Paul Roback and Julie Legler

GLMs and Multilevel Models

Broadening Your Statistical Horizons with Applications using R

Contents

Preface

GLMs and Multilevel Models: Broadening Your Statistical Horizons (BYSH) with Applications using R is intended to be accessible to undergraduate students who have successfully completed a regression course through, for example, a textbook like *Stat2* [Cannon et al., 2019]. We started teaching this course at St. Olaf in 2003 so students would be able to deal with the non-normal, correlated world we live in. It has been offered at St. Olaf every year since. Even though there is no mathematical prerequisite, we still introduce fairly sophisticated topics such as likelihood theory, zero-inflated Poisson, and parametric bootstrapping in an intuitive and applied manner. We believe strongly in case studies featuring real data and real research questions; thus, most of the data in the textbook and available at our GitHub repo[1] arises from collaborative research conducted by the authors and their students, or from student projects. Our goal is that, after working through this material, students will not necessarily be expert in these methods and associated theory, but that they will develop an expanded toolkit and a greater appreciation for the wider world of data and statistical modeling.

---

[1] https://github.com/proback/BYSH

1

cilia Noecker, Anna Johanson, Nicole Bettes, Kiegan Rice, Anna Wall, Jack Wolf, Josh Pelayo, Spencer Eanes, and Emily Patterson. Early editions of this book also benefitted greatly from feedback from instructors who used these materials in their classes, including Matt Beckman, Laura Boehm Vock, Beth Chance, Laura Chihara, Mine Dogucu, and Katie Ziegler-Graham. Finally, we have appreciated the support of two NSF grants (#DMS-1045015 and #DMS-0354308) and of our colleagues in Mathematics, Statistics, and Computer Science at St. Olaf.

1

Multilevel Data With More Than Two Levels

1.1 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.
- Apply a parametric bootstrap test of significance to appropriate situations with more than two levels.

```
# Packages required for Chapter 10
library(knitr)
library(gridExtra)
library(GGally)
library(mice)
library(nlme)
library(lme4)
library(mnormt)
library(boot)
library(HLMdiag)
library(kableExtra)
library(pander)
library(tidyverse)
```

1.2 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 [Philip 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], Philip 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. 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], Philip et al. [2004]), yet there are few explanations for this phenomenon. For instance, prairies reconstructed in different years (using the same seed combinations and dispersal techniques) have yielded disparate prairie communities.

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 [Angell, 2010]. 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 1.1.
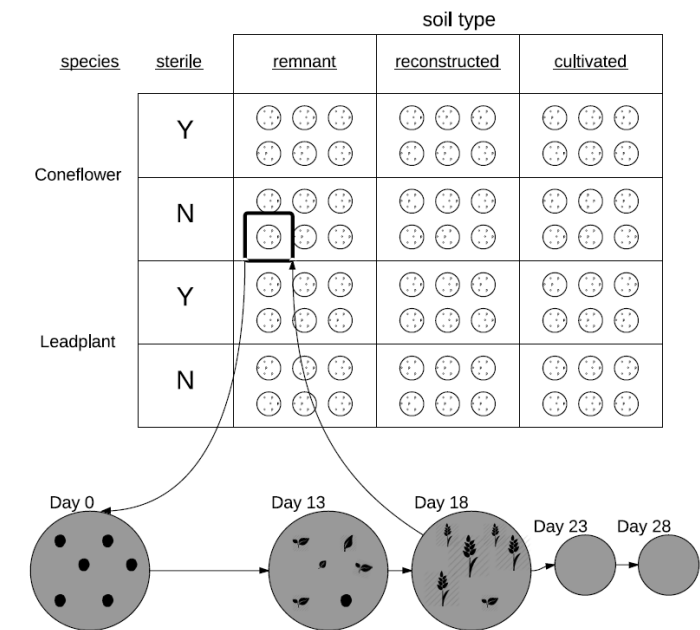

FIGURE 1.1: The design of the seed germination study

1.3 Initial Exploratory Analyses

1.3.1 Data Organization

Data for Case Study 1.2 in `seeds2.csv` contains the following variables:

TABLE 1.1: A snapshot of data (Plants 231-246) from the Seed Germination case study in wide format.

	pot	plant	soil	sterile	species	germin	hgt13	hgt18	hgt23	hgt28
135	23	231	CULT	N	C	Y	1.1	1.4	1.6	1.7
136	23	232	CULT	N	C	Y	1.3	2.2	2.5	2.7
137	23	233	CULT	N	C	Y	0.5	1.4	2.0	2.3
138	23	234	CULT	N	C	Y	0.3	0.4	1.2	1.7
139	23	235	CULT	N	C	Y	0.5	0.5	0.8	2.0
140	23	236	CULT	N	C	Y	0.1	NA	NA	NA
141	24	241	STP	Y	L	Y	1.8	2.6	3.9	4.2
142	24	242	STP	Y	L	Y	1.3	1.7	2.8	3.7
143	24	243	STP	Y	L	Y	1.5	1.6	3.9	3.9
144	24	244	STP	Y	L	Y	NA	1.0	2.3	3.8
145	24	245	STP	Y	L	N	NA	NA	NA	NA
146	24	246	STP	Y	L	N	NA	NA	NA	NA

- `pot` = Pot plant was grown in (1-72)
- `plant` = Unique plant identification number
- `species` = L for leadplant and C for coneflower
- `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 1.1). 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 $72 * 6 = 432$ 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

TABLE 1.2: A snapshot of data (Plants 236-242) from the Seed Germination case study in long format.

pot	plant	soil	sterile	species	germin	hgt	time13
23	236	CULT	N	C	Y	0.1	0
23	236	CULT	N	C	Y	NA	5
23	236	CULT	N	C	Y	NA	10
23	236	CULT	N	C	Y	NA	15
24	241	STP	Y	L	Y	1.8	0
24	241	STP	Y	L	Y	2.6	5
24	241	STP	Y	L	Y	3.9	10
24	241	STP	Y	L	Y	4.2	15
24	242	STP	Y	L	Y	1.3	0
24	242	STP	Y	L	Y	1.7	5
24	242	STP	Y	L	Y	2.8	10
24	242	STP	Y	L	Y	3.7	15

day) and were removed from analysis (for example, see rows 145-146 in Table 1.1). 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 1.2.

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 used thus far.

1.3.2 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 less correlated, we calculate mean height per plant (over all available time points) before generating exploratory plots investigating Level Three factors. Figure 1.2 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.
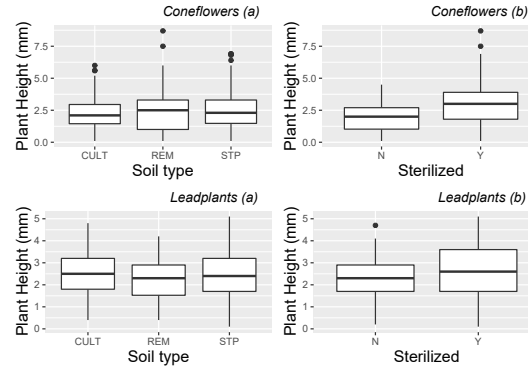
FIGURE 1.2: Plant height comparisons of (a) soil type and (b) sterilization within species. Each plant is represented by the mean height over all measurements at all time points for that plant.

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 1.3 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. Be-
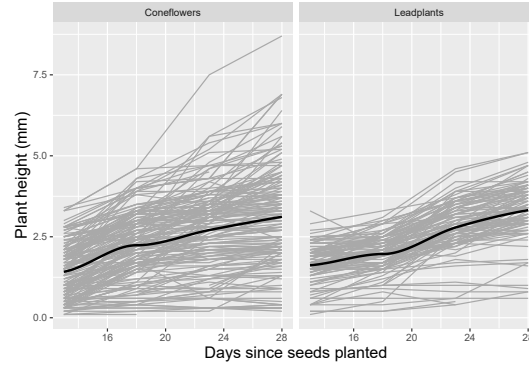
FIGURE 1.3: Spaghetti plot by species with loess fit. Each "line" represents one plant.

cause 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 distinctive 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 1.4 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 is 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 1.5). 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.

Spaghetti plots can also be an effective tool for examining the potential effects of soil type and sterilization on growth patterns of leadplants. Figure 1.6 and Figure 1.7 illustrate how the growth patterns of leadplants depend on soil type
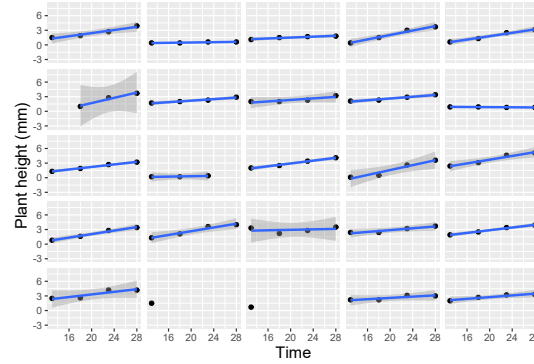
FIGURE 1.4: Lattice plot of linear trends fit to 24 randomly selected lead-plants. One plant with only a single height measurement has no associated regression line.
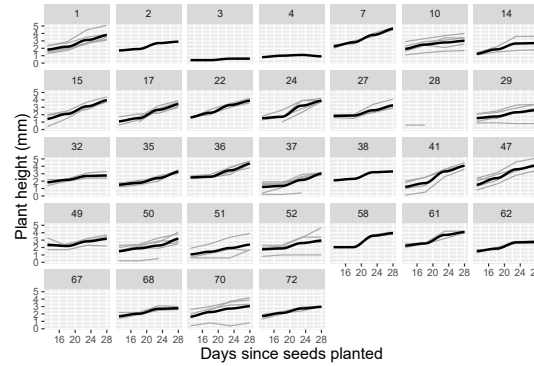
FIGURE 1.5: Spaghetti plot for leadplants by pot with loess fit.

