** using one or more **explanatory variables**. These examples have response
variables that are counts per some unit of time or space. We can explicitly define these responses and the corresponding explanatory variables in the following way.

- Y = number of motorcycle deaths by state
  X = helmet law indicator
- Y = number of rapes on a college campus in a school year
  X = private or private school indicator
- Y = number of asthma-related visits to an Emergency Room daily
  X = an air pollution index
- Y = number of deformed fish in a randomly selected square meter of a Minnesota lake
  X = trace mineral measurements

4.2.2 Poisson Random Variables

A Poisson random variable is often used to model counts. A description of key features of Poisson random variable follows.

<table>
<thead>
<tr>
<th>Properties of a Poisson random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Y = number of events per unit (time or space)</td>
</tr>
<tr>
<td>2. Possible values: 0, 1, 2, \ldots, \infty</td>
</tr>
<tr>
<td>3. Mean = \lambda</td>
</tr>
<tr>
<td>4. Variance = \lambda, which implies a standard deviation = \sqrt{\lambda}</td>
</tr>
<tr>
<td>5. P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}</td>
</tr>
<tr>
<td>6. The sum of Poisson variables is also Poisson.</td>
</tr>
</tbody>
</table>

The parameter of interest in Poisson modeling is \lambda, the rate of events per unit time or space. The observed count, Y, is made over a number of units of time such as year or space such as acre or volume such as square foot of lake water. The unit for the first example is state, the next is school year, for the ER example the unit is a day and for the last example, the unit is a square meter of a Minnesota lake. Define the unit so the rate, \lambda, is meaningful.

4.2.2.1 An example calculating Poisson probabilities

Let Y = the number of soccer goals scored in a game by DC United. Suppose that it is known that an average DC United scores 1.02 goals per game (\lambda = 1.02). How frequently would exactly two goals be scored during a game? With Property 5 and the information that \lambda = 1.02 goals per game, we can calculate the probability for any number of goals per game using:

\[ P(Y = y) = \frac{e^{-1.02} 1.02^y}{y!} \]

For our example, we expect to see exactly 2 goals in a game

\[ P(Y = 2) = \frac{e^{-1.02} 1.02^2}{2!} = 0.1876 \]

or nearly 20% of the time. Using the laws of probability it is also straightforward to calculate the chance of any events of interest such as seeing more than 2 goals in a game or at least one goal in a game. Conversely, statistical applications will have data such as Y=2, and the aim will be to estimate \lambda.
4.2.3 Modeling with Poisson variables

A Poisson random variable is a count so its minimum value is zero, and in theory the maximum is unbounded. We’d like to model the parameter, the average count per unit or \( \lambda \), as a function of one or more covariates. For an OLS model, the parameter, the average response, \( \mu_i \), is represented as a line in the case of one explanatory variable. It might seem reasonable to try to model the Poisson parameter \( \lambda_i \) as a linear function of an explanatory variable but there are some problems with this approach.

\[
\lambda_i = \beta_0 + \beta_1 x_i \quad \text{This doesn't work well for Poisson data.}
\]

A line is certain to yield negative values, but \( \lambda_i \) takes on values from 0 to \( \infty \). In addition the equal variance assumption is violated because as the mean rate for a Poisson variable increases the variance increases.

One way to avoid these problems is to use model the \( \log(\lambda_i) \) as a function of the covariates instead of \( \lambda_i \). The \( \log(\lambda_i) \) takes on values from \(-\infty \) to \( \infty \). We can also take into account the increase in the variance using this approach. (Note that throughout Broaden Your Statistical Horizons we use \( \log \) to represent the natural logarithm.)

\[
\log(\lambda_i) = \beta_0 + \beta_1 x_i \quad (4.1)
\]

where the observed values \( Y_i \sim \text{Poisson with } \lambda = \lambda_i \) for \( x_i \).

4.2.3.1 Poisson Regression Assumptions

Much like OLS, using Poisson regression to make inferences requires assumptions.
1. **Linearity** The log of the mean rate, \( \log(\lambda_i) \), must be a linear function of \( x \).
2. **Mean=Variance** By definition the mean of a Poisson random variable must be equal to its variance.
3. **Independence** The observations must be independent of one another.

4.2.4 A Graphical Look at Poisson Regression

Figure 4.1 illustrates a comparison of the OLS model for inference to Poisson regression using a log function of \( \lambda \).

1. The graphic displaying the ordinary least squares (OLS) inferential model appears in the left panel of Figure 4.1. It shows that for each level of \( X \), the responses appear to be approximately normal. The panel on the right side of Figure 4.1 depicts what a Poisson regression model looks
like. For each level of $X$, the responses follow a Poisson distribution. For Poisson regression small values of $\lambda$ are associated with a distribution that is noticeably skewed with lots of small values and only a few larger ones. As $\lambda$ increases the distribution of the responses begins to look more and more like a normal distribution.

2. In the case of OLS, the mean responses for each level of $X$, $\mu_{Y|X}$, fall on a line. In the case of the Poisson model, the mean values of $Y$ at each level of $X$, $\lambda_{Y|X}$, fall on a curve, not a line.

3. In the OLS regression model, the variation in $Y$ at each level of $X$, $\sigma^2$, is the same. For Poisson regression the responses at each level of $X$ become more variable with increasing means.

4.2.4.1 Poisson Models (Optional)

That single parameter, $\lambda$, specifies both the mean and variance of the distribution of counts. Thus as the mean increases the variance does as well. This is in contrast to OLS assumptions for inference which require the variance to be constant for all mean values of $Y$. While relaxing the assumption of equal variances is convenient, specifying that the variance be equal to the mean is rather restrictive. It would be nice to be able to uncouple the specification of the mean and variance instead of having them be exactly equal to one another. You will see later in this chapter that there are ways in which to modify the Poisson model to allow for more flexibility. These modifications reflect a principle for Broaden Your Statistical Horizons, modeling can be flexible. When modeling, you are not restricted to only a few options.

4.2.4.2 Poisson Process (Optional)

A Poisson model for counts can be thought of as being generated from a Poisson process. A time period or space can be broken into very small, equal size units. The probability of an event within a given unit is very small and independent of other units. For example, the probability of a rape within a very short period of time is very small. A Poisson model implies that the probability of a rape occurring during another very small, equal length time period is the same and independent of a rape occurring within another time period of equal length. Two related facts are worth noting. First, observations are independent. Second events occur during this process at a constant rate. The process can also be generated by using exponentially distributed waiting times between events.

4.3 Case Studies Overview

We take a look at this model in the context of a couple case studies. Each example illustrates different concepts encountered in Poisson regression. They are based on real data and real questions. Modeling the household size in Haiti introduces the idea of regression with a Poisson response along with its assumptions. A quadratic term is added to a model to determine an optimal size per household. The weekend drinking example uses a modification of a Poisson model to account for more zeros than would be expected for a Poisson random variable. These two case studies also provide context for some of the familiar concepts related to modeling such as exploratory data analysis, estimation and residual plots. An optional section demonstrates the use of likelihoods to fit Poisson models.

4.4 Case Study: Household Size in Haiti

How many other people live with you in your home? The number of people sharing a house differs from country to country and often from village to village. International agencies use household size when determining needs of populations. Household sizes determine the magnitude of the household needs. Much of the work for maintaining a household falls to the female head of the household. When can a female head of a household expect to be caring for the largest number of people?
Here we seek to determine when women in Haiti are likely to find the largest number of people in their household. We begin by explicitly defining our response, $Y =$ number of people in a household and the explanatory variables $X_1 =$ age of the female head of the household and $X_2 =$ type of floor (dirt, wood, cement, tile). Our response is a count so we consider a Poisson regression with the parameter of interest is $\lambda$, the average number of people per household. We model the average age of the female head of household controlling for the type of floor, an indicator of the relative wealth of a household.

This first case study introduces a number of key ideas. Here are the highlights.

1. **Modeling**: coefficients are estimated using likelihoods.
2. **Interpretation**: Exponentiate coefficients to interpret them.
3. **Assumptions**: Check the Poisson assumptions and Model Fitfit of the model.
   - **Linearity**: look at the log of mean responses vs $X$
   - **Mean=Variance**: look at the mean and variance of responses at levels of $X$
   - **Outliers**: look at residual plots
4. Compare models using deviances
5. Overdispersion: When a model is correctly specified and outliers considered, adjust for overdispersion if necessary.

### 4.4.1 Research Question

At what age are women in Haiti most likely to find the largest number in their household?

### 4.4.2 Data Collection

In the summer of 2010, 401 households in the region of Port Salut, Haiti were included in a public health survey to ascertain the health needs of the population. Trained community members completed 401 household surveys in an eight week period. A systematic sampling method was used in which every third (or fourth) house in a village was included in the survey sample. The survey was administered by Dr. Therese Zink of the University of Minnesota.

### 4.4.3 Data Organization

Each line of the data file refers to a household at the time of the survey:

- **age** = the age of the female head of household
- **numHouse** = the number in the household besides the female head of the household at the time of the survey (response)
- **floor** = the type of floor in the household (dirt(1), wood(2), cement(3), tile(4))

<table>
<thead>
<tr>
<th>index</th>
<th>agefHH</th>
<th>n.in.house</th>
<th>floor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>52</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>55</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

### 4.4.4 Exploratory Data Analysis

Poisson random variables typically have responses of no events or zeros. However, as constructed, the response of **number in the house** does not include any zeros because the female head of the
Figure 4.2: Haiti: Number in the house besides the female head of household.

Figure 4.3: (a) The raw data of number of inhabitants by age of female HH. (b) The log of the mean number in each household by grouped age of the female head of household.

household is also included. We can create a variable more appropriate for Poisson regression by subtracting one from the total in the household thereby describing the number in the household besides the female head of household. To simplify the discussion, we will refer to this variable as numHouse. Now we are ready to proceed.

Figure 4.2 displays the responses using a plot with spikes at each non-negative integer (0, 1, 2, ...). Counts cannot include values between integers hence the spikes. Figure 4.2 reveals a fair amount of variability in the number in each house ranging from 0 to 17 with many of the respondents reporting between 3 to 7 people in the house besides the female head of household. Like many Poisson distributions, this graph is right skewed. It clearly does not suggest that the numHouse is a normally distributed response.

Our research question concerns determining the age at which females have the most people in the house. A primary objective of this EDA is to determine whether a linear or quadratic model of the \( \log(\lambda) \) seems most plausible. If a quadratic model fits, we can determine the age for which the numHouse is a maximum with a little calculus.

A plot of the numHouse by the age of the female household (Figure 4.3) provides a look at the raw data. The spline on the plot exhibits some curvature.

For the remainder of the analysis, we’ll proceed in the following way.

Fit a linear model in age (using likelihoods).
1. Interpret the estimated coefficient for the age term and test its significance using a Wald-type test and a drop-in-deviance test.

2. Assess the assumptions of the linear model.
   (a) Linearity
   (b) Independence
   (c) Mean=Variance
   (d) A residual plot with no trends or outliers

**Fit a quadratic model.**

1. Test the significance of the coefficient of the age^2 term using a Wald-type test.
2. Add the factor Floor to the quadratic model, per household controlling for differences in Floor.
3. Introduce the drop-in-deviance test as a way of comparing models
   (a) Compare the linear to the quadratic models
   (b) Comparing the quadratic model with and with the Floor factor.
4. Examine the residual plot for the final model.
5. Use calculus to determine the age at which there is the maximum number of people.

**Adjust for overdispersion.**

### 4.5 Using Likelihoods to fit Poisson Regression Models (Optional)

Before we have a statistical package produce estimated regression coefficients for our Poisson regression model, let’s look under the hood to see how the model is fit. The least squares approach requires a linear relationship between the parameter, μi, and xi. However, it is \( \log(\lambda_i) \), not \( \lambda_i \), that is linearly related to X with the Poisson model. As noted, the assumptions of equal variance and normality also do not hold for Poisson regression. So the method of least squares will not be helpful here. Instead of OLS, we employ the likelihood principle to find estimates of our model coefficients.

We look for those coefficient estimates for which the likelihood of our data is maximized. These are the **maximum likelihood estimates**.

The likelihood for \( n \) independent observations is the product of the probabilities. For example, if we observe five households with \( \text{numHouse} = (4, 2, 8, 6, 1) \) the likelihood is:

\[
\text{Likelihood} = P(Y_1 = 4) * P(Y_2 = 2) * P(Y_3 = 8) * P(Y_4 = 6) * P(Y_5 = 1)
\]

Recall that the probability of a Poisson response can be written

\[
P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}
\]

So the likelihood can be written as

\[
\text{Likelihood} = \frac{e^{-\lambda} \lambda^4}{4!} * \frac{e^{-\lambda} \lambda^2}{2!} * \frac{e^{-\lambda} \lambda^8}{8!} * \frac{e^{-\lambda} \lambda^6}{6!} * \frac{e^{-\lambda} \lambda^1}{1!}
\]

As in the earlier likelihood chapter, it will be easier to find a maximum if we take the log of the likelihood and ignore the constant term resulting from the sum of the factorials:

\[
-logL = \lambda_{X=4} - 4 \log(\lambda_{X=4}) + \lambda_{X=2} - 2 \log(\lambda_{X=2}) + \lambda_{X=8} - 8 \log(\lambda_{X=8}) + \lambda_{X=6} - 6 \log(\lambda_{X=6}) + \lambda_{X=1} - \log(\lambda_{X=1})
\] (4.2)
Now if we had the age of the female head of the household for each house \((X)\), we consider the model:

\[
\log(\lambda_i) = \beta_0 + \beta_1 x_i
\]

This implies that \(\lambda\) differs for each age and can be determined using

\[
\lambda_i = e^{\beta_0 + \beta_1 x_i}.
\]

If the ages are \(X = (22, 21, 25, 24, 22)\) years of age and \textbf{numHouse} = \((4, 2, 8, 6, 1)\), respectively, our loglikelihood can be written:

\[
\begin{align*}
\log L_1 &= -[e^{\beta_0 + \beta_1 22} - 4(\beta_0 + \beta_1 22)] - [e^{\beta_0 + \beta_1 21} + 2(\beta_0 + \beta_1 21)] \\
&\quad -[e^{\beta_0 + \beta_1 25} + 8(\beta_0 + \beta_1 25)] - [e^{\beta_0 + \beta_1 24} + 6(\beta_0 + \beta_1 24)] \\
&\quad -[e^{\beta_0 + \beta_1 22} + (\beta_0 + \beta_1 22)]
\end{align*}
\] (4.3)

To see this, match the terms in Equation 4.2 with those in Equation 4.3 noting that \(\lambda_i\) has been replaced with \(e^{\beta_0 + \beta_1 x_i}\). It is Equation 4.3 that will be used to estimate the coefficients, \(\beta_0\) and \(\beta_1\). Although this looks a little more complicated than those loglikelihoods we saw previously, the fundamental ideas are the same. In theory, we try out different possible values of \(\beta_0\) and \(\beta_1\) until we find the two for which the loglikelihood is largest. Most statistical software packages have automated search algorithms to find those values for \(\beta_0\) and \(\beta_1\) that maximize the loglikelihood.

### 4.6 Modeling

We first consider a model for which the \(\log(\lambda)\) is linear in \textbf{age}. We check the Poisson assumptions to determine whether another model (one with a quadratic term in \textbf{age}) will provide a better fit.

#### 4.6.1 First Order Model

Although we intend to fit a quadratic model, we’ll start with a linear model. Output for the linear model appears below.

```r
# First order model
numHouse~age,family=Poisson
```

Coefficients:

| Estimate  | Std. Error | z value | Pr(>|z|) |
|-----------|------------|---------|----------|
| (Intercept) | 2.135759   | 0.057323 | 37.258   | <2e-16 *** |
| age        | -0.012095  | 0.001273 | -9.504   | <2e-16 *** |

(Dispersion parameter for Poisson family taken to be 1)

Null deviance: 988.78 on 401 degrees of freedom
Residual deviance: 894.60 on 400 degrees of freedom
AIC: 2158.4

\[
\exp(\text{model.age$coef[2]})
\]

\[
\text{age}
\]

0.9879783
4.6.2 Estimation and Inference

Most software packages report the estimated coefficients so that the estimated regression equation for the previous model is:

\[ \log(\lambda) = 2.14 - 0.0121 \times \text{age} \]

How can the coefficient estimates be interpreted in terms of this application? As done when interpreting slopes in the OLS models, we consider how the estimated mean number in the house, \( \lambda \), changes as the age of the household head increases by an additional year. But in place of looking at change in the mean number in the house, with a Poisson regression we consider the log of the mean number in the house.

\[
\begin{align*}
\log(\lambda_X) &= \beta_0 + \beta_1 X \\
\log(\lambda_{X+1}) &= \beta_0 + \beta_1 (X + 1) \\
\log(\frac{\lambda_{X+1}}{\lambda_X}) &= \beta_1 \\
\frac{\lambda_{X+1}}{\lambda_X} &= e^{\beta_1}
\end{align*}
\] (4.4)

These results suggest that by exponentiating the coefficient on age we obtain the factor by which the mean count changes. In this case, the mean number in the house changes by a factor of \( e^{-0.012} = 0.99 \) or decreases by a factor of 1% with each additional year older the household head is. The quantity on the left hand side of Equation 4.4 is referred to as a rate ratio or relative risk and conveys a percent change in the response for unit changes in X. Typically the standard errors for the estimated coefficients are included in Poisson regression output. Here the standard error for the estimated coefficient for age is 0.0013 so we can construct a confidence interval for \( \beta_1 \). A 95% CI provides a range of plausible values for the age coefficient and can be constructed:

\[-0.0120 - 1.96 \times 0.0013, -0.0120 + 1.96 \times 0.0013\]

\[(-0.0145, -0.0095).\]

Exponentiating the endpoints yields a confidence interval for the relative risk, i.e. annual percent change, (0.988,0.992) suggesting that the mean number in the house decreases by 1.2% (1.992) to as little as 0.8% (1.988) for each additional year older the head of household is. It is best to construct a CI for the coefficient and then exponentiate the endpoints because the estimated coefficients are closer to normal than the exponentiated coefficients. There are other approaches to constructing intervals in these circumstance including profile likelihood, the delta method and bootstrapping but we do not discuss them here. If the null model were true, there is no change in the rate for each additional year, \( \lambda_X \) is equal to \( \lambda_{X+1} \) and the ratio \( \lambda_{X+1}/\lambda + X \) is 1. Note that this interval, (0.986,0.991), does not include 1 so the model with age is preferred.

One way to test the significance of the age term is to calculate a Wald-type statistic. A Wald-type test statistic is the estimated coefficient divided by its standard error. When the true coefficient is 0, this test statistic follows a standard normal distribution for sufficiently large \( n \). The estimated coefficient associated with the linear term in age is \( \hat{\beta}_1 = -0.0121 \). The standard error (0.001273) and the value for the test statistic (-9.504) (p << 0.00001) indicate that the estimated coefficient differs significantly from 0.

Another way in which to assess the contribution of the age term is to perform a drop-in-deviance test.
### 4.6.3 Using Deviances to Compare Models

How helpful is it to include `age` in the model? Keeping in mind that the deviance describes the remaining unexplained variation, we calculate the **drop-in-deviance** when adding `age` to the null model. The null model includes only an overall mean denoted as $\beta_0$ and no covariates. A residual deviance for the first order model is reported as 894.60 with 400 df. The output also includes the deviance and df for the null model (988.78 with 401 df). The drop-in-deviance is $(988.78-894.60=94.18)$ with a difference of only 1 df. If the null model is true, we would expect the drop -in-deviance to follow a $\chi^2$ distribution with one df which is the difference in df for the two models.

In order to use the drop-in-deviance test, the models being compared must be nested, i.e. all the terms in the smaller model must appear in the larger model. Here the smaller model is the null model with the single term $\beta_0$ an the larger model has $\beta_0$ and $\beta_1$ so it can be said that they are nested. For nested models, we can compare the sum of squared deviances or residual deviances to determine whether the larger model provides a significantly better fit. The drop-in-deviance and the Wald-type test are usually consistent, however if there is a discrepancy the drop-in-deviance is preferred.

**Drop-in-deviance test to compare models**

- Compute the deviance for reduced model - deviance for the larger model.
- When the reduced model is true, the drop-in-deviance $\sim \chi^2_d$ where $d=$ the difference in the degrees of freedom associated with the two models.
- A large drop-in-deviance favors the larger model.

**Testing whether an estimated coefficient differs significantly from 0**

- Wald-type statistic = estimated coefficient/standard error
- When the true coefficient is 0, for sufficiently large $n$, the test statistic $\sim N(0,1)$.
- If the magnitude of the test statistic is large, there is evidence that the true coefficient is not 0.

The Wald-type test and the drop-in-deviance test suggest that a linear term in age is useful. But are the model assumptions satisfied?

### 4.6.4 Poisson Regression Assumptions

#### 4.6.4.1 Linearity

Recall that the Poisson regression model implies that $\log(\lambda_i)$, not the mean $\lambda_i$, is a linear function of $x_i$. Therefore to check the linearity assumption with Poisson regression we would like to plot $\log(\lambda_i)$ by `age`, however $\lambda_i$ is unknown. Our best guess of $\lambda_i$ is the observed mean number in the household for each `age` (level of $X$). For this data set we do not have many repeat observations at each $x_i$ so the range of values for the `age` variable was divided into 20 equal length intervals. Means of number in the household are calculated for each age interval. Because these means are computed for observed data, they are referred to as *empirical* means. Taking the logs of the empirical means, and plotting by the median of each age group provides a way to assess the linearity assumption. The smoothed curve added to Figure 4.3 suggests that there is a curvilinear, not linear, relationship between age and the log of the mean household size. This plot suggests that the linear model is not adequate for describing this data and adding a quadratic term should be considered. This finding is consistent with the researchers’ hypothesis that there is an age at which a maximum household size occurs. It is important to note that, unlike OLS, this plot does not directly compare to the model we
Table 4.1 The age group of the female head of household, the mean number in the household, the variance in the number in the household for that age head of household.

<table>
<thead>
<tr>
<th>age group</th>
<th>mean number in the house</th>
<th>variance of number in the house</th>
</tr>
</thead>
<tbody>
<tr>
<td>(11.9,16.2]</td>
<td>4.00</td>
<td>2.00</td>
</tr>
<tr>
<td>(16.2,20.5]</td>
<td>5.78</td>
<td>8.81</td>
</tr>
<tr>
<td>(20.5,24.8]</td>
<td>6.78</td>
<td>10.39</td>
</tr>
<tr>
<td>(24.8,29.1]</td>
<td>5.88</td>
<td>12.38</td>
</tr>
<tr>
<td>(29.1,33.5]</td>
<td>4.96</td>
<td>10.12</td>
</tr>
<tr>
<td>(33.5,37.8]</td>
<td>6.13</td>
<td>10.02</td>
</tr>
<tr>
<td>(37.8,42.1]</td>
<td>5.74</td>
<td>11.48</td>
</tr>
<tr>
<td>(42.1,46.4]</td>
<td>7.00</td>
<td>9.09</td>
</tr>
</tbody>
</table>

eventually fit. It does however provide an idea of the form of the relationship between log(λ) and X.

Figure 4.4 Residual plot for the Poisson model of number in the household besides the female HH by age of the female HH.

4.6.4.2 Mean = Variance

For Poisson random variables, the variance of Y (i.e. the square of the standard deviation of Y), is equal to its mean. As the mean increases, the variance increases. So, if the response is a count and the mean and variance are approximately equal for each group of X, a Poisson regression model may be a good choice. In Table 4.1 we display only a portion of the age groups to check to see if the means and variances of the number in the house are approximately equal for each age group. This provides us one way in which to check the Poisson assumption.

If there is a problem with this assumption most often we see variances much larger than means. Here we see variances that are consistently larger than the mean suggesting that we may have a problem with the mean=variance assumption.

4.6.4.3 Independence

We do not have enough information to assess the independence assumption with the information we are given. If selections of groups of households were made from different villages with differing
customs about living arrangements, the independence assumption would be violated. If this were the case, we could use a multilevel model like those discussed in later chapters with a village term.

### 4.6.5 Residual Plot

A residual plot can be of limited use when assessing Poisson regression model assumptions. It may provide some insight into linearity and outliers although the plots are not quite as useful here as they are for OLS. There are a couple of options for computing residuals and predicted values. Residuals may have the form of residuals for OLS models or as deviance residuals (defined in the next section) which when squared, sum to the deviance for the model. The predicted values can be estimates of the counts, $e^{b_0 + b_1X}$, or estimates of the log of the count, $\beta_0 + \beta_1X$. We choose to use the deviance residuals and predicted counts. More detail follows.

### 4.6.6 Residuals for Poisson Models (Optional)

Residuals for OLS are useful for comparing models. The residuals for OLS have the form:

\[
\text{OLS residual}_i = \text{obs}_i - \text{fit}_i = Y_i - \hat{\mu}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1X_i)
\]

Residual sum of squares (RSS) are formed by squaring and adding these residuals. We have several options for creating residuals for Poisson regression models. One is to create residuals in much the same way as we do in OLS. For Poisson residuals, the predicted values are denoted by $\hat{\lambda}_i$ in place of $\hat{\mu}_i$ and divide by the standard error, $\sqrt{\hat{\lambda}_i}$ These kinds of residuals are referred to as Pearson residuals and are used with Poisson regression. These residuals have the advantage that you are probably familiar with their meaning and the kinds of values you would expect. However, deviance residuals have some useful properties that make them a better choice for Poisson regression.

First, we define a deviance residual for an observation from a Poisson regression,

\[
deviance residual_i = \text{sign}(Y_i - \hat{\lambda}_i) \sqrt{2 \left[ Y_i \log \left( \frac{Y_i}{\hat{\lambda}_i} \right) - (Y_i - \hat{\lambda}_i) \right]} \tag{4.5}
\]

As its name implies, a deviance residual describes how the observed data deviates from the fitted model. Squaring and summing the deviances for the observations produces the residual deviance. Relatively speaking, observations for good fitting models will have small deviances, that is, the predicted values will deviate little from the observed. However, you can see that the deviance for an observation does not easily translate to a difference in observed and predicted responses as is the case with OLS models.

A careful inspection of the deviance formula reveals several places where the deviance compares $Y$ to $\hat{\lambda}$: the sign of the deviance is based on the difference between $Y$ and $\hat{\lambda}$, under the radical sign we see the ratio $Y/\hat{\lambda}$ and $Y - \hat{\lambda}$. These terms are 0 when $Y = \hat{\lambda}$, that is, when the model fits perfectly. So like the residuals in OLS, an observation that fits perfectly will not contribute to the sum of the squared deviances. This definition of a deviance depends on the likelihood for Poisson models. Other models will have different forms for the deviance depending on their likelihood.

A plot (Figure 4.6.4.1) of the residuals versus predicted responses for the first order model reveals several potential outliers for smaller predicted responses. These observations may be responsible for the larger variances for some of the smaller responses. Fitting the model without the outliers
only attenuates the effect of age but it remains significant with a decidedly larger p-value (p=0.02). Of note is the fact that fitting a curve to the residual plot exhibits curvature indicating that the model may be improved by adding a quadratic term. Other details related to residual plots can be found in a variety of sources including McCullagh and Nelder (1989).

Our analysis thus far suggests that a quadratic term in age might be useful to add to the model.

4.6.7 Second Order Model

As stated from the outset our intention is to examine a quadratic model to see if there exists an age where the number in the house is, on average, a maximum. The output appears below.

# Quadratic Model
numHouse ~ age + age2, family = Poisson)

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|
| (Intercept) 1.700e+00 | 1.407e-01 | 12.081 | < 2e-16 *** |
| age 9.815e-03 | 6.558e-03 | 1.497 | 0.134459 |
| age2 -2.350e-04 | 6.938e-05 | -3.387 | 0.000707 *** |

(Dispersion parameter for Poisson family taken to be 1)

Null deviance: 988.78 on 401 degrees of freedom
Residual deviance: 882.56 on 399 degrees of freedom
AIC: 2148.3

We can assess the importance of the quadratic term in two ways. The p-value for the Wald-type statistic for $age^2$ is highly significant. Another approach is to perform a drop-in-deviance test.

$H_0: \log(\lambda) = \beta_0 + \beta_1 age$

$H_1: \log(\lambda) = \beta_0 + \beta_1 age + \beta_2 age^2$

The first order model has a residual deviance = 894.60 and the second order model, the quadratic model, has a residual deviance= 882.56. The drop-in-deviance by adding the quadratic term to the linear model is $894.6 - 882.56 = 12.04$ which can be compared to a $\chi^2_{1,df}$. The p-value can be found comparing 12.04 to a $\chi^2$ distribution with 1 df. The p-value is less than .0006, so the observed drop of 12.04 suggests that the quadratic term is indeed useful.

Some statistical packages will summarize the drop-in-deviance tests in tabular form.

anova(null.model, model.agehousehold,model.agehousehold2,test="Chisq")

Analysis of Deviance Table

| Model     | Resid. Df | Resid. Dev | Df | Deviance | P(>|Chi|) |
|-----------|-----------|------------|----|----------|----------|
| Model 1: numHouse ~ 1 | 401 | 988.78 | | | |
| Model 2: numHouse ~ age | 400 | 894.60 | 1 | 94.172 | < 2.2e-16 *** |
| Model 3: numHouse ~ age + age2 | 399 | 882.56 | 1 | 12.043 | 0.0005199 *** |
4.6.8 Adding a covariate

Adding a covariate is in essence comparing two nested models. For example, the variable `Floor` may help to explain some of the remaining variability in the number of residents. Results from the fitted model appear below.

| Coefficients          | Estimate | Std. Error | z value | Pr(>|z|) |
|-----------------------|----------|------------|---------|----------|
| (Intercept)           | 1.611e+00| 1.442e-01  | 11.173  | < 2e-16  *** |
| agefHH                | 8.683e-03| 6.604e-03  | 1.315   | 0.18855  |
| agefHH2               | -2.204e-04| 6.996e-05 | -3.150  | 0.00163  ** |
| factor(Floor)2        | 1.914e-01| 7.500e-02  | 2.552   | 0.01071  * |
| factor(Floor)3        | 1.470e-01| 5.122e-02  | 2.869   | 0.00412  ** |
| factor(Floor)4        | -2.984e-01| 3.567e-01 | -0.837  | 0.40278  |

We must use the drop-in-deviance test because three, not one, terms are added when including the `Floor` variable. The Analysis of Deviance table with the model adding the three terms corresponding to the variable `Floor` below. The addition of the `Floor` variable to the quadratic model (Model 3) is significant (p=0.008). The model with `age`, `age^2` and `Floor` (Model 4) is preferred.

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Model 1: n.in.house_1 ~ 1</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 2: n.in.house_1 ~ agefHH</td>
<td>1</td>
<td>988.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3: n.in.house_1 ~ agefHH + agefHH2</td>
<td>2</td>
<td>894.60</td>
<td>1</td>
<td>94.172</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>Model 4: n.in.house_1 ~ agefHH + agefHH2 + factor(Floor)</td>
<td>3</td>
<td>882.56</td>
<td>1</td>
<td>12.043</td>
<td>0.0005199 ***</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>870.74</td>
<td>3</td>
<td>11.822</td>
<td>0.0080193 **</td>
</tr>
</tbody>
</table>

4.6.9 Residual Plot

As with the previous model, we can examine a residual plot for the quadratic model (not shown here). It is similar in appearance to the first order model. The same points appear as outliers although the scatter does not exhibit the strong curvature that the first order model does.

4.6.10 Overdispersion

The plot of the log of the means by the number in the household revealed some curvature. A squared term in `age` and the addition of the variable `Floor` were significant. The possible outliers did not affect the significance of either term in the model nor the magnitude of the deviance. Nonetheless, there remains a relatively large deviance that may be attributed to overdispersion. Overdispersion suggests that there is more variation in the response than the model implies. The comparison of the means and variances for the different age groups found larger variances than the means whereas for a Poisson model one would expect these to be about the same. We can take overdispersion into account in several different ways. The simplest is use an estimated dispersion factor to inflate standard errors. Another way is to include a random effect for each household. These kinds of models will be discussed later in the text.

For the first approach, we can estimate a dispersion parameter by dividing the model deviance by its corresponding degrees of freedom. If there is no overdispersion, this estimate should be close
to one. It will be larger than one in the presence of overdispersion. If we choose to use a dis-

persion parameter in our model, the model no longer can be fit using likelihoods, rather it is fit

using a quasilikelihood. We used the likelihood in previous examples and the variance was λ. With
the quasilikelihood approach, we multiply the variance by φ so that the resulting likelihood-like

function is not a bona-fide likelihood. The following output illustrates a use of a model fit using
quasipoisson.

Note that if you are using software that doesn’t make adjustments for overdispersion, you can
adjust the standard errors for the coefficients by multiplying by the square root of the estimated

dispersion parameter. For drop-in-deviance tests, you can divide the test statistic by the estimated
dispersion parameter and compare the result to an $F$ - distribution with the difference in the model
degrees of freedom for the numerator and the degrees of freedom for the larger model in the denomi-

nator.

\[
\text{TS} = 11.822 / (3 \times 2.154)
\]

1.829465

1-pf(TS,3,396)

0.1480795

(Dispersion parameter for quasipoisson family taken to be 2.153649)

In the absence of overdispersion we expect the dispersion parameter estimate to be 1.0. The

estimated dispersion parameter estimated is larger than 1.0 (2.154) indicating overdispersion. and
the larger estimated standard errors reflect the adjustment. The Wald-type test for the quadratic term

(t=-2.147, p=0.03) remains significant. To test for the significance of the Floor variable, we need

to adjust the drop-in-deviance test we did previously by dividing the test statistic, 11.822/3 by the

estimated dispersion parameter, 2.154. This result is then compared to an $F_{3,396}$ for an informal, ap-

proximate. The addition of the Floor no longer is significant (p=0.15).

\[
\text{TS} = 11.822 / (3 \times 2.154)
\]

1.829465

1-pf(TS,3,396)

0.1480795

### Accounting for Overdispersion

Use the estimated dispersion parameter, $\hat{\phi}$ to adjust

- Wald test statistics: multiply the standard errors by $\sqrt{\hat{\phi}}$
- CIs which use the adjusted standard errors and are thereby wider
- Drop-in-deviance test statistic comparing Model 1 (larger model with $p$ parameters) to
  Model 2 (smaller model with $q < p$ parameters) =

\[
\frac{1}{\hat{\phi}} \frac{D_2 - D_1}{p - q}
\]

(4.6)

where $D_1$ and $D_2$ are the residual deviances for models 1 and 2 respectively and $p - q$ is the
difference in the number of parameters for the two models. Note that both $D_1 - D_2$ and $p - q$
are positive. This test statistic is compared to an $F$-distribution with $p - q$ and $n - p_{\text{largest}}$
degrees of freedom.
Note that it was important to attain the correct model prior to considering adjustment for overdispersion. Adjusting prior to adding the quadratic term may lead to the wrong conclusion.

### 4.6.11 Using the Model

In the end, our interest centers on the age at which a female head of the household has the largest number of people living with her. The fitted model with **Floor** fixed at its reference level is \( \hat{Y} = 1.70 + 0.009815X - 0.000235X^2 \). To find at what age (X) the number in the house is a maximum, take the derivative with respect to X, set it equal to 0, and solve for X. This yields an age of 20.89 years of age. This result is consistent with the log means model presented earlier (Figure 4.3).

### 4.6.12 Least Squares Regression vs. Poisson Regression

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Poisson Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Response</strong></td>
<td>Normal</td>
<td>Counts, Poisson</td>
</tr>
<tr>
<td><strong>Variance</strong></td>
<td>Equal for each level of X</td>
<td>The same as the Mean for each level of X</td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>( \mu = \beta_0 + \beta_1X )</td>
<td>( \log(\lambda) = \beta_0 + \beta_1X )</td>
</tr>
<tr>
<td><strong>Fitting</strong></td>
<td>Least Squares</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td><strong>EDA</strong></td>
<td>plot X vs Y, add line</td>
<td>plot ( \log(\hat{y}) ) vs X for each subgroup</td>
</tr>
<tr>
<td><strong>Comparing models</strong></td>
<td>( y_i - \hat{y}_i )</td>
<td>deviance residual</td>
</tr>
<tr>
<td></td>
<td>Sum of squared residuals</td>
<td>Deviance</td>
</tr>
<tr>
<td></td>
<td>( F )-tests</td>
<td>( \chi^2 ) tests</td>
</tr>
<tr>
<td></td>
<td>Modified ( F )-tests</td>
<td>when overdispersed</td>
</tr>
<tr>
<td><strong>Interpreting coefficients</strong></td>
<td>( \beta_1 = \text{change in } \mu ) for unit change in X</td>
<td>( e^{\beta_1} = \text{percent change in } \lambda ) for unit change in X</td>
</tr>
<tr>
<td><strong>Overdispersed variance</strong></td>
<td>( \phi \lambda )</td>
<td></td>
</tr>
</tbody>
</table>

### 4.7 Case Study: Weekend drinking

Sometimes when analyzing Poisson data, you may see many more zeros in your data set than you would expect for a Poisson random variable. For example, an informal survey of students in an introductory statistics course included the question, “How many alcoholic drinks did you consume last weekend?”

This survey was done on a dry campus where no one is allowed alcohol even those of age. We anticipate that some portion of the respondents never drink. We expect these non-drinkers to report zero drinks. There will also be students who are drinkers reporting zero drinks because they just did not happen to drink during the past weekend. Our zeros are a mixture of responses from non-drinkers and drinkers who abstained. Ideally, we’d like to sort out the non-drinkers and drinkers when performing our analysis.

### 4.7.1 Research Question

The purpose of this survey is to explore factors related to drinking behavior on a dry campus. What proportion of students on this dry campus never drink? What factors, off campus living and sex are related to whether students never drink? Among those who do drink, to what extent is moving off campus associated with the number of drinks in a weekend? It is commonly assumed that males alcohol consumption is greater than females. Does our data bear that out among those who drink? Answering these questions would be a simple matter if we knew who was and was not a drinker in
our data set. Unfortunately, the non-drinkers did not identify themselves as such so we will need to use our model to estimate the proportion of drinkers and non-drinkers.

4.7.2 Data Organization

Each line of this file contains data provided by a student in an introductory statistics course. In this analysis, the response of interest is the respondent’s report of the number of alcoholic drinks they consumed the previous weekend, drinks, whether the student lives off campus, off.campus, and sex. Here is a sample of observations from this dataset:

<table>
<thead>
<tr>
<th>off.campus</th>
<th>drinks</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>on</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>on</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>on</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>on</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>on</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>off</td>
<td>4</td>
</tr>
</tbody>
</table>

4.7.3 Exploratory Data Analysis

As always we take stock of the amount of data, here there are 77 observations. Large sample sizes are preferred for the type of model we consider and n=77 is on the small side. We proceed with that in mind. If the data set is large or very large as more and more data sets tend to be, these types of models can be very helpful. A premise of this analysis is that we believe that those responding zero are coming from a mixture of non-drinkers and drinkers who abstained the weekend of the survey.

- **Non-drinkers**: respondents who never drink and would always reply with zero
- **Drinkers**: obviously this includes those responding with one or more drinks, but it also includes people who are drinkers but did not happen to imbibe the past weekend. These people reply zero, but are not counted among the non-drinkers.

Beginning the EDA with the response, number of drinks (drinks), we find that over 46% of the students reported no drinks during the past weekend. Figure 4.5 is a graph portraying the observed number of drinks reported by the students. The mean of the number of drinks reported is 2.013 drinks the past weekend. The data set consists of 74% females, 26% males. With respect to residence, only 9% live off campus.
Because our response, drinks, is a count it is natural to consider a Poisson regression model. You may recall that a Poisson distribution has only one parameter, its mean, \( \lambda \), which also equals the variance. Here we will have an additional parameter, \( \alpha \). We define \( \alpha \) to be the true proportion of the population that are non-drinkers.

The next step in the EDA is especially helpful if you suspect your data contains excess zeros. Figure ?? is what we might expect to see under a Poisson model. They represent the probabilities for a Poisson distribution (using the Poisson probability formula) with \( \lambda \) equal to the mean of the observed number of drinks, 2.013 drinks per weekend. It is an approximation to what the true distribution of the number of drinks might be for the student drinkers. Comparing this Poisson distribution to what we observed (Figure 4.5), it is clear that lots more zeros have been reported by the students than you would expect to see if the survey observations were coming from a Poisson distribution. This doesn’t surprise us because we had expected a subset of the survey respondents to be non-drinkers, i.e. they would not be included in this Poisson process. This circumstance actually arises in many Poisson regression settings. By tweaking our Poisson regression model, we can get an estimate for the proportion of true zeros(\( \alpha \)), here non-drinkers, as well as estimates of the effect of covariates on the Poisson distribution, here sex. This type of model is referred to as a zero-inflated model or zip model.

4.7.4 Modeling

Our model consists of two parts:

1. One part models the association of the number of drinks among drinkers with predictors sex and off camp.
2. The other part uses the predictors sex and off camp to obtain an estimate of the proportion of non-drinkers based on the reported zeros.

The form for each part of the model follows. The first part looks like an ordinary Poisson regression model:

\[
\log(\lambda) = \beta_0 + \beta_1(\text{off.campus}) + \beta_2(\text{sex})
\]

where \( \lambda \) is the mean number of drinks in a weekend among those who drink. The second part has the form

\[
\logit(\alpha) = \beta_0 + \beta_1(\text{off.campus}) + \beta_2(\text{sex})
\]

where \( \alpha \) is the probability of being in the non-drinkers group and \( \logit(\alpha) = \log(\alpha/(1 - \alpha)) \). We’ll provide more detail on the logit in the Logistic Regression chapter. There are many ways in which to structure this model, here we use both of our predictors to help in modeling who is and is not a drinker. It is not required to have all of the predictors for this part of the model - in fact, it is perfectly OK to have no predictors at all.

4.7.5 Fitting a ZIP Model

How is it possible to fit such a model? We cannot observed whether a respondent is a drinker or not (which probably would’ve been good to ask). The ZIP models are a special case of a more general type of statistical model referred to as a latent variable model. More specifically, it is a type of a mixture model where observations for one or more groups occur together and the group is unknown. Likelihood methods are at the core of this methodology, but fitting is an iterative process where it is necessary to start out with some guesses (or starting values). We have not done enough here to do any sophisticated analysis, but we do want you to know that these kinds of models exist and if you find yourself one day needing to use them, you can work with a statistician.

Optional. The following is a heuristic for fitting a zero-inflated model. Imagine that the graph
of the Poisson distribution Figure ?? is removed from the observed data distribution in Figure 4.5. Some zero responses will remain. These would correspond to non-drinkers and the proportion of all observations these zeros constitute might make a reasonable estimate for \( \alpha \), the proportion of non-drinkers. The likelihood is used and some iterating in the fitting process is involved because the Poisson distribution in Figure ?? is based on the mean of the observed data and we know that is not likely to be exact, i.e. it is not the average among drinkers. Furthermore, the likelihood incorporates the predictors, sex and off campus. So there is a little more to it than computing the proportion of zeros, but this heuristic should provide you a general idea of how these kinds of models are fit. Various software packages will fit these kinds of models. For example, the software package R along with the library(pscl) will fit a zip model for us. The command is zeroinfl(y x). Part of the output for our model appears below.

```r
> zip.m2=zeroinfl(drinks~off.campus+sex|off.campus+sex)
```

Count model coefficients (Poisson with log link):

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 0.7703 | 0.1439 | 5.354 | 8.60e-08 *** |
| off.campusTRUE | 0.3875 | 0.2109 | 1.838 | 0.0661 . |
| sexm | 1.0111 | 0.1762 | 5.737 | 9.63e-09 *** |

Zero-inflation model coefficients (binomial with logit link):

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | -0.1094 | 0.3262 | -0.335 | 0.737 |
| off.campusTRUE | -1.9309 | 1.4699 | -1.314 | 0.189 |
| sexm | -0.2553 | 0.5622 | -0.454 | 0.650 |

Log-likelihood: -140.9 on 6 Df

The top panel of this output that is labeled “Count Model” provides information on how the co-variates sex and off campus are related to the number of drinks among drinkers. As we have done with previous Poisson regression models, we exponentiate the coefficient for ease of interpretation, \( e^{0.1011} = 2.75 \). Thus, for those who drink, the number of drinks for males are 2.43 times the number for females (p<0.000001) given you are comparing people who live in comparable settings, i.e. either both on or both off campus. Among drinkers, the mean number of drinks for those living off campus is \( e^{0.3875} = 1.47 \) times that of those living on campus for those of the same sex (p=0.0661).

The second panel of the output is labeled “Zero-inflation model” and refers to separating drinkers from non-drinkers. We have

\[
\log(\alpha/(1-\alpha))) = -0.1094 - 1.9309(\text{off.campus}) - 0.2553(\text{male})
\]

However we are interested in \( \alpha \), the proportion of non-drinkers. With a little algebra (solving the equation with \( \log(\alpha/(1-\alpha)) \) for \( \alpha \), we have

\[
\hat{\alpha} = \frac{\exp(-0.1094-1.9309(\text{off.campus})-0.2553(\text{male}))}{1 + \exp(-0.1094-1.9309(\text{off.campus})-0.2553(\text{male}))}.
\]

For example, the chance that a female(male=0) who lives on campus(off.campus=0) is a non-drinker is

\[
e^{-0.1094} = 0.473
\]
or 47.3%. If you have seen logistic regression, this transformation is what is used to estimate a probability. More on this in the logistic regression chapter.
4.7.6 Comparing the ordinary Poisson regression model to the ZIP model

What kind of gain is there in using the more complex ZIP model? Is it really so much better than an ordinary Poisson regression model? One consideration is that estimating the proportion of the population belonging to the true zero group may be of interest. Here, we did want an estimate of the proportion of the students that are not drinkers. We can fit an ordinary Poisson regression model and compare it to the the zero-inflation model.

```r
> pois=glm(drinks~off.campus+sex,family=Poisson)
> summary(pois)
```

Coefficients:

|                 | Estimate | Std. Error | z value | Pr(>|z|) |
|-----------------|----------|------------|---------|----------|
| (Intercept)     | 0.1293   | 0.1241     | 1.041   | 0.298    |
| off.campusTRUE  | 0.8976   | 0.2008     | 4.470   | 7.83e-06 *** |
| sexm            | 1.1154   | 0.1611     | 6.925   | 4.36e-12 *** |

Null deviance: 294.54 on 76 degrees of freedom
Residual deviance: 230.54 on 74 degrees of freedom
AIC: 357.09

We cannot use the drop-in-deviance test we discussed earlier because these two models are not nested within one another. Vuong (1989) devised a test to make this comparison for the special case of comparing a zero-inflated model and ordinary regression model.

```r
> vuong(pois.m1,zip.m2)
```

Vuong Non-Nested Hypothesis Test-Statistic: -2.708384
(test-statistic is asymptotically distributed N(0,1) under the null that the models are indistinguishible)

in this case:
model2 > model1, with p-value 0.003380586

Here, we have structured the Vuong Test to compare Model 1: Ordinary Poisson Model to Model2: Zero-inflation Model. If the two models do not differ, the test statistic for Vuong would be asymptotically standard Normal and the p-value would be relatively large. Here the last line of the output indicates that the zero-inflation model is better with p<0.0034. Note that the test depends upon sufficiently large n for the Normal approximation. As we have already pointed out, the sample size is quite small here (n=77) and we need to interpret this result with caution.

4.7.7 Residual Plot

Fitted values ($\hat{y}$) and residuals can be computed for zero-inflation models and plotted. Here the plot (Figure 4.6) reveals that one observation appears to be extreme (Y=22 drinks during the past weekend). Is this a legitimate observation or was there a transcribing error? Without the original respondents we cannot settle this question. It might be worthwhile to get a sense of how influential this extreme observation is when fitting the model. Removing Y=22 and refitting the model can provide some idea about this.

In the exercises you will find application of ZIP modeling. One example seeks to determine the number of fish each group of visitors caught at a state park. There is not data, however, on whether a group was fishing or not. So a report of zero for a group could mean that no one in group was fishing or a report of 0 could come from some disgruntled fishermen who tried but did not catch any fish.
4.7.8 Caveats and Extensions

Given that you have progressed this far in your statistical education, the weekend drinking survey question should raise some red flags. What time period constitutes the “weekend” - will some students be thinking of only Saturday night while others include Friday night or possibly Sunday evening? What constitutes a drink - a bottle of beer? How many drinks will a respondent report for a bottle of wine? Precise definitions would vastly improve the quality of this data. Even better, a much larger sample would be desirable. There is also an issue related to confidentiality - if the data is collected in class, will the teacher be able to identify the respondent? What if it is a dry campus? Will respondents worry that a particular response will affect their grade in the class?

In addition to these concerns, there are a number of other caveats that should be noted here. Following the concern of whether this data represents a random sample of any population (it doesn’t), we also must be concerned with the size of this data set. ZIP models are not appropriate for small samples and this data set is not impressively large.

At times, a mixture of zeros occurs naturally. It may not come about because of neglecting to ask a critical question on a survey, but the information to the subpopulation may simply not be ascertainable. For example, a genetic predisposition that affects disease onset may require extensive, invasive data collection whereas a look at the observable data may reveal the presence of two subgroups. These kinds of circumstances occur often enough that ZIP models are becoming increasingly common. Actually there are many fewer ordinary Poisson regression applications in contrast toZIPs and other Poisson modeling approaches such as hurdle models and quasi-Poisson applications. So it is worth taking a look at these variations of Poisson regression models. Not enough detail is presented here to deal with the nuances of zero-inflated models, but we want you to know about models of this type. ZIP models demonstrate that modeling can be flexible and creative which we hope you see as a theme throughout this book.
4.8 R code

```r
glm(numHouse ~ age, family=Poisson)
```
This code fits a Poisson regression.

```r
exp(model.age$coef[2])
```
This code exponentiates the coefficient for linear term in the previous model.

```r
exp(model.age$coef[3])
```
This code exponentiates the coefficient for squared term term in the quadratic model. In general [2] refers to the first variable, [3] refers to the second and so on.

```r
ageHH2 = age*age
glm(numHouse ~ ageHH + ageHH2 + factor(Floor), family=quasipoisson)
```
This code fits adds a quadratic term and the variable Floor as a factor with an estimated dispersion parameter.

```r
anova(null.model, model.ageHH, model.ageHH2, full.model, test="Chisq")
```
This code produces an Analysis of Deviance table.

```r
zip.m2=zeroinfl(drinks ~ off.campus + sex ~ off.campus + sex)
```
This code fits a zero-inflated model which uses the variables off.campus and sex to estimate the proportion of non-drinkers as well as the rate for drinkers.

```r
vuong(pois.m1, zip.m2)
```
This code provides a comparison of model 1 (an ordinary Poisson model) with model 2 (a zero-inflated model.)

References


A Guide to F-Scale Damage Assessment(April 2003)
U.S. DEPARTMENT OF COMMERCE Donald L. Evans, Secretary
National Oceanic and Atmospheric Administration Vice Admiral Conrad C. Lautenbacher, Jr., Administrator National Weather Service John J. Kelly, Jr., Assistant Administrator
4.9 Exercises

4.9.1 Conceptual Exercises

1. What are features of inferential OLS models that make them less suitable for count data?
2. Models of the form \( Y_i = \beta_0 + \beta_1 X_i, \epsilon \sim iidN(0, \sigma) \) are fit using the method of least squares. What method is used to fit Poisson regression models?
3. What should be done before adjusting for overdispersion?
4. Why is quasilikelihood used and what effect does it have on corresponding models using likelihoods?
5. Why is the log of means, \( \log(\bar{Y}) \), not \( \bar{Y} \), plotted against \( X \) when assessing the assumptions for Poisson regression? How can this assumption be checked if there are not many repeated observations at each level of \( X \)?
6. Is it possible that a predictor is significant for a model fit using a likelihood, but not for a model for the same data fit using a quasilikelihood? Explain.

4.9.2 Properties of Poisson random variables

For the following exercises, refer to the properties of poisson random variables in section 4.2.2. Write out each in the context of the following studies. Why might a zero-inflated model be appropriate for the study?

Example: Fish or Poisson
A state wildlife biologist collected data from 250 park visitors as they left at the end of their stay. Each were asked to report the number of fish they caught during their one week stay. On average visitors caught 21.5 fish per week.

Solution:

1. \( Y = \) the number of fish caught by a visitor during a one week stay
2. The possible values are 0 fish, 1 fish, 2 fish, \( \ldots \) with no maximum.
3. \( \lambda \) is the true mean number of fish per visitor per week. The true mean can be estimated, \( \hat{\lambda} \), with the observed mean 21.5 fish per week.
4. The variation in the number of fish caught is also equal to \( \lambda \) and estimated to be 21.5.
5. To estimate the probability for a number of fish per week:

\[
P(Y = y) = \frac{e^{21.5^y}}{y!}
\]

6. The sum of all the fish caught by the 250 visitors is also a Poisson random variable, with mean = 250*21.5 fish per week.
7. The data may be zero-inflated if the zero responses are a mixture of visitors who do not fish and fishers that just didn’t happen to catch any fish.

Methadone Program Recidivism
Program facilitators keep track of the number of times their program’s patients return for more treatment or relapse within a year of initial treatment. Data on 100
patients yielded a mean number of 2.8 relapses per patient within the year of initial treatment.

**Clutch size** Nests were located and the number of eggs in each nest for the start of a season were counted. These counts were made following a particularly destructive storm and the mean number of eggs in 30 nests was only 1.7 eggs per nest.

### 4.9.3 Poisson regression

1. **Credit card use** A survey of 1,000 consumers asked respondents how many credit cards they use. Interest centers on the relationship between credit card use and income in 10,000. The estimated coefficient for income is 2.1 cards per respondent.
   
   (a) Identify the predictor and interpret the coefficient for the predictor in this context.
   
   (b) Describe how the assumption of linearity can be assessed in this example.

2. **Tornado Damage** - The number of deaths for 100 tornados was recorded along with the **storm rating**: F0, F1, F2, F4 or F5 and **region** of the country: Northeast, South, Midwest, West and Southwest.
   
   (a) Describe how the assumption of mean = variance could be assessed here.
   
   (b) How would the significance of the addition of **region** to a model with **storm rating** be determined? Be specific.

### 4.9.4 Poisson Approximation: Rare Events

For rare diseases, the probability of a case occurring, $p$, in a very large population, $n$, small probability With a small $p$ and large $n$, the random variable $Y$ = the number of cases out of $n$ people can be approximated using a Poisson random variable with $\lambda = np$. If the count of those with the disease is observed in several different populations independently of one another, the $Y_i$ represents the number of cases in the $i^{th}$ population and can be approximated using a Poisson random variable with $\lambda_i = n_i p_i$ where $p_i$ is the probability of a case for the $i^{th}$ population. Poisson regression can take into account the differences in the population sizes, $n_i$, using as an offset which is the log($n_i$). The coefficient of the offset is set at one, it is not estimated like the other coefficients. Thus the model statement has the form:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i + \log(n_i)$$

where $Y_i \sim \text{Poisson}(n_i \lambda_i)$. Note that $\lambda_i$ depends on $x_i$ which may differ for the different populations.

For each of the following examples, identify quantities which analogous to the following quantities in identify:

1. A case,
2. The population size, $n$,
3. The random variable $Y$ = the number of events out of $n$ trials,
4. The offset
5. The corresponding Poisson random variables, and
6. The predictor.

#### 4.9.4.1 Skin Cancer: an example.

Data from Scotto, Kopf, and Urbach (1974) reported the number of cases of non melanoma skin cancer for women by age group in two metropolitan areas (Minneapolis-St Paul and Dallas-Ft Worth); the year is unknown. The columns contain: number of cases of skin cancer, population size of the age group per city, age group (1=15-24, 2=25-34, 3=35-44, 4=45-54, 5=55-64, 6=65-74, 7=75-84, 8=85+), and metropolitan area (1=Minneapolis-St Paul, 2=Dallas-Ft Worth).
1. A case of non-melanoma skin cancer,

2. The population sizes vary according to each city-age group. The first few populations consist of 172675, 123065, 29007, and 7538 people.

3. \( Y \) = the number of cases out of each city by age group population,

4. The corresponding Poisson approximations are the number of cases for each city by age group population with \( \lambda_i = n_i p_i \) where \( n_i \) = the number of women in the age group by city population
\( p_i \) = the probability of a woman being a case in the \( i \)th age by city population

5. The predictors are age group and city.

### 4.9.4.2 Campus Crime

Researchers are interested in comparing regional differences in violent crime rates controlling for type of institution (University or College). Each row of this data set contains crime information from a post secondary institution, either a college or university. The variables include:
- **University** = an indicator for college(0) or university(1),
- **numViolent** = the number of violent crimes for that institution for the given year
- **numEnroll** = enrollment at the school.

The first few lines of the data set appear below.

<table>
<thead>
<tr>
<th>Region</th>
<th>University</th>
<th>numEnroll</th>
<th>numViolent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SE</td>
<td>U</td>
<td>5590</td>
</tr>
<tr>
<td>2</td>
<td>SE</td>
<td>C</td>
<td>540</td>
</tr>
<tr>
<td>3</td>
<td>W</td>
<td>U</td>
<td>35747</td>
</tr>
<tr>
<td>4</td>
<td>W</td>
<td>C</td>
<td>28176</td>
</tr>
<tr>
<td>5</td>
<td>SW</td>
<td>U</td>
<td>10568</td>
</tr>
<tr>
<td>6</td>
<td>SW</td>
<td>U</td>
<td>3127</td>
</tr>
</tbody>
</table>

Models that account for differences in the number of trials include an **offset**.

The model statements for the examples above in R look like:

- \( \text{numCases} \sim \text{age + city, family = Poisson, offset = log(ageXcitypop)} \)
- \( \text{numViolentCrimes} \sim \text{University + Region, family = Poisson, offset=log(enrolled)} \)

An offset can also be used to take into account differing surveillance time or space, sometimes referred to varying effort. In the example below the offset will be the log(Time online).

### 4.9.4.3 Dating Online

Researchers are interested in the number of dates respondents arranged on line and whether the rates differ by age group. Questions which illicit responses similar to this can be found in reports from the Pew survey concerning dating on line and relationships (Pew 2013). Each survey respondent was asked how many dates they have arranged on line in the past 3 months as well as the typical
amount of time, $t$, in hours, they spend on line weekly. Some rows of data appear below.

<table>
<thead>
<tr>
<th>Age</th>
<th>Time online</th>
<th>Number of dates arranged on line</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>35</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>38</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>55</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

The model has the form:

$$ \log(\lambda_i) = \beta_0 + \beta_1 x_i + \log(t_i) $$

where the observations, $Y_i$ come from a Poisson distribution with $\lambda_i = t_i p_i$. The corresponding code for R is:

```r
numDates ~ age, family = Poisson, offset = log(Time online)
```

4.9.4.4 Bus Driver Complaints

Source: Joseph Lang, U of Iowa

A bus company manager wishes to determine whether there is any relationship between a sensitivity training program and the number of filed complaints against bus drivers. In particular, the manager wanted to know whether the rate of complaints per driver tends to be lower for bus drivers who had more hours of sensitivity training. The manager was able to obtain records on 100 bus drivers. The records can be found in `busDrivers.csv` and include the following information:

- ID = Bus driver identifier
- DAYS = Days of sensitivity training
- COMPLAINTS = Number of filed complaints
- MONTHS = Number of months the bus driver has been driving.

