Solutions Manual
Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R

Paul Roback and Julie Legler

© 2021 by Taylor & Francis Group, LLC. Except as permitted under U.S. copyright law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by an electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

Contents

Preface

Three types of exercises are available for each chapter. **Conceptual Exercises** ask about key ideas in the contexts of case studies from the chapter and additional research articles where those ideas appear. **Guided Exercises** provide real data sets with background descriptions and lead students step-by-step through a set of questions to explore the data, build and interpret models, and address key research questions. Finally, **Open-Ended Exercises** provide real data sets with contextual descriptions and ask students to explore key questions without prescribing specific steps. A solutions manual with solutions to all exercises will be available to qualified instructors at our book's website[1].

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Acknowledgments. We would like to thank students of Stat 316 at St. Olaf College since 2010 for their patience as this book has taken shape with their feedback. We would especially like to thank these St. Olaf students for their summer research efforts which significantly improved aspects of this book: Cecilia 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 the Department of Mathematics, Statistics, and Computer Science at St. Olaf. We are also thankful to Samantha Roback for developing the cover image.

[1] www.routledge.com

1

Review of Multiple Linear Regression

```
# Packages required for Chapter 1
library(knitr)
library(mosaic)
library(gridExtra)
library(GGally)
library(kableExtra)
library(jtools)
library(rsample)
library(broom)
library(tidyverse)
```

1.1 Exercises

1.1.1 Conceptual Exercises

1. **Applications that do not violate assumptions for inference in LLSR**.

a) Response: Chirps/minute; Explanatory: Temperature.
L: Temperature is linearly associated with the number of chirps/minute.
I: Errors (deviations of actual chirps from predicted chirps based on a linear model) of each observation are unrelated.
N: Chirps/minute at a given temperature are normally distributed.
E: The variation in chirps/minute is approximately the same across all temperature levels.
Note: if chirps/minute are relatively low, some LINE assumptions may be violated, and this response might be better modeled as a Poisson random variable (see Chapter 4).

b) Response: Height; Explanatory: Shoe Size.

L: Height is linearly associated with shoe size for women 20 to 24 years old.
I: Random selection of women usually produces independence.
N: Height is normally distributed at any shoe size.
E: The variation of height is approximately equal at all shoe sizes.

2. **Applications that do violate assumptions for inference in LLSR**.

a) Response: Low birthweight (binary); Explanatory: Socioeconomic status and parental stability. Our normality assumption is violated. We have a binary response variable, which would not be normally distributed as it can only take 2 values. Further, we would not have a linear relationship. If we did, we would see predicted values outside of 0 or 1. Also, binary variables have higher variability when true probabilities are near 0.5.

b) Response: Number of patients getting relief; Explanatory: Dose level. Since our response (number of patients) is binomial (number of successes in a fixed number of trials), we would expect some skewness and discreteness (lack of normality), unequal variances (since variance is related to probability of success), and non-linearity (since else we'd get predicted counts outside our range of possible values).

c) Response: Number of trips made; Explanatory: Distance from Boundary Waters and socioeconomic status summarized by zip code. Our response is a count (number of trips per zip code). Such values would be bounded below by 0, likely violating our linearity assumption, and typically skewed right with variance increasing as counts increase.

d) Response: Depression score; Explanatory: Estrogen patch. The independence assumption is violated. We took multiple (6) measurements from the same subjects. These measurements are likely to be correlated with one another. Additionally, if depression scores often bump up against some upper or lower bound, our normality assumption would be violated.

3. **Kentucky Derby.** [Wikipedia contributors, 2018].

a) An advantage of using histograms is that we can get a good idea of the shape of distributions by group, as well as sample size. Boxplots are nice because they give us a good idea about where the center of each distribution is (if we used histograms, we would have to eyeball the mean of the distribution), and they allow easier group comparisons of center and spread of the middle 50 percent.

b) Looking at the scatterplot, we can see that the relationship may be quadratic. This is not apparent from the correlation coefficient, which assumes a linear relationship.

c) We could alter point size based on number of starters.

d) $\epsilon_i = Y_i - \widehat{Y}_i$, the difference between the observed and expected speed for the

ith observation. Y_i is the height of the point at Year_i, while $\hat{Y}_i$ is the height of the predictive line at Year_i.

e) We divide by the standard error to *standardize* our estimated slope and evaluate how many standard errors away from zero (or another null value) it lies. These standardized distances then follow a t-distribution under the null hypothesis.

f) Here, β_1 represents the difference in mean speeds between the two groups (fast and not fast) because a "1 unit change in X" with a binary predictor means going from 0 (not fast) to 1 (fast). This is the same as difference $\bar{Y}_{\text{fast}} - \bar{Y}_{\text{not fast}}$ that we would use in a t-test. When we use LLS regression, we assume equal variance in speed at all possible track conditions. Therefore, we are testing a difference of means where both means come from normal distributions with equal variance. The default t-test in R makes the Welch adjustment for unequal variance, so to match the test from LLSR, we need to require equal variances between groups.

g) Comparing two estimates $\hat{Y}_1$ and $\hat{Y}_2$ where $\text{Fast}_1 = 1$ and $\text{Fast}_2 = 0$, $\hat{Y}_1 - \hat{Y}_2 = \hat{\beta}_1(\text{Yearnew}_1 - \text{Yearnew}_2) + \hat{\beta}_2(1 - 0)$. Only when year is fixed ($\text{Year}_1 = \text{Year}_2$) does $\hat{\beta}_2$ correspond to the difference in estimates.

h) We are 95% confident that the true average winning speed in 1896 under non-fast conditions was between 50.61 and 51.22 ft/s.

i) We have statistically significant evidence ($t = 11.77, p < 0.001$) that average winning speed increases linearly over time, after controlling for track condition.

j) Equation (1.4) models $\hat{Y}_i$ (the predicted value), not Y_i (the true value).

k) The interaction coefficient β would represent the change in the yearly increase in speed corresponding to an increase of one starter. Alternatively, it would represent the change in the effect of a one starter over time. Specifically, we can fix one variable at the 25th and 75th percentiles and examine the effect of the second variable on speed in those two cases.

l) $\hat{\beta}_3$: After accounting for the quadratic effects of year and the number of starters, average winning speeds are 1.39 ft/s higher under fast conditions than under slow conditions. $\hat{\beta}_5$: Adjusting for the quadratic effect of the year and track conditions, for each additional starter, we expect a 0.025 ft/s decrease in the winning time.

4. **Moneyball.** [Farrar and Bruggink, 2011].

a) There are potential problems with the independence assumption in the player salary model. Players within the same teams likely have similar (correlated) salaries, for which an LLSR model would fail to account. Salaries are also notoriously right-skewed (non-normal), with greater variability at higher

salary levels. The team production model also features correlated data, since every team is represented by five years worth of data.

b) You certainly could argue in favor of Model 3. For example, Model 3 has a higher adjusted R^2 value than Model 1, even though neither interaction term is significant on its own. We could perform an extra sum of squares F-test to formally compare these two nested models.

c) While in the American League we would associate a 10 point increase in OBP with a 28.6 run increase, in the National League, it would be associated with a 28.6+2.75=31.35 run increase. Likewise, in the American League, a 10 point increase in SLG is associated with a 16.2 run increase, but a 16.2+2.41=18.61 run increase in the National League. Knowing that a "10 unit increase in individual OBP costs $370,500, and a 10 unit increase in individual SLG costs $369,800" (basically similar), we find that in the American League, investment in an OBP increase rather than a SLG increase leads to 77% more runs, while that same investment in the National League would lead to 68% more runs. Of course, this assumes a team can choose to add 10 points of OBP or 10 points of SLG while holding all else constant, which is almost impossible.

d) They could be using R^2 (the sum of squared residuals) as a measurement of fit. However, they used completely different sets of data and a completely different response variable, so any comparison is apples to oranges.

e) The term is not significant after accounting for all other predictors, however its inclusion still causes a significant drop in the squared sum of residuals. Adjusted R^2 or AIC just makes an ad-hoc adjustment for model complexity rather than a formal test based on assumed probability distributions.

f) The paper mainly discusses the already published Moneyball hypothesis. It moreso describes the current status of baseball rather than prescribing new tactics (as *Moneyball* did). In addition, this paper does not include factors such as defense and base running, or the fact that younger players' salaries are often controlled by a salary cap. Any estimates of effects must take into account that the run production model was analyzed at the team level. Finally, a team should pay a player based on projected future performance rather than past performance.

1.1.2 Guided Exercises

1. **Gender discrimination in bank salaries**. [Ramsey and Schafer, 2002].

```
banksalary <- read_csv("data/banksalary.csv")
bank <- banksalary %>%
  mutate(male = ifelse(banksalary$sex=="MALE", 1, 0))
```

a) Obs units = 93 bank workers; response = starting salary; explanatory = sex, experience, age, education, seniority

```
# Examine data frame
head(bank)       # print first 6 rows

# Generate relevant summary statistics for response variable
favstats(~ bsal, data = bank)

# Look at marginal relationship between sex and beginning salary

# Three options for getting summary statistics
favstats(~ bsal | sex, data = bank)

bank %>%
  group_by(sex) %>%
  summarize(mean = mean(bsal),
            median = median(bsal),
            sd = sd(bsal),
            iqr = IQR(bsal),
            n = n())

# Several options for plotting 1 categorical and 1 numeric variable
boxplot(bsal ~ sex, ylab="Beginning salary", data = bank)
bwplot(sex ~ bsal, data = bank)
ggplot(bank, aes(x = sex, y = bsal)) +
  geom_boxplot() + coord_flip()
ggplot(data = bank, mapping = aes(x = bsal, y = ..density..)) +
  geom_freqpoly(mapping = aes(colour = sex), binwidth = 250)
ggplot(data = bank, mapping = aes(x = bsal, y = ..density..)) +
  geom_density(mapping = aes(colour = sex))
ggplot(bank, aes(x=bsal, fill=sex)) +
  geom_density(alpha=0.4)
ggplot(bank, aes(x = sex, y = bsal)) +
  geom_violin()
ggplot(data = bank) +
  geom_histogram(mapping = aes(x = bsal), bins = 10) +
  facet_wrap(~ sex, nrow = 2)
ggplot(data = bank) +
  geom_histogram(mapping = aes(x = bsal, y = ..density..),
                 bins = 10) +
```

```
  facet_wrap(~ sex, nrow = 2)

# Initial analysis of sex vs salary, ignoring other covariates
t.test(bsal ~ sex, data = bank)

model0 = lm(bsal ~ sex, data = bank)
summary(model0)
t.test(bsal ~ male, var.equal = TRUE, data = bank)
```

b) First, we must consider variability – how much do male and female salaries vary? For example, the 25th percentile for males is approximately equal to the 75th percentile for females. A two-sample t-test suggests that any differences in mean salaries are above and beyond underlying variability among employees.

Next, and importantly, we must control for other factors which could explain higher starting salaries.

```
# How are covariates related to response?
ggplot(bank, aes(y = bsal, age)) +
  geom_point() +
  geom_smooth(method = lm)
ggplot(bank, aes(y = bsal, exper)) +
  geom_point() +
  geom_smooth(method = lm)
ggplot(bank, aes(y = bsal, educ)) +
  geom_point() +
  geom_smooth(method = lm)
ggplot(bank, aes(y = bsal, senior)) +
  geom_point() +
  geom_smooth(method = lm)
```

c) Higher age = higher salary (But r=.03. True just at lower end? Quadratic?). More experience = higher salary (But r=.17. True just at lower end? Quadratic?). More education = higher salary (Yes: r=.41).

It will be important to control for education. Do males earn more because they have more education?

d) Seniority allows us to control for inflation. More seniority = lower starting salary since longer ago (r = -.29).

```
# How are explanatory variables related to each other?
bank0 <- bank %>% select(bsal, age, exper, educ, senior, male)
pairs(bank0)     # matrix of scatterplots
cor(bank0)       # matrix of correlations

# How are explanatory variables related to sex?
favstats(~ age | sex, data = bank)
ggplot(bank, aes(x = sex, y = age)) +
  geom_boxplot() + coord_flip()

favstats(~ exper | sex, data = bank)
ggplot(bank, aes(x = sex, y = exper)) +
  geom_boxplot() + coord_flip()

favstats(~ educ | sex, data = bank)
ggplot(bank, aes(x = sex, y = educ)) +
  geom_boxplot() + coord_flip()

favstats(~ senior | sex, data = bank)
ggplot(bank, aes(x = sex, y = senior)) +
  geom_boxplot() + coord_flip()
```

e) Age vs experience has r=.80; since they are highly correlated, we may not need both in our model. Remember this does NOT imply an interaction between age and experience. There are no other extremely worrisome associations between predictors, although there is a tendency for males to be younger and more educated, which means it will be important to adjust for age and education when evaluating differences between males and females.

```
# Fit linear regression with single predictor (experience)
model1 <- lm(bsal ~ exper, data = bank)
summary(model1)

# Residual plots - run off the page unless redo margins
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(model1, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

f) Intercept: 5289.02 = mean starting salary if no experience

Slope: 1.30 = mean increase in starting salary for each 1 year increase in experience

R^2: 2.78% = 2.78% of person-to-person variability in starting salary can be explained by differences in experience

Significance: We do not have statistically significant evidence (t = 1.613, p = 0.11) that starting salary is associated with experience.

g) L = residuals vs fitted values has small curvature around 0 – slight evidence of quadratic effect. I = from data collection. Was everyone in the same year paid the same salary, etc.? N = normal QQ plot is straight line, except for one outlier on the high end. E = scale-location plot is basically flat. Plus, residuals vs. leverage values show no high leverage or potentially influential points

```
# Fit multiple regression model with four predictors
model2 = lm(bsal ~ senior + age + educ + exper, data = bank)
summary(model2)
anova(model1, model2, test="F")

# Examine AIC and BIC scores (lower is better)
AIC(model1)                      # AIC-model1
AIC(model2)                      # AIC-model2
AIC(model1, k=log(nrow(bank)))   # BIC-model1
AIC(model2, k=log(nrow(bank)))   # BIC-model2

# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(model2, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

h) R2 increased from .0278 to .3176. Adjusted R2 increased from .0171 to .2866. AIC decreased from 1487 to 1460 (good). BIC decreased from 1495 to 1476 (good). Based on a nested F-test, we have statistically significant evidence (F = 12.456, p < .001) that the larger model outperforms the smaller model. Finally, the residual plots show better adherence to the LINE conditions in the larger model.

i) We can likely remove age in future modeling steps. Age is not significantly associated with starting salary after accounting for seniority, education, and experience.

```
# Investigate interactions between age and experience
bank <- bank %>%
  mutate(experience = ifelse(exper > 100, "high", "low"))

ggplot(bank, aes(y = bsal, x = age, color = experience)) +
```

```
  geom_point() +
  geom_smooth(method = lm)

modelint = lm(bsal ~ age + experience + age:experience,
              data = bank)
summary(modelint)
```

j) The coded scatterplot shows some evidence of an interaction, that there's a negative association between age and starting salary for those with higher experience, but little association for those with lower experience. However, this interaction does not appear to be statistically significant.

```
# One potential final model
model3 <- lm(bsal ~ senior + educ + exper + male, data = bank)
summary(model3)

# Construct 95% CIs for model coefficients "by hand"...
betas = summary(model3)$coef[,1]   # store model betas
SEs = summary(model3)$coef[,2]    # store SEs of betas
tstar = qt(.975,model3$df)
lb = betas - tstar*SEs
ub = betas + tstar*SEs
cbind(lb,ub)

# ... or more simply:
confint(model3)

# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(model3, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

k) We can be 95% confident that males with similar education and experience earn between $488 and $956 more than females, after adjusting for inflation. All LINE assumptions appear to be met by model3.

l) Evidence of gender discrimination is pretty strong, yet we must be careful about assigning causation because this is an observational study. For example, we must be sure we have controlled for all important covariates in our analysis. In this case, there actually was a huge settlement in 1989 (first filed in 1977) that concluded similarly qualified women were given lower jobs.

```
# Should beginning salary be log transformed?
bank <- bank %>%
  mutate(logbsal = log(bsal))
hist1 <- ggplot(bank, aes(bsal)) + geom_histogram(bins = 10)
hist2 <- ggplot(bank, aes(logbsal)) + geom_histogram(bins = 10)
grid.arrange(hist1, hist2, ncol=2)

# Look at marginal relationship between sex and beginning salary
model3a = lm(logbsal ~ senior + educ + exper + male, data = bank)
summary(model3a)

# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(model3a, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

m) Log(`bsal`) is slightly more normal, but not a huge improvement. Often salaries in general are right-skewed, but not lower level starting salaries. Note that we cannot compare R^2 values since the response values are different in the two models.

n) Males had 13.8% (exp(-1.296) = 1.138) higher median salaries than females, after controlling for seniority, education, and experience. Residual plots look good for both 3 and 3a.

```
# Investigate interactions with sex
ggplot(bank, aes(y = bsal, x = age, color = sex)) +
  geom_point() +
  geom_smooth(method = lm)
ggplot(bank, aes(y = bsal, x = exper, color = sex)) +
  geom_point() +
  geom_smooth(method = lm)
ggplot(bank, aes(y = bsal, x = educ, color = sex)) +
  geom_point() +
  geom_smooth(method = lm)
ggplot(bank, aes(y = bsal, x = senior, color = sex)) +
  geom_point() +
  geom_smooth(method = lm)
```

o) There are no obvious interactions, so model3 is a pretty good final model, indicating that the amount of gender discrimination is pretty consistent across levels of age, experience, education, and seniority.

2. **Sitting and MTL thickness.** [Siddarth et al., 2018].

```
sitting <- read_csv("data/sitting.csv")
```

```
ggplot(data = sitting, aes(x = MET, y = sitting)) +
  geom_point() +
  geom_smooth()

sittingMod1 <- lm(sitting ~ MET, data = sitting)
summary(sittingMod1)
```

a) There seems to be almost no correlation between MET and sitting time, supporting the claim that sedentary behaviors may be independent from physical activity. Only 0.5% of the subject-to-subject variability in sitting time per day can be explained by knowing their metabolic units per week.

```
# Should MET be log transformed?
sitting <- sitting %>%
  mutate(logMET = log(MET))
hist1 <- ggplot(sitting, aes(MET)) + geom_histogram(bins = 8)
hist2 <- ggplot(sitting, aes(logMET)) + geom_histogram(bins = 8)
grid.arrange(hist1, hist2, ncol=2)
```

b) Logging MET seems like a good idea. The original histogram shows signficant right skewness, which mostly disappears after logging.

```
sittingMod2 <- lm(MTL ~ sitting, data = sitting)
summary(sittingMod2)

ggplot(sitting, aes(y = MTL, x = sitting)) +
  geom_point() +
  geom_smooth()

# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(sittingMod2, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

c) Other than a little evidence of unequal variance in the Scale-Location plot, LINE conditions look good.

```
sitting <- mutate(sitting, cAge = age - mean(age))
sittingMod3 <- lm(MTL ~ sitting + cAge, data = sitting)
summary(sittingMod3)
```

d) The mean MTL thickness for 60-year-olds who never sit is 2.69 mm. After controlling for age, each extra hour per day sitting is associated with an expected decrease of .021 mm in MTL thickness. After controlling for sitting time, each extra year in age is associated with a predicted increase of .00435 mm in MTL thickness.

```
sittingMod3 <- lm(MTL ~ sitting + logMET + age, data = sitting)
summary(sittingMod3)
confint(sittingMod3)
```

e) Yes - we produce the same coefficient estimates and CIs.

f) Yes. After controlling for physical activity, sitting is significantly negatively associated with MTL thickness, but after controlling for sitting, physical activity is not associated with MTL thickness. This assumes both are valid measures of activity and inactivity.

g) The headline seems greatly overstated (despite the qualifier "could"). Siddarth's study is observational, so ascribing cause-effect is a big leap; they do not specifically analyze the effect of standing at one's desk, and their responses variable is not smartness.

3. **Housing prices and log transformations.** [Kaggle, 2018a].

```
kcHouses <- read_csv("data/kingCountyHouses.csv")

hist1 <- ggplot(kcHouses, aes(price)) + geom_histogram(bins = 20)
hist2 <- ggplot(kcHouses, aes(sqft)) + geom_histogram(bins = 20)
grid.arrange(hist1, hist2, ncol=2)

summary(kcHouses$price)
summary(kcHouses$sqft)

ggplot(kcHouses, aes(y = price, x = sqft)) +
```

```
  geom_point(size = .25) +
  geom_smooth(method = lm)

cor(kcHouses$price, kcHouses$sqft)
```

a) Both price and sqft are heavily right-skewed, although positively associated.

```
kcMod1 <- lm(price ~ sqft, data = kcHouses)
summary(kcMod1)

# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(kcMod1, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

b) price $= -43580 + 280.6\text{sqft} + \epsilon$. We associate a 1 sqft increase in a house's interior with a \$280 increase in selling price. Most of our LINE conditions appear violated. L = some evidence of non-linearity in Residuals vs. Fitted (starts above 0, moves below 0, then back above). I = prices and residuals likely geographically correlated. N = Normal Q-Q shows skewness of residuals on the high side. E = Scale-Location shows increasing variability at higher predicted prices.

For reference, here are mathematical details on this interpretation of slope:

Interpreting Slopes: Y on X

$$\begin{aligned} mean(Y|X+1) &= \beta_0 + \beta_1(X+1) \\ mean(Y|X) &= \beta_0 + \beta_1(X) \\ mean(Y|X+1) - mean(Y|X) &= \beta_1 \end{aligned}$$

For every unit increase in X, the mean Y changes by β_1.

```
kcHouses <- mutate(kcHouses, logprice = log(price))

kcMod2 <- lm(logprice ~ sqft, data = kcHouses)
summary(kcMod2)
```

c) $\log(\hat{\text{price}}) = 12.2 + 0.00040\text{sqft}$.

d) Each 1 square foot increase is associated with a mean increase of .00040 (log dollars?) in the house's log price.

e) If Y_2 is the predicted price for a house with 1 more interior square foot than a house with predicted price Y_1, $\log(Y_2) - \log(Y_1) = 0.00040$, so $0.00040 = \log\left(\frac{Y_2}{Y_1}\right)$, and $e^{0.00040} = \frac{Y_2}{Y_1} = 1 + \frac{Y_2 - Y_1}{Y_1}$. Therefore, $Y_2 - Y_1 = Y_1(e^{0.00040} - 1)$. Thus, $e^{0.00040}$ is the multiplicative increase in Y for a 1 unit (additive) increase in X, and we can think of this as a $100(e^{0.0004} - 1) = 0.04$ percent increase in selling price corresponding to a 1 square foot interior increase.

Here are more mathematical details on this situation:

Interpreting Slopes: logY on X

$$\begin{aligned} mean(logY|X+1) &= \beta_0 + \beta_1(X+1) \\ mean(logY|X) &= \beta_0 + \beta_1(X) \\ mean(logY|X+1) - mean(logY|X) &= \beta_1 \end{aligned}$$

To ease interpretation, note that:

$$mean(log) \neq log(mean)$$

But

$$median(log(Y)) = log(median(Y))$$

So

$$\begin{aligned} log(median(Y|X+1)) - log(median(Y|X)) &= \beta_1 \\ log\left(\frac{median(Y|X+1)}{median(Y|X)}\right) &= \beta_1 \\ \frac{median(Y|X+1)}{median(Y|X)} &= e^{\beta_1} \end{aligned} \tag{1.1}$$

For each additional unit of X, the median of Y changes by a multiplicative factor of e^{β_1}.

```
# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(kcMod2, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

f) Conditions may not be met, as we observe some lack of linearity and unequal variance with higher predicted logprices.

```
kcHouses <- mutate(kcHouses, logsqft = log(sqft))
kcMod3 <- lm(price ~ logsqft, data = kcHouses)
summary(kcMod3)
```

g) $\widehat{\text{price}} = -3451377 + 528648 \log(\text{sqft})$.

h) Predicted `price` increases by \$528648, on average, for every 1 unit increase in `logsqft`.

i) Letting Y_2 be the fitted price of a house with 1% more square footage than a house with fitted price Y_1, $Y_2 - Y_1 = 528648(\log(1.01x) - \log(x)) = 528648 \log(1.01) = 5260.2$. So, we associate a 1% increase in `sqft` with a \$5260.2 mean increase in selling price.

Here are more mathematical details on this situation:

Interpreting Slopes: Y on logX

$$\begin{aligned} mean(Y|2X) &= \beta_0 + \beta_1(log(2X)) \\ mean(Y|X) &= \beta_0 + \beta_1(logX) \\ mean(Y|2X) - mean(Y|X) &= \beta_1(log2X - logX) \\ mean(Y|2X) - mean(Y|X) &= \beta_1(log2) \end{aligned}$$

When X is doubled, the mean Y changes by $\beta_1(log2)$.