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^2 values. Figures 1.8 and 1.9 provide just such an analysis, where Figure 1.8 shows results of fitting lines by plant, and Figure 1.9 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

FIGURE 1.6: Spaghetti plot for leadplants by soil type with loess fit.
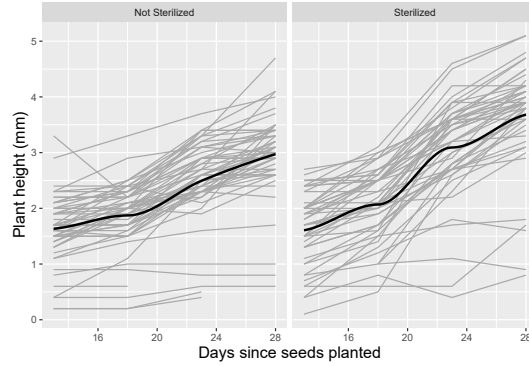
FIGURE 1.7: Spaghetti plot for leadplants by sterilization with loess fit.

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 pot-to-pot than we observed plant-to-plant (SD=0.46 for intercepts and SD=0.050 for slopes).

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
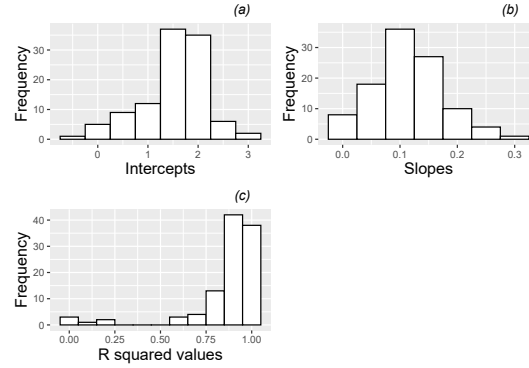
FIGURE 1.8: Histograms of (a) intercepts, (b) slopes, and (c) R-squared values for linear fits across all leadplants.
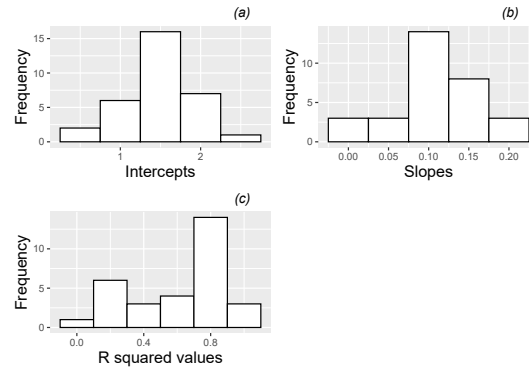
FIGURE 1.9: Histograms of (a) intercepts, (b) slopes, and (c) R-squared values for linear fits across all pots with leadplants.

(`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 1.10 complements Figure 1.3, 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 1.11 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 1.12 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.
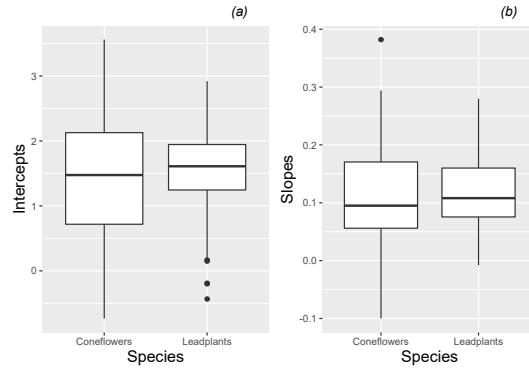
FIGURE 1.10: Boxplots of (a) intercepts and (b) slopes for all plants by species, based on a linear fit to height data from each plant.
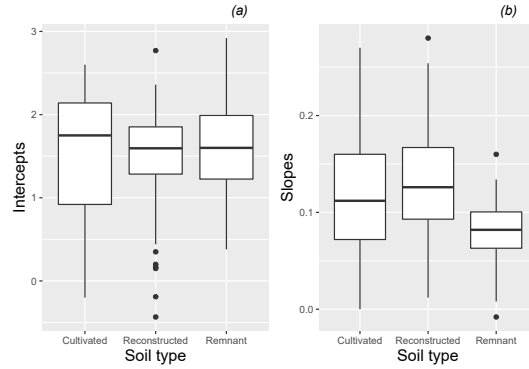
FIGURE 1.11: Boxplots of (a) intercepts and (b) slopes for all leadplants by soil type, based on a linear fit to height data from each plant.

Since we have time at Level One, any exploratory analysis of Case Study 1.2 should contain an investigation of the variance-covariance structure within plant. Figure 1.13 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 .77 to .91, while measurements ten days apart have
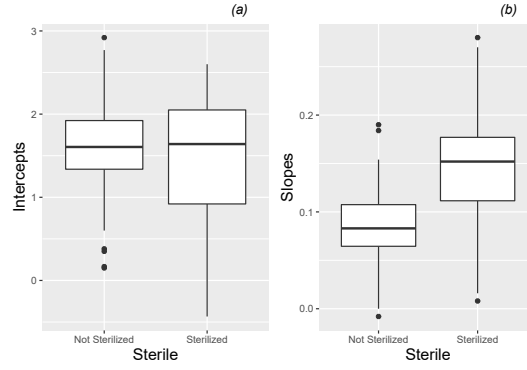
FIGURE 1.12: Boxplots of (a) intercepts and (b) slopes for all leadplants by sterilization, based on a linear fit to height data from each plant.

correlations of .62 and .70, and measurements fifteen days apart have correlation of .58.
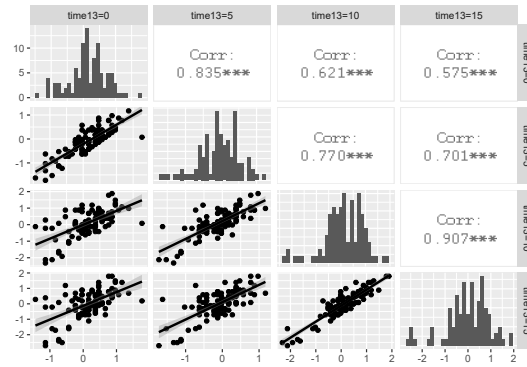
FIGURE 1.13: Correlation structure within plant. The upper right contains correlation coefficients between residuals at pairs of time points, the lower left contains scatterplots of the residuals at time point pairs, and the diagonal contains histograms of residuals at each of the four time points.

1.4 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 you 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 1.7).

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 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):

$$Y_{ijk} = a_{ij} + \epsilon_{ijk} \text{ where } \epsilon_{ijk} \sim N(0, \sigma^2) \tag{1.1}$$

- Level Two (plant within pot):

$$a_{ij} = a_i + u_{ij} \text{ where } u_{ij} \sim N(0, \sigma_u^2) \tag{1.2}$$

- Level Three (pot):

$$a_i = \alpha_0 + \tilde{u}_i \text{ where } \tilde{u}_i \sim N(0, \sigma_{\tilde{u}}^2) \tag{1.3}$$

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

- a_{ij} 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 a_{ij} at Level Two.
- a_i 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 a_i at Level Three.
- α_0 is a fixed effects model parameter representing the true mean height across all pots, plants, and time points
- ϵ_{ijk} describes how far an observed height Y_{ijk} is from the mean height for plant j from pot i
- u_{ij} describe how far the mean height of plant j from pot i is from the mean height of all plants from pot i
- $\tilde{u}_i$ describes how far the mean height of all observations from pot i is from the overall mean height across all pots, plants, and time points. None of the error terms $(\epsilon, u, \tilde{u})$ are considered model parameters; they simply account for differences between the observed data and expected values under our model.
- σ^2 is a variance component (random effects model parameter) that describes within-plant variability over time
- σ_u^2 is the variance component describing plant-to-plant variability within pot
- $\sigma_{\tilde{u}}^2$ is the variance component describing pot-to-pot variability.