4.9.5 A 3D Graphic of Poisson Regression

Let $X_i = 5, 10, 15, 20$. Calculate $\lambda_i$ for each level of $X_i$ using the model $\log(\lambda_i) = -1.7 + 0.2 X_i$ or $\lambda_i = e^{(-1.7+0.2 X_i)}$. Generate 1000 observations $Y_i$ for each $X_i$. This provides an estimate of the Poisson distribution at each level of $X_i$. Create a 3D plot with $X = 5, 10, 15, 20, Y =$ generated data (0, 1, 2, ...), and $Z =$ the estimated corresponding probability based on the relative proportions of the generated data. Mark and connect the $\lambda$s on the plot. Examine the scatterplot and assess the LINE assumptions.

4.9.6 Guided Exercises

**Elephant Mating** How does age affect male elephant mating patterns? An article by Poole (1989) investigated whether mating success in male elephants increases with age and whether there is a peak age for mating success. To address this question, the research team followed 41 elephants for one year and recorded both their ages and their number of matings. The data is found in `elephant.csv`, and relevant R code can be found under `elephantMating.R`.

The variables are:

- MATINGS: the number of matings in a given year
- AGE: the age of the elephant in years.

1. Create a histogram of MATINGS. Is there preliminary evidence that number of matings could be modeled as a Poisson response? Explain.
2. Plot MATINGS by AGE. Add a least squares line. Is there evidence that modeling matings using a linear regression with age might not be appropriate? Explain.

3. For each age, calculate the mean number of matings. Take the log of each mean and plot it by AGE.
   (a) What assumption can be assessed with this plot?
   (b) Is there evidence of a quadratic trend on this plot?

4. Fit a Poisson regression model with a linear term for AGE. Interpret the coefficient for AGE. Exponentiate and interpret the result.

5. Construct a 95% confidence interval for the slope. Then compute a 95% confidence interval for the rate of the number of matings for each additional year in age.

6. Are the number of matings significantly related to age? Test with
   (a) a Wald test and
   (b) a drop in deviance test.

7. Add a quadratic term in AGE to determine whether there is a maximum age for the number of matings for elephants. Is a quadratic model preferred to a linear model? To investigate this question, use
   (a) a Wald test and
   (b) a drop in deviance test.

8. What can we say about the goodness of fit of the model with age as the sole predictor? Compare the residual deviance for the linear model to a $\chi^2$ distribution with the residual model degrees of freedom.

9. Fit the linear model using a quasilikelihood. (Why?)
   (a) How do the estimated coefficients change?
   (b) How do the standard errors change?
   (c) What is the estimated dispersion parameter?
   (d) An estimated dispersion parameter greater than one suggests overdispersion. When adjusting for overdispersion, are you more or less likely to obtain a significant result when testing coefficients? Why?

4.9.6.1 Overdispersed Poisson

**Bus Drivers** Source: Joseph Lang, U of Iowa
A bus company manager wishes to determine whether there is any relationship between a sensitivity training program and the number of filed complaints against bus drivers. In particular, the manager wanted to know whether the rate of complaints per driver tends to be lower for bus drivers who had more hours of sensitivity training. The manager was able to obtain records on 100 bus drivers. The records can be found in `busDrivers.csv` and include the following information:
ID = Bus driver identifier
DAYS = Days of sensitivity training
COMPLAINTS = Number of filed complaints
MONTHS = Number of months the bus driver has been driving.

1. Plot a histogram of the number of complaints and comment.

2. The number of months, $M_i$, differs from driver to drive. Fit a Poisson regression model using $\log(M_i)$ as the offset. Note that the model has the form $\log(\lambda_i) = \beta_0 + \beta x_i + \log(M_i)$ where $x_i =$ days of sensitivity training for driver $i$, and $Y_i$'s are assumed to be independent Poisson random variables with means $M_i \lambda_i$, $i = 1, ..., 100$. State your conclusions.
Fit the previous model taking into account overdispersion by using quasilikelihood. That is, consider the model \( \log \lambda_i = \beta_0 + \beta x_i \) where \( x_i = x_i = \text{days of sensitivity training for driver } i \), and \( Y_i \)'s are assumed to be independent with means \( M_i \lambda_i \) and variances \( \phi M_i \lambda_i \), \( i = 1, \ldots, 100 \).

(a) How do the estimated coefficients for these two models compare?

(b) How do the standard errors for the estimated coefficients compare?

(c) Estimate \( \beta_1 \). Exponentiate it and interpret your result.

(d) Construct a confidence interval for \( \beta_1 \), exponentiate the endpoints and interpret your results.

### Campus Crime data
Students like to feel safe and secure when attending a college or university. In response to legislation, the US Department of Education seeks to provide data and reassurances to students and parents alike. All postsecondary institutions that participate in federal student aid programs are required the Campus Crime Statistics Act to collect and report data on crime occurring on campus to the Department of Education. In turn, this data is publicly available on the website of the Office of Postsecondary Education.

#### Research Question: Are there regional differences in violent crime on campus controlling for differences in the type of school, university or college?
Each row of this data set contains crime information from a post secondary institution, either a college or university. The variables include:

- **University** = an indicator for college(0) or university(1),
- **Region** = set of indicators for region (Central, Midwest, Northeast, Southeast, Southwest, and West)
- **numViolent** = the number of violent crimes for that institution for the given year, (the response)
- **numProp** = the number of property crimes,
- **numOverall** = the overall number of crimes, and
- **numEnroll** = enrollment at the school.

<table>
<thead>
<tr>
<th>Region</th>
<th>University</th>
<th>numEnroll</th>
<th>numProp</th>
<th>numViolent</th>
<th>numOverall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SE</td>
<td>U</td>
<td>5590</td>
<td>266</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>SE</td>
<td>C</td>
<td>540</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>W</td>
<td>U</td>
<td>35747</td>
<td>1233</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>W</td>
<td>C</td>
<td>28176</td>
<td>210</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>SW</td>
<td>U</td>
<td>10568</td>
<td>116</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>SW</td>
<td>U</td>
<td>3127</td>
<td>29</td>
<td>0</td>
</tr>
</tbody>
</table>

1. Plot the number of violent crimes. Does it appear to be a Poisson distribution? Why or why not?
2. Create a box plot for the number of violent crimes by region and type of institution. Comment.
3. Check the mean=variance assumption by calculating the mean and variance for each region by type combination. Does the assumption appear to be valid? What is unusual about Colleges in the Southwest?
4. The counts are not directly comparable because they come from different size schools. We cannot compare the 30 violent crimes from the first school in the data set to no violent crimes for the second school when their enrollments are vastly different; 5,590 for school 1 versus 540 for school 2. Fit a Poisson regression model with an interaction between region and type of institution. Include an offset for the number enrolled in each institution.
5. Note the huge standard error for the Colleges in the Southwest term. When the standard errors become unusually large, it suggests a problem with fitting a model with that term. A look back at the calculation of the means and variances provides a clue as to why this happened.
6. The box plots indicated quite different patterns for the colleges and universities. Because interest
EXERCISES

centers on the differences between regions, it may be best to fit separate models; one for colleges the other for universities.
7. Fit a Poisson regression model with the predictor Region and the offset, log(Enrolled) for colleges.
8. Check the mean=variance assumption for colleges.
9. Fit a Poisson regression model with the predictor Region and log(Enrolled) for colleges using a quasilikelihood.
10. What is the estimated dispersion parameter? How does it impact the standard errors for the previous model?
11. Estimate the effect for each region for colleges.
12. Repeat the analysis with universities.

4.9.7 Open Ended

Crab Satellites
Brockmann et al (Ethology 1996) carried out a study of nesting female horseshoe crabs. Female horseshoe crabs often have male crabs attached to a female’s nest known as satellites. One objective of the study was to determine which characteristics of the female were associated with the number satellites. Of particular interest is the relationship between the width of the female carapace and satellites.
The data can be found in crab.csv It includes:
NumSat = number of satellites
Width = carapace width
Wt = weight
Sp = spine condition and
C = color.
Use Poisson regression to investigate the research question. Be sure you work to obtain an appropriate model before considering overdispersion.

Doctor Visits Data on doctor visits on a sample of 5,190 from the 19771978 Australian Health Survey. The study seeks to explain the variation in doctor visits using one or more explanatory variables. The data is can be found in doctorVisits.csv. It contains the following variables:
visits Number of doctor visits in past 2 weeks.
gender Factor indicating gender.

age Age in years divided by 100.
income Annual income in tens of thousands of dollars.
ilness Number of illnesses in past 2 weeks.

reduced Number of days of reduced activity in past 2 weeks due to illness or injury. Larger scores are associated with greater reduction.

health General health questionnaire score using Goldberg’s method. Larger scores associated with poorer health.

private Factor. Does the individual have private health insurance?
freepoor Factor. Does the individual have free government health insurance due to low income?

freerepat Factor. Does the individual have free government health insurance due to old age, disability or veteran status?

nchronic Factor. Is there a chronic condition not limiting activity?

lchronic Factor. Is there a chronic condition limiting activity?

Explore the use of a zero-inflated model for this data. Begin with a histogram of the number of visits, some EDA and fitting several models. Summarize your results.

4.9.8 Other exercises

Fish
Child Asthma - Stata example
Children Ever Born G. Rodriguez. Revised September, 2007
Male Bull Frogs ZIP? pg 668 Sleuth
Caesarean Sections - data available
Learning Objectives

- Determine if a probability distribution can be expressed in one-parameter exponential family form.
- Identify canonical links for distributions of one parameter exponential family form.

5.1 One parameter exponential families

Thus far, we have expanded our repertoire of models from OLS to include Poisson regression. But in the early 1970s Nelder and Wedderburn identified a broader class of models that generalizes the multiple linear regression we considered in the introductory chapter and are referred to as generalized linear models (GLMs). GLMs have similar forms for the likelihoods, MLEs, and variance. This makes it easier to find model estimates and their corresponding uncertainty. To determine whether a model is a GLM, we consider the following properties. When a probability formula can be written in the form below

\[ f(y; \theta) = \exp(a(y)b(\theta) + c(\theta) + d(y)) \] (5.1)

and if the support (the set of possible values) does not depend upon \( \theta \), it is said to have a one-parameter exponential family form. We demonstrate that the Poisson distribution is a member of the one parameter exponential family by writing its probability mass function (pmf) in the form of Equation 5.1 and assessing its support.

5.1.1 One Parameter Exponential Family: Possion

Recall we begin with

\[ P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!} \text{ where } y = 0, 1, 2 \ldots \infty \]

A useful identity when attempting to establish exponential form is:

\[ \lambda^y = e^{y \log(\lambda)} \]

Determining whether the Poisson model is a member of the one-parameter exponential family is a matter of writing the Poisson pmf in the form of Equation 5.1 and checking that the support does not depend upon \( \lambda \). First, consider the condition concerning the support of the distribution. The set of possible values for any Poisson random variable is \( y = 0, 1, 2 \ldots \infty \) which does not depend on \( \lambda \). The support condition is met. Now we see if we can rewrite the pmf in one-parameter exponential family form.

\[ P(Y = y) = e^{-\lambda} e^{y \log(\lambda)} e^{-\log(y!)} \]
The exponent for Equation 5.1 must be the product of two factors, one solely a function of \( y \), \( a(y) \) and another, \( b(\lambda) \), a function of \( \lambda \) only. We must check the other two terms in the exponent to confirm that the Poisson distribution is in fact from the one-parameter exponential family. The middle term in the exponent must be a function of \( \lambda \) only, no \( y \)'s should appear. The last term has only \( y \)'s and no \( \lambda \). Since this appears to be the case here we can identify the different functions in this form.

\[
\begin{align*}
a(y) &= y \\
b(\lambda) &= \log(\lambda) \\
c(\lambda) &= -\lambda \\
d(y) &= -\log(y!) \\
\end{align*}
\]  

These functions have useful interpretations in statistical theory. We won’t be going in to this in detail, but we will note that function \( b(\lambda) \) or more generally \( b(\theta) \) will be particularly helpful in GLMs. The function \( b(\theta) \) is referred to as the **canonical link**. The canonical link is often a good choice to model as a linear function of the explanatory variables. In fact, there is a distinct advantage modeling the canonical link as opposed to other functions of \( \theta \), but it is also worth noting that other choices are possible and at times preferred depending upon the context of the application.

We’ll find that other distributions are members of the one parameter exponential family by writing their pdf or pmf in this manner and verifying the support condition. So for example, we’ll see that the binomial distribution meets these conditions so it is a member of the one parameter exponential family. The normal distribution is a special case. We note that we have two parameters, a mean and standard deviation. If it is possible, however, to assume that the standard deviation is known, we can show that a normal random variable is from a one parameter exponential family.

### 5.1.2 One parameter exponential family: Normal

Here we determine whether a normal distribution is a one parameter exponential family member. First we will need to assume that \( \sigma \) is known. Next, possible values for a normal random variable range from \(-\infty\) to \(\infty\), so we have no problems with the support. Finally, we’ll need to write the probability density function (pdf) in the one parameter exponential family form. We start with the familiar form:

\[
f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}
\]

Even writing \( 1/\sigma\sqrt{2\pi} \) as \( e^{-\log(\sigma)-\log(2\pi)/2} \) we still do not have the pdf written in one parameter exponential family form. We will first need to expand the exponent so that we have

\[
f(y) = e^{-\log(\sigma)-\log(2\pi)/2} e^{-\frac{y^2-2\mu+y^2}{2\sigma^2}}
\]

Assuming \( \sigma \) is known, we have

\[
f(y) \propto e^{-2\mu+y^2}
\]

From this result, we can see that the canonical link for a normal response is \( \mu \) which is consistent with what we’ve been doing with OLS. The simple linear regression model has the form:

\[
\mu_{Y|X} = \beta_0 + \beta_1 X.
\]
There are multiple benefits of identifying a response as being from a one parameter exponential family. Wedderburn’s observation has unearthed a unifying theory for regression modeling. By identifying the one parameter exponential family form, we can identify the function of the parameter (the canonical link) that can be modeled by a linear combination of the predictors. We can also use the one parameter exponential family form to determine the expected value, \( E(Y|X) \) and standard deviation of \( Y|X \). With statistical theory you can show that

\[
\mu = \frac{c'(\theta)}{b'(\theta)} \quad \text{and} \quad \sigma^2 = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}
\]

where differentiation is with respect to \( \theta \).

Verifying these results for the Poisson response:

\[
\mu = \frac{-1}{\lambda} \quad \text{and} \quad \sigma^2 = \frac{1/\lambda^2}{(1/\lambda)^3} = \lambda \quad \text{and} \quad = \lambda
\]

Verifying this result for the normal distribution is left as an exercise.

**Useful Facts for Writing Exponential Family Forms**

\[
a = e^{\log(a)}
\]

\[
a^x = e^{x\log(a)}
\]

\[
\log(ab) = \log(a) + \log(b)
\]

\[
\log\frac{a}{b} = \log(a) - \log(b)
\]

**5.2 Generalized Linear Modeling**

GLM theory suggests that the canonical link can be modeled as a linear combination of the explanatory variable(s). This approach unifies a number of modeling results used throughout the text. One example is that likelihoods will be used to compare models in the same way for parameter exponential family members.

We have now **generalized** our modeling to handle non-normal responses. For example, in addition to normally distributed responses, we are able to handle Poisson responses, binomial responses, and more. Writing a pmf or pdf for a response in one parameter exponential family form reveals the canonical link which can be modeled as a linear function of the predictors. This linear function of the predictors is the last piece of the puzzle for performing generalized linear modeling. But, in fact, it is really nothing new. We have used linear combinations when modeling normally distributed data.
### Three Components of a GLM

1. Distribution of Y (eg. Poisson)
2. Link Function (a function of the parameter, eg. log\(\lambda\) for Poisson)
3. Linear Predictor (choice of predictors, eg. \(\beta_1, \beta_2\))

In the chapter on Poisson modeling, we provided heuristic rationale for using the log() function as our link. That is, counts would be non-negative but a linear function inevitably goes negative. By taking the logarithm of our parameter \(\lambda\) we could use a linear predictor and not worry that it can take on negative values. Now we have theoretical justification for this choice, the log is the canonical link for poisson data. In the next chapter we encounter yet another type of response, a binary response, which calls for a different link function. Our work here suggests that we will model a logit(p) using a linear predictor.

Completing Table ?? is left as an exercise.

Note that *generalized linear models (GLMs)* differs from *General Linear Models*.\(^1\)

### 5.3 Exercises

1. For each distribution,
   - Write the pdf in one parameter exponential form, if possible.
   - Describe an example of a setting where this random variable might be used.
   - Identify the canonical link function, and
   - Compute \(\mu = -\frac{c'(\theta)}{b'(\theta)}\) and \(\sigma^2 = \frac{b''(\theta)c''(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}\)

   (a) Binary
      
      Y = 1 for a success, 0 for a failure

      \[ p(y) = p^y(1 - p)^{(1-y)} \]

---

\(^1\) The *general linear model* is a statistical linear model with multivariate vectors as responses. For example, each subject in a study may have their height, weight, and shoe size recorded and modeled as a function of age and sex. The response is a vector, \(Y = (\text{height}, \text{weight}, \text{shoe size})\), for each study participant. Age and sex are explanatory variables in the model. The residual is usually assumed to follow a multivariate normal distribution. If the residual is not a multivariate normal distribution, generalized linear models may be used to relax assumptions about \(Y\) and \(U\).
(b) Binomial
Y= number of successes in n independent, identical trials
\[ p(y) = \binom{n}{y} p^y (1-p)^{1-y} \]

(c) Poisson
Y=number of events occurring in a given time (or space) when the average event rate is \( \lambda \) per
time (or space)
\[ P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!} \]

(d) Normal
\[ f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \]

(e) Exponential
Y=time spent waiting for the first event for a process with an average rate of waiting 1/\( \lambda \) per
event
\[ f(y) = \lambda e^{-\lambda y} \]

(f) Gamma
Y=time spent waiting for the \( r^{th} \) event for a process with an average rate of waiting 1/\( \lambda \) per
event
\[ f(y; \lambda) = \frac{\lambda^r y^{r-1} e^{-\lambda y}}{\Gamma(r)} \]

(g) Geometric
Y=number of trials up to and including the first success
\[ p(y) = (1-p)^{(y-1)} p \]

(h) Negative Binomial (for fixed \( r \))
Y=number of trials up to and including the \( r^{th} \) success
\[ p(y) = \binom{y-1}{r-1} (1-p)^{(y-r)} p^r \]

(i) Pareto
\[ f(y) = \frac{\theta k^\theta}{y^{\theta+1}} \]
\[ y \geq k; \theta \geq 1 \]

2. Complete Table 5.1 containing your results of the preceding exercises.

5.4 References cited
Part 3, p. 370.
6.1 Learning Objectives

- Identify a binomial random variable and assess the validity of the binomial assumptions.
- Write the binomial probability mass function in one parameter exponential family form and identify the canonical link.
- Describe how to determine the maximum likelihood estimate (MLE) for $p$ using the likelihood.
- Write a generalized linear model for binomial responses in two forms, one as a function of the logit and one as a function of $p$.
- Explain how fitting a logistic regression differs from fitting an ordinary least squares (OLS) regression model.
- Interpret estimated coefficients.
- Use the residual deviance to compare models.
- Use the residual deviance to test for lack-of-fit when appropriate.
- Use residual deviances to check for unusual observations or needed transformations.

6.2 Introduction

6.2.1 Binary Responses

A binary response takes on only two values; for example, success ($Y=1$) or failure ($Y=0$), Yes ($Y=1$) or No ($Y=0$). Binary responses are ubiquitous. In fact, binary responses are one of the most common types of data statisticians encounter. Are students with poor grades more likely to binge drink? (Miller, J. W., et al. (2007)) What is the chance you are accepted into medical school given your GPA and MCAT scores? Does a single mom have a better chance of marrying the baby’s father if she has a boy? (Lundberg et al. 2003) Are students participating in sports in college more or less likely to graduate? Are there differences for males and females or by type of sport? (Stevenson 2010) Is exposure to a particular chemical associated with a cancer diagnosis?

Like earlier chapters, each of these examples calls for predicting a response using one or more explanatory variables. We can explicitly define these responses and the corresponding explanatory variables in the following way:

- $Y =$ a student will or will not binge drink
  $X =$ a measure of academic performance
- $Y =$ accepted to medical school or rejected
  $X =$ MCAT scores and GPA
- $Y =$ a mother marries baby’s father or not
  $X =$ sex of the baby
- $Y =$ a student graduates or not
X = participation in sports, type of sport, and gender

• Y = cancer diagnosis
  X = chemical exposure

### 6.2.2 Binomial Responses: sums of binary responses

The binary responses above can be summed to yield a binomial response. Binomial responses are characterized by the number of identical, independent trials, $n$, and $p$, the probability of success on a given trial. Our objective in modeling binomial responses is to quantify how the probability is associated with the covariates. When modeling a binomial response, responses with common covariate values are assumed to have the same probability of success. For example, consider the following 2 × 2 table with exposure status by cancer diagnosis.

<table>
<thead>
<tr>
<th></th>
<th>Exposed</th>
<th>Unexposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>$Y_e$</td>
<td>$Y_u$</td>
</tr>
<tr>
<td>Cancer-free</td>
<td>$n_e - Y_e$</td>
<td>$n_u - Y_u$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_e$</td>
<td>$n_u$</td>
</tr>
</tbody>
</table>

The number with cancer diagnoses among those who are exposed, $Y_e$, could be modeled using a binomial random variable $Y_e$ = the number with cancer diagnosis out of $n_e$ exposed. The variable $Y_e$ could be defined in a similar manner. Each of the $n_e$ exposed are assumed to have the same probability of cancer among the exposed, $p_e$. Likewise, each of the $n_u$ unexposed is thought to have had the same probability of cancer, $p_u$. A summary of features of a binomial random variable is given below.

<table>
<thead>
<tr>
<th>Binomial Random Variables Properties of a Binomial random variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. $Y$= number of successes out of $n$ independent, identical trials</td>
</tr>
<tr>
<td>2. $p$ = the probability of success on a single trial</td>
</tr>
<tr>
<td>3. Possible values: 0, 1, 2, ...n</td>
</tr>
<tr>
<td>4. Mean = $np$</td>
</tr>
<tr>
<td>5. Variance = $np(1 - p)$ implies a standard deviation = $\sqrt{np(1 - p)}$</td>
</tr>
<tr>
<td>6. $P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}$</td>
</tr>
</tbody>
</table>

For binary outcomes $n = 1$. The parameter of interest is $p$. Once $n$ and $p$ are known the variance is specified $np(1 - p)$. Like the Poisson, the variance changes as $p$ changes but it takes it maximum at $p = 0.5$, it does not always increase like Poisson random variables. Nonetheless, we are no longer bound by the equal variance assumption of OLS inference.

### 6.2.3 Bernoulli Process

A Bernoulli process is a sequence of independent trials each having the same probability of success, $p$. A binary random variable assigns 0 (failure) or 1 (success) A binomial random variable is created by summing $n$ trials. Key facts are that the trials are independent and the probability of success is the same.

### 6.2.4 An Example: Binge Drinking

Here we demonstrate how a probability can be calculated for a binomial random variable and define on the binomial assumptions. Suppose we know that among a college aged population the proba-
A MODELING FRAMEWORK

The probability of binge drinking is 0.05 and n = 10 students are randomly selected. What is the probability that at least one student selected binge drinks. With the equation above, we can calculate the probability of the any number of students from 0 to n.

\[
P(Y = y) = \binom{10}{y} 0.05^y 0.95^{10-y}
\]

Suppose we’d like to determine the probability of obtaining at least one binge drinker in our sample of 10.

To calculate the \(P(Y \geq 1)\), compute \(1 - P(Y = 0)\):

\[
P(Y \geq 1) = 1 - \binom{10}{0} \cdot 0.05^0 \cdot 0.95^{10} = 0.40
\]

The assumption for the binomial response is that the trials (sample selections) are independent of one another. Because the selections were made at random, there is no reason to believe they will not be independent. If instead different fraternities were selected and members of the fraternity questioned about their drinking habits, responses are not likely to be independent of one another.

6.2.5 A Graphical Look at Binomial Regression

Figure 6.1 illustrates a comparison of an ordinary least squares (OLS) used for inference to a regression for a binomial variable.

1. The graphic displaying OLS inferential model appears in the left panel of Figure 6.1. It shows that for each level of X, the responses appear to be approximately normal.

2. The panel on the right side of Figure 6.1 depicts what a binomial regression model looks like. For each level of X, the responses follow a binomial distribution.

3. In the case of OLS, the mean responses for each level of X, \(\mu_{Y|X}\), fall on a line. In the case of the binomial data, the mean values of Y at each level of X, \(np_{Y|X}\), fall on a curve, not a line.

4. The variation in Y at each level of X for the OLS example, \(\sigma^2_{Y|X}\), is the same. The responses at each level of X become more variable with increasing means. This is to be expected given that the in the case of binomial regression the variance changes with the mean, \(np_{Y|X} \cdot (1 - p_{Y|X})\).

5. For binomial regression small values of \(np_{Y|X}\), are associated with a distribution that is noticeably skewed. As \(np_{Y|X}\) increases the distribution of the responses begins to look more and more like a normal distribution. For the remainder of the chapter, we drop the \(Y|X\) subscripts but keep in mind that when modeling that the probability of success differs for different values of X.

6.3 A Modeling Framework

In the past with OLS, you have used a line to model the average response, \(\mu_Y\), as a linear function of an explanatory variable, X. Because the objective of modeling a binomial response is to associate the probability of success, \(p\), with the covariates, we may be tempted to try a linear model for the average binary response, \(p\).

\[p = \beta_0 + \beta_1 X \quad \text{This doesn’t work well for binomial data.}\]

Describing the parameter, \(p\), with a line has some serious drawbacks. One problem is that \(p\) should range from 0 to 1. A line, however, is certain to yield estimates for \(p\) that are less than 0 and greater than 1.

One way to avoid this problem is to model the odds instead of \(p\). Odds are calculated by taking
the ratio of the number of successes to the number of failures. Odds can take on non-negative values. The log of the odds is referred to as a logit, will take on values from $-\infty$ to $\infty$ and will be suitable for modeling with a linear function of the predictors.

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 X$$

(6.1)

Models of this form are referred to as binomial regression models, or more generally as logistic regression models. We take a look at these models in the context of a few case studies.

6.4 Case Studies Overview

After the first section of this chapter you should be able to recognize binary and binomial random variables and know the assumptions needed for inference. Next we consider three examples with some real data. The first two examples involve binomial data (Soccer Goalies and Reconstruction in Alabama). The last case uses binary data (Trying to Lose Weight). Here are the statistical concepts you will encounter for each Case Study.

The soccer goalie data can be written in the form of a $2 \times 2$ table. This example is used to describe some of the underlying theory for logistic regression. We demonstrate how binomial probability mass functions (pmfs) can be written in one parameter exponential family form from which we can identify the canonical link. Using the canonical link, we write a GLM for binomial counts and determine corresponding MLEs. Interpretation of the estimated parameter involves a fundamental concept, the odds ratio.

The Railroad Referenda is another binomial example which uses the concepts from the previous section and introduces the notion of deviances. Deviances are used in logistic regression to compare and assess models. We check the assumptions of logistic regression using empirical logit plots and deviance residuals. In addition, under certain circumstances deviances are used to determine if a model exhibits significant lack-of fit.

The last application addresses why teens try to lose weight. Here the response is a binary variable which allows us to analyze individual level data. The analysis builds on concepts from the pre-
<table>
<thead>
<tr>
<th></th>
<th>Behind</th>
<th>Not Behind</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saves</td>
<td>2</td>
<td>39</td>
<td>41</td>
</tr>
<tr>
<td>Scores</td>
<td>22</td>
<td>141</td>
<td>163</td>
</tr>
<tr>
<td>Total</td>
<td>24</td>
<td>180</td>
<td>204</td>
</tr>
</tbody>
</table>

Table 6.1 Soccer goalies’ saves when their team is and is not behind. Source: Roskes et al. 2011 Psychological Science

6.5 GLM Theory for Binomial Outcomes

6.5.1 Case Study: Soccer Goalie Saves

Roskes et al (2011) looked at penalty kicks in the men’s World Cup soccer championship from 1982 - 2010 and found data on 204 penalty shootouts.

6.5.2 Research Question

Does the probability of a save depend upon whether the goalie’s team is behind or not?

6.5.3 Data Organization

The data for this study is summarized in Table 6.1.

As noted in the section on the modeling framework, the odds are the number of successes divided by the number of failures. The log of the odds, the logit, makes a good choice for a linear model because it takes on values from $-\infty$ to $\infty$. Odds are one way to quantify a goalie’s performance. Here the odds that the goalie makes a save when his team is behind is 2 to 22 or 0.09 to 1. Or equivalently, the odds that a goal is scored on a penalty shot is 22 to 2 or 11 to 1. When the goalie’s team is not behind the odds a goal is scored is 141 to 39 or 3.61. We see that the odds of a goal scored on a penalty shot are better when the goalie’s team is behind than when it is not behind. We can compare these odds by calculating the odds ratio (OR), 11/3.61 or 3.05.

In our example, it is also possible to estimate the probability of a goal, $p$, for either circumstance. When the team is behind, we have the probability of a point scored is $p = 22/24$ or 0.833. It is easy to see that when the team is behind the ratio of the probability of a goal scored divided by the probability of no goal is $(22/24)/(2/24)$ or 11, the odds we had calculated above. The same calculation can be made when the goalie’s team is not behind. More generally, when it makes sense to estimate a probability from data, the odds can be written as $p/(1-p)$. But even if it doesn’t make sense to calculate a probability, odds are useful. It is the log of the odds, the logit, that we model in logistic regression. We present some rationale below using GLM theory.

We established that generalized linear models (GLMs) are a way in which to model a variety of different types of responses including binomial responses. In this chapter, we apply the general results of the GLMs to this specific application of binomial responses. Let $Y$ be the number scored out of $n$ penalty shots. The parameter, $p$, is the probability of a score on a penalty shot. Recall that the theory of GLM is based on the unifying notion of the one-parameter exponential family form:

$$f(y; \theta) = e^{a(y)b(\theta) + c(\theta) + d(y)}$$

To see that we can apply the general approach of GLMs to binomial responses, we first write an
expression for the probability of a binomial response and then use a little algebra to rewrite it until we can demonstrate that it, too, can be written in one-parameter exponential family form with \( \theta = p \).

This will provide a way in which to specify the canonical link and the form for the model. Additional theory allows us to read off the mean, standard deviation, and more from this form.

\[
P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}
\]

\[
= e^{y \log p + (n-y) \log (1-p) + \log \left( \frac{n}{y} \right)}
\]  

(6.3)

This pmf is not quite in one parameter exponential family form yet... note that there are two terms in the exponent which consist of a product of functions of \( y \) and \( p \). So more simplification is in order.

Writing the binomial pdf in one parameter exponential family form:

\[
P(Y = y) = e^{y \log \left( \frac{p}{1-p} \right) + n \log (1-p) + \log \left( \frac{n}{y} \right)}
\]  

(6.4)

Don’t forget to consider the support, is the set of possible values for this response dependent upon \( p \)? For any value of \( p \), \( 0 < p < 1 \) all integer values from 0 to \( n \) are possible so the support does not depend on \( p \).

Once the pmf is established to be of a one-parameter exponential family, the form for the model can be determined. The one parameter exponential family form for binomial responses suggests that the canonical link is \( \log \left( \frac{p}{1-p} \right) \) or logit. GLM theory suggests that constructing a model using the logit, the odds, \( \frac{p}{1-p} \), of a score, as a linear function of covariates is a reasonable approach. In our example we could define \( X = 0 \) for not behind and \( X = 1 \) for behind.

\[
\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 X
\]  

(6.5)

So shots taken when the goalie’s team is not behind,

\[
\log \left( \frac{p_0}{1-p_0} \right) = \log \left( \frac{p_0}{1-p_0} \right) = \beta_0
\]

and when the team is behind,

\[
\log \left( \frac{p_1}{1-p_1} \right) = \beta_0 + \beta_1.
\]

We can see that \( \beta_1 \) is the difference between the games behind logit equation and the games not behind logit equation.

\[
\log \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \beta_1.
\]  

(6.6)

Thus \( e^{\beta_1} \) is the ratio of the odds of scoring when the team is not behind compared to scoring when the team is behind. In general, exponentiated coefficients in logistic regression are ratios of odds or odds ratio (OR). A general interpretation of an OR is the odds of success for group A compared to the odds of success for group B.

The logit model can also be re-written in a probability form:
which can be re-written for games behind as

\[ p_1 = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \]

and for games when the goalie’s team is not behind as

\[ p_0 = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \]

We use the likelihood to estimate \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \). As we had done in the likelihood chapter, we can write the likelihood for this example in the following form:

\[ \text{Lik}(p_1, p_0) = p_1^{22} (1 - p_1)^2 p_0^{141} (1 - p_0)^39 \]

Our interest centers on estimating \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \), not \( p_1 \) or \( p_0 \). So we replace \( p_1 \) in the likelihood with an expression for \( p_1 \) in terms of \( \beta_0 \) and \( \beta_1 \) as in Equation 6.5.3. Similarly, \( p_0 \) in Equation 6.5.3 and involves only \( \beta_0 \). The new likelihood looks like:

\[ \text{Lik}(\beta_0, \beta_1) = \left( \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right)^{22} \left( 1 - \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right)^2 \]

\[ \times \left( \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right)^{141} \left( 1 - \frac{e^{\beta_0}}{1 + e^{\beta_0}} \right)^{39} \]  

Now what? Fitting the model means finding estimates of \( \beta_0 \) and \( \beta_1 \). We could consider all sorts of possible combinations of \( \beta_0 \) and \( \beta_1 \). We will let the principle of maximum likelihood guide us in our selection. That is, we will pick that pair of values for \( \beta_0 \) and \( \beta_1 \) that will yield the largest likelihood for our data. Trial and error to accomplish this tedious at best, more efficient numerical methods are used. The MLEs for the coefficients in the soccer goalie study are \( \hat{\beta}_0 = 0.1725 \) and \( \hat{\beta}_1 = 1.1127 \). Exponentiating \( \hat{\beta}_1 \) provides an estimate of the odds ratio (the odds of scoring when the goalie’s team is behind compared to the odds of scoring when the team is not behind) is 3.04 which is consistent with our calculations using the 2 \( \times \) 2 table. The odds of scoring when the goalie’s team is behind is over 3 times that of when the team is not behind.

Before we go on, you may be curious as to why there is no error term in our model statements for logistic or Poisson regression. One way to look at it is to consider that all models describe how observed values are generated. With the logistic model we assume that the observations are generated as a binomial variables. Each observation or realization of \( Y= \) number of successes in \( n \) independent and identical trials with a probability of success on any one trial of \( p \) is produced by \( Y \sim \text{Binomial}(n, p) \). So the randomness in this model is not introduced by an added error term but rather by appealing to a Binomial probability distribution.

### 6.6 Case Study: Reconstructing Alabama

This analysis is an example of a novel application of statistics, reconstruction era demographic and statistical research. You probably are aware that statistics are used in science experiments, surveys, and sports. This next analysis demonstrates how wide ranging applications of statistics can be. Many
would not associate statistics with history but this case study shows that it can be done. US Census data from 1870 helped historian Michael Fitzgerald of St. Olaf College to gain insight into important questions about how railroads were supported during the reconstruction era. In a paper entitled “Reconstructing Alabama: Reconstruction Era Demographic and Statistical Research,” Ben Bayer performs an analysis of data from 1870 to explain influences on voting on referendums related to railroad subsidies. Two theories suggest that positive votes are inversely proportional to the distance a voter is from the proposed railroad. The racial composition of a community as measured by the percentage of blacks, is hypothesized to be associated with voting behavior as well. Separate analyses of three counties in Alabama; Hale, Clarke and Dallas, were performed. We discuss Hale County here and leave the other two as exercises. This example differs from the soccer example in that it includes continuous covaraites.

6.6.1 Research Question
Was voting on railroad referenda related to distance from the proposed railroad line?

6.6.2 Data Collection
US Census data from 1870 for three counties in Alabama; Hale, Clarke and Dallas counties.

6.6.3 Data Organization
The unit of observation for this data is a location or community in a given county. The variables \( \text{percent black} \) and \( \text{distance} \) are the predictors. \( \text{Distance} \) is of primary interest, we control for the racial composition of a community using the variable \( \text{percent black} \).

\( N = 11 \) communities
\( Y_i = \) the number of “Yes” votes in the \( i^{\text{th}} \) community in Hale county
\( N_i = \) total number of in the \( i^{\text{th}} \) votes in a community in Hale county
\( \text{percent black} = \) the percentage of blacks in the community
\( \text{distance} = \) the distance, in miles, the location is from the community

To get a feel for the data, several rows are reproduced below.

<table>
<thead>
<tr>
<th>County</th>
<th>perBlk</th>
<th>distance</th>
<th>Yes (Y)</th>
<th>Num (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carthage</td>
<td>58.4</td>
<td>17</td>
<td>61</td>
<td>110</td>
</tr>
<tr>
<td>Cederville</td>
<td>92.4</td>
<td>7</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Greensboro</td>
<td>59.4</td>
<td>0</td>
<td>1790</td>
<td>1804</td>
</tr>
<tr>
<td>Havana</td>
<td>58.4</td>
<td>12</td>
<td>16</td>
<td>68</td>
</tr>
</tbody>
</table>

Table 6.2: Sample of the data for the Hale County, Alabama railroad subsidy vote.

6.6.4 Exploratory Data Analysis
We use a coded scatterplot to get a look at our data. Figure 6.2 portrays the relationship between \( \text{distance} \) and \( \text{percent black} \) coded by the \textit{in favor} status. All of those communities in favor are 60% or more black. All of those opposed are 7 miles or farther from the proposed line. The overall proportion in favor of the railroad is 88%.

Recall that for a model with two covaraites, the model has the form:

\[
\log(\text{odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.
\]
In logistic regression, we expect the logits, the log(odds), to be a linear function of X, the predictors. To assess the linearity assumption, we construct empirical logit plots. Empirical logits are computed using the log(number of successes/number of failures. See Figure 6.3. Plotting the empirical logits versus distance produces a plot that looks linear. In contrast, the empirical logits by percent black, reveals that Greensboro is quite a distance from the otherwise linear pattern. This suggests that it is possibly an outlier. Greensboro has a 99.2% voting yes with only 60% black. As a side note, it is important to realize that the empirical logits are based on the sample odds, are not what is modeled when performing logistic regression.

In addition to examining how the response correlates with the predictors, it is a good idea to determine whether the predictors correlate with one another. Here the correlation between “distance” and “percent black” is moderately high at -0.5. We’ll watch to see if the correlation affects the stability of our estimated odds ratio estimates.

6.6.5 Modeling Approach

We begin by fitting a logistic regression model with distance alone. Then the covariate percent Black is added. The Wald-type test and the drop-in-deviance provide strong support for the additional of percent Black to the model. The model with distance and percent Black has a large residual deviance suggesting an ill-fitting model. Possible reasons for the lack-of-fit are (1) omitting important covariates, (2) extreme observations or 3) overdispersion. We have used the covariates at hand and it may well be that the model is inadequate but we will proceed nonetheless. A look at the residuals indicates that Greensboro is an extreme observation. Models without Greensboro are fit and compared to our initial models. Seeing no appreciable improvement with Greensboro removed, we leave it in the model. There remains a large residual deviance so we attempt to account for it by using an estimated dispersion parameter. The final model includes distance and adjusts for overdispersion.
6.6.6 Results

6.6.6.1 Models

The first model includes only one covariate, distance.

Model 1: cbind(YesV1, NoV1) ~ distance, family = binomial

|              | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  | 3.33989  | 0.11263    | 29.65   | <2e-16 *** |
| distance     | -0.28933 | 0.01305    | -22.18  | <2e-16 *** |

Residual deviance: 321.18 on 10 degrees of freedom

\[
\exp(\text{coef(model.HaleD)})
\]

(Intercept) distance 28.2159217 0.7487635

The estimated odds ratio for distance, that is the exponentiated coefficient for distance, in this model is 0.75. It can be interpreted as follows: for each additional mile from the proposed railroad, the support declines by a factor of 25%.

The covariate percent black is then added to the first model. Despite the relatively high correlation between the percent Black and distance, the estimated odds ratio for distance remains approximately the same (OR=0.75). The confidence interval for the estimated odds ratio, (0.73, 0.77), is relatively narrow. If the model fit, the residual deviance should follow approximately a \( \chi^2 \)-distribution with \( n - p \) degrees of freedom, here 8. The large residual deviance for Model 2 indicates that it does not fit well.

Model 2: cbind(YesV1, NoV1) ~ perBlk + distance, family = binomial

|              | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  | 4.222021 | 0.296963   | 14.217  | < 2e-16 *** |
| perBlk       | -0.013227| 0.003897   | -3.394  | 0.000688 *** |
| distance     | -0.291735| 0.013100   | -22.270 | <2e-16 *** |

Residual deviance: 307.22 on 8 degrees of freedom

\[
\exp(\text{coef(Model 2)})
\]

(Intercept) perBlk distance 68.1711286 0.9868600 0.7469668

\[
\exp(\text{confint(Model 2)}) \# 95\% CI for exponentiated coefficients
\]

<table>
<thead>
<tr>
<th></th>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>38.2284603 122.6115988</td>
<td></td>
</tr>
<tr>
<td>perBlk</td>
<td>0.9793819 0.9944779</td>
<td></td>
</tr>
<tr>
<td>distance</td>
<td>0.7276167 0.7659900</td>
<td></td>
</tr>
</tbody>
</table>

A model with an interaction between distance and percent black is fit to determine whether the effect of distance differs for differing percent black. The interaction term is significant so it is included in the models that follow.

Model 3: distance + perBlk + perBlk * distance

Coefficients:

|              | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------|----------|------------|---------|----------|
| (Intercept)  |         |            |         |          |
| perBlk       |         |            |         |          |
| distance     |         |            |         |          |
6.6.6.2 Residuals for Binomial Regression

With OLS, residuals were used to assess model assumptions and identify outliers. For binomial regression, two different types of residuals are typically used. One residual, the Pearson residual, has a form similar to that used with OLS. Specifically, the Pearson residual is calculated using:

\[ \frac{Y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1 - \hat{p}_i) x_i}} \]

where \( m_i \) is the number of trials for the \( i^{th} \) observation and \( \hat{p}_i \) is the estimated probability of success for that same observation.

A deviance residual is an alternative residual for binomial regression based on the discrepancy between the observed values and those estimated using the likelihood. A deviance residual can be calculated for each observation and has the form:

\[ d_i = \text{sign}(Y_i - \hat{p}) \sqrt{-2[Y_i \ln(\hat{p}) + (1 - Y_i) \ln(1 - \hat{p})]} \]

When the number of trials is large for all of the observations and the models are appropriate, both sets of residuals should follow a standard normal distribution.

The sum of the residual deviances is referred to as the deviance or residual deviance. The deviance is used to assess the model. As the name suggests, a model with a small deviance is preferred. In the case of binomial regression, when the denominators, \( m_i \), are large and a model fits, the residual deviance follows a \( \chi^2 \) with \( n - p \) degrees of freedom (the residual degrees of freedom). Thus for a good fitting model the residual deviance should be approximately equal to its corresponding degrees of freedom. When binomial data meets these conditions, the deviance can be used for a goodness-of-fit test. The p-value for a lack-of-fit test is the proportion of values from a \( \chi^2 \) with \( n - p \) that are greater than the observed residual deviance.

Here, although we have included all of the covariates at hand as well as an interaction term, the residual deviance for this model remains large (274.23 with 7 df). Either the model is inadequate (missing important covariates) or there is an extreme observation(s) or there is overdispersion. Given the complexity of the response it is very likely that we are missing important covariates. For pedagogical purposes, we proceed assuming the model is adequate. To look for extreme observations, we examine the residuals.

The residuals for Model 3 are plotted against the fitted values in Figure 6.4. This kind of plot for binomial regression always produces two linear trends with slope -1. While we do observed a line with a slope of -1, we do not see two pronounced lines due to the small number of observations. See Figure 6.4. As we suspected from the EDA, the plot does suggest that Greensboro is an outlier. We will remove Greensboro and examine the impact that it has on the estimated coefficients. The residual plots will also be examined of clusters of residuals which may suggest a problem with the independence assumption. The extreme value for Greensboro makes the remaining observations appear to cluster. A single cluster is not a concern, it is when there exist several clusters that may
indicate a reason for a large deviance.

Figure 6.4 Fitted values by Residuals for the model with percent black and distance and interaction of percent black and distance from the proposed line for the Railroad Referendum data. Source: Bayer and Fitzgerald.

Once the suspected outlier is removed and the model refit (Model 4), a re-analysis results in a change in direction for the percent black predictor (OR from 0.99 to 1.02) and an attenuation of the distance predictor (OR from 0.747 to 0.91). The residual deviance suggests that there remains lack of fit although it is less extreme (deviance 274.23 with 7 df to 124.85 with 6 df. Technically these are not comparable because of the difference in the datasets, one with and one without Greensboro).

Model 4: distance + perBlk + distance*perBlk without Greensboro

| Coefficients: | Estimate | Std. Error | z value | Pr(>|z|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | -4.299344| 1.199362   | -3.585  | 0.000337 *** |
| perBlk        | 0.070101 | 0.014766   | 4.748   | 2.06e-06 *** |
| distance      | 0.230035 | 0.090741   | 2.535   | 0.011242 *  |
| perBlk:distance | -0.004403| 0.001207   | -3.649  | 0.000264 *** |

Residual deviance: 124.85 on 6 degrees of freedom

Because the residual deviance remains large despite removing the outlier and given we have included all of the covariates at hand as well as an interaction term, the observed binomial counts are overdispersed, exhibiting more variation than the model would suggest.

6.6.6.3 Overdispersion

Community members vote Yes (1) or No (0). Votes are summed in each community and recorded as the number of Yes votes. The number of Yes votes is a binomial variable if the community member’s votes are independent of one another. However, there is some concern that individual votes in a community are not independent of one another. For example, when there exist like-minded clusters within a community it may lead to overdispersion, a situation where the variation is actually greater than the binomial model might predict. If overdispersion is suspected, it is best to take it into account. Some statisticians would recommend that you always assume and correct for overdispersion, but there are differences of opinion on that point. Recall that the binomial model implies that the variance of a binomial variable is np(1 – p). With overdispersion there is extra-binomial variation, the variance will be greater than this. One way to adjust for overdispersion is to estimate a multiplier or a dispersion parameter , φ, for the variance that will inflate it and reflect
the reduction in the amount of information we would otherwise have with independent observations.

### Binomial Model

\[
E(Y_i) = p \\
\text{Var}(Y_i) = p(1 - p)
\]

### Overdispersed Model

\[
E(Y_i) = p \\
\text{Var}(Y_i) = \phi p(1 - p) \\
\hat{\phi} = \text{residual deviance}/ \text{deviance df}
\]

When overdispersion is adjusted for in this way, we no longer have a binomial model. Unlike the binomial model where we use maximum likelihood we instead use a quasi-likelihood to fit the model. Quasi-likelihood is similar in spirit to likelihood-based inference. But because the model uses the dispersion parameter it is not a binomial model, we do not have a true likelihood. Most statistical packages offer quasi-likelihood as an option when model fitting. The quasi-likelihood approach will yield the same coefficient estimates as maximum likelihood, however for \( \phi > 1 \), the variances will be larger in the presence of overdispersion. We will see other ways in which to deal with overdispersion and clusters in the remaining chapters in the book.

Output for a model that adjusts Model 3 for overdispersion appears below. The estimate of the dispersion parameter is 51.6. The estimated dispersion parameter is used to adjust the standard errors for the coefficients and the drop-in-deviance tests during model building. The standard errors are adjusted by multiplying each standard error by the square root of the estimated dispersion parameter. The standard errors for each estimated coefficient are inflated.

**Model 5: distance + perBlk + distance*perBlk fit using quasi-liklihood**

| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 7.550902 | 4.585464   | 1.647   | 0.144    |
| perBlk        | -0.064731| 0.065885   | -0.982  | 0.359    |
| distance      | -0.614005| 0.412171   | -1.490  | 0.180    |
| perBlk:distance| 0.005367 | 0.006453   | 0.832   | 0.433    |

(Dispersion parameter for quasibinomial family taken to be 51.5967)

Adjusting for overdispersion results in an insignificant interaction term. The interaction term is insignificant because taking into account overdispersion reduces the effective sample size, so the standard errors increase and it is less likely that a coefficient is significant. Removing the interaction term leaves `distance` and `perBlk` in the model. When we fit the model with `distance` and `perBlk` and found a large residual deviance. So for we refit the model with these two covariates using a quasi-likelihood to adjust for overdispersion and find the following.

**Model 6: distance + perBlk + distance fit using quasi-liklihood**

| Coefficients: | Estimate | Std. Error | t value | Pr(>|t|) |
|---------------|----------|------------|---------|----------|
| (Intercept)   | 4.22202  | 1.99031    | 2.121   | 0.0667   |
| perBlk        | -0.01323 | 0.02612    | -0.506  | 0.6262   |
| distance      | -0.29173 | 0.08780    | -3.323  | 0.0105 * |

(Dispersion parameter for quasibinomial family taken to be 44.9194)
\( \exp(\text{coef(model.HaleBDq)}) \)

(Intercept)  perBlk  distance
68.1711286  0.9868600  0.7469668

When using a quasi-likelihood, the estimated coefficients remain the same (see Model 2), but as noted the standard errors will be inflated by a factor of \( \hat{\phi} \). Here we find that the perBlk is no longer significant but the distance coefficient remains significant. The conclusions for this study are similar to those we found with Model 1. The estimated odds ratio for distance, that is the exponentiated coefficient for distance, in this model is 0.75. It can be interpreted as follows: for each additional mile from the proposed railroad, the support declines by a factor of 25%. There is no evidence that this association differs by the percentage of blacks in a community.

### 6.6.7 Least Squares Regression vs. Logistic Regression

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Variance</td>
<td>Normal</td>
<td>Binomial</td>
</tr>
<tr>
<td></td>
<td>Equal for each level of X</td>
<td>np(1 – p) for each level of X</td>
</tr>
<tr>
<td>Model Fitting</td>
<td>( \mu = \beta_0 + \beta_1 X )</td>
<td>log(odds) = ( \beta_0 + \beta_1 X )</td>
</tr>
<tr>
<td></td>
<td>Least Squares</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>EDA</td>
<td>plot X v Y, add line</td>
<td>plot log(odds) vs X for each subgroup</td>
</tr>
<tr>
<td>Comparing Models</td>
<td>Residual Sum of Squared Residuals</td>
<td>Deviance Residual ( \chi^2 ) tests</td>
</tr>
<tr>
<td>Interpreting Coefficients</td>
<td>( \beta_1 ) change in ( \mu )</td>
<td>( e^{\beta_1} ) percent change in the odds</td>
</tr>
<tr>
<td>Overdispersed Binomial</td>
<td>( \hat{\phi} np(1 – p) )</td>
<td>modified F-test</td>
</tr>
</tbody>
</table>

### 6.7 Case Study: Who wants to lose weight? Sex, Media, Sports, and BMI.

The next case study uses individual specific information so that our response is not the number of successes out of some number of trials. The responses are binary taking on values of 0 or 1.

#### 6.7.1 Background

What are the characteristics of young people who are trying to lose weight? The prevalence of obesity among US youth suggests that wanting to lose weight is sensible and desirable for some young people such as those with a high body mass index (BMI). On the flip side, there are young people who do not not need to lose weight but make ill advised attempts to do so nonetheless. A multitude of studies on weight loss focus specifically on youth and propose a variety of motivations for the young wanting to lose weight. Athletics and the media are two commonly cited sources of motivation for losing weight for young people.

Sports have been implicated as a reason for young people wanting to shed pounds. But not all studies are consistent with this idea. For example, a report on a study entitled *Disordered eating and dieting among adolescent elite athletes* by Martinsen et al. reported that, despite preconceptions to the contrary, there was a higher rate of self-reported eating disorders among controls (non-elite) as opposed to elite athletes. Interestingly, the kind of sport was not found to be a factor, participants in leanness sports (for example, distance running and swimming, gymnastics, dance, and diving) did
not differ in the proportion with eating disorders when compared to those in non-leanness sports. So in our analysis, we do not make a distinction between different sports in this analysis.

Other studies suggest that mass media is the culprit. They argue that students’ exposure to unrealistically thin celebrities may provide unhealthy motivation for some to try to slim down particularly young women. An examination and analysis of a large number of studies referred to as a meta-analysis (Gradbe et al. 2008) found a strong relationship between exposure to mass media and the amount of time that adolescents spend talking about what they see in the media, deciphering what it means and how they can be more like the celebrities.

6.7.2 Research Questions

The following alternative hypothesis are of interest.

- The odds that young females report trying to lose weight is greater that the odds that males do.
- Increasing BMI is associated with an interest in losing weight, regardless of sex.
- Sports participation is associated with the desire to lose weight although this may differ between men and women.
- Media exposure is associated with more interest in losing weight, more so for females.

6.7.3 Data Collection

A sample of 500 teens from data collected in 2009 through the U.S. Youth Risk Behavior Surveillance System (YRBSS). The YRBSS is an annual national school-based survey conducted by the Center for Disease Control (CDC) and state, territorial, local education and health agencies and tribal governments. The CDC states on their website that the “survey monitors six types of health-risk behaviors that contribute to leading causes of death and disability among youth and young adults,” including:

- Behaviors that contribute to unintentional injuries and violence
- Tobacco use
- Alcohol and other drug use
- Sexual risk behaviors
- Unhealthy dietary behaviors
- Physical inactivity

More information on this survey can be found at http://www.cdc.gov/HealthyYouth/yrbs/index.htm

6.7.4 Data Organization

Here are the four questions from the YRBSS we use for our investigation:

Q66. Which of the following are you trying to do about your weight?
A. Lose weight
B. Gain weight
C. Stay the same weight
D. I am not trying to do anything about my weight

Q81. On an average school day, how many hours do you watch TV?
A. I do not watch TV on an average school day
B. Less than 1 hour per day  
C. 1 hour per day  
D. 2 hours per day  
E. 3 hours per day  
F. 4 hours per day  
G. 5 or more hours per day

Q84. During the past 12 months, on how many sports teams did you play?  
(Include any teams run by your school or community groups.)  
A. 0 teams  
B. 1 team  
C. 2 teams  
D. 3 or more teams

Answers to Q66 are used to define our response variable: (A) trying to lose weight (Y=1). The other non-missing values are set to 0 (Y=0). Q84 provides information on students’ sports participation and is treated as numerical, 0 through 3 with 3 representing 3 or more. As a proxy for media exposure we use answers to Q81 as numerical values 0, 0.5, 1, 2, 3, 4, 5 using 5 to represent 5 or more. Media exposure and sports participation are also considered as factors, that is, responses having levels via indicator variables as opposed to their numerical values.

BMI is included in this study as the percentile for a given BMI for members of the same sex. This facilitates comparisons when modeling with males and females. We will use BMI and BMI percentile interchangeably with the understanding that we are always referring to the percentile.

With our sample, we use only the cases that include all of the data for these four questions. This is referred to as a complete case analysis. That brings our sample of 500 to 426. There are limitations of complete case analyses that we address in the Discussion.

6.7.5 Exploratory Data Analysis

Primary Outcome

Fully 50% of our sample of 426 youth report that they are trying to lose weight.

Covariates

Fifty-four percent of the sample are females. Among the 426 youth in our sample, approximately 55% play on one or more sports teams. Ten percent report that they do not watch any TV on school days whereas about 11% watched 5 or more hours each day.

The relationship between the Primary Outcome and Covariates

The most dramatic difference in the proportions of those who are trying to lose weight is by sex; 63% of the females want to lose weight in contrast to only 36% of the males. This provides strong support for the inclusion of a sex term in every model considered.

The next hypothesis addresses the question of whether those with higher BMI also tend to express attempting to lose weight regardless of sex. Table 6.3 displays the mean BMI of those wanting and not wanting to lose weight for males and females. The mean BMI is greater for those trying to lose weight compared to those not trying to lose weight, regardless of sex. The size of the difference is remarkably similar for the two sexes.

If we consider including a BMI term in our model(s) the logit should be linear in BMI. We can
investigate this assumption by constructing empirical logits. Because BMI is not discrete, we create BMI categories by dividing the BMI percentile variable into intervals, we choose intervals of length 10, and then count up the number of respondents in each interval and determine the proportion, \( \hat{p} \), that report that they want to lose weight. The empirical logit for each interval is \( \log(\hat{p}/(1 - \hat{p})) \). The choice of the interval length was somewhat arbitrary and there are legitimate definitions that may alter the results. We construct the empirical logits separately for males and females. Figure 6.5 presents the empirical logits for the BMI intervals by sex. Both males and females exhibit an increasing linear trend on the logit scale indicating that increasing BMI is associated a greater desire to lose weight. The slope for the females appears to be a little steeper than the males. This suggests that we consider an interaction term of BMI by sex in the model. It is also worth noting that there is a larger than expected logit for females in the 10th to 20th percentile. This may reflect undue pressure for girls to slim down unnecessarily.

Our next hypothesis involves sports participation. Of those that play sports, 48% want to lose weight whereas 54% want to lose weight among those who do not play sports. Table 6.4 contains the proportion of who want to lose weight by their sex and sport participation. The data support the hypothesis that sports participation is associated with less of a desire to lose weight contrary to what had originally been hypothesized. It suggests that the overall levels of those wanting to lose weight differs considerably between the sexes. However, the differences between those in and out of sports within sex appear to be very similar. A look at an empirical logit plot by number of teams (Figure ??) reinforces this result of similar patterns and nearly parallel slopes. A term for sports participation or number of teams will be considered but there is not compelling evidence that an interaction term will be needed.

It was posited that increased exposure to media, here measured as hours of TV daily, is associated with increased desire to lose weight particularly for females. Overall, the extremes among those watching TV are 45% of those watching 2 hours of TV per day want to lose weight in comparison to 56% watching one hour daily. There is very little variation in the proportion wanting to lose weight with both sexes combined. However, we are interested in differences between the sexes. Furthermore, we need a linear trend in the logit not \( p \). We create empirical logits using the proportion of the number of students trying to lose weight for each level of hours spent watching daily. We look at the trends in the logits separately for males and females. From Figure 6.6, there does not appear
<table>
<thead>
<tr>
<th>Sex</th>
<th>Sports Participation</th>
<th>Proportion wanting to lose weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>No sports</td>
<td>0.636</td>
</tr>
<tr>
<td></td>
<td>Sports</td>
<td>0.615</td>
</tr>
<tr>
<td>Male</td>
<td>No sports</td>
<td>0.368</td>
</tr>
<tr>
<td></td>
<td>Sports</td>
<td>0.359</td>
</tr>
</tbody>
</table>

Table 6.4: Mean BMI percentile by sex and desire to lose weight.