```
# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(kcMod3, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

j) Conditions are not met. We see lack of linearity (evident through the scatterplot and Residuals vs Fitted plot), lack of normality (Normal Q-Q), and unequal variance (Scale-Location).

```
kcMod4 <- lm(logprice ~ logsqft, data = kcHouses)
summary(kcMod4)
```

k) $\log(\widehat{\text{price}}) = 6.73 + 0.84 \log(\text{sqft})$.

l) Every 50% increase in square footage is associated with a multiplicative

increase of $1.50^{0.84} = 1.406$ in median selling price. So selling price increases by a factor of 1.406, which is 40.6%.

Here are more mathematical details on this situation:

Interpreting Slopes: logY on logX

$$\begin{aligned} log(median(Y|2X)) - log(median(Y|X)) &= \beta_1(log2) \\ log\left(\frac{median(Y|2X)}{median(Y|X)}\right) &= \beta_1(log2) \\ \frac{median(Y|2X)}{median(Y|X)} &= 2^{\beta_1} \end{aligned}$$

Doubling X changes the median of Y by a factor of 2^{β_1}

```
ggplot(kcHouses, aes(y = logprice, x = logsqft)) +
  geom_point(size = .25) +
  geom_smooth(method = lm)

# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(kcMod4, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

m) The scatterplot displays a linear relationship between `logsqft` and `logprice`; however, the Residuals vs Fitted plot hints at a violation of linearity and the possible utility of adding a quadratic term. We also have some lack of normality in the tails, as well as some evidence of unequal variance from the Scale-Location plot. And we could still have geographic correlation that violates independence. Overall, LINE conditions are a bit shaky.

```
kcMod5 <- lm(logprice ~ logsqft + waterfront, data = kcHouses)
summary(kcMod5)
```

n) A house on the waterfront sells for a predicted median price that is 2.10 ($e^{0.742}$) times larger than a house not on the waterfront, after adjusting for square footage of the house.

1.1.3 Open-Ended Exercises

1. **The Bechdel Test.** [Hickey, 2014].

```
bechdel <- read_csv("data/bechdel.csv")
```

As always, we begin by exploring our data. We see that `budget`, `totalGross`, `domGross`, `intGross`, and `totalROI` are all heavily skewed right.

```
ggplot(bechdel) +
  geom_histogram(aes(x = year), binwidth = 1)

ggplot(bechdel) +
  geom_histogram(aes(x = budget))

ggplot(bechdel) +
  geom_histogram(aes(x = totalGross))

ggplot(bechdel) +
  geom_histogram(aes(x = domGross))

ggplot(bechdel) +
  geom_histogram(aes(x = intGross))

ggplot(bechdel) +
  geom_histogram(aes(x = totalROI))
ggplot(bechdel) +
  geom_histogram(aes(x = totalROI), binwidth = 1) +
  coord_cartesian(xlim = c(0,100))
```

Examining some relationships with `pass`, we find that 862 movies in our dataset fail the Bechdel test while 753 pass it. There doesn't seem to be a clear relation between `year` and the proportion of movies passing the Bechdel test. Summary statistics show that movies failing the Bechdel test have higher mean budget ($65,877,024) than movies passing the test ($46,913,086). Movies which fail the test also have higher total, domestic, and international gross profits. Both groups have similar mean return on investment (5.26 for those failing vs. 5.15 for those passing).

```
table(bechdel$pass)

ggplot(bechdel) +
  geom_histogram(aes(x = year, fill=as.factor(pass)), binwidth=1)

ggplot(bechdel) +
  geom_histogram(aes(x = year, fill = as.factor(pass)),
                 position = "fill", binwidth = 1)

bechdel %>%
  split(.$pass) %>%
  map(summary)

ggplot(bechdel) +
  geom_density(aes(x = log(totalROI), color = as.factor(pass)))
```

After log transforming `totalROI`, it appears to be roughly normally distributed. However, we see no clear relation with `budget`.

```
ggplot(bechdel) +
  geom_point(aes(y = totalROI, x = budget))
ggplot(bechdel, aes(y = log(totalROI), x = budget)) +
  geom_point() +
  geom_smooth(method = "lm")
```

We find a relatively linear positive relationship between log(`budget`) and log(`totalGross`).

```
ggplot(bechdel) +
  geom_point(aes(x = log(budget), y = log(totalGross)))
```

We could first approach this question by modeling a movie's gross profits as a function of its budget and if it passed the Bechdel test. We could then theoretically discuss the relationship between profits and passing the Bechdel test *after accounting for* budget. Fitting this model, we find a significant negative relationship between a movie's total gross profits and passing the Bechdel test. That is to say, if a movie passes the test, we expect their logged total gross profits to be -0.22 lower. Equivalently, we expect profits of movies passing the Bechdel test to be 20% ($1-e^{-.22}$) smaller after accounting for the effect of budget.

```
bechdelMod1 <- lm(log(budget)~log(totalGross)+pass, data=bechdel)
summary(bechdelMod1)
100*(exp(-0.22399)-1)

ggplot(bechdel) +
  geom_point(aes(x = log(budget), y = log(totalGross),
                 color = as.factor(pass)), alpha = .5) +
  scale_color_manual(values = c("red","blue")) +
  geom_abline(intercept = 7.34194, slope = 0.54861,
              color = "red") +
  geom_abline(intercept = 7.43194-0.22399, slope = 0.54861,
              color = "blue")
```

Alternatively, as FiveThirtyEight did, we could model return on investment in place of gross profits. First, we could try a model with `pass` as the only predictor. Testing the significance of the β_1 coefficient corresponding to `pass`, we do not find evidence of a significant relationship ($t = -0.116, p = 0.91$). (Note that this is equivalent to a t-test of a difference of means.)

```
bechdelMod2 <- lm(totalROI~pass, data = bechdel)
summary(bechdelMod2)
```

We could experiment with covariates. For example, we could model log(`totalROI`) as a function of log(`budget`).

```
bechdelMod3 <- lm(log(totalROI)~log(budget), data = bechdel)
summary(bechdelMod3)
ggplot(bechdel, aes(x = log(budget), y = log(totalROI))) +
  geom_point() +
  geom_smooth(method = "lm")
```

Adding `pass` to this model, we find that after accounting for budget's effect, passing the Bechdel test still has no significant relationship with the return on investment of a film.

```
bechdelMod4 <- lm(log(totalROI)~log(budget)+pass, data = bechdel)
summary(bechdelMod4)
```

2. **Waitress tips.** [Dahlquist and Dong, 2011].

```
tips <- read_csv("data/TipData.csv")
```

a) We'll start with some Exploratory Data Analysis. Calculate tip percentage (`tippct` = 100*`Tip Percentage`) and consider this to be the primary response. Produce a plot and a set of summary statistics examining how each of the following potential explanatory variables is related to `tippct`: `Payment`, `Age`, `Meal`, `Bill`. Write one sentence summarizing how each explanatory variable is related to tip percentage.

```
# How is tippct related to explanatory vars?
head(tips)
tips <- tips %>%
  mutate(tippct = 100*`Tip Percentage`)

ggplot(tips, aes(x = tippct)) +
  geom_histogram()
favstats(~ tippct | Payment, data = tips)
favstats(~ tippct | Meal, data = tips)
favstats(~ tippct | Age, data = tips)
cor(tips$tippct, tips$Bill)

ggplot(tips, aes(x = Bill, y = tippct)) +
  geom_point() +
  geom_smooth() +
  labs(list(x = "Bill (dollars)", y = "Tip Percentage"))
ggplot(tips, aes(x = Payment, y = tippct)) +
  geom_boxplot() + coord_flip()
ggplot(tips, aes(x = Meal, y = tippct)) +
  geom_boxplot() + coord_flip()
ggplot(tips, aes(x = Age, y = tippct)) +
  geom_boxplot() + coord_flip()
```

Mean tip percentage is higher when credit cards are used (17.5%) than when just cash (17.0%) or credit with cash tip (15.6%) is used. [Note one patron with missing payment type; we'll leave them in our data for now.]

Mean tip percentage is higher during late night dining (21.1%) than during dinner (16.5%) or lunch (16.8%). [Note a possible outlier with tip of 80.2%; we'll leave them in our data for now.]

Mean tip percentage is lower for middle-aged patrons (16.7%) than for seniors (18.1%) or young adults (18.0%). [Note one missing age; we'll leave them in our data for now.]

Finally, tip percentage is negatively correlated with bill size (r=-.22).

b) Are there any interesting relationships among the predictors about which we should be aware? Find appropriate summary statistics relating `Bill` to `Payment`, `Age`, and `Meal`. Then use a table of conditional proportions to relate `Payment`, `Age`, and `Meal`–e.g. try `prop.table(table(Age,Meal),2)`. Summarize your findings in one or two sentences.

```
# Examine relationships among explanatory vars
ytable <- tally(~ Meal + Age, data = tips)
ytable
mosaicplot(ytable, color=c("blue", "light blue"))
prop.table(ytable, 1)

ytable <- tally(~ Meal + Payment, data = tips)
ytable
mosaicplot(ytable, color=c("blue", "light blue"))
prop.table(ytable, 1)

ytable <- tally(~ Age + Payment, data = tips)
ytable
mosaicplot(ytable, color=c("blue", "light blue"))
prop.table(ytable, 1)

favstats(~ Bill | Age, data = tips)
favstats(~ Bill | Meal, data = tips)
favstats(~ Bill | Payment, data = tips)
```

Late night meals have a much higher proportion of young adults (58%), while lunches have a relatively higher proportion of seniors (28%). Cash is most likely to be used late night (42%) and least likely to be used during dinner (30%). Young adults are most likely to use cash (44%), and middle-aged patrons are least likely (26%). The smallest bills are found during late night (mean $13.83), when young adults are dining (mean $20.25), and when cash is used (mean $27.98), or often some combination.

c) Model tip percentage as a linear function of bill size. Interpret your slope in the context of the problem.

```
# Model tippct as linear function of Bill
model0 <- lm(tippct ~ Bill, data = tips)
summary(model0)
```

The mean tip percentage decreases by 0.8% for each $10 increase in the bill size.

d) Create indicator variables for young patrons (1 if `Age` is "Yadult") and late dining (1 if `Meal` is "Late Night"). Model tip percentage as a function of `bill`, `late`, `young`, and the interaction between `bill` and `young`.

```
# Model tippct as function of late, young, bill, and young:bill
tips <- tips %>%
  mutate(credit = ifelse(Payment == "Credit", 1, 0),
         late = ifelse(Meal == "Late Night", 1, 0),
         young = ifelse(Age == "Yadult", 1, 0))
model1 <- lm(tippct ~ Bill + late + young + Bill:young,
             data = tips)
summary(model1)
confint(model1)
```

i. Is your model from (d) better than your model from (c)? Justify with summary statistics and/or a test of significance.

```
#create new model0 fit on same subset of data so we can
#  compare with model1
model0a <- lm(tippct ~ Bill, data = tips[!is.na(tips$Age),])

# Compare full vs reduced model using ESS F test and AIC/BIC

anova(model0a, model1, test="F")
AIC(model0a)                      # AIC-model0
AIC(model1)                       # AIC-model1
AIC(model0a, k=log(nrow(tips)))   # BIC-model0
AIC(model1, k=log(nrow(tips)))    # BIC-model1
```

Based on an extra sum of squares F test, model1 is significantly better than model0 ($F(3,417)=5.49$, $p=.001$). Model1 also has a higher adjusted r-squared (7.42% vs. 4.48%) and a lower AIC (2849 vs. 2859), although it does have a higher BIC (2873 vs. 2871).

ii. Create a plot illustrating the interaction between `bill` and `young`. Describe the general nature of the interaction.

```
# Illustrate young by bill interaction
ggplot(tips, aes(x = Bill, y = tippct,
                 color = as.factor(young))) +
  geom_point() +
  geom_smooth(method = "lm")
```

For young patrons, tip percentage decreases as bill size increases at a faster rate than older patrons, for whom tip percentage is less affected by bill size.

iii. Interpret each coefficient (including the intercept) in context. For `late`, interpret the 95% confidence interval rather than the estimated coefficient.

- Intercept: the mean tip percentage for older patrons eating lunch or dinner with no bill is 18.46%.
- Bill: for older patrons, every $10 increase in bill size is associated with an average decrease in tip percentage of 0.48%, after adjusting for meal.
- Late: we are 95% confident that the mean tip percentage for patrons eating late night is between 0.62% and 4.94% higher than those eating at other times, after adjusting for bill size and patron age.
- Young: when the bill size is $0, young patrons have mean tip percentage that is 2.29% higher than older patrons, after adjusting for meal.
- Interaction: the effect of every $10 increase in bill size on decreased tip percentage is 1.33 percentage points greater in young patrons than older ones, holding meal constant. For younger patrons, every $10 increase in bill size is associated with an average decrease in tip percentage of 1.81%, after adjusting for meal, compared to an average decrease of 0.48% in older patrons.

iv. The term for `young` is not significant. Should it be removed from our model? Why or why not?

No. We generally keep any main effect that is part of a significant interaction term, otherwise we are making assumptions about the intercept for the relationship between bill size and tip for young patrons.

v. Generate residual plots and comment on each of the model assumptions.

```
# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(model1, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

- (L) Linearity is reasonable based on no pattern around 0 in Resids vs. Fitted.
- (I) It is reasonable that the tipping of one patron does not affect that of another patron, although things like a poor cook one night can affect all the tips that day.
- (N) There is evidence of skewness, especially in the upper tails, as can be seen in the Normal QQ plot's deviation from linearity.
- (E) There is some evidence of increasing variability for the largest predicted tips, as seen in the increasing trend in Scale-Location and the "trumpet" shape to Resids vs. Fitted.

e) Now consider `Tip` as the primary response variable. Find the best model you can for `Tip` and explain how you arrived at that model. Do you prefer `Tip` or `tippct` as the response variable? Explain why.

```
# Fit new model with Tip as response
model2 <- lm(Tip ~ Bill + late + young + Bill:young + credit +
               credit:Bill + late:Bill, data = tips)
summary(model2)

# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(model2, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

I picked the model above using backward elimination after considering all main effects and two-way interactions. Three interactions with bill size were strongly significant, so I also included all main effects involved in those interactions, even if several were only borderline significant. Tip is an interesting response, because the coefficient for Bill is the estimated tip percentage. However, model assumptions are even more suspect than when using tip percentage as the response—greater evidence of increasing variability and non-linearity (especially with large predicted tips). Also, there's a lot of action in the interactions, which is interesting, but it means the predicted tip percentage is related to payment type, age, and meal type.

We must be careful with interpretations from this model. On the face of it, the size of a tip percentage is associated with several factors, including meal time, patron age, and method of payment. From the model coefficients, it appears that tip percentage is 3% higher for credit cards users (after controlling for age and meal time) and 5% higher for late night patrons (after controlling for age and payment type), but 4% lower for young patrons (after controlling for

payment type and meal time). However, this ignores the effect of intercepts—the fixed amount tipped for very low bills.

For example, according to the interactions, the best tippers (at 21% of the bill) should be older patrons eating late and paying with credit, while the worst tippers (at 9% of the bill) should be young patrons not eating late and paying with cash:

- Best tippers: Tip = -0.63 + 0.21*Bill
- Worst tippers: Tip = 1.53 + 0.09*Bill

However, if you isolate both subgroups, the "worst tippers" actually paid a higher tip percentage (14.5% vs. 14.0%), on a higher average bill ($29 vs. $14). The intercept makes a big difference, and prevents us from interpreting the multiplier in front of Bill as the tip percentage for a patron.

2

Beyond Least Squares: Using Likelihoods

```
# Packages required for Chapter 2
library(gridExtra)
library(knitr)
library(mosaic)
library(xtable)
library(kableExtra)
library(tidyverse)
```

2.1 Exercises

2.1.1 Conceptual Exercises

1. Large likelihood values are preferred, since it essentially measures the agreement between the data that was actually observed and potential values for p_B. The likelihood function, however, is not a probability function, so it does not have properties of probability density functions such as (a) the total area underneath it is 1, and (b) the area underneath it between any two points represents the probability of the interval between those points.

2. The parameter estimates for $p_{B|B_{bias}}$ and $p_{B|G_{bias}}$ would not be affected, but the estimate for $p_{B|N}$ would no longer include the only children (but would still include all first children as well as third children born into families with 1 boy and 1 girl). You could still perform a likelihood ratio test to compare sex unconditional and conditional models, but both models would have to be run on the same set of data (i.e., excluding only children).

3. You could, for example, create parameters for p_B that depend on the difference between previous boys and girls, expecting that the larger the difference, the more probabilities would tilt toward boys or girls.

2.1.2 Guided Exercises

1.

```
# Apply Model 1 to NLSY data (for families with 3 or fewer
#   children)
# possible values for prob a boy is born
pb <- seq(0, 1, length = 10001)
# loglik of getting obs data
loglik <- 5416 * log(pb) + 5256 * log(1 - pb)
# maximum loglikelihood over all values of pb
max(loglik)
# MLE of pb
pb[loglik==max(loglik)]
max_logL_m1_nlsy <- max(loglik)

# Model 0 is Model 1 with pb set equal to 0.5
maxloglik0 <- 5416 * log(0.5) + 5256 * log(1 - 0.5)
maxloglik0

# Model comparisons - Model 0 vs. Model 1
# likelihood ratio test statistic
lrt = 2 * (max(loglik) - maxloglik0)
lrt
# p-value for testing Ho: no diff between Models 0&1
1 - pchisq(lrt, df = 1)

# AIC and BIC values for Models 0 and 1
-2 * max(loglik) + 2 * 1   # aic1
-2 * max(loglik) + log(10672) * 1   # bic1
-2 * maxloglik0 + 2 * 0   # aic0
-2 * maxloglik0 + log(10672) * 0   # bic0
```

a) Write out the likelihood for your new model with NLSY data, recognizing that your new model should have 0 parameters to estimate.

Model 0: $L = 0.5^{5416}(1-0.5)^{5256}$

b) Compare your new model to Model 1 (Sex Unconditional) using a likelihood ratio test, along with AIC and BIC.

$LRT = 2 * [log(0.5^{5416}(1-0.5)^{5256}) - log(\hat{p}_B^{5416}(1-\hat{p}_B)^{5256})]$

There is not statistically significant evidence (LRT=2.40, df=1, p=.121) that allowing the probability of a boy to differ from 0.5 improves the model. Although AIC favors Model 1 slightly, BIC favors Model 0 with the probability

of boys and girls equal. For the sex unconditional model, allowing the probability of a boy to differ from .50 does not appear helpful.

2. **Case 3**

Find the MLE for p_B in two ways:

- a) Mathematically, using derivatives to maximize the log likelihood, and

$logL(p_B) = 6000log(p_B) + 4000log(1 - p_B)$

$\frac{d}{dp_B} logL(p_B) = \frac{6000}{p_B} - \frac{4000}{1-p_B} = 0$

$6000(1 - p_B) = 4000p_B \Rightarrow 6000 = 10000p_B \Rightarrow \hat{p}_B = \frac{6000}{10000} = .60$

- b) graphically, using a fine grid to find where the log likelihood function peaks

```
# possible values for prob a boy is born
pb <- seq(0, 1, length = 1000)
loglik <- 6000*log(pb) + 4000*log(1-pb)

# maximum likelihood over 1000 values of pb
max(loglik)
# value of pb where likelihood maximized
pb[loglik==max(loglik)]

# Repeat for Case 2
loglik <- 600*log(pb) + 400*log(1-pb)
max(loglik)
pb[loglik==max(loglik)]

# Repeat for Case 1
loglik <- 30*log(pb) + 20*log(1-pb)
max(loglik)
pb[loglik==max(loglik)]
```

As in Cases 1 and 2, the MLE of p_B is 0.60. The graph of the log-likelihood function in Case 3 would be even narrower than that in Figure 2.4(d), but still peaking at 0.60.

The log-likelihood at $p_B = .60$ for Case 3 is -6730, compared to -33.65 in Case 1 and -673 in Case 2. We cannot, however, compare these models with a LRT

because (a) they are fit using different data sets, and (b) they all use the same underlying model for the likelihood, so one is not a reduced version of another.

3.

a) $Lik(p_B = 0.5) = 0.5^{5416} * (1 - 0.5)^{5256}$ and $logLik(p_B = 0.5) = 5416log(0.5) + 5256log(1 - 0.5) = -7397$

b) $Lik(p_B = 0.45) = 0.45^{5416} * (1 - 0.45)^{5256}$ and $logLik(p_B = 0.45) = 5416log(0.45) + 5256log(1 - 0.45) = -7467$

c) $Lik(p_B = 0.55) = 0.55^{5416} * (1 - 0.55)^{5256}$ and $logLik(p_B = 0.55) = 5416log(0.55) + 5256log(1 - 0.55) = -7435$

d) $Lik(p_B = 0.5075) = 0.5075^{5416} * (1 - 0.5075)^{5256}$ and $logLik(p_B = 0.5075) = 5416log(0.5075) + 5256log(1 - 0.5075) = -7396$

$p_B = 0.5075$ would be our best estimate of these 4, since it produces the greatest (least negative) likelihood of getting the data we observed.

4.

$\widehat{p}_B$ is the total proportion of boys: $\widehat{p}_B = 5416/10672 = .5075$

$\widehat{p}_S$ is the proportion of times a couple stopped having children after a child was born, or the total number of families over the total number of children: $\widehat{p}_S = (930 + 951 + ...182 + 159)/10672 = 5626/10672 = .5272$

$p_{S|B1}$ is the proportion of times a couple stopped having children after having their first boy: $p_{S|B1} = .432$ from Section 2.7.

$p_{S|N}$ is the proportion of times a couple stopped having children after having a child that was not their first boy: $p_{S|N} = .584$ from Section 2.7.

The maximum likelihood for the Waiting for a Boy Model is -14661 from Table 2.12; for the Random Stopping model it is $5416log(.5075) + 5256log(1 - .5075) + 5626log(.5272) + 5046log(1 - .5272) = -14778$. The LRT is $-2(-14778 + 14661) = 234$ which produces a p-value of essentially 0. Thus we have statistically significant evidence that the Waiting for a Boy model outperforms the Random Stopping model, although it appears stopping after the first boy is less likely than stopping at random.

2.1.3 Open-Ended Exercises

1. **Another stopping rule model: balance-preference.**

a) Consider balance preference combined with a sex conditional model. Define your model parameters, and write out the likelihood contributions associated with family compositions in Table 2.3.

$B \Rightarrow p_{BN} p_{S|N}$ for 930 families.

$G \Rightarrow (1-p_{BN})p_{S|N}$ for 951 families.

$BB \Rightarrow p_{BN}(1-p_{S|N})p_{BB}p_{S|N}$ for 582 families.

$BG \Rightarrow p_{BN}(1-p_{S|N})(1-p_{BB})p_{S|B}$ for 666 families.

$GB \Rightarrow (1-p_{BN})(1-p_{S|N})p_{BG}p_{S|B}$ for 666 families.

$GGB \Rightarrow (1-p_{BN})(1-p_{S|N})(1-p_{BG})(1-p_{S|N})p_{BG}p_{S|B}$ for 125 families.

$GBB \Rightarrow (1-p_{BN})(1-p_{S|N})p_{BG}(1-p_{S|B})p_{BN}p_{S|B}$ for 151 families.

where $p_{S|N}$ is the probability of stopping when the family does NOT have at least one boy and one girl, $p_{S|B}$ is the probability of stopping when the family DOES have at least one boy and one girl, p_{BN} is the probability of a boy born into a neutral setting with equal boys and girls previously, p_{BB} is the probability of a boy born into a boy-biased setting with more boys than girls previously, and p_{BG} is the probability of a boy born into a girl-biased setting with fewer boys than girls previously.

b) Find the maximum likelihood estimates of your model parameters.

By taking partial derivatives of the loglikelihood, where the likelihood is given by:

$$\begin{aligned} L &= (p_{BN})^{3161}(1-p_{BN})^{3119}(p_{BB})^{1131}(1-p_{BB})^{1164}(p_{BG})^{1124}(1-p_{BG})^{973} \\ &\quad (p_{S|B})^{2288}(1-p_{S|B})^{654}(p_{S|N})^{3338}(1-p_{S|N})^{4392} \end{aligned}$$

we see that $\hat{p}_{BN} = .5033$, $\hat{p}_{BB} = .4928$, $\hat{p}_{BG} = .5360$, $\hat{p}_{S|B} = .778$, and $\hat{p}_{S|N} = .432$.

Note that:

$$\begin{aligned} 2288 &= 666+666+177+148+173+182+151+125 \\ 654 &= 148+173+182+151 \\ 3338 &= 930+951+582+530+186+159 \\ 4392 &= 582+666+666+530+186*2+177*2+148+173+182+ \\ &\quad 151+125*2+159*2 \end{aligned}$$