The three-level unconditional means model can also be expressed as a composite model:

$$Y_{ijk} = \alpha_0 + \tilde{u}_i + u_{ij} + \epsilon_{ijk} \tag{1.4}$$

and this composite model can be fit using statistical software:

```
# Model A - unconditional means
modelal = lmer(hgt ~ 1 + (1|plant) + (1|pot),
                REML=T, data=leaddata)
```

```
##  Groups    Name         Variance Std.Dev.
##  plant     (Intercept) 0.2782    0.527
##  pot       (Intercept) 0.0487    0.221
##  Residual              0.7278    0.853

##  Number of Level Two groups =  107
##  Number of Level Three groups =  32

##             Estimate Std. Error t value
## (Intercept)    2.388    0.07887   30.28
```

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

- $\hat{\alpha}_0 = 2.39 =$ the mean height (in mm) across all time points, plants, and pots.
- $\hat{\sigma}^2 = 0.728 =$ the variance over time within plants.
- $\hat{\sigma}_u^2 = 0.278 =$ the variance between plants from the same pot.
- $\hat{\sigma}_{\tilde{u}}^2 = 0.049 =$ 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:

- Level One (timepoint within plant):

$$Y_{ijk} = a_{ij} + b_{ij} time_{ijk} + \epsilon_{ijk} \tag{1.5}$$

- Level Two (plant within pot):

$$a_{ij} = a_i + u_{ij}$$
$$b_{ij} = b_i + v_{ij}$$

- Level Three (pot):

$$a_i = \alpha_0 + \tilde{u}_i$$
$$b_i = \beta_0 + \tilde{v}_i$$

or as a composite model:

$$Y_{ijk} = [\alpha_0 + \beta_0 time_{ijk}] + [\tilde{u}_i + v_{ij} + \epsilon_{ijk} + (\tilde{v}_i + v_{ij})time_{ijk}] \qquad (1.6)$$

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

$$\left[\begin{array}{c} u_{ij} \\ v_{ij} \end{array} \right] \sim N \left(\left[\begin{array}{c} 0 \\ 0 \end{array} \right], \left[\begin{array}{cc} \sigma_u^2 & \\ \sigma_{uv} & \sigma_v^2 \end{array} \right] \right),$$

and

$$\left[\begin{array}{c} \tilde{u}_i \\ \tilde{v}_i \end{array} \right] \sim N \left(\left[\begin{array}{c} 0 \\ 0 \end{array} \right], \left[\begin{array}{cc} \sigma_{\tilde{u}}^2 & \\ \sigma_{\tilde{u}\tilde{v}} & \sigma_{\tilde{v}}^2 \end{array} \right] \right).$$

In this model, at Level One the trajectory for plant j from pot i is assumed to be linear, with intercept a_{ij} (height on Day 13) and slope b_{ij} (daily growth rate between Days 13 and 28); the ϵ_{ijk} terms capture the deviation between the true growth trajectory of plant j from pot i and its observed heights. At Level Two, a_i represents the true mean intercept and b_i represents the true mean slope for all plants from pot i, while u_{ij} and v_{ij} 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 σ_{uv}. Finally, α_0 is the true mean intercept and β_0 is the true mean daily growth rate over the entire population of leadplants, while $\tilde{u}_i$ and $\tilde{v}_i$ 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 (σ_u^2 and $\sigma_{\tilde{u}}^2$) and variability in rates of change (σ_v^2 and $\sigma_{\tilde{v}}^2$).

Using the composite model specification, the unconditional growth model can be fit to the seed germination data:

```
# Model B - unconditional growth
modelbl = lmer(hgt ~ time13 + (time13|plant) + (time13|pot),
               REML=T, data=leaddata)
```

```
##  Groups    Name         Variance Std.Dev. Corr
##  plant     (Intercept)  0.29914  0.5469
##            time13       0.00119  0.0346   0.28
##  pot       (Intercept)  0.04423  0.2103
##            time13       0.00126  0.0355   -0.61
##  Residual               0.08216  0.2866
```

```
##  Number of Level Two groups =  107
##  Number of Level Three groups =  32


##              Estimate Std. Error t value
## (Intercept)   1.5377   0.070305   21.87
## time13        0.1121   0.007924   14.15
```

From this output, we obtain estimates of our nine model parameters (two fixed effects and seven variance components):

- $\hat{\alpha}_0 = 1.538$ = the mean height of leadplants 13 days after planting.
- $\hat{\beta}_0 = 0.112$ = the mean daily change in height of leadplants from 13 to 28 days after planting.
- $\hat{\sigma} = .287$ = the standard deviation in within-plant residuals after accounting for time.
- $\hat{\sigma}_u = .547$ = the standard deviation in Day 13 heights between plants from the same pot.
- $\hat{\sigma}_v = .0346$ = the standard deviation in rates of change in height between plants from the same pot.
- $\hat{\rho}_{uv} = .280$ = the correlation in plants' Day 13 height and their rate of change in height.
- $\hat{\sigma}_{\tilde{u}} = .210$ = the standard deviation in Day 13 heights between pots.
- $\hat{\sigma}_{\tilde{v}} = .0355$ = the standard deviation in rates of change in height between pots.
- $\hat{\rho}_{\tilde{u}\tilde{v}} = -.610$ = 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.

1.5 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`):

$$a_i = \alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \tilde{u}_i$$
$$b_i = \beta_0 + \beta_1 strl_i + \beta_2 cult_i + \beta_3 rem_i + \tilde{v}_i$$

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:

$$Y_{ijk} = (\alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \tilde{u}_i + u_{ij}) +$$
$$(\beta_0 + \beta_1 strl_i + \beta_2 cult_i + \beta_3 rem_i + \tilde{v}_i + v_{ij}) time_{ijk} + \epsilon_{ijk}$$

which, after combining fixed effects and random effects, can be rewritten as:

$$Y_{ijk} = [\alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \beta_0 time_{ijk} +$$
$$\beta_1 strl_i time_{ijk} + \beta_2 cult_i time_{ijk} + \beta_3 rem_i time_{ijk}] +$$
$$[\tilde{u}_i + u_{ij} + \epsilon_{ijk} + \tilde{v}_i time_{ijk} + v_{ij} time_{ijk}]$$

From the output below, the addition of Level Three covariates in Model C (`cult`, `rem`, `strl`, and their interactions with `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 - add covariates at pot level
modelc1 = lmer(hgt ~ time13 + strl + cult + rem +
    time13:strl + time13:cult + time13:rem + (time13|plant) +
    (time13|pot), REML=T, data=leaddata)
```

```
##  Groups    Name        Variance Std.Dev. Corr
##  plant    (Intercept) 0.298087 0.5460
```

```
##              time13      0.001208 0.0348    0.28
##  pot       (Intercept) 0.053118 0.2305
##              time13      0.000132 0.0115   -1.00
##  Residual              0.082073 0.2865


##  Number of Level Two groups =  107
##  Number of Level Three groups =  32


##              Estimate Std. Error t value
## (Intercept)  1.50289    0.126992 11.8345
## time13       0.10107    0.008292 12.1889
## strl        -0.07655    0.151361 -0.5058
## cult         0.13002    0.182715  0.7116
## rem          0.13840    0.176189  0.7855
## time13:strl  0.05892    0.010282  5.7300
## time13:cult -0.02977    0.012263 -2.4273
## time13:rem  -0.03586    0.011978 -2.9939
```

```
drop_in_dev <- anova(modelbl, modelcl, test = "Chisq")
```

```
        npar   AIC   BIC logLik   dev Chisq Df
modelbl    9 603.8 640.0 -292.9 585.8    NA NA
modelcl   15 583.6 643.9 -276.8 553.6  32.2  6
           pval
modelbl      NA
modelcl 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: β_0 and σ^2. As the intercept, β_0 can take on any value; any real number is "allowable". But, by definition, variance terms such as σ^2 must be non-negative; that is, $\sigma^2 \geq 0$. Under the Principle of Maximum Likelihood, maximum likelihood estimators for β_0 and σ^2 will be chosen to maximize the likelihood of observing our given data. The left plot in Figure 1.14 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 β_0 but only non-negative values for σ^2. This constrained search is illustrated in the right plot in Figure 1.14. 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**.
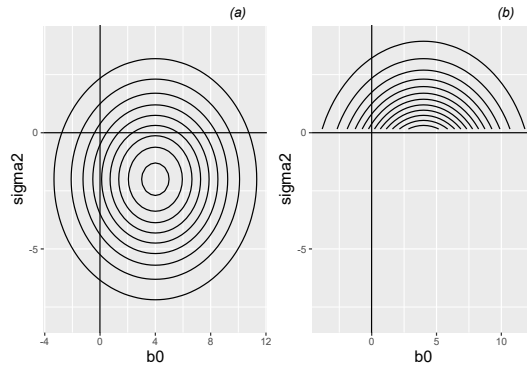
FIGURE 1.14: Left (a): hypothetical contours of the likelihood function $L(\beta_0, \sigma^2)$ with no restrictions on σ^2; the likelihood function is maximized at $(\hat{\beta}_0, \hat{\sigma}^2) = (4, -2)$. Right (b): hypothetical contours of the likelihood function $L(\beta_0, \sigma^2)$ with the restriction that $\sigma^2 \geq 0$; the constrained likelihood function is maximized at $(\hat{\beta}_0, \hat{\sigma}^2) = (4, 0)$

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 1.14) 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 $\rho_{\tilde{u}\tilde{v}} = 0$ rather than accepting the (constrained) maximum likelihood estimate of $\hat{\rho}_{\tilde{u}\tilde{v}} = -1$ that we saw in Model C).

```
# Try Model C without correlation between L2 errors
modelcl.noL2corr <- lmer(hgt ~ time13 + strl + cult + rem +
    time13:strl + time13:cult + time13:rem + (time13|plant) +
    (1|pot) + (0+time13|pot), REML=T, data=leaddata)
```

```
##  Groups    Name         Variance Std.Dev. Corr
##  plant     (Intercept) 0.294076 0.5423
##            time13       0.001206 0.0347    0.22
##  pot       (Intercept) 0.059273 0.2435
##  pot.1     time13       0.000138 0.0117
##  Residual               0.082170 0.2867

##  Number of Level Two groups =  107
##  Number of Level Three groups =  32

##              Estimate Std. Error t value
## (Intercept)  1.51209    0.129695 11.6588
## time13       0.10158    0.008361 12.1490
## strl        -0.08742    0.154113 -0.5673
## cult         0.13233    0.185687  0.7126
## rem          0.10657    0.179363  0.5942
## time13:strl  0.05869    0.010382  5.6529
## time13:cult -0.03065    0.012337 -2.4843
## time13:rem  -0.03810    0.012096 -3.1500
```