Figure 6.6: Empirical Logits for the odds of trying to lose weight by daily hours of TV and Sex.

to be a linear relationship for males or females. The most striking differences between the sexes are the high odds for attempted weight loss among females who watch little or no TV compared to the very low levels of desired weight loss among males with similar viewing habits. For one hour or more of viewing the patterns look similar for males and females although the level is higher for females overall.

6.7.6 Modeling

Our strategy for modeling is to use our hypotheses of interest and what we have learned in the exploratory data analysis for modeling. For each model we interpret the coefficient of interest, look at the corresponding Wald test (estimated coefficient/s.e.) and as a final step compare the deviances for the different models we considered.

6.7.6.1 Model 1: Sex

The estimated coefficient in the model with sex (0=male, 1=female) as the only covariate is 1.0811 with a standard error of 0.2016 and a Wald-type statistics of 5.362 implying a very low p-value(1.08e-11). We can interpret the coefficient by exponentiating \(e^{1.0811} = 2.95\) indicating that the odds of a female trying to lose weight is nearly three times that of the odds of a male trying to lose weight (95% CI: 1.99, 4.39). Most statistical packages will produce the confidence interval or it can be constructed using the point estimate and standard error. In the later case, you will need to construct the interval for the coefficient and then take exponents to obtain the CI for the OR. We retain sex in the model and consider adding the BMI percentile.
6.7.6.2 Model 2: sex + bmipct

The estimated coefficient for bmipct given that sex is in the model is 0.033 with a standard error of 0.004 (p < .0001) Clearly bmipct belongs in the model with sex. The coefficient can be interpreted using the $e^{0.033} = 1.03$ (95% CI: 1.02, 1.04) so there is an estimated 3% increase in the odds of wanting to lose weight for each additional percentile of BMI for members of the same sex. Just as we had done in other multiple regression models we need to interpret our coefficient given the other variables remain constant. An interaction term for BMI by sex was not significant (p=0.781).

6.7.6.3 Model 3: sex + bmipct + sport

Sports participation was considered for inclusion in the model in three ways. First an indicator of sports participation (0=no teams, 1=one or more teams), next the number of teams (0,1,2 or 3), and lastly, treating the number of teams as a factor. Sports teams were not significant in any of these models nor were interaction terms (sex by sports) and (bmipct by sports). Sports participation was no longer considered for inclusion in the model.

6.7.7 Final Model

6.7.7.1 Model 4: sex+media + sex*media + bmipct

Because interest centers on how media may affect attempts to lose weight and that it might affect females differently than males we fit a model with a media term and a sex by media interaction term. The interaction was statistically significant (p=0.04) suggesting that media exposure as measured by hours spent watching TV differs for males and females. In particular for males the odds of losing weight are $e^{0.145} = 1.15$ or 15% greater for each additional hour of TV watching daily for similar BMI. For females each additional hour watching TV reduces their odds of trying to lose weight by $1 - e^{0.145 - 0.295}$ or 14% for similar BMI.

6.7.7.2 Drop-in-deviance Tests

6.7.8 Results

The correlation between the predictors did not affect the estimated odds ratios form the model with distance alone compared to the model with distance and percent black (OR = 0.75 for both models). The results are similar with Greensboro removed. to compare the models is to examine the drop in the deviance with each added term.

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Model 1: lose.wt.l ~ sex.f</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 2: lose.wt.l ~ sex.f + bmipct</td>
<td>424</td>
<td>560.70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 3: lose.wt.l ~ sex.f + bmipct + as.numeric(sport4)</td>
<td>423</td>
<td>493.11</td>
<td>1</td>
<td>67.593</td>
</tr>
<tr>
<td>Model 4: lose.wt.l ~ sex.f + bmipct + as.factor(sport4)</td>
<td>422</td>
<td>493.04</td>
<td>1</td>
<td>0.075</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 1: lose.wt.l ~ sex.f</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
<th>Df</th>
<th>Deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 2: lose.wt.l ~ sex.f + bmipct</td>
<td>424</td>
<td>560.70</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The first set of deviances allows us to compare models involving sex, bmipct, and sports participation. Comparing models using differences in deviances requires that the models be nested, meaning each larger model contains all of the terms of the smaller model. These two sets of models are nested models so our approach is valid. Models 1, 2, and 4 are nested as are Models 1, 2, and 3.

There is a large drop-in-deviance adding bmipct to the model with sex. (Model 1 to Model 2, 67.593) which is clearly statistically significant when compared to a $\chi^2$ distribution with 1 df.

The drop-in-deviance for adding a linear term for sports to the model with sex and BMI is 493.11-493.04=0.075. There is a difference of a single parameter, so the drop-in-deviance would be compared to a $\chi^2$ distribution with 1 df. The resulting $p$-value is very large suggesting that adding a linear term in sports is not helpful.

Suppose we wanted to compare a model with sex + bmipct + sport to a model with sex + bmipct + media, we cannot use the drop-in-deviance because these two models are not nested. One approach is to look at the Akaike Information Criteria (AIC). The AIC is defined using the deviances or deviations from the model. Since less deviation is preferred, models with smaller AICs are preferred to models with larger AICs. Most statistics packages provide AICs.

The AIC is smallest for the model with media providing evidence that the latter model is preferred. Note that a similar comparison can be made using Bayesian Information Criteria (BIC).

6.7.9 Discussion

We found that the odds of wanting to lose weight are considerably greater for females compared to males. Appropriately respondents with greater BMI percentiles express a greater desire to lose weight for members of the same sex. Regardless of sex or BMI percentile, sports participation is not associated with lower odds for wanting to lose weight compared to those not participating in sports. It is clear that exposure to increasing amounts of TV is associated with trying to lose weight and the effect differs for males and females. In direct contradiction to what was hypothesized, the results do not provide support for a larger effect for females rather increasing TV watching for females was associated with decreasing attempts to lose weight and males with increasing efforts to lose weight.

A limitation of this analysis is that we used complete cases in place of a method of imputing responses or modeling missingness. This reduced our sample from 500 to 426 and it may have introduced bias. For example if respondents who watch a lot of TV were unwilling to reveal as much and if they differed with respect to their desire to lose weight from those respondents who reported watching lots of TV our inferences regarding the relationship between lots of TV and desire to lose weight may be biased.

Other limitations may result from definitions. Trying to lose weight is self-reported and may not correlate with any action undertaken to do so. The number of sports teams may not accurately reflect sports related pressures to lose weight. For example, elite athletes may focus on a single sport and be subject to pressures whereas athletes who casually participate in three sports may not feel any
pressure to lose weight. Hours spent watching TV is not likely to encompass the totality of media exposure particularly because exposure to celebrities occurs often online. Furthermore, this analysis does not explore in any detail maladaptions, inappropriate motivations for wanting to lose weight. For example, we did not focus our study on subsets of respondents with low BMI who are attempting to lose weight although one such group of females appears on the empirical logit plot of BMI by sex.

It would be instructive to use data science methodologies to explore the entire data set of 16,000 instead of sampling 500. However, the types of exploration and models used could translate to the larger sample size.

Finally a limitation may be introduced as a result of the acknowledged variation in the administration of the YRBS. States and local authorities are allowed to administer the survey as they see fit which at times results in significant variation in sample selection and response.

6.8 R code

```r
## Railroad Referenda
# To construct empirical logit plots
p = with(rrHale.df, Yes/(Yes+No))
elogit = log(p/(1-p))
plot(elogit~distance, data=rrHale.df)

## Lose Weight
# To draw a random sample from risk2009.data
# Creating a sample of 500 from risk2009.data
y_data = data.frame(risk2009.data)
y_id = id(y_data)
y_sample = with(y_data, y_data[y_id %in% sample(unique(y_id),500),])

# To obtain complete cases
y_ccind = with(y_sample, complete.cases(lose.wt,sex,bmipct,sport4,sport,media))
y_sample = data.frame(y_sample,y_ccind)

# y.cc is a data frame of complete cases
y_cc = data.frame(y_sample[y_ccind==TRUE,])
dim(y.cc)

# To obtain an analysis of the deviances
anova(model.1,model.12,model.123l,model.123f)
```

Works Cited

B Beadnell (2005) Condom Use, Frequency of Sex, and Number of Partners


Grabe, Shelly; Ward, L. Monica; Hyde, Janet Shibley (2008) The role of the media in body image concerns among women: A meta-analysis of experimental and correlational studies. Psycho-
6.9 Exercises

6.9.1 Interpret article abstracts

1. Interpret the odds ratios in the following abstract.

*Day Care Centers and Respiratory Health* Per Nafstad et al.

**Objective.** To estimate the effects of the type of day care on respiratory health in preschool children.

**Methods.** A population-based cross-sectional study of Oslo children born in 1992 was conducted at the end of 1996. A self-administered questionnaire inquired about day care arrangements, environmental conditions, and family characteristics (n = 3853; response rate, 79%).

**Results.** In a logistic regression controlling for confounding, children in day care centers had more often nightly cough (adjusted odds ratio, 1.89; 95%), and blocked or runny nose without common cold (1.55; 1.07, 1.61) during the past 12 months compared with children in home care. Poisson regression analysis showed an increased risk of the common cold (incidence rate ratio, 1.21; 1.12, 1.30) and otitis media (1.48; 1.22, 1.80), and the attributable proportion was 17.4% (95% confidence interval, 10.7, 23.1) for the common cold and 32.4% (18.044.4) for otitis media. Early starting age in the day care center increased the risk of developing recurrent otitis media. Also the lifetime risk of doctor-diagnosed asthma was higher in children who started day care center attendance during the first 2 years of life.

**Conclusions.** Attendance to day care centers increases the risk of upper respiratory symptoms and infections in 3- to 5-year-old children. The starting age seems to be an important determinant of recurrent otitis media as well as asthma. The effect of day care center attendance on asthma is limited to age up to 2 years. This effect is most likely mediated via early respiratory tract infections that are substantially more common in children in day care centers compared with children in home care.

*Pediatrics* 1999;103:753-758; children, respiratory health, asthma, otitis media, day care.
2. Construct a table and calculate the corresponding odds and odds ratios. Comment on the reported and calculated results.

**IVF and Birth Defects**

**Data Source:** CDC

In November, the Centers for Disease Control and Prevention published a paper reporting that babies conceived with IVF, or with a technique in which sperm are injected directly into eggs, have a slightly increased risk of several birth defects, including a hole between the two chambers of the heart, a cleft lip or palate, an improperly developed esophagus and a malformed rectum. The study involved 9,584 babies with birth defects and 4,792 babies without. Among the mothers of babies without birth defects, 1.1 percent had used IVF or related methods, compared with 2.4 percent of mothers of babies with birth defects. The findings are considered preliminary, and researchers say they believe IVF does not carry excessive risks. There is a 3 percent chance that any given baby will have a birth defect. **Skills:** Calculating and interpreting OR and probabilities and linking them to the logit form of the model. Also importance of comparing to a baseline.

6.9.2 Guided

1. Soccer goals on target

**Data Source:** Data from an article in *Psychological Science*, July 2011. The example in the text referred to all shots, good or bad. This data relates to shots on target, 18 out of 20 shots were scored when the goalie’s team was behind, 71 out of 90 shots were scored when the game was tied, and when the team was not behind 55 out of 75 shots were scored. Calculate the odds of scoring for games behind, games tied, and games not behind. Construct empirical odds ratio for scoring for behind versus tied and tied versus not behind.

2. Fit a model that with the predictor c("behind","tied","not behind") and interpret the exponentiated coefficients. How do they compare to the empirical odds ratios you calculated?

3. Would it be better to model this data using a negative binomial variable? Explain.

4. Is there a reason to believe that the shots may be correlated? Explain.

5. **Medical School Admissions**

The data for Medical School Admissions is in MedGPA.xls. **Data Source:** Undergraduates from a small liberal arts school over several years. Compare the relative effects of improving your MCAT score versus improving your GPA. **Skills:** EDA including empirical logit plots, interpreting multiple logistic coefficients, comparing multiple logistic coefficients.

6. **Medical School Admissions**

**Data Source:** Undergraduates from a small liberal arts school over several years.

Are there different majors that are more likely to gain medical school admission? Are there certain majors that perform better than others on the MCAT? **Skills:** Categorical covariates, binomial regression

**Skills:** Binomial regression

7. Titanic Survival

8. Bluebird Fledging

9. Fish Tumors

10. Cafe Senate Vote

6.9.3 Open-ended

1. **Presidential Voting in Minnesota counties**

Data is in MNVote.xls. **Data Source:** Response of Obama or McCain county-wide victory for each of the 87 counties in Minnesota along with county-level covariates such as percent of house-
LOGISTIC REGRESSION

holds in poverty, percent of the county that is rural, unemployment rate. Model the county-level Obama vote as a function of the percent of households in poverty, percent of the county that is rural, unemployment rate. Interpret your coefficients. **Skills: requires that covariates be interpreted.**

2. **Crime on Campus**
   Data in Crime.xls.
   In this dataset, crimes are characterized as violent or property crimes. Of the total number of crimes, does the proportion of violent crimes differ by regions of the country?

6.9.4 **Project Ideas**

1. **Spam** Perform a screen shot of an inbox. Identify which emails are spam. Have students come up with a model that predicts whether an email is spam or not. Try this model out on a new inbox.

2. **Trashball** Great for a rainy day! A fun way to generate overdispersed binomial data. Each student crumbles an 8-1/2 by 11 inch sheet and tosses it from three prescribed distances ten times each. 
   
   $Y_i$ = number of baskets out of 10 tosses keeping track of the distance.

3. **GSS** student generate research question(s)

4. **YRBS** student generate research question(s)

6.10 **References**

Psychological Science 22(11) 1403-1407 Marieke Roskes, Department of Psychology, University of Amsterdam, Weesperplein 4, 1018 XA Amsterdam, The Netherlands E-mail: m.roskes@uva.nl

Chapter 7

Correlated Data

7.1 Learning Objectives

- Given a data structure, be able to recognize when there is a potential for correlation.
- Identify a primary sampling unit and corresponding observations for a study.
- Provide reasons why correlated observations may cause problems when modeling.
- Describe overdispersion and why it might occur.
- Be able to use the overdispersion parameter estimate to adjust Wald tests, F-tests, and confidence intervals to account for correlation.
- Understand how correlation in data can be taken into account using random effects models.
- Be able to describe the differences in the kind of inferences that can be made using fixed versus random effects models.
- Use software to fit a model taking correlation into account using:
  - overdispersion parameter estimates (quasi-likelihood) or
  - random effects.

Introductory statistics courses typically require responses which are approximately normal and independent of one another. We saw from he first chapters in this book that there are models for non-normal responses so we have already broadened the types of applications we can handle. In this chapter, we relax an additional assumption from introductory statistics, that of independence. Here we introduce methods that allow correlation to be taken into account. When modeling, correlation can be considered for normal and non-normal responses. Taken together, we can handle many more applications than the average introductory student.

First, we focus on recognizing data structures that may imply correlation. Next, we consider potential problems correlated outcomes may cause and why we need to take correlation into account when modeling. Models which take correlation into account are then described and fit with the help of R.

7.2 Recognizing correlation

Correlated data is encountered in nearly every field. In education, student scores from a particular teacher are more similar than scores of other students who have had a different teacher. During a study measuring depression indices weekly over the course of a month, we may find that four measures for the same patient tend to be more similar than depression indices from other patients. In political polling, opinions from members of the same household may be more similar than opinions of members from other randomly selected households. The structure of these data sets suggest inherent patterns of similarities or correlation among outcomes. This kind of correlation specifically concerns correlation of observations within the same teacher or subject or household and is referred
To simplify our discussion, we borrow terminology from survey methodology. In our examples, teachers, patients, and households can be referred to as primary sampling units or PSUs. A PSU is sampled and multiple measurements or observations are made for each PSU, here are test scores, depression indices and political opinions are the measurements, respectively. Other examples appear in Table 7.1. Look for this structure when analyzing data. A description of this kind from the survey setting follows. It generalizes to the settings we consider.

In sample surveys, primary sampling unit (commonly abbreviated as PSU) arises in samples in which population elements are grouped into aggregates and the aggregates become units in sample selection. The aggregates are, due to their intended usage, called “sampling units.” Primary sampling unit refers to sampling units that are selected in the first (primary) stage of a multi-stage sample ultimately aimed at selecting individual elements. – James M. Lepkowski, SAGE

Mixtures including zero Inflated Poisson (zipcounts in R), and mixture of Normals are other correlated data structures. Several different types of data may apply to a single application. The type of response variable refers to the observations on the elements.

Table 7.1 Examples of data structures for studies with potential within PSU correlation. Elements of the PSUs with measurements to be made and type of measurement.

<table>
<thead>
<tr>
<th>Types of Data</th>
<th>Study</th>
<th>PSU</th>
<th>Elements</th>
<th>Response</th>
<th>Type of Response Variable*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustered Data</td>
<td>Four dams (pregnant rats) produce pups who have a defect or no defect.</td>
<td>Dam</td>
<td>Pups</td>
<td>No defect (0) or defect (1)</td>
<td>Binary</td>
</tr>
<tr>
<td>Repeated Measures</td>
<td>Ten newborns’ height is measured four times each.</td>
<td>Newborn</td>
<td>Measurement Attempts</td>
<td>Height measurement</td>
<td>Normal</td>
</tr>
<tr>
<td>Longitudinal Data</td>
<td>Twenty musicians record their level of anxiety for each of their performances throughout the season.</td>
<td>Musician</td>
<td>Performances</td>
<td>Anxiety scores</td>
<td>Normal</td>
</tr>
<tr>
<td>Survey Data</td>
<td>A random survey of county board members for 88 counties solicits their opinion on a tax referendum.</td>
<td>County</td>
<td>County board members</td>
<td>Opinions</td>
<td>Ordinal</td>
</tr>
<tr>
<td>Spatial Data</td>
<td>Particulate measurements are made at 25 locations within teen counties.</td>
<td>County</td>
<td>Measurement location</td>
<td>Particulate measurements</td>
<td>Poisson</td>
</tr>
</tbody>
</table>
7.3 Case Study: Dams and pups, Correlated Binary Outcomes

A **teratogen** is a substance or exposure that can result in harm to a developing fetus. An experiment is conducted to determine whether increasing levels of a potential teratogen results in increasing probability of defects in rat pups. Dams (pregnant rats) are exposed to one of two dose levels of a potential teratogen. A control group is not exposed. Dams produce litters and the number of pups with a defect is the outcome of interest.

### 7.3.1 Sources of Variability

In order to analyze the entire data set, let’s step back and look at the big picture. Statistics is all about analyzing variability, so let’s consider what sources of variability we have in the dams and pups example. There are several reasons why the counts of the number of defective pups differ from dam to dam and it is helpful to explicitly identify what they are in order to determine how the dose levels affects the pups in addition to accommodating correlation.

**Dose Effect** The dams and pups experiment is being carried out to determine whether different dose levels affect the development of defects differently. Of particular interest is determining whether a dose-response effect is present. A dose-response effect is evident when dams receiving higher dose levels produce higher proportions of pups with defects. The discovery of significant dose level effects is typically of interest within this experiment and beyond. Publishing the defect rates for each dose level in a paper in a journal, for example, would be of interest to other teratologists. For that reason, we refer to the dose level effects as **fixed effects**.

**Dams (Litter) Effect** In many settings like this, there is a litter effect as well. For example, some dams may exhibit a propensity to produce pups with defects while others rarely produce litters with defective pups. That is, observations on pups within the same litter are likely to be similar or correlated. Unlike the dose effect, teratologists reading experiment results are not interested in the estimated probability of defect for each dam in the study and we would not report these estimated probabilities in a paper. However, there may be interest in the variability in litter-specific defect probabilities. Often this kind of effect is modeled using the idea that randomly selected dams produce random effects. This provides one way in which to model correlated data. We elaborate on this idea throughout the remainder of the text after this next section on overdispersion.

**Pup-to-pup variability** The within litter pup-to-pup differences reflect random, unexplained variation in the model.

### 7.3.2 Analyzing a control group

We begin by focusing on a control group ignoring the dose group data. The control group for this study consists of four dams who each produce 15 rat pups in their litter. Each pup is examined for the presence or absence of a defect. (Dams typically produce 8 to 15 pups so it is unlikely that every dam produces 15 pups. Here we keep it simple with fifteen pups per litter.) Before proceeding any further with the analysis, take note of the data structure. We can think of this as a repeated measures study where dams are PSUs and the elements are the binary outcomes of defect or no defect for each pup. The total number of pups with a defect is the response which is binomial with $n=15$ and probability of defect $p$. This structure alerts us to the possibility that pup outcomes may be correlated within litters and we may not have 60 independent observations and we’ll want to take that into account when modeling.

Each dam $j$ produces a litter of 15 pups ($j=1, \ldots, 4$). Individual pup outcomes are either 0 or 1.
for “no defect” or “defect” and are denoted as $Y_{jk}$ for the $k^{th}$ pup from the $j^{th}$ dam, $k = 1, \ldots, 15$ and $j = 1, \ldots, 4$. The binomial counts, $Y_j$, the total number of defects for the $j^{th}$ litter, are recorded for each of the four litters. If dams do not differ in their propensity to produce defects, these counts follow a binomial distribution with $n=15$ and a common probability $p$. However, if dams’ probabilities of producing a pup with a defect differ from dam-to-dam, the number of pups with defects will not be coming from the same binomial distribution and the counts are likely to differ more than they would with a common $p$.

For illustration, we generate a (hypothetical) control group data set based on the notion that different dams are producing defective pups with different probabilities. Our simulation assumes that $p$, the true probability that dam 1 produces a pup with a defect, is 0.40. She produces 15 pups and has 8 pups with defects. Dam 2 has a true probability of 0.3 for the probability of a defective pup and among her 15 pups only 3 have defects. Dam 3 has a very unhealthy lifestyle and as a result produces pups with defects about 50% of the time. For her litter of 15, she has 7 pups with defects. Only 20% of the pups from the healthiest dam have defects and this litter has 3 pups with defects. We have made the variation in $p$ dramatic for illustration.

Overall the estimated probability of a defect in the control group is $(8+3+7+3)/(4\times15) = 0.35$. This would be a good estimate if all four dams had the same probability of producing a defective pup. If we proceeded with this naive assumption, recall that the variance of a binomial random variable is $np(1-p)$ so we would expect the variation in the number of defective pups to be approximately $15\times 0.35 \times (1-0.35)$ or 3.41. However, we calculate the variance for our four litters ($s^2 = \sum (x_i - \bar{x})^2/(n-1)$) and find 6.92. Comparing the model-based variance $np(1-p)$ to the observed variance ($s^2$), we find a lot more variation in the number of defects among the control group litters than we would expect if in fact they were all coming from the same binomial distribution. If we hadn’t generated the data, making this comparison would lead us to suspect that there is a dam effect and the observations within litters are correlated. In the case of binomial data, this is referred to as extra-binomial variation and more generally as overdispersion.

We found that overdispersion can be a clue that observations are correlated within PSUs. Why should we worry about it? If there is correlation within litters, there are not 60 independent pieces of information, that is the effective sample size is less than 60. Standard errors are based on the reciprocal of the sample size, so standard errors for the naive model will be artificially small resulting in more significant results and narrower confidence intervals than there would be with the larger standard errors based on the effective sample size. Overdispersion suggests that our naive binomial model is incorrect and may well result in significant lack-of-fit. Luckily there are ways in which we can modify the naive model to improve this situation. Had we been attentive to the data structure, we would not be surprised to find overdispersion. How can we take this correlation into account?

The model-based variance of the counts for the four litters is artificially small (3.41 instead of 6.92). Using a model-based variance to calculate standard errors for Wald tests and confidence intervals will lead to too many significant results or confidence intervals which are too narrow. One way to correct for overdispersion is to inflate the model-based variance by using a factor (an estimated dispersion parameter often denoted $\hat{\phi}$) so the model uses $\phi np(1-p)$ as the variance instead of $np(1-p)$ where $\phi > 1$. We will demonstrate this using software with an expanded data set in a subsequent section.

7.3.3 Control and Dose Groups

We expand our look at the teratology data to include eight more dams in two dose groups. The sum of the binary observations (defect or not) for each litter yields a binomial count (the number of
Table 7.2 Data on number of pups with defects out of 15 pups for each of 12 dams randomly assigned to a control or dose group.

The simulated data is displayed in Table 7.2 and Figure 7.1. The means and proportions suggest a dose-response effect. The plot of the empirical logits suggests that a model with logits linear in dose is plausible. We need to ascertain whether the observed group differences are statistically significant.

7.3.3.1 Exploratory Data Analysis of the Teratology Data

The simulated data is displayed in Table 7.2 and Figure 7.1. The means and proportions suggest a dose-response effect. The plot of the empirical logits suggests that a model with logits linear in dose is plausible. We need to ascertain whether the observed group differences are statistically significant.

In addition to the issue of differences between the groups, we also want to consider the variability within dose groups. If the binomial model holds we expect the variances of the counts within dose groups to be \(15 \times p(1 - p)\). Table 7.3 contains group specific model-based variances calcu-
### Table 7.3: Mean number of defects, variance and model-based variance for controls and dose groups.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Account for correlation?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. $\chi^2$ test</td>
<td>Construct a table of counts of defects and non-defects for each dose level. Use the marginal proportions to obtain expected values for each table entry. Calculate the $\chi^2$ test statistic.</td>
<td>No</td>
</tr>
<tr>
<td>B. GLM model</td>
<td>Use a likelihood to fit a GLM binomial model with the logit of the number of defects as the response and dose group as the predictor.</td>
<td>No</td>
</tr>
<tr>
<td>C. Quasi-Likelihood Model</td>
<td>Use a quasi-binomial likelihood to fit a model with the logit of the number of defects as the response and dose group as the predictor using a logit link using a .</td>
<td>Yes</td>
</tr>
<tr>
<td>D. GLM model with random effects</td>
<td>Fit a GLM binomial model with fixed effects for dose and a random effect for dam.</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 7.4: Some Models for the Dams and Pups example.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Account for correlation?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. $\chi^2$ test</td>
<td>Construct a table of counts of defects and non-defects for each dose level. Use the marginal proportions to obtain expected values for each table entry. Calculate the $\chi^2$ test statistic.</td>
<td>No</td>
</tr>
<tr>
<td>B. GLM model</td>
<td>Use a likelihood to fit a GLM binomial model with the logit of the number of defects as the response and dose group as the predictor.</td>
<td>No</td>
</tr>
<tr>
<td>C. Quasi-Likelihood Model</td>
<td>Use a quasi-binomial likelihood to fit a model with the logit of the number of defects as the response and dose group as the predictor using a logit link using a .</td>
<td>Yes</td>
</tr>
<tr>
<td>D. GLM model with random effects</td>
<td>Fit a GLM binomial model with fixed effects for dose and a random effect for dam.</td>
<td>Yes</td>
</tr>
</tbody>
</table>

7.3.3.2 Modeling Options

Given the form of this data, there are a several options for modeling. Several possibilities are described in Table 7.4. One approach is to perform a $\chi^2$ test. Another more flexible approach would be to construct a generalized linear model. A GLM model would allow for more covariates, however in and of itself a GLM model does not account for correlation. There are ways in which to modify a GLM model to take into account correlation within litter. One way is to incorporate a dispersion parameter. This model would not be a GLM and would not be fit using a likelihood. Instead it is using a quasi-likelihood. Another approach is to use a random effects model. We discuss the GLM model and the quasi-likelihood model in this section and postpone discussion of the random effects models for later chapters.

Model B can be thought of as the naive model. The data structure implying correlation is ignored. How do these results compare to the model C which takes the correlation into account?
7.3.3.3 A Naive Model

A portion of the output for the naive model (B) appears below. Initially we fit a model linear in the logit without taking into account correlation. Note that the dispersion parameter is assumed to be 1.0 implying we are (incorrectly) not taking into account correlation.

```
# Model B GLM with Binomial Model
# assuming a logit linear effect in dose

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -0.7172  | 0.2488     | -2.883  | 0.003939 ** |
| dose           | 0.6932   | 0.1937     | 3.580   | 0.000344 *** |
```

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 30.773 on 11 degrees of freedom
Residual deviance: 17.179 on 10 degrees of freedom
AIC: 57.295

The estimated dose coefficient is highly significant providing support for the model of the logit model with a linear term of dose. But how well does this model fit? Is there evidence of overdispersion, specifically does there appear to be a significant dam effect? Recall that the residual deviance provides an indication of how much the data differs from the model. When the model fits poorly, the residual deviance will be large relative to the corresponding degrees of freedom. When it fits well, we expect the residual deviance to be close to the residual degrees of freedom. Here we have a residual deviance larger than the residual degrees of freedom. We can test whether this difference is significant by performing a goodness-of-fit test. Comparing the observed residual deviance to a $\chi^2$-distribution, we find evidence of lack-of-fit ($p=0.0704$).

Three reasons for lack-of-fit may be:
1. outliers or extreme observations,
2. an inadequate model which is missing important covariates, or
3. overdispersion.
First we check the residuals. Figure 7.2 does not implicate any particular observation as extreme. Next we can safely assume that there are no critical covariates missing because rat dams were randomly assigned to dose group. So dispersion is the remaining concern and the data structure suggests that we should have taken correlation into account at the start. Model C does this with the following results. Theoretically the foundation for fitting these models is likelihood methods, however it is not a true likelihood (hence quasi-likelihood) because of our adjustment of the variance.

The data are based on a (simulated) experiment which has dose group as its only covariate. Hence we have not omitted any important covariates. Our final consideration is overdispersion. The results of fitting a model linear in dose using a quasi-likelihood to accommodate a dispersion parameter appears below.

### Model C GLM with quasibinomial option

> # accounting for overdispersion

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -0.7172 | 0.3243 | -2.212 | 0.0514 . |
| dose      | 0.6932 | 0.2524 | 2.746 | 0.0206 * |

(Dispersion parameter for quasibinomial family taken to be 1.69886)

Null deviance: 30.773 on 11 degrees of freedom
Residual deviance: 17.179 on 10 degrees of freedom
AIC: NA

A couple of points can be made comparing Model B to Model C results. First, note that the estimated coefficients are the same. What differs are the standard errors. The standard errors for the the coefficients from the quasi-likelihood model are larger than the naive model. This is not surprising given the results of our exploratory analysis. The variances of the counts were much larger than the model-based variances. With the larger standard errors, the significance of the coefficients will be smaller and the confidence intervals wider. Also note that unlike the naive model which (incorrectly in this case) assumes that the dispersion parameter is one, the quasi-likelihood model estimates a dispersion parameter. When the estimated dispersion parameter if greater than one there is evidence of overdispersion. Here it is greater than 1.0 ($\hat{\phi} = 1.7$) consistent with our exploratory results.

The estimated dispersion parameter has been used to adjust the standard errors in the following way.

$$se\text{adj}(\hat{\beta}) = \sqrt{\hat{\phi}}se(\hat{\beta})$$

So for the dose term, the standard error changes from 0.1937 to $\sqrt{1.7} \times 0.1937$ or 0.2524. The standard error for the remaining coefficients are adjusted similarly. This results in larger p-values and a more conservative analysis. In addition to adjusting standard errors the drop-in-deviance test can also be adjusted in the following way:

$$F\text{-statistic} = \frac{\text{DropInDeviance}/d}{\hat{\phi}}.$$ 

where $d$ is the difference in the number of terms for the two models. The statistic itself is compared to an $F$-distribution with $d$ and $n - d$ degrees of freedom. There is no solid theory to
7.4 Case Study: Tree Growth, Random Effects Model

A student research team at St. Olaf College contributed to the efforts of biologist, Kathy Shea, to investigate a rich data set concerning forestation in the surrounding land. Much of south-central Minnesota was comprised of maple-basswood forest prior to agricultural development and settlement. Currently, the forests that remain are highly fragmented and disturbed by human activities. Original land surveys describe the forest, also known as the Big Woods ecosystem, as dominated by elm, maple, oak, and basswood. In order to recover the loss of ecosystem services that forests provide, many new forested areas have been established through restoration.

Tubes were placed on trees in some locations or transects but not in others. Interest centers on whether tree growth in the first year is affected by the presence of tubes. We see that this analysis has a structure similar to the dams and pups. A diagram in Figure 7.3 depicts the study design for tree growth. This diagram illustrates how the tree data is very similar in structure to the teratology data set. Some transects were assigned to have tubes on all of their trees and other transects had no tubes just as dams were assigned to dose groups. Within a transect, trees’ first year of growth was measured much like the presence or absence of a defect was noted for pups within dams. As we did for the dams and pups, let’s examine the sources of variability in first year growth for this study.

7.4.1 Format of the data set

Each row for the data set for this analysis contains information for a unique tree:

- **id**: a unique identifier for each black walnut tree within transect
- **transect**: denotes a numbered transect for the location in which the tree grew
- **tubes**: an indicator variable for the presence or absence of tubes for a given transect
- **height91**: first year height for each tree in meters
- **height90**: baseline height for each tree in meters
- **growth.yr1**: height91 - height90, in meters

A small portion of the data set appears below.

<table>
<thead>
<tr>
<th>transect</th>
<th>tube</th>
<th>growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.2140809</td>
</tr>
</tbody>
</table>
This portion of the data indicates that the four trees in transect one do not have tubes, while the two trees in transect two have tubes. The concern with this kind of data structure is that trees from the same transect may be more similar or correlated with one another in contrast to trees from other transects. This could be true for a variety of reasons, some transects may receive less sun than others or irrigation of the soil may differ from transect to transect. These unmeasured but possibly influential factors may imply that transects induce a correlation among trees within transects and in fact we do not have independent pieces of information, but fewer than our sample size would suggest. To determine the extent to which this may occur and to perform our final analysis we examine the sources of variability of the first year tree growth.

7.4.2 Sources of variability

First year tree growth may vary because of

**Tube effects** A purpose of this analysis is to determine whether tubes affect first year tube growth. Differences in the mean growth based on the presence or absence of tubes would be of interest and
Table 7.5: Summary Statistics for Black Walnut Tree Growth by Transect

<table>
<thead>
<tr>
<th>Transect</th>
<th>Mean Growth</th>
<th>Var Growth</th>
<th>Number of Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2031</td>
<td>0.01203</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>0.5172</td>
<td>0.01145</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>0.3893</td>
<td>0.01553</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>0.5818</td>
<td>0.02168</td>
<td>10</td>
</tr>
</tbody>
</table>

would be included in a publication for this analysis. For this reason, tube effects are referred to as **fixed effects**. This is analogous to the dose effect in the dams and pups example.

**Transect effects** For some of the factors previously mentioned such as sun exposure or water availability, first year growth may vary. A difference in the mean growth in heights per transect is not of interest and would not appear in a publication of this study. These **random effects** are analogous to dam effects which were not of interest.

**Tree-to-tree variability within transects** There is inherent variability in tree growth even when they are subject to the same transect and dose effects. This variability remains unexplained in our model.

In this case study, the data consists of heights of Black Walnut trees in transects with and without tubes. Note that height measurements are continuous in contrast to the binary observations we had for the pups. The structure of this data requires that we use models that account for correlation. Instead of using the estimated overdispersion parameter as we did with the dams and pups, here with the continuous tree growth measurements we will use random effects.

There are multiple tree growth measurements per PSU (transects). Data sets with this kind of structure are often referred to as **multilevel data** and the remaining chapters delve into models for multilevel data in gory detail. We postpone the analysis that considers more species until later when more elaborate multilevel models are presented and begin with an analysis of first year growth data for the Black Walnut species only with or without the tubes.

### 7.4.3 Exploratory Data Analysis

If the trees within a transect are more similar than those in other transects, we would expect the trees within a transect to vary less that the variation we would see using all the transects. Here we see that the variance for the ten trees in transect one is \( \text{var}(c(0.214, 0.345, \ldots, 0.445)) = 0.012 \) whereas the variability in tree growth for all of the 10 trees in transect 3 is 0.0155. This provides evidence of correlation among outcomes within transects. With this continuous data, we can calculate an **intraclass correlation coefficient** which quantifies an indicator of the extent of the correlation within classes.

\[
\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}
\]

The means, standard deviations and number for tree growth measurements for black walnut trees without tubes appears in Table 7.4.3. This data is in turn used to calculate the intraclass correlation coefficient. Using software, we can calculate of the intraclass correlation coefficient for the black walnut data without tubes. We will encounter this quantity again in the multilevel chapters.

```r
# intraclass correlation
```
library(ICC)
> ICCbare(transect, growth, data=sim.df)
$ICC
  0.634672

7.4.4 Fitting Models

In order to take into account the correlation in the data, we include a random effect for each transect since they are not of general interest. Note that variation in transects mean tree growth may be of interest and we are able to recover an estimate of the variation in tree growth across transects with the following models.

# Naive Model: fixed effects only
> model.naive = lm(growth~tube)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.29618    0.03121   9.49   1.43e-11 ***
tube        0.25335    0.04414   5.74   1.29e-06 ***
---

> # Model accounting for correlation: random effects

> transect.f = as.factor(transect)
> model.re = model.re=lmer(growth~1+ tube + (1|transect.f),data = sim.df)

Random effects:
  Groups     Name Variance Std.Dev.
  transect.f (Intercept)  0.008194  0.09052
  Residual              0.015169  0.12316
Number of obs: 40, groups: transect.f, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.29618    0.06968   4.250
tube         0.25335    0.09855   2.571

Correlation of Fixed Effects:
  (Intr)
tube -0.707

Like we saw in the case of the binary outcomes, the standard error for the coefficients is larger when we take correlation into account. The t-statistic for tubes is smaller reducing our enthusiasm for the tubes effect. This conservative approach occurs because the observations within a transect are correlated and therefore not independent as assumed in the naive model. The variation in the estimated random effects is reported to be 0.0082. In succeeding chapters, we address the meaning of this result and other results related to the random effects.
7.5 Summary

The most important idea from this chapter is that structures of datasets may imply that outcomes are correlated. Correlated outcomes provide less information than independent outcomes resulting in effective sample sizes that are less than the number observed. Neglecting to take into account correlation may lead to underestimating the standard errors, overstating significance and precision. In summary, when should correlation be accounted for? Correlation is likely and should be accounted for if:

1. Each primary sampling unit (PSU) generates multiple responses.
2. PSU characteristics vary.
3. Many PSU characteristics are unmeasured.
4. Responses are related to PSU characteristics.

We have taken a look at two ways in which to account for correlation: incorporate a dispersion parameter and include random effects. There are other ways in which to account for correlation including inflating the variance using Huber-White estimator, aka Sandwich estimator, and producing corrected variances using bootstrapping. These are beyond the scope of this text.

7.6 R Code

R-code for Dams and Pups example

# Model B
model.b <- glm(cbind(defects,noDefects)~dose, family=binomial, data=corr.df)
summary(model.b)
1-pchisq(model.b$deviance,model.b$df.residual)

# Model C
model.c <- glm(cbind(defects,noDefects)~dose, family=quasibinomial, data=corr.df)
summary(model.c)

7.7 Exercises

7.7.1 Conceptual

1. Types of Data Structures For each of the following studies, identify the type of data structure (clustered, repeated measures, longitudinal, or spatial). Identify the PSU, the elements, and the responses measured. Write a sentence describing the within-PSU correlation. Explain.

(a) Example: Blood pressure is measured weekly for participants in a study.

   Answer:
   i. This could be considered longitudinal data because outcomes are collected over time.
   ii. Participants are the PSUs.
   iii. The elements are the weeks.
   iv. The responses are blood pressure measurements.
   v. Blood pressures on the same subject are likely to be correlated.

(b) Study: Voting Voter turnout for a recent presidential election was recorded for every polling place in precincts in Florida.

(c) Study: Car accidents The monthly number of car accidents at fifteen intersections during 2013.
### Table 7.6: Child growth data. Need citation-Fitzmaurice?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sex</th>
<th>Age8</th>
<th>Age10</th>
<th>Age12</th>
<th>Age14</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>21.0</td>
<td>20.0</td>
<td>21.5</td>
<td>23.0</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>21.0</td>
<td>21.5</td>
<td>24.0</td>
<td>25.5</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>20.5</td>
<td>24.0</td>
<td>24.5</td>
<td>26.0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>M</td>
<td>22.0</td>
<td>21.5</td>
<td>23.5</td>
<td>25.0</td>
</tr>
</tbody>
</table>

(d) **Study: Census** Census researchers randomly selects citizens from census tracts to report their income.

(e) **Study: Skaters** Each figure skater receives scores for 5 judges.

(f) **Study: Water supply** The water supply for a village comes from three different streams each which are distributed by five pumps. Fecal coliform measurements are made on two samples from each pump.

2. **Type of Responses** For each of the following situations, decide whether the best model for the data would be ordinary linear regression, logistic regression, or Poisson regression, and justify your decision. Should the models account for correlation? Why or why not?

(a) A researcher wants to investigate whether the number of hours a person works in a month affects the number of over-the-counter medications he or she uses. Participants are recruited from a variety of types of employment. They agree to keep a log of time at work and number of times and types of over-the-counter medication they used, for one month.

(b) A researcher wants to investigate the relationship between a wife’s education level and divorce. Using marriage records, she selects a sample of couples who were married in 1990-1995, locates them, and confirms who is still married. She also finds out how many years of education the wife had at the time of marriage.

(c) The same researcher as in (b) decides to also study the relationship between the length of marriage (in years) and the number of years of education the wife had at the time of marriage, using the same data.

(d) A randomly selected group of high school students are observed through one flu season (November through March) to see if there is any relationship between susceptibility to flu and time spent in team sports.

3. **More dams and pups** Describe how to generalize the pup and dam example by allowing for different size litters.

#### 7.7.2 Guided

1. **More dams and pups** Interpret the estimated coefficient for Models B and C. Compare the t-ratios for the coefficient. Construct a confidence interval for the coefficient for Models B and C and compare.

2. **Child Health Data needed.** Growth measurements were taken on 11 girls and 16 boys, at 4 different times; age=8, 10, 12, and 14. (See complete data listing.)

3. **Crime** Police Stop by Ethnic Group, pages 112-113 in G & H.
   (a) Interpret coefficients.
   (b) Drop in Deviance.
   (c) Exposure number of arrests- others possible.
   (d) Overdispersion.
EXERCISES 141

(e) Negative binomial.

4. **Fish Tanks and Tumors**

Research Description: Aflatoxin and Fish Tumors in Trout. An experiment at the Marine/Freshwater Biomedical Sciences Center at Oregon State University investigated the carcinogenic effects of aflatoxical, a metabolite of Aflatoxin B1, which is a toxic by-product produced by a mold that infects cottonseed meal, peanuts, and grains. Twenty tanks of rainbow trout embryos were exposed to one of five doses of aflatoxical for one hour. The data in fishtumor.xls contains the dose of aflatoxical, the number of fish with tumors, and the total number of fish in each tank. *Source: Baily and Hendricks.*

(a) The data appear below.

(b) Plot the empirical logit of probability of a tumor by dose. How does it look?

```r
> cbind(dose,tumor,total)
dose  tumor total
[1,] 0.010 9 87
[2,] 0.010 5 86
[3,] 0.010 2 89
[4,] 0.010 9 85
[5,] 0.025 30 86
[6,] 0.025 41 86
[7,] 0.025 27 86
[8,] 0.025 34 88
[9,] 0.050 54 89
[10,] 0.050 53 86
[11,] 0.050 64 90
[12,] 0.050 55 88
[13,] 0.100 71 88
[14,] 0.100 73 89
[15,] 0.100 65 88
[16,] 0.100 72 90
[17,] 0.250 66 86
[18,] 0.250 75 82
[19,] 0.250 72 81
[20,] 0.250 73 89
```

(c) Fit a GLM model with the log odds of tumor as the dependent variable and logdose as the explanatory variable. You will need to identify "total" as the number of trials variable.

(d) Test the goodness-if-fit of the model.

Ho: No lack-of-fit
Ha: Lack-of-fit

Test Statistic:
p-value=
Conclusion:

(e) What is a possible cause of this lack-of-fit?

i. Plot the deviance residuals against logdose. What do you see?

ii. Create a \((logdose)^2\) variable. Fit the model with \(logdose\) and \((logdose)^2\) as explanatory variables. Now test for lack-of-fit.

Ho: No lack-of-fit
Ha: Lack-of-fit

Test Statistic:
p-value=

Conclusion:

iii. Check the deviance residuals from this new model. How does it look? Can we attribute any problems with goodness-of-fit to a particular troublesome point or two or to overdispersion?

iv. Let’s adjust this last model for overdispersion (even though it is a close call). You can adjust for overdispersion by re-running the model with family = quasibinomial.

A. Find the estimated dispersion parameter on the output, \( \hat{\phi} \).

B. How do the coefficients from the previous model compare to this model?

C. How do the standard errors for the coefficients from the previous model compare to this model?

D. How do the Wald Test statistics and corresponding p-values from the previous model compare to this model?

E. Compute confidence intervals for the parameters from the previous model compare to this model.

F. Speculate on why the standard errors and confidence intervals changed from the GLM to quasi-likelihood model.

G. Perform a Drop-in-deviance test for both of the logdose terms by with and without adjusting for overdispersion. How do the results compare?

v. Continuous example with random effects and intraclass correlation calculation and interpretation.

vi. Extra Credit: There are other ways to adjust for extra-binomial variation. Two possibilities we will discuss later in the course are:

A. Robust variance estimator aka ”An unbiased sandwich estimator”.

B. Bootstrap estimator. Use 500 reps.

5. Simulating Correlated Binary Outcomes: Beta-Binomial

(a) Simulating Control Data

We begin by creating two datasets using the dams and pups example. One dataset has independent observations, the other has correlated observations. The two datasets consist of 6 dams who each have litters of 10 pups.

i. Model-based: The number of defects per litter is binomial with \( n = 10 \) and each dam produces defective pups with a common probability, \( p = 0.119 \). The observations are independent of one another.

ii. Overdispersed: Each dam gets her own probability of producing defective pups. The observations are correlated.

(b) Produce the model-based dataset of six independent binomial observations using the rbinom command with \( n=10 \) and \( p = 0.119 \).
(c) Produce the overdispersed dataset of correlated observations by selecting different \( p \)'s for each dam. The following will help you to do this.

For the overdispersed dataset we need to select 6 different probabilities for the 6 dams. This could be done in several ways, but selecting probabilities from a beta distribution has some useful theoretical properties. First we investigate the properties of a beta distribution to see why it is a good choice for selecting \( p \)'s. This information can also be found in the chapter on distributions.

6. **Detour: Learning about Beta Distributions**

(a) Plotting beta densities

```r
p = seq(0, 1, by=0.05)
# To plot a beta density use dbeta
# Here I selected a=5, b=1
density = dbeta(p, 5, 1)
plot(p, density, type="l")
# Try some other values for a>0 and b>0
# to see other forms for the Beta distribution.
```

(b) What values do Beta random variables take on?

(c) What do these values represent for the dams and pups simulation?

(d) Do the possible values depend on \( a \) or \( b \)?

(e) What is a feature of the Beta density when \( a = b \)?

(f) What happens to the density when \( a \neq b \)?

(g) How does the magnitude of \( a \) or \( b \) affect the density?

(h) How does the difference between \( a \) and \( b \) affect the density?

(i) If you wanted to simulate dams with mostly low probabilities of defects and a few with very high probabilities, how would you do it?

(j) If you wanted to simulate dams with mostly high probabilities of defects and a few with very low probabilities, how would you do it?

(k) If you wanted to simulate a population of dams where half of the probabilities of defects are very high and half are very low, how would you do it?

(l) How might you decide on values for \( a \) and \( b \) if you have run a preliminary experiment and gathered data on the number of dams with deformed pups?

Returning to the second simulated sample of counts of defective pups.

(a) Select the beta distribution of your choice by selecting values for its parameters \( a \geq 0 \) and \( b \geq 0 \).

(b) Plot the density for this distribution and describe what it would imply for deformation rates in contrast to other choices for \( a \) and \( b \).

(c) Produce six counts for the six dams by selecting six \( p \)'s from a beta distribution of your choosing, that is, choose \( a \) and \( b \). These counts are correlated.

```r
p.dam = rbeta(6, a, b)
defects = rbinom(6, n=10, p.dam)
```

Now create another overdispersed dataset using the beta distribution with \( a = 0.27 \) and \( b = 2 \) to select 6 values of \( p \). The mean of a beta distribution is \( a/(a+b) \) or here, 0.119. Produce 6 counts of the number of defective pups using the \( p \) values you selected from the this beta distribution.

(d) Calculate the variance of the six counts \( (s^2) \) and the model based variance \( 10 \times 0.119^2 (1 - 0.119) \) and compare.
7. **Simulating Control and Dose Data**

Next we add the dose groups to the controls and produce two data sets, again model-based data and overdispersed data. For these simulations we will have six dams with 10 pups exposed to one of the dose levels of 1/3, 2/3, 1 and a control group who experiences no exposure. The effect of the dose will be generated using

\[
\text{logit} = -2 + 4 \times \text{dose}
\]

This implies that the dose-specific probabilities of a defective pup are 0.119, 0.339, 0.661, and 0.881 for the controls, dose group levels 0, 1/3, 2/3, 1 respectively.

(a) **Model-based Data:** Create a dataset of 24 counts using the dose-specific probabilities of the form \(\text{rbinom}(6, 10, 0.119)\), \(\text{rbinom}(6, 10, 0.339)\), etc.

(b) **Overdispersed Data:** Select dam-specific \(p\)'s from a beta distribution with \(b=2\) and \(a=b\times\text{dose-specific } p / (1-\text{dose-specific } p)\) (This formulation of \(a\) and \(b\) will provide averages equal to the dose-specific probabilities.) Each dose level will have a different beta distribution defined by \(b=2\) and \(a=b\times\text{dose-specific } p / (1-\text{dose-specific } p)\). Select 6 \(p\)'s for each of these dose-specific beta distribution and generate the number of defects for the 6 dams at that dose level.

(c) Fit a model with the logit of the number of defects linear on dose for each data set. Compare the t-ratios for the dose coefficient and comment.

(d) Construct a confidence interval for the coefficient for dose using the Model-based Data and the Overdispersed Data. Do the intervals capture the true dose coefficient used to generate the data? Comment on the differences of the two confidence intervals.

7.7.3 **Note on Correlated Binary Outcomes**

The correlated binomial counts simulated here are in fact beta-binomial random variables like those simulated in the section in Chapter 3 Beta-Binomial. In fact, we could use the form of a beta-binomial pdf to model overdispersed binomial variables. Unlike the more generic form of accounting for correlation using dispersion parameter estimates, beta-binomial models are more specific and highly parameterized. This approach involves more assumptions but may also yield more information than the quasi-likelihood approach. However if the beta-binomial model is incorrect, our results may be misleading. That said, the beta-binomial structure is quite flexible and conforms to many situations.

7.8 **References**

(a) James Lepkowski SAGE

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.