```
# fix parameters at MLEs
pbn <- .5033
pbb <- .4928
pbg <- .5360
psb <- .778
psn <- .432
max_logL_balpref <- 3161 * log(pbn) + 3119 * log(1 - pbn) +
```

TABLE 2.1: Data for Open-ended Exercise 2. (B = made basket. M = missed basket.)

Game	First 5 shots	Likelihood (No Hot Hand)	Likelihood (Hot Hand)
1	BMMBB	$p_B^3(1-p_B)^2$	$p_B^2(1-p_B)^1 p_{B\|B}^1(1-p_{B\|B})^1$
2	MBMBM	$p_B^2(1-p_B)^3$	$p_B^2(1-p_B)^1 p_{B\|B}^0(1-p_{B\|B})^2$
3	MMBBB	$p_B^3(1-p_B)^2$	$p_B^1(1-p_B)^2 p_{B\|B}^2(1-p_{B\|B})^0$
4	BMMMB	$p_B^2(1-p_B)^3$	$p_B^2(1-p_B)^2 p_{B\|B}^0(1-p_{B\|B})^1$
5	MMMMM	$p_B^0(1-p_B)^5$	$p_B^0(1-p_B)^5 p_{B\|B}^0(1-p_{B\|B})^0$
Total		$p_B^{10}(1-p_B)^{15}$	$p_B^7(1-p_B)^1 1 p_{B\|B}^3(1-p_{B\|B})^4$

```
  1131 * log(pbb) + 1164 * log(1 - pbb) +
  1124 * log(pbg) + 973 * log(1 - pbg) +
  2288 * log(psb) + 654 * log(1 - psb) +
  3338 * log(psn) + 4392 * log(1 - psn)
max_logL_balpref

ps <- (2288+3338) / 10672
max_logL_randomstop <- 3161 * log(pbn) + 3119 * log(1 - pbn) +
  1131 * log(pbb) + 1164 * log(1 - pbb) +
  1124 * log(pbg) + 973 * log(1 - pbg) +
  (2288+3338) * log(ps) + (654+4392) * log(1 - ps)
max_logL_randomstop

# perform LRT
# likelihood ratio test statistic
lrt = 2 * (max_logL_balpref - max_logL_randomstop)
lrt
# p-value for testing Ho: no diff between models
1 - pchisq(lrt, df = 1)
```

We have statistically significant evidence (LRT = 1074, p < .001) that the balance preference model (wait for at least one boy and one girl) is better than the random stopping model. However, our analysis is biased because we don't consider families with more than 3 kids; for example, not all families that reach GGB decide to stop at that point.

2. **The hot hand in basketball.**

a) See Table 2.1.

b) Because $\text{Lik}(p_B = .40) = (.40)^{10}(1-.40)^{15} = 4.93^{-8}$ is greater than $\text{Lik}(p_B = .30) = (.30)^{10}(1-.30)^{15} = 2.80^{-8}$.

c)

```
pb_mle_noHH = 10 / 15
pb_mle_HH = 7 / 18
pbb_mle_HH = 3 / 7

max_logL_noHH <- 10 * log(pb_mle_noHH) +
  15 * log(1 - pb_mle_noHH)
max_logL_noHH
max_logL_HH <- 7 * log(pb_mle_HH) + 11 * log(1 - pb_mle_HH) +
  3 * log(pbb_mle_HH) + 4 * log(1 - pbb_mle_HH)
max_logL_HH

lrt = 2 * (max_logL_HH - max_logL_noHH)
lrt
# p-value for testing Ho: no diff between models
1 - pchisq(lrt, df = 1)
```

We have statistically significant evidence (LRT = 7.45, p = .006) that the hot hand model is better than the no hot hand model. And in this case, the shooter made a higher percentage of shots after makes than after misses.

3

Distribution Theory

```
# Packages required for Chapter 3
library(gridExtra)
library(knitr)
library(kableExtra)
library(tidyverse)
```

3.1 Exercises

3.1.1 Conceptual Exercises

1. σ is largest when $p = 0.5$, and smallest (0, in fact) when $p = 0$ or $p = 1$

2. Both binomial and hypergeometric random variables count the number of successes in n trials, but binomial RVs sample with replacement while hypergeometric RVs sample without replacement. Given n, both variables can take on values no more than n, but hypergeometric RVs may be bounded by a smaller value m, if $m < n$, where m is the total number of "successes" among N objects.

3. Exponential RVs can be used to model "wait time" for an event in a Poisson process, where a Poisson RV counts the number of events in a fixed interval of that same Poisson process.

4. They both can be used to model phenomena ending after 1 event—exponential the wait time until the first event from a Poisson process, and geometric the number of failures until the first success from a Bernoulli process. Thus, geometric RVs are discrete while exponential RVs are continuous. They are also both special cases of other random variables from this chapter (negative binomial for geometric and gamma for exponential).

5. A hypergeometric distribution would be useful ($n = 5,\ N = 35,\ m = 10$).

6. A Poisson distribution—such distributions model counts of events within a unit of time or space, e.g., counts of pollutants per cubic foot.

7. We could use a binary distribution as it only takes on the values 0 and 1.

8. A Poisson distribution makes sense. We have a count per unit of time, and our data is likely to be skewed right (just like Poisson distributions).

9. Let Y be the time in seconds before 100 people click an online advertisement.

3.1.2 Guided Exercises

1. **Beta-binomial distribution.**

```
#generate 1000 binomial observations, n = 10, p = .8
binomYs <- rbinom(1:1000, 10, .8)
binom <- tibble(x = 1:1000, binomYs)
p1 <- ggplot(data = binom, aes(x = binomYs)) +
  geom_histogram(aes(y=..count../sum(..count..)),
                 binwidth = .25) +
  labs(x = "number of events", y = "probability",
       title = "Binomial, p = 0.8, n = 10") +
  xlim(-1,11)

#generate 1000 p_i from a beta distribution with a = 4, b = 1
pis <- rbeta(1:1000, 4, 1)
#generate 1000 binom variables, n = 10, p = p_i
betabinomYs <- rbinom(1:1000, 10, pis)
betabinom <- tibble(x = 1:1000, betabinomYs)
p2 <- ggplot(data = betabinom, aes(x = betabinomYs)) +
  geom_histogram(aes(y=..count../sum(..count..)),
                 binwidth = .25) +
  labs(x = "number of events", y = "probability",
       title = expression(paste("Beta-Binomial, ",
                  alpha, "=4, ", beta, "=1, n = 10"))) +
  xlim(-1,11)

mean(binom$binomYs)
mean(betabinom$betabinomYs)

sd(binom$binomYs)
sd(betabinom$betabinomYs)

grid.arrange(p1, p2, ncol = 1)
```

Both distributions take on integers between 0 and 10, and both have means around 8, but the beta-binomial has much more variability (SD around 1.99 compared to 1.29 for binomial). The binomial distribution is left-skewed but peaks at 8, while the beta-binomial actually peaks at 10 with a longer left-tail

2. **Gamma-Poisson mixture I.**

```
#normal poisson, lambda = 1.5
# generate 10,000 random poisson observations with lambda = 1.5
pois <- tibble(x = 1:10000, y = rpois(10000, 1.5))
p1 <- ggplot(data = pois, aes(x = y)) +
  geom_histogram(aes(y=..count../sum(..count..)),
                 binwidth = .25) +
  labs(x = "number of events", y = "probability",
       title = "Poisson Distribution") +
  xlim(-1,8)

#mixture distribution
# generate 10,000 poisson observations where lambda follows
# a gamma distribution
mix <- tibble(x = 1:10000,
              y = rpois(10000, rgamma(10000, shape = 3,
                                      rate = 2)))
p2 <- ggplot(data = mix, aes(x = y)) +
  geom_histogram(aes(y=..count../sum(..count..)),
                 binwidth = .25) +
  labs(x = "number of events", y = "probability",
       title = "Mixture Distribution") +
  xlim(-1,8)

grid.arrange(p1, p2, ncol = 1)

mean(pois$y)
mean(mix$y)

sd(pois$y)
sd(mix$y)
```

Both distributions have mean near 1.5, but the gamma-Poisson mixture has a greater SD than the plain Poisson (approximately 1.47 vs. 1.22).

3. **Gamma-Poisson mixture II.**

```
#negative binomial, r = 3, p = 2/3
NBinom <- tibble(x = 1:10000, y = rnbinom(10000, 3, 2/3))

p3 <- ggplot(data = NBinom, aes(x = y)) +
  geom_histogram(aes(y=..count../sum(..count..)),
                 binwidth = .25) +
  labs(x = "number of events", y = "probability",
       title = "Negative Binomial") +
  xlim(-1,8)

grid.arrange(p2, p3, ncol = 1)

mean(NBinom$y)
mean(mix$y)

sd(NBinom$y)
sd(mix$y)
```

The plots, means, and SDs for a negative binomial distribution with $r = 3$ and $p = 2/3$ looks equivalent to a gamma-Poisson mixture where the λ for the Poisson distribution is sampled from a gamma distribution with $r = 3$ and $\lambda' = 2 = \frac{p}{1-p}$.

4. **Mixture of two normal distributions**

```
#create likelihood function
lik <- function(mu1, sigma1, mu2, sigma2, alpha){
  x = faithful$waiting
  likelihoods = log(alpha*dnorm(x, mu1, sigma1) +
                      (1-alpha)*dnorm(x, mu2, sigma2))
  sum(likelihoods)
}

lik(55, 5, 80, 5, .33) #loglikelihood -1042


#using a package to estimate the MLE
library(mixtools)
multinorm <- normalmixEM(faithful$waiting, k=2)
multinorm$lambda
multinorm$mu
multinorm$sigma
plot(multinorm,which=2)
```

```
lines(density(faithful$waiting), lty=2)
lik(54.61, 5.87, 80.09, 5.87, .36)
#MLE estimate loglikelihood -1034
```

The values $\mu_1 = 55$, $\sigma_1 = 5$, $\mu_2 = 80$, $\sigma_2 = 5$, and $\alpha = .33$ produces a loglikelihood of -1042, which is greater than -1050. The function `normalmixEM()` from the package `mixtools` produces MLEs of $\mu_1 = 54.61$, $\sigma_1 = 5.87$, $\mu_2 = 80.09$, $\sigma_2 = 5.87$, and $\alpha = .36$ with a loglikelihood of -1034.

4

Poisson Regression

```
# Packages required for Chapter 4
library(maxLik)
library(gridExtra)
library(knitr)
library(kableExtra)
library(mosaic)
library(xtable)
library(pscl)
library(multcomp)
library(pander)
library(MASS)
library(tidyverse)
```

4.1 Exercises

4.1.1 Conceptual Exercises

1. Y = number of motorcycle deaths in a given year by state; X = helmet law indicator

2. Y = number of employers conducting interviews in a year; X = private or private school indicator

3. Y = number of daily asthma-related visits to an Emergency Room; Xs = air pollution indices

4. Y = number of deformed fish per square meter of a Minnesota lake; Xs = trace mineral change measurements

5. Models of the form

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

are fit using the method of maximum likelihood.

6. See if variance is different from the mean. Also, use goodness of fit tests.

7. Quasi-Poisson models adjust for overdispersion (variances which are greater than the mean) using quasi-likelihood methods by inflating the standard error of coefficients with an overdispersion parameter. The result is greater standard errors, smaller test statistics, and larger p-values for coefficients than regular Poisson regression using conventional likelihood methods.

8. In Poisson regression, we assume that the log of the mean count (λ) is linearly related to our predictors. Thus, we estimate $log(\lambda)$ with $log(\bar{Y})$.

9. Look at mean and variance of responses at different levels of X. You may need to bin observations together based on their X-values (e.g. 1-5, 6-10, etc.).

10. Yes, using quasilikelihood introduces the overdispersion parameter, which inflates the standard error for each parameter estimate, decreasing their t-values, and therefore increases their p-values, making them less significant.

11. **Fish (or, as they say in French, poisson).**

a) Number of fish caught during one week

b) 0 to number of fish in the park

c) The average number of fish caught per week and the variance from person to person's number of fish caught.

d) We would consider a ZIP model if there is a large number of people who did not catch fish, or who did not fish. True zeroes would be people who never fish in the park.

12. **Methadone program recidivism.**

a) Number of relapses

b) 0 to maximum number of relapses

c) The mean number of relapses and variance from person to person in number of relapses

d) This probably isn't a situation with a true zero, unless patients are lost to follow-up.

13. **Clutch size.**

a) Number of eggs in a nest

b) 0 to maximum number of eggs that can be laid

c) The mean number of eggs per nest and the variance from nest to nest in number of eggs

d) If there were many nests that had no eggs after the storm, we would use

ZIP. The true zeros are the nests wiped out by the storm, as opposed to nests not affected by the storm where no eggs were laid.

14. **Credit card use.**

Predictor: income. Response: Credit cards that they use.

For every $10,000 increase in income, the estimated mean number of credit cards used increases by a factor of 8.17 ($e^{2.1}$)—i.e., is 8.17 times greater.

Linearity can be assessed by breaking income into intervals and then plotting median income in an interval vs. log of the average number of credit cards in that interval.

To assess equal mean and variance assumption we could bin income into intervals and measure mean and variance of number of credit cards for each bin.

15. **Dating online.**

Response: Number of dates arranged online. Predictor: Age.

Offset: Time Online. An offset may make sense here, as time online should correspond with sampling effort in terms of arranging dates, so we're thinking of dates arranged per hour.

Our model would be:

$$log(\lambda_i) = \beta_0 + \beta_1 Age_i + log(TimeOnline_i)$$

where λ_i is the number of dates for subject i and may differ by age after adjusting for time online.

The true zeros would be those who create an online account, but do not participate in online dating.

16. **Poisson approximation: rare events.**

- A case: someone with non-melanoma skin cancer.
- The population size, n_i: the total population within a city and age group (the offset).
- Probability, p_i: the probability a random person in a specific city and age group has non-melanoma skin cancer.
- Poisson parameter, λ_i: the mean number of skin cancer cases per year based on city and age group.
- The random variable, Y_i: the number of cases of skin cancer within a specific city and age group.
- The predictors (2): the city and age group.

4.1.2 Guided Exercises

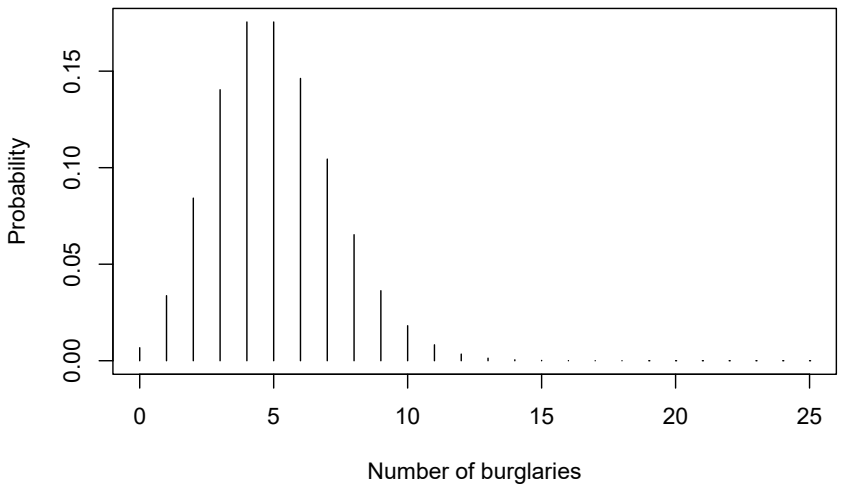
1. **College burglaries.**

a) Y is a count variable (taking on only integer values at or above 0) which represents the number of events per unit of time or space (time in this case).

b) Enrollment, location variables (urban/rural, surrounding neighborhood demographics), academic profile (average SAT, selectivity), church affiliation, proportion in campus housing, etc.

c)

```
x = 0:25
y = dpois(x, lambda=5)
plot(x,y,type="h",xlab="Number of burglaries",ylab="Probability")
```




```
burgs = rpois(10000,lambda=5)
mean(burgs)
```

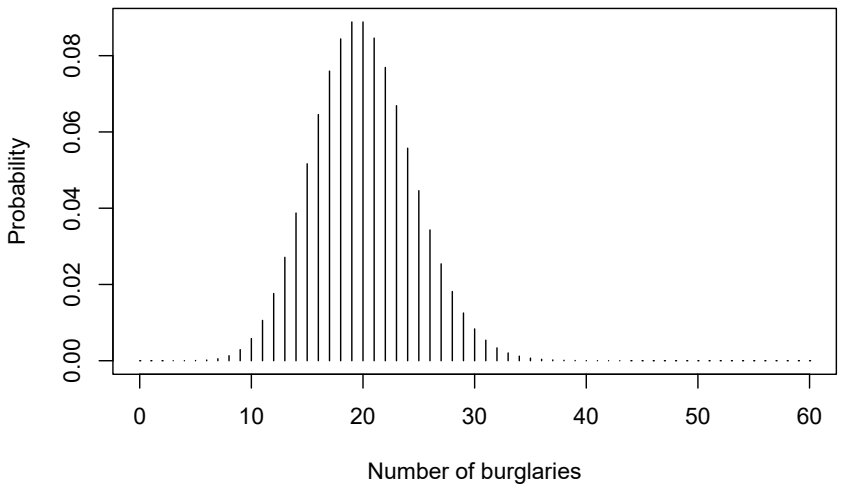
```
## [1] 5.019
```

```
var(burgs)
```

```
## [1] 5.177
```

d)

```
x = 0:60
y = dpois(x, lambda=20)
plot(x,y,type="h",xlab="Number of burglaries",ylab="Probability")
```




```
burgs = rpois(10000,lambda=20)
mean(burgs)
```

```
## [1] 19.97
```

```
var(burgs)
```

```
## [1] 19.98
```

2. **Elephant mating.** Article: [Poole, 1989]. Data source: [Ramsey and Schafer, 2002].

```
elephant <- read_csv("data/elephant.csv")
```

```
ggplot(elephant, aes(MATINGS)) + geom_histogram(binwidth = .25) +
  xlab("Number of Matings in a Year") +
  ylab("Count of Matings")
```

a) The data looks fairly skewed right, so there is evidence that using Poisson modeling may be useful in this case.

```
ggplot(elephant, aes(x=AGE, y=MATINGS)) +
  geom_point() +
  geom_smooth()

model_lm <- lm(MATINGS ~ AGE, data = elephant)

# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(model_lm, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

b) It appears that the equal variance condition may not be met for linear regression since the Scale-Location plot shows increasing variability as predicted values increase, and there's a small "funnel effect" in the scatterplot.

```
elephant1 <- elephant %>%
  group_by(AGE) %>%
  summarize(MATINGS = mean(MATINGS)) %>%
  mutate(log_matings = log(MATINGS+0.5))

ggplot(elephant1, aes(x=AGE, y=log_matings)) +
  geom_point() +
  geom_smooth(method = "loess", size = 1.5)
```

c) We can address the assumption of linearity—our predictor (AGE) should be linearly related to the log of mean responses. Quadratic terms are usually associated with max or mins, or faster growth in certain intervals.

```
modela <- glm(MATINGS ~ AGE, family = poisson, data = elephant)
summary(modela)
exp(modela$coefficients)
```

d) The mean number of matings increase by a factor of 1.07, or 7%, for every additional year older the elephant is.

```
exp(confint(modela))
```

e) We are 95% confident that the true mean number of elephant matings for an elephant of age 0 is between 0.07 and 0.59. We are 95% confident that average matings increase between 4.3% and 10.0% for every one year increase in elephant age.

```
model0 <- glm(MATINGS ~ 1, family = poisson, data = elephant)
anova(model0, modela, test = "Chisq")
```

f) We have statistically significant evidence ($Z = 4.997, p = 5.81e-7$ from a Wald test; $\chi^2 = 24.36, p = 7.991e-7$ from a drop in deviance test) that the average number of matings is related to age.

```
elephant <- elephant %>% mutate(agesq = AGE^2)
modelb <- glm(MATINGS ~ AGE + agesq, family = poisson,
              data = elephant)
summary(modelb)
anova(modela, modelb, test = "Chisq")
```

g) We do not have statistically significant evidence ($Z = -0.427, p = 0.669$ from a Wald test; $\chi^2 = 0.1854, p = 0.6667$ from a drop in deviance test) that a quadratic model outperforms a linear model.

```
# Goodness-of-fit test
gof.pvalue = 1 - pchisq(modela$deviance, modela$df.residual)
gof.pvalue
```

h) We do not have statistically significant evidence of lack of fit at the .05 level (residual deviance = 51.102, $p = Pr(\chi^2_{39} > 51.102) = .094$). LOF can be caused by (1) inadequate models – need new terms or transformed terms; (2) outliers; (3) overdispersion (variance greater than the mean). The GOF test is not powerful here – it's best with 5 or more reps at each combination of covariates.

```
modelq = glm(MATINGS ~ AGE, family = quasipoisson, data = elephant)
summary(modelq)
```

i) Since our goodness of fit test was a bit underpowered and our p-value was marginally significant, we fit a quasi-Poisson model to see if our model improves. The estimated coefficients do not change, but the standard errors increase in the quasi-Poisson model compared to the standard Poisson model. The estimated dispersion parameter is 1.157; since this is greater than 1, we are not as likely to find that coefficients are significantly different from 0.

3. **Smoking at work and home.**

a)

$$L(\lambda) = \frac{e^{-\lambda}\lambda^3}{3!}\frac{e^{-\lambda}\lambda^0}{0!}\frac{e^{-\lambda}\lambda^0}{0!}\frac{e^{-\lambda}\lambda^1}{1!}\frac{e^{-\lambda}\lambda^2}{2!}\frac{e^{-\lambda}\lambda^1}{1!} = \frac{e^{-6\lambda}\lambda^7}{3!0!0!1!2!1!}$$

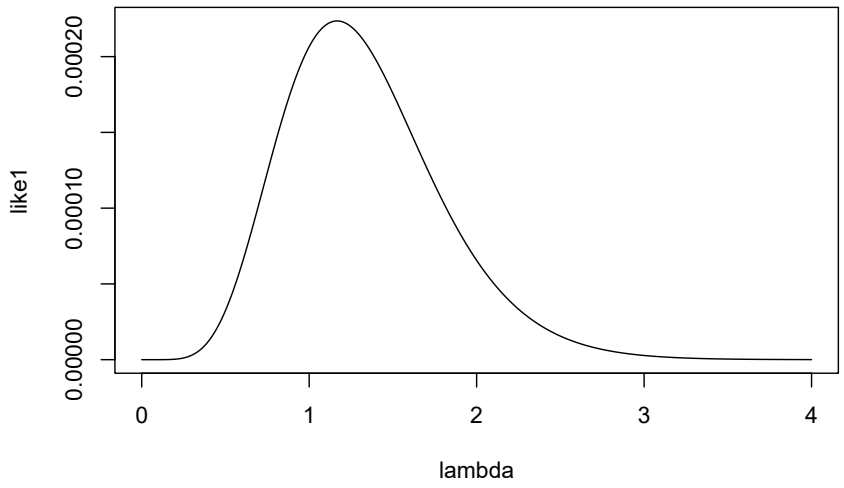
$$\begin{aligned} logL(\lambda) &= (-\lambda + 3log\lambda - log(3!)) + ... + (-\lambda + 1log\lambda - log(1!)) \\ &= (-6\lambda + 7log\lambda - log(3!0!0!1!2!1!)) \end{aligned}$$

b) Since λ represents the average number of cigarettes smoked in a 2-hour period, a natural estimate for λ is the average number of cigarettes smoked over the 6 subjects = 7/6 = 1.17.

c)

```
# Mock data
cigs=c(3,0,0,1,2,1)
work=c(0,1,1,1,0,0)

# Model 1: no difference between home and work

# Find lambda to maximize likelihood
lambda = seq(0,4,length=10001)
like1 = c()
for (i in 1:10001)  {
```