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 it is possible to remove boundary constraints through reasonable model reparameterizations, that is typically the preferred route. In this case, one option we might consider is simplifying Model C by setting $\sigma^2_{\tilde{v}} = \sigma_{\tilde{u}\tilde{v}} = 0$. We can then write our new model (Model C.1) in level-by-level formulation:

- Level One (timepoint within plant):

$$Y_{ijk} = a_{ij} + b_{ij}time_{ijk} + \epsilon_{ijk} \tag{1.7}$$

- Level Two (plant within pot):

$$a_{ij} = a_i + u_{ij}$$
$$b_{ij} = b_i + v_{ij}$$

- Level Three (pot):

$$a_i = \alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \tilde{u}_i$$
$$b_i = \beta_0 + \beta_1 strl_i + \beta_2 cult_i + \beta_3 rem_i$$

Note that there is no longer an error term associated with the model for mean growth rate b_i 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: $\tilde{u}_i \sim N(0, \sigma^2_{\tilde{u}})$. By removing one of our two Level Three error terms ($\tilde{v}_i$), we effectively removed two parameters – the variance for $\tilde{v}_i$ and the correlation between $\tilde{u}_i$ and $\tilde{v}_i$. Fixed effects remain similar, as can be seen in the output below:

```
# Try Model C without time at pot level
modelc10 <- lmer(hgt ~ time13 + strl + cult + rem +
    time13:strl + time13:cult + time13:rem + (time13|plant) +
    (1|pot), REML=T, data=leaddata)
```

```
##  Groups    Name          Variance Std.Dev. Corr
##  plant     (Intercept) 0.29471  0.5429
##            time13        0.00133  0.0364    0.19
##  pot       (Intercept) 0.05768  0.2402
##  Residual                0.08222  0.2867


##  Number of Level Two groups =   107
##  Number of Level Three groups =   32


##              Estimate Std. Error t value
## (Intercept)  1.51223    0.129016 11.7212
## time13        0.10109    0.007452 13.5648
## strl         -0.08752    0.153451 -0.5703
## cult          0.13287    0.184898  0.7186
## rem           0.10653    0.178612  0.5964
## time13:strl   0.05926    0.009492  6.2437
## time13:cult -0.03082    0.011353 -2.7151
## time13:rem  -0.03624    0.011101 -3.2649
```

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=2.089 on 2 df, p=0.3519) that $\sigma_{\tilde{v}}^2$ or $\sigma_{\tilde{u}\tilde{v}}$ is non-zero, so it is advisable to use the simpler Model C.1. However, Section 1.6 describes why this test may be misleading and prescribes a potentially better approach.

```
drop_in_dev <- anova(modelcl0, modelcl, test = "Chisq")
```

	npar	AIC	BIC	logLik	dev	Chisq	Df	pval
modelcl0	13	581.6	633.9	-277.8	555.6	NA	NA	NA
modelcl	15	583.6	643.9	-276.8	553.6	2.089	2	0.3519

1.6 Parametric bootstrap testing

As in Section **??**, when testing variance components at the boundary (such as $\sigma_{\tilde{v}}^2 = 0$), a method like the **parametric bootstrap** should be used, since using a chi-square distribution to conduct a likelihood ratio test produces p-values that tend to be too large.

TABLE 1.3: Original data for Plants 11 and 12 from Pot 1

pot	plant	soil	sterile	species	germin	hgt13	hgt18	hgt23	hgt28
1	11	STP	Y	L	Y	2.3	2.9	4.5	5.1
1	12	STP	Y	L	Y	1.9	2.0	2.6	3.5

Under the parametric bootstrap, we must simulate data under the null hypothesis many times. Here are the basic steps for running a parametric bootstrap procedure to compare Model C.1 with Model C:

- Fit Model C.1 (the null model) to obtain estimated fixed effects and variance components (this is the "parametric" part.)
- Use the estimated fixed effects and variance components 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

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):

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 a_i and b_i, the true intercept and slope from Pot i, in Model C.1 are:

$$a_i = \alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \tilde{u}_i$$
$$b_i = \beta_0 + \beta_1 strl_i + \beta_2 cult_i + \beta_3 rem_i$$

All the α and β terms will be fixed at their estimated values, so the one term that will change for each bootstrapped data set is $\tilde{u}_i$. As we obtain a numeric value for $\tilde{u}_i$ for each pot, we will fix the subscript. For example, if $\tilde{u}_i$ is set to -.192 for Pot #1, then we would denote this by $\tilde{u}_1 = -.192$. Similarly, in

the context of Model C.1, a_1 represents the mean height at Day 13 across all plants in Pot #1, where $\tilde{u}_1$ 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 $\tilde{u}_i$ is sampled from a normal distribution with mean 0 and standard deviation .240 (note that the standard deviation σ_u^2 is also fixed at its estimated value from Model C.1, given in Section 1.5). That is, a random component to the intercept for Pot #1 ($\tilde{u}_1$) would be sampled from a normal distribution with mean 0 and SD .240; say, for instance, $\tilde{u}_1 = -.192$. We would sample $\tilde{u}_2, ..., \tilde{u}_{72}$ in a similar manner. Then we can produce a model-based intercept and slope for Pot #1:

$$a_1 = 1.512 - .088(1) + .133(0) + .107(0) - .192 = 1.232$$
$$b_1 = .101 + .059(1) - .031(0) - .036(0) = .160$$

Notice a couple of features of the above derivations. First, all of the coefficients from the above equations ($\alpha_0 = 1.512$, $\alpha_1 = -.088$, etc.) come from the estimated fixed effects from Model C.1 reported in Section 1.5. 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 .160 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 .160 mm/day because of our assumption that there is no pot-to-pot variability in growth rate (i.e., $\tilde{v}_i = 0$).

Level Two

We next proceed to Level Two, where our equations for Model C.1 are:

$$a_{ij} = a_i + u_{ij}$$
$$b_{ij} = b_i + v_{ij}$$

We will initially focus on Plant #11 from Pot #1. Notice that the intercept (Day 13 height $= a_{11}$) for Plant #11 has two components: the mean Day 13 height for Pot #1 (a_1) which we specified at Level Three, and an error term (u_{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 $= b_{11}$) for Plant #11 similarly has two components. Since both a_1 and b_1 were determined at Level Three, at this point we need to find the two error terms for Plant #11: u_{11} and v_{11}. According to our multilevel model, we can sample u_{11} and v_{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 $u_{11} = .336$ and $v_{11} = .029$. Then we can produce a model-based intercept and slope for Plant #11:

$$a_{11} = 1.232 + .336 = 1.568$$
$$b_{11} = .160 + .029 = .189$$

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 ϵ_{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 1.5. Suppose we obtain residuals of $\epsilon_{111} = -.311$, $\epsilon_{112} = .119$, $\epsilon_{113} = .241$, and $\epsilon_{114} = -.066$. In that case, our parametrically generated data for Plant #11 from Pot #1 would look like:

$$
\begin{array}{rclcl}
Y_{111} & = & 1.568 + .189(0) - .311 & = & 1.257 \\
Y_{112} & = & 1.568 + .189(5) + .119 & = & 2.632 \\
Y_{113} & = & 1.568 + .189(10) + .241 & = & 3.699 \\
Y_{114} & = & 1.568 + .189(15) - .066 & = & 4.337
\end{array}
$$

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, $a_1 = 1.232$ and $b_1 = .160$, at Level Three. We would, however, sample new residuals u_{12} and v_{12} at Level Two, producing a different intercept a_{12} and slope b_{12} than those observed for Plant #11. Four new independent residuals ϵ_{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 – $\sigma_{\tilde{v}}^2$ and $\sigma_{\tilde{u}\tilde{v}}$ – that were not actually used when generating the responses

```
# Generate 1 set of bootstrapped data and run chi-square test
#  (will also work if use REML models, but may take longer)
set.seed(3333)
d <- drop(simulate(modelcl0.ml))
m2 <- refit(modelcl.ml, newresp=d)
m1 <- refit(modelcl0.ml, newresp=d)
drop_in_dev <- anova(m2, m1, test = "Chisq")
```

```
   npar   AIC   BIC logLik   dev Chisq Df   pval
m1   13 588.5 640.8 -281.2 562.5    NA NA     NA
m2   15 591.1 651.4 -280.5 561.1 1.405  2 0.4953
```

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 both models to the "bootstrapped" data. Since the data was generated using Model C.1, we would expect the two extra terms in Model C ($\sigma_{\tilde{v}}^2$ and $\sigma_{\tilde{u}\tilde{v}}$) 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 (-280.54 vs. -281.24), but this small difference is not statistically significant based on a chi-square test with 2 degrees of freedom (p=.4953).

However, we are really only interested in saving the likelihood ratio test statistic from this bootstrapped sample $(2*(-280.54-(-281.24) = 1.40)$. 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 $\sigma_{\tilde{v}}^2 = \sigma_{\tilde{u}\tilde{v}} = 0$, rather than making the (dubious) assumption that its behavior is described by a chi-square distribution with 2 degrees of freedom. Figure 1.15 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 (2.089). The parametric bootstrap provides a more reliable p-value in this case (.088 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, but we nearly had enough evidence to favor the more complex model.

```
bootstrapAnova(mA=modelcl.ml, m0=modelcl0.ml, B=1000)
```

```
   Df logLik    dev Chisq ChiDf pval_boot
m0 13 -277.8 555.6    NA    NA        NA
mA 15 -276.8 553.6 2.089     2     0.088
```
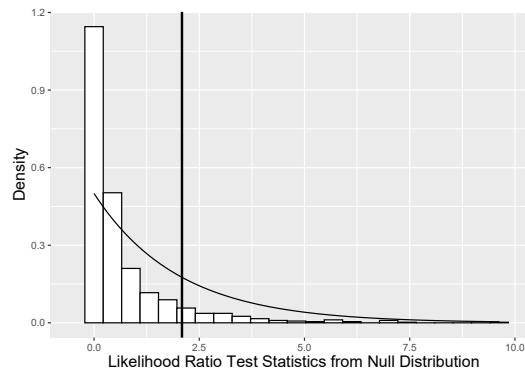

FIGURE 1.15: Null distribution of likelihood ratio test statistic derived using parametric bootstrap (histogram) compared to a chi-square distribution with 2 degrees of freedom (smooth curve). The horizontal line represents the observed likelihood ratio test statistic.