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.
```markdown
<table>
<thead>
<tr>
<th>id</th>
<th>diary</th>
<th>perform_type</th>
<th>memory</th>
<th>audience</th>
<th>na</th>
<th>gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Solo</td>
<td>Unspecified</td>
<td>Instructor</td>
<td>11</td>
<td>Female</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Large Ensemble</td>
<td>Memory</td>
<td>Public Performance</td>
<td>19</td>
<td>Female</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Large Ensemble</td>
<td>Memory</td>
<td>Public Performance</td>
<td>14</td>
<td>Female</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>495</td>
<td>43</td>
<td>Solo</td>
<td>Score</td>
<td>Instructor</td>
<td>13</td>
<td>Female</td>
</tr>
<tr>
<td>496</td>
<td>43</td>
<td>Small Ensemble</td>
<td>Memory</td>
<td>Juried Recital</td>
<td>19</td>
<td>Female</td>
</tr>
<tr>
<td>497</td>
<td>43</td>
<td>Solo</td>
<td>Score</td>
<td>Instructor</td>
<td>11</td>
<td>Female</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>diary</th>
<th>instrument</th>
<th>mpqab</th>
<th>mpqpem</th>
<th>mpqnem</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>voice</td>
<td>16</td>
<td>52</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>voice</td>
<td>16</td>
<td>52</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>voice</td>
<td>16</td>
<td>52</td>
<td>16</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>495</td>
<td>43</td>
<td>voice</td>
<td>31</td>
<td>64</td>
<td>17</td>
</tr>
<tr>
<td>496</td>
<td>43</td>
<td>voice</td>
<td>31</td>
<td>64</td>
<td>17</td>
</tr>
<tr>
<td>497</td>
<td>43</td>
<td>voice</td>
<td>31</td>
<td>64</td>
<td>17</td>
</tr>
</tbody>
</table>

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

- Lower music performance anxiety will be associated with greater number of years of study.

8.3 Initial Exploratory Analyses

8.3.1 Data Organization

Our examination of the data from Sadler and Miller (2010) in this chapter will focus on the following key variables:
- id = unique musician identification number
- diary = cumulative total of diaries filled out by musician
- perform 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), the same 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.
We can also summarize categorical Level One covariates across all (possibly correlated) observations to get a rough relative comparison of trends. 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.

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 the variation in the response differs at different levels of the covariate, and (3) if subjects at certain levels of the covariate tend to have similar mean responses.

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 the next chapter. 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 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.

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

Armed with any statistical software package, it would be relatively simple to take our complete data set of 497 observations and run an OLS multiple linear 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...
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.

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.

Instrument? and, Was the performance a large ensemble?) is presented below. Do you see any problems with this approach?

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 15.7212  | 0.3591     | 43.778  | < 2e-16  *** |
| orch           | 1.7887   | 0.5516     | 3.243   | 0.00126  ** |
| large          | -0.2767  | 0.7910     | -0.350  | 0.72662  |
| orch:large     | -1.7087  | 1.0621     | -1.609  | 0.10831  |
Residual standard error: 5.179 on 493 degrees of freedom  
Multiple R-squared: 0.02782, Adjusted R-squared: 0.0219  
F-statistic: 4.702 on 3 and 493 DF, p-value: 0.003012

Other than somewhat skewed residuals, residual plots (not shown) do not indicate any major problems with the OLS multiple regression 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 her first performance, chances are good that she will have relatively high anxiety levels before subsequent performances. Thus, OLS multiple linear 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 OLS 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 OLS multiple linear regression applied to only Musician #22’s data, using appropriate indicator variables for factors of interest.

Let $Y_j$ 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:

$$Y_j = \lambda_0 + \lambda_1 \text{LargeEns}_j + \varepsilon_j \text{ where } \varepsilon_j \sim N(0, \sigma^2) \text{ and}$$

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

The parameters in this model ($\lambda_0$, $\lambda_1$, and $\sigma^2$) can be estimated through OLS methods. $\lambda_0$ represents the true intercept for Musician #22—the expected anxiety score for Musician #22 when performance type is a Solo or Small Ensemble ($\text{LargeEns} = 0$), or the true average anxiety for Musician #22 over all Solo or Small Ensemble performances he may conceivably give. $\lambda_1$ 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 $\epsilon_j$ 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 $\sigma^2$. 

For Subject 22, we estimate $\hat{\lambda}_0 = 24.5, \hat{\lambda}_1 = -7.8$, 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 ($t(13)=-3.10, p=.009$).

<table>
<thead>
<tr>
<th>Coefficients:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate Std. Error t value Pr(&gt;</td>
</tr>
<tr>
<td>(Intercept) 24.500 1.960 12.503 1.28e-08 ***</td>
</tr>
<tr>
<td>large -7.833 2.530 -3.097 0.0085 **</td>
</tr>
</tbody>
</table>

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 4.8 on 13 degrees of freedom
Multiple R-squared: 0.4245, Adjusted R-squared: 0.3802
F-statistic: 9.588 on 1 and 13 DF, p-value: 0.008504

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-square 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.
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 OLS regression 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 ($\lambda_{0i}$) and fitted slopes ($\lambda_{1i}$) for Subject $i$:

$$\lambda_{0i} = \beta_{00} + \beta_{10} \text{Orch}_i + \epsilon_{0i} \quad (8.2)$$
$$\lambda_{1i} = \beta_{10} + \beta_{11} \text{Orch}_i + \epsilon_{1i} \quad (8.3)$$

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, 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 slight 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: $\hat{\beta}_{00} = 16.3$, $\hat{\beta}_{01} = 1.4$, $\hat{\beta}_{10} = -0.8$, and $\hat{\beta}_{11} = -1.4$. Thus, the predicted intercept ($\lambda_{0i}$) and predicted slope ($\lambda_{1i}$) for Subject $i$ can be modeled as:

$$\hat{\lambda}_{0i} = 16.3 + 1.4Orch_i + e_{0i}$$

$$\hat{\lambda}_{1i} = -0.8 - 1.4Orch_i + e_{1i}$$

where $\lambda_{0i}$ is the true mean negative affect when Subject $i$ is playing solos or small ensembles, and $\lambda_{1i}$ 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 large ensembles, vocalists and instrumentalists have performance anxiety which is 0.8 points lower, on average, than before solos and small ensembles (15.5), while subjects playing orchestral instruments experience an average difference of 2.2 points, producing an average performance anxiety of 15.5 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 ($t=-1.168$, $p=0.253$).

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 over the remainder of this section.
8.5 Two level modeling: a unified approach

8.5.1 The framework we’ll use

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:**
  \[ Y_{ij} = \lambda_{0i} + \lambda_{1i} \text{LargeEns}_{ij} + \epsilon_{ij} \]  
  (8.6)

- **Level Two:**
  \[ \lambda_{0i} = \beta_{00} + \beta_{01} \text{Orch}_i + \epsilon_{0i} \]
  \[ \lambda_{1i} = \beta_{10} + \beta_{11} \text{Orch}_i + \epsilon_{1i} \]

In this system, there are 4 key **fixed effects** to estimate: $\beta_{00}$, $\beta_{01}$, $\beta_{10}$, and $\beta_{11}$. Fixed effects are the fixed but unknown population effects associated with certain covariates. The intercepts and slopes for each subject from Level One, $\lambda_{0i}$ and $\lambda_{1i}$, 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 $\lambda_{0i}$ and $\lambda_{1i}$:

\[ Y_{ij} = \left[ \beta_{00} + \beta_{01} \text{Orch}_i + \beta_{10} \text{LargeEns}_{ij} + \beta_{11} \text{Orch}_i \text{LargeEns}_{ij} \right] + \left[ \epsilon_{0i} + \epsilon_{1i} \text{LargeEns}_{ij} + \epsilon_{ij} \right] \]

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 ($\text{Orch}=0$) and those who play orchestral instruments ($\text{Orch}=1$):

- **Keyboardists and Vocalists ($\text{Orch}=0$)**
  \[ \lambda_{0i} = \beta_{00} + \epsilon_{0i} \]
  \[ \lambda_{1i} = \beta_{10} + \epsilon_{1i} \]

- **Orchestral instrumentalists ($\text{Orch}=1$)**
  \[ \lambda_{0i} = \left( \beta_{00} + \beta_{01} \right) + \epsilon_{0i} \]
  \[ \lambda_{1i} = \left( \beta_{10} + \beta_{11} \right) + \epsilon_{1i} \]

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:** $\beta_{00}$
- **Keyboardists or vocalists playing large ensembles:** $\beta_{00} + \beta_{10}$
- **Orchestral instrumentalists playing solos or small ensembles:** $\beta_{00} + \beta_{01}$
- **Orchestral instrumentalists playing large ensembles:** $\beta_{00} + \beta_{01} + \beta_{10} + \beta_{11}$
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 $e_{ij}$, $e_{0i}$, and $e_{1i}$ 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: the multivariate normal distribution

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: $e_{ij} \sim N(0, \sigma^2)$. At Level Two, we have one error term ($e_{0i}$) associated with subject-to-subject differences in intercepts, and one error term ($e_{1i}$) associated with subject-to-subject differences in slopes. That is, $e_{0i}$ represents the deviation of Subject $i$ from the mean performance anxiety before solos and small ensembles after accounting for his or her instrument, and $e_{1i}$ represents the deviation of Subject $i$ from the mean difference in performance anxiety between large ensembles and other performance types after accounting for his or her instrument.

In modeling the random behavior of $e_{0i}$ and $e_{1i}$, 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).

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:

$$\begin{bmatrix} e_{0i} \\ e_{1i} \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \rho_{01} \sigma_0 \sigma_1 \\ \rho_{01} \sigma_0 \sigma_1 & \sigma_1^2 \end{bmatrix} \right)$$

Note that the correlation $\rho_{01}$ between the error terms is simply the covariance $\sigma_{01}$ converted to a $[-1, 1]$ scale through the relationship:

$$\rho_{01} = \frac{\sigma_{01}}{\sigma_0 \sigma_1} \quad (8.7)$$
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 (\( \sigma_0^2 \) and \( \sigma_1^2 \)) and 1 covariance term (\( \sigma_{01} \)).

The multivariate normal distribution with \( n = 2 \) is illustrated in Figure 8.12 for two cases: (a) the error terms are uncorrelated (\( \sigma_{01} = \rho_{01} = 0 \)), and (b) the error terms are positively correlated (\( \sigma_{01} = \rho_{01} > 0 \)). In general, if the errors in intercepts (\( e_{0i} \)) are placed on the x-axis and the errors in slopes (\( e_{1i} \)) are placed on the y-axis, then \( \sigma_0^2 \) measures spread in the x-direction and \( \sigma_1^2 \) measures spread in the y-direction, while \( \sigma_{01} \) measures tilt. Positive tilt (\( \sigma_{01} > 0 \)) indicates a tendency for errors from the same subject to both be positive or both be negative, while negative tilt (\( \sigma_{01} < 0 \)) indicates a tendency for one error from a subject to be positive and the other to be negative. In Figure 8.12, \( \sigma_{01}^2 = 4 \) and \( \sigma_1^2 = 1 \), so both contour plots show a greater range of errors in the x-direction than the y-direction. Internal ellipses in the contour plot indicate pairs of \( e_{0i} \) and \( e_{1i} \) that are more likely. In Figure 8.12(a) \( \sigma_{01} = \rho_{01} = 0 \), so the axes of the contour plot correspond to the x- and y-axes, but in Figure 8.12(b) \( \sigma_{01} = 1.5 \), so the contour plot tilts up, reflecting a tendency for high values of \( \sigma_0^2 \) to be associated with high values of \( \sigma_1^2 \).

### 8.5.4 Technical issues when estimating and testing parameters (Optional)

Now, our two-level model has 8 parameters that need to be estimated: 4 fixed effects (\( \beta_{00}, \beta_{01}, \beta_{10}, \) and \( \beta_{11} \)), and 4 variance components (\( \sigma^2, \sigma_0^2, \sigma_1^2, \) and \( \sigma_{01} \)). 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 maximum likelihood (ML) was introduced
in Chapter 2, where parameter estimates are chosen to maximize the value of the likelihood function based on observed data. Restricted maximum likelihood (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, but references can be found at the end of this chapter (p.177 MS; p.85 SW).

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. 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 10.6, 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. Once again, additional details can be found in the references at the end of this chapter.
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:

A) Formula: na ~ orch + large + orch:large + (large | id)

B) AIC  BIC  logLik  deviance  REMLdev
   3003  3037  -1493   2991   2987

Random effects:
Grp  Name  Variance  Std.Dev.  Corr
C)  id  (Intercept)  5.65534  2.37810
D)  large  0.45206  0.67236  -0.635
E)  Residual  21.80740  4.66984
F)  Number of obs: 497, groups: id, 37

Fixed effects:
Estimate  Std. Error  t value
G)  (Intercept) 15.9297  0.6415  24.833
H)   orch  1.6926  0.9452  1.791
I)   large -0.9106  0.8452 -1.077
J)  orch:large -1.4239  1.0992 -1.295

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 chapters, 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.
- C) Estimated variance components (\( \hat{\sigma}_0^2 \) and \( \hat{\sigma}_{00} \)) associated with the intercept equation in Level Two.
- D) Estimated variance components (\( \hat{\sigma}_1^2 \) and \( \hat{\sigma}_1 \)) associated with the large ensemble effect equation in Level Two, along with the estimated correlation (\( \hat{\rho}_{01} \)) between the two Level Two error terms.
- E) Estimated variance components (\( \hat{\sigma}_2^2 \) 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 (\( \hat{\beta}_0 \)) for 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 (\( \hat{\beta}_{01} \)) for the orchestral instrument effect, along with its standard error and t-value.
- I) Estimated fixed effect (\( \hat{\beta}_{10} \)) for the large ensemble effect, along with its standard error and t-value.
- J) Estimated fixed effect (\( \hat{\beta}_{11} \)) 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:**
  - \( \hat{\beta}_{00} = 15.9 \). The estimated mean performance anxiety for solos and small ensembles (Large=0) for keyboard players and vocalists (Orch=0) is 15.9.
  - \( \hat{\beta}_{01} = 1.7 \). Orchestral instrumentalists have an estimated mean performance anxiety for solos and small ensembles which is 1.7 points higher than keyboard players and vocalists.
  - \( \hat{\beta}_{10} = -0.9 \). 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.
  - \( \hat{\beta}_{11} = -1.4 \). Orchestral instrumentalists have an estimated mean decrease in performance anxiety of 2.3 points when playing in large ensembles instead of solos or small ensembles, 1.4 points greater than the mean decrease among keyboard players and vocalists.

- **Variance components:**
  - \( \hat{\sigma}_0 = 2.4 \). The estimated standard deviation of performance anxiety levels for solos and small ensembles is 2.4 points, after controlling for instrument played.
  - \( \hat{\sigma}_1 = 0.7 \). 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.
  - \( \hat{\rho}_{01} = -0.64 \). 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.
  - \( \hat{\sigma} \) from Level One = 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.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Independence</th>
<th>Two-stage</th>
<th>LVCF</th>
<th>Multilevel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>15.72 (0.36)</td>
<td>16.28 (0.67)</td>
<td>15.20 (1.25)</td>
<td>15.93 (0.64),</td>
</tr>
<tr>
<td>Orch</td>
<td>1.79 (0.55)</td>
<td>1.41 (0.99)</td>
<td>1.45 (1.84)</td>
<td>1.69 (0.95)</td>
</tr>
<tr>
<td>Large</td>
<td>-0.28 (0.79)</td>
<td>-0.77 (0.85)</td>
<td>-</td>
<td>-0.91 (0.85)</td>
</tr>
<tr>
<td>Orch*Large</td>
<td>-1.71 (1.06)</td>
<td>-1.41 (1.20)</td>
<td>-</td>
<td>-1.42 (1.10)</td>
</tr>
</tbody>
</table>

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 Byrk (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 (YEAR), 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: unconditional means or 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:**
  \[ Y_{ij} = \lambda_{0i} + \varepsilon_{ij} \text{ where } \varepsilon_{ij} \sim N(0, \sigma^2) \]  
  (8.8)
- **Level Two:**
  \[ \lambda_{0i} = \beta_{00} + \varepsilon_{0i} \text{ where } \varepsilon_{0i} \sim N(0, \sigma_{0}^2) \]  
  (8.9)

or as a composite model:

\[ Y_{ij} = \beta_{00} + \varepsilon_{0i} + \varepsilon_{ij} \]  
(8.10)

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 \( \lambda_{0i} \) is the true mean response of all observations for subject \( i \). On the other hand, \( \beta_{00} \) 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—\( \sigma^2 \) is the within-person variability, while \( \sigma_{0}^2 \) is the between-person variability. The name random intercepts model then arises from the Level Two equation for \( \lambda_{0i} \): each subject’s intercept is assumed to be a random value from a normal distribution centered at \( \beta_{00} \) with variance \( \sigma_{0}^2 \).
Using the composite model specification, the unconditional means model can be fit to the music performance anxiety data using statistical software:

Formula: na ~ 1 + (1 | id)

AIC  BIC  logLik  deviance  REMLdev
3012  3024 -1503  3006  3006

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>(Intercept)</td>
<td>4.9501</td>
<td>2.2249</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>22.4596</td>
<td>4.7392</td>
</tr>
</tbody>
</table>

Number of obs: 497, groups: id, 37

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>16.2370</td>
<td>0.4279</td>
</tr>
</tbody>
</table>

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

- $\hat{\beta}_0 = 16.2$ = the estimated mean performance anxiety score across all performances and all subjects.
- $\hat{\sigma}^2 = 22.5$ = the estimated variance in within-person deviations.
- $\hat{\sigma}_{0}^2 = 5.0$ = the estimated variance in between-person deviations.

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

$$\hat{\rho} = \frac{\text{Between-person variability}}{\text{Total variability}} = \frac{\hat{\sigma}_{0}^2}{\hat{\sigma}_{0}^2 + \hat{\sigma}^2} = \frac{5.0}{5.0 + 22.5} = 0.182.$$  (8.11)

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 0.182. As $\rho$ approaches 0, responses from an individual are essentially independent and accounting for the multilevel structure of the data becomes less crucial. However, as $\rho$ approaches 1, repeated observations from the same individual essentially provide no additional information and accounting for the multilevel structure becomes very important. With $\rho$ 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 $\rho$ 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 ??, we will include a binary covariate for performance type:

$$LargeEns_{ij} = \begin{cases} 
1 & \text{if perform type = Large Ensemble} \\
0 & \text{if perform type = Solo or Small Ensemble}.
\end{cases}$$

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
BINARY COVARIATES AT LEVEL ONE AND LEVEL TWO

Ensemble performances.) The resulting model, which we will denote as Model B, can be specified either using formulations at both levels:

- **Level One:**
  \[
  Y_{ij} = \lambda_{0i} + \lambda_{1i}LargeEns_{ij} + \varepsilon_{ij}
  \]  
  (8.12)

- **Level Two:**
  \[
  \begin{align*}
  \lambda_{0i} &= \beta_{00} + \varepsilon_{0i} \\
  \lambda_{1i} &= \beta_{10} + \varepsilon_{1i}
  \end{align*}
  \]

or as a composite model:

\[
Y_{ij} = [\beta_{00} + \beta_{10}LargeEns_{ij}] + [\varepsilon_{0i} + \varepsilon_{1i}LargeEns_{ij} + \varepsilon_{ij}]
\]

(8.13)

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

\[
\begin{bmatrix}
\varepsilon_{0i} \\
\varepsilon_{1i}
\end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\
0
\end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\
\sigma_{01} & \sigma_1^2
\end{bmatrix}\right).
\]

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 \(\varepsilon_{ij}\) terms capture the deviation between the true performance anxiety levels for subjects (based on performance type) and their observed anxiety levels. \(\beta_{00}\) is then the true mean performance anxiety level for Solos and Small Ensembles, and \(\beta_{10}\) is the true mean difference in performance anxiety for Large Ensembles compared to other performance types. As before, \(\sigma^2\) 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 (\(\sigma_0^2\)) and variability in differences with Large Ensembles (\(\sigma_1^2\)).

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

Formula: na ~ large + (large | id)

AIC  BIC logLik deviance REMLdev
3006  3031  -1497  2995  2994

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>(Intercept)</td>
<td>6.33295</td>
<td>2.5165</td>
<td></td>
</tr>
<tr>
<td></td>
<td>large</td>
<td>0.74288</td>
<td>0.8619</td>
<td>-0.759</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>21.77122</td>
<td>4.6660</td>
<td></td>
</tr>
</tbody>
</table>
Number of obs: 497, groups: id, 37

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>16.7297</td>
<td>34.09</td>
</tr>
<tr>
<td>large</td>
<td>-1.6761</td>
<td>-3.09</td>
</tr>
</tbody>
</table>

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

- \(\hat{\beta}_{00} = 16.7 = \) the mean performance anxiety level before solos and small ensemble performances.
\[ \hat{\beta}_{10} = -1.7 \] the mean decrease in performance anxiety before large ensemble performances.

\[ \hat{\sigma}^2 = 21.8 \] the variance in within-person deviations.

\[ \hat{\sigma}_0^2 = 6.3 \] the variance in between-person deviations in performance anxiety scores before solos and small ensembles.

\[ \hat{\sigma}_1^2 = 0.7 \] the variance in between-person deviations in increases (or decreases) in performance anxiety scores before large ensembles.

\[ \rho_{01} = -0.76 \] 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 the identical slope (0), but with varying intercepts (with variance 5.0).

![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.](image)

Figures 8.13 and 8.14 use **empirical Bayes estimates** for the intercepts (\( \hat{\lambda}_{0i} \)) and slopes (\( \hat{\lambda}_{1i} \)) 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 $\lambda_0$ involves both fixed and random components; further detail can be found in Raudenbush and Byrk (2002) and Singer and Willett (2003).

8.7.2 Pseudo $R^2$ 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-square value:

$$Pseudo \ R^2_{L1} = \frac{\hat{\sigma}^2(\text{Model A}) - \hat{\sigma}^2(\text{Model B})}{\hat{\sigma}^2(\text{Model A})} = \frac{22.5 - 21.8}{22.5} = 0.031$$ (8.14)

Values of $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ cannot be compared with 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, $\hat{\rho}_{01} = -0.76$ 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 $R^2$ 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 $R^2$ values differ or are negative. For this reason, pseudo $R^2$ 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:
  \[ Y_{ij} = \lambda_{0i} + \lambda_{1i} \text{LargeEns}_{ij} + \epsilon_{ij} \]  

- Level Two:
  \[ \lambda_{0i} = \beta_{00} + \beta_{01} \text{Orch}_i + \nu_{0i} \]
  \[ \lambda_{1i} = \beta_{10} + \beta_{11} \text{Orch}_i + \nu_{1i} \]

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

\[
\begin{bmatrix}
\epsilon_{0i} \\
\epsilon_{1i}
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_0 & \sigma_{01} \\ \sigma_{01} & \sigma^2_1 \end{bmatrix} \right).
\]

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 discernable 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-square values for both intercepts and slopes to evaluate the impact on between-person variability of adding instrument to Model B.

\[
PseudoR^2_{L2_0} = \frac{\hat{\sigma}^2_0 (\text{Model B}) - \hat{\sigma}^2_0 (\text{Model C})}{\hat{\sigma}^2_0 (\text{Model B})} = \frac{6.33 - 5.66}{6.33} = 0.106 \quad (8.16)
\]

\[
PseudoR^2_{L2_1} = \frac{\hat{\sigma}^2_1 (\text{Model B}) - \hat{\sigma}^2_1 (\text{Model C})}{\hat{\sigma}^2_1 (\text{Model B})} = \frac{0.74 - 0.45}{0.74} = 0.392 \quad (8.17)
\]

Pseudo \( R^2_{L2_0} \) describes the improvement in Model C over Model B in explaining subject-to-subject variability in intercepts, and Pseudo \( R^2_{L2_1} \) 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:
  \[ Y_{ij} = \lambda_{0i} + \lambda_{1i} \text{LargeEns}_{ij} + \epsilon_{ij} \]  

- Level Two:
  \[ \lambda_{0i} = \beta_{00} + \beta_{01} \text{Orch}_i + \nu_{0i} \]
  \[ \lambda_{1i} = \beta_{10} + \beta_{11} \text{Orch}_i, \]
where $e_{ij} \sim N(0, \sigma^2)$ and $e_{0i} \sim N(0, \sigma_0^2)$.

The output below contains REML estimates of our 6 parameters from this simplified version of Model C:

Formula: na ~ orch + large + orch:large + (1 | id)

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>REMLdev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2999</td>
<td>3025</td>
<td>-1494</td>
<td>2992</td>
<td>2987</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>(Intercept)</td>
<td>5.1311</td>
<td>2.2652</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>21.8824</td>
<td>4.6779</td>
</tr>
</tbody>
</table>

Number of obs: 497, groups: id, 37

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>15.9026</td>
<td>0.6187</td>
</tr>
<tr>
<td>orch</td>
<td>1.7100</td>
<td>0.9130</td>
</tr>
<tr>
<td>large</td>
<td>-0.8918</td>
<td>0.8415</td>
</tr>
<tr>
<td>orch:large</td>
<td>-1.4650</td>
<td>1.0880</td>
</tr>
</tbody>
</table>

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 Additional covariates: model comparison and interpretability

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 ??, 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:**
  \[
  Y_{ij} = \lambda_{0i} + \lambda_{1i} \text{LargeEns}_{ij} + \epsilon_{ij}
  \] (8.19)

- **Level Two:**
  \[
  \lambda_{0i} = \beta_{00} + \beta_{01} \text{Orch}_i + \beta_{02} \text{MPQnem}_i + \epsilon_{0i}
  \]
  \[
  \lambda_{1i} = \beta_{10} + \beta_{11} \text{Orch}_i + \beta_{12} \text{MPQnem}_i + \epsilon_{1i},
  \]

or as a composite model:

\[
Y_{ij} = [\beta_{00} + \beta_{01} \text{Orch}_i + \beta_{02} \text{MPQnem}_i + \beta_{10} \text{LargeEns}_{ij} \\
+ \beta_{11} \text{Orch}_i \text{LargeEns}_{ij} + \beta_{12} \text{MPQnem}_i \text{LargeEns}_{ij} \\
+ \epsilon_{0i} + \epsilon_{1i} \text{LargeEns}_{ij} + \epsilon_{ij}]
\]
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.

**Formula:** na ~ orch + mpqnem + large + orch:large + mpqnem:large + (large | id)

<table>
<thead>
<tr>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>REMLdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>3002</td>
<td>3044</td>
<td>-1491</td>
<td>2977</td>
<td>2982</td>
</tr>
</tbody>
</table>

**Random effects:**

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>(Intercept)</td>
<td>3.28569</td>
<td>1.81265</td>
<td></td>
</tr>
<tr>
<td></td>
<td>large</td>
<td>0.55654</td>
<td>0.74601</td>
<td>-0.379</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>21.81142</td>
<td>4.67027</td>
<td></td>
</tr>
</tbody>
</table>

**Number of obs: 497, groups: id, 37**

**Fixed effects:**

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>11.56800</td>
<td>1.22058</td>
</tr>
<tr>
<td>orch</td>
<td>1.00068</td>
<td>0.81713</td>
</tr>
<tr>
<td>mpqnem</td>
<td>0.14823</td>
<td>0.03808</td>
</tr>
<tr>
<td>large</td>
<td>-0.28008</td>
<td>1.83411</td>
</tr>
<tr>
<td>orch:large</td>
<td>-0.94919</td>
<td>1.10620</td>
</tr>
<tr>
<td>mpqnem:large</td>
<td>-0.03018</td>
<td>0.05246</td>
</tr>
</tbody>
</table>

### 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:

- $\hat{\beta}_{00} = 11.57$. The estimated mean performance anxiety for solos and small ensembles ($\text{large}=0$) is 11.57 for keyboard players and vocalists ($\text{orch}=0$) with negative emotionality of 0 at baseline ($\text{mpqnem}=0$). Since the minimum negative emotionality score in this study was 11, this interpretation, while technically correct, is not practically meaningful.

- $\hat{\beta}_{01} = 1.00$. 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.

- $\hat{\beta}_{02} = 0.15$. 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.

- $\hat{\beta}_{10} = -0.28$. Keyboard players and vocalists ($\text{orch}=0$) with baseline negative emotionality levels of 0 ($\text{mpqnem}=0$) have an estimated mean decrease in anxiety level of 0.28 points before large ensemble performances compared to other performance types.
ADDITIONAL COVARIATES: MODEL COMPARISON AND INTERPRETABILITY

- $\hat{b}_{11} = -0.95$. Orchestral instrument players have an estimated mean difference in anxiety levels between large ensembles and other performance types that is 0.95 points lower than the difference for keyboard players and vocalists, controlling for the effects of baseline negative emotionality. For example, orchestral instrumentalists with $\text{mpqnm}=0$ have a larger estimated mean decrease in anxiety level of 1.23 points before large ensemble performances, rather than the mean decrease of 0.28 in those who don’t play orchestral instruments.

- $\hat{b}_{12} = -0.03$. A one point increase in baseline negative emotionality is associated with an estimated .03 mean decrease in the difference in anxiety levels for large ensembles and other performance types, after controlling for instrument.

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.

- $\hat{b}_{01}$. For solos and small ensembles ($\text{LargeEns} = 0$), the following equations describe the fixed effects portion of the composite model for negative affect score for vocalists and keyboardsists ($\text{Orch} = 0$) and orchestral instrumentalists ($\text{Orch} = 1$):

  $\text{Orch} = 0 :$
  
  $$Y_{ij} = \beta_{00} + \beta_{02}\text{MPQnem}_{i}$$

  $\text{Orch} = 1 :$
  
  $$Y_{ij} = (\beta_{00} + \beta_{11}) + \beta_{02}\text{MPQnem}_{i}$$

Regardless of the subjects’ baseline negative emotionality ($\text{MPQnem}$), $\hat{b}_{01}$ represents the 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 $\hat{b}_{01} + \hat{b}_{11}$, holding $\text{MPQnem}$ constant (Show!).

- $\hat{b}_{10}$. Because $\text{LargeEns}$ interacts with both $\text{Orch}$ and $\text{MPQnem}$ in Model C, $\hat{b}_{10}$ only describes the difference between large ensembles and other performance types when both $\text{Orch} = 0$ and $\text{MPQnem} = 0$, thus removing the effects of the interaction terms. If, for instance, $\text{Orch} = 1$ and $\text{MPQnem} = 20$, then the difference between large ensembles and other performance types is given by $\hat{b}_{10} + \hat{b}_{11} + 20\hat{b}_{12}$.

- $\hat{b}_{11}$. As with $\hat{b}_{01}$, we consider equations describing the fixed effects portion of the composite model for negative affect score for vocalists and keyboardsists ($\text{Orch} = 0$) and orchestral instrumentalists ($\text{Orch} = 1$), except here we leave $\text{LargeEns}$ as an unknown rather than restricting the model to solos and small ensembles:

  $\text{Orch} = 0 :$
  
  $$Y_{ij} = \beta_{00} + \beta_{02}\text{MPQnem}_{i} + \beta_{10}\text{LargeEns}_{ij}$$
  $$+ \beta_{12}\text{MPQnem}_{i}\text{LargeEns}_{ij}$$

  $\text{Orch} = 1 :$
  
  $$Y_{ij} = (\beta_{00} + \beta_{11}) + \beta_{02}\text{MPQnem}_{i} + (\beta_{10} + \beta_{11})\text{LargeEns}_{ij}$$
  $$+ \beta_{12}\text{MPQnem}_{i}\text{LargeEns}_{ij}$$

As long as baseline negative emotionality is held constant (at any level, not just 0), then $\hat{b}_{11}$ represents the 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 (drop in deviance) test, as we’ve used in earlier chapters and 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 methods when random effects same
> anova(model.d,model.c)

|                  | model.c: na ~ orch + large + orch:large + (large | id) |
|------------------|-----------------------------------------------------|
| Df               | 8                                                   |
| AIC              | 3007.3                                              |
| BIC              | 3041.0                                              |
| logLik           | -1495.6                                             |
| Chisq            |                                                    |
| Chi Df           |                                                    |
| Pr(>Chisq)       |                                                    |

<table>
<thead>
<tr>
<th></th>
<th>model.d: na ~ orch + mpqnem + large + orch:large +</th>
</tr>
</thead>
<tbody>
<tr>
<td>(large</td>
<td>id)</td>
</tr>
<tr>
<td>Df</td>
<td>10</td>
</tr>
<tr>
<td>AIC</td>
<td>2996.7</td>
</tr>
<tr>
<td>BIC</td>
<td>3038.8</td>
</tr>
<tr>
<td>logLik</td>
<td>-1488.4</td>
</tr>
<tr>
<td>Chisq</td>
<td>14.572</td>
</tr>
<tr>
<td>Chi Df</td>
<td>2</td>
</tr>
<tr>
<td>Pr(&gt;Chisq)</td>
<td>0.0006849 ***</td>
</tr>
</tbody>
</table>

Two models, whether they are nested or not, can be compared using AIC and BIC measures, which were first seen in Chapter 2 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.

> model.d1 <- lmer(na ~ orch + mpqnem + large + orch:large + |
| (large | id), data=music)
| > anova(model.d,model.d1)

<table>
<thead>
<tr>
<th></th>
<th>model.d1: na ~ orch + mpqnem + large + orch:large +</th>
</tr>
</thead>
<tbody>
<tr>
<td>(large</td>
<td>id)</td>
</tr>
<tr>
<td>Df</td>
<td>9</td>
</tr>
<tr>
<td>AIC</td>
<td>2995.0</td>
</tr>
<tr>
<td>BIC</td>
<td>3032.8</td>
</tr>
<tr>
<td>logLik</td>
<td>-1488.5</td>
</tr>
<tr>
<td>Chisq</td>
<td></td>
</tr>
<tr>
<td>Chi Df</td>
<td></td>
</tr>
<tr>
<td>Pr(&gt;Chisq)</td>
<td></td>
</tr>
</tbody>
</table>

|                  | model.d: na ~ orch + mpqnem + large + orch:large + |
| (large | id)          |
| Df               | 10                                                  |
| AIC              | 2996.7                                              |
| BIC              | 3038.8                                              |
| logLik           | -1488.4                                             |
| Chisq            | 0.2598                                              |
| Chi Df           | 1                                                   |
| Pr(>Chisq)       | 0.6102                                              |

8.9 Center 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 MPQ NEM scores of 0 at baseline (as in $\hat{\beta}_{10}$), 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 = 12; GPA = 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

\[
\text{centeredbaselineNEM} = \text{cmpqnem} = \text{mpqnem} - \text{mean(mpqnem)} = \text{mpqnem} - 31.63
\]

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

\[
\text{Formula: na} \sim \text{orch} + \text{cmpqnem} + \text{large} + \text{orch:large} + \text{cmpqnem:large} + \text{large} | \text{id}
\]

\[
\text{AIC BIC logLik deviance REMLdev}
\]

3002 3044 -1491 2977 2982

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>(Intercept)</td>
<td>3.28569</td>
<td>1.81265</td>
<td></td>
</tr>
<tr>
<td></td>
<td>large</td>
<td>0.55654</td>
<td>0.74601</td>
<td>-0.379</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>21.81142</td>
<td>4.67027</td>
<td></td>
</tr>
</tbody>
</table>

Number of obs: 497, groups: id, 37

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>16.25681</td>
<td>0.54757</td>
</tr>
<tr>
<td>orch</td>
<td>1.00068</td>
<td>0.81713</td>
</tr>
<tr>
<td>cmpqnem</td>
<td>0.14823</td>
<td>0.03808</td>
</tr>
<tr>
<td>large</td>
<td>-1.23487</td>
<td>0.84319</td>
</tr>
<tr>
<td>orch:large</td>
<td>-0.94919</td>
<td>1.10620</td>
</tr>
<tr>
<td>cmpqnem:large</td>
<td>-0.03018</td>
<td>0.05246</td>
</tr>
</tbody>
</table>

As you compare Model D to Model E, you should notice that only two things change — \( \hat{\beta}_{00} \) and \( \hat{\beta}_{10} \). 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. \( \hat{\beta}_{00} \) and \( \hat{\beta}_{10} \) 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:

- \( \hat{\beta}_{00} = 16.26 \). 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).
- \( \hat{\beta}_{10} = -1.23 \). 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 potential 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 chapters, 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:
  \[ Y_{ij} = \lambda_{0i} + \lambda_{1i} \text{previous}_{ij} + \lambda_{2i} \text{students}_{ij} + \lambda_{3i} \text{juried}_{ij} + \lambda_{4i} \text{public}_{ij} + \lambda_{5i} \text{solo}_{ij} + \varepsilon_{ij} \quad (8.20) \]

- Level Two:
  \[
  \begin{align*}
  \lambda_{0i} &= \beta_{00} + \beta_{01} \text{mpqem}_{i} + \beta_{02} \text{mpqa}_{i} + \beta_{03} \text{orch}_{i} + \beta_{04} \text{mpqen}_{i} + \varepsilon_{0i} \\
  \lambda_{1i} &= \beta_{10} + \varepsilon_{1i} \\
  \lambda_{2i} &= \beta_{20} + \varepsilon_{2i} \\
  \lambda_{3i} &= \beta_{30} + \varepsilon_{3i} \\
  \lambda_{4i} &= \beta_{40} + \varepsilon_{4i} \\
  \lambda_{5i} &= \beta_{50} + \beta_{51} \text{mpqen}_{i} + \varepsilon_{5i},
  \end{align*}
\]

where previous is the number of previous diary entries filled out by that individual \((\text{diary} - 1)\); 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 is the performance was either a small or large ensemble. In addition, we assume the following variance-covariance structure at Level Two:

\[
\begin{bmatrix}
\varepsilon_{0i} \\
\varepsilon_{1i} \\
\varepsilon_{2i} \\
\varepsilon_{3i} \\
\varepsilon_{4i} \\
\varepsilon_{5i}
\end{bmatrix}
\sim N
\begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}
,
\begin{bmatrix}
\sigma_{0}^{2} & \sigma_{01} & \sigma_{02} & \sigma_{03} & \sigma_{04} & \sigma_{05} \\
\sigma_{01} & \sigma_{1}^{2} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\
\sigma_{02} & \sigma_{12} & \sigma_{2}^{2} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\
\sigma_{03} & \sigma_{13} & \sigma_{23} & \sigma_{3}^{2} & \sigma_{34} & \sigma_{35} \\
\sigma_{04} & \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{4}^{2} & \sigma_{45} \\
\sigma_{05} & \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_{5}^{2}
\end{bmatrix}
\]

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 chapters 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 esti-

mates 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 ad-

dition, 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:

- $\hat{\beta}_{01} = 0.11$. 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.

- $\hat{\beta}_{51} = 0.08$. When musicians play solos, a one point increase in baseline level of negative emo-
tionality 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.

Formula: na ~ previous + students + juried + public + solo + mpqpem +
mpqab + orch + mpqem + mpqem:solo + (previous + students +
juried + public + solo | id)

AIC BIC logLik deviance REMLdev
2948 3087 -1441 2862 2882

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>(Intercept)</td>
<td>14.46761</td>
<td>3.80363</td>
<td></td>
</tr>
<tr>
<td></td>
<td>previous</td>
<td>0.07078</td>
<td>0.26604</td>
<td>-0.649</td>
</tr>
<tr>
<td></td>
<td>students</td>
<td>8.21114</td>
<td>2.86551</td>
<td>-0.627-0.001</td>
</tr>
<tr>
<td></td>
<td>juried</td>
<td>18.30422</td>
<td>4.27834</td>
<td>-0.643-0.117 0.835</td>
</tr>
<tr>
<td></td>
<td>public</td>
<td>12.79116</td>
<td>3.57647</td>
<td>-0.827 0.333 0.659 0.575</td>
</tr>
<tr>
<td></td>
<td>solo</td>
<td>0.76381</td>
<td>0.87396</td>
<td>-0.675 0.475 0.490 0.207 0.899</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>15.28478</td>
<td>3.90958</td>
<td></td>
</tr>
</tbody>
</table>

Number of obs: 497, groups: id, 37

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>8.36988</td>
<td>1.91373</td>
</tr>
<tr>
<td>previous</td>
<td>-0.14304</td>
<td>0.06248</td>
</tr>
<tr>
<td>students</td>
<td>3.61114</td>
<td>0.76788</td>
</tr>
<tr>
<td>juried</td>
<td>4.07337</td>
<td>1.03110</td>
</tr>
<tr>
<td>public</td>
<td>3.06419</td>
<td>0.89240</td>
</tr>
<tr>
<td>solo</td>
<td>0.51372</td>
<td>1.39644</td>
</tr>
<tr>
<td>mpqpem</td>
<td>-0.08310</td>
<td>0.02407</td>
</tr>
</tbody>
</table>
8.11 Modeling the multilevel structure: is it really necessary?

Before going too much further, we should really consider if this multilevel structure has gained us anything over ordinary 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 book entitled *Multilevel Analysis for Applied Research: It's Just Regression* (2007), 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." 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 OLS 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 Hypothetical data from 4 subjects relating number of previous performances to negative affect. The solid black line depicts the overall relationship between previous performances and negative affect as determined by a multilevel model, while the dashed black line depicts the overall relationship as determined by an OLS regression model.

Figure 8.15 is based on a simulated 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 solid black line)
that's consistent with the individual subjects–slope around -0.5 with an intercept that appears to average the 4 subjects. Fitting an OLS regression model, however, produces an overall relationship (the dashed 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 OLS regression analysis has gotten the estimated slope of the overall relationship backwards, producing a continuous data version of Simpson’s Paradox.

Figure 8.16  Density plots of parameter estimates for the four fixed effects of Model C under both a multilevel model and OLS 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) $\beta_{00}$, (b) $\beta_{01}$, (c) $\beta_{10}$, and (d) $\beta_{11}$ were obtained using both a multilevel and an OLS regression model. Each plot then shows a density plot for the 1000 estimates of the corresponding $\beta$-term using multilevel modeling vs. a similar density plot for the 1000 estimates using OLS regression.

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 an OLS regression model, and the estimated fixed effects ($\hat{\beta}_{00}$, $\hat{\beta}_{01}$, $\hat{\beta}_{10}$, and $\hat{\beta}_{11}$) and their standard errors were saved. Figure 8.16 shows density plots comparing the 1000 estimated fixed effects for each $\beta$-term from the two modeling approaches; in general, estimates from multilevel modeling and OLS regression tend to agree pretty well, without noticeable bias. Based on coefficient estimates alone, there appears to be no reason to favor multilevel modeling over OLS regression 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 $\beta$-term from the two modeling approaches; in general, standard errors are markedly larger with multilevel modeling than OLS regression. This is not unusual, since OLS regression assumes all 497 observations are independent while multilevel modeling acknowledges that, with correlated data within subject, there are fewer than 497 independent pieces of data. Therefore, OLS 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 OLS regression are then not reflective of the true structure.
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 OLS 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) $\text{SE}(\beta_{00})$, (b) $\text{SE}(\beta_{01})$, (c) $\text{SE}(\beta_{10})$, and (d) $\text{SE}(\beta_{11})$ were obtained using both a multilevel and an OLS regression model. Each plot then shows a density plot for the 1000 estimates of the corresponding $\text{SE}(\beta)$-term using multilevel modeling vs. a similar density plot for the 1000 estimates using OLS regression.

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). For summarizing variables at Level Two, a data frame with one observation per subject was created (see R code below); fitted intercepts and slopes by subject could then be added to this data frame.

```r
# Create Level2 data set by picking off one observation per subject,
# which would be easier if every subject had a diary entry labeled '1'
music.lev2 = keydata[diary==2 | (id==42 & diary==3),c(1,7,8,9,10,11)]
dim(music.lev2) # should be 37 rows and 6 columns (one per L2 variable)

# Add average across all performances for each subject for EDA plots
meanbysubj = by(music$na,music$id,mean,na.rm=T)
music.lev2 = data.frame(music.lev2,meanbysubj=as.numeric(meanbysubj))

# Find intercepts by subject
# "by" line runs "coefficients" function for each subject
# "coefficients" puts int and slope from regression of na on large into list
# "[[1]]" picks off the first value from the list (the intercept)
int <- by(music, id, function(data)
          coefficients(lm(na ~ large, data = data))[[1]])
summary(int)  # summary statistics for 37 intercepts
music.lev2$int = int[1:37]

# Find fitted rate of change
# same as above, except second coefficient (slope) picked off
rate <- by(music, id, function(data)
            coefficients(lm(na ~ large, data = data))[[2]])
summary(rate)
music.lev2$rate = rate[1:37]

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 $b_{01}$, $b_{10}$, and $b_{11}$, along with the intercept $b_{00}$ 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 ($\epsilon_0$) and one corresponding to the large ensemble effect ($\epsilon_1$); 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)

Finally, the R code below produced the simulated data for the two examples in Section 8.11.

# Simulation 1 - OLS coefficient is wrong direction
subject = c(rep(1,10),rep(2,10),rep(3,10),rep(4,10))
lambda0 = c(rep(10,10),rep(20,10),rep(30,10),rep(40,10))
lambda1 = rep(-0.5,40)
previj = c(1:10,4:13,7:16,10:19)
eij = rnorm(40,0,1)
yij = lambda0 + lambda1*previj + eij
simdata = data.frame(subject=subject,lambda0=lambda0,lambda1=lambda1,previj=previj,eij=eij,yij=yij)

# Simulation 2 - use estimated parameters from Model C
# - SEs usually higher with OLS
library(MASS)
orch.sim = music$orch
large.sim = music$large
subj.sim = music$id
reps.sim = as.numeric(table(music$id))
b00 = 15.9297
b01 = 1.6926
b10 = -0.9106
b11 = -1.4239
sig01 = (-0.635)*2.3781*0.67236
Sigma = matrix(c(2.3781^2,sig01,sig01,0.67236^2),2,2)
e0e1 = mvrnorm(37,c(0,0),Sigma)
e0i = rep(e0e1[,1],times=reps.sim)
e1i = rep(e0e1[,2],times=reps.sim)
lam0i = b00 + b01*orch.sim + e0i
lam1i = b10 + b11*orch.sim + e1i
Eij = rnorm(497,0,4.66984)
Yij = lam0i + lam1i*large.sim + Eij
data.sim = data.frame(subj.sim=subj.sim,orch.sim=orch.sim,large.sim=large.sim,
e0i=e0i,e1i=e1i,lam0i=lam0i,lam1i=lam1i,Eij=Eij,Yij=Yij)

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 house 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. In Equations 8.2 and 8.3, explain why the subscript of the term $Orch_i$ has a period.

6. Explain the difference between $\lambda_{0i}$ in Equation 8.2 and $\hat{\lambda}_{0i}$ in Equation 8.4.

7. 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?

8. 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?

9. 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?

10. 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,
EXERCISES

- 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.

11. 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).

12. The interpretation of $\hat{\beta}_{01}$ 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 $\hat{\beta}_{01} + \hat{\beta}_{11}$, holding MPQem.” Show that this claim is true.

13. Explain how the interpretations of the following parameter estimates change (or don’t change) as we change our model:
   - $\hat{\beta}_{00}$ from Model A to B to C to D to E
   - $\hat{\beta}_{10}$ from Model B to C to D to E
   - $\hat{\beta}_{01}$ from Model C to D to E
   - $\hat{\beta}_{11}$ from Model C to D to E
   - $\hat{\sigma}_{0}$ from Model A to B to C to D to E
   - $\hat{\sigma}_{1}$ from Model B to C to D to E

14. When moving from Model B to Model C in Section 8.7.3, $\hat{\sigma}_{20}^2$ increases slightly. Why might this have occurred?

15. Interpret other estimated parameters from Model F beyond those interpreted in Section 8.10: $\hat{\beta}_{00}, \hat{\beta}_{02}, \hat{\beta}_{03}, \hat{\beta}_{10}, \hat{\beta}_{20}, \hat{\beta}_{30}, \sigma^2, \sigma^2_0$, and $\sigma^2_1$.

16. Explain Figure 8.15 in your own words. Why would OLS regression produce a misleading analysis in this case, but multilevel models would not?

17. Summarize Figures 8.16 and 8.17 in your own words.

8.13.2 Guided Exercise

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

   (a) 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.

   (b) 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.
(c) Report estimated fixed effects and variance components from Model B, using proper notation from this chapter. Interpret your MLE estimates for \( \hat{\beta}_{00} \) (the intercept), \( \hat{\beta}_{10} \) (the instructor indicator), and \( \hat{\sigma}_0 \) (the Level Two standard deviation for the intercept). Also report and interpret an appropriate pseudo-Rsquare value.

(d) 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?

(e) Report and interpret the following parameter estimates from Model C: \( \hat{\beta}_{00} \), \( \hat{\beta}_{01} \), \( \hat{\beta}_{20} \), \( \hat{\beta}_{11} \), \( \hat{\sigma}_0 \), \( \hat{\sigma}_1 \), and \( \rho_{01} \). Interpretations for variance components should be done in terms of standard deviations and correlation coefficients.

(f) 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.

(g) Also report and interpret the following parameter estimates from Model D: \( \hat{\beta}_{02} \) and \( \hat{\beta}_{12} \).

(h) Use a 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-Rsquare value(s) and with AIC and BIC statistics.

8.13.3 Open-ended Exercises

1. **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 (Lampert, Friedrich, Sturtz) collected 2012 data on Rockies pitchers from FanGraphs to evaluate Project 5183. 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 taking 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 results the Rockies hoped to observe was that their pitchers pitch better (allow fewer runs to the opponent) with a pitch limit.

The data set 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.)
Chapter 9

Two Level Longitudinal Data

9.1 Learning objectives

After finishing this chapter, you should be able to:

- Recognize longitudinal data as a special case of multilevel data, with time at Level One.
- Consider patterns of missingness and implications of that missing data on multilevel analyses.
- Apply exploratory data analysis techniques specific to longitudinal data.
- Build and understand a taxonomy of models for longitudinal data.
- Interpret model parameters in multilevel models with time at Level One.
- Compare models, both nested and not, with appropriate statistical tests and summary statistics.
- Consider different ways of modeling the variance-covariance structure in longitudinal data.

9.2 Case study: Charter schools

Charter schools were first introduced in the state of Minnesota in 1991 (www.uscharterschools.org). Since then, charter schools have begun appearing all over the United States. While publicly funded, a unique feature of charter schools is their independence from many of the regulations that are present in the public school systems of their respective city or state. Thus, charters will often extend the school days or year and tend to offer non-traditional techniques and styles of instruction and learning.

One example of this unique schedule structure is the KIPP (Knowledge is Power Program) Stand Academy in Minneapolis, MN. KIPP stresses longer days and better partnerships with parents, and they claim that 80% of their students go to college from a population where 87% qualify for free and reduced lunch and 95% are African-American or Latino (www.kipp.org). However, the larger question is whether or not charter schools are out-performing non-charter public schools in general. Because of the relative youthfulness of charter schools, data has just begun to be collected to evaluate the performance of charter versus non-charter schools and some of the factors that influence a school’s performance. Along these lines, we will examine data collected by the Minnesota Department of Education for all Minnesota schools during the years 2008-2010.

Comparisons of student performance in charter schools versus public schools have produced conflicting results, potentially as a result of the strong differences in the structure and population of the student bodies that represent the two types of schools. A study by the Policy and Program Studies Service of five states found that charter schools are less likely to meet state performance standards than conventional public schools (US Department of Education, 2004). However, Witte et al. (2007) performed a statistical analysis comparing Wisconsin charter and non-charter schools and found that average achievement test scores were significantly higher in charter schools compared to non-charter schools, after controlling for demographic variables such as the percentage of white students. In addition, a study of California students who took the Stanford 9 exam from 1998 through 2002 found that charter schools, on average, were performing at the same level as conventional public schools (Buddin and Zimmer, 2005). Although school performance is difficult to quantify with a
single measure, for illustration purposes in this chapter, we will focus on that aspect of school performance measured by the math portion of the Minnesota Comprehensive Assessment (MCA-II) data for 6th grade students enrolled in 618 different Minnesota schools during the years 2008, 2009, and 2010. Similar comparisons could obviously be conducted for other grade levels or modes of assessment.

As described in Green et al. (2003), it is very challenging to compare charter and public non-charter schools, as charter schools are often designed to target or attract specific populations of students. Without accounting for differences in student populations, comparisons lose meaning. With the assistance of multiple school-specific predictors, we will attempt to model sixth grade math MCA-II scores of Minnesota schools, focusing on the differences between charter and public non-charter school performances. In the process, we hope to answer the following research questions:

- Which factors most influence a school’s performance in MCA testing?
- How do the average math MCA-II scores for 6th graders enrolled in charter schools differ from scores for students who attend non-charter public schools? Do these differences persist after accounting for differences in student populations?
- Are there differences in yearly improvement between charter and non-charter public schools?

9.3 Initial Exploratory Analyses

9.3.1 Data organization

Key variables which we will examine to address the research questions above are:
- schoolid = includes district type, district number, and school number
- schoolName = name of school
- urban = is the school in an urban (1) or rural (0) location?
- charter = is the school a charter school (1) or a non-charter public school (0)?
- schPctnonw = proportion of non-white students in a school (based on 2010 figures)
- schPctsped = proportion of special education students in a school (based on 2010 figures)
- schPctfree = proportion of students who receive free or reduced lunches in a school (based on 2010 figures). This serves as a measure of poverty among school families.
- MathAvgScore.0 = average MCA-II math score for all sixth grade students in a school in 2008
- MathAvgScore.1 = average MCA-II math score for all sixth grade students in a school in 2009
- MathAvgScore.2 = average MCA-II math score for all sixth grade students in a school in 2010

This data is stored in WIDE format, with one row per school, as illustrated in Table 9.1.

For most statistical analyses, it will be advantageous to convert WIDE format to LONG format, with one row per year per school. To make this conversion, we will have to create a time variable, which under the LONG format is very flexible—each school can have a different number of and differently-spaced time points, and they can even have predictors which vary over time. Details for making this conversion in R can be found in Section 9.8, and the form of the LONG data in this study is exhibited in the next section.

9.3.2 Missing data

In this case, before we convert our data to LONG form, we should first address problems with missing data. Missing data is a common phenomenon in longitudinal studies. For instance, it could arise if a new school was started during the observation period, a school was shut down during the observation period, or no results were reported in a given year. Dealing with missing data in
INITIAL EXPLORATORY ANALYSES

Table 9.1: The first six observations in the wide data set for the Charter Schools case study.

Table 9.2: A frequency table of missing data patterns. The number of schools with a particular missing data pattern are listed in the left column; the remaining columns of 0’s and 1’s describe the missing data pattern, with 0 indicating a missing value. Some covariates that are present for every school are not listed. The bottom row gives the number of schools with missing values for specific variables; the last entry indicates that 121 total observations were missing.

A statistical analysis is not trivial, but fortunately many multilevel packages (including the lme4 package in R) are adept at handling missing data.

First, we must understand the extent and nature of missing data in our study. Table 9.2, where 1 indicates presence of a variable and 0 indicates a missing value for a particular variable, is a helpful starting point. Among our 618 schools, 540 had complete data (all covariates and math scores for all three years), 25 were missing a math score for 2008, 35 were missing math scores in both 2008 and 2009, etc.

Statisticians have devised different strategies for handling missing data; a few common approaches are described briefly here:

- Include only schools with complete data. This is the cleanest approach analytically; however, ignoring data from 12.6% of the study’s schools (since 78 of the 618 schools had incomplete data) means that a large amount of potentially useful data is being thrown away. In addition, this approach creates potential issues with informative missingness. Informative missingness occurs when a school’s lack of scores is not a random phenomenon but provides information about the effectiveness of the school type (e.g., a school closes because of low test scores).
TWO LEVEL LONGITUDINAL DATA

<table>
<thead>
<tr>
<th>schoolName</th>
<th>charter</th>
<th>schPctsped</th>
<th>schPctfree</th>
<th>year08</th>
<th>MathAvgScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIPPLESIDE ELEMENTARY</td>
<td>0</td>
<td>0.1176471</td>
<td>0.3627451</td>
<td>0</td>
<td>652.8</td>
</tr>
<tr>
<td>RIPPLESIDE ELEMENTARY</td>
<td>0</td>
<td>0.1176471</td>
<td>0.3627451</td>
<td>1</td>
<td>656.6</td>
</tr>
<tr>
<td>RIPPLESIDE ELEMENTARY</td>
<td>0</td>
<td>0.1176471</td>
<td>0.3627451</td>
<td>2</td>
<td>652.6</td>
</tr>
<tr>
<td>WRENSHALL ELEMENTARY</td>
<td>0</td>
<td>0.1515152</td>
<td>0.4242424</td>
<td>0</td>
<td>646.9</td>
</tr>
<tr>
<td>WRENSHALL ELEMENTARY</td>
<td>0</td>
<td>0.1515152</td>
<td>0.4242424</td>
<td>1</td>
<td>645.3</td>
</tr>
<tr>
<td>WRENSHALL ELEMENTARY</td>
<td>0</td>
<td>0.1515152</td>
<td>0.4242424</td>
<td>2</td>
<td>651.9</td>
</tr>
</tbody>
</table>

Table 9.3 The first six observations in the long data set for the Charter Schools case study; these lines correspond to the first two observations from the wide data set illustrated in Table 9.1.

- Last observation carried forward. Each school’s last math score is analyzed as a univariate response, whether the last measurement was taken in 2008, 2009, or 2010. With this approach, data from all schools can be used, and analyses can be conducted with traditional methods assuming independent responses. This approach is sometimes used in clinical trials because it tends to be conservative, setting a higher bar for showing that a new therapy is significantly better than a traditional therapy. Of course, we must assume that a school’s 2008 score is representative of their 2010 score. In addition, information about trajectories over time is thrown away.

- Imputation of missing observations. Many methods have been developed for sensibly “filling in” missing observations, using imputation models which base imputed data on subjects with similar covariate profiles and on typical observed time trends. Once an imputed data set is created (or several imputed data sets), analyses can proceed with complete data methods that are easier to apply. Risks with the imputation approach include misrepresenting missing observations and overstating precision in final results.

- Apply multilevel methods, which use available data to estimate patterns over time by school and then combine those school estimates in a way that recognizes that time trends for schools with complete data are more precise than time trends for schools with fewer measurements. Laird (1988) demonstrates that multilevel models are valid under the fairly unrestrictive condition that the probability of missingness cannot depend on any unobserved predictors or the response. This is the approach we will follow in the remainder of the text.

Now, we are ready to create our LONG data set. Fortunately, many packages (including R) have built-in functions for easing this conversion. The resulting LONG data set is shown in Table 9.3, where year08 measures the number of years since 2008:

9.3.3 Exploratory analyses for general multilevel models

Notice the longitudinal structure of our data—we have up to three measurements of test scores at different time points for each of our 618 schools. With this structure, we can address questions at two levels:

- Within school—changes over time
- Between schools—effects of school-specific covariates (charter or non-charter, urban or rural, percent free and reduced lunch, percent special education, and percent non-white) on 2008 math scores and rate of change between 2008 and 2010.

As with any statistical analysis, it is vitally important to begin with graphical and numerical summaries of important variables and relationships between variables. We’ll begin with initial exploratory analyses that we introduced in the previous chapter, noting that we have no Level One covariates other than time at this point (potential covariates at this level may have included measures of the number of students tested or funds available per student). We will, however, consider the Level Two variables of charter or non-charter, urban or rural, percent free and reduced lunch, percent special education, and percent non-white. Although covariates such as percent free and re-
duced lunch may vary slightly from year to year within a school, the larger and more important differences tend to occur between schools, so we used mean percent free and reduced lunch over 3 years for a school as a Level Two variable.

As in Chapter 8, we can conduct initial investigations of relationships between Level Two covariates and test scores in two ways. First, we can use all 1733 observations to investigate relationships of Level Two covariates with test scores; although these plots will contain dependent points, since each school is represented by up to three years of test score data, general patterns exhibited in these plots tend to be real. Second, we can calculate mean scores across all years for each of the 618 schools; while we lose some information with this approach, we can more easily consider each plotted point to be independent. Typically, both types of exploratory plots illustrate similar relationships, and in this case, both approaches are so similar that we will only show plots using the second approach, with one observation per school.

Figure 9.1 shows the distribution of MCA math test scores as somewhat left skewed. MCA test scores for sixth graders are scaled to fall between 600 and 700, where scores above 650 for individual students indicate “meeting standards”. Thus, schools with averages below 650 will often have increased incentive to improve their scores the following year. When we refer to the “math score” for a particular school in a particular year, we will assume that score represents the average for all sixth graders at that school. In Figure 9.2, we see that test scores are generally higher for both schools in rural areas and for public non-charter schools. Note that in this data set there are 237 schools in rural areas and 381 schools in urban areas, as well as 545 public non-charter schools and 73 charter schools. In addition, we can see in Figure ?? that schools tend to have lower math scores if they have higher percentages of students with free and reduced lunch, with special education needs, or who are non-white.

Figure 9.1 *Histogram of mean sixth grade MCA math test scores over the years 2008-2010 for 618 Minnesota schools.*

### 9.3.4 Exploratory analyses for longitudinal data

In addition to the initial exploratory analyses above, longitudinal data—multilevel data with time at Level One—calls for further plots and summaries that describe time trends within and across individuals. For example, we can examine trends over time within individual schools. Figure 9.4 provides a *lattice plot* illustrating trends over time for the first 24 schools in the data set. We note differences among schools in starting point (test scores in 2008), slope (change in test scores over
Figure 9.2 Boxplots of categorical Level Two covariates vs. average MCA math scores. Plot (a) shows charter vs. public non-charter schools, while plot (b) shows urban vs. rural schools.

Figure 9.3 Scatterplots of average MCA math scores by (a) percent free and reduced lunch, (b) percent special education, and (c) percent non-white in a school.

the three year period), and form of the relationship. These differences among schools are nicely illustrated in so-called spaghetti plots such as Figure 9.5, which overlays the individual schools’ time trends (for the math test scores) from Figure 9.4 on a single set of axes. In order to illustrate the overall time trend without making global assumptions about the form of the relationship, we overlaid in bold a nonparametric fitted curve through a loess smoother. LOESS comes from “locally weighted scatterplot smoother”, in which a low-degree polynomial is fit to each data point using weighted regression techniques, where nearby points receive greater weight according to a weight function. LOESS is a computationally intensive method which performs especially well with larger sets of data, although ideally there would be a greater diversity of x-values than the three time points we have. In this case, the loess smoother follows very closely to a linear trend, indicating
that assuming a linear increase in test scores over the three year period is probably a reasonable simplifying assumption. To further examine the hypothesis that linearity would provide a reasonable approximation to the form of the individual time trends in most cases, Figure 9.6 shows a lattice plot containing linear fits through ordinary least squares rather than connected time points as in Figure 9.4.

![Lattice plot by school of math scores over time for the first 24 schools in the data set.](image)

**Figure 9.4** Lattice plot by school of math scores over time for the first 24 schools in the data set.

![Spaghetti plot of math scores over time by school, for all the charter schools and a random sample of public non-charter schools, with overall fit using loess (bold).](image)

**Figure 9.5** Spaghetti plot of math scores over time by school, for all the charter schools and a random sample of public non-charter schools, with overall fit using loess (bold).

Just as we explored the relationship between our response (average math scores) and important covariates in Section 9.3.3, we can now examine the relationships between time trends by school and important covariates. For instance, Figure 9.7 shows that charter schools had math scores that were lower on average than public non-charter schools and more variable. This type of plot is sometimes called a **trellis graph**, since it displays a grid of smaller charts with consistent scales, where each
Figure 9.6 Lattice plot by school of math scores over time with linear fit for the first 24 schools in the data set.

Figure 9.7 Spaghetti plot showing time trends for each school by school type, for a random sample of public non-charter schools (left) and all charter schools (right), with overall fits using loess (bold).

smaller chart represents a condition—an item in a category. Trends over time by school type are denoted by bold loess curves. Public non-charter schools have higher scores across all years; both school types show little growth between 2008 and 2009, but greater growth between 2009 and 2010, especially charter schools. Exploratory analyses like this can be repeated for other covariates, such as percent free and reduced lunch in Figure 9.8. The trellis plot automatically divides schools into four groups based on quartiles of their percent free and reduced lunch, and we see that schools with lower percentages of free and reduced lunch students tend to have higher math scores and less variability. Across all levels of free and reduced lunch, we see greater gains between 2009 and 2010 than between 2008 and 2009.
Figure 9.8 Spaghetti plot showing time trends for each school by quartiles of percent free and reduced lunch, with loess fits.

### 9.4 Preliminary two-stage modeling

#### 9.4.1 Linear trends within schools

Even though we know that every school’s math test scores were not strictly linearly increasing or decreasing over the observation period, a linear model for individual time trends is often a simple but reasonable way to model data. One advantage of using a linear model within school is that each school’s data points can be summarized with two summary statistics — an intercept and a slope (obviously, this is an even bigger advantage when there are more observations over time per school). For instance, we see in Figure 9.6 that sixth graders from the school depicted in the top right slot slowly increased math scores over the three year observation period, while students from the school depicted in the fourth column of the top row generally experienced decreasing math scores over the same period. As a whole, the linear model fits individual trends pretty well, and many schools appear to have slowly increasing math scores over time, as researchers in this study may have hypothesized.

Another advantage of assuming a linear trend at Level One (within schools) is that we can examine summary statistics across schools. Both the intercept and slope are meaningful for each school: the intercept conveys the school’s math score in 2008, while the slope conveys the school’s average yearly increase or decrease in math scores over the three year period. Figure 9.9 shows that point estimates and uncertainty surrounding individual estimates of intercepts and slopes vary considerably. In addition, we can generate summary statistics and histograms for the 618 intercepts and slopes produced by fitting linear regression models at Level One, in addition to R-square values which describe the strength of fit of the linear model for each school (Figure 9.10). For our 618 schools, the mean math score for 2008 was 651.4 (SD=7.28), and the mean yearly rate of change in math scores over the three year period was 1.30 (SD=2.51). We can further examine the relationship between schools’ intercepts and slopes. Figure 9.11 shows a general decreasing trend, suggesting that schools with lower 2008 test scores tend to have greater growth in scores between 2008 and 2010 (potentially because those schools have more room for improvement); this trend is supported with a correlation coefficient of -0.32 between fitted intercepts and slopes. Note that, with only 3 or fewer observations for each school, extreme or intractable values for the slope and R-square are possible. For example, slopes cannot be estimated for those schools with just a single test score,
R-square values cannot be calculated for those schools with no variability in test scores between 2008 and 2010, and R-square values must be 1 for those schools with only two test scores.