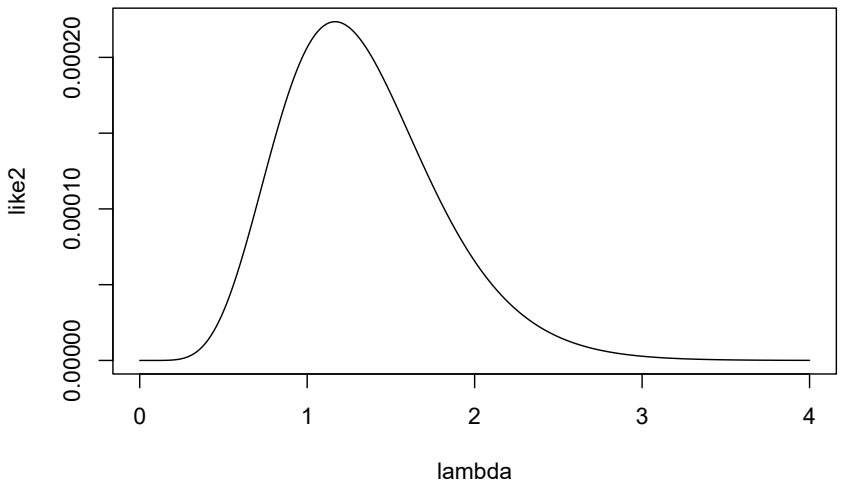
```
  like1[i] = prod( exp(-lambda[i]) * lambda[i]^cigs /
                     factorial(cigs) ) }
plot(lambda,like1,type="l")
```




```
lambda[like1==max(like1)]  # Method 1
```

```
## [1] 1.167
```

```
like2 = c()
for (i in 1:10001)  {
  like2[i] = prod( dpois(cigs,lambda[i]) ) }
plot(lambda,like2,type="l")
```




```
lambda[like2==max(like2)]  # Method 2
```

```
## [1] 1.167
```

```
f <- function (lambda) prod( exp(-lambda) * lambda^cigs /
  factorial(cigs) )
lambda.mle <- optimize(f, c(0, 4), tol = 0.0001, maximum=TRUE)
lambda.mle  # Method 3
```

```
## $maximum
## [1] 1.167
##
## $objective
## [1] 0.0002236
```

d)

$$
\begin{aligned}
L(\lambda_W, \lambda_H) &= \frac{e^{-\lambda_H}\lambda_H^3}{3!}\frac{e^{-\lambda_W}\lambda_W^0}{0!}\frac{e^{-\lambda_W}\lambda_W^0}{0!}\frac{e^{-\lambda_W}\lambda_W^1}{1!}\frac{e^{-\lambda_H}\lambda_H^2}{2!}\frac{e^{-\lambda_H}\lambda_H^1}{1!} \\
&= \frac{e^{-3\lambda}\lambda^6}{3!2!1!}\frac{e^{-3\lambda}\lambda^1}{0!0!1!}
\end{aligned}
$$

$logL(\lambda_W, \lambda_H)$

$= (-\lambda_H + 3log\lambda_H - log(3!)) + ... + (-\lambda_H + 1log\lambda_H - log(1!))$

$= (-3\lambda_H + 6log\lambda_H - log(3!2!1!)) + (-3\lambda_W + 1log\lambda_W - log(0!0!1!))$

e) As in (b), a natural estimate for λ_H is the average number of cigarettes smoked in a 2-hour period for the 3 subjects who were at home = 6/3 = 2.0, while a natural estimate for λ_W is the average number of cigarettes smoked in a 2-hour period for the 3 subjects who were at work = 1/3 = 0.33.

f) See associated R code.

```
# Model 2: separate lambdas for work and home

# Find lambdaW and lambdaH to maximize log-likelihood
lw = seq(0,1,length=101)
lh = seq(1,3,length=101)
xy = expand.grid(lw,lh)
z = c()
for (i in 1:(101*101) )  {
  lambdas = c(xy[i,2],xy[i,1],xy[i,1],xy[i,1],xy[i,2],xy[i,2])
  z[i] = prod( dpois(cigs,lambdas) )  }
zmat <- matrix(z,ncol=101)
persp(lw,lh,zmat,xlab="lambda-work",ylab="lambda-home",
      zlab="log-likelihood")
contour(lw,lh,zmat,xlab="lambda-work",ylab="lambda-home")
#image(lw,lh,zmat,xlab="lambda-work",ylab="lambda-home")
max(z)
xy[z==max(z),1:2]   # Method 1

loglik <- function(lambdas) {
  lw <- lambdas[1]
  lh <- lambdas[2]
  lambda6 = c(lh,lw,lw,lw,lh,lh)
  ll = prod( dpois(cigs,lambda6) )
  ll
}
mles = maxLik(loglik, start=c(1,1))
mles$maximum
mles   # Method 2
```

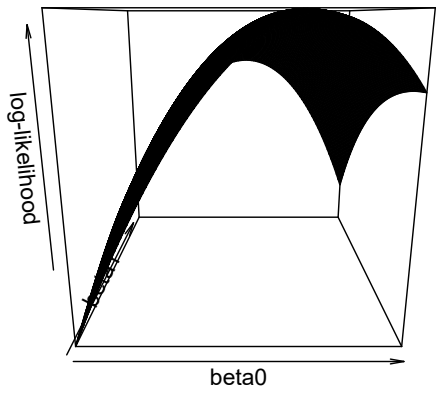
4. **Smoking at work and home (continued).**

a)

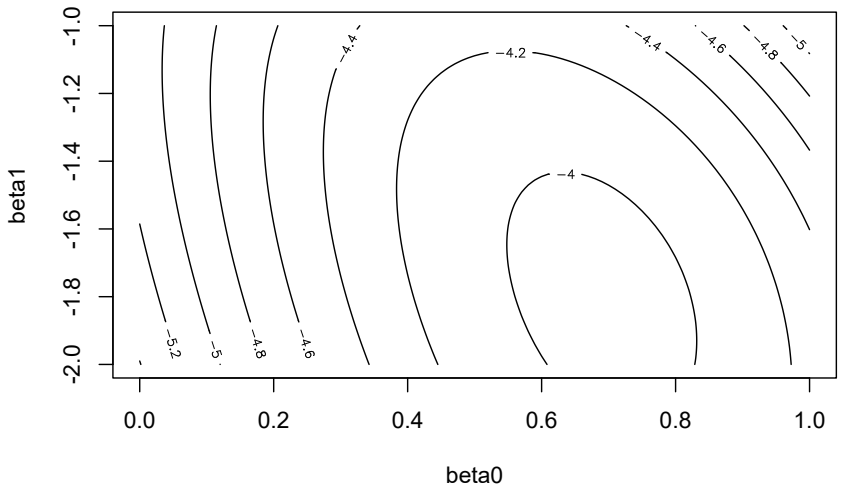
$$\begin{aligned} logL(\beta_0, \beta_1) &= (-e^{\beta_0+\beta_1 X_1} + 3(\beta_0 + \beta_1 X_1) - log(3!)) \\ &+ \ ... + (-e^{\beta_0+\beta_1 X_6} + 1(\beta_0 + \beta_1 X_6) - log(1!)) \end{aligned}$$

b)

```
# Model 3: include effect of work vs home

# Find beta0 and beta1 to maximize log-likelihood
b0 = seq(0,1,length=101)
b1 = seq(-2,-1,length=101)
xy = expand.grid(b0,b1)
z = c()
for (i in 1:(101*101) )  {
  z[i] = sum( -exp(xy[i,1]+xy[i,2]*work) +
                cigs*(xy[i,1]+xy[i,2]*work) ) }
zmat <- matrix(z,ncol=101)
persp(b0,b1,zmat,xlab="beta0",ylab="beta1",
      zlab="log-likelihood")
```




```
contour(b0,b1,zmat,xlab="beta0",ylab="beta1")
```




```
#image(b0,b1,zmat,xlab="beta0",ylab="beta1")
max(z)
```

```
## [1] -3.94
```

```
xy[z==max(z),1:2]   # Method 1
```

```
##      Var1  Var2
## 2191 0.69 -1.79
```

```
loglik <- function(beta) {
  b0 <- beta[1]
  b1 <- beta[2]
  ll = sum( -exp(b0+b1*work) + cigs*(b0+b1*work) -
              log(factorial(cigs)) )
  ll
}
mles = maxLik(loglik, start=c(0,0))
mles$maximum
```

```
## [1] -6.425
```

```
mles   # Method 2
```

```
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 5 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: -6.425 (2 free parameter(s))
## Estimate(s): 0.6931 -1.792
```

c)

```
# Confirm Model 1 and Model 3 estimates with glm()
fit1=glm(cigs~1, family=poisson)
summary(fit1)
exp(coef(fit1))

fit2=glm(cigs~work, family=poisson)
summary(fit2)
```

$log(\lambda) = \beta_0 + \beta_1(0) = \beta_0 = 0.69$ for subjects at home; thus, $\hat{\lambda}_H = e^{0.69} = 2.0$. Similarly, $log(\lambda) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1 = -1.10$ for subjects at work; thus, $\hat{\lambda}_W = e^{-1.10} = 0.33$.

d) $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 > 0$. We believe that the mean number of cigarettes smoked at work will be smaller than the number smoked at home (hypothetically due to more restrictive rules instituted at the workplace in MN).

e) No offset is needed here because every observation was taken over a two-hour time interval.

f)

```
exp(coef(fit2))
```

$\hat{\beta}_0 = 0.69$ and $e^{\hat{\beta}_0} = 2.00$. The mean number of cigarettes smoked (per 2-hour period) at home is 2.00.

$\hat{\beta}_1 = -1.79$ and $e^{\hat{\beta}_1} = 0.167$. The mean number of cigarettes smoked (per 2-hour period) at work is just 16.7% of the number smoked at home. In other

words, the mean number of cigarettes smoked at work is $e^{-1.10} = 0.33$, which is 83.3% lower than the mean number at home.

g)

```
confint(fit2)
exp(confint(fit2))
1/exp(confint(fit2))
```

A 95% CI for β_1 is (-4.730, -0.025). It can be interpreted in terms of exponentiated bounds $(e^{-4.730}, e^{-0.025}) = (.009, .976)$. We can be 95% confident that the mean number of cigarettes smoked at work (for each 2-hour period) is between 0.9% and 97.6% of the number smoked at home (i.e., 2.4% to 99.1% lower). Or, based on inverses, we can be 95% confident that the mean number of cigarettes smoked is between 1.025 and 113.3 times higher at home than at work.

h)

```
# Compare Model 1 to Model 3
anova(fit1, fit2, test="Chisq")
```

Test 1 – Wald-type test from `glm()`. We do not have significant evidence (at the .05 level) that the mean number of cigarettes smoked at work differs from the number smoked at home (t=-1.659, p=.0971). Although the one-sided test would be significant.

Test 2 – Drop-in-deviance test from `anova()`. We have significant evidence (at the .05 level) that the mean number of cigarettes smoked at work differs from the number smoked at home (drop-in-dev=3.9624, p=.04653). The drop-in-deviance test is probably a bit more reliable here, since it's less reliant on large sample sizes than Wald tests.

i)

```
# Assessing the goodness of fit of Model 2
gof <- 1-pchisq(fit2$deviance, fit2$df.residual)
gof
```

We have no significant evidence (chi-square goodness-of-fit = 3.244, df=4, p=.518) of lack-of-fit for Model 3; however, this test is not very powerful, especially with such a small sample size.

j) Although conceivably the study involved a random sample of Minnesota smokers (although this is not entirely clear), and our data set represents a random subset of the study sample, our small sample size leads to large standard errors, making inferences difficult. Plus small sample sizes are less likely to be representative of their populations.

k) There is no information suggesting that smokers were randomized to the home or work location. A randomization would be required to make a causal statement, otherwise confounding factors could be responsible for observed differences.

l) Some examples of improvements would be to gather data on the same smokers at both locations, use a longer surveillance time than 2 hours, and collect additional covariates such as the extent of the smoker's habit, their age, and the type of workplace.

5. **Campus crime.**

```
crime <- read_csv("data/campuscrime09.csv")
with(crime, summary(burg09))
summarize(crime, mean=mean(burg09))
summarize(crime, var=var(burg09))

ggplot(crime, aes(x = burg09)) +
  geom_histogram(bins = 15)

plt1 <- ggplot(crime, aes(x = total, y = burg09)) +
  geom_point() + geom_smooth(method = "lm")
plt2 <- ggplot(crime, aes(x = sat.tot, y = burg09)) +
  geom_point() + geom_smooth(method = "lm")
plt3 <- ggplot(crime, aes(x = act.comp, y = burg09)) +
  geom_point() + geom_smooth(method = "lm")
plt4 <- ggplot(crime, aes(x = tuition, y = burg09)) +
  geom_point() + geom_smooth(method = "lm")
plt5 <- ggplot(crime, aes(x = pct.male, y = burg09)) +
  geom_point() + geom_smooth(method = "lm")
plt6 <- ggplot(crime, aes(x = state, y = burg09)) +
  geom_boxplot()
grid.arrange(plt1, plt2, plt3, plt4, plt5, plt6, nrow = 2)

cor(as.matrix(dplyr::select(crime, burg09, total, sat.tot,
                            act.comp, tuition, pct.male)))
pairs(as.matrix(dplyr::select(crime, burg09, total, sat.tot,
                              act.comp, tuition, pct.male)))
```

a) Burglaries in 2009 count the number of events per unit of time, and its

distribution is right skewed, which suggests a Poisson model. However, its variance (1535) is much greater than its mean (44).

There seems to be a moderate positive correlation between number of burglaries and ACT/SAT scores (r=.26, .25) and total number of students (r=.30). There is a smaller negative correlation between burglaries and tuition (r=-.18), and minimal correlation between burglaries and percent of men (r=.08).

```
fitlm <- lm(burg09 ~ act.comp + tuition + pct.male + total,
            data=crime)
summary(fitlm)

# Residual plots
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
plot(fitlm, which = c(1, 2, 3, 5))
par(mfrow = c(1, 1), mar = c(5.1, 4.1, 4.1, 2.1))
```

b) Based on the residual plot, there is concern about model assumptions, including linearity (based on pattern around 0 in residuals vs. fitted values), normality (especially in the upper tail of the QQ plot), and equal variance (based on the slight upward trend in the scale-location plot). Plus, with linear regression, there's always the possibility of negative predicted responses.

```
fit0 <- glm(burg09 ~ 1, family = poisson, data=crime)
summary(fit0)

fit1 <- glm(burg09 ~ act.comp + tuition + pct.male + total,
            family=poisson, data=crime)
summary(fit1)
exp(coef(fit1))
```

c) Each $1 increase in tuition is associated with a .0039% (1-.999961) decrease in mean number of burglaries, after controlling for ACT scores, percent male, and total number of students. Or, each $1 decrease in tuition is associated with a 0.0039% (1/.999961) increase in mean number of burglaries, after controlling for ACT scores, percent male, and total number of students.

Each 1 point increase in the percentage of males is associated with a 1.3% (1-.987) decrease in the mean number of burglaries, after controlling for ACT scores, tuition, and total number of students. Or, a 1 point decrease in percent of males is associated with a 1.3% increase in burglaries (1/.987 = 1.013).

```
crime <- mutate(crime, tuition.thous = tuition/1000)
fit1a <- glm(burg09 ~ act.comp + tuition.thous + pct.male +
  total, family=poisson, data=crime)
summary(fit1a)
exp(coef(fit1a))
```

d) The new model is exactly the same in terms of overall performance (AIC, residual deviance) and Wald tests for individual terms. The only difference is that the coefficient and SE for tuition are both 1000 times greater.

Each $1000 increase in tuition is associated with a 3.82% (1-.9618) decrease in mean number of burglaries, after controlling for ACT scores, percent male, and total number of students. Or, each $1000 decrease in tuition is associated with a 3.97% (1/.9618) increase in mean number of burglaries, after controlling for ACT scores, percent male, and total number of students.

```
fit2 <- glm(burg09 ~ act.comp + tuition.thous + pct.male +
              offset(log(total)), family = poisson, data=crime)
summary(fit2)

fit2a <- glm(burg09 ~ act.comp + tuition.thous + pct.male +
               log(total), family = poisson, data=crime)
summary(fit2a)
confint(fit2a)
```

e) If each student has a certain probability of being burglarized, then more students would lead to more expected burglaries. Although that's more of a binomial idea than Poisson, where offsets usually refer to changes in amount of time or space over which events are counted.

The new model seems to be a worse fit, as measured by AIC and residual deviance.

We would expect the coefficient to be 1, yet the 95% confidence interval lies entirely below 1 (.500, .892), providing significant evidence that 1 is not the true coefficient for log(total).

```
crime <- mutate(crime, total.thous = total/1000)
fit3 <- glm(burg09 ~ act.comp + tuition.thous +
            total.thous + act.comp:tuition.thous +
            act.comp:total.thous, family = poisson,
            data = crime)
```

```
summary(fit3)
exp(coef(fit3))
exp(confint(fit3))
summary(crime$act.comp)
```

f) For a campus with a 0 for its average ACT composite score, and after controlling for total number of students, each additional $1000 in tuition is associated with a 15.6% increase in mean burglaries.

To interpret the interaction, we can evaluate the model at the Q1 of ACT composite (21.75) and Q3 (26.00):

- Q1: $log(\lambda)$ = 4.172 - .0127`tuition` - .0164`total`
- Q3: $log(\lambda)$ = 3.598 - .0435`tuition` + .0230`total`

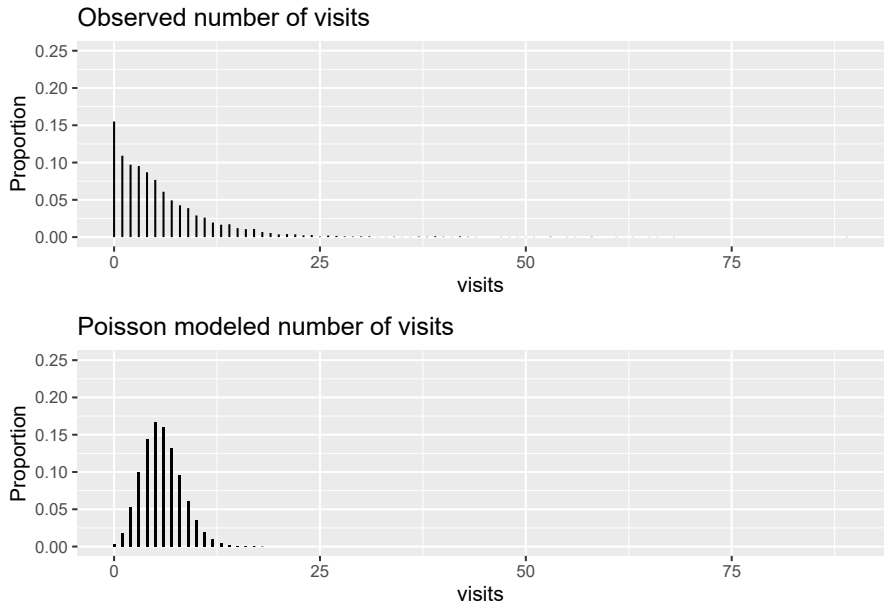
Thus, for campuses at the first quartile of ACT scores, each $1000 decrease in tuition is associated with a 1.3% increase (1/exp(-.0127) = 1/.9874) in mean burglaries, after adjusting for total number of students, while for campuses at the third quartile of ACT scores, each $1000 decrease in tuition is associated with a 4.4% increase (1/exp(-.0435) = 1/.9574) in mean burglaries.

6. **U.S. National Medical Expenditure Survey.**

a)

```
library(AER)
data(NMES1988)

p.table <- NMES1988 %>%
  group_by(visits) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n))

gmn.model <- glm(visits ~ 1, family = poisson, data = NMES1988)
# summary(gmn.model)
lambda <- exp(coef(gmn.model))
#  modeled pois probabilities
p.model <- dpois(NMES1988$visits, lambda)
nmes <- mutate(NMES1988, p.model = p.model)

linehist1 <- ggplot(p.table, aes(x = visits, xend = visits,
                                 y = 0, yend = freq)) +
  geom_segment() + ylim(0,.25) +
```

```
  labs(y = "Proportion", title = "Observed number of visits")
linehist2 <- ggplot(nmes, aes(x = visits, xend = visits,
                              y = 0, yend = p.model)) +
  geom_segment() + ylim(0,.25) +
  labs(y = "Proportion", title = "Poisson modeled number of visits")
grid.arrange(linehist1, linehist2, ncol=1)
```



b)

```
zip.m2 <- zeroinfl(visits ~ chronic + health + insurance |
                     chronic + insurance, data = nmes)
summary(zip.m2)
exp(coef(zip.m2))
1 / exp(coef(zip.m2))
.688 / (1 + .688)
```

- `chronic` in the Poisson part of the model

The expected number of physician visits increases by 12.6% for each additional chronic condition a person has, among older adults inclined to visit a physician, after controlling for health status and private insurance.

- poor `health` in the Poisson part of the model

A person in poor health is expected to have 34.3% more physician visits than a person in average health, among older adults inclined to visit a physician, after controlling for chronic conditions and private insurance.

- the Intercept in the logistic part of the model

The odds a person with no chronic conditions and no private insurance tends not to visit a physician is .688, which corresponds to a probability of .408.

- `insurance` in the logistic part of the model

A person with private insurance has 2.42 times greater odds of tending to visit a physician compared to a person without private insurance, after controlling for number of chronic conditions.

c)

```
pois.m1 <- glm(visits ~ chronic + health + insurance,
               family = poisson, data = nmes)
summary(pois.m1)
vuong(pois.m1,zip.m2)
```

Based on a Vuong Test, we have statistically significant evidence (Z = -18.2, p $<$ .001) that the ZIP model is an improvement over a simple Poisson regression model.

7. **Going vague: ambiguity in political issue statements.** [Chapp et al., 2018].

```
ambiguity <- read_csv("data/ambiguity.csv")

p.table <- ambiguity %>%
  group_by(totalIssuePages) %>%
  summarise (n = n()) %>%
  mutate(freq = n / sum(n))

ggplot(p.table, aes(x = totalIssuePages, xend = totalIssuePages,
                    y = 0, yend = freq)) +
  geom_segment() +
  labs(y = "Proportion", title = "Observed total issue pages")
```

a) The zeros here are not a mixture - they are only those candidates who have purposefully chosen not to post any issue statements on their webpage. But we can model number of issue pages among those who post at least one.

```
logodds_table <- ambiguity %>%
  filter(!is.na(ideology), !is.na(totalIssuePages)) %>%
  mutate(one_or_more_issues = totalIssuePages > 0,
         ideology_groups = cut_number(ideology, 10)) %>%
  group_by(ideology_groups) %>%
  summarise(p = mean(one_or_more_issues),
            odds = p / (1-p),
            log_odds = log(odds),
            median_ideology = median(ideology))

ggplot(data = logodds_table, aes(x = median_ideology,
                                 y = log_odds)) +
  geom_point() +
  geom_smooth(method = "lm")
```

b) In the logistic part of the hurdle model, there appears to be a positive linear relationship between ideology and the log odds of posting at least one issue statement.

```
logissue_table <- ambiguity %>%
  filter(!is.na(ideology), !is.na(totalIssuePages),
         totalIssuePages > 0) %>%
  mutate(ideology_groups = cut_number(ideology, 10)) %>%
  group_by(democrat, ideology_groups) %>%
  summarise(log_mean_issues = log(mean(totalIssuePages)),
            median_ideology = median(ideology))

ggplot(data = logissue_table, aes(x = median_ideology,
          y = log_mean_issues, color = as.factor(democrat))) +
  geom_point() +
  geom_smooth(method = "lm")
```

c) In the Poisson part of the hurdle model, there appears to be an interaction between ideology and party when modeling log mean issue statements among those who post at least one.

```
hurdle_model1 <- hurdle(totalIssuePages ~ democrat + ideology |
  democrat + ideology, data = ambiguity, dist="poisson",
  zero="binomial")
summary(hurdle_model1)
exp(coef(hurdle_model1))
1/.9941155
```

d) For each 1 unit decrease in ideology (more liberal), the mean number of issue pages (for candidates with at least one issue page) increases by 0.6%, holding party constant.

For each 1 unit increase in ideology (more conservative), the odds a candidate has at least one issue page increases by 77.6%, holding party constant.

```
hurdle_model2 <- hurdle(totalIssuePages ~ democrat + ideology +
  democrat:ideology | democrat + ideology + democrat:ideology,
  data = ambiguity, dist="poisson", zero="binomial")
summary(hurdle_model2)
exp(1.3667)
exp(-1.3995+1.3667)
```

e) For each 1 unit increase in ideology (more conservative), the odds a Republican candidate has at least one issue page is 3.92 times greater, while the odds a Democrat candidate has at least one issue page is 3.2% lower.

```
hurdle_model3 <- hurdle(totalIssuePages ~ democrat + incumbent +
  demHeterogeneity + mismatch + attHeterogeneity + distLean +
  ideology + ideology:democrat | democrat + incumbent +
  demHeterogeneity + mismatch + attHeterogeneity + distLean +
  ideology + ideology:democrat,
  data = ambiguity, dist="poisson", zero="binomial")
summary(hurdle_model3)
AIC(hurdle_model3)
BIC(hurdle_model3)
AIC(hurdle_model3, k=2)
AIC(hurdle_model3, k=log(870))
```

f) Candidates for US House were more likely to have at least one issue page if they were not an incumbent, if voter demographics were more homogeneous, if their ideology agreed with voter ideology, and if they were a republican

with right-leaning ideology or a democrat with left-leaning ideology. Of those candidates who had at least one issue page, the same profile marked candidates with a greater average number of issue pages.