Another way of testing whether or not we should stick with the reduced model or reject it in favor of the larger model is by generating parametric bootstrap samples, and then using those samples to produce 95% confidence intervals for both $\rho_{\tilde{u}\tilde{v}}$ and $\sigma_{\tilde{v}}$. From the output below, the 95% bootstrapped confidence interval for $\rho_{\tilde{u}\tilde{v}}$ (-1, 1) contains 0, and the interval for $\sigma_{\tilde{v}}$ (.0000044, .0198) nearly contains 0, providing further evidence that the larger model is not needed.

```
confint(modelcl, method="boot", oldNames = F)
```

```
##                                2.5 %    97.5 %
## sd_(Intercept)|plant         4.307e-01  0.62069
## cor_time13.(Intercept)|plant -4.852e-02  0.71047
## sd_time13|plant              2.315e-02  0.04016
## sd_(Intercept)|pot           0.000e+00  0.33743
## cor_time13.(Intercept)|pot  -1.000e+00  1.00000
## sd_time13|pot                4.789e-06  0.02053
```

```
## sigma                           2.575e-01  0.31698
## (Intercept)                     1.276e+00  1.73778
## time13                          8.516e-02  0.11725
## strl                          -3.480e-01  0.22389
## cult                          -2.611e-01  0.47057
## rem                           -1.763e-01  0.46096
## time13:strl                     3.908e-02  0.07812
## time13:cult                   -5.490e-02 -0.00562
## time13:rem                    -5.802e-02 -0.01245
```

1.7 Exploding variance components

Our modeling task in Section 1.5 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 param-
eters: 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 equa-
tions for a_{ij} and b_{ij} 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_plus) could be written as follows:

- Level One (timepoint within plant):

$$Y_{ijk} = a_{ij} + b_{ij}time_{ijk} + \epsilon_{ijk} \tag{1.8}$$

- Level Two (plant within pot):

$$a_{ij} = a_i + c_i seedsize_{ij} + u_{ij}$$
$$b_{ij} = b_i + d_i seedsize_{ij} + v_{ij}$$

- Level Three (pot):

$$a_i = \alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \tilde{u}_i$$
$$b_i = \beta_0 + \beta_1 strl_i + \beta_2 cult_i + \beta_3 rem_i + \tilde{v}_i$$
$$c_i = \gamma_0 + \gamma_1 strl_i + \gamma_2 cult_i + \gamma_3 rem_i + \tilde{w}_i$$
$$d_i = \delta_0 + \delta_1 strl_i + \delta_2 cult_i + \delta_3 rem_i + \tilde{z}_i$$

or as a composite model:

$$
\begin{aligned}
Y_{ijk} = & [\alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \gamma_0 seedsize_{ij} + \\
& \beta_0 time_{ijk} + \beta_1 strl_i time_{ijk} + \beta_2 cult_i time_{ijk} + \beta_3 rem_i time_{ijk} + \\
& \gamma_1 strl_i seedsize_{ij} + \gamma_2 cult_i seedsize_{ij} + \gamma_3 rem_i seedsize_{ij} + \\
& \delta_0 seedsize_{ij} time_{ijk} + \delta_1 strl_i seedsize_{ij} time_{ijk} + \\
& \delta_2 cult_i seedsize_{ij} time_{ijk} + \delta_3 rem_i seedsize_{ij} time_{ijk}] + \\
& [\tilde{u}_i + u_{ij} + \epsilon_{ijk} + \tilde{w}_i seedsize_{ij} + \tilde{v}_i time_{ijk} + v_{ij} time_{ijk} + \\
& \tilde{z}_i seedsize_{ij} time_{ijk}]
\end{aligned}
$$

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

$$
\left[\begin{array}{c} u_{ij} \\ v_{ij} \end{array} \right] \sim N \left(\left[\begin{array}{c} 0 \\ 0 \end{array} \right], \left[\begin{array}{cc} \sigma_u^2 & \\ \sigma_{uv} & \sigma_v^2 \end{array} \right] \right),
$$

and

$$
\left[\begin{array}{c} \tilde{u}_i \\ \tilde{v}_i \\ \tilde{w}_i \\ \tilde{z}_i \end{array} \right] \sim N \left(\left[\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \end{array} \right], \left[\begin{array}{cccc} \sigma_{\tilde{u}}^2 & & & \\ \sigma_{\tilde{u}\tilde{v}} & \sigma_{\tilde{v}}^2 & & \\ \sigma_{\tilde{u}\tilde{w}} & \sigma_{\tilde{v}\tilde{w}} & \sigma_{\tilde{w}}^2 & \\ \sigma_{\tilde{u}\tilde{z}} & \sigma_{\tilde{v}\tilde{z}} & \sigma_{\tilde{w}\tilde{z}} & \sigma_{\tilde{z}}^2 \end{array} \right] \right).
$$

We would have 16 fixed effects from the four equations at Level Three, each with 4 fixed effects to estimate. 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 C_plus 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 Levels 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 (C_plus) can be simplified to this 9-parameter model:

- Level One:

$$Y_{ijk} = a_{ij} + b_{ij}time_{ijk} + \epsilon_{ijk} \tag{1.9}$$

- Level Two:

$$a_{ij} = a_i + c_i seedsize_{ij} + u_{ij}$$
$$b_{ij} = b_i$$

- Level Three:

$$a_i = \alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \tilde{u}_i$$
$$b_i = \beta_0$$
$$c_i = \gamma_0$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$, $u_{ij} \sim N(0, \sigma_u^2)$, and $\tilde{u}_i \sim N(0, \sigma_{\tilde{u}}^2)$. Or, in terms of a composite model:

$$Y_{ijk} = [\alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \gamma_0 seedsize_{ij} + \beta_0 time_{ijk}] +$$
$$[\tilde{u}_i + u_{ij} + \epsilon_{ijk}]$$

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 ($\sigma_{\tilde{u}}^2$) after controlling for sterilization and soil type, a random effect for plant within pot (σ_u^2) after controlling for seed size, and a random effect for error about the time trend for individual plants (σ^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 proscriptive, 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 Levels Two and Three can produce better estimates of fixed effects of interest and their standard errors.

1.8   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:

$$Y_{ijk} = a_{ij} + b_{ij}time_{ijk} + \epsilon_{ijk} \qquad (1.10)$$

- Level Two:

$$a_{ij} = a_i + u_{ij}$$
$$b_{ij} = b_i + v_{ij}$$

- Level Three:

$$a_i = \alpha_0 + \alpha_1 strl_i + \alpha_2 cult_i + \alpha_3 rem_i + \alpha_4 strl_i rem_i + \alpha_5 strl_i cult_i + \tilde{u}_i$$
$$b_i = \beta_0 + \beta_1 strl_i + \beta_2 cult_i + \beta_3 rem_i + \beta_4 strl_i rem_i + \beta_5 strl_i cult_i$$

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.

```
# Model D - add interactions to Model C.1
modeldl0 <- lmer(hgt ~ time13 + strl + cult + rem +
  time13:strl + time13:cult + time13:rem + strl:cult +
  strl:rem + time13:strl:cult + time13:strl:rem +
  (time13|plant) + (1|pot), REML=T, data=leaddata)
```

```
##  Groups   Name        Variance Std.Dev. Corr
##  plant    (Intercept) 0.29271  0.5410
##           time13      0.00128  0.0357   0.20
##  pot      (Intercept) 0.07056  0.2656
##  Residual             0.08214  0.2866
```

```
##  Number of Level Two groups =  107
##  Number of Level Three groups =  32
```

```
##                 Estimate Std. Error t value
## (Intercept)      1.53125  0.153905  9.9493
## time13           0.09492  0.008119 11.6917
## strl            -0.12651  0.222671 -0.5682
```

```
## cult             -0.06185    0.329926 -0.1875
## rem               0.12425    0.239595  0.5186
## time13:strl       0.07279    0.012005  6.0633
## time13:cult      -0.02262    0.020836 -1.0858
## time13:rem       -0.02047    0.013165 -1.5552
## strl:cult         0.27618    0.407111  0.6784
## strl:rem         -0.07139    0.378060 -0.1888
## time13:strl:cult -0.01608    0.024812 -0.6481
## time13:strl:rem  -0.05199    0.023951 -2.1707
```

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). This means we may have a **three-way interaction** in our final composite model between sterilization, remnant prairies, and time. Three-way interactions show that the size of an interaction between two predictors differs depending on the level of a third predictor. Whew!