![Graphs showing point estimates and 95% confidence intervals for intercepts and slopes by school.](image)

**Figure 9.9** Point estimates and 95% confidence intervals for (a) intercepts and (b) slopes by school, for the first 24 schools in the data set.

![Histograms for intercepts, slopes, and R-square values.](image)

**Figure 9.10** Histograms for (a) intercepts, (b) slopes, and (c) R-square values from fitted regression lines by school.

### 9.4.2 Effects of level two covariates on linear time trends

Summarizing trends over time within schools is typically only a start, however. Most of the primary research questions from this study involve comparisons among schools, such as: (a) are there significant differences between charter schools and public non-charter schools, and (b) do any differences between charter schools and public schools change with percent free and reduced lunch, percent special education, or location? These are Level Two questions, and we can begin to explore
these questions by graphically examining the effects of school-level variables on schools’ linear time trends. By school-level variables, we are referring to those covariates that differ by school but are not dependent on time. For example school type (charter or public non-charter), urban or rural location, percent non-white, percent special education, and percent free and reduced lunch are all variables which differ by school but which don’t change over time, at least as they were assessed in this study. Variables which would be time-dependent include quantities such as per pupil funding and reading scores.

Figure 9.12 shows differences in the average time trends by school type, using estimated intercepts and slopes to support observations from the spaghetti plots in Figure 9.7. Based on intercepts, charter schools have lower math scores, on average, in 2008 than public non-charter schools. Based on slopes, however, charter schools tend to improve their math scores at a slightly faster rate than public schools, especially at the seventy-fifth percentile and above. By the end of the three year observation period we would nevertheless expect charter schools to have lower average math scores than public schools. For another exploratory perspective on school type comparisons, we can examine differences between school types with respect to math scores in 2008 and math scores in 2010. As expected, boxplots by school type (Figure 9.13) show clearly lower math scores for charter schools in 2008, but differences are slightly less dramatic in 2010.

Any initial exploratory analyses should also investigate effects of potential confounding variables such as school demographics and location. If we discover, for instance, that those schools with higher levels of poverty (measured by the percentage of students receiving free and reduced lunch) display lower test scores in 2008 but greater improvements between 2008 and 2010, then we might be able to use percentage of free and reduced lunch in statistical modeling of intercepts and slopes, leading to more precise estimates of the charter school effects on these two outcomes. In addition, we should also look for any interaction with school type—any evidence that the difference between charter and non-charter schools changes based on the level of a confounding variable. For example, do charter schools perform better relative to non-charter schools when there is a large percentage of non-white students at a school?

With a continuous confounding variable such as percentage of free and reduced lunch, we will treat this variable as continuous to produce the most powerful exploratory analyses. We can begin by examining boxplots of free and reduced lunch percentage against school type (Figure 9.14). We
observe that charter schools tend to have greater percentages of free and reduced lunch students as well as greater school-to-school variability. Next, we can use scatterplots to graphically illustrate the relationships between free and reduced lunch percentages and significant outcomes such as intercept and slope (also Figure 9.14). In this study, it appears that schools with higher levels of free and reduced lunch (i.e., greater poverty) tend to have lower math scores in 2008, but there is little evidence of a relationship between levels of free and reduced lunch and improvements in test scores between 2008 and 2010. These observations are supported with correlation coefficients between percent free and reduced lunch and intercepts ($r=-0.61$) and slopes ($r=-0.06$).

A less powerful but occasionally informative way to look at the effect of a continuous confounder on an outcome variables is by creating a categorical variable out of the confounder. For instance, we could classify any school with a percentage of free and reduced lunch students above the median as having a high percentage of free and reduced lunch students and all other schools as
having a low percentage of free and reduced lunch students. Then we could examine a possible interaction between percent free and reduced lunch and school type through a series of four boxplots (Figure 9.15). In fact, these boxplots suggest that the gap between charter and public non-charter schools in 2008 was greater in schools with a high percentage of free and reduced lunch students, while the difference in rate of change in test scores between charter and public non-charter schools appeared similar for high and low levels of free and reduced lunch. We will investigate these trends more thoroughly with statistical modeling.

![Boxplot of percent free and reduced lunch by school type (charter vs. public non-charter), along with scatterplots of (b) intercepts and (c) slopes from fitted regression lines by school vs. percent free and reduced lunch.](Figure 9.14)

![Boxplots of (a) intercepts and (b) slopes from fitted regression lines by school vs. school type (charter vs. public non-charter), separated by high and low levels of percent free and reduced lunch.](Figure 9.15)

The effect of other confounding variables (e.g., percent non-white, percent special education, urban or rural location) can be investigated in a similar fashion to free and reduced lunch percentage,
both in terms of main effect (variability in outcomes such as slope and intercept which can be explained by the confounding variable) and interaction with school type (ability of the confounding variable to explain differences between charter and public non-charter schools). We leave these explorations as an exercise.

9.4.3 Error structure within schools

Finally, with longitudinal data it is important to investigate the error variance-covariance structure of data collected within a school (the Level Two observational unit). In multilevel data, as in the examples we introduced in Chapter ??, we suspect observations within group (like a school) to be correlated, and we strive to model that correlation. When the data within group is collected over time, we often see distinct patterns in the residuals that can be modeled—correlations which decrease systematically as the time interval increases, variances that change over time, correlation structure that depends on a covariate, etc. A first step in modeling the error variance-covariance structure is the production of an exploratory plot such as Figure 9.16. To generate this plot, we begin by modeling MCA math score as a linear function of time using all 1733 observations and ignoring the school variable. This population (marginal) trend is illustrated in Figure 9.5 and is given by:

\[
\hat{Y}_{ij} = 651.69 + 1.20\text{Time}_{ij},
\]  

(9.1)

where \(\hat{Y}_{ij}\) is the predicted math score of the \(i^{th}\) school at time \(j\), where time \(j\) is the number of years since 2008. In this model, the predicted math score will be identical for all schools at a given time point \(j\). Residuals \(Y_{ij} - \hat{Y}_{ij}\) are then calculated for each observation, measuring the difference between actual math score and the average overall time trend. Figure 9.16 then combines three pieces of information: the upper right triangle contains correlation coefficients for residuals between pairs of years, the diagonal contains histograms of residuals at each time point, and the lower left triangle contains scatterplots of residuals from two different years. In our case, we see that correlation between residuals from adjacent years is strongly positive (0.81-0.83) and does not drop off greatly as the time interval between years increases.

Figure 9.16 Correlation structure within school. The upper right contains correlation coefficients between residuals at pairs of time points, the lower left contains scatterplots of the residuals at time point pairs, and the diagonal contains histograms of residuals at each of the three time points.
9.5 Initial models

Throughout the exploratory analysis phase, our original research questions have guided our work, and now with modeling we return to familiar questions such as:

- are differences between charter and public non-charter schools (in intercept, in slope, in 2010 math score) statistically significant?
- are differences between school types statistically significant, even after accounting for school demographics and location?
- do charter schools offer any measurable benefit over non-charter public schools, either overall or within certain subgroups of schools based on demographics or location?

As you might expect, answers to these questions will arise from proper consideration of variability and properly identified statistical models. As in Chapter 8, we will begin model fitting with some simple, preliminary models, in part to establish a baseline for evaluating larger models. Then, we can build toward a final model for inference by attempting to add important covariates, centering certain variables, and checking assumptions.

9.5.1 Unconditional means model

In the multilevel context, we almost always begin with the unconditional means model, in which there are no predictors at any level. The purpose of the unconditional means model is to assess the amount of variation at each level, and to compare variability within school to variability between schools. Define $Y_{ij}$ as the MCA-II math score from school $i$ and year $j$. Using the composite model specification from Chapter 8:

$$Y_{ij} = \beta_{00} + \epsilon_{0i} + \epsilon_{ij} \sim N(0, \sigma_0^2) \text{ and } \epsilon_{ij} \sim N(0, \sigma^2)$$

the unconditional means model can be fit to the MCA-II data:

**Formula:** MathAvgScore ~ 1 + (1 | schoolid)

**AIC** 10536 10552
**BIC** logLik deviance REMLdev
10536 10552 -5265 10529 10530

**Random effects:**

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>schoolid</td>
<td>(Intercept)</td>
<td>41.869</td>
<td>6.4706</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>10.571</td>
<td>3.2514</td>
</tr>
</tbody>
</table>

**Number of obs:** 1733, groups: schoolid, 618

**Fixed effects:**

<table>
<thead>
<tr>
<th>Estimation</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>652.7460</td>
<td>0.2726</td>
</tr>
</tbody>
</table>

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

- $\hat{\beta}_{00} = 652.7 = \text{the mean math score across all schools and all years}$
- $\hat{\sigma}^2 = 10.6 = \text{the variance in within-school deviations between individual scores and the school mean across all years}$
- $\hat{\sigma}_0^2 = 41.9 = \text{the variance in between-school deviations between school means and the overall mean across all schools and all years}$

Based on the intraclass correlation coefficient:

$$\hat{\rho} = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_0^2 + \hat{\sigma}^2} = \frac{41.869}{41.869 + 10.571} = 0.798$$

79.8 percent of the total variation in math scores is attributable to difference among schools rather than changes over time within schools. We can also say that the average correlation for any pair of responses from the same school is 0.798.
9.5.2 Unconditional growth model

The second model in most multilevel contexts introduces a covariate at Level One (see Model B in Chapter 8). With longitudinal data, this second model introduces time as a predictor at Level One, but there are still no predictors at Level Two. This model is then called the unconditional growth model. The unconditional growth model allows us to assess how much of the within-school variability can be attributed to systematic changes over time.

At the lowest level, we can consider building individual growth models over time for each of the 618 schools in our study. First, we must decide upon a form for each of our 618 growth curves. Based on our initial exploratory analyses, assuming that an individual school’s MCA-II math scores follow a linear trend seems like a reasonable starting point. Under the assumption of linearity, we must estimate an intercept and a slope for each school, based on their 1-3 test scores over a period of three years. Compared to time series analyses of economic data, most longitudinal data analyses have relatively few time periods for each subject (or school), and the basic patterns within subject are often reasonably described by simpler functional forms.

Let $Y_{ij}$ be the math score of the $i^{th}$ school in year $j$. Then we can model the linear change in math test scores over time for school $i$ according to Model B:

$$Y_{ij} = \lambda_{0i} + \lambda_{1i} \text{Year}_{ij} + \epsilon_{ij} \text{ where } \epsilon_{ij} \sim N(0, \sigma^2)$$

The parameters in this model ($\lambda_{0i}, \lambda_{1i},$ and $\sigma^2$) can be estimated through OLS methods. $\lambda_{0i}$ represents the true intercept for School $i$—i.e., the expected test score level for School $i$ when time is zero (2008)—while $\lambda_{1i}$ represents the true slope for School $i$—i.e., the expected yearly rate of change in math score for School $i$ over the three year observation period. The $\epsilon_{ij}$ terms represent the deviation of School $i$’s actual test scores from the expected results under linear growth—the part of school $i$’s test score at time $j$ that is not explained by linear changes over time. The variability in these deviations from the linear model is given by $\sigma^2$. In Figure 9.17, which illustrates a linear growth model for Norwood Central Middle School, $\lambda_{0i}$ is estimated by the $y$-intercept of the fitted regression line, $\lambda_{1i}$ is estimated by the slope of the fitted regression line, and $\sigma^2$ is estimated by the variability in the vertical distances between each point (the actual math score in year $j$) and the line (the predicted math score in year $j$).

![Figure 9.17: Linear growth model for Norwood Central Middle School](image)

In a multilevel model, we let intercepts ($\lambda_{0i}$) and slopes ($\lambda_{1i}$) vary by school and build models for these intercepts and slopes using school-level variables at Level Two. An unconditional growth
model features no predictors at Level Two and can be specified either using formulations at both levels:

- **Level One:**
  \[ Y_{ij} = \lambda_{0i} + \lambda_{1i} Year08_{ij} + \epsilon_{ij} \]

- **Level Two:**
  \[ \lambda_{0i} = \beta_{00} + \epsilon_{0i} \]
  \[ \lambda_{1i} = \beta_{10} + \epsilon_{1i} \]

or as a composite model:

\[ Y_{ij} = \beta_{00} + \beta_{10} Year08_{ij} + \epsilon_{0i} + \epsilon_{1i} Year08_{ij} + \epsilon_{ij} \]

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

\[
\begin{bmatrix}
\epsilon_{0i} \\
\epsilon_{1i}
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right).
\]

As before, \( \sigma^2 \) quantifies the within-school variability (the scatter of points around schools’ linear growth trajectories), while now the between-school variability is partitioned into variability in initial status (\( \sigma_0^2 \)) and variability in rates of change (\( \sigma_1^2 \)).

Using the composite model specification, the unconditional growth model can be fit to the MCA-II test data:

**Formula:** MathAvgScore ~ year08 + (year08 | schoolid)

**Random effects:**

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>schoolid</td>
<td>(Intercept)</td>
<td>39.44156</td>
<td>6.28025</td>
<td></td>
</tr>
<tr>
<td></td>
<td>year08</td>
<td>0.11046</td>
<td>0.33236</td>
<td>0.723</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>8.82005</td>
<td>2.96986</td>
<td></td>
</tr>
</tbody>
</table>

**Number of obs:** 1733, **groups:** schoolid, 618

**Fixed effects:**

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>651.40765</td>
<td>0.27934</td>
</tr>
<tr>
<td>year08</td>
<td>1.26496</td>
<td>0.08997</td>
</tr>
</tbody>
</table>

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

- \( \hat{\beta}_{00} = 651.4 \) = the mean math score for the population of schools in 2008.
- \( \hat{\beta}_{10} = 1.26 \) = the mean yearly change in math test scores for the population during the three year observation period.
- \( \hat{\sigma}^2 = 8.82 \) = the variance in within-school deviations.
- \( \hat{\sigma}_0^2 = 39.4 \) = the variance between schools in 2008 scores.
- \( \hat{\sigma}_1^2 = 0.11 \) = the variance between schools in rates of change in math test scores during the three year observation period.
- \( \hat{\rho}_{01} = 0.72 \) = the correlation in schools’ 2008 math score and their rate of change in scores between 2008 and 2010.

We see that schools had a mean math test score of 651.4 in 2008 and their mean test scores tended to increase by 1.26 points per year over the three year observation period, producing a mean test score at the end of three years of 653.9. According to the t-value (14.1), the increase in mean test scores noted during the three year observation period is statistically significant.
The estimated within-school variance $\hat{\sigma}^2$ decreased by about 17% from the unconditional means model, implying that 17% of within-school variability in test scores can be explained by a linear increase over time:

$$\text{Psuedo } R^2_{L1} = \frac{\sigma^2(\text{uncondmeans}) - \sigma^2(\text{uncondgrowth})}{\sigma^2(\text{uncondmeans})} = \frac{10.571 - 8.820}{10.571} = 0.17$$

### 9.5.3 Modeling other trends over time

While modeling linear trends over time is often a good approximation of reality, it is by no means the only way to model the effect of time. One alternative is to model the quadratic effect of time, which implies adding terms for both time and the square of time. Typically, to reduce the correlation between the linear and quadratic components of the time effect, the time variable is often centered first. Of course, in our case study, there are only 3 possible time points (2008, 2009, and 2010), so a centered time variable would have values -1, 0, and 1, centering on the year 2009. Modifying Model B to produce an **unconditional quadratic growth model** would take the following form:

- **Level One:**
  $$Y_{ij} = \lambda_0 + \lambda_1 \text{Year}08_{ij} + \lambda_2 \text{Year}08_{ij}^2 + \epsilon_{ij}$$

- **Level Two:**
  $$\lambda_0 = \beta_0 + \epsilon_{0i}$$
  $$\lambda_1 = \beta_1 + \epsilon_{1i}$$
  $$\lambda_2 = \beta_2 + \epsilon_{2i}$$

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

$$\begin{bmatrix} \epsilon_{0i} \\ \epsilon_{1i} \\ \epsilon_{2i} \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{bmatrix} \right).$$

With the extra term at Level One for the quadratic effect, we now have 3 equations at Level Two, and 6 variance components at Level Two (3 variance terms and 3 covariance terms). Using the composite model specification, the unconditional quadratic growth model can be fit to the MCA-II test data:

**Formula:** MathAvgScore ~ yearc + yearc2 + (yearc + yearc2 | schoolid)

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>REMLdev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10289</td>
<td>10343</td>
<td>-5134</td>
<td>10263</td>
<td>10269</td>
</tr>
</tbody>
</table>

**Random effects:**

- **Groups** Name     Variance Std.Dev. Corr
  - schoolid (Intercept) 50.6044 7.1137
  - yearc 1.0563 1.0278 0.285
  - yearc2 1.4741 1.2141 -0.681 -0.341
- **Residual** 7.0442 2.6541

**Number of obs:** 1733, groups: schoolid, 618

**Fixed effects:**

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>651.92268</td>
<td>0.30864</td>
<td>2112.3</td>
</tr>
<tr>
<td>yearc</td>
<td>1.24767</td>
<td>0.09041</td>
<td>13.8</td>
</tr>
<tr>
<td>yearc2</td>
<td>1.10155</td>
<td>0.14618</td>
<td>7.5</td>
</tr>
</tbody>
</table>
INITIAL MODELS

From this output, we see that the quadratic effect is positive and significant (t=7.5), in this case indicating that increases in test scores are greater between 2009 and 2010 than between 2008 and 2009. Based on AIC and BIC values, the quadratic growth model outperforms the linear growth model (AIC: 10289 vs. 10352; BIC: 10343 vs. 10384) despite having 4 additional parameters to estimate in the quadratic growth model (1 extra fixed effect and 3 extra variance components). It would be possible to model the quadratic time effect with fewer variance components—for instance, by only using an error term on the intercept at Level Two:

\[
\begin{align*}
\lambda_{0i} &= \beta_{00} + \epsilon_{0i} \\
\lambda_{1i} &= \beta_{10} \\
\lambda_{2i} &= \beta_{20}
\end{align*}
\]

Models like this are frequently used in practice—they allow for a separate overall effect on test scores for each school while minimizing parameters that must be estimated. However, this model does not allow linear and quadratic effects to differ by school, and it’s not clear the simplification provides a significant improvement (AIC favors the larger model, while BIC favors the smaller model). In addition, fixed effect estimates under the simplified error structure remain very similar. Thus, we have reported results from the fuller unconditional quadratic growth model.

Another frequently used approach to modeling time effects is the piecewise linear model. In this model, the complete time span of the study is divided into two or more segments, with a separate slope relating time to the response in each segment. In our case study there is only one piecewise option—fitting separate slopes in 2008-09 and 2009-10. To fit the first segment (2008-09), we zero out the time variable in all other years (only 2010 here), and to fit the second segment (2009-10) we again zero out the time variable in all other years (only 2008). Thus, we now have two time variables at Level One:

- Year0809 = 0 if Year08 = 0, 1 if Year08 = 1, and 0 if Year08 = 2
- Year0910 = 0 if Year08 = 0, 0 if Year08 = 1, and 1 if Year08 = 2

With only 3 time points, creating a piecewise linear model is a bit simplified, but it can still be generalized to segments with more than two years each. Like the model for quadratic growth, we now have an extra equation at Level Two and four additional parameters to estimate (1 extra fixed effect and 3 extra variance components) with full error structure at Level Two:

- Level One:
  \[
  Y_{ij} = \lambda_{0i} + \lambda_{1i} \text{Year0809}_{ij} + \lambda_{2i} \text{Year0910}_{ij} + \epsilon_{ij}
  \]

- Level Two:
  \[
  \begin{align*}
  \lambda_{0i} &= \beta_{00} + \epsilon_{0i} \\
  \lambda_{1i} &= \beta_{10} + \epsilon_{1i} \\
  \lambda_{2i} &= \beta_{20} + \epsilon_{2i}
  \end{align*}
  \]

where error distributions are the same as in the unconditional quadratic growth model. Using the composite model specification, the piecewise linear growth model can be fit using statistical software:

**Formula:** MathAvgScore ~ year0809 + year0910 + (year0809 + year0910 | schoolid)

**Random effects:**

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>schoolid</td>
<td>(Intercept)</td>
<td>39.4730</td>
<td>6.2828</td>
<td></td>
</tr>
<tr>
<td></td>
<td>year0809</td>
<td>6.2379</td>
<td>2.4976</td>
<td>0.201</td>
</tr>
<tr>
<td></td>
<td>year0910</td>
<td>7.0817</td>
<td>2.6611</td>
<td>-0.0130.661</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>5.6159</td>
<td>2.3698</td>
<td></td>
</tr>
</tbody>
</table>

**Number of obs:** 1733, **groups:** schoolid, 618
Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>651.7765</td>
<td>0.2746</td>
</tr>
<tr>
<td>year0809</td>
<td>0.1461</td>
<td>0.1771</td>
</tr>
<tr>
<td>year0910</td>
<td>2.4954</td>
<td>0.1808</td>
</tr>
</tbody>
</table>

The performance of this model is very similar to the quadratic growth model by AIC and BIC measures, and the story told by fixed effects estimates is also very similar. While the mean yearly increase in math scores was 0.1 points between 2008 and 2009, it was 2.5 points between 2009 and 2010.

Despite the good performances of the quadratic growth and piecewise linear models on our three-year window of data, we will continue to use linear growth assumptions in the remainder of this chapter. Not only is a linear model easier to interpret and explain, but it’s probably a more reasonable assumption in years beyond 2010. Predicting future performance is more risky by assuming a steep one year rise or a non-linear rise will continue, rather than by using the average increase over two years.

### 9.6 Building to a final model

#### 9.6.1 Uncontrolled effects of school type

Initially, we can consider whether or not there are significant differences in individual school growth parameters (intercepts and slopes) based on school type. From a modeling perspective, we would build a system of two Level Two models:

\[
\begin{align*}
\lambda_{0i} &= \beta_{00} + \beta_{01} Charter_i + \varepsilon_{0i} \\
\lambda_{1i} &= \beta_{10} + \beta_{11} Charter_i + \varepsilon_{1i}
\end{align*}
\]

where \( Charter_i = 1 \) if School \( i \) is a charter school and \( Charter_i = 0 \) if School \( i \) is a non-charter public school. In addition, the error terms at Level Two are assumed to follow a multivariate normal distribution:

\[
\begin{bmatrix}
\varepsilon_{0i} \\
\varepsilon_{1i}
\end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}\right)
\]

With a binary predictor at Level Two such as school type, we can write out what our Level Two model looks like for public non-charter schools and charter schools.

- Public schools
  \[
  \begin{align*}
  \lambda_{0i} &= \beta_{00} + \varepsilon_{0i} \\
  \lambda_{1i} &= \beta_{10} + \varepsilon_{1i}
  \end{align*}
  \]

- Charter schools
  \[
  \begin{align*}
  \lambda_{0i} &= (\beta_{00} + \varepsilon_{0i}) \\
  \lambda_{1i} &= (\beta_{10} + \varepsilon_{1i})
  \end{align*}
  \]

Writing the Level Two model in this manner helps us interpret the model parameters from our two-level model. We can use statistical software (such as the lmer() function from the lme4 package in R) to obtain parameter estimates using our 1733 observations, after first converting our Level One and Level Two models into a composite model (Model C) with fixed effects and variance components separated:

\[
Y_{ij} = \lambda_{0i} + \lambda_{1i} Year08_{ij} + \varepsilon_{ij}
\]

\[
= (\beta_{00} + \varepsilon_{0i}) + (\beta_{10} + \beta_{11} Charter_i + \varepsilon_{1i}) Year08_{ij} + \varepsilon_{ij}
\]

\[
= [\beta_{00} + \beta_{10} Year08_i + \beta_{01} Charter_i + \beta_{11} Charter_i Year08_{ij}] + [\varepsilon_{0i} + \varepsilon_{1i} Year08_{ij} + \varepsilon_{ij}]
\]
BUILDING TO A FINAL MODEL

Formula: MathAvgScore ~ charter + year08 + charter:year08 + (year08 | schoolid)

AIC  BIC  logLik  deviance  REMLdev
10308  10351 -5146  10289  10292

Random effects:
Groups   Name   Variance  Std.Dev.  Corr
schoolid  (Intercept)  35.83184  5.98597
            year08    0.13115  0.36214  0.880
Residual         8.78447  2.96386

Number of obs: 1733, groups: schoolid, 618

Fixed effects:
     Estimate Std. Error  t value
(Intercept)     652.05843    0.28449   2292.0
charter          -6.01843    0.86562    -7.0
year08           1.19709    0.09427    12.7
charter:year08   0.85571    0.31429     2.7

Armed with our parameter estimates, we can offer concrete interpretations:

• Fixed effects:
  – $\hat{\beta}_{00} = 652.1$. The estimated mean test score for 2008 for non-charter public schools is 652.1.
  – $\hat{\beta}_{01} = -6.02$. Charter schools have an estimated test score in 2008 which is 6.02 points lower than public non-charter schools.
  – $\hat{\beta}_{10} = 1.20$. Public non-charter schools have an estimated mean increase in test scores of 1.20 points per year.
  – $\hat{\beta}_{11} = 0.86$. Charter schools have an estimated mean increase in test scores of 2.06 points per year over the three year observation period, 0.86 points higher than the mean yearly increase among public non-charter schools.

• Variance components:
  – $\hat{\sigma}_0 = 5.99$. The estimated standard deviation of 2008 test scores is 5.99 points, after controlling for school type.
  – $\hat{\sigma}_1 = 0.36$. The estimated standard deviation of yearly changes in test scores during the three year observation period is 0.36 points, after controlling for school type.
  – $\rho_{01} = 0.88$. The estimated correlation between 2008 test scores and yearly changes in test scores is 0.88, after controlling for school type.
  – $\hat{\sigma} = 2.96$. The estimated standard deviation in residuals for the individual growth curves is 2.96 points.

Based on t-values reported by R, the effects of time and school type both appear to be statistically significant, and there is also significant evidence of an interaction between time and school type. Public schools had a significantly higher mean math score in 2008, while charter schools had significantly greater improvement in scores between 2008 and 2010 (although the mean score of charter schools still lagged behind that of public schools in 2010, as indicated in the graphical comparison of models B and C in Figure 9.18). Based on pseudo R-square values, the addition of a charter school indicator to the unconditional growth model has decreased unexplained school-to-school variability in 2008 math scores by 4.7%, while unexplained variability in yearly improvement actually increased slightly. Obviously, it makes little sense that introducing an additional predictor would reduce the amount of variability in test scores explained, but this is an example of the limitations in the pseudo $R^2$ values discussed in Section 8.7.2.
9.6.2 Add percent free and reduced lunch as a covariate

Although we will still be primarily interested in the effect of school type on both 2008 test scores and rate of change in test scores (as we observed in Model C), we can try to improve our estimates of school type effects through the introduction of meaningful covariates. In this study, we are particularly interested in Level Two covariates—those variables which differ by school but which remain basically constant for a given school over time—such as urban or rural location, percentage of special education students, and percentage of students with free and reduced lunch. In Section 9.4, we investigated the relationship between percent free and reduced lunch and a school’s test score in 2008 and their rate of change from 2008 to 2010.

Based on these analyses, we will begin by adding percent free and reduced lunch as a Level Two predictor for both intercept and slope (Model D):

- Level One:
  \[ Y_{ij} = \lambda_0 + \lambda_1 \text{Year08}_{ij} + \epsilon_{ij} \]

- Level Two:
  \[ \lambda_0 = \beta_{00} + \beta_{01} \text{Charter}_i + \beta_{02} \text{SchPctFree}_i + \epsilon_{0i} \]
  \[ \lambda_1 = \beta_{10} + \beta_{11} \text{Charter}_i + \beta_{12} \text{SchPctFree}_i + \epsilon_{1i} \]

The composite model is then:
\[ Y_{ij} = [\beta_{00} + \beta_{01} \text{Charter}_i + \beta_{02} \text{SchPctFree}_i + \beta_{10} \text{Year08}_i + \beta_{11} \text{Charter}_i \text{Year08}_i + \beta_{12} \text{SchPctFree}_i \text{Year08}_i] + [\epsilon_{0i} + \epsilon_{1i} \text{Year08}_i + \epsilon_{ij}] \]

where error terms are defined as in Model C.

Formula: MathAvgScore ~ charter + SchPctFree + year08 + charter:year08 + SchPctFree:year08 + (year08 | schoolid)
AIC  BIC  logLik  deviance  REMLdev
9988 10043  -4984  9947  9968

Random effects:

- Groups  Name  Variance  Std.Dev.  Corr
- schoolid  (Intercept)  19.1318  4.37400
- year08  0.1603  0.40037  0.515
Residual 8.7981 2.96616  
Number of obs: 1733, groups: schoolid, 618

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>659.278481</td>
<td>0.444688</td>
</tr>
<tr>
<td>charter</td>
<td>-3.439943</td>
<td>0.712834</td>
</tr>
<tr>
<td>SchPctFree</td>
<td>-0.166539</td>
<td>0.008907</td>
</tr>
<tr>
<td>year08</td>
<td>1.641369</td>
<td>0.189498</td>
</tr>
<tr>
<td>charter:year08</td>
<td>0.980762</td>
<td>0.318582</td>
</tr>
<tr>
<td>SchPctFree:year08</td>
<td>-0.010409</td>
<td>0.003839</td>
</tr>
</tbody>
</table>

Compared to Model C, the introduction of school-level poverty based on percentage of students receiving free and reduced lunch in Model D leads to similar conclusions about the significance of the charter school effect on both the intercept and the slope, although the magnitude of these estimates change after controlling for poverty levels. The estimated gap in test scores between charter and non-charter schools in 2008 is smaller in Model D, while estimates of improvement between 2008 and 2010 increase for both types of schools. Inclusion of free and reduced lunch reduces the unexplained variability between schools in 2008 math scores by 27%, while unexplained variability in rates of change between schools again increases slightly based on pseudo $R^2$ values. A likelihood ratio test using maximum likelihood estimates illustrates that adding free and reduced lunch as a Level Two covariate significantly improves our model ($\chi^2 = 341.5, df = 2, p < .001$). Specific fixed effect parameter estimates are given below:

- $\hat{\beta}_{00} = 659.3$. The estimated mean math test score for 2008 is 659.3 for non-charter public schools with no students receiving free and reduced lunch.
- $\hat{\beta}_{01} = -3.44$. Charter schools have an estimated mean math test score in 2008 which is 3.44 points lower than non-charter public schools, controlling for effects of school-level poverty.
- $\hat{\beta}_{02} = -0.17$. Each 10% increase in the percentage of students at a school receiving free and reduced lunch is associated with a 1.7 point decrease in mean math test scores for 2008, after controlling for school type.
- $\hat{\beta}_{10} = 1.64$. Public non-charter schools with no students receiving free and reduced lunch have an estimated mean increase in math test score of 1.64 points per year during the three years of observation.
- $\hat{\beta}_{11} = 0.98$. Charter schools have an estimated mean yearly increase in math test scores over the three year observation period of 2.62, which is 0.98 points higher than the annual increase for public non-charter schools, after controlling for school-level poverty.
- $\hat{\beta}_{12} = -0.010$. Each 10% increase in the percentage of students at a school receiving free and reduced lunch is associated with a 0.10 point decrease in rate of change over the three years of observation, after controlling for school type.

### 9.6.3 A potential final model with three Level Two covariates

We now begin iterating toward a “final model” for these data, on which we will base conclusions. Being cognizant of typical features of a “final model” as outlined in Chapter 8, we offer one possible final model for this data—Model F:

- **Level One:**
  \[ Y_{ij} = \lambda_{0i} + \lambda_{1i} \text{Year08}_{ij} + \epsilon_{ij} \]
- **Level Two:**
  \[ \lambda_{0i} = \beta_{00} + \beta_{01} \text{Charter}_i + \beta_{02} \text{urban}_i + \beta_{03} \text{schpctsped}_i + \beta_{04} \text{schpctfree}_i + \epsilon_{0i} \]
  \[ \lambda_{1i} = \beta_{10} + \beta_{11} \text{Charter}_i + \beta_{12} \text{urban}_i + \beta_{13} \text{schpctsped}_i + \epsilon_{1i} \]
where we find the effect of charter schools on 2008 test scores after adjusting for urban or rural location, percentage of special education students, and percentage of students that receive free or reduced lunch, and the effect of charter schools on yearly change between 2008 and 2010 after adjusting for urban or rural location and percentage of special education students. We can use AIC and BIC criteria to compare Model F with Model D, since the two models are not nested. By both criteria, Model F is significantly better than Model D: AIC of 9855 vs. 9988, and BIC of 9956 vs. 10043. Based on the R output below, we offer interpretations for estimates of model fixed effects:

Formula: MathAvgScore ~ charter + urban + SchPctFree + SchPctSped +
  charter:year08 + urban:year08 + SchPctSped:year08 + year08 +
  (year08 | schoolid)

AIC  BIC  logLik deviance REMLdev
9885 9956  -4929   9831    9859

Random effects:
Groups   Name   Variance   Std.Dev.   Corr
schoolid (Intercept)   16.9468107   4.116650
  year08   0.0035586   0.059654   0.981
Residual            8.8230706   2.970365
Number of obs: 1733, groups: schoolid, 618

Fixed effects:
                        Estimate   Std. Error    t value
(Intercept)        661.010466   0.512887   1288.8
charter           -3.222819   0.698546    -4.6
urban              -1.113834   0.427567    -2.6
SchPctFree         -0.152814   0.008096   -18.9
SchPctSped         -0.117702   0.020612    -5.7
year08             2.144274   0.200854    10.7
charter:year08     1.030854   0.315139     3.3
urban:year08       -0.527489   0.186467    -2.8
SchPctSped:year08  -0.046740   0.010165   -4.6

- \hat{\beta}_{00} = 661.0. The estimated mean math test score for 2008 is 661.0 for public schools in rural areas with no students qualifying for special education or free and reduced lunch.
- \hat{\beta}_{01} = -3.22. Charter schools have an estimated mean math test score in 2008 which is 3.22 points lower than non-charter public schools, after controlling for urban or rural location, percent special education, and percent free and reduced lunch.
- \hat{\beta}_{02} = -1.11. Schools in urban areas have an estimated mean math score in 2008 which is 1.11 points lower than schools in rural areas, after controlling for school type, percent special education, and percent free and reduced lunch.
- \hat{\beta}_{03} = -0.118. A 10% increase in special education students at a school is associated with a 1.18 point decrease in estimated mean math score for 2008, after controlling for school type, urban or rural location, and percent free and reduced lunch.
- \hat{\beta}_{04} = -0.153. A 10% increase in free and reduced lunch students at a school is associated with a 1.53 point decrease in estimated mean math score for 2008, after controlling for school type, urban or rural location, and percent special education.
- \hat{\beta}_{10} = 2.14. Public non-charter schools in rural areas with no students qualifying for special education have an estimated increase in mean math test score of 2.14 points per year over the three year observation period, after controlling for percent of students receiving free and reduced lunch.
- \hat{\beta}_{11} = 1.03. Charter schools have an estimated mean annual increase in math score that is 1.03
points higher than public non-charter schools over the three year observation period, after controlling for urban or rural location, percent special education, and percent free and reduced lunch.

- $\hat{\beta}_{12} = -0.53$. Schools in urban areas have an estimated mean annual increase in math score that is 0.53 points lower than schools from rural areas over the three year observation period, after controlling for school type, percent special education, and percent free and reduced lunch.
- $\hat{\beta}_{13} = -0.047$. A 10% increase in special education students at a school is associated with an estimated mean annual increase in math score that is 0.47 points lower over the three year observation period, after controlling for school type, urban or rural location, and percent free and reduced lunch.

From this model, we again see that 2008 sixth grade math test scores from charter schools were significantly lower than similar scores from public non-charter schools, after controlling for school location and demographics. However, charter schools showed significantly greater improvement between 2008 and 2010 compared to public non-charter schools, although charter school test scores were still lower than public school scores in 2010, on average. We also tested several interactions between Level Two covariates and charter schools and found none to be significant, indicating that the 2008 gap between charter schools and public non-charter schools was consistent across demographic subgroups. The faster improvement between 2008 and 2010 for charter schools was also consistent across demographic subgroups. Controlling for school location and demographic variables provided more reliable and nuanced estimates of the effects of charter schools, while also providing interesting insights. For example, schools in rural areas not only had higher test scores than schools in urban areas in 2008, but the gap grew larger over the study period given fixed levels of percent special education, percent free and reduced lunch, and school type. In addition, schools with higher levels of poverty lagged behind other schools and showed no signs of closing the gap, and schools with higher levels of special education students had both lower test scores in 2008 and slower rates of improvement during the study period, again given fixed levels of other covariates.

As we demonstrated in this case study, applying multilevel methods to two-level longitudinal data yields valuable insights about our original research questions while properly accounting for the structure of the data.

9.7 Covariance structure among observations

Part of our motivation for framing our model for multilevel data was to account for the correlation among observations made on the same school (the Level Two observational unit). Our two-level model, through error terms on both Level One and Level Two variables, actually implies a specific within-school covariance structure among observations, yet we have not (until now) focused on this imposed structure; for example:

- What does our two-level model say about the relative variability of 2008 and 2010 scores from the same school?
- What does it say about the correlation between 2008 and 2009 scores from the same school?

In this section, we will describe the within-school covariance structure imposed by our two-level model and offer alternative covariance structures that we might consider, especially in the context of longitudinal data. In short, we will discuss how we might decide if our implicit covariance structure in our two-level model is satisfactory for the data at hand. Then, in the succeeding optional section, we provide derivations of the imposed within-school covariance structure for our standard two-level model using results from probability theory.
9.7.1 Standard covariance structure

We will use Model C (uncontrolled effects of school type) to illustrate covariance structure within subjects. Recall that, in composite form, Model C is:

\[ Y_{ij} = \lambda_{0i} + \lambda_{1i} \text{Year08}_{ij} + \epsilon_{ij} \]
\[ = (\beta_{00} + \beta_{01} \text{Charter}_i + \epsilon_{0i}) + (\beta_{10} + \beta_{11} \text{Charter}_i + \epsilon_{1i}) \text{Year08}_{ij} + \epsilon_{ij} \]
\[ = [\beta_{00} + \beta_{01} \text{Charter}_i + \beta_{10} \text{Year08}_i + \beta_{11} \text{Charter}_i \text{Year08}_{ij}] + [\epsilon_{0i} + \epsilon_{1i} \text{Year08}_{ij} + \epsilon_{ij}] \]

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

\[
\begin{bmatrix}
\epsilon_{0i} \\
\epsilon_{1i}
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\
\sigma_{01} & \sigma_1^2 \end{bmatrix} \right).
\]

For School \( i \), the covariance structure for the three time points has general form:

\[
\text{Cov}(Y) = \begin{bmatrix}
\text{Var}(Y_1) & \text{Cov}(Y_1, Y_2) & \text{Cov}(Y_1, Y_3) \\
\text{Cov}(Y_1, Y_2) & \text{Var}(Y_2) & \text{Cov}(Y_2, Y_3) \\
\text{Cov}(Y_1, Y_3) & \text{Cov}(Y_2, Y_3) & \text{Var}(Y_3)
\end{bmatrix}
\]

where, for instance, \( \text{Var}(Y_1) \) is the variability in 2008 test scores (time \( j = 1 \)). \( \text{Cov}(Y_1, Y_2) \) is the covariance between 2008 and 2009 test scores (times \( j = 1 \) and \( j = 2 \), etc. Since covariance measures the tendency of two measures to move together, we expect positive values for all three covariance terms in \( \text{Cov}(Y) \), since schools with relatively high test scores in 2008 are likely to also have relatively high test scores in 2009 or 2010. The correlation between two measures then scales covariance terms to values between -1 and 1, so by the same rationale, we expect correlation coefficients between two years to be near 1. If observations within school were independent—that is, knowing a school had relatively high scores in 2008 tells nothing about whether that school will have relatively high scores in 2009 or 2010—then we would expect covariance and correlation values near 0.

It is important to notice that the error structure at Level Two is not the same as the within-school covariance structure among observations. That is, the relationship between \( \epsilon_{0i} \) and \( \epsilon_{1i} \) from the Level Two equations is not the same as the relationship between test scores from different years at the same school (e.g., the relationship between \( Y_{1i} \) and \( Y_{2i} \)). In other words,

\[
\text{Cov}(Y) \neq \begin{bmatrix}
\epsilon_{0i} \\
\epsilon_{1i}
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\
\sigma_{01} & \sigma_1^2 \end{bmatrix} \right).
\]

Yet, the error structure and the covariance structure are connected to each other, as we will now explore.

Using results from probability theory (see Section 9.7.5), we can show that:

\[
\text{Var}(Y_j) = \sigma_0^2 + t_j^2 \sigma_1^2 + \sigma^2 + 2t_j \sigma_{01},
\]
\[
\text{Cov}(Y_j, Y_k) = \sigma_0^2 + t_j t_k \sigma_1^2 + (t_j + t_k) \sigma_{01}.
\]

where our time variable (Year08) has values \( t_1 = 0, t_2 = 1, \) and \( t_3 = 2 \). Intuitively, these formulas are sensible. For instance, the uncertainty (variability) around a school’s score in 2008 increases as the uncertainty in intercepts and slopes increases, as the uncertainty around that school’s linear time trend increases, and as the covariance between intercept and slope residuals increases (since if one is off, the other one is likely off as well). Also, the covariance between 2008 and 2009 scores does not depend on level one error. Thus, in the 3-by-3 within-school covariance structure of the charter schools case study, our standard two-level model determines all 6 covariance matrix elements.
through the estimation of three parameters and the imposition of a specific structure related to time.

To obtain estimated variances for individual observations and covariances between two time points from the same school, we can simply plug estimated variance components from our two-level model along with time points from our data collection into the equations above. For instance, in Section 9.6.1, we obtained the following estimates of variance components: $\hat{\sigma}^2 = 8.69$, $\hat{\sigma}_0^2 = 35.97$, $\hat{\sigma}_1^2 = 0.23$, and $\hat{\rho}_{01} = \hat{\rho}_{00}\hat{\sigma}_1 = 1.81$. Therefore, our estimated within-school variances for the three time points would be:

$$
\begin{align*}
\text{Var}(Y_1) &= 35.97 + 0.23 + 8.69 + 2(0)1.81 = 44.66 \\
\text{Var}(Y_2) &= 35.97 + 1^20.23 + 8.69 + 2(1)1.81 = 48.50 \\
\text{Var}(Y_3) &= 35.97 + 2^20.23 + 8.69 + 2(2)1.81 = 52.80 \\
\end{align*}
$$

and our estimated within-school covariances between different time points would be:

$$
\begin{align*}
\hat{\text{Cov}}(Y_1, Y_2) &= 35.97 + (0)(1)0.23 + (0 + 1)1.81 = 37.77 \\
\hat{\text{Cov}}(Y_1, Y_3) &= 35.97 + (0)(2)0.23 + (0 + 2)1.81 = 39.58 \\
\hat{\text{Cov}}(Y_2, Y_3) &= 35.97 + (1)(2)0.23 + (1 + 2)1.81 = 41.84 \\
\end{align*}
$$

Written in matrix form, our two-level model implicitly imposes this estimated covariance structure on within-school observations:

$$
\hat{\text{Cov}}(Y) = \begin{bmatrix} 44.66 & 37.77 & 39.58 \\
37.77 & 48.50 & 41.84 \\
39.58 & 41.84 & 52.80 \end{bmatrix}
$$

and this estimated covariance matrix can be converted into an estimated within-school correlation matrix using the identity $\hat{\text{Corr}}(Y_1, Y_2) = \hat{\text{Cov}}(Y_1, Y_2) / \sqrt{\text{Var}(Y_1)\text{Var}(Y_2)}$:

$$
\hat{\text{Corr}}(Y) = \begin{bmatrix} 1 & .812 & .815 \\
.812 & 1 & .827 \\
.815 & .827 & 1 \end{bmatrix}
$$

A couple of features of these two matrices can be highlighted that offer insights into implications of our standard two-level model on the covariance structure among observations at Level One from the same school:

- Many longitudinal data sets show higher correlation for observations that are closer in time. In this case, we see that correlation is very consistent between all pairs of observations from the same school; the correlation between test scores separated by two years (.815) is approximately the same as the correlation between test scores separated by a single year (.812 for 2008 and 2009 scores; .827 for 2009 and 2010 scores).

- Many longitudinal data sets show similar variability at all time points. In this case, the variability in 2010 (52.80) is about 18% greater than the variability in 2008 (44.66), while the variability in 2009 is in between (48.50).

- Our two-level model actually imposes a quadratic structure on the relationship between variance and time; note that the equation for $\text{Var}(Y_t)$ contains both $t^2$ and $t$. The variance is therefore minimized at $t = -\frac{\hat{\sigma}_1}{\hat{\sigma}_0}$. With the charter school data, the variance in test scores is minimized when $t = -\frac{\hat{\sigma}_1}{\hat{\sigma}_0} = -\frac{1.81}{0.23} = -8.0$; that is, the smallest within-school variance in test scores is expected 8.0 years prior to 2008 (i.e., in 2000), and the variance increases parabolically from there. In
general, cases in which $\sigma_1^2$ and $\sigma_{01}$ are relatively small have little curvature and fairly consistent variability over time.

- There is no requirement that time points within school need to be evenly spaced or even that each school has an equal number of measurements over time, which makes the two-level model structure nicely flexible.

9.7.2 Alternative covariance structures

The standard covariance structure that’s implied by our multilevel modeling structure provides a useful model in a wide variety of situations—it provides flexibility to modeling Level One variability with a relatively small number of parameters, and it has sufficient flexibility to accommodate irregular time intervals as well as subjects with different number of observations over time. However, there may be cases in which a better fitting model requires additional parameters, or when a simpler model with fewer parameters still provides a good fit to the data. Here is an outline of a few alternative error structures:

- **Unstructured** - Every variance and covariance term for observations within a school is a separate parameter and is therefore estimated uniquely; no patterns among variances or correlations are assumed. This structure offers maximum flexibility but is most costly in terms of parameters estimated.

- **Compound symmetry** - Assume variance is constant across all time points and correlation is constant across all pairs of time points. This structure is highly restrictive but least costly in terms of parameters estimated.

- **Autoregressive** - Assume variance is constant across all time points, but correlation drops off in a systematic fashion as the gap in time increases. Autoregressive models expand compound symmetry by allowing for a common structure where points closest in time are most highly correlated.

- **Toeplitz** - Toeplitz is similar to the autoregressive model, except that it does not impose any structure on the decline in correlation as time gaps increase. Thus, it requires more parameters to be estimated than the autoregressive model while providing additional flexibility.

- **Heterogeneous variances** - The assumption that variances are equal across time points found in the compound symmetry, autoregressive, and Toeplitz models can be relaxed by introducing additional parameters to allow unequal (heterogeneous) variances.

When the focus of an analysis is on stochastic parameters (variance components) rather than fixed effects, parameter estimates are typically based on restricted maximum likelihood (REML) methods; model performance statistics then reflect only the stochastic portion of the model. Models with the same fixed effects but different covariance structures can be compared as usual—with AIC and BIC measures when models are not nested and with likelihood ratio tests when models are nested. However, using a chi-square distribution to conduct a likelihood ratio test in these cases can often produce a conservative test, with p-values that are too large and not rejected enough (Raudenbush and Bryk, Singer and Willett, Faraway). In Section 10.6, we introduce the parametric bootstrap as a potentially better way of testing models nested in their random effects.
9.7.3 Covariance structure in non-longitudinal multilevel models

Careful modeling and estimation of the Level One covariance matrix is especially important and valuable for longitudinal data (with time at Level One) and as we’ve seen, our standard two-level model has several nice properties for this purpose. The standard model is also often appropriate for non-longitudinal multilevel models as discussed in Chapter 8, although we must remain aware of the covariance structure implicitly imposed. In other words, the ideas in this section generalize even if time isn’t a Level One covariate.

As an example, in Case Study 8.2 where Level One observational units are musical performances rather than time points, the standard model implies the following covariance structure for Musician $i$ in Model C, which uses an indicator for large ensembles as a Level One predictor:

$$
\begin{align*}
\text{Var}(Y_{ij}) &= \sigma_0^2 + \text{Large}_{ij}\sigma_1^2 + \sigma_3^2 + 2\text{Large}_{ij}\sigma_{01} \\
&= \begin{cases} 
\sigma_0^2 + \sigma_1^2 & \text{if } \text{Large}_{ij} = 0 \\
\sigma_0^2 + \sigma_3^2 + 2\sigma_{01} & \text{if } \text{Large}_{ij} = 1
\end{cases}
\end{align*}
$$

and

$$
\begin{align*}
\text{Cov}(Y_{ij}, Y_{ik}) &= \sigma_0^2 + \text{Large}_{ij}\text{Large}_{ik}\sigma_1^2 + (\text{Large}_{ij} + \text{Large}_{ik})\sigma_{01} \\
&= \begin{cases} 
\sigma_0^2 + \sigma_{01} & \text{if } \text{Large}_{ij} = \text{Large}_{ik} = 0 \\
\sigma_0^2 + \sigma_{01} & \text{if } \text{Large}_{ij} = 0 \text{ and } \text{Large}_{ik} = 1 \text{ or vice versa} \\
\sigma_0^2 + 2\sigma_{01} & \text{if } \text{Large}_{ij} = \text{Large}_{ik} = 1
\end{cases}
\end{align*}
$$

Note that, in the Music Performance Anxiety case study, each subject will have a unique Level One variance-covariance structure, since each subject has a different number of performances and a different mix of large ensemble and small ensemble or solo performances.

9.7.4 Final thoughts regarding covariance structures

In the charter school example, as is often true in multilevel models, the choice of covariance matrix does not greatly affect estimates of fixed effects. The choice of covariance structure could potentially impact the standard errors of fixed effects, and thus the associated test statistics, but the impact appears minimal in this particular case study. In fact, the standard model typically works very well. So is it worth the time and effort to accurately model the covariance structure? If primary interest is in inference regarding fixed effects, and if the standard errors for the fixed effects appear robust to choice of covariance structure, then extensive time spent modeling the covariance structure is not advised. However, if researchers are interested in predicted random effects and estimated variance components in addition to estimated fixed effects, then choice of covariance structure can make a big difference. For instance, if researchers are interested in drawing conclusions about particular schools rather than charter schools in general, they may more carefully model the covariance structure in this study.

9.7.5 Details of covariance structures (Optional)

Using Model C as specified in Section 9.7.1, we specified the general covariance structure for School $i$ as:

$$
\text{Cov}(Y_i) = \begin{bmatrix}
\text{Var}(Y_{i1}) & \text{Cov}(Y_{i1}, Y_{i2}) & \text{Cov}(Y_{i1}, Y_{i3}) \\
\text{Cov}(Y_{i1}, Y_{i2}) & \text{Var}(Y_{i2}) & \text{Cov}(Y_{i2}, Y_{i3}) \\
\text{Cov}(Y_{i1}, Y_{i3}) & \text{Cov}(Y_{i2}, Y_{i3}) & \text{Var}(Y_{i3})
\end{bmatrix}
$$

If $Y_1 = a_1X_1 + a_2X_2 + a_3$ and $Y_2 = b_1X_1 + b_2X_2 + b_3$ where $X_1$ and $X_2$ are random variables and $a_i$ and $b_i$ are constants for $i = 1, 2, 3$, then we know from probability theory that:

$$
\text{Var}(Y_1) = a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + 2a_1a_2\text{Cov}(X_1, X_2)
$$
\[ \text{Cov}(Y_1, Y_2) = a_1 b_1 \text{Var}(X_1) + a_2 b_2 \text{Var}(X_2) + (a_1 b_2 + a_2 b_1) \text{Cov}(X_1, X_2) \]

Applying these identities to Model C, we first see that we can ignore all fixed effects, since they do not contribute to the variability. Thus,

\[ \text{Var}(Y_{ij}) = \text{Var}(\varepsilon_0 + \varepsilon_i \text{Year}_08_{ij} + \varepsilon_{ij}) = \text{Var}(\varepsilon_0) + \text{Var}(\varepsilon_i \text{Year}_08_{ij}) + 2 \text{Year}_08_{ij} \text{Cov}(\varepsilon_0, \varepsilon_{ij}) \]

\[ = \sigma_0^2 + \text{Year}_08_{ij} \sigma_i^2 + \sigma^2 + 2 \text{Year}_08_{ij} \sigma_{01} \]

where the last line reflects the fact that observations were taken at the same time points for all schools. We can derive the covariance terms in a similar fashion:

\[ \text{Cov}(Y_{ij}, Y_{ik}) = \text{Cov}(\varepsilon_0 + \varepsilon_i \text{Year}_08_{ij} + \varepsilon_{ij}, \varepsilon_0 + \varepsilon_i \text{Year}_08_{ik} + \varepsilon_{ik}) \]

\[ = \text{Var}(\varepsilon_0) + \text{Var}(\varepsilon_i 	ext{Year}_08_{ij}) + \text{Var}(\varepsilon_i \text{Year}_08_{ik}) + (\text{Year}_08_{ij} + \text{Year}_08_{ik}) \text{Cov}(\varepsilon_0, \varepsilon_{ij}) \]

\[ = \sigma_0^2 + \text{Year}_08_{ij} \sigma_i^2 + \sigma^2 + 2 \text{Year}_08_{ij} \sigma_{01} \]

In Model C, we obtained the following estimates of variance components: \( \hat{\sigma}^2 = 8.69, \hat{\sigma}_0^2 = 35.97, \hat{\sigma}_1^2 = 0.23, \) and \( \hat{\sigma}_{01} = \beta \hat{\sigma}_0 \hat{\sigma}_1 = 1.81. \) Therefore, our two level model implicitly imposes this covariance structure on within subject observations:

\[
\text{Cov}(Y_{ij}) = \begin{bmatrix}
44.66 \\
37.77 \\
39.58
\end{bmatrix}
\]

and this covariance matrix can be converted into a within-subject correlation matrix:

\[
\text{Corr}(Y_{ij}) = \begin{bmatrix}
1 & .812 & .815 \\
.812 & 1 & .827 \\
.815 & .827 & 1
\end{bmatrix}
\]

### 9.8 Notes on Using R (Optional)

Data for Case Study 9.2 can be found in our data repository. This data is stored in WIDE format, with one row per subject. The R code below performs the data organization tasks described in Section 9.3: read in the data in WIDE format, investigate patterns of missingness using the mice package, and use the reshape() function to create our LONG data set.

```r
# chart_wide_condense.csv 618x11
chart.wide = read.csv(file=file.choose())
l library(mice)
md.pattern(chart.wide)

# Create data frame in LONG form (one obs per school-year)
# chart.long is 1854x10 with 121 NAs for MathAvgScore
chart.long0=reshape(chart.wide,idvar="schoolid",
 varying=c("MathAvgScore.0", "MathAvgScore.1", "MathAvgScore.2"),
 timevar="year08", time=c(0,1,2),
direction="long")

# Add average across all years for each school for EDA plots
mean3yr = by(chart.long$MathAvgScore,chart.long$schoolid,mean,na.rm=T)
urban0 = ifelse(chart.wide$urban==1, "urban", "rural")
```
charter0 = ifelse(chart.wide$charter==1,"charter","public non-charter")
chart.wide = data.frame(chart.wide,mean3yr=as.numeric(mean3yr),
                       urban0=urban0,charter0=charter0)

The model below is our final model with $\sigma_{01}$ set to 0—i.e., we have added the restriction that Level Two error terms are uncorrelated. Motivation for this restriction came from repeated estimates of correlation in different versions of the final model near 1, when empirically a slightly negative correlation might be expected. As we will describe in Chapter ??, inclusion of the Level Two correlation as a model parameter appears to lead to boundary constraints—maximum likelihood parameter estimates near the maximum or minimum allowable value for a parameter. A likelihood ratio test using full maximum likelihood estimates confirms that the inclusion of a correlation term does not lead to an improved model (LRT test statistic = .223 on 1 df, $p = .637$). Estimates of fixed effects and their standard errors are extremely consistent with the full model in Section 9.6.3; only the estimates of the variability in $\sigma_1$ is noticeably higher.

# Modified final model
model.f2a <- lmer(MathAvgScore ~ charter + urban + SchPctFree +
                   SchPctSped + charter:year08 + urban:year08 +
                   SchPctSped:year08 + year08 +
                   (1|schoolid) + (0+year08|schoolid), REML=T, data=chart.long)
summary(model.f2a)

# R output
Linear mixed model fit by REML
Formula: MathAvgScore ~ charter + urban + SchPctFree + SchPctSped +
        charter:year08 + urban:year08 + SchPctSped:year08 + year08 +
        (1 | schoolid) + (0 + year08 | schoolid)
Data: chart.long
AIC BIC logLik deviance REMLdev
9883 9948 -4929 9831 9859

Random effects:
 Groups    Name  Variance Std.Dev.          
schoolid (Intercept) 17.35510 4.16659
schoolid year08      0.11391 0.33751
Residual             8.71628 2.95233
Number of obs: 1733, groups: schoolid, 618

Fixed effects:
(Intercept)     Estimate  Std. Error   t value
charter         -3.22467   0.70317   -4.60
urban           -1.11664   0.43042   -2.60
SchPctFree      -0.15295   0.00809   -18.9
SchPctSped      -0.11777   0.02074   -5.70
year08          2.14269    0.20208   10.60
charter:year08  1.03341    0.31716    3.20
urban:year08    -0.52442    0.18767   -2.80
SchPctSped:year08 -0.04672    0.01022   -4.60

# LRT comparing final model in chapter (model.f2ml) with maximum
# likelihood estimates to modified final model (model.f2aml)
# with uncorrelated Level Two errors.
anovamodel.f2ml, model.f2aml

Df AIC BIC logLik Chisq Chi Df Pr(>Chisq)
model.f2aml 12 9855.5 9920.9 -4915.7
model.f2ml 13 9857.2 9928.2 -4915.6 0.2231 1 0.6367

9.9 Exercises

9.9.1 Conceptual Exercises

1. Parenting and Gang Activity. Walker-Barnes and Mason (2001) describe “Ethnic differences in the effect of parenting on gang involvement and gang delinquency: a longitudinal, hierarchical linear modeling perspective”. In this study, 300 ninth graders from one high school in an urban southeastern city were assessed at the beginning of the school year about their gang activity, the gang activity of their peers, behavior of their parents, and their ethnic and cultural heritage. Then, information about their gang activity was collected at 7 additional occasions during the school year. For this study: (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. Describe the difference between the wide and long formats for longitudinal data in this study.

3. Describe scenarios or research questions in which a lattice plot would be more informative than a spaghetti plot, and other scenarios or research questions in which a spaghetti plot would be preferable to a lattice plot.

4. Walker-Barnes and Mason summarize their analytic approach in the following way:

The first series [of analyses] tested whether there was overall change and/or significant individual variability in gang [activity] over time, regardless of parenting behavior, peer behavior, or ethnic and cultural heritage. Second, given the well documented relation between peer and adolescent behavior . . . HLM analyses were conducted examining the effect of peer gang [activity] on [initial gang activity and] changes in gang [activity] over time. Finally, four pairs of analyses were conducted examining the role of each of the four parenting variables on [initial gang activity and] changes in gang [activity].