4.1.3 Open-Ended Exercises

1. **Airbnb in NYC.** [Awad et al., 2017].

```
NYC <- read_csv("data/NYCairbnb.csv") %>%
  mutate(logprice = log(price),
         logreviews = log(number_of_reviews+0.5))
summary(NYC)

# to make EDA plots quicker
NYCsamp <- sample(NYC, 1000)

ggplot(NYCsamp,aes(x=number_of_reviews)) + geom_histogram()
ggplot(NYCsamp,aes(x=logreviews)) + geom_histogram()
NYCsamp %>% ggplot(aes(x=logprice)) + geom_histogram()

NYCsamp %>%
  filter(days < 10000) %>%
  ggplot(aes(days, logreviews)) +
    geom_point() +
    geom_smooth(method=lm)

ggplot(NYCsamp, aes(logprice,logreviews)) +
  geom_point() +
  geom_smooth(method=lm)
ggplot(NYCsamp, aes(logprice,logreviews,
                color=as.factor(room_type))) +
  geom_point() +
  geom_smooth(method=lm)

ggplot(NYCsamp, aes(bedrooms,logreviews)) +
  geom_point() +
  geom_smooth(method=lm)
ggplot(NYCsamp, aes(bathrooms,logreviews)) +
  geom_point() +
  geom_smooth(method=lm)

ggplot(NYCsamp, aes(review_scores_cleanliness,logreviews)) +
  geom_point() +
```

```
  geom_smooth(method=lm)
ggplot(NYCsamp, aes(review_scores_location,logreviews)) +
  geom_point() +
  geom_smooth(method=lm)
ggplot(NYCsamp, aes(review_scores_value,logreviews)) +
  geom_point() +
  geom_smooth(method=lm)

#instantly bookable
NYC %>%
  group_by(room_type) %>%
  summarise(n = n(),
            mean_reviews = mean(number_of_reviews),
            median_reviews = median(number_of_reviews),
            mean_price = mean(price))
#room type
NYC %>%
  group_by(instant_bookable) %>%
  summarise(n = n(),
            mean_reviews = mean(number_of_reviews),
            median_reviews = median(number_of_reviews),
            mean_price = mean(price))
```

```
pois1 <- glm(number_of_reviews ~ logprice + room_type,
             offset = log(days), family = poisson, data=NYC)
summary(pois1)

pois2 <- glm(number_of_reviews ~ logprice + room_type +
             bedrooms + bathrooms, offset = log(days),
             family = poisson, data=NYC)
summary(pois2)

pois3 <- glm(number_of_reviews ~ logprice + room_type +
             bedrooms + bathrooms + instant_bookable,
             offset = log(days), family = poisson, data=NYC)
summary(pois3)

pois4 <- glm(number_of_reviews ~ logprice + room_type +
             bedrooms + bathrooms + instant_bookable,
             offset = log(days), family = poisson, data=NYC)
summary(pois4)
```

```
pois5 <- glm(number_of_reviews ~ bedrooms + logprice +
  room_type + review_scores_cleanliness +
  review_scores_location + review_scores_value +
  instant_bookable + bathrooms, offset = log(days),
  family = poisson, data = NYC)
summary(pois5)

pois_final <- glm(number_of_reviews ~ bedrooms + logprice +
  room_type + review_scores_cleanliness +
  review_scores_location + review_scores_value +
  instant_bookable + bathrooms, offset = log(days),
  family = quasipoisson, data = NYC)
summary(pois_final)
exp(coef(pois_final))
```

An offset of `days` should be included because we expect units that have been listed longer have more time to accumulate reviews, thus we must adjust for `days`. Furthermore, we notice in `pois1` - `pois3` that all the z-values are extremely large because of our large sample size (despite losing about 1/4 of our data by including review scores data). We use quasi-Poisson regression to adjust for lack of fit due to extra-Poisson variation and find a dispersion parameter of 43.6, thus we should use quasi-Possion regression over Poisson regression to adjust for overdispersion.

The following interpretations focus on $e^{\hat{\beta}}$, except `logprice` which uses $2^{\hat{\beta}}$ since X is logged:

`bedrooms`: 1.060 - An increase in bedrooms of one corresponds to an increase in mean number of reviews by 6.0%, when controlling for price, room type, cleanliness, location, value, instant bookable, and number of bathrooms.

`logprice`: 1.041 - A doubling of price corresponds to a 4.1% increase in mean number of reviews, when controlling for number of bedrooms, room type, cleanliness, location, value, instant bookable, and number of bathrooms.

`room_typePrivate room`: 1.094 - A private room has on average 9.4% more reviews than an entire home/apartment, when controlling for number of bedrooms, price, cleanliness, location, value, instant bookable, and number of bathrooms.

`room_typeShared room`: 1.004 - A shared room has on average 0.4% more reviews than an entire home/apartment when controlling for number of bedrooms, price, cleanliness, location, value, instant bookable, and number of bathrooms.

`review_scores_cleanliness`: 1.116 - A one point increase in cleanliness score

corresponds to a mean increase in number of reviews by 11.6% when controlling for number of bedrooms, price, room type, location, value, instant bookable, and number of bathrooms.

`review_scores_location`: 0.911 - A one point increase in location score corresponds to a mean decrease in number of reviews by 8.9% when controlling for number of bedrooms, price, room type, cleanliness, value, instant bookable, and number of bathrooms.

`review_scores_value`: 0.918 - A one point increase in value score corresponds to a mean decrease in number of reviews by 8.2% when controlling for number of bedrooms, price, room type, cleanliness, location, instant bookable, and number of bathrooms.

`instant_bookableTRUE`: 1.769 - A room that is instantly bookable has on average 76.9% more reviews than a room that is not instantly bookable when controlling for number of bedrooms, price, room type, cleanliness, location, value, and number of bathrooms.

`bathrooms`: 0.915 - Each additional bathroom corresponds to a mean decrease in number of reviews by 8.5% when controlling for number of bedrooms, price, room type, cleanliness, location, value, and instant bookable.

2. **Crab satellites.** [Brockmann, 1996].

```
crab <- read_csv("data/crab.csv")
summary(crab)
cor(crab)  # Width and Weight highly correlated (r=.89)

crab %>% count(Spine)
crab %>% count(Color)
crab %>% count(Satellite)
ggplot(crab, aes(x = Satellite)) + geom_histogram()

crab <- crab %>%
  mutate(Color = Color - 1,
         Color_fac = dplyr::recode(Color, `1` = "light medium",
           `2` = "medium", `3` = "dark medium", `4` = "dark"),
         Spine_fac = dplyr::recode(Spine, "both good",
            "one worn", "both worn"),
         Weight_fac = ifelse(Weight < 2350, "low weight",
            "high weight"))
ggplot(data = crab, aes(x = Color_fac, fill = Spine_fac)) +
  geom_bar(position = "fill")

ggplot(crab, aes(x = Width, y = Satellite)) +
```

```
  geom_point() +
  geom_smooth(method=lm)
ggplot(crab, aes(x = Weight, y = Satellite)) +
  geom_point() +
  geom_smooth(method=lm)
ggplot(crab, aes(x = Satellite, color = Spine_fac)) +
  geom_density()
ggplot(crab, aes(x = Satellite, color = Color_fac)) +
  geom_density()

ggplot(crab, aes(x = Width, y = Satellite,
                 color = Spine_fac)) +
  geom_point() +
  geom_smooth(method=lm)
ggplot(crab, aes(x = Width, y = Satellite,
                 color = Color_fac)) +
  geom_point() +
  geom_smooth(method=lm)
ggplot(crab, aes(x = Width, y = Satellite,
                 color = Weight_fac)) +
  geom_point() +
  geom_smooth(method=lm)

mod1 <- glm(Satellite ~ Width + Weight + Spine_fac +
            Color_fac, family = poisson, data = crab)
summary(mod1)
mod2 <- glm(Satellite ~ Width + Spine_fac + Color_fac,
            family = poisson, data = crab)
summary(mod2)
mod3 <- glm(Satellite ~ Width + Spine_fac + Color_fac +
              Width:Spine_fac + Width:Color_fac,
            family = poisson, data = crab)
summary(mod3)
mod4 <- glm(Satellite ~ Width + Color_fac + Width:Color_fac,
            family = poisson, data = crab)
summary(mod4)
mod4_quasi <- glm(Satellite ~ Width + Color_fac +
  Width:Color_fac, family = quasipoisson, data = crab)
summary(mod4_quasi)
mod5_quasi <- glm(Satellite ~ Width + Color_fac,
                  family = quasipoisson, data = crab)
summary(mod5_quasi)
mod5a_quasi <- glm(Satellite ~ Width + Spine_fac,
                  family = quasipoisson, data = crab)
```

```
summary(mod5a_quasi)
mod6_quasi <- glm(Satellite ~ Width,
                  family = quasipoisson, data = crab)
summary(mod6_quasi)

hurdle1 <- hurdle(Satellite ~ Width + Color_fac + Spine_fac |
  Width + Color_fac + Spine_fac,
  data = crab, dist = "poisson", zero = "binomial")
summary(hurdle1)
hurdle2 <- hurdle(Satellite ~ Width + Color_fac +
  Width:Color_fac | Width + Color_fac + Width:Color_fac,
  data = crab, dist = "poisson", zero = "binomial")
summary(hurdle2)
hurdle3 <- hurdle(Satellite ~ Width + Color_fac |
  Width + Color_fac,
  data = crab, dist = "poisson", zero = "binomial")
summary(hurdle3)
```

From the EDA, we discover that satellites increase with both increasing width and weight, but that width and weight are highly correlated, so we'll only include our explanatory variable of interest (width) in models. We also see that spines with both good and light medium color have the most satellites, while dark color has the least. However, light medium crabs are also highly likely to have both good spines, so we will probably not use both spine and color together in a model.

Modeling number of satellites using Poisson regression, adjusting for overdispersion becomes necessary, and after adjusting for extra-Poisson variation, width is the only significant predictor (greater width is associated with more satellites). A better approach appears to be using a hurdle model—modeling those crabs with and without any satellites, and then also modeling satellite counts among those with at least one. In that case, after adjusting for color, we find width to be significantly associated ($Z = 4.434$, $p < .001$) with having at least one satellite, but only marginally associated ($Z = 1.783$, $p = .0747$) with more satellites among those with at least one. Each extra cm in width is associated with 59.7% greater odds of having at least one satellite and with a 4.0% mean increase in number of satellites among crabs with at least one.

3. **Doctor visits.** [Cameron and Trivedi, 1986].

```
#Install the package AER.
library(AER)
data("DoctorVisits")
```

```
doc <- DoctorVisits
head(doc)

summary(doc)
doc %>%
  dplyr::select(visits, age, income, illness,
                reduced, health) %>%
  cor()
favstats(visits ~ gender, data = doc)
favstats(visits ~ private, data = doc)
favstats(visits ~ freepoor, data = doc)
favstats(visits ~ freerepat, data = doc)
favstats(visits ~ nchronic, data = doc)
favstats(visits ~ lchronic, data = doc)

ggplot(doc, aes(x=visits)) + geom_histogram()

ggplot(doc, aes(y=visits, x=reduced)) +
  geom_point() +
  geom_smooth(method=lm)

ggplot(doc, aes(x = visits, color = lchronic)) +
  geom_density()

# All predictors
mod1 <- glm(visits ~ gender + age + income + illness + reduced +
  health + private + freepoor + freerepat + nchronic + lchronic,
  family = poisson, data = doc)
summary(mod1)

# Backward selection and quasipoisson
mod2 <- glm(visits ~ gender + age + income + illness + reduced +
  health + freepoor, family = quasipoisson, data = doc)
summary(mod2)

# All predictors in both parts
zip.model1<-zeroinfl(visits ~ gender + age + income + illness +
  reduced + health + private + freepoor + freerepat + nchronic +
  lchronic | gender + age + income + illness + reduced +
  health + private + freepoor + freerepat + nchronic + lchronic,
  data = doc)
summary(zip.model1)

# Backward selection
```

```
zip.model2 <- zeroinfl(visits ~ income + illness + reduced +
  health + freepoor + freerepat | gender + age + illness +
  reduced + health + private + freerepat, data = doc)
summary(zip.model2)
```

With a ZIP model, we assume the zeros are comprised of individuals who don't visit doctors (true zeros) and those who would visit a doctor but just didn't over the past two weeks. Under this assumption, the true zeros are more likely to be male, younger, have fewer illnesses and days of reduced activity, and have no private insurance and no free insurance due to old age or veteran status. Among those who would visit a doctor, mean visits increase with lower income, more illnesses and days of reduced activity, worse general health, and no free insurance for low income or old age or veteran status.

Quasi-Poisson models without a zero-inflation piece highlight similar associations with more visits: sicker, poorer, older, female, and less free insurance.

4. **More fish.** [UCLA Statistical Consulting Group, 2018].

```
fish2 <- read_csv("data/fish2.csv") %>%
  mutate(camper_fac = ifelse(camper == 1, "camper",
                             "no camper"),
         other_ppl = persons-1,
         logcount = log(count + 0.5),
         logLOS = log(LOS),
         adults = persons - child)
head(fish2)
summary(fish2)

ggplot(fish2, aes(x=count)) + geom_histogram()

ggplot(fish2, aes(x=persons, y=logcount, color=camper_fac)) +
  geom_point() +
  geom_smooth(method=lm)

ggplot(fish2, aes(x=child, y=logcount)) +
  geom_point() +
  geom_smooth(method=lm)

ggplot(fish2, aes(x=logLOS, y=logcount)) +
  geom_point() +
  geom_smooth(method=lm)
```

```
zip1 <- zeroinfl(count ~ camper + adults + child  |
  camper + adults + child , offset = log(LOS), data = fish2)
summary(zip1)

zip2 <- zeroinfl(count ~ adults |
  camper + adults + child , offset = log(LOS), data = fish2)
summary(zip2)
exp(coefficients(zip2))
1/ exp(coefficients(zip2))
```

A ZIP model is natural here, since the true zeros would represent camping parties where no one fished, and there would still be parties where they fished but didn't catch anything. In the final ZIP model, parties without a camper have 3.29 times greater odds of not fishing after controlling for adults and children in the party, each 1 fewer adult is associated with 2.63 times higher odds of not fishing after controlling for campers and number of children, and each extra child is associated with 3.23 times greater odds of not fishing after controlling for campers and number of adults. Among those who fished, each additional adult is associated with a mean increase of 60.4% in number of fish caught. No other factors were signficantly related to number of fish caught.

5

Generalized Linear Models: A Unifying Theory

5.1 Exercises

1a) Binary: Y = 1 for a success, 0 for a failure

$$
\begin{aligned}
f(y) &= p^y(1-p)^{1-y} \\
&= \exp\left[\log\left(p^y(1-p)^{1-y}\right)\right] \\
&= \exp\left[y\log(p) + (1-y)\log(1-p)\right] \\
&= \exp\left[y\log(p) + \log(1-p) - y\log(1-p)\right] \\
&= \exp\left[y\log\left(\frac{p}{1-p}\right) + \log(1-p)\right]
\end{aligned}
$$

Thus,

$a(y) = y$

$b(p) = \log\left(\frac{p}{1-p}\right)$ = canonical link

$c(p) = \log(1-p)$

$d(y) = 0$

This distribution could be useful for modeling probabilities of heart disease.

$$\begin{aligned}\mu &= -c'(p)/b'(p)\\ &= -\left(\frac{-1}{1-p}\right)\Big/\left(\frac{1}{p-p^2}\right)\\ &= \left(\frac{1}{1-p}\right)\left(\frac{p-p^2}{1}\right)\\ &= \frac{(1-p)p}{1-p}\\ &= p\end{aligned}$$

$$\begin{aligned}\sigma^2 &= (b''(p)c'(p) - c''(p)b'(p))/b'(p)^3\\ &= \left[\left(\frac{2p-1}{(p-1)^2p^2}\right)\left(\frac{-1}{1-p}\right) - \left(\frac{-1}{(1-p)^2}\right)\left(\frac{1}{p-p^2}\right)\right]\Big/\left(\frac{1}{p-p^2}\right)^3\\ &= \cdots = p(1-p)\end{aligned}$$

b) Binomial (for fixed n): Y = number of successes in n independent, identical trials

$$\begin{aligned}f(y) &= \binom{n}{y}p^y(1-p)^{n-y}\\ &= \exp\left(\log\left(\binom{n}{y}p^y(1-p)^{n-y}\right)\right)\\ &= \exp\left(\log\binom{n}{y} + y\log(p) + (n-y)\log(1-p)\right)\\ &= \exp\left(\log\binom{n}{y} + y\log(p) - y\log(1-p) + n\log(1-p)\right)\\ &= \exp\left(\log\binom{n}{y} + y\log\left(\frac{p}{1-p}\right) + n\log(1-p)\right)\\ &= \exp\left(y\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{y}\right)\end{aligned}$$

Thus,

$a(y) = y$

$b(p) = \log\left(\frac{p}{1-p}\right) =$ canonical link

$c(p) = n\log(1-p)$

$d(y) = \log\binom{n}{y}$

We could use a binomial distribution to model the number of free throws a basketball player makes out of a fixed number of shots.

$\mu = -c'(p)/b'(p) = np$

$\sigma^2 = (b''(p)c'(p) - c''(p)b'(p))/b'(p)^3 = np(1-p)$

c) Poisson: Y = number of events occurring in a given time (or space) when the average event rate is λ per unit of time (or space)

$$\begin{aligned} f(y) &= e^{-\lambda}\lambda^y/y! \\ &= \exp\left(\log\left(e^{-\lambda}\lambda^y/y!\right)\right) \\ &= \exp\left(-\lambda\log(e) + y\log(\lambda) - \log(y!)\right) \\ &= \exp\left(y\log(\lambda) - \lambda - \log(y!)\right) \end{aligned}$$

Hence,

$a(y) = y$

$b(\lambda) = \log(\lambda) =$ canonical link

$c(\lambda) = -\lambda$

$d(y) = -\log(y!)$

The Poisson distribution could be used to model the number of potholes per mile on Minnesota roads.

$\mu = -c'(\lambda)/b'(\lambda) = \lambda$

$\sigma^2 = (b''(\lambda)c'(\lambda) - c''(\lambda)b'(\lambda))/b'(\lambda)^3 = \lambda$

d) Normal (with fixed σ – could set $\sigma = 1$ without loss of generality)

$$\begin{aligned} f(y) &= \frac{e^{-(y-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \\ &= \exp\left[\log\left(\frac{e^{-(y-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}\right)\right] \\ &= \exp\left[-(y-\mu)^2/(2\sigma^2)\log(e) - \log\left(\sqrt{2\pi\sigma^2}\right)\right] \\ &= \exp\left[\frac{-(y^2 - 2y\mu + \mu^2)}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right)\right] \\ &\propto \exp\left[\left(\frac{y}{\sigma^2}\right)\mu - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2}\right] \end{aligned}$$

So, with $\sigma = 1$,

$a(y) = y$

$b(\mu) = \mu =$ canonical link

$c(\mu) = -\frac{1}{2}\mu^2$

$d(y) = -\frac{1}{2}y^2$

The normal distribution could be used to model weights of cereal boxes coming off an assembly line; here we know the variability based on past machine performance and are interested in the true mean weight.

$\mu = -c'(\mu)/b'(\mu) = \mu$

$\sigma^2 = (b''(\mu)c'(\mu) - c''(\mu)b'(\mu))/b'(\mu)^3 = \sigma^2 = 1$

e) Normal (with fixed μ – could set $\mu = 0$ without loss of generality)

$$
\begin{aligned}
f(y) &= \frac{e^{-(y-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \\
&= \exp\left[\log\left(\frac{e^{-(y-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}\right)\right] \\
&= \exp\left[-(y-\mu)^2/(2\sigma^2)\log(e) - \log\left(\sqrt{2\pi\sigma^2}\right)\right] \\
&= \exp\left[\frac{-(y^2 - 2y\mu + \mu^2)}{2\sigma^2} - \log\left(\sqrt{2\pi\sigma^2}\right)\right] \\
&\propto \exp\left[-\frac{y^2}{2\sigma^2} - log\sigma\right]
\end{aligned}
$$

So, with $\mu = 0$,

$a(y) = -\frac{1}{2}y^2$

$b(\sigma) = \frac{1}{\sigma^2} =$ canonical link

$c(\sigma) = -log\sigma$

$d(y) = 0$

The normal distribution could be used to model weights of cereal boxes coming off an assembly line; here we know the mean based on machine settings and are interested in the true variability in weight.

$\mu = -c'(\sigma)/b'(\sigma) = \mu = 0$

$\sigma^2 = (b''(\sigma)c'(\sigma) - c''(\sigma)b'(\sigma))/b'(\sigma)^3 = \sigma^2$

f) Exponential: Y = time spent waiting for the first event in a Poisson process with an average rate of λ events per unit of time

$$
\begin{aligned}
f(y) &= \lambda e^{-\lambda y} \\
&= \exp\left[\log\left(\lambda e^{-\lambda y}\right)\right] \\
&= \exp\left[\log(\lambda) - \lambda y \log(e)\right] \\
&= \exp\left[-\lambda y + \log(\lambda)\right]
\end{aligned}
$$

Then,

$a(y) = -y$
$b(\lambda) = \lambda =$ canonical link
$c(\lambda) = log(\lambda)$
$d(y) = 0$

The exponential distribution can be used to model the number of miles traveled until encountering the first pothole on a Minnesota road.

$\mu = -c'(\lambda)/b'(\lambda) = 1/\lambda$
$\sigma^2 = (b''(\lambda)c'(\lambda) - c''(\lambda)b'(\lambda))/b'(\lambda)^3 = 1/\lambda^2$

g) Gamma (for fixed r): Y = time spent waiting for the r^{th} event in a Poisson process with an average rate of λ events per unit of time

$$
\begin{aligned}
f(y) &= \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y} \\
&= \exp\left[\log\left(\frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y}\right)\right] \\
&= \exp\left[r\log(\lambda) - \log\left(\Gamma(r)\right) + (r-1)\log(y) - \lambda y \log(e)\right] \\
&\propto \exp\left[-\lambda y + r\log(\lambda) + (r-1)\log(y)\right]
\end{aligned}
$$

Thus,

$a(y) = -y$
$b(\lambda) = \lambda =$ canonical link
$c(\lambda) = r\log(\lambda)$
$d(y) = (r-1)\log(y)$

The gamma distribution can be used to model the number of miles traveled until encountering 10 potholes on a Minnesota road.

$\mu = -c'(\lambda)/b'(\lambda) = r/\lambda$
$\sigma^2 = (b''(\lambda)c'(\lambda) - c''(\lambda)b'(\lambda))/b'(\lambda)^3 = r/\lambda^2$

h) Geometric: Y = number of failures before the first success in a Bernoulli process

$$
\begin{aligned}
p(y) &= (1-p)^y p \\
&= \exp\left[\log\left((1-p)^y p\right)\right] \\
&= \exp\left[y\log(1-p) + \log(p)\right]
\end{aligned}
$$

Hence,

$a(y) = y$

$b(p) = \log(1-p) =$ canonical link

$c(p) = \log(p)$

$d(y) = 0$

A geometric distribution can be used to model the number of random people you call who decline before someone agrees to complete a survey.

$$\mu = -c'(p)/b'(p) = \frac{1-p}{p}$$

$$\sigma^2 = (b''(p)c'(p) - c''(p)b'(p))/b'(p)^3 = \frac{1-p}{p^2}$$

i) Negative Binomial (for fixed r): Y = number of failures prior to the r^{th} success in a Bernoulli process

$$
\begin{aligned}
p(y) &= \frac{\Gamma(y+r)}{\Gamma(r)y!}(1-p)^y(p)^r \\
&= \exp\left[\log\left(\frac{\Gamma(y+r)}{\Gamma(r)y!}(1-p)^y(p)^r\right)\right] \\
&= \exp\left[\log\left(\frac{\Gamma(y+r)}{\Gamma(r)y!}\right) + y\log(1-p) + r\log(p)\right] \\
&= \exp\left[y\log(1-p) + r\log(p) + \log\left(\frac{\Gamma(y+r)}{\Gamma(r)y!}\right)\right]
\end{aligned}
$$

So,

$a(y) = y$

$b(p) = \log(1-p) =$ canonical link

$c(p) = r\log(p)$

$$d(y) = \log\left(\frac{\Gamma(y+r)}{\Gamma(r)y!}\right)$$

A negative binomial distribution can be used to model the number of random people you call who decline before 100 people agree to complete a survey.

$$\mu = -c'(p)/b'(p) = \frac{r(1-p)}{p}$$

$$\sigma^2 = (b''(p)c'(p) - c''(p)b'(p))/b'(p)^3 = \frac{r(1-p)}{p^2}$$

j) Pareto (for fixed k):

$$f(y) = \frac{\theta k^\theta}{y^{(\theta+1)}}$$
$$= \exp\left[\log\left(\frac{\theta k^\theta}{y^{(\theta+1)}}\right)\right]$$
$$= \exp\Big[\log(\theta k^\theta) - (\theta+1)\log(y)\Big] \quad = \exp\Big[-\theta\log(y) + \log(\theta k^\theta) - \log(y)\Big]$$

Therefore,

$a(y) = -\log(y)$

$b(\theta) = \theta =$ canonical link

$c(\theta) = \log(\theta k^\theta)$

$d(y) = -\log(y)$

A Pareto distribution could be used to model incomes where we expect a skewed distribution with a long right tail.

$$\mu = -c'(\theta)/b'(\theta) = \frac{k\theta}{\theta - 1}$$

$$\sigma^2 = (b''(\theta)c'(\theta) - c''(\theta)b'(\theta))/b'(\theta)^3 = \frac{k^2\theta}{(\theta-1)^2(\theta-2)} \text{ for } \theta > 2$$

2) Fill in the table with answers from the previous exercises.