Our final model (Model F), with its constraints on Level Three error terms, can be expressed level-by-level as:

- Level One:

$$Y_{ijk} = a_{ij} + b_{ij}time_{ijk} + \epsilon_{ijk} \qquad (1.11)$$

- Level Two:

$$a_{ij} = a_i + u_{ij}$$
$$b_{ij} = b_i + v_{ij}$$

- Level Three:

$$a_i = \alpha_0 + \tilde{u}_i$$
$$b_i = \beta_0 + \beta_1 strl_i + \beta_2 rem_i + \beta_3 strl_i rem_i$$

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

$$\left[\begin{array}{c} u_{ij} \\ v_{ij} \end{array} \right] \sim N \left(\left[\begin{array}{c} 0 \\ 0 \end{array} \right], \left[\begin{array}{cc} \sigma_u^2 & \\ \sigma_{uv} & \sigma_v^2 \end{array} \right] \right),$$

and $\tilde{u}_i \sim N(0, \sigma_{\tilde{u}}^2)$.

In composite form, we have:

$$Y_{ijk} = [\alpha_0 + \beta_0 time_{ijk} + \beta_1 strl_i time_{ijk} + \beta_2 rem_i time_{ijk}$$
$$+ \beta_3 strl_i rem_i time_{ijk}] + [\tilde{u}_i + u_{ij} + \epsilon_{ijk} + v_{ij} time_{ijk}]$$

```
# Model F - simplify and move toward "final model"
modelfl0 = lmer(hgt ~ time13 + time13:strl +
  time13:rem + time13:strl:rem +
  (time13|plant) + (1|pot), REML=T, data=leaddata)
```

```
## Groups    Name         Variance Std.Dev. Corr
## plant     (Intercept) 0.29404  0.5423
##           time13      0.00143  0.0378   0.16
## pot       (Intercept) 0.04871  0.2207
## Residual              0.08205  0.2864
```

```
##  Number of Level Two groups =  107
##  Number of Level Three groups =  32
```

```
##                 Estimate Std. Error t value
## (Intercept)      1.52909   0.071336  21.435
## time13           0.09137   0.007745  11.798
## time13:strl      0.05967   0.010375   5.751
## time13:rem      -0.01649   0.013238  -1.246
## time13:strl:rem -0.03939   0.023847  -1.652
```

A likelihood ratio test shows no significant difference between Models D and F (chi-square test statistic = 11.15 on 7 df, p=.1323), supporting the use of simplified Model F.

```
drop_in_dev <- anova(modelfl0, modeldl0, test = "Chisq")
```

```
          npar    AIC   BIC logLik   dev Chisq Df   pval
modelfl0    10  581.2 621.4 -280.6 561.2    NA NA     NA
modeldl0    17  584.0 652.4 -275.0 550.0 11.15  7 0.1323
```

We mentioned in Section 1.6 that we could also compare Model F and Model D, which differ only in their fixed effects terms, using the parametric bootstrap approach. In fact, in Section 1.6 we suggested that the p-value using the chi-square approximation (.1323) 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 .201. In Figure 1.16, 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.
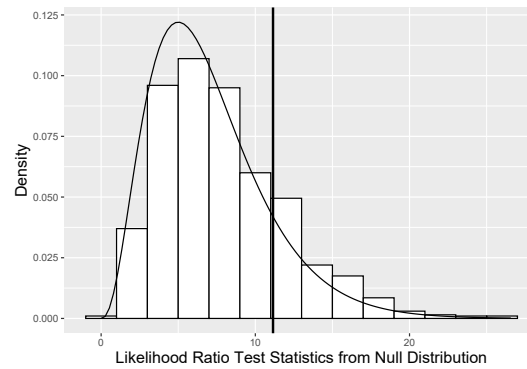
FIGURE 1.16: Null distribution of likelihood ratio test statistic derived using parametric bootstrap (histogram) compared to a chi-square distribution with 7 degrees of freedom (smooth curve). The horizontal line represents the observed likelihood ratio test statistic.

```
bootstrapAnova(mA=modeldl0, m0=modelfl0, B=1000)
```

```
   Df logLik   dev Chisq ChiDf pval_boot
m0 10 -280.6 561.2   NA    NA        NA
mA 17 -275.0 550.0 11.15    7     0.201
```

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 = 9.40 on 2 df, p=.0090). Thus, we will use Model F as our "Final Model" for generating inference.

```
drop_in_dev <- anova(modelel0, modelfl0, test = "Chisq")
```

```
          npar   AIC   BIC logLik   dev Chisq Df
modelel0     8 586.6 618.8 -285.3 570.6    NA NA
modelfl0    10 581.2 621.4 -280.6 561.2 9.401  2
             pval
modelel0       NA
modelfl0 0.009091
```

Estimates of model parameters can be interpreted in the following manner:

- $\hat{\sigma} = .287$ = the standard deviation in within-plant residuals after accounting for time.
- $\hat{\sigma}_u = .543$ = the standard deviation in Day 13 heights between plants from the same pot.
- $\hat{\sigma}_v = .037$ = the standard deviation in rates of change in height between plants from the same pot.
- $\hat{\rho}_{uv} = .157$ = the correlation in plants' Day 13 height and their rate of change in height.
- $\hat{\sigma}_{\tilde{u}} = .206$ = the standard deviation in Day 13 heights between pots.
- $\hat{\alpha}_0 = 1.529$ = the mean height for leadplants 13 days after planting.
- $\hat{\beta}_0 = 0.091$ = the mean daily change in height from 13 to 28 days after planting for leadplants from reconstructed prairies or cultivated lands (`rem`=0) with no sterilization (`strl`=0) .
- $\hat{\beta}_1 = 0.060$ = 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.
- $\hat{\beta}_2 = -0.017$ = 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.
- $\hat{\beta}_3 = -0.039$ = the decrease in mean daily change in height for leadplants from sterilized soil from remnant prairies, compared to the expected daily change based on $\hat{\beta}_1$ and $\hat{\beta}_2$. 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 just .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.

1.9   Covariance structure (Optional)

As in Chapter **??**, 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 $\tilde{v}_i$ at Level Three, and then we will evaluate the resulting covariance structure that results from removing $\tilde{v}_i$, thereby restricting $\sigma_{\tilde{v}}^2 = \sigma_{\tilde{u}\tilde{v}} = 0$. The composite version of Model B has been previously expressed as:

$$Y_{ijk} = [\alpha_0 + \beta_0 time_{ijk}] + [\tilde{u}_i + u_{ij} + \epsilon_{ijk} + (\tilde{v}_i + v_{ij})time_{ijk}] \qquad (1.12)$$

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

$$\left[\begin{array}{c} u_{ij} \\ v_{ij} \end{array} \right] \sim N \left(\left[\begin{array}{c} 0 \\ 0 \end{array} \right], \left[\begin{array}{cc} \sigma_u^2 & \\ \sigma_{uv} & \sigma_v^2 \end{array} \right] \right),$$

and

$$\left[\begin{array}{c} \tilde{u}_i \\ \tilde{v}_i \end{array} \right] \sim N \left(\left[\begin{array}{c} 0 \\ 0 \end{array} \right], \left[\begin{array}{cc} \sigma_{\tilde{u}}^2 & \\ \sigma_{\tilde{u}\tilde{v}} & \sigma_{\tilde{v}}^2 \end{array} \right] \right).$$

In order to assess the implied covariance structure from our standard model, we must first derive variance and covariance terms for related observations (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 $t_k = time_{ijk}$, since all plants were planned to be observed on the same 4 days.

The variance for an individual observation can be expressed as:

$$Var(Y_{ijk}) = (\sigma^2 + \sigma_u^2 + \sigma_{\tilde{u}}^2) + 2(\sigma_{uv} + \sigma_{\tilde{u}\tilde{v}})t_k + (\sigma_v^2 + \sigma_{\tilde{v}}^2)t_k^2, \qquad (1.13)$$

and the covariance between observations taken at different timepoints (k and k') from the same plant (j) is:

$$Cov(Y_{ijk}, Y_{ijk'}) = (\sigma_u^2 + \sigma_{\tilde{u}}^2) + (\sigma_{uv} + \sigma_{\tilde{u}\tilde{v}})(t_k + t_{k'}) + (\sigma_v^2 + \sigma_{\tilde{v}}^2)t_k t_{k'}, \ (1.14)$$

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:

$$Cov(Y_{ijk}, Y_{ij'k'}) = \sigma_{\tilde{u}}^2 + \sigma_{\tilde{u}\tilde{v}}(t_k + t_{k'}) + \sigma_{\tilde{v}}^2 t_k t_{k'}. \qquad (1.15)$$

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:

$$Cov(\mathbf{Y}_{ij}) = \left[\begin{array}{cccc} \tau_1^2 & & & \\ \tau_{12} & \tau_2^2 & & \\ \tau_{13} & \tau_{23} & \tau_3^2 & \\ \tau_{14} & \tau_{24} & \tau_{34} & \tau_4^2 \end{array} \right],$$

where $\tau_k^2 = Var(Y_{ijk})$ and $\tau_{kk'} = Cov(Y_{ijk}, Y_{ijk'})$. Note that τ_k^2 and $\tau_{kk'}$ are both independent of i and j so that $Cov(\mathbf{Y}_{ij})$ 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 (1.13) through (1.15).