The last series of analyses controlled for peer gang activity and ethnic and cultural heritage, in addition to examining interactions between parenting and ethnic and cultural heritage.

Although the authors examined four parenting behaviors—behavioral control, lax control, psychological control, and parental warmth—they did so one at a time, using four separate multilevel models. Based on their description, write out a sample model from each step in the series. For each model, (a) write out the two-level model for predicting gang activity, (b) write out the corresponding composite model, and (c) determine how many model parameters (fixed effects and variance components) must be estimated.

5. The table below shows a portion of Table 2: Results of Hierarchical Linear Modeling Analyses Modeling Gang Involvement from Walker-Barnes and Mason (2001). Provide interpretations of significant coefficients in context.

6. Charter Schools. Differences exist in both sets of boxplots in Figure 9.12. What do these differences imply for multilevel modeling?

7. What implications do the scatterplots in Figure 9.14 (b) and (c) have for multilevel modeling? What implications does the boxplot in Figure 9.14(a) have?

8. What are the implications of Figure 9.15 for multilevel modeling?

9. Sketch a set of boxplots to indicate an obvious interaction between percent special education and percent non-white in modeling 2008 math scores. Where would this interaction appear in the multilevel model?
Table 9.4 A portion of Table 2: Results of Hierarchical Linear Modeling Analyses Modeling Gang Involvement from Walker-Barnes and Mason (2001). These columns focus on the parenting behavior of psychological control.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\gamma$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept (initial status)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base (intercept for predicting int term)</td>
<td>-.219</td>
<td>.160</td>
</tr>
<tr>
<td>Peer behavior</td>
<td>.252**</td>
<td>.026</td>
</tr>
<tr>
<td>Black ethnicity</td>
<td>.671*</td>
<td>.289</td>
</tr>
<tr>
<td>White/Other ethnicity</td>
<td>.149</td>
<td>.252</td>
</tr>
<tr>
<td>Parenting</td>
<td>.076</td>
<td>.050</td>
</tr>
<tr>
<td>Black ethnicity X Parenting</td>
<td>-.161+</td>
<td>.088</td>
</tr>
<tr>
<td>White/Other ethnicity X Parenting</td>
<td>-.026</td>
<td>.082</td>
</tr>
<tr>
<td><strong>Slope (change)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base (intercept for predicting slope term)</td>
<td>.028</td>
<td>.030</td>
</tr>
<tr>
<td>Peer behavior</td>
<td>-.011*</td>
<td>.005</td>
</tr>
<tr>
<td>Black ethnicity</td>
<td>-.132*</td>
<td>.054</td>
</tr>
<tr>
<td>White/Other ethnicity</td>
<td>-.059</td>
<td>.046</td>
</tr>
<tr>
<td>Parenting</td>
<td>-.015+</td>
<td>.009</td>
</tr>
<tr>
<td>Black ethnicity X Parenting</td>
<td>.048**</td>
<td>.017</td>
</tr>
<tr>
<td>White/Other ethnicity X Parenting</td>
<td>.016</td>
<td>.015</td>
</tr>
</tbody>
</table>

Note: Table reports values for $\gamma$s in the final model with all variables entered.
$p < .05; ** p < .01; + p < .10$

10. In Model A, $\sigma^2$ is defined as the variance in within-school deviations and $\sigma^2_0$ is defined as the variance in between-school deviations. Give potential sources of within-school and between-school deviations.

11. In Chapter 8 Model B is called the “random slopes and intercepts model”, while in this chapter Model B is called the “unconditional growth model”. Are these models essentially the same or systematically different? Explain.

12. In Section 9.5.2, why don’t we examine the psuedo $R^2$ value for Level Two?

13. If we have test score data from 2001-2010, explain how we’d create new variables to fit a piece-wise model.

14. In Section 9.6.2 could we have used percent free and reduced lunch as a Level One covariate rather than 2008 percent free and reduced lunch as a Level Two covariate? If so, explain how interpretations would have changed. What if we had used average percent free and reduced lunch over all three years or 2010 percent free and reduced lunch instead of 2008 percent free and reduced lunch - how would this have changed the interpretation of this term?

15. In Section 9.6.2, why do we look at a 10 percent increase in the percentage of students receiving free and reduced lunch when interpreting $\beta_{02}$?

16. In Section 9.6.3, if the gap in 2008 math scores between charter and non-charter schools differed for schools of different poverty levels (as measured by percent free and reduced lunch), how would the final model have differed?

17. Explain in your own words why “the error structure at Level Two is not the same as the within-school covariance structure among observations”.

18. Here is the estimated unstructured covariance matrix for Model C:

$$
\text{Cov}(\mathbf{Y}) = \begin{bmatrix}
41.87 & 36.46 & 48.18 \\
36.46 & 39.84 & 45.77 \\
48.18 & 39.84 & 45.77
\end{bmatrix}
$$

Explain why this matrix cannot represent an estimated covariance matrix with a compound symmetry, autoregressive, or Toeplitz structure. Also explain why it cannot represent our standard two-level model.
9.9.2 Guided Exercise

1. Teen Alcohol Use. Curran, Stice, and Chassin (Journal of Consulting and Clinical Psychology, 1997) collected data on 82 adolescents at three time points starting at age 14 to assess factors that affect teen drinking behavior. Key variables in the data set (source: Singer and Willett, 2003) are as follows:

- **id** = numerical identifier for subject
- **age** = 14, 15, or 16
- **coa** = 1 if the teen is a child of an alcoholic parent; 0 otherwise
- **male** = 1 if male; 0 if female
- **peer** = a measure of peer alcohol use, taken when each subject was 14. This is the square root of the sum of two 6-point items about the proportion of friends who drink occasionally or regularly.
- **alcuse** = the primary response. Four items—(a) drank beer or wine, (b) drank hard liquor, (c) 5 or more drinks in a row, and (d) got drunk—were each scored on an 8-point scale, from 0 = "not at all" to 7 = "every day". Then alcuse is the square root of the sum of these four items.

Primary research questions included:

- do trajectories of alcohol use differ by parental alcoholism?
- do trajectories of alcohol use differ by peer alcohol use?

(a) Identify Level One and Level Two predictors.

(b) Perform a quick EDA. What can you say about the shape of alcuse, and the relationship between alcuse and coa, male, and peer? Appeal to plots and summary statistics in making your statements.

(c) Generate a plot as in Figure 9.4 with alcohol use over time for all 82 subjects. Comment.

(d) Generate three spaghetti plots with loess fits similar to Figure 9.7 (one for coa, one for male, and one after creating a binary variable from peer). Comment on what you can conclude from each plot.

(e) Fit a linear trend to the data from each of the 82 subjects using age as the time variable. Generate histograms as in Figure 9.10 showing the results of these 82 linear regression lines, and generate pairs of boxplots as in Figure 9.10 for coa and male. No commentary necessary. [Hint: to produce Figure 9.10, you will need a data frame with one observation per subject.]

(f) Repeat (e) using centered age (age14=age-14) as the time variable. Also generate a pair of scatterplots as in Figure 9.14 for peer alcohol use. Comment on trends you observe in these plots. [Hint: after forming age14, append it to your current data frame.]

(g) Discuss similarities and differences between (e) and (f). Why does using age14 as the time variable make more sense in this example?

(h) (Model A) Run an unconditional means model. Report and interpret the intraclass correlation coefficient.

(i) (Model B) Run an unconditional growth model with age14 as the time variable at Level One. Report and interpret estimated fixed effects, using proper notation. Also report and interpret a pseudo-R-square value.

(j) (Model C) Build upon the unconditional growth model by adding the effects of having an alcoholic parent and peer alcohol use in both Level Two equations. Report and interpret all estimated fixed effects, using proper notation.

(k) (Model D) Remove the child of an alcoholic indicator variable as a predictor of slope in Model C (it will still be a predictor of intercept). Write out Model D as both a two-level and a composite model using proper notation (including error distributions); how many parameters (fixed effects and variance components) must be estimated? Compare Model D to Model C using an appropriate method and state a conclusion.
9.9.3 Open-ended Exercises

1. **UCLA Nurse Blood Pressure Study.** A study by Goldstein and Shapiro (2000) collected information from 203 registered nurses in the Los Angeles area between 24 and 50 years of age on blood pressure and potential factors that contribute to hypertension. This information includes family history, including whether the subject had one or two hypertensive parents, as well as a wide range of measures of the physical and emotional condition of each nurse throughout the day. Researchers sought to study the links between blood pressure and family history, personality, mood changes, working status, and menstrual phase.

Data from this study provided by Weiss (2005) includes observations (40-60 per nurse) repeatedly taken on the 203 nurses over the course of a single day. The first blood pressure measurement was taken half an hour before the subject’s normal start of work, and was measured approximately every 20 minutes for the rest of the day. At each blood pressure reading, the nurses also rate their mood on several dimensions, including how stressed they feel at the moment the blood pressure is taken. In addition, the activity of each subject during the 10 minutes before each reading was measured using an actigraph worn on the waist. Each of the variables is described below:

- **SNUM:** subject identification number
- **SYS:** systolic blood pressure (mmHg)
- **DIA:** diastolic blood pressure (mmHg)
- **HRT:** heart rate (beats per minute)
- **MNACT5:** activity level (frequency of movements in 1-minute intervals, over a 10-minute period)
- **PHASE:** menstrual phase (follicular—beginning with the end of menstruation and ending with ovulation, or luteal—beginning with ovulation and ending with pregnancy or menstruation)
- **DAY:** workday or non-workday
- **POSTURE:** position during blood pressure measurement—either sitting, standing, or reclining
- **STR, HAP, TIR:** self-ratings by each nurse of their level of stress, happiness and tiredness at the time of each blood pressure measurement on a 5-point scale, with 5 being the strongest sensation of that feeling and 1 the weakest
- **AGE:** age in years
- **FH123:** coded as either NO (no family history of hypertension), YES (1 hypertensive parent), or YESYES (both parents hypertensive)
- **time:** in minutes from midnight
- **timept:** number of the measurement that day (approximately 50 for each subject)
- **timepass:** time in minutes beginning with 0 at time point 1

Using systolic blood pressure as the primary response, write a short report detailing factors that are significantly associated with higher systolic blood pressure. Be sure to support your conclusions with appropriate exploratory plots and multilevel models. In particular, how are work conditions—activity level, mood, and work status—related to trends in blood pressure levels? As an appendix to your report, describe your modeling process—how did you arrive at your final model, which covariates are Level One or Level Two covariates, what did you learn from exploratory plots, etc.

Potential alternative directions: consider diastolic blood pressure or heart rate as the primary response variable, or even try modeling emotion rating using a multilevel model.
2. **Completion Rates at US Colleges.** Education researchers wonder which factors most affect the completion rates at US colleges. Using the IPEDS database containing data from 1310 institutions over the years 2002-2009, the following variables were assembled:

- **id** = unique identification number for each college or university
- **rate** = completion rate (number of degrees awarded per 100 students enrolled)

**Level 1 predictors:**

- **year**
- **instpct** = percentage of students who receive an institutional grant
- **instamt** = typical amount of an institutional grant among recipients (in $1000s)

**Level 2 predictors:**

- **faculty** = mean number of full-time faculty per 100 students in between 2002-2009
- **tuition** = mean yearly tuition in between 2002 and 2009 (in $1000s)

Perform exploratory analyses and run multilevel models to determine significant predictors of baseline (2002) completion rates and changes in completion rates between 2002 and 2009. In particular, is the percentage of grant recipients or the average institutional grant awarded related to completion rate?
Chapter 10

Multilevel Data With More Than Two Levels

10.1 Learning Objectives

After finishing this chapter, you should be able to:

- Extend the standard multilevel model to cases with more than two levels.
- Apply exploratory data analysis techniques specific to data from more than two levels.
- Formulate multilevel models including the variance-covariance structure.
- Build and understand a taxonomy of models for data with more than two levels.
- Interpret parameters in models with more than two levels.
- Develop strategies for handling an exploding number of parameters in multilevel models.
- Recognize when a fitted model has encountered boundary constraints and understand strategies for moving forward.
- Understand how a parametric bootstrap test of significance works and when it might be useful.

10.2 Case Studies: Seed Germination

It is estimated that 82-99% of historic tallgrass prairie ecosystems have been converted to agricultural use (Baer et al. 2002). A prime example of this large scale conversion of native prairie to agricultural purposes can be seen in Minnesota, where less than 1% of the prairies that once existed in the state still remain (Camill et al. 2004). Such large scale alteration of prairie communities has been associated with numerous problems. For example, erosion and decomposition that readily take place in cultivated soils have increased atmospheric CO2 levels and increased nitrogen inputs to adjacent waterways (Baer et al. 2002, Camill et al. 2004, Knops and Tilman 2000). In addition, cultivation practices are known to affect rhizosphere composition as tilling can disrupt networks of soil microbes (Allison et al. 2005). The rhizosphere is the narrow region of soil that is directly influenced by root secretions and associated soil microorganisms; much of the nutrient cycling and disease suppression needed by plants occur immediately adjacent to roots (http://en.wikipedia.org/wiki/Rhizosphere). It is important to note that microbial communities in prairie soils have been implicated with plant diversity and overall ecosystem function by controlling carbon and nitrogen cycling in the soils (Zak et al. 2003).

There have been many responses to these claims, but one response in recent years is reconstruction of the native prairie community. These reconstruction projects provide new habitat for a variety of native prairie species, yet it is important to know as much as possible about the outcomes of prairie reconstruction projects in order to ensure that a functioning prairie community is established. The ecological repercussions resulting from prairie reconstruction are not well known. For example, all of the aforementioned changes associated with cultivation practices are known to affect the subsequent reconstructed prairie community (Baer et al. 2002, Camill et al. 2004), yet there are few explanations for this phenomenon. For instance, prairies reconstructed in different years (using
Researchers at a small Midwestern college decided to experimentally explore the underlying causes of variation in reconstruction projects in order to make future projects more effective. Introductory ecology classes were organized to collect longitudinal data on native plant species grown in a greenhouse setting, using soil samples from surrounding lands. We will examine their data to compare germination and growth of two species of prairie plants—leadplants (Amorpha canescens) and coneflowers (Ratibida pinnata)—in soils taken from a remnant (natural) prairie, a cultivated (agricultural) field, and a restored (reconstructed) prairie. Additionally, half of the sampled soil was sterilized to determine if rhizosphere differences were responsible for the observed variation, so we will examine the effects of sterilization as well.

The data we’ll examine was collected through an experiment run using a 3x2x2 factorial design, with 3 levels of soil type (remnant, cultivated, and restored), 2 levels of sterilization (yes or no), and 2 levels of species (leadplant and coneflower). Each of the 12 treatments (unique combinations of factor levels) was replicated in 6 pots, for a total of 72 pots. Six seeds were planted in each pot (although a few pots had 7 or 8 seeds), and initially student researchers recorded days to germination (defined as when two leaves are visible), if germination occurred. In addition, the height of each germinated plant (in mm) was measured at 13, 18, 23, and 28 days after planting. The study design is illustrated in Figure 10.1.

![Figure 10.1: The design of the seed germination study.](image)

### Initial Exploratory Analyses

#### Data Organization

Data for Case Study 10.2 can be found under seeds2.csv in our data repository. Examination of the data reveals the following variables:

- **pot** = Pot plant was grown in (1-72)
- **plant** = Unique plant identification number
- **species** = L for leadplant and C for coneflower
Table 10.1: A snapshot of data (Plants 231-246) from the Seed Germination case study in wide format.

- soil = STP for reconstructed prairie, REM for remnant prairie, and CULT for cultivated land
- sterile = Y for yes and N for no
- germin = Y if plant germinated, N if not
- hgt13 = height of plant (in mm) 13 days after seeds planted
- hgt18 = height of plant (in mm) 18 days after seeds planted
- hgt23 = height of plant (in mm) 23 days after seeds planted
- hgt28 = height of plant (in mm) 28 days after seeds planted

This data is stored in wide format, with one row per plant (see 12 sample plants in Table 10.1). As we have done in previous multilevel analyses, we will convert to long format (one observation per plant-time combination) after examining the missing data pattern and removing any plants with no growth data. In this case, we are almost assuredly losing information by removing plants with no height data at all four time points, since these plants did not germinate, and there may well be differences between species, soil type, and sterilization with respect to germination rates. We will handle this possibility by analyzing germination rates separately (see Chapter ??); the analysis in this chapter will focus on effects of species, soil type, and sterilization on initial growth and growth rate among plants that germinate.

Although the experimental design called for 72 x 6 = 432 plants, the wide data set has 437 plants because a few pots had more than six plants (likely because two of the microscopically small seeds stuck together when planted). Of those 437 plants, 154 had no height data (did not germinate by the 28th day) and were removed from analysis (for example, see rows 145-146 in Table 10.1). 248 plants had complete height data (e.g., rows 135-139 and 141-143), 13 germinated later than the 13th day but had complete heights once they germinated (e.g., row 144), and 22 germinated and had measurable height on the 13th day but died before the 28th day (e.g., row 140). Ultimately, the long data set contains 1132 unique observations where plants heights were recorded; representation of plants 236-242 in the long data set can be seen in Table 10.2.

Notice the three-level structure of this data. Treatments (levels of the three experimental factors) were assigned at the pot level, then multiple plants were grown in each pot, and multiple measurements were taken over time for each plant. Our multilevel analysis must therefore account for pot-to-pot variability in height measurements (which could result from factor effects), plant-to-plant variability in height within a single pot, and variability over time in height for individual plants. In order to fit such a three-level model, we must extend the two-level model which we have
Table 10.2: A snapshot of data (Plants 236-242) from the Seed Germination case study in long format.

<table>
<thead>
<tr>
<th>pot</th>
<th>plant</th>
<th>soil</th>
<th>sterile</th>
<th>species</th>
<th>germin</th>
<th>time</th>
<th>hgt</th>
<th>time13</th>
</tr>
</thead>
<tbody>
<tr>
<td>236.1</td>
<td>23</td>
<td>236</td>
<td>CULT</td>
<td>N</td>
<td>C</td>
<td>Y</td>
<td>13</td>
<td>0.1</td>
</tr>
<tr>
<td>236.2</td>
<td>23</td>
<td>236</td>
<td>CULT</td>
<td>N</td>
<td>C</td>
<td>Y</td>
<td>18</td>
<td>NA</td>
</tr>
<tr>
<td>236.3</td>
<td>23</td>
<td>236</td>
<td>CULT</td>
<td>N</td>
<td>C</td>
<td>Y</td>
<td>23</td>
<td>NA</td>
</tr>
<tr>
<td>236.4</td>
<td>23</td>
<td>236</td>
<td>CULT</td>
<td>N</td>
<td>C</td>
<td>Y</td>
<td>28</td>
<td>NA</td>
</tr>
<tr>
<td>241.1</td>
<td>24</td>
<td>241</td>
<td>STP</td>
<td>Y</td>
<td>L</td>
<td>Y</td>
<td>13</td>
<td>1.8</td>
</tr>
<tr>
<td>241.2</td>
<td>24</td>
<td>241</td>
<td>STP</td>
<td>Y</td>
<td>L</td>
<td>Y</td>
<td>18</td>
<td>2.6</td>
</tr>
<tr>
<td>241.3</td>
<td>24</td>
<td>241</td>
<td>STP</td>
<td>Y</td>
<td>L</td>
<td>Y</td>
<td>23</td>
<td>3.9</td>
</tr>
<tr>
<td>241.4</td>
<td>24</td>
<td>241</td>
<td>STP</td>
<td>Y</td>
<td>L</td>
<td>Y</td>
<td>28</td>
<td>4.2</td>
</tr>
<tr>
<td>242.1</td>
<td>24</td>
<td>242</td>
<td>STP</td>
<td>Y</td>
<td>L</td>
<td>Y</td>
<td>13</td>
<td>1.3</td>
</tr>
<tr>
<td>242.2</td>
<td>24</td>
<td>242</td>
<td>STP</td>
<td>Y</td>
<td>L</td>
<td>Y</td>
<td>18</td>
<td>1.7</td>
</tr>
<tr>
<td>242.3</td>
<td>24</td>
<td>242</td>
<td>STP</td>
<td>Y</td>
<td>L</td>
<td>Y</td>
<td>23</td>
<td>2.8</td>
</tr>
<tr>
<td>242.4</td>
<td>24</td>
<td>242</td>
<td>STP</td>
<td>Y</td>
<td>L</td>
<td>Y</td>
<td>28</td>
<td>3.7</td>
</tr>
</tbody>
</table>

10.3.2 Exploratory Analyses

We start by taking an initial look at the effect of Level Three covariates (factors applied at the pot level: species, soil type, and sterilization) on plant height, pooling observations across pot, across plant, and across time of measurement within plant. First, we observe that the initial balance which existed after randomization of pot to treatment no longer holds. After removing plants that did not germinate (and therefore had no height data), more height measurements exist for coneflowers (n=704, compared to 428 for leadplants), soil from restored prairies (n=524, compared to 288 for cultivated land and 320 for remnant prairies), and unsterilized soil (n=612, compared to 520 for sterilized soil). This imbalance indicates possible factor effects on germination rate; we will take up those hypotheses in Chapter ???. In this chapter, we will focus on the effects of species, soil type, and sterilization on the growth patterns of plants that germinate.

Because we suspect that height measurements over time for a single plant are highly correlated, while height measurements from different plants from the same pot are relatively uncorrelated, we calculate mean height per plant (over all available time points) before generating exploratory plots investigating Level Three factors. Figure 10.2 then examines the effects of soil type and sterilization separately by species. Sterilization seems to have a bigger benefit for coneflowers, while soil from remnant prairies seems to lead to smaller leadplants and taller coneflowers.

We also use spaghetti plots to examine time trends within species to see (a) if it is reasonable to assume linear growth between Day 13 and Day 28 after planting, and (b) if initial height and rate of growth is similar in the two species. Figure 10.3 illustrates differences between species. While both species have similar average heights 13 days after planting, coneflowers appear to have faster early growth which slows later, while leadplants have a more linear growth rate which culminates in greater average heights 28 days after planting. Coneflowers also appear to have greater variability in initial height and growth rate, although there are more coneflowers with height data.

Exploratory analyses such as these confirm the suspicions of biology researchers that leadplants and coneflowers should be analyzed separately. Because of biological differences, it is expected that these two species will show different growth patterns and respond differently to treatments such as fertilization. Coneflowers are members of the aster family, growing up to 4 feet tall with their dis-
Figure 10.2. Plant height comparisons of (a) soil type and (b) sterilization within species. Each plant is represented by the mean height over all measurements at all time points for that plant.

Figure 10.3: Spaghetti plot by species with loess fit. Each “line” represents one plant.

tinctive gray seed heads and drooping yellow petals. Leadplants, on the other hand, are members of the bean family, with purple flowers, a height of 1 to 3 feet, and compound greyish green leaves which look to be dusted with white lead. Leadplants have deep root systems and are symbiotic N-fixers, which means they might experience stifled growth in sterilized soil compared with other species. For the remainder of this chapter, we will focus on leadplants and how their growth patterns are affected by soil type and sterilization. You will have a chance to analyze coneflower data later in the Exercises section.

Lattice plots, illustrating several observational units simultaneously, each with fitted lines where appropriate, are also valuable to examine during the exploratory analysis phase. Figure 10.4 shows height over time for 24 randomly selected leadplants that germinated in this study, with a fitted
linear regression line. Linearity appears reasonable in most cases, although there some variability in the intercepts and a good deal of variability in the slopes of the fitted lines. These intercepts and slopes by plant, of course, will be potential parameters in a multilevel model which we will fit to this data. Given the three-level nature of this data, it is also useful to examine a spaghetti plot by pot (Figure 10.5). While linearity appears to reasonably model the average trend over time within pot, we see differences in the plant-to-plant variability within pot, but some consistency in intercept and slope from pot to pot.

Figure 10.4 Lattice plot of linear trends fit to 24 randomly selected leadplants. One plant with only a single height measurement has no associated regression line.

Figure 10.5 Spaghetti plot for leadplants by pot with loess fit. Four pots only had a single plant that germinated.

Spaghetti plots can also be an effective tool for examining the potential effects of soil type and sterilization of growth patterns of leadplants. Figure 10.6 and Figure 10.7 illustrate how the growth
INITIAL EXPLORATORY ANALYSES

patterns of leadplants depend on soil type and sterilization. In general, we observe slower growth in soil from remnant prairies and soil that has not been sterilized.

We can further explore the variability in linear growth among plants and among pots by fitting regression lines and examining the estimated intercepts and slopes, as well as the corresponding R-squared values. Figures 10.8 and 10.9 provide just such an analysis, where Figure 10.8 shows results of fitting lines by plant, and Figure 10.9 shows results of fitting lines by pot. Certain caveats accompany these summaries. In the case of fitted lines by plant, each plant is given equal weight regardless of the number of observations (2-4) for a given plant, and in the case of fitted lines by pot, a line is estimated by simply pooling all observations from a given pot, ignoring the plant from which the observations came, and equally weighting pots regardless of how many plants germinated and
survived to Day 28. Nevertheless, the summaries of fitted lines provide useful information. When fitting regression lines by plant, we see a mean intercept of 1.52 (SD=0.66), indicating an estimated average height at 13 days of 1.5 mm, and a mean slope of 0.114 mm per day of growth from Days 13 to 28 (SD=0.059). Most R-squared values were strong (e.g., 84% were above 0.8). Summaries of fitted regression lines by pot show similar mean intercepts (1.50) and slopes (0.107), but somewhat less variability from pot-to-pot than we observed from plant-to-plant (SD=0.46 for intercepts and SD=0.050 for slopes).

![Figure 10.8](image1)

Figure 10.8 Histograms of (a) intercepts, (b) slopes, and (c) R-squared values for linear fits across all lead-plants.

![Figure 10.9](image2)

Figure 10.9 Histograms of (a) intercepts, (b) slopes, and (c) R-squared values for linear fits across all pots with leadplants.

Another way to examine variability due to plant vs. variability due to pot is through summary statistics. Plant-to-plant variability can be estimated by averaging standard deviations from each pot
(.489 for intercepts and .039 for slopes), while pot-to-pot variability can be estimated by finding the standard deviation of average intercept (.478) or slope (.051) within pot. Based on these rough measurements, variability due to plants and pots is comparable.

Fitted lines by plant and pot are modeled using a centered time variable (time13), adjusted so that the first day of height measurements (13 days after planting) corresponds to time13=0. This centering has two primary advantages. First, the estimated intercept becomes more interpretable. Rather than representing height on the day of planting (which should be 0 mm, but which represents a hefty extrapolation from our observed range of days 13 to 28), the intercept now represents height on Day 13. Second, the intercept and slope are much less correlated \(r=-0.16\) than when uncentered time is used, which improves the stability of future models.

Fitted intercepts and slopes by plant can be used for an additional exploratory examination of factor effects to complement those from the earlier spaghetti plots. Figure 10.10 complements Figure 10.3, again showing differences between species – coneflowers tend to start smaller and have slower growth rates, although they have much more variability in growth patterns than leadplants. Returning to our focus on leadplants, Figure 10.11 shows that plants grown in soil from cultivated fields tends to be taller at Day 13, and plants grown in soil from remnant prairies tend to grow more slowly than plants grown in other soil types. Figure 10.12 shows the strong tendency for plants grown in sterilized soil to grow faster than plants grown in non-sterilized soil. We will soon see if our fitted multilevel models support these observed trends.

![Boxplots of intercepts and slopes for all plants by species, based on a linear fit to height data from each plant.](image)

Since we have time at Level One, any exploratory analysis of Case Study 10.2 should contain an investigation of the variance-covariance structure within plant. Figure 10.13 shows the potential for an autocorrelation structure in which the correlation between observations from the same plant diminishes as the time between measurements increases. Residuals five days apart have correlations ranging from .79 to .90, while measurements ten days apart have correlations of .66 and .67, and measurements fifteen days apart have correlation of .58.
10.4 Initial models: unconditional means and unconditional growth

The structure and notation for three-level models will closely resemble the structure and notation for two-level models, just with extra subscripts. Therein lies some of the power of multilevel models – extensions are relatively easy and allow the analyst to control for many sources of variability, obtaining more precise estimates of important parameters. However, the number of variance component parameters to estimate can quickly mushroom as covariates are added at lower levels, so implementing simplifying restrictions will often become necessary (see Section 10.7).

We once again begin with the unconditional means model, in which there are no predictors at any level, in order to assess the amount of variation at each level. Here, Level Three is pot, Level
Figure 10.13 Correlation structure within plant. Histograms labeled “lmres” contain residuals at each time point after fitting a linear trend to each plant’s data over time; the lower triangular region contains scatterplots examining residuals from different time points; and, the upper triangular region contains the corresponding correlation coefficients for residuals at different time points.

Two is plant within pot, and Level One is time within plant. Using model formulations at each of the three levels, the unconditional means three-level model can be expressed as:

- **Level One (timepoint within plant):**
  \[ Y_{ijk} = \lambda_{0ij} + \varepsilon_{ijk} \text{ where } \varepsilon_{ijk} \sim N(0, \sigma^2) \]  \[ (10.1) \]

- **Level Two (plant within pot):**
  \[ \lambda_{0ij} = \lambda_{00i} + \varepsilon_{0ij} \text{ where } \varepsilon_{0ij} \sim N(0, \sigma_0^2) \]  \[ (10.2) \]

- **Level Three (pot):**
  \[ \lambda_{00i} = \beta_{000} + \varepsilon_{00i} \text{ where } \varepsilon_{00i} \sim N(0, \sigma_{00}^2) \]  \[ (10.3) \]

where the heights of plants from different pots are considered independent, but plants from the same pot are correlated as well as measurements at different times from the same plant.

Keeping track of all the model terms, especially with three subscripts, is not a trivial task, but it’s worth spending time thinking through. Here is a quick guide to the meaning of terms found in our three-level model:

- **\( Y_{ijk} \)** is the height (in mm) of plant \( j \) from pot \( i \) at time \( k \)
- **\( \lambda_{0ij} \)** is the true mean height for plant \( j \) from pot \( i \) across all time points. This is not considered a model parameter, since we further model \( \lambda_{0ij} \) at Level Two.
- **\( \lambda_{00i} \)** is the true mean height for pot \( i \) across all plants from that pot and all time points. This is also not considered a model parameter, since we further model \( \lambda_{00i} \) at Level Three.
- **\( \beta_{000} \)** is a fixed effects model parameter representing the true mean height across all pots, plants, and time points
- **\( \varepsilon_{ijk} \)** describes how far an observed height \( Y_{ijk} \) is from the mean height for plant \( j \) from pot \( i \)
- **\( \varepsilon_{0ij} \)** describe how far the mean height of plant \( j \) from pot \( i \) is from the mean height of all plants from pot \( i \)
The three-level unconditional means model can also be expressed as a composite model:

\[ Y_{ijk} = \beta_{000} + \varepsilon_{00i} + \varepsilon_{0ij} + \varepsilon_{ijk} \]  

(10.4)

This composite model can be fit using statistical software:

```
# Model A - Unconditional means #
Formula: hgt ~ 1 + (1 | plant) + (1 | pot)
```

AIC  BIC  logLik deviance  REMLdev
1157 1173  -574.4     1146     1149

Random effects:
Groups     Name   Variance  Std.Dev.
plant     (Intercept) 0.278169 0.52742
pot       (Intercept) 0.048728 0.22074
Residual          0.727823 0.85313

Number of obs: 413, groups: plant, 107; pot, 32

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.38808</td>
<td>0.07886</td>
</tr>
</tbody>
</table>

From this output, we obtain estimates of our four model parameters:
- \( \hat{\beta}_{000} = 2.39 \) = the mean height (in mm) across all time points, plants, and pots.
- \( \hat{\sigma}^2 = 0.728 \) = the variance over time within plants.
- \( \hat{\sigma}_{0}^2 = 0.278 \) = the variance between plants from the same pot.
- \( \hat{\sigma}_{00}^2 = 0.049 \) = the variance between pots.

From the estimates of variance components, 69.0% of total variability in height measurements is due to differences over time for each plant, 26.4% of total variability is due to differences between plants from the same pot, and only 4.6% of total variability is due to difference between pots. Accordingly, we will next explore whether the incorporation of time as a linear predictor at Level One can reduce the unexplained variability within plant.

The unconditional growth model introduces time as a predictor at Level One, but there are still no predictors at Levels Two or Three. The unconditional growth model allows us to assess how much of the within-plant variability (the variability among height measurements from the same plant at different time points) can be attributed to linear changes over time, while also determining how much variability we see in the intercept (Day 13 height) and slope (daily growth rate) from plant-to-plant and pot-to-pot. Later, we can model plant-to-plant and pot-to-pot differences in intercepts and slopes with Level Two and Three covariates.

The three-level unconditional growth model (Model B) can be specified either using formulations at each level:
INITIAL MODELS: UNCONDITIONAL MEANS AND UNCONDITIONAL GROWTH

- Level One (timepoint within plant):
  \[ Y_{ijk} = \lambda_{0ij} + \lambda_{1ij} \text{time}_{ijk} + \epsilon_{ijk} \tag{10.5} \]

- Level Two (plant within pot):
  \[
  \begin{align*}
  \lambda_{0ij} &= \lambda_{00i} + \epsilon_{0ij} \\
  \lambda_{1ij} &= \lambda_{10i} + \epsilon_{1ij}
  \end{align*}
  \]

- Level Three (pot):
  \[
  \begin{align*}
  \lambda_{00i} &= \beta_{000} + \epsilon_{00i} \\
  \lambda_{10i} &= \beta_{100} + \epsilon_{10i}
  \end{align*}
  \]

or as a composite model:
\[
Y_{ijk} = [\beta_{000} + \beta_{100} \text{time}_{ijk}] + [\epsilon_{00i} + \epsilon_{0ij} + \epsilon_{ijk} + (\epsilon_{10i} + \epsilon_{1ij}) \text{time}_{ijk}] \tag{10.6}
\]

where \( \epsilon_{ijk} \sim N(0, \sigma^2) \),
\[
\begin{bmatrix}
  \epsilon_{0ij} \\
  \epsilon_{1ij}
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix} \right),
\]
and
\[
\begin{bmatrix}
  \epsilon_{00i} \\
  \epsilon_{10i}
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{00}^2 & \sigma_{0011} \\ \sigma_{0011} & \sigma_{11}^2 \end{bmatrix} \right).
\]

In this model, at Level One the trajectory for plant \( j \) from pot \( i \) is assumed to be linear, with intercept \( \lambda_{0ij} \) (height on Day 13) and slope \( \lambda_{1ij} \) (daily growth rate between Days 13 and 28); the \( \epsilon_{ijk} \) terms capture the deviation between the true growth trajectory of plant \( j \) from pot \( i \) and its observed heights. At Level Two, \( \lambda_{00i} \) represents the true mean intercept and \( \lambda_{10i} \) represents the true mean slope for all plants from pot \( i \), while \( \epsilon_{0ij} \) and \( \epsilon_{1ij} \) capture the deviation between plant \( j \)'s true growth trajectory and the mean intercept and slope for pot \( i \). The deviations in intercept and slope at Level Two are allowed to be correlated through the covariance parameter \( \sigma_{01} \). Finally, \( \beta_{000} \) is the true mean intercept and \( \beta_{100} \) is the true mean daily growth rate over the entire population of lead-plants, while \( \epsilon_{00i} \) and \( \epsilon_{10i} \) capture the deviation between pot \( i \)'s true overall growth trajectory and the population mean intercept and slope. Note that between-plant and between-pot variability are both partitioned now into variability in initial status (\( \sigma_0^2 \) and \( \sigma_{00}^2 \)) and variability in rates of change (\( \sigma_1^2 \) and \( \sigma_{11}^2 \)).

Using the composite model specification, the unconditional growth model can be fit to the seed germination data:

# Model B - unconditional growth#

Formula: ght ~ time13 + (time13 | plant) + (time13 | pot)

<table>
<thead>
<tr>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>REMLdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>615.2</td>
<td>651.4</td>
<td>-298.6</td>
<td>585.8</td>
<td>597.2</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std. Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>(Intercept)</td>
<td>0.2991646</td>
<td>0.546959</td>
<td></td>
</tr>
<tr>
<td></td>
<td>time13</td>
<td>0.0011942</td>
<td>0.034557</td>
<td>0.280</td>
</tr>
<tr>
<td>pot</td>
<td>(Intercept)</td>
<td>0.0442185</td>
<td>0.210282</td>
<td></td>
</tr>
<tr>
<td></td>
<td>time13</td>
<td>0.0012609</td>
<td>0.035508</td>
<td>-0.610</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>0.0821572</td>
<td>0.286631</td>
<td></td>
</tr>
</tbody>
</table>
From this output, we obtain estimates of our nine model parameters (two fixed effects and seven variance components):

- $\hat{b}_{000} = 1.538$ is the mean height of leadplants 13 days after planting.
- $\hat{b}_{100} = 0.112$ is the mean daily change in height of leadplants from 13 to 28 days after planting.
- $\hat{\sigma} = .287$ is the standard deviation in within-plant residuals after accounting for time.
- $\hat{\sigma}_0 = .547$ is the standard deviation in Day 13 heights between plants from the same pot.
- $\hat{\sigma}_1 = .0346$ is the standard deviation in rates of change in height between plants from the same pot.
- $\hat{\rho}_{01} = .280$ is the correlation in plants’ Day 13 height and their rate of change in height.
- $\hat{\sigma}_{00} = .210$ is the standard deviation in Day 13 heights between pots.
- $\hat{\sigma}_{11} = .0355$ is the standard deviation in rates of change in height between pots.
- $\hat{\rho}_{0011} = -.610$ is the correlation in pots’ Day 13 height and their rate of change in height.

We see that, on average, leadplants have a height of 1.54 mm thirteen days after planting (pooled across pots and treatment groups), and their heights tend to grow by 0.11 mm per day, producing an average height at the end of the study (Day 28) of 3.22 mm. According to the t-values listed in R, both the Day 13 height and the growth rate are statistically significant. The estimated within-plant variance $\hat{\sigma}^2$ decreased by 88.7% from the unconditional means model (from 0.728 to 0.082), implying that 88.7% of within-plant variability in height can be explained by linear growth over time.

### 10.5 Encountering boundary constraints

Typically, with models consisting of three or more levels, the next step after adding covariates at Level One (such as time) is considering covariates at Level Two. In the seed germination experiment, however, there are no Level Two covariates of interest, and the treatments being studied were applied to pots (Level Three). We are primarily interested in the effects of soil type and sterilization on the growth of leadplants. Since soil type is a categorical factor with three levels, we can represent soil type in our model with indicator variables for cultivated lands ($cult$) and remnant prairies ($rem$), using reconstructed prairies as the reference level. For sterilization, we create a single indicator variable ($strl$) which takes on the value 1 for sterilized soil.

Our Level One and Level Two models will look identical to those from Model B; our Level Three models will contain the new covariates for soil type ($cult$ and $rem$) and sterilization ($strl$):

$$
\lambda_{0i} = \beta_{000} + \beta_{001}strl_i + \beta_{002}cult_i + \beta_{003}rem_i + \epsilon_{00i}
$$

$$
\lambda_{10i} = \beta_{100} + \beta_{101}strl_i + \beta_{102}cult_i + \beta_{103}rem_i + \epsilon_{10i}
$$

where the error terms at Level Three follow the same multivariate normal distribution as in Model B. In our case, the composite model can be written as:

$$
Y_{ijk} = (\beta_{000} + \beta_{001}strl_i + \beta_{002}cult_i + \beta_{003}rem_i + \epsilon_{00i} + \epsilon_{0ij}) +
(\beta_{100} + \beta_{101}strl_i + \beta_{102}cult_i + \beta_{103}rem_i + \epsilon_{10i} + \epsilon_{1ij})time_{ijk} + \epsilon_{ijk}
$$
ENCOUNTERING BOUNDARY CONSTRAINTS

which, after combining fixed effects and random effects, can be rewritten as:

\[ Y_{ijk} = \left[ \beta_{000} + \beta_{001} \text{strl}_i + \beta_{002} \text{cult}_i + \beta_{003} \text{rem}_i + \beta_{100} \text{time}_{ijk} + \right. \\
\left. \beta_{101} \text{strl}_i \text{time}_{ijk} + \beta_{102} \text{cult}_i \text{time}_{ijk} + \beta_{103} \text{rem}_i \text{time}_{ijk} \right] + \\
\left. \varepsilon_{00} + \varepsilon_{0i} + \varepsilon_{ij} + \varepsilon_{ijk} + \varepsilon_{100} \text{time}_{ijk} + \epsilon_{ij} \text{time}_{ijk} \right] \\

From the output below, the addition of Level Three covariates in Model C (\text{cult}, \text{rem}, \text{strl}, and their interactions with \text{time}) appears to provide a significant improvement (likelihood ratio test statistic = 32.2 on 6 df, \( p < .001 \)) to the unconditional growth model (Model B).

# Model C #

Formula: hgt \sim \text{time13} + \text{strl} + \text{cult} + \text{rem} + \text{time13:strl} + \text{time13:cult} + \\
\text{time13:rem} + (\text{time13} | \text{plant}) + (\text{time13} | \text{pot})

AIC BIC logLik deviance REMLdev
623.2 683.6 -296.6 553.8 593.2

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>(Intercept)</td>
<td>0.29802073</td>
<td>0.545913</td>
<td></td>
</tr>
<tr>
<td></td>
<td>time13</td>
<td>0.00120813</td>
<td>0.034758</td>
<td>0.284</td>
</tr>
<tr>
<td>pot</td>
<td>(Intercept)</td>
<td>0.05315051</td>
<td>0.230544</td>
<td></td>
</tr>
<tr>
<td></td>
<td>time13</td>
<td>0.00013174</td>
<td>0.011478</td>
<td>-1.000</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>0.08207687</td>
<td>0.286491</td>
<td></td>
</tr>
</tbody>
</table>

Number of obs: 413, groups: plant, 107; pot, 32

Fixed effects:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.502891</td>
<td>0.127001</td>
<td>11.834</td>
</tr>
<tr>
<td>time13</td>
<td>0.101069</td>
<td>0.008292</td>
<td>12.188</td>
</tr>
<tr>
<td>strl</td>
<td>-0.076544</td>
<td>0.151368</td>
<td>-0.506</td>
</tr>
<tr>
<td>cult</td>
<td>0.130013</td>
<td>0.182723</td>
<td>0.712</td>
</tr>
<tr>
<td>rem</td>
<td>0.138392</td>
<td>0.176196</td>
<td>0.785</td>
</tr>
<tr>
<td>time13:strl</td>
<td>0.058917</td>
<td>0.010283</td>
<td>5.730</td>
</tr>
<tr>
<td>time13:cult</td>
<td>-0.029765</td>
<td>0.012263</td>
<td>-2.427</td>
</tr>
<tr>
<td>time13:rem</td>
<td>-0.035860</td>
<td>0.011978</td>
<td>-2.994</td>
</tr>
</tbody>
</table>

> Df AIC  BIC  logLik  Chisq  Chi Df Pr(>Chisq)
modelB 9 603.76 639.97 -292.88
modelC 15 583.55 643.90 -276.78 32.202 6 1.493e-05 ***

However, Model C has encountered a boundary constraint with an estimated Level 3 correlation between the intercept and slope error terms of -1. “Allowable” values of correlation coefficients run from -1 to 1; by definition, it is impossible to have a correlation between two error terms below -1. Thus, our estimate of -1 is right on the boundary of the allowable values. But how did this happen, and why is it potentially problematic?

Consider a model in which we have two parameters that must be estimated: \( \beta_0 \) and \( \sigma^2 \). As the intercept, \( \beta_0 \) can take on any value; any real number is “allowable”. But, by definition, variance terms such as \( \sigma^2 \) must be non-negative; that is, \( \sigma^2 \geq 0 \). Under the Principle of Maximum Likelihood, maximum likelihood estimators for \( \beta_0 \) and \( \sigma^2 \) will be chosen to maximize the likelihood of observing our given data. The left plot in Figure 10.14 shows hypothetical contours of the
likelihood function \( L(\beta_0, \sigma^2) \); the likelihood is clearly maximized at \((\hat{\beta}_0, \hat{\sigma}^2) = (4, -2)\). However, variance terms cannot be negative! A more sensible approach would have been to perform a constrained search for MLEs, considering any potential values for \(\beta_0\) but only non-negative values for \(\sigma^2\). This constrained search is illustrated in the right plot in Figure 10.14. In this case, the likelihood is maximized at \((\hat{\beta}_0, \hat{\sigma}^2) = (4, 0)\). Note that the estimated intercept did not change, but the estimated variance is simply set at the smallest allowable value – at the boundary constraint.

Graphically, in this simple illustration, the effect of the boundary constraint is to alter the likelihood function from a nice hill (in the left plot in Figure 10.14) with a single peak at \((4, -2)\), to a hill with a huge cliff face where \(\sigma^2 = 0\). The highest point overlooking this cliff is at \((4, 0)\), straight down the hill from the original peak.

In general, then, boundary constraints occur when the maximum likelihood estimator of at least one model parameter occurs at the limits of allowable values (such as estimated correlation coefficients of -1 or 1, or estimated variances of 0). Maximum likelihood estimates at the boundary tend to indicate that the likelihood function would be maximized at non-allowable values of that parameter, if an unconstrained search for MLEs was conducted. Most software packages, however, will only report maximum likelihood estimates with allowable values. Therefore, boundary constraints would ideally be avoided, if possible.

What should you do if you encounter boundary constraints? Often, boundary constraints signal that your model needs to be reparameterized – i.e., you should alter your model to feature different parameters or ones that are interpreted differently. This can be accomplished in several ways:
- remove parameters, especially those variance and correlation terms which are being estimated on their boundaries
- fix the values of certain parameters; for instance, you could set two variance terms equal to each other, thereby reducing the number of unknown parameters to estimate by one
- transform covariates. Centering variables, standardizing variables, or changing units can all help
stabilize a model. Numerical procedures for searching for and finding maximum likelihood estimates can encounter difficulties when variables have very high or low values, extreme ranges, outliers, or are highly correlated.

Although it is worthwhile attempting to reparameterize models to remove boundary constraints, sometimes they can be tolerated if (a) you are not interested in estimates of those parameters encountering boundary issues, and (b) removing those parameters does not affect conclusions about parameters of interest. For example, in the output below we explore the implications of simply removing the correlation between error terms at the pot level (i.e., assume $\rho_{0011} = 0$ rather than accepting the (constrained) maximum likelihood estimate of $\hat{\rho}_{0011} = -1$ that we saw in Model C).

# Model C without the correlation between Level 3 error terms #

Formula: hgt ~ time13 + strl + cult + rem + time13:strl + time13:cult + time13:rem + (time13 | plant) + (1 | pot) + (0 + time13 | pot)

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>REMLdev</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>623.2</td>
<td>679.5</td>
<td>-297.6</td>
<td>556.2</td>
<td>595.2</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>plant</td>
<td>(Intercept)</td>
<td>0.29410146</td>
<td>0.542311</td>
<td></td>
</tr>
<tr>
<td></td>
<td>time13</td>
<td>0.00120624</td>
<td>0.034731</td>
<td>0.220</td>
</tr>
<tr>
<td>pot</td>
<td>time13</td>
<td>0.00013751</td>
<td>0.011726</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Intercept)</td>
<td>0.05916199</td>
<td>0.243232</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>0.08217416</td>
<td>0.286660</td>
<td></td>
</tr>
</tbody>
</table>

Number of obs: 413, groups: plant, 107; pot, 32

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.51208</td>
<td>0.12964</td>
</tr>
<tr>
<td>time13</td>
<td>0.10158</td>
<td>0.00836</td>
</tr>
<tr>
<td>strl</td>
<td>-0.08743</td>
<td>0.15406</td>
</tr>
<tr>
<td>cult</td>
<td>0.13236</td>
<td>0.18562</td>
</tr>
<tr>
<td>rem</td>
<td>0.10662</td>
<td>0.17930</td>
</tr>
<tr>
<td>time13:strl</td>
<td>0.05869</td>
<td>0.01038</td>
</tr>
<tr>
<td>time13:cult</td>
<td>-0.03065</td>
<td>0.01234</td>
</tr>
<tr>
<td>time13:rem</td>
<td>-0.03810</td>
<td>0.01209</td>
</tr>
</tbody>
</table>

Note that the estimated variance components are all very similar to Model C, and the estimated fixed effects and their associated t-statistics are also very similar to Model C. Therefore, in this case we could consider simply reporting the results of Model C despite the boundary constraint.

However, when removing boundary constraints through reasonable model reparameterizations is possible, that is typically the preferred route. In this case, one option we might consider is simplifying Model C by setting $\sigma_{11}^2 = \sigma_{0011} = 0$. We can then write our new model (Model C.1) in level-by-level formulation:

- **Level One (timepoint within plant):**
  \[
  Y_{ijk} = \lambda_{0ij} + \lambda_{1ij}time_{ijk} + \epsilon_{ijk}
  \]

- **Level Two (plant within pot):**
  \[
  \lambda_{0ij} = \lambda_{0i0} + \epsilon_{0ij} \\
  \lambda_{1ij} = \lambda_{1i0} + \epsilon_{1ij}
  \]
MULTILEVEL DATA WITH MORE THAN TWO LEVELS

- Level Three (pot):

\[
\lambda_{00i} = \beta_{000} + \beta_{001}strl_i + \beta_{002}cult_i + \beta_{003}rem_i + \epsilon_{00i}
\]

\[
\lambda_{10i} = \beta_{100} + \beta_{101}strl_i + \beta_{102}cult_i + \beta_{103}rem_i
\]

Note that there is no longer an error term associated with the model for mean growth rate \(\lambda_{10i}\) at the pot level. The growth rate for pot \(i\) is assumed to be fixed, after accounting for soil type and sterilization; all pots with the same soil type and sterilization are assumed to have the same growth rate. As a result, our error assumption at Level Three is no longer bivariate normal, but rather univariate normal: \(\epsilon_{00i} \sim N(0, \sigma^2_{00})\). By removing one of our two Level Three error terms (\(\epsilon_{10i}\)), we effectively removed two parameters – the variance for \(\epsilon_{10i}\) and the correlation between \(\epsilon_{00i}\) and \(\epsilon_{10i}\).

Fixed effects remain similar, as can be seen in the output below:

```
# Model C.1 #
Formula: hgt ~ time13 + strl + cult + rem + time13:strl + time13:cult +
         time13:rem + (time13 | plant) + (1 | pot)

AIC BIC logLik deviance REMLdev
621.7 674 -297.8 556 595.7

Random effects:
Groups     Name    Variance Std.Dev.    Corr
plant     (Intercept) 0.2947178 0.542879
          time13  0.0013266 0.036423  0.194
pot       (Intercept) 0.0576535 0.240111
          Residual  0.0822210 0.286742

Number of obs: 413, groups: plant, 107; pot, 32

Fixed effects:
                     Estimate Std. Error t value
(Intercept)       1.512224   0.128998  11.723
          time13  0.101091   0.007453  13.565
            strl -0.087519   0.153433  -0.570
            cult  0.132880   0.184877   0.719
            rem  0.106543   0.178592   0.597
     time13:strl  0.059264   0.009492   6.244
     time13:cult -0.030824   0.011353  -2.715
     time13:rem  -0.036244   0.011102  -3.265
```

We now have a more stable model, free of boundary constraints. In fact, we can attempt to determine whether or not removing the two variance component parameters for Model C.1 provides a significant reduction in performance. Based on a likelihood ratio test (see below), we do not have significant evidence (chi-square test statistic=2.1471 on 2 df, \(p=0.3418\)) that \(\sigma^2_{11}\) or \(\sigma_{0011}\) is nonzero, so it is advisable to use the simpler Model C.1. However, Section 10.6 describes why this test may be misleading and prescribes a potentially better approach.

```
Df   AIC   BIC logLik Chisq Chi Df Pr(>Chisq)
modelC1 13 581.97 634.27 -277.98
modelC 15 583.82 644.17 -276.91 2.1471 2   0.3418
```
10.6 Parametric bootstrap testing

When testing random effects at the boundary (such as \( \sigma^2_{11} = 0 \)), using a chi-square distribution to conduct a likelihood ratio test like we did in Section 10.5 is not appropriate. In fact, this will produce a conservative test, with p-values that are too large and not rejected enough (Raudenbush and Bryk, Singer and Willett, Faraway). For example, we should suspect that the p-value (.3418) produced by the likelihood ratio test comparing Models C and C.1 in Section 10.5 is too large, that the real probability of getting a likelihood ratio test statistic of 2.1471 or greater when Model C.1 is true is smaller than .3418. Researchers often use the **parametric bootstrap** to better approximate the distribution of the likelihood test statistic and produce more accurate p-values by simulating data under the null hypothesis.

Here are the basic steps for running a parametric bootstrap procedure to compare Model C.1 with Model C (see associated diagram in Figure 10.15):

- Fit Model C.1 (the null model) to obtain estimated fixed and random effects (This is the “parametric” part.)
- Use the estimated fixed and random effects from the null model to regenerate a new set of plant heights with the same sample size (\( n = 413 \)) and associated covariates for each observation as the original data (This is the “bootstrap” part.)
- Fit both Model C.1 (the reduced model) and Model C (the full model) to the new data
- Compute a likelihood ratio statistic comparing Models C.1 and C
- Repeat the previous 3 steps many times (e.g., 1000)
- Produce a histogram of likelihood ratio statistics to illustrate its behavior when the null hypothesis is true
- Calculate a p-value by finding the proportion of times the bootstrapped test statistic is greater than our observed test statistic

![Diagram of parametric bootstrap procedure](image)

**Figure 10.15:** *The steps in conducting a parametric bootstrap test comparing Models C and C.1.*

Let’s see how new plant heights are generated under the parametric bootstrap. Consider, for instance, \( i = 1 \) and \( j = 1, 2 \). That is, consider Plants #11 and #12. These plants are found in Pot #1, which was randomly assigned to contain sterilized soil from a restored prairie (STP):

<table>
<thead>
<tr>
<th>pot</th>
<th>plant</th>
<th>soil</th>
<th>sterile</th>
<th>species</th>
<th>germin</th>
<th>hgt13</th>
<th>hgt18</th>
<th>hgt23</th>
<th>hgt28</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>sterile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Level Three

One way to see the data generation process under the null model (Model C.1) is to start with Level Three and work backwards to Level One. Recall that our Level Three models for $\lambda_{00i}$ and $\lambda_{10i}$, the true intercept and slope from Pot $i$, in Model C.1 are:

$$\lambda_{00i} = \beta_{000} + \beta_{001} \text{strl}_i + \beta_{002} \text{cult}_i + \beta_{003} \text{rem}_i + \varepsilon_{00i}$$
$$\lambda_{10i} = \beta_{100} + \beta_{101} \text{strl}_i + \beta_{102} \text{cult}_i + \beta_{103} \text{rem}_i$$

All the $\beta$ terms will be fixed at their estimated values, so the one term that will change for each bootstrapped data set is $\varepsilon_{00i}$. As we obtain a numeric value for $\varepsilon_{00i}$ for each pot, our notation will change slightly. Specifically, a comma will be added to our subscript to clearly separate the part of the subscript designating the model term (before the comma) and the part designating the observational unit (after the comma). For example, if $\varepsilon_{00i}$ is set to -.192 for Pot #1, then we would denote this by $\varepsilon_{00,1} = -.192$. Similarly, in the context of Model C.1, $\lambda_{00,1}$ represents the mean height at Day 13 across all plants in Pot #1, where $\varepsilon_{00,1}$ quantifies how Pot #1’s Day 13 height relates to other pots with the same sterilization and soil type.

According to Model C.1, each $\varepsilon_{00i}$ is sampled from a normal distribution with mean 0 and standard deviation .240 (note that the standard deviation $\sigma^2_{\varepsilon0}$ is also fixed at its estimated value from Model C.1, given in Section 10.5). For example, a random component to the intercept for Pot #1 ($\varepsilon_{00,1}$) would be sampled from the just-mentioned normal distribution; say, for instance, $\varepsilon_{00,1} = -.192$. Then we can produce a model-based intercept and slope for Pot #1:

$$\lambda_{00,1} = 1.512 - .088(1) + .133(0) + .107(0) - .192 = 1.232$$
$$\lambda_{10,1} = .101 + .059(1) - .031(0) - .036(0) = .160$$

Notice a couple of features of the above derivations. First, all of the coefficients from the above equations ($\beta_{000} = 1.512$, $\beta_{001} = -.088$, etc.) come from the estimated fixed effects from Model C.1 reported in Section 10.5. Second, restored prairie is the reference level for soil type, so that indicators for cultivated land and remnant prairie are both 0. Third, the mean intercept (Day 13 height) for observations from sterilized restored prairie soil is $1.512 - 0.088 = 1.424$ mm across all pots, while the mean daily growth is $1.60$ mm. Pot #1 therefore has mean Day 13 height that is $.192$ mm below the mean for all pots with sterilized restored prairie soil, but every such pot is assumed to have the same growth rate of $1.160$ mm/day because of our assumption that there is no pot-to-pot variability in growth rate (i.e., $\sigma^{2}_{11} = 0$).

Level Two

We next proceed to Level Two, where our equations for Model C.1 are:

$$\lambda_{0ij} = \lambda_{00i} + \varepsilon_{0ij}$$
$$\lambda_{1ij} = \lambda_{10i} + \varepsilon_{1ij}$$

We will initially focus on Plant #11 from Pot #1. Notice that the intercept (Day 13 height = $\lambda_{0,11}$) for Plant #11 has two components: the mean Day 13 height for Pot #1 ($\lambda_{00,1}$) which we specified at Level Three, and an error term ($\varepsilon_{0,11}$) which indicates how the Day 13 height for Plant #11 differs from the overall average for all plants from Pot #1. The slope (daily growth rate = $\lambda_{1,11}$) for Plant #11 similarly has two components. Since both $\lambda_{0,11}$ and $\lambda_{1,11}$ were determined at Level Three, at this point we need to find the two error terms for Plant #11: $\varepsilon_{0,11}$ and $\varepsilon_{1,11}$. According to our multilevel model, we can sample $\varepsilon_{0,11}$ and $\varepsilon_{1,11}$ from a bivariate normal distribution with means both...
equal to 0, standard deviation for the intercept of .543, standard deviation for the slope of .036, and correlation between the intercept and slope of .194.

For instance, suppose we sample $\varepsilon_{0,11} = .336$ and $\varepsilon_{1,11} = .029$. Then we can produce a model-based intercept and slope for Plant #11:

$$\lambda_{0,11} = 1.232 + .336 = 1.568$$
$$\lambda_{1,11} = .160 + .029 = .189$$

Although plants from Pot #1 have a mean Day 13 height of 1.232 mm, Plant #11’s mean Day 13 height is .336 mm above that. Similarly, although plants from Pot #1 have a mean growth rate of .160 mm/day (just like every other pot with sterilized restored prairie soil), Plant #11’s growth rate is .029 mm/day faster.