6

Logistic Regression

```
# Packages required for Chapter 6
library(gridExtra)
library(mnormt)
library(lme4)
library(knitr)
library(pander)
library(tidyverse)
```

6.1 Exercises

6.1.1 Conceptual Exercises

1. List the explanatory and response variable(s) for each research question.

a) Response: a student will or will not binge drink. Explanatory: a measure of academic performance.

b) Response: accepted to medical school or rejected. Explanatory: MCAT scores and GPA.

c) Response: a mother marries baby's father or not. Explanatory: sex of the baby.

d) Response: a student graduates or not. Explanatory: participation in sports, type of sport, and gender.

e) Response: cancer diagnosis (yes or no). Explanatory: chemical exposure (yes or no, or amount of exposure).

2. The odds of having a nightly cough are 89% greater if the child is in a day care center than when the child is in home care, after controlling for environmental conditions and family characteristics. Also, the odds of a blocked or

TABLE 6.1: Birth Defect Rates

Situation	Defect	NoDefect	Total
IVF	230	53	283
Non-IVF	9354	4739	14093
Total	9584	4792	14376

runny nose without common cold are 55% greater if the child is in a day care center than when the child is in home care, after controlling for environmental conditions and family characteristics.

3. The odds a baby from a mother who used IVF has a birth defect are 2.20 times higher than for a baby from a mother who didn't use IVF. This does not seem consistent with the statement that, "IVF does not carry excessive risks".

```
## [1] 4.34
```

```
## [1] 1.974
```

```
## [1] 2.199
```

4. **Turbine wheels**

$\text{logL}(p_L, p_H) = Pr(Y_1 = 1, Y_2 = 2, Y_3 = 1, Y_4 = 0)$
$= \binom{3}{1} p_L^1 (1-p_L)^2 \binom{3}{2} p_L^2 (1-p_L)^1 \binom{3}{1} p_H^1 (1-p_H)^2 \binom{3}{0} p_H^0 (1-p_H)^3$
$= 27 p_L^3 (1-p_L)^3 p_H^1 (1-p_H)^5$

6.1.2 Guided Exercises

1. **Soccer goals on target.** [Roskes et al., 2011].

a)

- odds of a successful PK when behind: 18/2 = 9
- odds of a successful PK when not behind: 55/20 = 2.75
- odds of a successful PK when tied: 71/19 = 3.74
- odds ratio behind vs. tied = 9 / 3.74 = 2.41
- odds ratio tied vs. ahead = 3.74 / 2.75 = 1.36

TABLE 6.2: Soccer goalkeepers' Saves

Situation	Saves	Scores	Total
Behind	2	18	20
Ahead	20	55	75
Tied	19	71	90
Total	41	144	185

b)

```
table1 <- as_tibble(table1[-4,]) %>%
  mutate(Situation <- relevel(as_factor(Situation), ref=3),
         Behind = ifelse(Situation == "Behind", 1, 0),
         Ahead = ifelse(Situation == "Ahead", 1, 0))

model1 <- glm(cbind(Scores, Saves) ~ Behind + Ahead,
              family = binomial, data = table1)
summary(model1)
exp(coef(model1))
1/exp(coef(model1))
```

The exponentiated coefficients (or their inverse) match up perfectly with the odds ratios we calculated in (a):

- Intercept - Odds of a successful PK in a tied situation is 3.74; that is, the probability of a successful PK in a tied situation is $3.74/(1+3.74) = 0.79 = 71/90$.
- Behind - Odds of a successful PK when the goalkeeper's team is behind is 2.4 times higher than when the game is tied.
- Ahead - Odds of a successful PK when the goalkeeper's team is ahead are 26.4% (1 - .736) lower than when the game is tied. In other words, the odds of a successful PK are 35.9% higher (1 / .736 = 1.359) when the game is tied than when the goalkeeper's team is ahead.

2. **Medical school admissions.**

Note: a good argument could be made for removing row 54 since their WS is missing and, as a result, their MCAT score is artificially low. But we leave row 54 in for these analyses since results are minimally affected.

a) Every 1 point increase in GPA is associated with odds of getting into med school that are 232.8 ($e^{5.45}$) times higher. Every 1 point increase in MCAT

is associated with odds of getting into med school that are 1.28 ($e^{.246}$) times higher. It appears as though GPA has a higher impact on acceptance, compared to MCAT scores, but a 1.0 increase in GPA is much harder to achieve than a 1 point increase in an MCAT score. You could reframe the effect of GPA as every 0.1 point increase in GPA is associated with odds of getting into med school that are 1.725 ($e^{.545}$) times higher. When both GPA and MCAT are placed into a model together, so we get the effect of one after controlling for the other, it appears that GPA has a greater effect than MCAT (Z = 2.85 vs. Z = 1.59).

```
MedGPA <- read_csv("data/MedGPA.csv")

MedGPA <- MedGPA %>%
  mutate(MCATcuts =  cut(MCAT,
    breaks = quantile(MCAT, probs=seq(0,1,by=.25)),
    include.lowest = T))

ggplot(MedGPA, aes(x=GPA, y=MCAT)) +
  geom_point()  +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)

m1 <- glm(Acceptance ~ GPA, family = binomial, data = MedGPA)
summary(m1)

m2 <- glm(Acceptance ~ MCAT, family = binomial, data = MedGPA)
summary(m2)

m3 <- glm(Acceptance ~ GPA + MCAT, family = binomial,
          data = MedGPA)
summary(m3)
```

b) There is not statistically significant evidence (Z = 0.57, p = .565) that there is an association between number of applications and odds of getting into med school, after accounting for GPA and MCAT scores.

```
m4 <- glm(Acceptance ~ GPA + MCAT + Apps, family = binomial,
          data = MedGPA)
summary(m4)
```

c) Exploratory analyses show that the biological science and physical science scores are most closely related to acceptance. Placing all subscales into a single model shows that BS is most significant (Z = 3.08, p = .002) after controlling

for the other subscales, while WS and PS are both marginally significant (and WS is negative).

```
ggplot(data = MedGPA, aes(x = VR, fill = Accept)) +
  geom_density(position = 'fill', alpha = 0.5)
favstats(VR ~ Accept, data = MedGPA)

ggplot(data = MedGPA, aes(x = PS, fill = Accept)) +
  geom_density(position = 'fill', alpha = 0.5)
favstats(PS ~ Accept, data = MedGPA)

ggplot(data = MedGPA, aes(x = WS, fill = Accept)) +
  geom_density(position = 'fill', alpha = 0.5)
favstats(WS ~ Accept, data = MedGPA)

ggplot(data = MedGPA, aes(x = BS, fill = Accept)) +
  geom_density(position = 'fill', alpha = 0.5)
favstats(BS ~ Accept, data = MedGPA)

m5 <- glm(Acceptance ~ VR + PS + WS + BS, family = binomial,
          data = MedGPA)
summary(m5)
```

d) Even though there is some evidence of an MCAT-by-Sex interaction in exploratory plots, there is not statistically significant evidence that the effect of MCAT score (Z = 0.87, p = 0.38) or GPA (Z = -0.46, p = .645) differs for males and females (tests based on the interaction term, not controlling for other predictors).

```
MedGPA <- MedGPA %>%
  mutate(MCAT_cuts2 = cut_number(MCAT, 4))

emplogit3 <- MedGPA %>%
  group_by(MCAT_cuts2, Sex) %>%
  summarise(num_yes = sum(Acceptance),
            n = n(),
            prop_yes = num_yes / n,
            midpoint = median(MCAT)) %>%
  mutate(prop_yes = ifelse(prop_yes == 1 | prop_yes == 0,
                           (num_yes + 0.5) / (n + 1), prop_yes),
         odds = prop_yes / (1 - prop_yes),
         emplogit = log(odds))
```

```
emplogit3

ggplot(emplogit3, aes(x = midpoint, y = emplogit,
                      color = Sex)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlab("Midpoint of MCAT group") + ylab("Empirical logits")

MedGPA <- MedGPA %>%
  mutate(GPA_cuts2 = cut_number(GPA, 4))

emplogit4 <- MedGPA %>%
  group_by(GPA_cuts2, Sex) %>%
  summarise(num_yes = sum(Acceptance),
            n = n(),
            prop_yes = num_yes / n,
            midpoint = median(GPA)) %>%
  mutate(prop_yes = ifelse(prop_yes == 1 | prop_yes == 0,
                           (num_yes + 0.5) / (n + 1), prop_yes),
         odds = prop_yes / (1 - prop_yes),
         emplogit = log(odds))
emplogit4

ggplot(emplogit4, aes(x = midpoint, y = emplogit,
                      color = Sex)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlab("Midpoint of GPA group") + ylab("Empirical logits")

m6 <- glm(Acceptance ~ GPA + Sex + GPA:Sex, family = binomial,
          data = MedGPA)
summary(m6)

m7 <- glm(Acceptance ~ MCAT + Sex + MCAT:Sex, family = binomial,
          data = MedGPA)
summary(m7)
```

3. **Moths.** Article: [Bishop, 1972]; data source: [Ramsey and Schafer, 2002].

a) Here, an empirical logit is an estimated log-odds of removal for all moths of one morph at a single distance (log (# removed / # not removed)). Our logistic regression model assumes predictors are linearly related to log-odds.

```
moth <- read_csv("data/moth.csv")
moth <- mutate(moth,
               notremoved = PLACED - REMOVED,
               logit1 = log(REMOVED / notremoved),
               prop1 = REMOVED / PLACED,
               dark = ifelse(MORPH=="dark",1,0) )

# Plot empirical logits vs. distance separately by morph
ggplot(data = moth, aes(x = DISTANCE, y = logit1, color = MORPH,
                        shape = MORPH)) +
  geom_point() +
  geom_smooth(method="lm", alpha = 1/4)
```

b) The plot shows that a binomial model with interaction may fit well—linearity within morph is reasonable and slopes run in opposite directions for light and dark morph moths.

```
moth1 <- glm(prop1 ~ DISTANCE + dark, weights = PLACED,
             family = binomial, data = moth)
summary(moth1)
```

c) $e^{-1.137} = 0.32$: estimated odds of removal for light morph in Liverpool (probability of removal $= \frac{0.32}{1+.32} = .24$).

$e^{.0053(10)} = 1.054$: estimated odds of removal increases by 5.4% for every 10 km from Liverpool for a fixed morph.

$e^{\cdot}404 = 1.50$: estimated odds of removal is 50% higher for dark morphs compared to light morphs at a given distance

```
moth2 <- glm(prop1 ~ DISTANCE + dark + DISTANCE:dark,
             weights = PLACED, family = binomial, data = moth)
summary(moth2)
```

d) $e^{-0.72} = 0.49$: estimated odds of removal for light morph in Liverpool (probability of removal $= \frac{0.49}{1+.49} = .33$).

$1/e^{-.009287(10)} = 1/.9113 = 1.097$: estimated odds of removal of a light morph increases by 9.7% for every 10 km closer to Liverpool. Or, estimated odds of removal for a light morph is 1-.9113 = .0887 = 8.9% lower for each 10 km from Liverpool.

$1/e^{0.41} = 1/0.663 = 1.51$: estimated odds of removal is 51% higher for light morphs compared to dark morphs in Liverpool (distance = 0). Or, estimated odds of removal for dark morphs is 1-.663 = 33.7% lower in Liverpool.

$e^{(-00929+0.02779)10} = 1.203$: estimated odds of removal of a dark morph increases by 20.3% for every 10 km from Liverpool, compared to a decrease of 8.9% for light morphs.

```
anova(moth1, moth2)
```

e) We have statistically significant evidence (Z = 3.44, p = .00059 from Wald test; D = 11.931, p = .00055 from drop-in-deviance / Likelihood ratio test) that the relative odds of removal of light and dark morphs changes with distance to Liverpool.

```
gof = 1-pchisq(moth2$deviance, moth2$df.residual)
gof
```

f) No evidence of lack of fit (Residual deviance = 13.230, p = .21). Either the model is adequate (model well-specified, binomial variance good, no outliers) or test is underpowered.

```
moth_quasi <- glm(prop1 ~ DISTANCE + dark + DISTANCE:dark,
  weights = PLACED, family = quasibinomial, data = moth)
summary(moth_quasi)
```

g) Yes, there is evidence of overdispersion, which we can see in the model above having a dispersion coefficient larger than one. Moth removal is not likely to be independent or have constant probability, which could lead to overdispersion (e.g. if one moth at a location is removed, chances increase that others will also be removed). Note that all coefficients, and thus all interpretations, remain the same from (d); however, standard errors grow and thus p-values also grow.

```
confint((moth2))
confint(moth_quasi)
```

h) Quasi-Poisson produces wider intervals.

```
mtemp1 = rep(moth$dark[1],moth$REMOVED[1])
dtemp1 = rep(moth$DISTANCE[1],moth$REMOVED[1])
rtemp1 = rep(1,moth$REMOVED[1])
mtemp1 = c(mtemp1,rep(moth$dark[1],
                      moth$PLACED[1]-moth$REMOVED[1]))
dtemp1 = c(dtemp1,rep(moth$DISTANCE[1],
                      moth$PLACED[1]-moth$REMOVED[1]))
rtemp1 = c(rtemp1,rep(0,moth$PLACED[1]-moth$REMOVED[1]))
for(i in 2:14)  {
  mtemp1 = c(mtemp1,rep(moth$dark[i],moth$REMOVED[i]))
  dtemp1 = c(dtemp1,rep(moth$DISTANCE[i],moth$REMOVED[i]))
  rtemp1 = c(rtemp1,rep(1,moth$REMOVED[i]))
  mtemp1 = c(mtemp1,rep(moth$dark[i],
                        moth$PLACED[i]-moth$REMOVED[i]))
  dtemp1 = c(dtemp1,rep(moth$DISTANCE[i],
                        moth$PLACED[i]-moth$REMOVED[i]))
  rtemp1 = c(rtemp1,rep(0,moth$PLACED[i]-moth$REMOVED[i]))  }
newdata = data.frame(removed=rtemp1,dark=mtemp1,dist=dtemp1)
newdata[1:25,]
cdplot(as.factor(rtemp1)~dtemp1)
```

```
moth_log <- glm(removed ~ dark + dist + dark:dist,
                family = binomial, data = newdata)
summary(moth_log)
summary(moth2)
```

i) Relatively the same: coefficients, standard errors, z-values, and p-values are all the same. However, the logistic model has larger null (1207.9) and residual deviances (1185.7), and has more degrees of freedom. Advantages of binomial regression: we can estimate overdispersion and adjust analyses, we can test GOF, and convenience (only 14 observations).

4. **Birdkeeping and lung cancer: a retrospective observational study.** Article: [Holst et al., 1988]; data source: [Ramsey and Schafer, 2002].

```
birds <- read_csv("data/birdkeeping.csv") %>%
  mutate(sex = ifelse(female == 1, "Female", "Male"),
         socioecon_status = ifelse(highstatus == 1,
                                   "High", "Low"),
         keep_bird = ifelse(bird == 1, "Keep Bird", "No Bird"),
```

```
        lung_cancer = ifelse(cancer == 1, "Cancer",
                             "No Cancer")) %>%
  mutate(years_factor = cut(yrsmoke,
                            breaks = c(-Inf, 0, 25, Inf),
           labels = c("No smoking", "1-25 years",
                      "Over 25 years")))
```

```
# age
ggplot(data = birds, aes(y = age, x = lung_cancer)) +
  geom_boxplot() +
  coord_flip()

ggplot(data = birds, aes(x = age, fill = lung_cancer)) +
  geom_density(position = 'fill', alpha = 0.5)

favstats(age ~ lung_cancer, data = birds)

# yrsmoke
ggplot(data = birds, aes(y = yrsmoke, x = lung_cancer)) +
  geom_boxplot() +
  coord_flip()

ggplot(data = birds, aes(x = yrsmoke, fill = lung_cancer)) +
  geom_density(position = 'fill', alpha = 0.5)

favstats(yrsmoke ~ lung_cancer, data = birds)

# cigsday
ggplot(data = birds, aes(y = cigsday, x = lung_cancer)) +
  geom_boxplot() +
  coord_flip()

ggplot(data = birds, aes(x = cigsday, fill = lung_cancer)) +
  geom_density(position = 'fill', alpha = 0.5, adjust = 2)

favstats(cigsday ~ lung_cancer, data = birds)

# female
two_by_two <- table(birds$sex, birds$lung_cancer)
addmargins(two_by_two)
table_of_props <- prop.table(two_by_two, margin = 1)
round(table_of_props, 3)
```

```
ggplot(data = birds, aes(x = sex, fill = lung_cancer)) +
  geom_bar(position = "fill") +
  labs(y = "Percentage")

# highstatus
two_by_two <- table(birds$socioecon_status, birds$lung_cancer)
addmargins(two_by_two)
table_of_props <- prop.table(two_by_two, margin = 1)
round(table_of_props, 3)

ggplot(data = birds, aes(x = socioecon_status,
                         fill = lung_cancer)) +
  geom_bar(position = "fill") +
  labs(y = "Percentage")

# bird
two_by_two <- table(birds$keep_bird, birds$lung_cancer)
addmargins(two_by_two)
table_of_props <- prop.table(two_by_two, margin = 1)
round(table_of_props, 3)

ggplot(data = birds, aes(x = keep_bird, fill = lung_cancer)) +
  geom_bar(position = "fill") +
  labs(y = "Percentage")
```

a) Age is unrelated to lung cancer; average age is nearly identical (56.9 vs 57.0) for patients with and without cancer. As number of years smoked increases, the proportion with cancer increases; those with cancer have smoked an average of 33.6 years, while those without cancer only averaged 25.0 years. Cigarettes per day seems to be related to cancer less strongly than years smoked; mean cigarettes per day is lower for those without cancer (18.8 vs. 14.2), although the largest effect seems to be with the lowest levels of cigarettes.

Sex is unrelated to lung cancer; both males and females had 1/3 of patients with cancer. There is a moderate relationship between socioeconomic status and lung cancer, with high status patients less likely to have cancer (26.7% vs. 36.3%). Birdkeepers were much more likey to have cancer than non-birdkeepers (49.3% vs. 20.0%).

b) Researchers designed their study this way. Each case (cancer patient) was matched with 2 controls that had the same sex and similar age. So the age and sex distributions in both cases and controls should look exactly the same. In terms of modeling, we should not have to include terms for age and sex, since the study design has already controlled for these factors (although some

statisticians recommend including them simply because they are part of the study design). We may still want to check for interactions with age and sex, though.

[Note: because we have a case-control study, we do not have a random sample from any population of interest; rather, we controlled the number of lung cancer and non-lung cancer subjects in our study. As a result, we cannot trust the interpretations of intercepts or predicted values or even coefficients of the variables we matched on. We can, however, trust the coefficient interpretations for other factors, along with interactions of other factors with matching variables.]

```
two_by_two <- table(birds$keep_bird, birds$lung_cancer)
addmargins(two_by_two)

odds_bird <- 33 / 34
odds_nobird <- 16 / 64
OR <- odds_bird / odds_nobird
OR

prob_bird <- 33 / 67
prob_nobird <- 16 / 80
RR <- prob_bird / prob_nobird
RR
```

c) OR = (33/34) / (16/64) = 3.88. The odds that a birdkeeper gets lung cancer are 3.88 times higher than the odds a non-birdkeeper gets lung cancer. Note that this odds ratio does not control for any other factors.

RR = (33/67) / (16/80) = 2.46. The probability that a birdkeeper gets lung cancer is 2.46 times higher than the probability a non-birdkeeper gets lung cancer.

```
birds <- birds %>%
  mutate(yrsmoke_cuts = cut_number(yrsmoke, 8))

emplogit2 <- birds %>%
  group_by(yrsmoke_cuts) %>%
  summarise(num_with = sum(cancer),
            n = n(),
            prop_with = num_with / n,
            midpoint = median(yrsmoke)) %>%
  mutate(prop_with = ifelse(prop_with == 1 | prop_with == 0,
```

```
                      (num_with + 0.5) / (n + 1), prop_with),
         odds = prop_with / (1 - prop_with),
         emplogit = log(odds))
emplogit2

ggplot(emplogit2, aes(x = midpoint, y = emplogit)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("Midpoint of years smoked group") +
  ylab("Empirical logits")

library(Stat2Data)
emplogitplot1(cancer ~ yrsmoke, data = births, ngroups = 8,
              out = TRUE)
```

d) Yes, it seems as if the linearity assumption in logistic regression – that log odds are linearly related to our predictor – holds for years smoked.

```
int_summary <- favstats(yrsmoke ~ lung_cancer + keep_bird,
                        data = births)
int_summary

int_plot <- births %>% group_by(lung_cancer, keep_bird) %>%
  summarise(mean_years = mean(yrsmoke))

ggplot(data = int_plot, aes(y = mean_years, x = lung_cancer,
    color = keep_bird, linetype = keep_bird)) +
  geom_point() +
  geom_line(aes(group = keep_bird))


emplogit3 <- births %>%
  group_by(yrsmoke_cuts, keep_bird) %>%
  summarise(num_with = sum(cancer),
            n = n(),
            prop_with = num_with / n,
            midpoint = median(yrsmoke)) %>%
  mutate(prop_with = ifelse(prop_with == 1 | prop_with == 0,
                      (num_with + 0.5) / (n + 1), prop_with),
         odds = prop_with / (1 - prop_with),
         emplogit = log(odds))
emplogit3
```

```
ggplot(emplogit3, aes(x = midpoint, y = emplogit,
                      color = keep_bird)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlab("Midpoint of years smoked group") +
  ylab("Empirical logits")

# This doesn't seem to work
# emplogitplot2(cancer ~ yrsmoke + keep_bird,
#   data = birds, out = TRUE)
```

e) There is little evidence of an interaction between years smoked and bird-keeping from either the interaction plot (focusing on mean years smoked) or the empirical logit plot (focusing on log odds of developing cancer). The effect of years smoked (slope) on log odds of developing cancer is very similar for birdkeepers and non-birdkeepers.

```
model1 <- glm(cancer ~ age + yrsmoke + cigsday + female +
  highstatus + bird, family = binomial, data = birds)

model2 <- glm(cancer ~ yrsmoke + cigsday + highstatus + bird,
              family = binomial, data = birds)

model4 <- glm(cancer ~ yrsmoke + bird,
              family = binomial, data = birds)

birds <- birds %>%
  mutate(yrsmoke_sq = yrsmoke^2)

model5 <- glm(cancer ~ yrsmoke + bird + yrsmoke_sq + yrsmoke:bird,
              family = binomial, data = birds)

model6 <- glm(cancer ~ yrsmoke + bird + yrsmoke:bird,
              family = binomial, data = birds)
```

```
summary(model1)
summary(model2)
anova(model2, model1, test = "Chisq")
```

f) Based on a nested G / drop in deviance / likelihood ratio test, we do not have

significant evidence (G = 2.5257 on 2 df, p = .2828) that adding age and sex to model1 improves our ability to explain variability in subjects' probability of a lung cancer diagnosis. After accounting for years smoked, cigarettes per day, socioeconomic status, and birdkeeping, there is no significant value in accounting for age and sex as well.

```
summary(model4)
summary(model5)
anova(model4, model5, test = "Chisq")
```

g) We do not have significant evidence (G = 1.0568 on 2 df, p=.5895) that a complete second order model featuring years smoked and birdkeeping improves upon the linear model with those two predictors. Note that although we usually add 3 terms to produce the complete second order model, here bird-squared is the same as bird, since it's an indicator variable (so that 0-squared = 0 and 1-squared = 1).

```
summary(model6)
exp(model6$coefficients)
exp(model6$coefficients[1]) / (1 + exp(model6$coefficients)[1])
```

h) Note that we have separate equations for birdkeepers and non-birdkeepers:

- bird=0: log-odds = -3.00 + .053 x yrsmoke
- bird=1: log-odds = (-3.00 + 1.18) + (.053 + .009) x yrsmoke = -1.82 + .061 x yrsmoke

Coefficient interpretations can then be stated as:

- $\hat{\beta}_0 = -3.00$. The predicted odds of having lung cancer for a non-birdkeeper who has never smoked is .050 ($e^{-3.00}$). In other words, the predicted probability that a non-birdkeeper who has never smoked has lung cancer is .048 ($\frac{.050}{1+.050}$). [Note: while algorithmically true, because we have a case-control study we cannot put any faith in the interpretations of intercept terms.]
- $\hat{\beta}_1 = .053$. Each additional year smoked increases the odds of having lung cancer by 5.4% ($e^{.053} = 1.054$) for non-birdkeepers.
- $\hat{\beta}_2 = 1.18$. The predicted odds of having lung cancer for subjects who have never smoked is 3.25 ($e^{1.18}$) times higher for birdkeepers than non-birdkeepers.