Similarly, the covariance matrix between observations from plants j and j' from pot i can be expressed as this 4x4 matrix:

$$Cov(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'}) = \begin{bmatrix} \tilde{\tau}_{11} & & & \\ \tilde{\tau}_{12} & \tilde{\tau}_{22} & & \\ \tilde{\tau}_{13} & \tilde{\tau}_{23} & \tilde{\tau}_{33} & \\ \tilde{\tau}_{14} & \tilde{\tau}_{24} & \tilde{\tau}_{34} & \tilde{\tau}_{44} \end{bmatrix},$$

where $\tilde{\tau}_{kk} = Cov(Y_{ijk}, Y_{ij'k}) = \sigma_{\tilde{u}}^2 + 2\sigma_{\tilde{u}\tilde{v}}t_k + \sigma_{\tilde{v}}^2 t_k^2$ and $\tilde{\tau}_{kk'} = Cov(Y_{ijk}, Y_{ij'k'})$ as derived above. As we saw with $Cov(\mathbf{Y}_{ij})$, $\tilde{\tau}_{kk}$ and $\tilde{\tau}_{kk'}$ are both independent of i and j so that $Cov(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'})$ 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 $Cov(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'})$ 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 1.4 produced the following estimates for variance components: $\hat{\sigma}^2 = .0822$, $\hat{\sigma}_u^2 = .299$, $\hat{\sigma}_v^2 = .00119$, $\hat{\sigma}_{uv} = \hat{\rho}_{uv}\sqrt{\hat{\sigma}_u^2 \hat{\sigma}_v^2} = .00528$, $\hat{\sigma}_{\tilde{u}}^2 = .0442$, $\hat{\sigma}_{\tilde{v}}^2 = .00126$, $\hat{\sigma}_{\tilde{u}\tilde{v}} = \hat{\rho}_{\tilde{u}\tilde{v}}\sqrt{\hat{\sigma}_{\tilde{u}}^2 \hat{\sigma}_{\tilde{v}}^2} = -.00455$. Based on these estimates and the derivations above, the within-plant correlation structure over time is estimated to be:

$$Corr(\mathbf{Y}_{ij}) = \begin{bmatrix} 1 & & & \\ .76 & 1 & & \\ .65 & .82 & 1 & \\ .54 & .77 & .88 & 1 \end{bmatrix}$$

for all plants j and all pots i, and the correlation structure between different plants from the same pot is estimated to be:

$$Corr(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'}) = \begin{bmatrix} .104 & & & \\ .047 & .061 & & \\ -.002 & .067 & .116 & \\ -.037 & .068 & .144 & .191 \end{bmatrix}.$$

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., $\sigma_{\tilde{v}}^2 = \sigma_{\tilde{u}\tilde{v}} = 0$). Using the same REML variance components estimates as above except that $\hat{\sigma}_{\tilde{v}}^2 = 0$ rather than .00126 and $\hat{\sigma}_{\tilde{u}\tilde{v}} = \hat{\rho}_{\tilde{u}\tilde{v}}\sqrt{\hat{\sigma}_{\tilde{u}}^2\hat{\sigma}_{\tilde{v}}^2} = 0$ rather than $-.00455$, the within-plant correlation structure is estimated to be:

$$Corr(\mathbf{Y}_{ij}) = \begin{bmatrix} 1 & & & \\ .80 & 1 & & \\ .75 & .84 & 1 & \\ .70 & .82 & .88 & 1 \end{bmatrix}$$

for all plants j and all pots i, and the correlation structure between different plants from the same pot is estimated to be:

$$Corr(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'}) = \begin{bmatrix} .104 & & & \\ .095 & .087 & & \\ .084 & .077 & .068 & \\ .073 & .067 & .059 & .052 \end{bmatrix}.$$

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 1.13 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.

1.9.1 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:

$$Var(Y_{ijk}) = Var(\epsilon_{ijk} + u_{ij} + \tilde{u}_i + (v_{ij} + \tilde{v}_i)time_{ijk})$$
$$= (\sigma^2 + \sigma_u^2 + \sigma_{\tilde{u}}^2) + 2(\sigma_{uv} + \sigma_{\tilde{u}\tilde{v}})t_k + (\sigma_v^2 + \sigma_{\tilde{v}}^2)t_k^2$$

The covariance between observations taken at different timepoints from the same plant is:

$$Cov(Y_{ijk}, Y_{ijk'}) = Cov(\epsilon_{ijk} + u_{ij} + \tilde{u}_i + (v_{ij} + \tilde{v}_i)t_k,$$
$$\epsilon_{ijk'} + u_{ij} + \tilde{u}_i + (v_{ij} + \tilde{v}_i)t_{k'})$$
$$= (\sigma_u^2 + \sigma_{\tilde{u}}^2) + (\sigma_{uv} + \sigma_{\tilde{u}\tilde{v}})(t_k + t_{k'}) + (\sigma_v^2 + \sigma_{\tilde{v}}^2)t_k t_{k'}$$

The covariance between observations taken from different plants from the same pot is:

$$Cov(Y_{ijk}, Y_{ij'k'}) = Cov(\epsilon_{ijk} + u_{ij} + \tilde{u}_i + (v_{ij} + \tilde{v}_i)t_k,$$
$$\epsilon_{ij'k'} + u_{ij'} + \tilde{u}_i + (v_{ij'} + \tilde{v}_i)t_{k'})$$
$$= \sigma_{\tilde{u}}^2 + \sigma_{\tilde{u}\tilde{v}}(t_k + t_{k'}) + \sigma_{\tilde{v}}^2 t_k t_{k'}$$

Based on these variances and covariances and the expressions for $Cov(\mathbf{Y}_{ij})$ and $Cov(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'})$ in Section 1.9, 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):

$$Cov(\mathbf{Y}_i) = \begin{bmatrix} Cov(\mathbf{Y}_{i1}) & & & & & \\ Cov(\mathbf{Y}_{i1}, \mathbf{Y}_{i2}) & Cov(\mathbf{Y}_{i2}) & & & & \\ Cov(\mathbf{Y}_{i1}, \mathbf{Y}_{i3}) & Cov(\mathbf{Y}_{i2}, \mathbf{Y}_{i3}) & Cov(\mathbf{Y}_{i3}) & & & \\ Cov(\mathbf{Y}_{i1}, \mathbf{Y}_{i4}) & Cov(\mathbf{Y}_{i2}, \mathbf{Y}_{i4}) & Cov(\mathbf{Y}_{i3}, \mathbf{Y}_{i4}) & Cov(\mathbf{Y}_{i4}) & & \\ Cov(\mathbf{Y}_{i1}, \mathbf{Y}_{i5}) & Cov(\mathbf{Y}_{i2}, \mathbf{Y}_{i5}) & Cov(\mathbf{Y}_{i3}, \mathbf{Y}_{i5}) & Cov(\mathbf{Y}_{i4}, \mathbf{Y}_{i5}) & Cov(\mathbf{Y}_{i5}) & \\ Cov(\mathbf{Y}_{i1}, \mathbf{Y}_{i6}) & Cov(\mathbf{Y}_{i2}, \mathbf{Y}_{i6}) & Cov(\mathbf{Y}_{i3}, \mathbf{Y}_{i6}) & Cov(\mathbf{Y}_{i4}, \mathbf{Y}_{i6}) & Cov(\mathbf{Y}_{i5}, \mathbf{Y}_{i6}) & Cov(\mathbf{Y}_{i6}) \end{bmatrix}.$$

A covariance matrix for our entire data set, therefore, would be block diagonal, with $Cov(\mathbf{Y}_i)$ 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 $Cov(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'})$ 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 $Cov(\mathbf{Y}_{ij}, \mathbf{Y}_{ij'})$, the diagonal terms in a correlation matrix are formed by $Corr(Y_{ijk}, Y_{ij'k}) = \frac{\tilde{\tau}_{kk}}{\sqrt{Var(Y_{ijk})Var(Y_{ij'k})}} = \frac{\tilde{\tau}_{kk}}{\tau_k^2}$ and the off-diagonal terms are formed by $Corr(Y_{ijk}, Y_{ij'k'}) = \frac{\tilde{\tau}_{kk'}}{\sqrt{Var(Y_{ijk})Var(Y_{ij'k'})}} = \frac{\tilde{\tau}_{kk'}}{\tau_k \tau_{k'}}$.

We calculated estimated covariance and correlation matrices within plant and between plants in Section 1.9 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.

1.10 Notes on Using R (Optional)

When fitting three level models in `lmer()`, note that an estimated variance at Level One (σ^2) comes for "free", but variance terms at Level Two (σ_u^2) and Level Three ($\sigma_{\tilde{u}}^2$) 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 (σ_u^2 comes for "free") and slope (σ_v^2 comes from the `time13` term) at the plant level, along with a covariance or correlation (σ_{uv} or ρ_{uv}) between those two error terms. To restrict $\sigma_{uv} = \rho_{uv} = 0$, you could specify each error term separately: (`1 | plant`) + (`0+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.

In Section 1.6 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 `anova()` function in R, which created two problems: (a) the `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.

1.11 Exercises

1.11.1 Conceptual Exercises

1. **Seed Germination.** In Sections 1.3.1 and 1.3.2, 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 1.2, would using mean heights by pot be reasonable as well? How about for Figure 1.3?

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 lead-plant data at the end of Section 1.3.2.

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 $\hat{\sigma}_u^2$ 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. How does Figure **??** show that a likelihood ratio test using a chi-square distribution would be biased?

13. In Section 1.7, 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 1.7; (c) what assumptions have you made in your simplification in (b)?

14. In Section 1.8, Model F (the null model) is compared to Model D using a parametric bootstrap test. As in Section 1.6, 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.

15. Section 1.8 contains an interpretation for the coefficient of a three-way interaction term, $\hat{\beta}_3$. Provide an alternative interpretation for $\hat{\beta}_3$ by focusing on how the sterilization-by-soil type interaction differs over time.