Level One

Finally we proceed to Level One, where the height of Plant #11 is modeled as a linear function of time $(1.568 + .189time_{11k})$ with a normally distributed residual $\varepsilon_{11k}$ at each time point $k$. Four residuals (one for each time point) are sampled independently from a normal distribution with mean 0 and standard deviation .287 – the standard deviation again coming from parameter estimates from fitting Model C.1 to the actual data as reported in Section 10.5. Suppose we obtain residuals of $\varepsilon_{111} = -.311$, $\varepsilon_{112} = .119$, $\varepsilon_{113} = .241$, and $\varepsilon_{114} = -.066$. In that case, our parametrically generated data for Plant #11 from Pot #1 would look like:

$$Y_{111} = 1.568 + .189(0) - .311 = 1.257$$
$$Y_{112} = 1.568 + .189(5) + .119 = 2.632$$
$$Y_{113} = 1.568 + .189(10) + .241 = 3.699$$
$$Y_{114} = 1.568 + .189(15) - .066 = 4.337$$

We would next turn to Plant #12 from Pot #1 ($i = 1$ and $j = 2$). Fixed effects would remain the same, as would coefficients for Pot #1, $\lambda_{00,1} = 1.232$ and $\lambda_{10,1} = .160$, at Level Three. We would, however, sample new residuals $\varepsilon_{0,12}$ and $\varepsilon_{1,12}$ at Level Two, producing a different intercept $\lambda_{0,12}$ and slope $\lambda_{1,12}$ than those observed for Plant #11. Four new independent residuals $\varepsilon_{12k}$ would also be selected at Level One, from the same normal distribution as before with mean 0 and standard deviation .287.

Once an entire set of simulated heights for every pot, plant, and time point have been generated based on Model C.1, two models are fit to this data:

- Model C.1 – the correct (null) model that was actually used to generate the responses
- Model C – the incorrect (full) model that contains two extra variance components – $\sigma^2_{11}$ and $\sigma_{0011}$ – that were not actually used when generating the responses

A likelihood ratio test statistic is calculated comparing Model C.1 to Model C. For example, after continuing as above to generate new $Y_{ijk}$ values corresponding to all 413 leadplant height measurements, we fit the “bootstrapped” data using both models:

```
# Model comparison for first bootstrapped sample based on Model C.1 #
modelC1: hgt ~ time13 + strl + cult + rem + time13:strl + time13:cult +
        time13:rem + (time13 | plant) + (1 | pot)
modelC:  hgt ~ time13 + strl + cult + rem + time13:strl + time13:cult +
         time13:rem + (time13 | plant) + (time13 | pot)

Df  AIC  BIC logLik  Chisq Chi Df Pr(>Chisq)
modelC1 13 566.24 618.55  -270.12
modelC 15 569.70 630.05  -269.85  0.544 2 0.7619
```

Since the data was generated using Model C.1, we would expect the two extra terms in Model C ($\sigma^2_{11}$
and $\sigma_{0011}$) to contribute very little to the quality of the fit; Model C will have a slightly larger likelihood and loglikelihood since it contains every parameter from Model C.1 plus two more, but the difference in the likelihoods should be due to chance. In fact, that is what the output above shows. Model C does have a larger loglikelihood than Model C.1 (-269.85 vs. -270.12), but this small difference is not statistically significant based on a chi-square test with 2 degrees of freedom ($p = .7619$).

However, we are really only interested in saving the likelihood ratio test statistic from this bootstrapped sample ($2 * (-269.85 - (-270.12)) = 0.544$). By generating (“bootstrapping”) many sets of responses based on estimated parameters from Model C.1 and calculating many likelihood ratio test statistics, we can observe how this test statistic behaves under the null hypothesis of $\sigma_{011}^2 = \sigma_{0011} = 0$, rather than making the (dubious) assumption that its behavior is described by a chi-square distribution with 2 degrees of freedom. Figure 10.16 illustrates the null distribution of the likelihood ratio test statistic derived by the parametric bootstrap procedure as compared to a chi-square distribution. A p-value for comparing our full and reduced models can be approximated by finding the proportion of likelihood ratio test statistics generated under the null model which exceed our observed likelihood ratio test (2.1471). The parametric bootstrap provides a more reliable p-value in this case (.122 from table below); a chi-square distribution puts too much mass in the tail and not enough near 0, leading to overestimation of the p-value. Based on this test, we would still choose our simpler Model C.1.

<table>
<thead>
<tr>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Pr_boot(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>modelC1</td>
<td>13 581.97</td>
<td>634.27</td>
<td>-277.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>modelC</td>
<td>15 583.82</td>
<td>644.17</td>
<td>-276.91</td>
<td>2.1471</td>
<td>0.122</td>
</tr>
</tbody>
</table>

Figure 10.16 Null distribution of likelihood ratio test statistic derived using parametric bootstrap (histogram) compared to a chi-square distribution with 2 degrees of freedom (smooth curve). The horizontal line represents the observed likelihood ratio test statistic.

In this section, we have offered the parametric bootstrap as a noticeable improvement over the likelihood ratio test with an approximate chi-square distribution for testing random effects near the boundary. And typically when we conduct hypothesis tests involving variance terms we are testing at the boundary, since we are asking if the variance term is really necessary (i.e., $H_0: \sigma^2 = 0$ vs. $H_A: \sigma^2 > 0$). However, what if we are conducting a hypothesis test about a fixed effect? For the typical test of whether or not a fixed effect is significant – i.e., $H_0: \beta = 0$ vs. $H_A: \beta \neq 0$ – we are not testing at the boundary, since most fixed effects have no bounds on allowable values. We have
often used a likelihood ratio test with an approximate chi-square distribution in these settings – does that provide accurate p-values? Although some research (e.g., Faraway, 2006) shows that p-values of fixed effects from likelihood ratio tests can tend to be anti-conservative (too high), in general the approximation is not bad. We will continue to use the likelihood ratio test with a chi-square distribution for fixed effects, but you could always check your p-values using a parametric bootstrap approach.

10.7 Exploding variance components

Our modeling task in Section 10.5 was simplified by the absence of covariates at Level Two. As multilevel models grow to include three or more levels, the addition of just a few covariates at lower levels can lead to a huge increase in the number of parameters (fixed effects and variance components) that must be estimated throughout the model. In this section, we will examine when and why the number of model parameters might explode, and we will consider strategies for dealing with these potentially complex models.

For instance, consider Model C, where we must estimate a total of 15 parameters: 8 fixed effects plus 7 variance components (1 at Level One, 3 at Level Two, and 3 at Level Three). By adding just a single covariate to the equations for \( \lambda_{0ij} \) and \( \lambda_{1ij} \) at Level Two in Model C (say, for instance, the size of each seed), we would now have a total of 30 parameters to estimate! The new multilevel model (Model C+) could be written as follows:

- Level One (timepoint within plant):

  \[
  Y_{ijk} = \lambda_{0ij} + \lambda_{ij} time_{ijk} + \epsilon_{ijk}
  \]  

  (10.7)

- Level Two (plant within pot):

  \[
  \begin{align*}
  \lambda_{0ij} & = \lambda_{00i} + \lambda_{01i} seed_{ij} + \epsilon_{0ij} \\
  \lambda_{1ij} & = \lambda_{10i} + \lambda_{11i} seed_{ij} + \epsilon_{1ij}
  \end{align*}
  \]

- Level Three (pot):

  \[
  \begin{align*}
  \lambda_{00i} & = \beta_{000} + \beta_{001} strl_i + \beta_{002} cult_i + \beta_{003} rem_i + \epsilon_{00i} \\
  \lambda_{01i} & = \beta_{010} + \beta_{011} strl_i + \beta_{012} cult_i + \beta_{013} rem_i + \epsilon_{01i} \\
  \lambda_{10i} & = \beta_{100} + \beta_{101} strl_i + \beta_{102} cult_i + \beta_{103} rem_i + \epsilon_{10i} \\
  \lambda_{11i} & = \beta_{110} + \beta_{111} strl_i + \beta_{112} cult_i + \beta_{113} rem_i + \epsilon_{11i}
  \end{align*}
  \]

or as a composite model:

\[
Y_{ijk} = [\beta_{000} + \beta_{001} strl_i + \beta_{002} cult_i + \beta_{003} rem_i + \beta_{010} seed_{ij} + \\
\beta_{011} strl_i seed_{ij} + \beta_{012} cult_i seed_{ij} + \beta_{013} rem_i seed_{ij} + \\
\beta_{100} time_{ijk} + \beta_{101} strl_i time_{ijk} + \beta_{102} cult_i time_{ijk} + \beta_{103} rem_i time_{ijk} + \\
\beta_{110} seed_{ij} time_{ijk} + \beta_{111} strl_i seed_{ij} time_{ijk} + \beta_{112} cult_i seed_{ij} time_{ijk} + \\
\beta_{113} rem_i seed_{ij} time_{ijk} + ] + \\
[ \epsilon_{00i} + \epsilon_{0ij} + \epsilon_{10i} + \epsilon_{1ij} + \epsilon_{ij} + \epsilon_{1ij} + \epsilon_{11i} + \epsilon_{11ij} + \epsilon_{ij} + \epsilon_{1ij} + \epsilon_{11ij} + ]
\]

where \( \epsilon_{ijk} \sim N(0, \sigma^2) \),

\[
\begin{bmatrix}
\epsilon_{0ij} \\
\epsilon_{1ij}
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix} , \begin{bmatrix} \sigma_0^2 & \rho \sigma_0 \sigma_1 \\
\rho \sigma_0 \sigma_1 & \sigma_1^2 \end{bmatrix} \right),
\]
MULTILEVEL DATA WITH MORE THAN TWO LEVELS

We would have 16 fixed effects from the four equations at Level Three, each with 4 fixed effects to estimate if we use the same 3 indicators for soil type and sterilization for each λ. And, with four equations at Level Three, the error covariance matrix at the pot level would be 4x4 with 10 variance components to estimate; each error term (4) has a variance associated with it, and each pair of error terms (6) has an associated correlation. The error structure at Levels One (1 variance term) and Two (2 variance terms and 1 correlation) would remain the same, for a total of 14 variance components.

Now consider adding an extra Level One covariate to the model in the previous paragraph. How many model parameters would now need to be estimated? (Try writing out the multilevel models and counting parameters...) The correct answer is 52 total parameters! There are 24 fixed effects (from 6 Level Three equations) and 28 variance components (1 at Level One, 6 at Level Two, and 21 at Level Three).

Estimating even 30 parameters as in Model Cplus from a single set of data is an ambitious task and computationally very challenging. Essentially we (or the statistics package we are using) must determine which combination of values for the 30 parameters would maximize the likelihood associated with our model and the observed data. Even in the absolute simplest case, with only two options for each parameter, there would be over one billion possible combinations to consider. But if our primary interest is in fixed effects, we really only have 5 covariates (and their associated interactions) in our model. What can be done to make fitting a 3-level model with 1 covariate at Level One, 1 at Level Two, and 3 at Level Three more manageable? Reasonable options include:

- Reduce the number of variance components by assuming all error terms to be independent; that is, set all correlation terms to 0.
- Reduce the number of variance components by removing error terms from certain Level Two and Three equations. Often, researchers will begin with a random intercepts model, in which only the first equation at Level Two and Three has an error term. With the leadplant data, we would account for variability in Day 13 height (intercept) among plants and pots, but assume that the effects of time and seed size are constant among pots and plants.
- Reduce the number of fixed effects by removing interaction terms that are not expected to be meaningful. Interaction terms between covariates at different levels can be eliminated simply by reducing the number of terms in certain equations at Levels Two and Three. There is no requirement that all equations at a certain level contain the same set of predictors. Often, researchers will not include covariates in equations beyond the intercept at a certain level unless there’s a compelling reason.

By following the options above, our potential 30-parameter model can be simplified to this 9-parameter model:

- **Level One:**
  \[ Y_{ijk} = \lambda_{0ij} + \lambda_{1ij}time_{ijk} + \epsilon_{ijk} \]

- **Level Two:**
  \[ \lambda_{0ij} = \lambda_{00i} + \lambda_{01j}seedsize_{ij} + \epsilon_{0ij} \]
  \[ \lambda_{1ij} = \lambda_{10i} \]

- **Level Three:**
  \[ \lambda_{00i} = \beta_{000} + \beta_{001}strl_i + \beta_{002}cult_i + \beta_{003}rem_i + \epsilon_{00i} \]
BUILDING TO A FINAL MODEL

\[
\lambda_{0i} = \beta_{010} \\
\lambda_{10i} = \beta_{100}
\]

where \( \varepsilon_{ijk} \sim N(0, \sigma^2) \), \( \varepsilon_{0ij} \sim N(0, \sigma_{00}^2) \), and \( \varepsilon_{00i} \sim N(0, \sigma_{000}^2) \). Or, in terms of a composite model:

\[
Y_{ijk} = [\beta_{000} + \beta_{001} \text{strl}_i + \beta_{002} \text{cult}_i + \beta_{003} \text{rem}_i + \beta_{010} \text{seedsize}_{ij} + \beta_{100} \text{time}_{ijk}] + \\
[\varepsilon_{00i} + \varepsilon_{0ij} + \varepsilon_{ijk}]
\]

According to the second option, we have built a random intercepts model with error terms only at the first (intercept) equation at each level. Not only does this eliminate variance terms associated with the missing error terms, but it also eliminates correlation terms between errors (as suggested by Option 1) since there are no pairs of error terms that can be formed at any level. In addition, as suggested by Option 3, we have eliminated predictors (and their fixed effects coefficients) at every equation other than the intercept at each level.

The simplified 9-parameter model essentially includes a random effect for pot \((\sigma_{000}^2)\) after controlling for sterilization and soil type, a random effect for plant within pot \((\sigma_{0}^2)\) after controlling for seed size, and a random effect for error about the time trend for individual plants \((\sigma^2)\). We must assume that the effect of time is the same for all plants and all pots, and it does not depend on seed size, sterilization, or soil type. Similarly, we must assume that the effect of seed size is the same for each pot and does not depend on sterilization or soil type. While somewhat prescriptive, a random intercepts model such as this can be a sensible starting point, since the simple act of accounting for variability of observational units at Level Two or Three can produce better estimates of fixed effects of interest and their standard errors.

10.8 Building to a final model

In Model C we considered the main effects of soil type and sterilization on leadplant initial height and growth rate, but we did not consider interactions—even though biology researchers expect that sterilization will aid growth in certain soil types more than others. Thus, in Model D we will build Level Three interaction terms into Model C.1:

- Level One:
  \[
  Y_{ijk} = \lambda_{0ij} + \lambda_{1ij} \text{time}_{ijk} + \varepsilon_{ijk}
  \]

- Level Two:
  \[
  \lambda_{0ij} = \lambda_{00i} + \varepsilon_{0ij} \\
  \lambda_{1ij} = \lambda_{10i} + \varepsilon_{1ij}
  \]

- Level Three:
  \[
  \lambda_{00i} = \beta_{000} + \beta_{001} \text{strl}_i + \beta_{002} \text{cult}_i + \beta_{003} \text{rem}_i + \beta_{005} \text{strl}_i \text{cult}_i + \varepsilon_{00i} \\
  \lambda_{10i} = \beta_{100} + \beta_{101} \text{strl}_i + \beta_{102} \text{cult}_i + \beta_{103} \text{rem}_i + \beta_{105} \text{strl}_i \text{cult}_i
  \]

where error terms are defined as in Model C.1.

From the output below, we see that the interaction terms were not especially helpful, except possibly for a differential effect of sterilization in remnant and reconstructed prairies on the growth rate of leadplants. But it’s clear that Model D can be simplified through the removal of certain fixed effects with low t-ratios.
To form Model F, we begin by removing all covariates describing the intercept (Day 13 height), since neither sterilization nor soil type nor their interaction appear to be significantly related to initial height. However, sterilization, remnant prairie soil, and their interaction appear to have significant influences on growth rate, although the effect of cultivated soil on growth rate did not appear significantly different from that of restored prairie soil (the reference level). A likelihood ratio test shows no significant difference between Models D and F (chi-square test statistic = 10.52 on 7 df, p=.1612), supporting the use of simplified Model F.

# Model F #

(time13 | plant) + (1 | pot)

AIC BIC logLik deviance REMLdev
614.1 654.3 -297.1 561.3 594.1

Random effects:

Groups Name Variance Std.Dev. Corr
plant (Intercept) 0.2940436 0.542258
   time13 0.0022624 0.047602 0.155
pot (Intercept) 0.0487102 0.220704
   Residual 0.0820448 0.286435

Number of obs: 413, groups: plant, 107; pot, 32
We mentioned in Section 10.6 that we could compare Model F and Model D, which differ only in their fixed effects terms, using the parametric bootstrap approach. In fact, in Section 10.6 we suggested that the p-value using the chi-square approximation (.1612) may be a slight under-estimate of the true p-value, but probably in the ballpark. In fact, when we generated 1000 bootstrapped samples of plant heights under Model F, and produced 1000 simulated likelihood ratios comparing Models D and F, we produced a p-value of .199. In Figure 10.17, we see that the chi-square distribution has too much area in the peak and too little area in the tails, although in general it approximates the parametric bootstrap distribution of the likelihood ratio pretty nicely.

The effects of remnant prairie soil and the interaction between remnant soil and sterilization appear to have marginal benefit in Model F, so we remove those two terms to create Model E. A likelihood ratio test comparing Models E and F, however, shows that Model F significantly outperforms Model E (chi-square test statistic = 9.36 on 2 df, p=.0093). Thus, we will use Model F as our “Final Model” for generating inference.

# Model E #
Formula: hgt ~ time13 + time13:strl + (time13 | plant) + (1 | pot)

AIC   BIC  logLik  deviance  REMLdev
606.5 638.7 -295.2   570.6   590.5

Random effects:
Groups     Name    Variance   Std.Dev.   Corr
plant (Intercept)  0.2944091  0.542595
               time13  0.0015899  0.039873  0.138
pot (Intercept)  0.0471188  0.217069
Residual        0.0819900  0.286339

Number of obs: 413, groups: plant, 107; pot, 32

Fixed effects:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.528771</td>
<td>0.070962</td>
</tr>
<tr>
<td>time13</td>
<td>0.085682</td>
<td>0.006524</td>
</tr>
<tr>
<td>time13:strl</td>
<td>0.058645</td>
<td>0.009364</td>
</tr>
</tbody>
</table>

Df  AIC   BIC  logLik  Chisq Chi Df  Pr(>Chisq)
modelE 8 586.63 618.82 -285.31 
modelF 10 581.27 621.50 -280.63 9.3576 2 0.00929 **

Our final model (Model F), with its constraints on Level Three error terms, can be expressed level-by-level as:

- Level One:
  \[ Y_{ijk} = \lambda_{0ij} + \lambda_{1ij}time_{ijk} + e_{ijk} \]

- Level Two:
  \[ \lambda_{0ij} = \lambda_{00i} + \epsilon_{0ij} \]
  \[ \lambda_{1ij} = \lambda_{10i} + \epsilon_{1ij} \]

- Level Three:
  \[ \lambda_{00i} = \beta_{00i} + \epsilon_{00i} \]
  \[ \lambda_{10i} = \beta_{10i} + \beta_{101}strl + \beta_{102}rem + \beta_{103}strl:rem \]

where \( e_{ijk} \sim N(0, \sigma^2) \),
\[
\begin{bmatrix}
  \epsilon_{0ij} \\
  \epsilon_{1ij}
\end{bmatrix}
\sim N\left(\begin{bmatrix} 0 \\
0 \end{bmatrix}, \begin{bmatrix} \sigma^2_0 & \sigma^1_0 \\
\sigma^1_0 & \sigma^2_1 \end{bmatrix}\right),
\]
and \( \epsilon_{00i} \sim N(0, \sigma^2_{00}) \).

In composite form, we have:
\[
Y_{ijk} = [\beta_{00i} + \beta_{10i}time_{ijk} + \beta_{101}strl:time_{ijk} + \beta_{102}rem:time_{ijk} + \beta_{103}strl:rem:time_{ijk}] +
[\epsilon_{00i} + \epsilon_{0ij} + \epsilon_{1ij} + \epsilon_{11j}time_{ijk}]
\]

Estimates of model parameters can be interpreted in the following manner:

- \( \sigma = .287 \) = the standard deviation in within-plant residuals after accounting for time.
COVARIANCE STRUCTURE (OPTIONAL) 249

- $\hat{s}_{0} = 0.543$ = the standard deviation in Day 13 heights between plants from the same pot.
- $\hat{s}_{1} = 0.037$ = the standard deviation in rates of change in height between plants from the same pot.
- $\hat{\rho}_{10} = 0.157$ = the correlation in plants’ Day 13 height and their rate of change in height.
- $\hat{\sigma}_{00} = 0.206$ = the standard deviation in Day 13 heights between pots.
- $\hat{\beta}_{100} = 1.529$ = the mean height for leadplants 13 days after planting.
- $\hat{\beta}_{101} = 0.091$ = the mean daily change in height for leadplants from reconstructed prairies or cultivated lands (rem=0) with no sterilization (strl=0) from 13 to 28 days after planting.
- $\hat{\beta}_{102} = 0.060$ = the increase in mean daily change in height for leadplants from using sterilized soil instead of unsterilized soil in reconstructed prairies or cultivated lands. Thus, leadplants grown in sterilized soil from reconstructed prairies or cultivated lands have an estimated daily increase in height of 0.151 mm.
- $\hat{\beta}_{103} = -0.017$ = the decrease in mean daily change in height for leadplants from using unsterilized soil from remnant prairies, rather than unsterilized soil from reconstructed prairies or cultivated lands. Thus, leadplants grown in unsterilized soil from remnant prairies have an estimated daily increase in height of 0.074 mm.
- $\hat{\beta}_{101} = -0.039$ = the decrease in mean daily change in height for leadplants from sterilized soil from remnant prairies, compared to the expected daily change based on $\hat{\beta}_{101}$ and $\hat{\beta}_{102}$. Three-way interactions show that the size of an interaction between two predictors differs depending on the level of a third predictor. In this case, we might focus on how the interaction between remnant prairies and time differs for unsterilized and sterilized soil. Specifically, the negative effect of remnant prairies on growth rate (compared to reconstructed prairies or cultivated lands) is larger in sterilized soil than unsterilized; in sterilized soil, plants from remnant prairie soil grow .056 mm/day slower on average than plants from other soil types (.095 vs .151 mm/day), while in unsterilized soil, plants from remnant prairie soil grow .017 mm/day slower than plants from other soil types (.074 vs .091 mm/day). Note that the difference between .056 and .017 is our three-way interaction coefficient. Through this three-way interaction term, we also see that leadplants grown in sterilized soil from remnant prairies have an estimated daily increase in height of 0.095 mm.

Based on t-values produced by Model F, sterilization has the most significant effect on leadplant growth, while there is some evidence that growth rate is somewhat slower in remnant prairies, and that the effect of sterilization is also somewhat muted in remnant prairies. Sterilization leads to an estimated 66% increase in growth rate of leadplants from Days 13 to 28 in soil from reconstructed prairies and cultivated lands, and an estimated 28% increase in soil from remnant prairies. In unsterilized soil, plants from remnant prairies grow an estimated 19% slower than plants from other soil types.

10.9 Covariance structure (Optional)

As in Chapter 9, it is important to be aware of the covariance structure implied by our chosen models (focusing initially on Model B). Our three-level model, through error terms at each level, defines a specific covariance structure at both the plant level (Level Two) and the pot level (Level Three). For example, our standard model implies a certain level of correlation among measurements from the same plant and among plants from the same pot. Although three-level models are noticeably more complex than two-level models, it is still possible to systematically determine the implication of our standard model; by doing this, we can evaluate whether our fitted model agrees with our exploratory analyses, and we can also decide if it’s worth considering alternative covariance structures.

We will first consider Model B with $\epsilon_{10}$ at Level Three, and then we will evaluate the resulting covariance structure that results from removing $\epsilon_{10}$, thereby restricting $\sigma_{11}^2 = \sigma_{00} = 0$. The
As we saw with the covariance between observations taken at potentially different times (i.e., same timepoint and same plant, different timepoints but same plant, different plants but same pot). Each derivation will rely on the random effects portion of the composite model, since there is no variability associated with fixed effects. For ease of notation, we will let \( t_k = \text{time}_{ijk} \), since all plants were planned to be observed on the same 4 days.

The variance for an individual observation can be expressed as:

\[
Var(Y_{ijk}) = \sigma_0^2 + \sigma_{ij}^2 + \sigma_{ijk}^2 + 2(\sigma_{01} + \sigma_{001} + \sigma_{0011})t_k + (\sigma_{11}^2 + \sigma_{111}^2)t_k^2,
\]

the covariance between observations taken at different timepoints (\( k \) and \( k' \)) from the same plant (\( j \)) is:

\[
Cov(Y_{ijk}, Y_{ijk'}) = \sigma_{00}^2 + (\sigma_{01} + \sigma_{001})(t_k + t_{k'}) + (\sigma_{11}^2 + \sigma_{111}^2)t_k t_{k'},
\]

and the covariance between observations taken at potentially different times (\( k \) and \( k' \)) from different plants (\( j \) and \( j' \)) from the same pot (\( i \)) is:

\[
Cov(Y_{ijk}, Y_{ijk'}) = \sigma_{00}^2 + \sigma_{0011}(t_k + t_{k'}) + \sigma_{11}^2 t_k t_{k'}.
\]

Based on these variances and covariances, the covariance matrix for observations over time from the same plant (\( j \)) from pot \( i \) can be expressed as the following 4x4 matrix:

\[
Cov(Y_{ij}) = \begin{bmatrix}
\sigma_{00}^2 & \sigma_{01} & \sigma_{001} & \sigma_{0011} \\
\sigma_{01} & \sigma_{11}^2 & \sigma_{111} & \sigma_{1111} \\
\sigma_{001} & \sigma_{111} & \sigma_{1111} & \sigma_{11111} \\
\sigma_{0011} & \sigma_{1111} & \sigma_{11111} & \sigma_{111111}
\end{bmatrix},
\]

where \( \sigma_{ij}^2 = Var(Y_{ij}) \) and \( \sigma_{kk'} = Cov(Y_{ij}, Y_{ijk'}) \). Note that \( \sigma_{ij}^2 \) and \( \sigma_{kk'} \) are both independent of \( i \) and \( j \) so that \( Cov(Y_{ij}) \) will be constant for all plants from all pots. That is, every plant from every pot will have the same set of variances over the four timepoints and the same correlations between heights at different timepoints. But, the variances and correlations can change depending on the timepoint under consideration as suggested by the presence of \( t_k \) terms in Equations 10.9 through 10.11.

Similarly, the covariance matrix between observations from plants \( j \) and \( j' \) from pot \( i \) can be expressed as this 4x4 matrix:

\[
Cov(Y_{ij}, Y_{ij'}) = \begin{bmatrix}
w_{11} & w_{12} & w_{13} & w_{14} \\
w_{12} & w_{22} & w_{23} & w_{24} \\
w_{13} & w_{23} & w_{33} & w_{34} \\
w_{14} & w_{24} & w_{34} & w_{44}
\end{bmatrix},
\]

where \( w_{kk} = Cov(Y_{ij}, Y_{ijk}) = \sigma_{00}^2 + 2\sigma_{0011}t_k + \sigma_{111}^2 t_k^2 \) and \( w_{kk'} = Cov(Y_{ij}, Y_{ijk'}) \) as derived above. As we saw with \( Cov(Y_{ij}) \), \( w_{kk} \) and \( w_{kk'} \) are both independent of \( i \) and \( j \) so that \( Cov(Y_{ij}, Y_{ij'}) \) will
be constant for all pairs of plants from all pots. That is, any pair of plants from the same pot will have the same correlations between heights at any two timepoints. As with any covariance matrix, we can convert Cov(Y_i j, Y_i j') into a correlation matrix if desired.

Now that we have the general covariance structure implied by the standard multilevel model in place, we can examine the specific structure suggested by the estimates of variance components in Model B. Restricted maximum likelihood (REML) in Section 10.4 produced the following estimates for variance components: \( \hat{\sigma}^2 = .0822, \hat{\sigma}^2_0 = .299, \hat{\sigma}^2_1 = .00119, \hat{\rho}_{01} = \rho_{01} \sqrt{\hat{\sigma}^2_0 \hat{\sigma}^2_1} = .00528, \hat{\sigma}^2_{00} = .0442, \hat{\sigma}^2_{11} = .00126, \hat{\sigma}^2_{0011} = \rho_{0011} \sqrt{\hat{\sigma}^2_{00} \hat{\sigma}^2_{11}} = -.00455. \) Based on these estimates and the derivations above, the within-plant correlation structure over time is estimated to be:

\[
\text{Corr}(Y_{ij}) = \begin{bmatrix}
1 & .76 & 1 \\
.65 & .82 & 1 \\
.54 & .77 & .88 & 1
\end{bmatrix}
\]

for all plants \( j \) and all pots \( i \), and the correlation structure between different plants from the same pot is estimated to be:

\[
\text{Corr}(Y_{ij}, Y_{ij'}) = \begin{bmatrix}
.104 & .047 & .061 \\
-.002 & .067 & .116 \\
-.037 & .068 & .144 & .191
\end{bmatrix}.
\]

The within-plant correlation structure suggests that measurements taken closer in time tend to be more highly correlated than those with more separation in time, and later measurements tend to be more highly correlated than earlier measurements. Examination of standard deviation terms by timepoint suggests that variability increases over time within a plant (estimated SDs of .652, .703, .828, and .999 for days 13, 18, 23, and 28, respectively). The correlation structure between plants from the same pot depicts a fairly low level of correlation; even for measurements taken at the same timepoint, the largest correlation between plants from the same pot occurs on Day 28 (\( r = .191 \)) while the smallest correlation occurs on Day 18 (\( r = .061 \)).

We can use these results to estimate within-plant and within-pot correlation structure after imposing the same constraints on Model B that we did on Model F (i.e., \( \sigma^2_{11} = \sigma_{0011} = 0 \)). Using the same REML variance components estimates as above except that \( \hat{\sigma}^2_{11} = 0 \) rather than \(.00126 \) and \( \hat{\sigma}^2_{0011} = \rho_{0011} \sqrt{\hat{\sigma}^2_{00} \hat{\sigma}^2_{11}} = 0 \) rather than \(-.00455 \), the within-plant correlation structure is estimated to be:

\[
\text{Corr}(Y_{ij}) = \begin{bmatrix}
1 & .80 & 1 \\
.75 & .84 & 1 \\
.70 & .82 & .88 & 1
\end{bmatrix}
\]

for all plants \( j \) and all pots \( i \), and the correlation structure between different plants from the same pot is estimated to be:

\[
\text{Corr}(Y_{ij}, Y_{ij'}) = \begin{bmatrix}
.104 & .095 & .087 \\
.084 & .077 & .068 \\
.073 & .067 & .059 & .052
\end{bmatrix}.
\]
Our model restrictions produced slightly higher estimated within-plant correlations, especially for observations separated by longer periods of time. Standard deviation terms by timepoint are very similar (estimated SDs of .652, .713, .806, and .923 for days 13, 18, 23, and 28, respectively). In terms of the relationship between heights of different plants from the same pot, our model restrictions produced slightly higher correlation estimates with Day 13 height, but slightly lower correlation estimates associated with heights at Days 23 and 28. For measurements taken at the same timepoint, the largest correlation between plants from the same pot now occurs on Day 13 (r=.104) while the smallest correlation now occurs on Day 28 (r=.052). None of the differences in the covariance structure, however, should have a large impact on final conclusions, especially regarding fixed effects, so our strategy for dealing with boundary constraints appears very reasonable. In addition, the covariance structure implied by our standard 3-level model appears to model the correlation structure we observed in Figure 10.13 during our exploratory analyses very nicely. Even the variability over time implied by the standard model matches well with the raw observed variability in height by time period (respective standard deviations of .64, .64, .86, and .91). Thus, we feel well justified in fitting models based on the standard covariance structure.

10.9.1 Details of covariance structures

In this section, we present additional details regarding implications of our standard covariance structure for 3-level models. We will focus on Model B; derivations for Model F would proceed in a similar fashion.

The variance for an individual observation can be derived as:

\[
\text{Var}(Y_{ijk}) = \text{Var}(\epsilon_{ijk} + \epsilon_{0ij} + \epsilon_{0ik} + (\epsilon_{1ij} + \epsilon_{10i}) \text{time}_{ijk})
\]

\[
= (\sigma^2 + \sigma_0^2 + \sigma_{01}^2) + 2(\sigma_{01} + \sigma_{001})I_k + (\sigma_1^2 + \sigma_{11}^2)I_k^2
\]

The covariance between observations taken at different timepoints from the same plant is:

\[
\text{Cov}(Y_{ijk}, Y_{ij'k'}) = \text{Cov}(\epsilon_{ijk} + \epsilon_{0ij} + \epsilon_{0ik} + (\epsilon_{1ij} + \epsilon_{10i})t_k, \epsilon_{ij'k'} + \epsilon_{0ij'} + \epsilon_{0ik'} + (\epsilon_{1ij'} + \epsilon_{10i})t_{k'})
\]

\[
= (\sigma^2 + \sigma_0^2 + \sigma_{01}^2) + (\sigma_{01} + \sigma_{001})(t_k + t_{k'}) + (\sigma_1^2 + \sigma_{11}^2)t_k t_{k'}
\]

The covariance between observations taken from different plants from the same pot is:

\[
\text{Cov}(Y_{ijk}, Y_{ij'k'}) = \text{Cov}(\epsilon_{ijk} + \epsilon_{0ij} + \epsilon_{0ik} + (\epsilon_{1ij} + \epsilon_{10i})t_k, \epsilon_{ij'k'} + \epsilon_{0ij'} + \epsilon_{0ik'} + (\epsilon_{1ij'} + \epsilon_{10i})t_{k'})
\]

\[
= \sigma_0^2 + \sigma_{001}(t_k + t_{k'}) + \sigma_1^2 t_k t_{k'}
\]

Based on these variances and covariances and the expressions for \(\text{Cov}(Y_{ij})\) and \(\text{Cov}(Y_{ij}, Y_{ij'})\) in Section 10.9, the complete covariance matrix for observations from pot \(i\) can be expressed as the following 24x24 matrix (assuming 4 observations over time for each of 6 plants):

\[
\text{Cov}(Y_{ij}) = \begin{bmatrix}
\text{Cov}(Y_{ij1}) & \text{Cov}(Y_{ij2}) & \text{Cov}(Y_{ij3}) & \text{Cov}(Y_{ij4}) \\
\text{Cov}(Y_{ij1}, Y_{ij3}) & \text{Cov}(Y_{ij2}, Y_{ij3}) & \text{Cov}(Y_{ij4}, Y_{ij3}) & \text{Cov}(Y_{ij4}, Y_{ij3}) \\
\text{Cov}(Y_{ij1}, Y_{ij4}) & \text{Cov}(Y_{ij2}, Y_{ij4}) & \text{Cov}(Y_{ij4}, Y_{ij4}) & \text{Cov}(Y_{ij4}, Y_{ij4}) \\
\text{Cov}(Y_{ij1}, Y_{ij5}) & \text{Cov}(Y_{ij2}, Y_{ij5}) & \text{Cov}(Y_{ij4}, Y_{ij5}) & \text{Cov}(Y_{ij4}, Y_{ij5}) \\
\end{bmatrix}
\]

A covariance matrix for our entire data set, therefore, would be block diagonal, with \(\text{Cov}(Y_{ij})\) matrices along the diagonal reflecting within pot correlation and 0’s off-diagonal reflecting the assumed independence of observations from plants from different pots. As with any covariance matrix, we can convert the \(\text{Cov}(Y_{ij}, Y_{ij'})\) blocks for two different plants from the same
pot into correlation matrices by dividing covariance terms by the product of corresponding standard deviations. Specifically, for $Cov(Y_{ij}, Y_{ij}')$, the diagonal terms in a correlation matrix are formed by $Corr(Y_{ijk}, Y_{ijk}') = \frac{w_{ikk}}{\sqrt{Var(Y_{ijk})Var(Y_{ijk}')}} = \frac{w_{ik}}{v_k^2}$ and the off-diagonal terms are formed by $Corr(Y_{ijk}, Y_{ijk}') = \frac{w_{ik'}v_k}{\sqrt{Var(Y_{ijk})Var(Y_{ijk}')}} = \frac{w_{ik'}}{v_kv_k'}$.

We calculated estimated covariance and correlation matrices within plant and between plants in Section 10.9 based on the standard covariance structure for three-level models. However, it can sometimes be helpful to consider alternative covariance structures and evaluate the robustness of results to changes in assumptions. A couple of natural covariance structures to fit in the Seed Germination case study, given the observed structure in our data, are the heterogeneous compound symmetry and heterogeneous AR(1) models. We fit both structures, along with the toeplitz structure, and compared the resulting models with our standard 3-level model. In all cases, the AIC and BIC from the standard model (615.2 and 651.4, respectively) are considerably lower than the corresponding performance measures from the models with alternative covariance structures. Thus, we feel justified in fitting models based on the standard covariance structure.

**10.10 Notes on Using R (Optional)**

The R code below fits Models A-C.1 from Sections 10.4 and 10.5. Note that, in `lmer()`, an estimated variance at Level One ($\sigma^2$) comes for “free”, but variance terms at Level Two ($\sigma_0^2$) and Level Three ($\sigma_{01}^2$) must be specified separately in Model A through “(1|plant)” and “(1|pot)”. Specifying “(time13|plant)” in Model B is equivalent to specifying “(1+time13|plant)” and produces a total of 3 variance components to estimate: variances for error terms associated with the intercept ($\sigma_0^2$ comes for “free”) and slope ($\sigma_1^2$ comes from the time13 term) at the plant level, along with a covariance or correlation ($\sigma_{01}$ or $\rho_{01}$) between those two error terms. To restrict $\sigma_{01} = \rho_{01} = 0$, you could specify each error term separately: “(1|plant) + (time13|plant)”.

Also note that to fit Model C.1 in R, the random error components are written as “(time13|plant) + (1|pot)”, indicating you’d like error terms associated with the intercept and slope at the plant level, but only for the intercept term at the pot level. The fixed effects in Model C.1 just reflect all fixed effects in the composite model.

```r
library(lme4)

# Model A - unconditional means
modelal = lmer(hgt ~ 1 + (1|plant) + (1|pot), REML=T, data=leaddata)
summary(modelal)

# Model B - unconditional growth
modelbl = lmer(hgt ~ time13 + (time13|plant) + (time13|pot),
               REML=T, data=leaddata)
summary(modelbl)

# Model C - add covariates at pot level
modelcl = lmer(hgt ~ time13 + strl + cult + rem + time13:strl +
               time13:cult + time13:rem + (time13|plant) + (time13|pot),
               REML=T, data=leaddata)
summary(modelcl) # get corr=-1 at pot level

# Model C.1 - remove two variance components at Level Two from Model C
modelcl0 = lmer(hgt ~ time13 + strl + cult + rem + time13:strl +
                time13:cult + time13:rem + (time13|plant) + (1|pot),
`
In Section 10.6 we sought to perform a significance test comparing Models C and C.1, where Model C.1 restricted two variance components from Model C to be 0. Our initial attempt used the \texttt{anova()} function in R, which created two problems: (a) the \texttt{anova()} function uses full maximum likelihood estimates rather than REML estimates of model parameters and performance, which is fine when two models differ in fixed effects but not, as in this case, when two models differ only in random effects; and, (b) the likelihood ratio test statistic is often not well approximated by a chi-square distribution. Therefore, we implemented the parametric bootstrap method to simulate the distribution of the likelihood ratio test statistic and obtain a more reliable p-value, also illustrating that the chi-square distribution would produce an artificially large p-value.

\begin{verbatim}
anova(modelcl0,modelcl)  # go with Model C.0 (fewer varcomps)

# run bootstrapAnova function = parametric bootstrap code for lme4-models
# http://stats.stackexchange.com/questions/4858/
#  longitudinal-data-analysis-by-multilevel-modeling-in-r/4870#4870
#  Posted on Nov 24, 2010, by Fabian Scheipl

#m0 is the lmer model under the null hypothesis (i.e. the smaller model)
#mA is the lmer model under the alternative

bootstrapAnova <- function(mA, m0, B=1000){
    oneBootstrap <- function(m0, mA){
        d <- drop(simulate(m0))
        m2 <- refit(mA, newresp=d)
        m1 <- refit(m0, newresp=d)
        return(anova(m2,m1)$Chisq[2])
    }
    nulldist <- replicate(B, oneBootstrap(m0, mA))
    ret <- anova(mA, m0)
    names(ret)[7] <- "Pr_boot(>Chisq)"
    attr(ret, "heading") <- c(attr(ret, "heading")[1],
        paste("Parametric bootstrap with", B,"samples."),
        attr(ret, "heading")[1])
    attr(ret, "nulldist") <- nulldist
    return(ret)
}

bRLRT = bootstrapAnova(mA=modelcl, m0=modelcl0)
nullLRT = attr(bRLRT,"nulldist")
hist(nullLRT,prob=T)
abline(v=2.1471,col=2)  # 2.1471 is the observed LRT statistic
x=seq(0,max(nullLRT),length=1000)
y=dchisq(x,2)
lines(x,y)
\end{verbatim}
10.11 Exercises

10.11.1 Conceptual Exercises

1. **Seed Germination.** In Sections 10.3.1 and 10.3.2, why must we be careful excluding plants with no height data? Why can’t we simply leave those plants in our 3-level analysis with all missing heights set to 0 mm?

2. Give an example of a Level Two covariate that might have been recorded in this study.

3. In Figure 10.2, would using mean heights by pot be reasonable as well? How about for Figure 10.3?

4. Explain how “plant-to-plant variability can be estimated by averaging standard deviations from each pot … while pot-to-pot variability can be estimated by finding the standard deviation of average intercept or slope within pot.”

5. Shouldn’t we subtract the mean number of days rather than 13 to calculate centered time? Why does the lower correlation between intercepts and slopes produced by centered time result in a more stable model?

6. Explain why an autoregressive error structure is suggested for leadplant data at the end of Section 10.3.2.

7. The experimental factors of interest in the seed germination study are Level Three covariates, yet the unconditional means model shows only 4.6% of total variability in plant heights to be at Level Three. Does that mean a multilevel model will not be helpful? Would it be better to find the mean height for each pot and just use those 72 values to examine the effects of experimental factors?

8. Explain why a likelihood ratio test is appropriate for comparing Models B and C.

9. Should we be concerned that $\hat{\sigma}_0^2$ increased from Model A to B? Why or why not?

10. Explain the idea of boundary constraints in your own words. Why can it be a problem in multilevel models?

11. In Model C, we initially addressed boundary constraints by removing the Level Three correlation between error terms from our multilevel model. What other model adjustments might we have considered?

12. Recall Model C from Chapter 9. Describe the steps for performing a parametric bootstrap test on $\sigma_{01} = \sigma_1^2 = 0$.

13. How does Figure 10.16 show that a likelihood ratio test using a chi-square distribution would be biased?

14. In Section 10.7, a model with 52 parameters is described: (a) illustrate that the model does indeed contain 52 parameters; (b) explain how to minimize the total number of parameters using ideas from Section 10.7; (c) what assumptions have you made in your simplification in (b)?

15. In Section 10.8, Model F (the null model) is compared to Model D using a parametric bootstrap test. As in Section 10.6, show in detail how bootstrapped data would be generated under Model F for, say, Plant # 1 from Pot # 1. For the random parts, tell what distribution the random pieces are coming from and then select a random value from that distribution. Finally, explain how the parametric bootstrap test would be carried out.

16. Section 10.8 contains an interpretation for the coefficient of a three-way interaction term, $\hat{\beta}_{103}$. Provide an alternative interpretation for $\hat{\beta}_{103}$ by focusing on how the sterilization-by-soil type interaction differs over time.

17. **Collective Efficacy and Violent Crime.** In a 1997 Science article, Sampson, Raudenbush, and Earls studied the effects on violent crime of a neighborhood’s collective efficacy, defined as “social cohesion among neighbors combined with their willingness to intervene on behalf of the common good.” Multiple items related to collective efficacy were collected from 8782 Chicago
residents from 343 neighborhood clusters. For this study, give the observational units at Levels One, Two, and Three.

18. Table 10.3 shows Table 3 from Sampson et al. (1997). Provide interpretations of the bolded coefficients in context.

Table 10.3: Table 3: Correlates of collective efficacy from Sampson et al. (1997).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE</th>
<th>t ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.523</td>
<td>0.013</td>
<td>263.20</td>
</tr>
<tr>
<td>Person-level predictors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.012</td>
<td>0.015</td>
<td>-0.76</td>
</tr>
<tr>
<td>Married</td>
<td>-0.005</td>
<td>0.021</td>
<td>-0.25</td>
</tr>
<tr>
<td>Separated or divorced</td>
<td>-0.045</td>
<td>0.026</td>
<td>-1.72</td>
</tr>
<tr>
<td>Single</td>
<td>-0.026</td>
<td>0.024</td>
<td>-1.05</td>
</tr>
<tr>
<td>Homeowner</td>
<td>0.122</td>
<td>0.020</td>
<td>6.04</td>
</tr>
<tr>
<td>Latino</td>
<td>0.042</td>
<td>0.028</td>
<td>1.52</td>
</tr>
<tr>
<td>Black</td>
<td>-0.029</td>
<td>0.030</td>
<td>-0.98</td>
</tr>
<tr>
<td>Mobility</td>
<td>-0.025</td>
<td>0.007</td>
<td>-3.71</td>
</tr>
<tr>
<td>Age</td>
<td>0.0021</td>
<td>0.006</td>
<td>3.47</td>
</tr>
<tr>
<td>Years in neighborhood</td>
<td>0.0006</td>
<td>0.008</td>
<td>0.78</td>
</tr>
<tr>
<td>SES</td>
<td>0.035</td>
<td>0.008</td>
<td>4.64</td>
</tr>
<tr>
<td>Neighborhood-level predictors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concentrated disadvantage</td>
<td>-0.172</td>
<td>0.016</td>
<td>-10.74</td>
</tr>
<tr>
<td>Immigrant concentration</td>
<td>-0.037</td>
<td>0.014</td>
<td>-2.66</td>
</tr>
<tr>
<td>Residential stability</td>
<td>0.074</td>
<td>0.013</td>
<td>5.61</td>
</tr>
<tr>
<td>Variance components</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within neighborhoods</td>
<td>0.320</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between neighborhoods</td>
<td>0.026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of variance explained</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within neighborhoods</td>
<td>3.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between neighborhoods</td>
<td>70.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

19. Based on Table 10.3, let $Y_{ijk}$ be the response of person $j$ from neighborhood $k$ to item $i$ regarding collective efficacy; these are (difficulty-adjusted) responses to 10 items per person about collective efficacy. Then (a) write out the three-level model that likely produced this table, and (b) write out the corresponding composite model. Assume there was also an unreported variance component estimating item-to-item variability within person.

20. Suggest valuable exploratory data analysis plots to complement Table 10.3.

21. If the model suggested by Table 10.3 were expanded to include all potential variance components, how many total parameters would need to be estimated? If it were expanded further to include the 3 neighborhood-level covariates as predictors in all Level Three equations, how many total parameters would need to be estimated?

22. At the bottom of Table 10.3, the percent of variance explained is given within and between neighborhoods. Explain what these values likely represent and how they were calculated.

23. Table 10.4 shows a portion of Table 4 from Sampson et al. (1997). Describe the multilevel model that likely produced this table. State the primary result from this table in context. [Note that collective efficacy is a Level Three covariate in this table, summarized over an entire neighborhood.]

10.11.2 Guided Exercise

1. Kentucky Math Scores. Data was collected from 48,058 eighth graders from Kentucky who took the California Basic Educational Skills Test (Bickel 2007). These students attended 235 different middle schools from 132 different districts, and the following variables were collected:
   - dis_id = District Identifier
   - sch_id = School Identifier
Table 10.4 A portion of Table 4: Neighborhood correlates of perceived neighborhood violence, violent victimization, and 1995 homicide events from Sampson et al. (1997).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1: social composition</th>
<th>Model 2: social composition and collective efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>SE</td>
</tr>
<tr>
<td>Perceived neighborhood violence*</td>
<td>0.277</td>
<td>0.021</td>
</tr>
<tr>
<td>Concentrated disadvantage</td>
<td>0.041</td>
<td>0.017</td>
</tr>
<tr>
<td>Immigrant concentration</td>
<td>-0.102</td>
<td>0.015</td>
</tr>
<tr>
<td>Residential stability</td>
<td>-0.618</td>
<td>0.104</td>
</tr>
<tr>
<td>Collective efficacy</td>
<td>-0.618</td>
<td>0.104</td>
</tr>
</tbody>
</table>

* Estimates of neighborhood-level coefficients control for gender, marital status, homeownership, ethnicity, mobility, age, years in neighborhood, and SES of those interviewed. Model 1 accounts for 70.5% of the variation between neighborhoods in perceived violence, whereas model 2 accounts for 77.8% of the variation.

The primary research questions are whether or not math scores of eighth graders in Kentucky differ based on gender or ethnicity, and if any gender gap differs by ethnicity, or if any ethnicity gap differs by gender. Researchers wanted to be sure to adjust for important covariates at the school and district levels (e.g. school/district size and socio-economic status).

(a) Conduct a short exploratory data analysis. Which variables are most strongly related to math scores? Illustrate with summary statistics and plots. [Hints: your plots should accommodate the large number of observations in the data, and your summary statistics will have to handle occasional missing values.]

(b) (Model A) Run an unconditional means model with 3 levels and report the percentage of variability in math scores that can be explained at each level.

(c) (Model B) Add female, nonwhite, and their interaction at Level One.
   i. Write out the complete three-level model. How many parameters must be estimated?
   ii. Run the model (be patient – it may take a few minutes!). Report and interpret a relevant pseudo-Rsquare value. Is there evidence (based on the t-value) of a significant interaction? In layman’s terms, what can you conclude based on the test for interaction?

(d) (Model C) Subtract the female-by-nonwhite interaction from Model B and add sch_ses, where sch_ses is a predictor in all Level Two equations. How many parameters must be estimated? Break your count down into fixed effects + variance components at Level 1 + varcomps at Level 2 + varcomps at Level 3. (No need to write out the model or run the model unless you want to...)

(e) (Model D) Subtract nonwhite from Model C and add dis_size, where dis_size is a predictor in all Level Three equations.
   i. Write out the complete three-level model. Also write out the composite model. How many parameters must be estimated?
ii. Run the model (be patient!), and then re-run it with an error term only on the intercept equation at Level Three. What are the implications of using this error structure? Does it change any conclusions about fixed effects? Which of the two models would you choose and why?

iii. (Optional) Explore the nature of the 3-way interaction between female, sch_ses, and dis_size by finding the predicted math score based on Model D with an error term only on the intercept equation at Level Three in 5 cases. Comment on trends you observe.

A. Female vs. male with average sch_ses and dis_size
B. Female vs. male with sch_ses at Q1 and dis_size at Q1
C. Female vs. male with sch_ses at Q1 and dis_size at Q3
D. Female vs. male with sch_ses at Q3 and dis_size at Q1
E. Female vs. male with sch_ses at Q3 and dis_size at Q3

iv. (Optional) Create two scatterplots to illustrate the 3-way interaction between female, sch_ses, and dis_size.

10.11.3 Open-ended Exercises

1. **Seed Germination: Coneflowers** Repeat the exploratory data analyses and model fitting from this chapter using coneflowers rather than leadplants.

2. **Mudamalai Leaf Growth** Plant growth is influenced by many environmental factors, among them sunlight and water availability. A study conducted in the Mudamalai Wildlife Sanctuary in India in October of 2008 had as its purpose to “broaden the specific knowledge of certain tree species and climate zones located in the Nilgiri Hills of South India, and to enhance the overall knowledge of the dynamic relationship between plant growth and its environment” (Pray, 2009).

Study researchers collected 1,960 leaves from 45 different trees (5 trees from each of 3 species within each of 3 climate zone). Within each tree, 3 branches were randomly chosen from each of 3 strata (high, medium, and low), and 5 leaves were randomly selected from each branch. Three different descriptive climatic zones were chosen for analysis—dry thorn, dry deciduous, and moist deciduous—and three different species of trees were analyzed—*Cassia fistula* (golden shower tree), *Anogeissus latifolia* (axlewood tree), and *Diospyros montana* (mountain ebony). Height and girth were measured for each tree, and length was assessed for each branch. Measurements taken on each leaf included length, width, surface area (determined carefully for 25 leaves per species and then by linear regression using length and width as predictors for the rest), pedial length (the length of the stem), pedial width, and percent herbivory (an eyeball estimate of the percent of each leaf that had been removed or eaten by herbivores). In addition, stomata density was measured using a compound scope to examine nail polish impressions of leaves for each strata of each tree (135 measurements).

Here is a description of available variables:

- **Species** = tree species (Cassia fistula, Anogeissus latifolia, or Diospyros montana)
- **Zone** = climate zone (dry thorn, dry deciduous, or moist deciduous)
- **Tree** = tree number (1-5) within climate zone
- **Tree.height** = tree height (m)
- **Tree.girth** = tree girth (cm)
- **Strata** = height of branch from ground (high, medium, or low)
- **Branch** = branch number (1-3) within tree and strata
- **Branch.length** = length of branch (cm)
- **Length** = length of leaf (cm)
Biology researchers were interested in determining “optimal physical characteristics for growth and survival of these trees in the various areas” to further conservation efforts. Construct a multi-level model to address the researchers’ questions, focusing on leaf area as the response of interest. Defend your final model, and interpret fixed effect and variance component estimates produced by your model.
11.1 Learning Objectives

After finishing this chapter, students should be able to:

- Recognize when generalized linear multilevel models (GLMMs) are appropriate.
- Understand how GLMMs essentially combine ideas from earlier chapters.
- Apply exploratory data analysis techniques to multilevel data with non-normal responses.
- Recognize when crossed random effects are appropriate and how they differ from nested random effects.
- Write out a generalized linear multilevel statistical model, including assumptions about variance components.
- Interpret model parameters (including fixed effects and variance components) from a GLMM.
- Generate and interpret random effect estimates.

11.2 Case Study: College Basketball Referees

An article by Anderson and Pierce (2009) describes empirical evidence that officials in NCAA men’s college basketball tend to “even out” foul calls over the course of a game, based on data collected in 2004-2005. Using logistic regression to model the effect of foul differential on the probability that the next foul called would be on the home team (controlling for score differential, conference, and whether or not the home team had the lead), Anderson and Pierce found that “the probability of the next foul being called on the visiting team can reach as high as 0.70.” More recently, Moskowitz and Wertheim, in their book *Scorecasting* (2012), argue that the number one reason for the home field advantage in sports is referee bias. Specifically, in basketball, they demonstrate that calls over which referees have greater control—offensive fouls, loose ball fouls, ambiguous turnovers such as palming and traveling—were more likely to benefit the home team than more clearcut calls, especially in crucial situations.

Data have been gathered from the 2009-2010 college basketball season for three major conferences to investigate the following questions (Noecker and Roback 2013):

- Does evidence that college basketball referees tend to “even out” calls still exist in 2010 as it did in 2005?
- How do results change if our analysis accounts for the correlation in calls from the same game and the same teams?
- Is the tendency to even out calls stronger for fouls over which the referee generally has greater control? Fouls are divided into offensive, personal, and shooting fouls, and one could argue that referees have the most control over offensive fouls (typically where the player with the ball knocks over a stationary defender) and the least control over shooting fouls (where an offensive player is fouled in the act of shooting).
11.3 Initial Exploratory Analyses

11.3.1 Data organization

Data for Case Study 11.2 can be found under basketball0910.csv in our data repository. Examination of the data reveals the following key variables:

- game = unique game identification number
- date = date game was played (YYYYMMDD)
- visitor = visiting team abbreviation
- hometeam = home team abbreviation
- foul.num = cumulative foul number within game
- foul.home = indicator if foul was called on the home team
- foul.vis = indicator if foul was called on the visiting team
- foul.diff = the difference in fouls before the current foul was called (home - visitor)
- score.diff = the score differential before the current foul was called (home - visitor)
- lead.vis = indicator if visiting team has the lead
- lead.home = indicator if home team has the lead
- previous.foul.home = indicator if previous foul was called on the home team
- previous.foul.vis = indicator if previous foul was called on the visiting team
- foul.type = categorical variable if current foul was offensive, personal, or shooting
- shooting = indicator if foul was a shooting foul
- personal = indicator if foul was a personal foul
- offensive = indicator if foul was an offensive foul
- time = number of minutes left in the first half when foul called

Data was collected for 4972 fouls over 340 games from the Big Ten, ACC, and Big East conference seasons during 2009-2010. We focus on fouls called during the first half to avoid the issue of intentional fouls by the trailing team at the end of games. Table 11.1 illustrates key variables from the first 10 rows of the data set.

For our initial analysis, our primary response variable is foul.home, and our primary hypothesis concerns evening out foul calls. We hypothesize that the probability a foul is called on the home team is inversely related to the foul differential; that is, if more fouls have been called on the home team than the visiting team, the next foul is less likely to be on the home team.

The structure of this data suggests a couple of familiar attributes combined in an unfamiliar way. With a binary response variable, a generalized linear model is typically applied, especially one with a logit link function (indicating logistic regression). But, with covariates at multiple levels—some at the individual foul level and others at the game level—a multilevel model would also be sensible. So what we need is a multilevel model with a non-normal response; in other words, a generalized linear multilevel model (GLMM). We will investigate what such a model might look like in the next section, but we will still begin by exploring the data with initial graphical and numerical summaries.

As with other multilevel situations, we will begin with broad summaries across all 4972 foul calls from all 340 games. Most of the variables we have collected can vary with each foul called; these Level One variables include:
**11.3.2 Exploratory analyses**

In Figure 11.1, we see histograms for the continuous Level One covariates (time remaining, foul differential, and score differential). These plots treat each foul within a game as independent even though we expect them to be correlated, but they provide a sense for the overall patterns. We see that time remaining is reasonably uniform, except during the first five minutes of the first half when relatively fewer fouls are called. Score differential and foul differential are both bell-shaped, with a mean slightly favoring the home team in both cases – on average, the home team leads by 2.04 points (SD 7.24) and has 0.36 fewer previous fouls (SD 2.05) at the time a foul is called.

Summaries of the categorical response (whether the foul was called on the home team) and categorical Level One covariates (whether the home team has the lead and what type of foul was called) can be provided through tables of proportions. More fouls are called on visiting teams (52.1%) than home teams, the home team is more likely to hold a lead (57.1%), and personal fouls are most likely to be called (51.6%), followed by shooting fouls (38.7%) and then offensive fouls (9.7%).

For an initial examination of Level Two covariates (the home and visiting teams), we can take the number of times, for instance, Minnesota (MN) appears in the long data set (with one row per foul called as illustrated in Table 11.1) as the home team and divide by the number of unique games in which Minnesota is the home team. This ratio (12.1), found in Table 11.2, shows that Minnesota is among the bottom three teams in the average total number of fouls in the first halves of games in which it is the home team. That is, games at Minnesota have few total fouls relative to games played elsewhere. Accounting for the effect of home and visiting team will likely be an important part of our model, since some teams tend to play in games with twice as many fouls called as others, and other teams see a noticeable disparity in the total number of fouls depending on if they are home or
Next, we inspect numerical and graphical summaries of relationships between Level One model covariates and our binary model response. As with other multilevel analyses, we will begin by observing broad trends involving all 4972 fouls called, even though fouls from the same game may be correlated. The conditional density plots in the first row of Figure 11.2 examine continuous Level One covariates. Figure 11.2a provides support for our primary hypothesis about evening out foul calls, indicating a very strong trend for fouls to be more often called on the home team at points in the game when more fouls had previously been called on the visiting team. Figures 11.2b and 11.2c then show that fouls were somewhat more likely to be called on the home team when the home team’s lead was greater and (very slightly) later in the half. Conclusions from the conditional density plots in Figure 11.2a-c are supported with associated empirical logit plots in Figure 11.2d-f. If a logistic link function is appropriate, these plots should be linear, and the stronger the linear association, the more promising the predictor. We see in Figure 11.2d further confirmation of our primary hypothesis, with lower log-odds of a foul called on the home team the greater number of previous fouls the home team had accumulated compared to the visiting team. Figure 11.2e shows that game score may play a role in foul trends, as the log-odds of a foul on the home team grows as
the home team accumulates a bigger lead on the scoreboard, and Figure 11.2f shows a very slight tendency for greater log-odds of a foul called on the home team as the half proceeds (since points on the right are closer to the beginning of the game).

Conclusions about continuous Level One covariates are further supported by summary statistics calculated separately for fouls called on the home team and those called on the visiting team. For instance, when a foul is called on the home team, there is an average of 0.64 additional fouls on the visitors, compared to an average of 0.10 additional fouls on the visitors when a foul is called on the visiting team. Similarly, when a foul is called on the home team, they are in the lead by an average of 2.7 points, compared to an average home lead of 1.4 points when a foul is called on the visiting team. As expected, the average time remaining the half is very similar for home and visitor fouls (9.2 vs. 9.5 minutes, respectively).

The segmented barcharts in Figure 11.3 examine categorical Level One covariates, indicating that fouls were more likely to be called on the home team when the home team was leading, when the previous foul was on the visiting team, and when the foul was a personal foul rather than a shooting foul or an offensive foul. 51.8% of calls go against the home team when it is leading the game, compared to only 42.9% of calls when the it is behind; 51.3% of calls go against the home team when the previous foul went against the visitors, compared to only 43.8% of calls when the previous foul went against the home team; and, 49.2% of personal fouls are called against the home team, compared to only 46.9% of shooting fouls and 45.7% of offensive fouls. Eventually we will want to examine the relationship between foul type (personal, shooting, or offensive) and foul differential, examining our hypothesis that the tendency to even out calls will be even stronger for calls over which the referees have greater control (personal fouls and especially offensive fouls).

![Figure 11.2 Conditional density and empirical logit plots of the binary model response (foul called on home or visitor) vs. the three continuous Level One covariates (foul differential, score differential, and time remaining). The dark shading in a conditional density plot shows the proportion of fouls called on the home team for a fixed value of (a) foul differential, (b) score differential, and (c) time remaining. In empirical logit plots, estimated log odds of a home team foul are calculated for each distinct foul (d) and score (e) differential, except for differentials at the high and low extremes with insufficient data; for time (f), estimated log odds are calculated for two-minute time intervals and plotted against the midpoints of those interval.](image-url)
seeing univariate trends, and noting bivariate relationships between variable pairs. However, our important research questions (a) involve the effect of foul differential after adjusting for other significant predictors of which team is called for a foul, (b) account for potential correlation between foul calls within a game (or within a particular home or visiting team), and (c) determine if the effect of foul differential is constant across game conditions. In order to address research questions such as these, we need to consider multilevel, multivariate statistical models for a binary response variable.

11.4 Two level Modeling with a Generalized Response

11.4.1 A multiple generalized linear model approach (correlation not accounted for)

One quick and dirty approach to analysis might be to run a multiple logistic regression model on the entire long data set of 4972 fouls. In fact, Anderson and Pierce ran such a model in their 2009 paper, using the results of their multiple logistic regression model to support their primary conclusions, while justifying their approach by confirming a low level of correlation within games and the minimal impact on fixed effect estimates that accounting for clustering would have. Output from one potential multiple logistic regression model is shown below; this initial modeling attempt shows significant evidence that referees tend to even out calls (i.e., that the probability of a foul called on the home team decreases as total home fouls increase compared to total visiting team fouls) after accounting for score differential and time remaining (Z=-3.078, p=.002). The extent of the effect of foul differential also appears to grow (in a negative direction) as the first half goes on, based on an interaction between time remaining and foul differential (Z=-2.485, p=.013). We will compare this model with others that formally account for clustering and correlation patterns in our data.