- $\hat{\beta}_3 = .009$. The effect of an additional year smoked on the odds of having lung cancer is 0.9% greater for birdkeepers than non-birdkeepers. For birdkeepers, the odds increase by 6.4% ($e^{.062} = 1.064$) for each additional year smoked, compared to just 5.4% in non-birdkeepers.

```
birds <- birds %>%
  mutate(years_center = yrsmoke - mean(yrsmoke))

model6a <- glm(cancer ~ years_center + bird + years_center:bird,
               family = binomial, data = birds)
summary(model6a)
```

i) These elements would change: $\hat{\beta}_0$ and $\hat{\beta}_2$ and their associated p-values. Everything else will remain the same.

```
exp(confint(model4))
```

j) We are 95% confident that each additional year smoked increases the odds of having lung cancer between 2.8% and 9.8%, controlling for birdkeeping status (and some would add age and sex, since we controlled them in the design process).

We are 95% confident that birdkeepers have between 2.05 and 9.75 times greater odds of having lung cancer than non-birdkeepers, controlling for smoking history (and some would add age and sex, since we controlled them in the design process).

```
summary(model4)
exp(model4$coefficients)
```

k) The adjusted odds ratio (4.37) is slightly higher than the unadjusted odds ratio (3.88). Based on the Wald test, birdkeeping is associated with a significant increase in the odds of developing lung cancer (Z = 3.727, p = .000194) after adjusting for smoking history (and age and sex by design).

```
birds <- birds %>%
  mutate(years_factor = cut(yrsmoke, breaks = c(-Inf, 0, 25, Inf),
      labels = c("No smoking", "1-25 years", "Over 25 years")))
```

```
model4a <- glm(cancer ~ years_factor + bird,
               family = binomial, data = birds)
summary(model4a)
exp(model4a$coefficients)
```

l) After accounting for birdkeeping status, patients who have been smoking over 25 years have 15.6 times greater odds of developing lung cancer than patients who have never smoked.

I would prefer model4 – it's simpler (fewer terms), you don't lose information from categorizing a numeric variable, and it's got a smaller AIC.

m) We should be careful when generalizing outside of The Netherlands – this study was conducted at a time and place where birdkeeping was fairly common and smoking was widespread. Although the authors offer a convincing scientific argument why keeping birds could increase the risk of lung cancer, these risks could be different for different birds, different caging systems, and a different population. Because we have an observational study, we cannot be sure that we've controlled for all important covariates, so the observed association between birdkeeping and lung cancer could still be explained by other factors, such as air quality, genetic factors, etc.

It turns out that with 1:2 pairing based on age and sex, statisticians will often use conditional logistic regression, which focuses on factor effects within groups and accounts for (but doesn't specifically estimate) differences in baseline log-odds of having lung cancer for different age and sex combinations. Conditional logistic regression estimates and standard errors are both often slightly larger than those we found using unconditional logistic regression, but primary conclusions usually remain the same.

n) Some points you might have made:

- Researchers also collected data on beta carotene, vitamin C, and alcohol consumption.
- Their adjusted odds ratio was 6.7.
- Their final model included birdkeeping, smoking, and vitamin C.
- They concluded that the 14 patients who were eligible but did not participate did not bias the results (robustness check).
- People who keep birds are inhaling and expectorating excess allergens and dust particles, which causes dysfunction of lung macrophages and local deficiency in humoral and cellular immunity.
- Netherlands has more than 1 bird for every 2 inhabitants.

5. **2016 Election.** [Blakeman et al., 2018]

```
electiondata <- read_csv("data/electiondata.csv")
electiondata <- electiondata %>%
  mutate(Vote = ifelse(Vote=="Other", "Other", "Trump"),
         Vote01 = ifelse(Vote=="Trump", 1, 0),
         pctforeign = 100*propforeign,
         evangelical = ifelse(pew_bornagain == 2, 1, 0))
electiondata %>%
  mutate(EconWorseYN = ifelse(EconWorse == 1, "Yes", "No")) %>%
  ggplot(aes(x = EconWorseYN, fill = Vote)) +
    geom_bar(position = "fill")
```

a) Voters who believed the economy had gotten worse were much more likely to vote for Trump.

```
electiondata %>%
  mutate(EconWorseYN = ifelse(EconWorse == 1, "Yes", "No"),
         RepYN = ifelse(republican == 1, "Republican",
                        "non-Republican")) %>%
  ggplot(aes(x = EconWorseYN, fill = Vote)) +
    geom_bar(position = "fill") +
    facet_wrap(~ RepYN, nrow = 1)
```

b) This relationship holds for both republicans and non-republicans, although the gap appears larger in non-republicans.

```
electiondata %>%
  group_by(inputstate) %>%
  summarise(n = n_distinct(medinc)) %>%
  arrange(desc(n))

electiondata %>%
  group_by(inputstate) %>%
  summarise(trumpvotes = sum(Vote01==1),
            othervotes = sum(Vote01==0),
            logodds_trump = log(trumpvotes / othervotes),
            state_median_faminc = median(medinc)) %>%
  ggplot(aes(x = state_median_faminc, y = logodds_trump)) +
    geom_point() +
    geom_smooth()
```

c) There is a strong correlation between lower income states and higher log odds of voting for trump.

```
model1a <- glm(Vote01 ~ zfaminc + zmedinc + EconWorse +
    EducStatus + republican + EducStatus:republican +
    EconWorse:zfaminc + EconWorse:republican,
    family = binomial, data = electiondata)
summary(model1a)
exp(coef(model1a))
```

d) For each point increase in state median income z-score, the the odds of voting for Trump decreased 11% (1-.89), when controlling for familial income, beliefs about the economy, if the voter has a bachelor's degree, political party, and some associated interactions.

e) For voters who did not have at least a bachelor's degree and did not believe that the economy got worse in the last four years, the odds they voted for Trump were 29.6 times greater if they were a republican instead of a non-republican, when controlling for familial and state income.

f) Non-republicans with average family income who believed the economy had gotten worse had 10.7 ($e^{2.37}$) times greater odds of voting for Trump than those who didn't believe the economy was worse; in contrast, republicans with average family income who believed the economy had gotten worse had 4.7 ($e^{2.37-0.83}$) times greater odds of voting for Trump. This is true after controlling for state income and education status. Thus, there seems to be evidence to support Theory 1, with a stronger effect of economic beliefs among non-republicans.

```
electiondata %>%
  mutate(ImmPolicy = ifelse(Noimmigrants == 1, "Anti-immigrant",
                            "Not Anti-immigrant")) %>%
  ggplot(aes(x = ImmPolicy, fill = Vote)) +
    geom_bar(position = "fill")

electiondata %>%
  mutate(ImmPolicy = ifelse(Noimmigrants == 1, "Anti-immigrant",
                            "Not Anti-immigrant"),
         RepYN = ifelse(republican == 1, "Republican",
                        "non-Republican")) %>%
  ggplot(aes(x = ImmPolicy, fill = Vote)) +
    geom_bar(position = "fill") +
    facet_wrap(~ RepYN, nrow = 1)
```

```
model2a <- glm(Vote01 ~ Noimmigrants + pctforeign + evangelical +
  republican + Noimmigrants:republican, family = binomial,
  data = electiondata)
summary(model2a)
exp(coef(model2a))
```

g) Exploratory plots show that supporters of anti-immigrant policies were more likely to vote for Trump, with the gap fairly similar among both republicans and non-republicans.

After controlling for proportion foreign-born in a state and if a voter is evangelical, non-republicans who support anti-immigration policies had 10.2 ($e^{2.318}$) times greater odds of voting for Trump than those who didn't support anti-immigrant policies, compared to 5.4 ($e^{2.318-0.629}$) times greater odds among republicans. Again, we see empirical support for Theory 2, with the effect being greater among non-republicans.

```
library(lme4)

# try as multilevel model (not much diff)
model1b <- glmer(Vote01 ~ zfaminc + zmedinc + EconWorse +
  EducStatus + republican + EducStatus:republican +
  EconWorse:zfaminc + EconWorse:republican + (1|inputstate),
  family = binomial, data = electiondata)
summary(model1b)

# try with two levels (not much diff)
model2b <- glmer(Vote01 ~ Noimmigrants + pctforeign +
  evangelical + republican + Noimmigrants:republican +
  (1|inputstate), family = binomial, data = electiondata)
summary(model2b)
```

h) There is some concern that there is correlation within states – that voters from the same state are more correlated than voters from different states. This is multilevel data – we have some variables measured at the individual level and some at the state level. In the R code, we show how this could be modeled using concepts from chapters 8 through 11, although results are similar in this case.

6.1.3 Open-Ended Exercises

1. **2008 Presidential voting in Minnesota counties.**

```
mn08 <- read_csv("data/mn08.csv") %>%
  mutate(pct_Obama1 = pct_Obama / 100,
         deltaAge = medAge2007-medAge2000,
         medHHinc1000 = medHHinc / 1000,
         result = ifelse(Obama > McCain, 1, 0))
summary(mn08)
```

```
# Note that Gini_Index, pct_native, and medAge2007 all have
#   40 NAs among n=87 counties

# Decent model with only 47 counties
mnbin47 <- glm(pct_Obama1 ~ pct_rural + pct_poverty +
  Gini_Index + deltaAge + pct_rural:pct_poverty,
  weights = (Obama+McCain), family = quasibinomial, data = mn08)
summary(mnbin47)
exp(coef(mnbin47))

# Decent model with all 87 counties
mnbin87 <- glm(pct_Obama1 ~ medHHinc1000 + unemp_rate +
  pct_poverty, weights = (Obama+McCain),
  family = quasibinomial, data = mn08)
summary(mnbin87)
exp(coef(mnbin87))
```

From a binomial model with only 47 of the 87 counties (because 40 counties have missing values for Gini_Index, pct_native, and medAge2007) with percent Obama vote as the response:

- pct_rural - For a county with zero percent poverty, a one point increase in percent rural increases odds of voting for Obama by 0.7%, controlling for Gini index and change in median age.
- pct_poverty - For a county with zero percent rural, a one point increase in percent poverty increases odds of voting for Obama by 4.9%, controlling for Gini index and change in median age.
- Gini_Index - For a 0.1 increase in Gini index, odds of voting for Obama increase by 47.7% ($e^{3.902/10}$), controlling for percent rural and poverty, and change in median age.
- deltaAge - For each one point increase in the difference in median age from 2007 compared to 2000 (that is as median age in the county has increased in the last 7 years), the odds of voting for Obama increase 6.9%, controlling for percent rural and poverty, and Gini Index.

- pct_rural:pct_poverty - Together the estimated effects of `pct_rural` and `pct_poverty` overstate the odds of voting for Obama and must be adjusted by the interaction term. One way we can describe this term is by looking at the effect of `pct_poverty` among low and high `pct_rural` counties. For counties at the 25th percentile of pct_rural (46.0%), every 1 percent increase in percent poverty is associated with a 0.7% increase ($e^{.0480-.0009*46}$) in the odds of voting for Obama, after controlling for Gini index and change in median age. However, for counties at the 75th percentile of pct_rural (83.0%), every 1 percent increase in percent poverty is associated with a 2.6% decrease ($e^{.0480-.0009*83}$) in the odds of voting for Obama, after controlling for Gini index and change in median age. So the effect of poverty is different in mostly rural counties compared to less rural counties.

From a binomial model with all 87 counties with percent Obama vote as the response:

- pct_poverty - A one point increase in percent poverty increases odds of voting for Obama by 10.5%, controlling for median household income and unemployment rate.
- unemp_rate - A one point increase in unemployment rate decreases odds of voting for Obama by 7.9%, controlling for median household income and percent poverty.
- medHHinc1000 - A $1000 increase in median household income increases odds of voting for Obama by 1.7%, controlling for percent poverty and unemployment rate.

```
mnlog87 <- glm(result ~ unemp_rate + pct_poverty,
               family = binomial, data = mn08)
summary(mnlog87)
exp(coef(mnlog87))
```

When we fit a logistic model with whether or not Obama carried the county as the response, we lose power and very little is significant, even when restricting ourselves to predictors with no missing values. One potential final model is similar to the binomial model for all 87 counties, with a negative effect of unemployment rate and positive effect of percent poverty:

- pct_poverty - A one point increase in percent poverty increases odds that Obama wins a county by 20.2%, controlling for unemployment rate.
- unemp_rate - A one point increase in unemployment rate decreases odds that Obama wins a county by 26.0%, controlling for percent poverty.

2. **Crime on campus.**

```
c.data <- read_csv("data/c_data2.csv") %>%
  mutate(region2 = fct_recode(region, "S" = "SW", "S"="SE"),
         viol_prop = num_viol / total_crime,
         enroll1000 = Enrollment/1000)
summary(c.data)
ggplot(c.data, aes(x = region2, y = viol_prop, fill = type)) +
  geom_boxplot()
```

```
model1 <- glm(viol_prop ~ region2, family = binomial,
              weights=total_crime, data = c.data)
summary(model1)

model1a <- glm(viol_prop ~ region2, family = quasibinomial,
               weights=total_crime, data = c.data)
summary(model1a)
```

It does not appear as though there is a significant difference in violent crimes based on region, once overdispersion has been accounted for.

```
model2 <- glm(viol_prop ~ type, family = binomial,
              weights=total_crime, data = c.data)
summary(model2)

model2a <- glm(viol_prop ~ type, family = quasibinomial,
               weights=total_crime, data = c.data)
summary(model2a)
```

Type of institution (college or university) also doesn't seem to be associated with the rate of violent crime after accounting for overdispersion.

```
model3 <- glm(viol_prop ~ enroll1000, family = binomial,
              weights=total_crime, data = c.data)
summary(model3)

model3a <- glm(viol_prop ~ enroll1000, family = quasibinomial,
               weights=total_crime, data = c.data)
```

```
summary(model3a)
exp(coef(model3a))

model3b <- glm(viol_prop ~ enroll1000 + type + enroll1000:type,
               family = quasibinomial,
               weights=total_crime, data = c.data)
summary(model3b)
exp(coef(model3b))
```

It appears that larger institutions have a lower proportion of violent crimes. For each increase in school enrollment of 1000 students, odds of a violent crime decrease 2.3%. There is also marginally significant evidence that the negative effect of enrollment is less pronounced in universities rather than colleges.

3. **NBA data.** [Kaggle, 2018b].

```
nba <- read_csv("data/NBA1718team.csv") %>%
  mutate(FG_pct = 100* FG / FGA)
summary(nba)

# a few representative EDA plots - should also look at summary
#   statistics and empirical logit plots
ggplot(nba, aes(x = FG_pct, y = win_pct)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("FG%") + ylab("Win Percentage")

ggplot(nba, aes(x = FT_pct, y = win_pct)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("FT%") + ylab("Win Percentage")

ggplot(nba, aes(x = avg_3P_pct, y = win_pct)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("3 point pct") + ylab("Win Percentage")

ggplot(nba, aes(x = attempts_3P, y = win_pct)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("3P Attemps") + ylab("Win Percentage")

ggplot(nba, aes(x = AST, y = win_pct)) +
```

```
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("Average assists") + ylab("Win Percentage")
```

```
# removed FG, PTS, REB, FGA
#   since can be calculated with others

# Binomial model with all predictors - everything significant
model_bin <- glm(cbind(Win, Loss) ~ FT_pct + TOV +
  attempts_2P + attempts_3P + avg_3P_pct + OREB + DREB +
  AST + STL + BLK + PF + FG_pct,
  weights = (Win+Loss), family = binomial, data = nba)
summary(model_bin)

# Quasi-binomial model after backward elimination
model_qbin <- glm(cbind(Win, Loss) ~ TOV + attempts_3P +
  avg_3P_pct + OREB + DREB + AST + STL + FG_pct,
  weights = (Win+Loss), family = quasibinomial, data = nba)
summary(model_qbin)
exp(coef(model_qbin))

# Linear model with all predictors
model_lin <- lm(win_pct ~ FT_pct + TOV + attempts_2P +
  attempts_3P + avg_3P_pct + OREB + DREB + AST + STL +
  BLK + PF + FG_pct, data = nba)
summary(model_lin)

# Backward elimination - linear model
model_lin2 <- lm(win_pct ~ TOV + attempts_3P + OREB + DREB +
  STL + FG_pct, data = nba)
summary(model_lin2)
```

A quasi-binomial model shows that the odds of winning increase with increasing 3-point attempts, 3-point percentage, offensive rebounds, defensive rebounds, steals, and overall FG percentage, and decreasing turnovers and assists. For example, holding all else constant, the odds of winning increase by 5.0% for each additional 3 point attempt on average.

Given the distribution of winning percentage and constant number of games per team, we could also reasonably fit a linear regression model to predict changes in winning percentage. A model with turnovers, 3-point attempts,

offensive rebounds, defensive rebounds, steals, and overall FG percentage explains 88.9% of team-to-team variability in winning percentage.

We should also examine residual plots to identify unusual teams.

4. **Trashball.**

Here is a sample assignment based on the data set `TrashballF19.csv` containing the data collected from a class during Fall Semester 2019. After loading in the data, create the following additional variables:

```
trashball <- read_csv("data/TrashballF19.csv")
trashball <- trashball %>%
  mutate(missed = shots - made,
         logit1 = log((made+0.5) / (missed+0.5)),
         prop1 = made / shots,
         senior = ifelse(year == "2020", 1, 0),
         class = as.factor(year),
         location = as.factor(distance),
         playhoops = as.factor(basketball),
         under = ifelse(underhand == "Y", 1, 0),
         tennisfirst = ifelse(order=="tennis-pingpong", 1, 0),
         tennisball = ifelse(ball=="tennis", 1, 0))
```

1. Why did I add 0.5 to the number of made and missed baskets when calculating the log-odds?

The 0.5 is arbitrary, but it prevents undefined logits when either the numerator or denominator is 0.

2. Create scatterplots of `logit1` vs. continuous predictors (`distance`, `height`, `HSsports`, `awake`, `sleep`, and `shots`) and boxplots of `logit1` vs. categorical variables (`senior`, `location`, `glasses`, `playhoops`, `order`, `ball`, `underhand`). Summarize important trends in one or two sentences.

```
# EDA
plt1 <- ggplot(trashball, aes(x = distance, y = logit1)) +
  geom_point() + geom_smooth(method = "lm")
plt2 <- ggplot(trashball, aes(x = height, y = logit1)) +
  geom_point() + geom_smooth(method = "lm")
plt3 <- ggplot(trashball, aes(x = HSsports, y = logit1)) +
  geom_point() + geom_smooth(method = "lm")
```

```
plt4 <- ggplot(trashball, aes(x = awake, y = logit1)) +
  geom_point() + geom_smooth(method = "lm")
plt5 <- ggplot(trashball, aes(x = sleep, y = logit1)) +
  geom_point() + geom_smooth(method = "lm")
plt6 <- ggplot(trashball, aes(x = shots, y = logit1)) +
  geom_point() + geom_smooth(method = "lm")
grid.arrange(plt1, plt2, plt3, plt4, plt5, plt6, nrow = 2)

plt7 <- ggplot(trashball, aes(y = logit1, x = playhoops)) +
  geom_boxplot()
plt8 <- ggplot(trashball, aes(y = logit1, x = order)) +
  geom_boxplot()
plt9 <- ggplot(trashball, aes(y = logit1, x = ball)) +
  geom_boxplot()
plt10 <- ggplot(trashball, aes(y = logit1, x = class)) +
  geom_boxplot()
plt11 <- ggplot(trashball, aes(y = logit1, x = glasses)) +
  geom_boxplot()
plt12 <- ggplot(trashball, aes(y = logit1, x = underhand)) +
  geom_boxplot()
plt13 <- ggplot(trashball, aes(y = logit1, x = location)) +
  geom_boxplot()
grid.arrange(plt7, plt8, plt9, plt10, plt11, plt12, plt13,
             nrow = 3)
```

From scatterplots of continuous predictors, there is a strong tendency for log odds of making a shot to increase as distance decreases, height increases, and sleep time increases. From boxplots, there seems to be higher log odds of making a shot for those who shot tennis balls first, shot overhand, and those without glasses or contacts; the effect of distance also appears mostly linear.

3. Create a correlation matrix featuring `logit1`, `distance`, `shots`, `height`, `basketball`, `HSsports`, `awake`, `sleep`, `senior`, `under`, `tennisfirst`, and `tennisball`. Summarize important trends in one or two sentences.

```
cor(as.matrix(dplyr::select(trashball, logit1, distance,
  shots, height, basketball, HSsports, awake, sleep, senior,
  under, tennisfirst, tennisball)))
#pairs(as.matrix(dplyr::select(trashball, logit1, distance,
#  shots, height, basketball, HSsports, awake, sleep, senior,
#  under, tennisfirst, tennisball)))
```

Distance, order, sleep, height, underhand, and hours awake are confirmed as the strongest predictors of making a shot. There is some risk of correlated predictors, such as those with less sleep being awake longer. Interestingly, taller individuals were more likely to shoot at longer distances and less likely to shoot underhand, and those using tennis balls first were more likely to have more sleep.

4. Create a graph with empirical logits vs. `distance` plotted separately by `ball`. What might you conclude from this plot?

```
trashball %>%
  ggplot(aes(x = distance, y = logit1, colour = ball)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    labs(x = "Distance from basket (ft)",
         y = "Log odds of baskets made")
```

In general, the odds of making a basket with a ping pong ball decreases as distance increases, but it does not change with distance for tennis balls.

5. Create two binomial regression models, one with `distance` as the sole predictor, and one with `location` as the sole predictor. Which model performs better? What are the pros and cons to using `distance` rather than `location` to model the probability of making a basket?

```
breg1a <- glm(prop1 ~ distance, weights = shots,
              family = binomial, data = trashball)
summary(breg1a)
exp(coef(breg1a))

breg1b <- glm(prop1 ~ location, weights = shots,
              family = binomial, data = trashball)
summary(breg1b)
exp(coef(breg1b))
```

In the first model, we assume that log odds of making a basket is linearly related to distance, while in the second model we allow the log odds to decrease in a non-linear fashion (e.g., a smaller drop from 7 to 10 feet vs. 10 to 13 feet). By AIC, it appears the first model performs slightly better, so

assuming a linear effect of distance on the log odds of making a shot seems reasonable. With location we're able to relax the linearity assumption (at the cost of 1 additional degree of freedom), while the model with distance is more generalizable to distances not studied.

6. Fit a binomial regression model with `distance`, `ball`, and their interaction.

```
breg2 <- glm(prop1 ~ distance + ball + distance:ball,
  weights = shots, family = binomial, data = trashball)
summary(breg2)
exp(coef(breg2))

breg2a <- glm(prop1 ~ distance + ball,
  weights = shots, family = binomial, data = trashball)
summary(breg2a)
anova(breg2a, breg2, test = "Chisq")
```

- a. Interpret each of the 4 coefficients in context.

The odds of making a basket with a ping pong ball from 0 feet are 8.61; that is, the probability of making a basket with a ping pong ball from 0 feet is 8.61 / (1 + 8.61) = .896.

For ping pong balls, the odds of making a basket decrease by 20.4% (1-.796) for each additional foot in distance. Or, the odds increase by 25.6% (1/.796) for each foot closer.

At 0 feet, the odds of making a basket with a tennis ball are 82.4% lower than with a ping pong ball. Or, the odds of making a basket with a ping pong ball are 5.68 times higher.

For tennis balls, the odds of making a basket decrease by 5.0% (1-.950) for each additional foot in distance, compared to 20.4% with ping pong balls. Or, the odds for tennis balls increase by 5.3% (1/.950) for each foot closer. Note that .950 = exp(-.228+.177).

- b. Is there significant evidence that the effect of distance is reduced with tennis balls? Support your answer with two different hypothesis tests.

No, at least not at the strict .05 level of significance, although there is marginal evidence that points in that direction. The ball by distance interaction term is not statistically significant (Z=1.908, p=.0563 by Wald test; Dev=3.678, p=.0551 by drop in deviance test).

7. Is there significant evidence of a height effect after accounting for hours awake in addition to our two experimental variables of distance and order of balls? Support your answer with a properly interpreted 95% confidence interval.

```
breg3 = glm(prop1 ~ distance + tennisfirst + awake + height,
          weights = shots, family = binomial, data = trashball)
summary(breg3)
exp(confint(breg3))
```

Yes. After accounting for hours awake, order of balls, and distance, the odds of making a basket increases between 5.4% to 22.0% for each additional inch of height.

8. What can you conclude about the goodness of fit in your model from (7)? Name one or two factors that might cause any lack of fit.

```
gof = 1-pchisq(breg3$deviance, breg3$df.residual)
gof
```

There is no significant evidence of lack of fit (p=.122) in our model from (7) despite lack of independence—each person is represented twice in the data set. In addition, we could be missing important predictors; for instance, we have no interaction terms.