16. **Collective Efficacy and Violent Crime.** In a 1997 *Science* article, Sampson et al. [1997] 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.

17. Table 1.4 shows Table 3 from Sampson et al. [1997]. Provide interpretations of the **bolded** coefficients in context.

18. Based on Table 1.4, let Y_{ijk} be the response of person j from neighborhood i to item k 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.

19. Suggest valuable exploratory data analysis plots to complement Table 1.4.

TABLE 1.4: Table 3: Correlates of collective efficacy from Sampson et al. (1997)

Variable	Coefficient	SE	t ratio
Intercept	3.523	0.013	263.20
Person-level predictors			
Person-Level predictors			
Female	-0.012	0.015	-0.76
Married	-0.005	0.021	-0.25
Separated or divorced	-0.045	0.026	-1.72
Single	**-0.026**	0.024	-1.05
Homeowner	0.122	0.020	6.04
Latino	0.042	0.028	1.52
Black	-0.029	0.030	-0.98
Mobility	**-0.025**	0.007	-3.71
Age	0.0021	0.0006	3.47
Years in neighborhood	**6e-04**	0.0008	0.78
Neighborhood-level predictors			
SES	0.035	0.008	4.64
Neighborhood-level predictors			
Concentrated disadvantage	**-0.172**	0.016	-10.74
Variance components			
Immigrant concentration	-0.037	0.014	-2.66
Residential stability	0.074	0.013	5.61
Percent of variance explained			
Variance Components			
Within neighborhoods	0.32		
Between neighborhoods	0.026		
Percent of variance explained			
Within neighborhoods	3.2		
Between neighborhoods	70.3		

TABLE 1.5: A portion of Table 4: Neighborhood correlates of perceived neighborhood violence from Sampson et al. (1997).

| | Model 1 | | | Model 2 | | |
| | Social composition | | | Social comp and collective efficacy | | |
Variable	Coefficient	SE	t	Coefficient	SE	t
Concentrated disadvantage	**0.277**	**0.021**	**13.30**	**0.171**	**0.024**	**7.24**
Immigrant concentration	0.041	0.017	2.44	0.018	0.016	1.12
Residential stability	-0.102	0.015	-6.95	-0.056	0.016	-3.49
Collective efficacy				-0.618	0.104	-5.95

20. If the model suggested by Table 1.4 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?

21. At the bottom of Table 1.4, the percent of variance explained is given within and between neighborhoods. Explain what these values likely represent and how they were calculated.

22. Table 1.5 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.] 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.

1.11.2 Guided Exercises

1. **Tree Tubes** A student research team at St. Olaf College contributed to the efforts of biologist Dr. Kathy Shea to investigate a rich data set concerning forestation in the surrounding land [Eisinger et al., 2011]. Tubes were placed on trees in some locations or *transects* but not in others. Interest centers on whether tree growth is affected by the presence of tubes. The data is currently stored in long format in `treetube.csv`. Each row represents one tree in a given year. Key variables include:

   - `TRANSECT`: The id of the transect housing the tree
   - `TUBEX`: 1 if the tree had a tube, 0 if not

- **ID**: The tree's unique id
- **SPECIES**: The tree's species
- **YEAR**: Year of the observation
- **HEIGHT**: The tree's height in meters

a. Perform basic exploratory data analysis. For example, which variables are correlated with the heights of the trees?

b. Explore patterns of missing data. Consider looking for patterns between transects. If you found any patterns of missing data, how might this affect your modeling?

c. We wish to fit a three level model for a tree's height. What would be observational units at level three? At level two? At level one? What are our level three variables? Level two variables? Level one variables?

d. Generate spaghetti plots showing how height varies over time for trees which had tubes and for trees which did not have tubes. What do you notice?

e. Fit Model A, a three level, unconditional means model for height. Write out the model at levels three, two, and one, as well as the composite model.

f. Create a new variable, TIME, which represents the number of years since 1990. Use this to fit Model B, an unconditonal growth model for height. Write out this model at levels three, two, and one. What is the correlation of random effects at level two? What does this mean? (Section 1.5 may help.)

g. In response to part (f), fit a new unconditional growth model (Model C) with $\rho_{uv} = 0$ and $\rho_{\hat{u}\hat{v}} = 0$ (uncorrelated random effects at levels two and three). Write out the model at levels two and three.

h. If we wanted to compare Model B and Model C, it would not be appropriate to use a likelihood ratio test. Why is this?

i. Use the parametric bootstrap to compare Model B and Model C. Which model does the test favor?

j. Regardless of your answer to (i), we will build off of Model C for the rest of the analysis. With that in mind, test the hypothesis that trees' growth rates are affected by tubes by adding an interaction between TUBEX and TIME (Model D). Interpret the fitted estimate of this interaction.

k. Perform a likelihood ratio test comparing Model C and Model D. Interpret the results of this test in the context of this study.

2. **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 mid-

dle schools from 132 different districts, and the following variables were collected and can be found in `kentucky.csv`:

- `dis_id` = District Identifier
- `sch_id` = School Identifier
- `stud_nm` = Student Identifier
- `female` = Coded 1 if Female and 0 if Male
- `nonwhite` = Coded 0 if White and 1 otherwise
- `readn` = California Test of Basic Skills Reading Score
- `mathn` = California Test of Basic Skills Math Score
- `sch_size` = School-Level Size (centered natural log)
- `sch_ses` = School-Level SES (socio-economic status - centered)
- `dis_size` = District-Level Size (centered natural log)
- `dis_ses` = District-Level SES (socio-economic status - centered)

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 `mathn`? 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.

- Write out the complete three-level model. How many parameters must be estimated?

- 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.

   - Write out the complete three-level model. Also write out
     the composite model. How many parameters must be esti-
     mated?
   - 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?
   - (Optional) Explore the nature of the 3-way interaction be-
     tween female, sch_ses, and dis_size by finding the pre-
     dicted math score for 5 cases based on Model D with an
     error term only on the intercept equation at Level Three.
     Comment on trends you observe.
       – Female vs. male with average `sch_ses` and `dis_size`
       – Female vs. male with `sch_ses` at Q1 and `dis_size` at
         Q1
       – Female vs. male with `sch_ses` at Q1 and `dis_size` at
         Q3
       – Female vs. male with `sch_ses` at Q3 and `dis_size` at
         Q1
       – Female vs. male with `sch_ses` at Q3 and `dis_size` at
         Q3
   - (Optional) Create two scatterplots to illustrate the 3-way
     interaction between `female`, `sch_ses`, and `dis_size`.

1.11.3 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 knowl-
   edge 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 in `mudamalai.csv`:

- `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)
- `Width` = width of leaf (cm)
- `Area` = surface area of leaf (sq cm)
- `Pedial.length` = length of the leaf stem (mm)
- `Pedial.width` = width of the leaf stem (mm)
- `Herbivory` = percent of leaf removed or eaten by herbivores
- `Stomata` = density of specialized openings that allow for gas exchange on leaves

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 multilevel 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.

Bibliography

E. H. Allison, W. N. Adger, M. C. Badjeck, K. Brown, D. Conway, N. K. Dulvy, A. Halls, A. Perry, and J. D. Reynolds. Effects of climate change on the sustainability of capture and enhancement fisheries important to the poor, 2005.

Diane Angell. Effects of soil type and sterilization on the growth of coneflowers and leadplants, 2010.

S. G. Baer, D. J. Kitchen, J. M. Blair, and C. W. Rice. Changes in ecosystem structure and function along a chronosequence of restored grasslands. *Ecological Applications*, 12(6):1688–1701, 2002. doi: 10.1890/1051-0761(2002)012[1688:CIESAF]2.0.CO;2.

Robert Bickel. *Multilevel Analysis for Applied Research*. Guilford Publications, New York, New York, 2007.

Ann Cannon, George Cobb, Brad Hartlaub, Julie Legler, Robin Lock, Tom Moore, Allan Rossman, and Jeff Witmer. *Stat2*. Macmillan, 2019.

Robert Eisinger, Amanda Elling, and J.R. Stamp. Tree growth rates and mortality. In *Proceedings of the National Conference on Undergraduate Research (NCUR)*, Ithaca College, New York, 2011. Center for Interdisciplinary Research. March 31,2011 – April 2, 2011.

Johannes M. H. Knops and David Tilman. Dynamics of soil nitrogen and carbon accumulation for 61 years after agricultural abandonment. *Ecology*, 81 (1):88–98, 2000. doi: 10.1890/0012-9658(2000)081[0088:DOSNAC]2.0.CO;2.

Camill Philip, McKone Mark J., Sturges Sean T., Severud William J., Ellis Erin, Limmer Jacob, Martin Christopher B., Navratil Ryan T., Purdie Amy J., Sandel Brody S., Talukder Shano, and Trout Andrew. Community- and ecosystem-level changes in a species-rich tallgrass prairie restoration. *Ecological Applications*, 14(6):1680–1694, 2004. doi: 10.1890/03-5273.

Ian Pray. Effects of rainfall and sun exposure on leaf characteristics, 2009.

Robert J. Sampson, Stephen W. Raudenbush, and Felton Earls. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328):918–924, 1997. ISSN 0036-8075. doi: 10.1126/science.277.5328. 918.

Donald R. Zak, William E. Holmes, David C. White, Aaron D. Peacock, and David Tilman. Plant diversity, soil microbial communities, and ecosystem function. *Ecology*, 84(8):2042–2050, 2003. doi: 10.1890/02-0433.