```
glm(formula = foul.home ~ foul.diff + score.diff + lead.home + time + foul.diff:time + lead.home:time, family = binomial)
```

Coefficients:
11.4.2 A two-stage modeling approach (provides the basic idea for multilevel modeling)

As we saw in Section 8.4.2, to avoid clustering we could consider fitting a separate regression model to each unit of observation at Level Two (each game in this case). Since our primary response variable is binary (Was the foul called on the home team or not?), we would fit a logistic regression model to the data from each game. For example, consider the 14 fouls called during the first half of the March 3, 2010, game featuring Virginia at Boston College (Game 110).

Is there evidence from this game that referees tend to “even out” foul calls when one team starts to accumulate more fouls? Is the score differential associated with the probability of a foul on the home team? Is the effect of foul differential constant across all foul types during this game? We can address these questions through multiple logistic regression applied to only data from Game 110.

First, notation must be defined. Let \( Y_{ij} \) be an indicator variable recording if the \( j^{th} \) foul from Game \( i \) was called on the home team (1) or the visiting team (0). We can consider \( Y_{ij} \) to be a Bernoulli random variable with parameter \( p_{ij} \), where \( p_{ij} \) is the true probability that the \( j^{th} \) foul from Game \( i \) was called on the home team. As in Chapter ??, we will begin by modeling the logit function:
of \( p_{ij} \) for Game \( i \) as a linear function of Level One covariates. For instance, if we were to consider a simple model with foul differential as the sole predictor, we could model the probability of a foul on the home team in Game \( i \) with the model:

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \lambda_{0i} + \lambda_{1i} \text{foul.diff}_{ij}
\]  

(11.1)

where \( i \) is fixed. Note that there is no separate error term or variance parameter, since the variance is a function of \( p_{ij} \) with a Bernoulli random variable.

Maximum likelihood estimators for the parameters in this model (\( \lambda_{0i} \) and \( \lambda_{1i} \)) can be obtained through statistical software. \( e^{\lambda_{0i}} \) represents the odds that a foul is called on the home team when the foul totals in Game \( i \) are even, and \( e^{\lambda_{1i}} \) represents the multiplicative change in the odds that a foul is called on the home team for each additional foul for the home team relative to the visiting team during the first half of Game \( i \). For example, if we compared the situation when the home team has been called for 3 more fouls than the visiting team to when the home team has been called for 4 more fouls than the visitors, we’d estimate \( e^{\lambda_{1i}} \) to be the multiplicative difference in odds of a foul called on the home team (the odds should decrease from 3 to 4 additional home fouls).

For Game 110, we estimate \( \hat{\lambda}_{0i} = -5.67 \) and \( \hat{\lambda}_{1i} = -2.11 \) (see output below). Thus, according to our simple logistic regression model, the odds that a foul is called on the home team when both teams have an equal number of fouls in Game 110 is \( e^{-5.67} = 0.0035 \); that is, the probability that a foul is called on the visiting team (0.9966) is 1/0.0035 = 289 times higher than the probability a foul is called on the home team (0.0034) in that situation. While these parameter estimates seem quite extreme, reliable estimates are difficult to obtain with 14 observations and a binary response variable, especially in a case like this where the fouls were only even at the start of the game. Also, as the gap between home and visiting fouls increases by 1, the odds that a foul is called on the visiting team increases by a multiplicative factor of more than 8 (since \( 1/e^{-2.11} = 8.25 \)). In Game 110, this trend toward referees evening out foul calls is statistically significant at the 0.10 level (Z=-1.851, \( p = .0642 \)).

\[
\text{glm(formula = foul.home} \sim \text{foul.diff, family = binomial, data = game110)}
\]

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -5.668   | 3.131      | -1.810  | 0.0703   |
| foul.diff      | -2.110   | 1.140      | -1.851  | 0.0642   |

Null deviance: 19.121 on 13 degrees of freedom
Residual deviance: 11.754 on 12 degrees of freedom
AIC: 15.754

As in Section 8.4.2, we can proceed to fit similar logistic regression models for each of the 339 other games in our data set. Each model will yield a new estimate of the intercept and the slope from Equation 11.1. These estimates are summarized graphically in Figure 11.4. There is noticeable variability among the 340 games in the fitted intercepts and slopes, with an IQR for intercepts ranging from -1.45 to 0.70, and an IQR for slopes ranging from -1.81 to -0.51. The majority of intercepts are below 0 (median = -0.41), so that in most games a foul is less likely to be called on the home team when foul totals are even. In addition, almost all of the estimated slopes are below 0 (median = -0.89), indicating that as the home team’s foul total grows in relation to the visiting team’s foul total, the odds of a foul on the home team continues to decrease.

At this point, you might imagine expanding model building efforts in a couple of directions: (a)
TWO LEVEL MODELING WITH A GENERALIZED RESPONSE

Figure 11.4 Histograms of (a) intercepts and (b) slopes from fitting simple logistic regression models by game. Several extreme outliers have been cut off in these plots for illustration purposes.

continue to improve the Level One model in Equation 11.1 by controlling for covariates and adding potential interaction terms, or (b) build Level Two models to explain the game-to-game variability in intercepts or slopes using covariates which remain constant from foul to foul within a game (like the teams playing). While we could pursue these two directions independently, we can accomplish our modeling goals in a much cleaner and more powerful way by proceeding as in Chapter 8 and building a unified multilevel framework under which all parameters are estimated simultaneously and we remain faithful to the correlation structure inherent in our data.

11.4.3 A unified multilevel approach (the framework we’ll use)

As in Chapters 8 and 9, we will write out a composite model after first expressing Level One and Level Two models. That is, we will create Level One and Level Two models as in Section 11.4.2, but we will then combine those models into a composite model and estimate all model parameters simultaneously. Once again $Y_{ij}$ is an indicator variable recording if the $j$th foul from Game $i$ was called on the home team (1) or the visiting team (0), and $p_{ij}$ is the true probability that the $j$th foul from Game $i$ was called on the home team. Our Level One model with foul differential as the sole predictor is given by Equation 11.1:

$$
\log\left( \frac{p_{ij}}{1-p_{ij}} \right) = \lambda_{0i} + \lambda_{1i} \text{foul.diff}_{ij}
$$

Then we include no fixed covariates at Level Two, but we include error terms to allow the intercept and slope from Level One to vary by game, and we allow these errors to be correlated:

$$
\begin{align*}
\lambda_{0i} &= \beta_{00} + \epsilon_{0i} \\
\lambda_{1i} &= \beta_{10} + \epsilon_{1i},
\end{align*}
$$

where the error terms at Level Two can be assumed to follow a multivariate normal distribution:

$$
\begin{bmatrix}
\epsilon_{0i} \\
\epsilon_{1i}
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{0}^2 & \sigma_{01} \\ \sigma_{01} & \sigma_{1}^2 \end{bmatrix} \right)
$$
Our composite model then looks like:

\[
\log\left( \frac{p_{ij}}{1-p_{ij}} \right) = \lambda_{0i} + \lambda_{i,foul.diff_{ij}}
\]

\[
= (\beta_{00} + \epsilon_{0i}) + (\beta_{10} + \epsilon_{1i})foul.diff_{ij}
\]

\[
= [\beta_{00} + \beta_{10}foul.diff_{ij}] + [\epsilon_{0i} + \epsilon_{1i}foul.diff_{ij}]
\]

The major changes when moving from a normally distributed response to a binary response are the form of the response variable (a logit function) and the absence of an error term at Level One.

Again, we can use statistical software to obtain parameter estimates for this unified multilevel model using all 4972 fouls recorded from the 340 games. For example, the glmer() function from the lme4 package in R extends the lmer() function to handle generalized responses and to account for the fact that fouls are not independent within games. Results are given below for the two-level model with foul differential as the sole covariate and Game as the Level Two observational unit.

Formula: foul.home ~ foul.diff + (foul.diff | game)

<table>
<thead>
<tr>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>6791</td>
<td>6824</td>
<td>-3391</td>
<td>6781</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>game</td>
<td>(Intercept)</td>
<td>0.2941075</td>
<td>0.542317</td>
<td></td>
</tr>
<tr>
<td></td>
<td>foul.diff</td>
<td>0.0012351</td>
<td>0.035145</td>
<td>-1.000</td>
</tr>
</tbody>
</table>

Number of obs: 4972, groups: game, 340

Fixed effects:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -0.15684 | 0.04210    | -3.725  | 0.000195 *** |
| foul.diff      | -0.28533 | 0.01765    | -16.162 | < 2e-16 *** |

When parameter estimates from the multilevel model above are compared with those from the naive logistic regression model assuming independence of all observations (below), there are noticeable differences. For instance, each additional foul for the visiting team is associated with a 33% increase \((1/e^{-2.85})\) in the odds of a foul called on the home team under the multilevel model, but the single level model estimates the same increase as only 14% \((1/e^{-1.30})\). Also, estimated standard errors for fixed effects are greater under generalized linear multilevel modeling, which is not unusual after accounting for correlated observations, which effectively reduces the sample size.

glm(formula = foul.home ~ foul.diff, family = binomial)

Coefficients:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -0.13005 | 0.02912    | -4.466  | 7.98e-06 *** |
| foul.diff      | -0.13047 | 0.01426    | -9.148  | < 2e-16 *** |

Null deviance: 6884.3 on 4971 degrees of freedom
Residual deviance: 6798.1 on 4970 degrees of freedom
AIC: 6802.1
11.5 Crossed Random Effects

In the College Basketball Referees case study, our two primary Level Two covariates are home team and visiting team. In Section 11.3.2 we showed evidence that the probability a foul is called on the home team changes if we know precisely who the home and visiting teams are. However, if we were to include an indicator variable for each distinct team, we would need 38 indicator variables for home teams and 38 more for visiting teams. That’s a lot of degrees of freedom to spend! And adding those indicator variables would complicate our model considerably, despite the fact that we’re not very interested in specific coefficient estimates for each team – we just want to control for teams playing to draw stronger conclusions about referee bias (focusing on the \( \text{foul.diff} \) variable).

One way, then, of accounting for teams is by treating them as random effects rather than fixed effects. For instance, we can assume the effect that Minnesota being the home team has on the log odds of foul call on the home team is drawn from a normal distribution with mean 0 and variance \( \sigma^2_{0,h} \). Maybe the estimated random effect for Minnesota as the home team is -0.5, so that Minnesota is slightly less likely that the typical home team to have fouls called on it, and the odds of a foul on the home team are somewhat correlated whenever Minnesota is the home team. In this way, we account for the complete range of team effects, while only having to estimate a single parameter \( (\sigma^2_{0,h}; \lambda_h) \) rather than coefficients for 38 indicator variables. The effect of visiting teams can be similarly modeled.

How will treating home and visiting teams as random effects change our multilevel model? Another way we might view this situation is by considering that Game is not the only Level Two observational unit we might have selected. What if we instead decided to focus on Home Team as the Level Two observational unit? That is, what if we assumed that fouls called on the same home team across all games must be correlated? In this case, we could redefine our Level One model from Equation 11.1. Let \( Y_{hj} \) be an indicator variable recording if the \( j^{th} \) foul from Home Team \( h \) was called on the home team (1) or the visiting team (0), and \( p_{hj} \) be the true probability that the \( j^{th} \) foul from Home Team \( h \) was called on the home team. Now, if we were to consider a simple model with foul differential as the sole predictor, we could model the probability of a foul on the home team for Home Team \( h \) with the model:

\[
\log\left( \frac{p_{hj}}{1 - p_{hj}} \right) = \lambda_{0h} + \lambda_{1h}\text{foul.diff}_{hj}
\] (11.2)

In this case, \( e^{\lambda_{0h}} \) represents the odds that a foul is called on the home team when total fouls are equal between both teams in a game involving Home Team \( h \), and \( e^{\lambda_{1h}} \) represents the multiplicative change in the odds that a foul is called on the home team for every extra foul on the home team compared to the visitors in a game involving Home Team \( h \). After fitting logistic regression models for each of the 39 teams in our data set, we see in Figure 11.5 variability in fitted intercepts (mean=-0.15, sd=0.33) and slopes (mean=-0.22, sd=0.12) among the 39 teams, although much less variability than we observed from game-to-game. Of course, each logistic regression model for a home team was based on about 10 times more foul calls than each model for a game, so observing less variability from team-to-team was not unexpected.

From a modeling perspective, accounting for clustering by game and by home team (not to mention by visiting team) brings up an interesting issue we have not yet considered—can we handle random effects that are not nested? Since each foul called is associated with only one game (or only one home team and one visiting team), foul is considered nested in game (or home or visiting team). However, a specific home team is not associated with a single game; that home team appears in several games. Therefore, any effects of game, home team, and visiting team are considered crossed random effects.

A two-level model which accounts for variability among games, home teams, and visiting teams
would take on a slightly new look. First, the full subscripting would change a bit. Our primary response variable would now be written as $Y_{i[gh]j}$, an indicator variable recording if the $j^{th}$ foul from Game $i$ was called on the home team (1) or the visiting team (0), where Game $i$ pitted Visiting Team $g$ against Home Team $h$. Square brackets are introduced since $g$ and $h$ are essentially at the same level as $i$, whereas we have assumed (without stating so) throughout this book that subscripting without square brackets implies a movement to lower levels as the subscripts move left to right (e.g., $ij$ indicates $i$ units are at Level Two, while $j$ units are at Level One, nested inside Level Two units). We can then consider $Y_{i[gh]j}$ to be a Bernoulli random variable with parameter $p_{i[gh]j}$, where $p_{i[gh]j}$ is the true probability that the $j^{th}$ foul from Game $i$ was called on Home Team $h$ rather than Visiting Team $g$. We will include the crossed subscripting only where necessary.

Typically, with the addition of crossed effects, the Level One model will remain familiar and changes will be seen at Level Two, especially in the equation for the intercept term. In the model formulation below we allow, as before, the slope and intercept to vary by game:

- **Level One:**

  \[
  \log\left(\frac{p_{i[gh]j}}{1-p_{i[gh]j}}\right) = \lambda_{0i} + \lambda_{1i}\text{foul.diff}_{ij}
  \]  

  \[11.3\]

- **Level Two:**

  $\lambda_{0i} = \beta_{00} + \epsilon_{0i} + \eta_{0h} + v_{0g}$

  $\lambda_{1i} = \beta_{10}$

Therefore, at Level Two, we assume that $\lambda_{0i}$, the log odds of a foul on the home team when the home and visiting teams in Game $i$ have an equal number of fouls, depends on four components:

- $\beta_{00}$ is the population average across all games and fouls (fixed)
- $\epsilon_{0i}$ is the effect of Game $i$ (random)
- $\eta_{0h}$ is the effect of Home Team $h$ (random)
- $v_{0g}$ is the effect of Visiting Team $g$ (random)

where error terms (random effects) at Level Two can be assumed to follow independent normal
distributions:

\[ \varepsilon_{0i} \sim N(0, \sigma_{0,\varepsilon}^2) \]
\[ \eta_{0h} \sim N(0, \sigma_{0,\eta}^2) \]
\[ v_{0g} \sim N(0, \sigma_{0,v}^2) \].

We could include terms that vary by home or visiting team in other Level Two equations, but often adjusting for these random effects on the intercept is sufficient. The advantages to including additional random effects are three-fold. First, by accounting for additional sources of variability, we should obtain more precise estimates of the other model parameters, including key fixed effects. Second, we obtain estimates of variance components, allowing us to compare the relative sizes of game-to-game and team-to-team variability. Third, as outlined in Section 11.8, we can obtain estimated random effects which allow us to compare the effects on the log-odds of a home foul of individual home and visiting teams.

Our composite model then looks like:

\[
\log\left(\frac{p_{i[gh]j}}{1 - p_{i[gh]j}}\right) = [\beta_{00} + \beta_{10}\text{foul}.\text{diff}_{ij}] + [\varepsilon_{0i} + \eta_{0h} + v_{0g}] .
\]

We will refer to this as Model A, where we look at the effect of foul differential on the odds a foul is called on the home team, while accounting for three crossed random effects at Level Two (game, home team, and visiting team). Parameter estimates for Model A are given below:

### Model A ###

Formula: foul.home ~ foul.diff + (1 | game) + (1 | hometeam) + (1 | visitor)

<table>
<thead>
<tr>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>6780</td>
<td>6813</td>
<td>-3385</td>
<td>6770</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>game</td>
<td>(Intercept)</td>
<td>0.171651</td>
<td>0.41431</td>
</tr>
<tr>
<td>visitor</td>
<td>(Intercept)</td>
<td>0.023235</td>
<td>0.15243</td>
</tr>
<tr>
<td>hometeam</td>
<td>(Intercept)</td>
<td>0.068098</td>
<td>0.26096</td>
</tr>
</tbody>
</table>

Number of obs: 4972, groups: game, 340; visitor, 39; hometeam, 39

Fixed effects:

|                   | Estimate | Std. Error | z value | Pr(>|z|) |
|-------------------|----------|------------|---------|----------|
| (Intercept)       | -0.18781 | 0.06136    | -3.061  | 0.00221  ** |
| foul.diff         | -0.26388 | 0.01710    | -15.433 | < 2e-16  *** |

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

- \( \hat{\beta}_{00} = -0.188 = \) the mean log odds of a home foul at the point where total fouls are equal between teams. In other words, when fouls are balanced between teams, the odds that a foul is called on the visiting team is 20.7% higher than the odds a foul is called on the home team (\(1/e^{-0.188} = 1.207\)).

- \( \hat{\beta}_{10} = -0.264 = \) the decrease in mean log odds of a home foul for each 1 foul increase in the foul differential. More specifically, the odds the next foul is called on the visiting team rather than the home team increases by 30.2% with each additional foul called on the home team (\(1/e^{-0.264} = 1.302\)).

- \( \sigma_{0,\varepsilon}^2 = 0.172 = \) the variance in intercepts from game-to-game
Based on the t-value (-15.433) and p-value ($p < .001$) associated with foul differential in this model, we have significant evidence of a negative association between foul differential and the odds of a home team foul. That is, we have significant evidence that the odds that a foul is called on the home team shrinks as the home team has more total fouls compared with the visiting team. Thus, there seems to be preliminary evidence in the 2009-10 data that college basketball referees tend to even out foul calls over the course of the first half. Of course, we have yet to adjust for other significant covariates.

11.6 Model Comparisons Using the Parametric Bootstrap

Our estimates of variance components provide evidence of the relative variability in the different Level Two random effects. For example, an estimated 65.4% of variability in the intercepts is due to differences from game-to-game, while 25.9% is due to differences among home teams, and 8.7% is due to differences among visiting teams. At this point, we could reasonably ask: if we use a random effect to account for differences among games, does it really pay off to also account for which team is home and which is the visitor?

To answer this, we could compare models with and without the random effects for home and visiting teams (i.e., Model A, the full model, vs. Model A.0, the reduced model) using a likelihood ratio test. In Model A.0, Level Two now looks like:

$$
\lambda_{0i} = \beta_{00} + \epsilon_{0i}
$$
$$
\lambda_{1i} = \beta_{10}
$$

The likelihood ratio test (see below) provides significant evidence (LRT=16.074, df=2, $p=.0003$) that accounting for variability among home teams and among visiting teams improves our model.

Models:

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A.0: foul.home ~ foul.diff + (1</td>
<td>game)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: foul.home ~ foul.diff + (1</td>
<td>game) + (1</td>
<td>hometeam) + (1</td>
<td>visitor)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Df</td>
<td>AIC</td>
<td>BIC</td>
<td>logLik</td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>A.0</td>
<td>3</td>
<td>6792.5</td>
<td>6812.1</td>
<td>-3393.3</td>
</tr>
<tr>
<td>A</td>
<td>5</td>
<td>6780.5</td>
<td>6813.0</td>
<td>-3385.2</td>
</tr>
</tbody>
</table>

In Section 8.5.4 we noted that REML estimation is typically used when comparing models that differ only in random effects for multilevel models with normal responses, but with generalized linear multilevel models full maximum likelihood (ML) estimation procedures are typically used regardless of the situation (including in the function glmer() in R). So the likelihood ratio test above is based on ML methods—can we trust its accuracy? In Section 10.6 we introduced the parametric bootstrap as a more reliable method in many cases for conducting hypothesis tests compared to the likelihood ratio test, especially when comparing full and reduced models that differ in their variance components. What would a parametric bootstrap test say about testing $H_0: \sigma^2_{0,\eta} = \sigma^2_{0,v} = 0$ vs. $H_A: \text{at least one of } \sigma^2_{0,\eta} \text{ or } \sigma^2_{0,v} \text{ is not equal to } 0$? Under the null hypothesis, since the two variance terms are being set to a value (0) on the boundary constraint, we would not expect the chi-square distribution to adequately approximate the behavior of the likelihood ratio test statistic.

Figure 11.6 illustrates the null distribution of the likelihood ratio test statistic derived by the parametric bootstrap procedure with 1000 samples as compared to a chi-square distribution. As we
observed in Section 10.6, the parametric bootstrap provides a more reliable p-value in this case (< .001 from table below) because a chi-square distribution puts too much mass in the tail and not enough near 0. However, the parametric bootstrap is computationally intensive, and it can take a long time to run even with moderately complex models. With this data, we would select our full Model A based on a parametric bootstrap test.

Parametric bootstrap with 1000 samples.

Models:
A.0: foul.home ~ foul.diff + (1 | game)
A: foul.home ~ foul.diff + (1 | game) + (1 | hometeam) + (1 | visitor)

<table>
<thead>
<tr>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr_boot(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.0</td>
<td>3</td>
<td>6792.5</td>
<td>6812.1</td>
<td>-3393.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>5</td>
<td>6780.5</td>
<td>6813.0</td>
<td>-3385.2</td>
<td>16.074</td>
<td>2</td>
</tr>
</tbody>
</table>

We might also reasonably ask: is it helpful to allow slopes (coefficients for foul differential) to vary by game, home team, and visiting team as well? Again, since we are comparing models that differ in random effects, and since the null hypothesis involves setting random effects at their boundaries, we use the parametric bootstrap. Formally, we are comparing Model A to Model A.1, which has the same Level One equation as Model A:

\[
\log \left( \frac{P_{ghij}}{1 - P_{ghij}} \right) = \lambda_{0i} + \lambda_{1i} \text{foul.diff}_{ij}
\]

but more variance components to estimate at Level Two:

\[
\begin{align*}
\lambda_{0i} & = \beta_{00} + \varepsilon_{0i} + \eta_{0i} + \omega_{0i} \\
\lambda_{1i} & = \beta_{10} + \varepsilon_{1i} + \eta_{1i} + \nu_{1i}
\end{align*}
\]

where error terms (random effects) at Level Two can be assumed to follow independent normal distributions:

\[
\varepsilon_{0i} \sim N(0, \sigma_{0i}^2)
\]

Figure 11.6 *Null distribution of likelihood ratio test statistic comparing Models A and A.0 derived using parametric bootstrap with 1000 samples (histogram) compared to a chi-square distribution with 2 degrees of freedom (smooth curve).*
Thus our null hypothesis for comparing Model A vs. Model A.1 is

$$
H_0: \sigma^2_{e_i} = \sigma^2_{\eta_{ih}} = \sigma^2_{\eta_{ig}} = 0.
$$

We do not have significant evidence (LRT=0.349, df=3, p=.437 by parametric bootstrap) of variability among slopes, so we will only include random effects for game, home team, and visiting team for the intercept going forward. Figure 11.7 illustrates the null distribution of the likelihood ratio test statistic derived by the parametric bootstrap procedure as compared to a chi-square distribution, again showing that the tails are too heavy in the chi-square distribution.

![Figure 11.7](image_url)

**Figure 11.7** Null distribution of likelihood ratio test statistic comparing Models A and A.1 derived using parametric bootstrap with 1000 samples (histogram) compared to a chi-square distribution with 3 degrees of freedom (smooth curve).

### Model A.1 ###

Formula: foul.home ~ foul.diff + (1 | game) + (1 | hometeam) +

(1 | visitor) + (0 + foul.diff | game) + (0 + foul.diff | hometeam) +

(0 + foul.diff | visitor)

<table>
<thead>
<tr>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>6786</td>
<td>6838</td>
<td>-3385</td>
<td>6770</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>game</td>
<td>foul.diff</td>
<td>7.102e-05</td>
<td>8.4274e-03</td>
</tr>
<tr>
<td>game</td>
<td>(Intercept)</td>
<td>1.6982e-01</td>
<td>4.1209e-01</td>
</tr>
<tr>
<td>visitor</td>
<td>foul.diff</td>
<td>1.2584e-11</td>
<td>3.5474e-06</td>
</tr>
<tr>
<td>hometeam</td>
<td>foul.diff</td>
<td>1.3993e-03</td>
<td>3.7407e-02</td>
</tr>
<tr>
<td>visitor</td>
<td>(Intercept)</td>
<td>2.3307e-02</td>
<td>1.5267e-01</td>
</tr>
</tbody>
</table>
A POTENTIAL FINAL MODEL FOR EXAMINING REFEREE BIAS

hometeam (Intercept) 6.7347e-02 2.5951e-01
Number of obs: 4972, groups: game, 340; visitor, 39; hometeam, 39

Fixed effects:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|---------|
| (Intercept) -0.1853 0.0613 -3.023 0.0025 ** |
| foul.diff -0.2668 0.0182 -14.663 <2e-16 *** |

Parametric bootstrap with 1000 samples.

Models:
mA: foul.home ~ foul.diff + (1 | game) + (1 | hometeam) + (1 | visitor)
mA1: foul.home ~ foul.diff + (1 | game) + (1 | hometeam) +
(1 | visitor) + (0 + foul.diff | game) + (0 + foul.diff | hometeam) +
(0 + foul.diff | visitor)

Df AIC BIC logLik Chisq Chi Df Pr_boot(>Chisq)
mA 5 6780.5 6813.0 -3385.2
mA1 8 6786.1 6838.2 -3385.1 0.349 3 0.437

Note that we could have also allowed for a correlation between the error terms for the intercept and slope by game, home team, or visiting team – i.e., assume, for example:

\[
\begin{bmatrix}
\varepsilon_{0i} \\
\varepsilon_{1i}
\end{bmatrix}
\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\
\sigma_{01} & \sigma_1^2 \end{bmatrix}\right)
\]

while error terms by game, home team, or visiting team are still independent. Here, the new model would have 6 additional parameters when compared with Model A (3 variance terms and 3 covariance terms). By the parametric bootstrap, there is no significant evidence that the model with 6 additional parameters is necessary (LRT=6.49, df=6, p=.093 by parametric bootstrap). The associated p-value based on a likelihood ratio test with approximate chi-square distribution (and restricted maximum likelihood estimation) is .370, reflecting once again the overly heavy tails of the chi-square distribution.

11.7 A Potential Final Model for Examining Referee Bias

In constructing a final model for this college basketball case study, we are guided by several considerations. First, we want to estimate the effect of foul differential on the odds of a home foul, after adjusting for important covariates. Second, we wish to find important interactions with foul differential, recognizing that the effect of foul differential might depend on the game situation and other covariates. Third, we want to account for random effects associated with game, home team, and visiting team. What follows is one potential final model that follows these guidelines:

- **Level One:**

  \[
  \log\left(\frac{p_{ij|h,j}}{1-p_{ij|h,j}}\right) = \lambda_0 + \lambda_{1i} \text{foul.diff}_{ij} + \lambda_{2i} \text{score.diff}_{ij} + \lambda_{3i} \text{lead.home}_{ij} + \lambda_{4i} \text{time}_{ij} + \lambda_{5i} \text{offensive}_{ij} + \lambda_{6i} \text{personal}_{ij} + \lambda_{7i} \text{foul.diff*offensive}_{ij} + \lambda_{8i} \text{foul.diff*personal}_{ij} + \lambda_{9i} \text{foul.diff*time}_{ij} + \lambda_{10i} \text{lead.home*time}_{ij}
  \]

- **Level Two:**

  \[
  \lambda_{0i} = \beta_{00} + \varepsilon_{0i} + \eta_{0i} + \nu_{0i}
  \]
\[ \lambda_{1i} = \beta_{10} \]
\[ \lambda_{2i} = \beta_{20} \]
\[ \lambda_{3i} = \beta_{30} \]
\[ \lambda_{4i} = \beta_{40} \]
\[ \lambda_{5i} = \beta_{50} \]
\[ \lambda_{6i} = \beta_{60} \]
\[ \lambda_{7i} = \beta_{70} \]
\[ \lambda_{8i} = \beta_{80} \]
\[ \lambda_{9i} = \beta_{90} \]
\[ \lambda_{10i} = \beta_{10,0} \]

where error terms at Level Two can be assumed to follow independent normal distributions:
\[ \varepsilon_{0h} \sim N(0, \sigma_{0,e}^2) \]
\[ \eta_{0h} \sim N(0, \sigma_{0,\eta}^2) \]
\[ v_{0g} \sim N(0, \sigma_{0,v}^2) \]

Our composite model then looks like:
\[
\log \left( \frac{p_{ij|hj}}{1 - p_{ij|hj}} \right) = (\beta_{00} + \beta_{10}\text{foul.diff}_{ij} + \beta_{20}\text{score.diff}_{ij} + \beta_{30}\text{lead.home}_{ij} + \beta_{40}\text{time}_{ij} + \beta_{50}\text{offensive}_{ij} + \beta_{60}\text{personal}_{ij} + \beta_{70}\text{foul.diff}\times\text{offensive}_{ij} + \beta_{80}\text{foul.diff}\times\text{personal}_{ij} + \beta_{90}\text{foul.diff}\times\text{time}_{ij} + \beta_{10,0}\text{lead.home}\times\text{time}_{ij} + \varepsilon_{0h} + \eta_{0h} + v_{0g})
\]

Using the composite form of this generalized linear multilevel model, the parameter estimates for our 11 fixed effects and 3 variance components are given in the output below:

### Model F ###

Formula: foul.home ~ foul.diff + score.diff + lead.home + time + offensive + personal + foul.diff:offensive + foul.diff:personal + foul.diff:time + lead.home:time + (1 | game) + (1 | hometeam) + (1 | visitor)

<table>
<thead>
<tr>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
</tr>
</thead>
<tbody>
<tr>
<td>6731</td>
<td>6822</td>
<td>-3352</td>
<td>6703</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>game</td>
<td>(Intercept)</td>
<td>0.184850 0.42994</td>
</tr>
<tr>
<td>visitor</td>
<td>(Intercept)</td>
<td>0.043151 0.20773</td>
</tr>
<tr>
<td>hometeam</td>
<td>(Intercept)</td>
<td>0.078336 0.27989</td>
</tr>
</tbody>
</table>

Number of obs: 4972, groups: game, 340; visitor, 39; hometeam, 39

Fixed effects:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | -0.246814 | 0.129518 | -1.906  0.05670 . |
| foul.diff  | -0.171648 | 0.003563 | -5.114 3.15e-07 *** |
| score.diff | 0.033507  | 0.007795 | 4.299 1.72e-05 *** |
| lead.home  | -0.150105 | 0.174695 | -0.859 0.39021 |
| time       | -0.008714 | 0.008386 | -1.039 0.29877 |
In general, we see a highly significant negative effect of foul differential—a strong tendency for referees to even out foul calls when one team starts amassing more fouls than the other. Important covariates to control for (because of their effects on the odds of a home foul) include score differential, whether the home team held the lead, time left in the first half, and the type of foul called.

Furthermore, we see that the effect of foul differential depends on type of foul called and time left in the half—the tendency for evening out foul calls is stronger earlier in the half, and when offensive and personal fouls are called instead of shooting fouls. The effect of foul type supports the hypothesis that if referees are consciously or subconsciously evening out foul calls, the behavior will be more noticeable for calls over which they have more control, especially offensive fouls (which are notorious judgment calls) and then personal fouls (which don’t affect a player’s shot, and thus a referee can choose to let them go uncalled). Evidence like this can be considered dose response, since higher “doses” of referee control are associated with a greater effect of foul differential on their calls. A dose response effect provides even stronger indication of referee bias.

Analyses of data from 2004-05 (Noecker and Roback, 2013) showed that the tendency to even out foul calls was stronger when one team had a large lead, but we found no evidence of a foul differential by score differential interaction in the 2009-10 data, although home team fouls are more likely when the home team has a large lead, regardless of the foul differential.

Here are specific interpretations of key model parameters:

- \( \exp(\hat{\beta}_{0}) = \exp(-0.247) = 0.781 \). The odds of a foul on the home team is 0.781 at the end of the first half when the score is tied, the fouls are even, and the referee has just called a shooting foul. In other words, only 43.9% of shooting fouls in those situations will be called on the home team.
- \( \exp(\hat{\beta}_{1}) = \exp(-0.172) = 0.842 \). Also, 0.842\(^{-1} = 1.188 \). As the foul differential decreases by 1 (the visiting team accumulates another foul relative to the home team), the odds of a home foul increase by 18.8%. This interpretation applies to shooting fouls at the end of the half, after controlling for the effects of score differential and whether or not the home team has the lead.
- \( \exp(\hat{\beta}_{2}) = \exp(0.034) = 1.034 \). As the score differential increases by 1 (the home team accumulates another point relative to the visiting team), the odds of a home foul increase by 3.4%, after controlling for foul differential, type of foul, whether or not the home team has the lead, and time remaining in the half. Referees are more likely to call fouls on the home team when the home team is leading, and vice versa. Note that a change in the score differential could result in the home team gaining the lead, so that the effect of score differential experiences a non-linear “bump” at 0, where the size of the bump depends on the time remaining (this would involve the interpretation for \( \hat{\beta}_{1} \)).
- \( \exp(\hat{\beta}_{3}) = \exp(-0.103) = 0.902 \). Also, 0.902\(^{-1} = 1.109 \). The effect of foul differential increases by 10.9% if a foul is an offensive foul rather than a shooting foul, after controlling for score differential, whether the home team has the lead, and time remaining. As hypothesized, the effect of foul differential is greater for offensive fouls, over which referees have more control when compared with shooting fouls. For example, midway through the half (time=10), the odds that a shooting foul is on the home team increase by 29.6% for each extra foul on the visiting team, while the odds that an offensive foul is on the home team increase by 43.6%.
- \( \exp(\hat{b}_{80}) = \exp(-0.056) = 0.946 \). Also, \( 0.946^{-1} = 1.057 \). The effect of foul differential increases by 5.7% if a foul is a personal foul rather than a shooting foul, after controlling for score differential, whether the home team has the lead, and time remaining.

- \( \exp(\hat{b}_{90}) = \exp(-0.0087) = 0.991 \). Also, \( 0.991^{-1} = 1.009 \). The effect of foul differential increases by 0.9% for each extra minute that is remaining in the half, after controlling for foul differential, score differential, whether the home team has the lead, and type of foul. Thus, the tendency to even out foul calls is strongest earlier in the game. For example, midway through the half (time=10), the odds that a shooting foul is on the home team increase by 29.6% for each extra foul on the visiting team, while at the end of the half (time=0) the odds increase by 18.8% for each extra visiting foul.

- \( \sigma^2_{0,e} = 0.185 \) = the variance in log-odds intercepts from game-to-game after controlling for all other covariates in the model.

11.8 Estimated Random Effects

Our final model includes random effects for game, home team, and visiting team, and thus our model accounts for variability due to these three factors without estimating fixed effects to represent specific teams or games. However, we may be interested in examining the relative level of the random effects used for different teams or games. The presence of random effects allows different teams or games to begin with different baseline odds of a home foul. In other words, the intercept—the log odds of a home foul at the end of the first half when the score is tied, the fouls are even, and a shooting foul has been called—is adjusted by a random effect specific to the two teams playing and to the game itself. Together, the estimated random effects for, say, home team, should follow a normal distribution as we assumed: centered at 0 and with variance given by \( \sigma^2_{0,h} \). It is sometimes interesting to know, then, which teams are typical (with estimated random effects near 0) and which fall in the upper and lower tails of the normal distribution.

Figure 11.8 shows the 39 estimated random effects associated with each home team (see the discussion on Empirical Bayes estimates in Chapter 8). As expected, the distribution is normal with standard deviation near \( \hat{\sigma}_{0,h} = 0.28 \). To compare individual home teams, Figure 11.9 shows the estimated random effect and associated prediction interval for each of the 39 home teams. Although there is a great deal of uncertainty surrounding each estimate, we see that, for instance, DePaul and Seton Hall have higher baseline odds of home fouls than Purdue or Syracuse. Similar histograms and prediction intervals plots can be generated for random effects due to visiting teams and specific games.

11.9 Notes on Using R (Optional)

The R code below fits Model A.1 from Section 11.6. Note that, in glmer() or lmer(), if you have two equations at Level Two and want to fit a model with an error term for each equation, but you also want to assume that the two error terms are independent, the error terms must be requested separately. For example, “(1|hometeam)” allows the Level One intercept to vary by home team, while “(0+foul.diff|hometeam)” allows the Level One effect of foul.diff to vary by home team. Under this formulation, the correlation between those two error terms is assumed to be 0; a non-zero correlation could have been specified with “(1+foul.diff|hometeam)”.

```R
library(lme4)

# Model A.1 (Multilevel model with only foul.diff and all 3 random effects
# applied to both intercepts and slopes but no correlations)
model.a3 <- glmer(foul.home ~ foul.diff + (1|game) +
```
Figure 11.8: Histogram of estimated random effects for 39 home teams in Model F.

Figure 11.9: Estimated random effects and associated prediction intervals for 39 home teams in Model F.

(1|hometeam)+(1|visitor) + (0+foul.diff|game)+(0+foul.diff|hometeam)+(0+foul.diff|visitor),family=binomial,data=refdata)

The R code below shows how fixef() can be used to extract the estimated fixed effects from a multilevel model. Even more, it shows how ranef() can be used to illustrate estimated random effects by Game, Home Team, and Visiting Team, along with prediction intervals for those random effects. These estimated random effects are sometimes called Empirical Bayes estimators. In this case, random effects are placed only on the “[(Intercept)]” term; the phrase “Intercept” could be replaced with other Level One covariates whose values are allowed to vary by game, home team, or visiting team in our model.
# Model F - final model?
model.f=glmer(foul.home~foul.diff+score.diff+lead.home+time+offensive+
          personal+foul.diff:offensive+foul.diff:personal+foul.diff:time+
          lead.home:time+(1|game)+(1|hometeam)+(1|visitor),family=binomial,
          data=refdata)
summary(model.f)
exp(fixef(model.f))

# Get estimated random effects based on Model F
re.int = ranef(model.f)$'game'[['(Intercept)']]
hist(re.int,xlab="Random Effect",main="Random Effects for Game")
Home.re=ranef(model.f)$'hometeam'[['(Intercept)']]
hist(Home.re,xlab="Random Effect",main="Random Effects for Home Team")
Visiting.re=ranef(model.f)$'visitor'[['(Intercept)']]
hist(Visiting.re,xlab="Random Effect",
     main="Random Effects for the Visiting Team",xlim=c(-0.5,0.5))
cbind(Home.re,Visiting.re) # 39x2 matrix of REs by team

# Prediction intervals for random effects based on Model F
ranef1=dotplot(ranef(model.f, postVar = TRUE), strip = FALSE)
print(ranef1[[3]], more = TRUE) # HOME
print(ranef1[[2]], more = TRUE) # VIS
print(ranef1[[1]], more = TRUE)

11.10 Exercises

11.10.1 Conceptual Exercises

1. Give an example of a data set and an associated research question that would best be addressed with a multilevel model for a Poisson response.

2. **College Basketball Referees.** Explain to someone unfamiliar with the plots in Figure 11.2 how to read both a conditional density plot and an empirical logit plot. For example, explain what the dark region in a conditional density plot represents, what each point in an empirical logit plot represents, etc.

3. With the strength of the evidence found in Figure 11.2, plots (a) and (d), is there any need to run statistical models to convince someone that referee tendencies are related to the foul differential?

4. In Section 11.4.2, why don’t we simply run a logistic regression model for each of the 340 games, collect the 340 intercepts and slopes, and fit a model to those intercepts and slopes?

5. Explain in your own words the difference between crossed and nested random effects (Section 11.5).

6. In the context of Case Study 9.2, describe a situation in which crossed random effects would be needed.

7. Assume that we added $\epsilon_{i1}$, $\eta_{i1}$, and $v_{i1}$ to the Level Two equation for $l_{ij}$ in Equations 11.3. (a) Give interpretations for those 3 new random effects. (b) How many additional parameters would need to be estimated (and name the new model parameters)?

8. In Section 11.6, could we use a likelihood ratio test to determine if it would be better to add either a random effect for home team or a random effect for visiting team to the model with a random effect for game (assuming we’re going to add one or the other)? Why or why not?
9. Describe how we would obtain the 1000 values representing the parametric bootstrap distribution in the histogram in Figure 11.6.

10. In Figure 11.6, why is it a problem that “a chi-square distribution puts too much mass in the tail” when using a likelihood ratio test to compare models?

11. What would be the implications of using the R expression “(foul.diff | game) + (foul.diff | home-team) + (foul.diff | visitor)” in Model A.1 (Section 11.6) to obtain our variance components?

12. Explain the implications of having no error terms in any Level Two equations in Section 11.7 except the one for $\lambda_0$.

13. In the interpretation of $\hat{\beta}_{10}$, explain why this applies to “shooting fouls at the end of the half”. Couldn’t we just say that we controlled for type of foul and time elapsed?

14. In the interpretation of $\hat{\beta}_{20}$, explain the idea that “the effect of score differential experiences a non-linear ‘bump’ at 0, where the size of the bump depends on time remaining.” Consider, for example, the effect of a point scored by the home team with 10 minutes left in the first half, depending on whether the score is tied or the home team in ahead by 2.

15. In the interpretation of $\hat{\beta}_{70}$, verify the odds increases of 29.6% for shooting fouls and 43.6% for offensive fouls. Where does the stated 10.9% increase factor in?

16. We could also interpret the interaction between two quantitative variables described by $\hat{\beta}_{90}$ as “the effect of time remaining increases by 0.9% for each extra foul on the visiting team, after controlling for...” Numerically illustrate this interpretation by considering foul differentials of 2 (the home team has 2 more fouls than the visitors) and -2 (the visitors have 2 more fouls than the home team).

17. Provide interpretations in context for $\hat{\beta}_{40}$, $\hat{\beta}_{50}$, and $\hat{\beta}_{10.0}$.

18. In Section 11.8, why isn’t the baseline odds of a home foul for DePaul considered a model parameter?

19. **Heart attacks in Aboriginal Australians.** Randall, et al., published a 2014 article in *Health and Place* entitled “Exploring disparities in acute myocardial infarction events between Aboriginal and non-Aboriginal Australians: Roles of age, gender, geography and area-level disadvantage.” They used multilevel Poisson models to compare rates of acute myocardial infarction (AMI) in the 199 Statistical Local Areas (SLAs) in New South Wales. Within SLA, AMI rates (number of events over population count) were summarized by subgroups determined by year (2002-2007), age (in 10-year groupings), sex, and Aboriginal status; these are then our Level One variables. Analyses also incorporated remoteness (classified by quartiles) and socio-economic status (classified by quintiles) assessed at the SLA level (Level Two). For this study, give the observational units at Levels One and Two.

20. Table 11.4 shows Table 2 from Randall et al. (2014). Let $Y_{ij}$ be the number of acute myocardial infarctions in subgroup $j$ from SLA $i$; write out the multilevel model that likely produced Table 11.4. How many fixed effects and variance components must be estimated?

21. Provide interpretations in context for the bolded rate ratios, confidence intervals, and p-values in Table 11.4.

22. Given the rate ratio and 95% confidence interval reported for Aboriginal Australians in Table 11.4, find the estimated model fixed effect for Aboriginal Australians from the multilevel model along with its standard error.

23. How might the p-value for Age have been produced?

24. Randall et al. report that “we identified a significant interaction between Aboriginal status and age group ($p < 0.01$) and Aboriginal status and sex ($p < 0.01$), but there was no significant interaction between Aboriginal status and year ($p=0.94$).” How would the multilevel model associated with Table 11.4 need to have been adjusted to allow these interactions to be tested?
Table 11.4 *Adjusted rate ratios for individual-level variables from the multilevel Poisson regression model with random intercept for area from Table 2 in Randall et al. (2014).*

<table>
<thead>
<tr>
<th>Variable</th>
<th>RR</th>
<th>95% CI</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aboriginal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No (ref)</td>
<td>1.00</td>
<td></td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Yes</td>
<td>2.10</td>
<td>1.98-2.23</td>
<td></td>
</tr>
<tr>
<td>Age group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25-34 (ref)</td>
<td>1.00</td>
<td></td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>35-44</td>
<td>6.01</td>
<td>5.44-6.64</td>
<td></td>
</tr>
<tr>
<td>45-54</td>
<td>19.36</td>
<td>17.58-21.31</td>
<td></td>
</tr>
<tr>
<td>55-64</td>
<td>40.29</td>
<td>36.67-44.26</td>
<td></td>
</tr>
<tr>
<td>65-74</td>
<td>79.92</td>
<td>72.74-87.80</td>
<td></td>
</tr>
<tr>
<td>75-84</td>
<td>178.75</td>
<td>162.70-196.39</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male (ref)</td>
<td>1.00</td>
<td></td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Female</td>
<td>0.45</td>
<td>0.44-0.45</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002 (ref)</td>
<td>1.00</td>
<td></td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>2003</td>
<td>1.00</td>
<td>0.98-1.03</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>0.97</td>
<td>0.95-0.99</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>0.91</td>
<td>0.89-0.94</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>0.88</td>
<td>0.86-0.91</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>0.88</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

25. Table 11.5 shows Table 3 from Randall et al. (2014). Describe the changes to the multilevel model for Table 11.4 that likely produced this new table.

26. Provide interpretations in context for the bolded rate ratios, confidence intervals, and p-values in Table 11.5.

Table 11.5 *Adjusted rate ratios for area-level variables from the multilevel Poisson regression model with random intercept for area from Table 3 in Randall et al. (2014).*

<table>
<thead>
<tr>
<th>Variable</th>
<th>RR</th>
<th>95% CI</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remoteness of residence*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major city</td>
<td>1.00</td>
<td></td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Inner regional</td>
<td>1.16</td>
<td>1.04-1.28</td>
<td></td>
</tr>
<tr>
<td>Outer regional</td>
<td>1.11</td>
<td>1.01-1.23</td>
<td></td>
</tr>
<tr>
<td>Remote/very remote</td>
<td>1.22</td>
<td>1.02-1.45</td>
<td></td>
</tr>
<tr>
<td>SES quintile*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 least disadvantaged</td>
<td>1.00</td>
<td></td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>2</td>
<td>1.26</td>
<td>1.11-1.43</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.40</td>
<td>1.24-1.58</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.46</td>
<td>1.30-1.64</td>
<td></td>
</tr>
<tr>
<td>5 most disadvantaged</td>
<td>1.70</td>
<td>1.52-1.91</td>
<td></td>
</tr>
</tbody>
</table>

* Area-level factors added one at a time to the fully adjusted individual-level model (adjusted for Aboriginal status, age, sex and year) due to being highly associated.

27. Randall et al. also report the results of a single-level Poisson regression model: “After adjusting for age, sex and year of event, the rate of AMI events in Aboriginal people was 2.3 times higher than in non-Aboriginal people (95% CI: 2.17-2.44).” Compare this to the results of the multilevel Poisson model; what might explain any observed differences?

28. Randall et al. claim that “our application of multilevel modelling techniques allowed us to account for clustering by area of residence and produce ‘shrunken’ small-area estimates, which are not as prone to random fluctuations as crude or standardised rates.” Explain what they mean by this statement.
1. **Seed Germination** We will return to the data from Case Study 10.2, but whether or not a seed germinated will be considered the response variable rather than the heights of germinated plants. We will use the wide data set (one row per plant) for this analysis as described in Section 10.3.1, although we will ignore plant heights over time and focus solely on if the plant germinated at any time. Use generalized linear multilevel models to determine the effects of soil type and sterilization on germination rates; perform separate analyses for coneflowers and leadplants, and describe differences between the two species. Support your conclusions with well-interpreted model coefficients and insightful graphical summaries.

2. **Book Banning** The risk of literature censorship is constant, with provocative literature risking being challenged and subsequently being banned. Reasons for censorship have changed over time, with modern day challenges more likely to be for reasons relating to obscenity and child protection rather than overt political and religious objections (Jenkins, 2006). Many past studies have addressed chiefly the reasons listed by those who challenge a book, but few examine the overall context in which books are banned – for example, the characteristics of the states in which they occur.

A team of students (Fast and Hegland, 2011) assembled a data set by starting with information on book challenges from the American Library Society. These book challenges - an initiation of the formal procedure for book censorship - occurred in US States between January of 2000 and November of 2010. In addition, state-level demographic information was obtained from the US Census Bureau and the Political Value Index (PVI) was obtained from the Cook Political Report.

We will consider a data set with 931 challenges and 18 variables. All book challenges over the nearly 11 year period are included except those from the State of Texas; Texas featured 683 challenges over this timeframe, nearly 5 times the number in the next largest state. Thus, the challenges from Texas have been removed and could be analyzed separately. Here, then, is a description of available variables:

- book = unique book ID number
- booktitle = name of book
- author = name of author
- state = state where challenge made
- removed = 1 if book was removed (challenge was successful); 0 otherwise
- pvi2 = state score on the Political Value Index, where positive indicates a Democratic leaning, negative indicates a Republican leaning, and 0 is neutral
- cpherhs = percentage of high school graduates in a state (grand mean centered)
- cmedin = median state income (grand median centered)
- cperba = percentage of college graduates in a state (grand mean centered)
- days2000 = date challenge was made, measured by number of days after January 1, 2000
- obama = 1 if challenge was made while Barack Obama was President; 0 otherwise
- freqchal = 1 if book was written by a frequently challenged author (10 or more challenges across the country); 0 otherwise
- sexexp = 1 if reason for challenge was sexually explicit material; 0 otherwise
- antifamily = 1 if reason for challenge was antifamily material; 0 otherwise
- occult = 1 if reason for challenge was material about the occult; 0 otherwise
- language = 1 if reason for challenge was inappropriate language; 0 otherwise
- homosexuality = 1 if reason for challenge was material about homosexuality; 0 otherwise
• violence = 1 if reason for challenge was violent material; 0 otherwise

The primary response variable is removed. Certain potential predictors are measured at the state level (e.g., pvi2 and cperhs), at least one is at the book level (freqchal), and several are specific to an individual challenge (e.g., obama and sexexp). In addition, note that the same book can be challenged for more than one reason, in different states, or even at different times in the same state.

Perform exploratory analyses and then run multilevel models to examine significant determinants of successful challenges. Write a short report comparing specific reasons for the challenge to the greater context in which a challenge was made.
Index

<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lack-of-fit</td>
<td>130</td>
</tr>
<tr>
<td>alternative covariance structures</td>
<td>212</td>
</tr>
<tr>
<td>Bernoulli process</td>
<td>104</td>
</tr>
<tr>
<td>binomial regression models</td>
<td>106</td>
</tr>
<tr>
<td>boundary constraints</td>
<td>235</td>
</tr>
<tr>
<td>centering covariates</td>
<td>173</td>
</tr>
<tr>
<td>composite model</td>
<td>157</td>
</tr>
<tr>
<td>crossed random effects</td>
<td>271</td>
</tr>
<tr>
<td>dams</td>
<td>129</td>
</tr>
<tr>
<td>deviance residual</td>
<td>113</td>
</tr>
<tr>
<td>dispersion parameter</td>
<td>114</td>
</tr>
<tr>
<td>dose-response</td>
<td>129</td>
</tr>
<tr>
<td>extra-binomial variation</td>
<td>130</td>
</tr>
<tr>
<td>fixed effect</td>
<td>157, 158</td>
</tr>
<tr>
<td>fixed effects</td>
<td>129</td>
</tr>
<tr>
<td>generalized linear multilevel model (GLMM)</td>
<td>262</td>
</tr>
<tr>
<td>intraclass correlation coefficient</td>
<td>137, 164</td>
</tr>
<tr>
<td>Pearson residual</td>
<td>113</td>
</tr>
<tr>
<td>primary sampling units</td>
<td>128</td>
</tr>
<tr>
<td>pseudo R-square value</td>
<td>167</td>
</tr>
<tr>
<td>Quasi-likelihood Model</td>
<td>134</td>
</tr>
<tr>
<td>random effect</td>
<td>158</td>
</tr>
<tr>
<td>random effects</td>
<td>129</td>
</tr>
<tr>
<td>random intercepts model</td>
<td>163</td>
</tr>
<tr>
<td>random slopes and intercepts model</td>
<td>164</td>
</tr>
<tr>
<td>reasons for lack-of-fit for binomial models</td>
<td>113</td>
</tr>
<tr>
<td>restricted maximum likelihood</td>
<td>159</td>
</tr>
<tr>
<td>spaghetti plots</td>
<td>190</td>
</tr>
<tr>
<td>standard covariance structure</td>
<td>210, 249</td>
</tr>
<tr>
<td>tertogen</td>
<td>129</td>
</tr>
<tr>
<td>unconditional growth model</td>
<td>200, 232</td>
</tr>
<tr>
<td>unconditional means model</td>
<td>163, 199, 230</td>
</tr>
<tr>
<td>variance components</td>
<td>159</td>
</tr>
<tr>
<td>maximum likelihood</td>
<td>159</td>
</tr>
<tr>
<td>missing data</td>
<td>186</td>
</tr>
<tr>
<td>multivariate normal distribution</td>
<td>158</td>
</tr>
<tr>
<td>naive model</td>
<td>133</td>
</tr>
<tr>
<td>odds</td>
<td>107</td>
</tr>
<tr>
<td>odds ratio</td>
<td>107</td>
</tr>
<tr>
<td>overdispersion</td>
<td>114, 130</td>
</tr>
<tr>
<td>parametric bootstrap</td>
<td>239, 275</td>
</tr>
</tbody>
</table>