9. Fit a final model using any combination of predictors you find useful, and then generate predicted probabilities. Did you exceed expectations with your own performance or not? Confirm with actual and predicted probabilities.

```
# Construct table of predicted values and prediction errors
breg4 = glm(prop1 ~ distance + tennisfirst + height,
          weights = shots, family = binomial, data = trashball)
summary(breg4)

bin.prop = trashball$prop1
model.est = predict(breg4, type="response")
dev.resid = residuals(breg4, type="deviance")
```

```
display1 = data.frame(id = trashball$id,
  distance = trashball$distance,
  order = trashball$order,
  ball = trashball$ball,
  height = trashball$height,
  bin.prop, model.est, dev.resid)
display1
```

The table generated by the R code above is based on the model in (7) without hours awake. In this case, student #12, for example, performed better than expected with a ping pong ball but worse than expected with a tennis ball.

10. Use quasi-likelihood methods to include an overdispersion parameter in your model from (6). Focus on the `distance:ball` term:

```
# Adjusting for extra-binomial variation (overdispersion)
breg2q <- glm(prop1 ~ distance + ball + distance:ball,
  weights = shots, family = quasibinomial, data = trashball)
summary(breg2q)
phihat = sum(residuals(breg2, type="pearson")^2) /
  breg2$df.residual
phihat
sqrt(phihat)
exp(coef(breg2q))
exp(confint(breg2q))   # profile likelihood CIs

# Generate CIs by hand after adjusting for
#   extra-binomial variation
qlcoef = summary(breg2q)$coefficients[,1]
qlse = summary(breg2q)$coefficients[,2]
lb = qlcoef - qt(.975,breg2q$df.residual)*qlse
ub = qlcoef + qt(.975,breg2q$df.residual)*qlse
cbind(exp(lb),exp(ub))

# Evaluate interaction after adjusting for overdispersion
breg2qa = glm(prop1 ~ distance + ball,
  weights = shots, family = quasibinomial, data = trashball)
summary(breg2qa)
anova(breg2qa, breg2q, test = "F")

breg2a = glm(prop1 ~ distance + ball,
```

```
  weights = shots, family = binomial, data = trashball)
summary(breg2a)

dropindev <- breg2a$deviance - breg2$deviance
d <- breg2a$df.residual - breg2$df.residual
Fstat = (dropindev/d) / phihat
pval = 1-pf(Fstat, d, breg2$df.residual)
Fstat
pval
beta = summary(breg2q)$coefficients[4,1]
sebeta = summary(breg2q)$coefficients[4,2]
tstat = beta / sebeta
pval = 2*(1-pt(abs(tstat), breg2q$df.residual))
tstat
pval
```

- a. Report new estimates for the `distance:ball` coefficient, its standard error, and the overdispersion parameter.

The coefficient is unchanged (0.177) but its SE has grown to .093 from .115. The overdispersion parameter is 1.532.

- b. Report test statistics and p-values for the interaction term from an adjusted Wald test and Drop in Deviance test.

Adjusted Wald test: t(36)=1.542, p=.132 Adjusted drop in deviance test: F(1,36)=2.400, p=.130

- c. Report a new 95% confidence interval for the exponentiated interaction coefficient. Compare with your CI from the model in (6).

New CI after adjusting for overdispersion is (0.954, 1.498), which is slightly wider than the CI from (6). [The adjusted Wald-type CI is (0.946, 1.506).]

7

Correlated Data

```
# Packages required for Chapter 7
library(gridExtra)
library(knitr)
library(kableExtra)
library(lme4)
library(ICC)
library(knitr)
library(tidyverse)
```

Thought Questions

```
set.seed(2)  # to get the same simulated results as reported here
```

```
pi_1a <- rep(0.5, 24)
count_1a <- rbinom(24, 10, pi_1a)

pi_1b <- rbeta(24,.5,.5)
count_1b <- rbinom(24, 10, pi_1b)
```

```
theoretical_pi <- tibble(x = 1:250000,
                         p1 = rbeta(x, 0.5, 0.5))

tibble(x = 1:24, pi_1b) %>%
  ggplot() +
    geom_histogram(bins = 5, aes(x = pi_1b, y = ..density..),
                   color = "black", fill = "white") +
    coord_cartesian(xlim = c(0,1)) +
```

```
    geom_density(data = theoretical_pi, aes(x = p1),
                 bw = 0.025, linetype = 3) +
    geom_vline(xintercept = 0.5, color = "black", lwd = 2) +
    labs(title = "", x = "Probability of Deformity")
```

```
scenario_1 <- tibble(pi_1a, count_1a, pi_1b, count_1b) %>%
  mutate(phat_1a = count_1a / 10, phat_1b = count_1b / 10)

hist_1a <- ggplot(data = scenario_1, aes(x = count_1a)) +
  geom_histogram(bins = 5, color = "black", fill = "white") +
  coord_cartesian(xlim = c(0, 10)) +
  labs(title = "Scenario 1a: Binomial, p = 0.5",
       x = "Count of deformed pups per dam")

hist_1b <- ggplot(data = scenario_1, aes(x = count_1b)) +
  geom_histogram(bins = 5, color = "black", fill = "white") +
  coord_cartesian(xlim = c(0, 10)) +
  labs(title = "Scenario 1b: Binomial, p ~ Beta(0.5, 0.5)",
       x = "Count of deformed pups per dam")

grid.arrange(hist_1a, hist_1b, ncol=1)
```

```
scenario_1 %>%
  summarise(mean_1a = mean(count_1a), sd_1a = sd(count_1a),
            mean_1b = mean(count_1b), sd_1b = sd(count_1b) )
```

1. Will the counts of deformed pups for dams in Scenario 1a behave like a binomial distribution with $n = 10$ and $p = 0.5$ (that is, like counting heads in 10 flips of a fair coin)? Why or why not?

Yes. Every dam is assumed to have exactly 10 pups, and each pup is assumed to have a probability of exactly 0.5 of being deformed, regardless of their dam.

2. Will the counts of deformed pups for dams in Scenario 1b behave like a binomial distribution with $n = 10$ and $p = 0.5$ (that is, like counting heads in 10 flips of a fair coin)? If not, extend the coin flipping analogy to Scenario 1b.

No. Although every dam has exactly 10 pups, each dam has a unique probability of having deformed pups. So some pups have a much higher chance of being deformed and some much lower, depending on their dam. The probabilities across all dams just happen to average out to 0.5. In terms of coins, we could envision this as each dam having a unique weighted coin, so that each pup from a specific dam has the same weighted coin flipped for them.

3. Is Scenario 1b realistic? Why might some dams have higher probabilities than others?

Sure. Some dams might be prone to having deformed pups based on genetics, diet, environment, etc.

```
fit_1a_binom <- glm(phat_1a ~ 1, family=binomial,
                    weight=rep(10,24), data = scenario_1)
summary(fit_1a_binom)
# estimated odds of deformity
exp(coef(fit_1a_binom))
# estimated prob of deformity
exp(coef(fit_1a_binom)) / (1+exp(coef(fit_1a_binom)))

confint(fit_1a_binom)
exp(confint(fit_1a_binom))
exp(confint(fit_1a_binom)) / (1 + exp(confint(fit_1a_binom)))

gof <- 1-pchisq(fit_1a_binom$deviance, fit_1a_binom$df.residual)
gof    # not in textbook Rmd
```

```
fit_1a_quasi = glm(phat_1a ~ 1, family=quasibinomial,
                   weight=rep(10,24), data=scenario_1)
summary(fit_1a_quasi)
# estimated odds of deformity
exp(coef(fit_1a_quasi))
# estimated prob of deformity
exp(coef(fit_1a_quasi)) / (1+exp(coef(fit_1a_quasi)))

confint(fit_1a_quasi)
exp(confint(fit_1a_quasi))
exp(confint(fit_1a_quasi)) / (1 + exp(confint(fit_1a_quasi)))
```

```
fit_1b_binom <- glm(phat_1b ~ 1, family=binomial,
                    weight=rep(10,24), data = scenario_1)
summary(fit_1b_binom)
# estimated odds of deformity
exp(coef(fit_1b_binom))
# estimated prob of deformity
exp(coef(fit_1b_binom)) / (1+exp(coef(fit_1b_binom)))

confint(fit_1b_binom)
exp(confint(fit_1b_binom))
exp(confint(fit_1b_binom)) / (1 + exp(confint(fit_1b_binom)))

gof <- 1-pchisq(fit_1b_binom$deviance, fit_1b_binom$df.residual)
gof    # not in textbook Rmd
```

```
fit_1b_quasi = glm(phat_1b ~ 1, family=quasibinomial,
                   weight=rep(10,24), data=scenario_1)
summary(fit_1b_quasi)
# estimated odds of deformity
exp(coef(fit_1b_quasi))
# estimated prob of deformity
exp(coef(fit_1b_quasi)) / (1+exp(coef(fit_1b_quasi)))

confint(fit_1b_quasi)
exp(confint(fit_1b_quasi))
exp(confint(fit_1b_quasi)) / (1 + exp(confint(fit_1b_quasi)))
```

4. Describe how the quasibinomial analysis of Scenario 1b differs from the binomial analysis of the same simulated data. Refer to Table 7.1 when answering this question; you will need to run the R code in the R markdown file for this chapter to completely fill out the table. Do confidence intervals contain the true model parameters?

There is a great deal of overdispersion in Scenario 1b, so the SE in the quasi-binomial analysis is 2.6 times greater than in the binomial analysis. Thus, even though the estimated intercept is the same, the t-statistic in the quasi-binomial analysis is much smaller and the p-value much larger, and the confidence interval much wider. In the binomial analysis, the p-value was incorrectly significant (since the true intercept is exactly 0) and the confidence interval incorrectly did not contain 0.5. Once we adjust for extra-binomial variation in the quasi-binomial analysis, the p-value is correctly above 0.05 and the confidence interval correctly contains 0.5.

5. Why are differences between quasibinomial and binomial models of Scenario 1a less noticeable than the differences in Scenario 1b?

There is no extra-binomial variation to adjust for in Scenario 1a, since each dam has the same probability so that pups from different dams behave similarly in terms of their chance of being deformed. Thus, the binomial and quasi-binomial analyses for Scenario 1a are very similar, with the estimated overdispersion parameter differing from 1 slightly just by chance.

```
x <- 0:3
p_2 <- exp(-2+4/3*x)/(1+exp(-2+4/3*x))
p_2
```

```
set.seed(1)

dose <- c(rep(0,6),rep(1,6),rep(2,6),rep(3,6))

pi_2a <- exp(-2+4/3*dose)/(1+exp(-2+4/3*dose))
count_2a <- rbinom(24, 10, pi_2a)

b <- 2
a <- b*pi_2a / (1-pi_2a)
pi_2b <- rbeta(24, a, b)
count_2b <- rbinom(24, 10, pi_2b)
```

```
scenario_2 <- tibble(dose, pi_2a, count_2a, pi_2b, count_2b)
theoretical_pi <- tibble(x = 1:50000,
  p1 = rbeta(x, shape1 = 2*p_2[1]/(1-p_2[1]), shape2 = 2),
  p2 = rbeta(x, shape1 = 2*p_2[2]/(1-p_2[2]), shape2 = 2),
  p3 = rbeta(x, shape1 = 2*p_2[3]/(1-p_2[3]), shape2 = 2),
  p4 = rbeta(x, shape1 = 2*p_2[4]/(1-p_2[4]), shape2 = 2))

hist1 <- ggplot() +
  geom_histogram(data = scenario_2[1:6,], bins = 5,
                 aes(x = pi_2b, y = ..density..),
                 color = "black", fill = "white") +
  coord_cartesian(xlim = c(0,1)) +
  geom_density(data = theoretical_pi, aes(x = p1),
               bw = 0.05, linetype = 3) +
```

```
  #stat_function(fun = dbeta,
  #   args = list(shape1 = 2*0.119/(1-0.119), shape2 = 2),
  #   xlim = c(0.01,1)) +
  geom_vline(xintercept = p_2[1], color = "black", lwd = 2) +
  labs(title = "Dosage = 0 mg", x = "Probability of Deformity")


hist2 <- ggplot() +
  geom_histogram(data = scenario_2[7:12,],
                 aes(x = pi_2b, y = ..density..), bins = 5,
                 color = "black", fill = "white") +
  coord_cartesian(xlim = c(0,1)) +
  geom_density(data = theoretical_pi, aes(x = p2),
               bw = 0.05, linetype = 3) +
  geom_vline(xintercept = p_2[2], color = "black", lwd = 2) +
  labs(title = "Dosage = 1 mg", x = "Probability of Deformity")

hist3 <- ggplot() +
  geom_histogram(data = scenario_2[13:18,],
                 aes(x = pi_2b, y = ..density..), bins = 5,
                 color = "black", fill = "white") +
  coord_cartesian(xlim = c(0,1)) +
  geom_density(data = theoretical_pi, aes(x = p3),
               bw = 0.05, linetype = 3) +
  geom_vline(xintercept = p_2[3], color = "black", lwd = 2) +
  labs(title = "Dosage = 2 mg", x = "Probability of Deformity")

hist4 <- ggplot() +
  geom_histogram(data = scenario_2[19:24,],
                 aes(x = pi_2b, y = ..density..), bins = 5,
                 color = "black", fill = "white") +
  coord_cartesian(xlim = c(0,1)) +
  geom_density(data = theoretical_pi, aes(x = p4),
               bw = 0.05, linetype = 3) +
  geom_vline(xintercept = p_2[4], color = "black", lwd = 2) +
  labs(title = "Dosage = 3 mg", x = "Probability of Deformity")

grid.arrange(hist1, hist2, hist3, hist4, ncol=1)
```

```
scenario_2 %>%
  summarise(mean_2a = mean(count_2a), sd_2a = sd(count_2a),
            mean_2b = mean(count_2b), sd_2b = sd(count_2b) )
```

```
scenario2Tab <- scenario_2 %>%
          group_by(dose) %>%
          summarise(mean_2a_pi = round(mean(pi_2a),3),
                    sd_2a_pi = round(sd(pi_2a),3),
                    mean_2a_cnt = round(mean(count_2a),3),
                    sd_2a_cnt = round(sd(count_2a),3),
                    mean_2b_pi = round(mean(pi_2b),3),
                    sd_2b_pi = round(sd(pi_2b),3),
                    mean_2b_cnt = round(mean(count_2b),3),
                    sd_2b_cnt = round(sd(count_2b),3)) %>%
          as.data.frame()
colnames(scenario2Tab) <- c("Dosage","Mean p", "SD p",
    "Mean Count", "SD Count", "Mean p", "SD p",
    "Mean Count", "SD Count")
kable(scenario2Tab, booktabs = T,
    caption="Summary statistics of Scenario 2 by dose.") %>%
    add_header_above(c(" " = 1, "Scenario 2a" = 4,
                       "Scenario 2b" = 4)) %>%
    kable_styling(latex_options = "scale_down") %>%
    column_spec(c(4:5,8:9), width = "1cm")
```

6. Compare and contrast the probabilities associated with the 24 dams under Scenarios 1a and 1b to the probabilities under Scenarios 2a and 2b.

In Scenario 1a, all 24 dams had fixed probability of 0.5 of having a deformed pup, while in Scenario 2a 6 dams have fixed probability of .119, 6 have fixed probability of .339, 6 have .661, and 6 have .881. Note that the fixed probabilities in Scenario 2a average out to 0.5 across all 24 dams. In Scenario 1b, each dam has a unique probability, where probabilities average 0.5 but are more likely to be near 0 or 1. In Scenario 2b, 6 dams have a unique probability selected from a distribution with a mean of .119, 6 dams have a unique probability suggested from a distribution with a mean of .339, etc.

7. In Scenario 2a, dams produced 4.79 deformed pups on average, with standard deviation 3.20. Scenario 2b saw an average of 4.67 with standard deviation 3.58. Explain why comparisons by dose are more meaningful than these overall comparisons. You might refer to the results in Table 7.2.

As in Scenarios 1a and 1b, we expect similar mean counts but greater variability at each fixed dose in Scenario 2b compared to 2a. But when all doses are lumped together, these trends get muddled.

8. In Table 7.1, predict what you'll see in the column headed "CI_odds_ratio". Among the 4 entries: What can you say about the center and the width of the confidence intervals? Which will be similar and why? Which will be different and how?

We expect the center to be similar in all 4 entries, since estimated coefficients should be similar in 2a and 2b. We expect CIs for binomial and quasi-binomial in 2a to be similar in width, since there should be no extra-binomial variation to adjust for, but we expect the quasi-binomial CI in 2b to be wider than the binomial CI, since there is correlation with dam and thus extra-binomial variation to adjust for. Our quasi-binomial CI in 2b should have closer to a 95% probability of including the true odds ratio across many replications.

```
scenario_2 <- scenario_2 %>%
  mutate(dose = dose,
         phat_2a = count_2a / 10, phat_2b = count_2b / 10,
       logit_2a = log ((count_2a + 0.5) / (10 - count_2a + 0.5)),
       logit_2b = log ((count_2b + 0.5) / (10 - count_2b + 0.5)) )

fit_2a_binom = glm(phat_2a ~ dose, family=binomial,
                   weight=rep(10,24), data=scenario_2)
summary(fit_2a_binom)
exp(coef(fit_2a_binom))   # not in textbook Rmd
exp(confint(fit_2a_binom))

gof<-1-pchisq(fit_2a_binom$deviance,fit_2a_binom$df.residual)
gof    # not in textbook Rmd
```

```
fit_2a_quasi = glm(phat_2a ~ dose, family=quasibinomial,
                   weight=rep(10,24), data=scenario_2)
summary(fit_2a_quasi)
exp(coef(fit_2a_quasi))   # not in textbook Rmd
exp(confint(fit_2a_quasi))
```

```
fit_2b_binom = glm(phat_2b ~ dose, family=binomial,
                   weight=rep(10,24), data=scenario_2)
summary(fit_2b_binom)
exp(coef(fit_2b_binom))   # not in textbook Rmd
exp(confint(fit_2b_binom))
```

```
gof <- 1-pchisq(fit_2b_binom$deviance, fit_2b_binom$df.residual)
gof    # not in textbook Rmd
```

```
fit_2b_quasi = glm(phat_2b ~ dose, family=quasibinomial,
                   weight=rep(10,24), data=scenario_2)
summary(fit_2b_quasi)
exp(coef(fit_2b_quasi))   # not in textbook Rmd
exp(confint(fit_2b_quasi))
```

9. Describe how the quasibinomial analysis of Scenario 2b differs from the binomial analysis of the same simulated data. Refer to Table 7.1 when answering this question; you will need to run the R code in the R markdown file for this chapter to completely fill out the table. Do confidence intervals contain the true model parameters?

Both analyses give the same estimated coefficient and odds ratio, but the SE for the quasi-binomial analysis is larger, leading to a lower t-statistic, higher p-value, and wider CI. In this case, both CIs contain the true odds ratio (3.79), but we'd expect across many simulations the quasi-binomial CI would have closer to 95% coverage than the binomial CI.

10. Why are differences between quasibinomial and binomial models of Scenario 2a less noticeable than the differences in Scenario 2b?

The estimated ϕ is smaller for Scenario 2a, which is what we'd expect (see next answer).

11. Why does Scenario 2b contain correlated data that we must account for, while Scenario 2a does not?

There is structurally no extra-binomial variation to adjust for in Scenario 2a, since all pups at a single dose behave similarly, regardless of dam (all dams at a single dose have the same probability of deformity); any difference of ϕ from 1 is due to sampling variability. In Scenario 2b, there is extra-binomial variation to adjust for, since the results from pups at a single dose depend on their dam and the specific probability associated with that dam.

TABLE 7.1: Summary of simulations for Dams and Pups case study.

Scenario	Model	Model Name	β_0	SE β_0	t	p value	ϕ	Est prob	CI prob	Mean count	SD count	GOF p value
1a	Binomial	fit_1a_binom	.067	.129	0.52	.61	1	.517	(.454, .579)	5.17	1.49	.568
	Quasibinomial	fit_1a_quasi	.067	.122	0.55	.59	.89	.517	(.457, .576)	X	X	X
1b	Binomial	fit_1b_binom	.268	.130	2.06	.039	1	.567	(.504, .628)	5.67	4.10	0
	Quasibinomial	fit_1b_quasi	.268	.341	0.79	.44	6.86	.567	(.402, .722)	X	X	X
Scenario	**Model**	**Model Name**	β_1	**SE** β_1	t	**p value**	ϕ	**Est odds ratio**	**CI odds ratio**	**Mean count Dose=1**	**SD count Dose=1**	**GOF p value**
2a	Binomial	fit_2a_binom	1.26	.164	7.72	1.5e-11	1	3.54	(2.61, 4.96)	3.17	1.84	.093
	Quasibinomial	fit_2a_quasi	1.26	.184	6.87	6.8e-07	1.27	3.54	(2.51, 5.19)	X	X	X
2b	Binomial	fit_2b_binom	1.46	.180	8.11	5.0e-16	1	4.31	(3.09, 6.27)	3.50	2.88	.00098
	Quasibinomial	fit_2b_quasi	1.46	.250	5.84	7.1e-06	1.93	4.31	(2.74, 7.35)	X	X	X

7.1 Exercises

7.1.1 Conceptual Exercises

1. **Examples with correlated data.**

a) *Nurse stress study.*

- Nurses
- Ward and hospital
- Job-related stress (probably normal)
- We expect correlation in stress levels between nurses on the same ward, and between wards at the same hospital due to patients served, environment, etc.
- fixed = experience, age (nurse level), type (ward level), size (hospital level); random = ward, hospital

b) *Epilepsy study.*

- Patient visit
- Patient
- Number of seizures (Poisson)

- We expect correlation in seizure counts between visits from the same patient (some tend to have lots of seizures, some tend to have fewer).
- fixed = visit number / time (visit level), age and sex (patient level); random = patient

c) *Cockroaches!*

- Room within apartment unit within building
- Apartment, building (assuming multiple apartment units per building)
- Number of cockroaches caught (Poisson)
- We expect correlation between cockroach counts in different rooms from the same apartment, and correlation between counts from different units in the same building.
- fixed = room type (room level), tenant income (apartment level), building age (building level); random = apartment, building

d) *Prairie restoration.*

- weekly measurement per plant
- plant, pot
- plant height (normal), did a plant germinate (binary)
- We expect correlation between measurements at different times from the same plant, and between measurements between different plants from the same pot.
- fixed = time (time point within plant), species, soil type, and sterilization (pot level); random = plant, pot

e) *Radon in Minnesota.*

- level of a specific home
- home, county
- radon measurement (possibly normal after log transformation)
- we expect radon concentrations in different levels from the same home to be correlated, and we also expect correlation between nearby homes (e.g., from the same county)

- fixed = upper/lower level (level of home), uranium (county); random = home, county

f) *Teen alcohol use.*

- time point within teen
- teen
- alcohol use (probably normal)
- we expect measurements over time from the same teen to be correlated (some tend to drink more, some less)
- fixed = age (time level), coa, male, and peer (teen level); random = teen

2. **More dams and pups** Since we are fitting binomial regression models to the count of deformed pups for each dam, we can simply weight each count by the litter size.

7.1.2 Guided Exercises

1. **Exploring Beta distributions.**

a) 0 to 1

b) The probability that an individual pup from a dam is deformed.

c) Both. The mean is $E(Y) = \alpha/(\alpha + \beta)$ and the standard deviation is $SD(Y) = \sqrt{\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$.

d) The mean is 0.5 and the distribution is symmetric around 0.5.

e) The density peaks above 0.5 if $\alpha > \beta$ and vice versa.

f) Larger α or β decreases the SD.

g) Big differences lead to highly skewed distributions, with mass concentrated near 0 (if β is larger) or 1 (if α is larger).

h) Small α and large β.

i) Large α and small β.

j) Equal α and β where both are very small (near 0).

k) You could find the mean and SD of the proportions of deformed pups across dams in the experiment, and then use the formulas from (c) to solve for α and β after substituting the sample mean for E(Y) and the sample

standard deviation for SD(Y). Or you could determine two equations and two unknowns in other ways—25th and 75th percentiles, etc.

2. **Dams and pups (continued)**.

a) Simply change the line `pi_1b <- rbeta(24,.5,.5)` by replacing α (the first 0.5) and β (the second 0.5) with any other equal values and re-run the code.

b) As one example, for Scenario 1a, fix π at 0.2 for all 24 dams, and in Scenario 1b, π is randomly chosen from Beta (1, 4) which has mean 0.2.

```
## Scenario 1: log (pi / 1 - pi) = 0
##    a) pi fixed at 0.2 for all 24 dams
##    b) pi randomly chosen from Beta (1, 4) which has mean 0.2

# Generate pi for each of 24 dams in sample under
#   Scenarios 1a and 1b
set.seed(53)
pi_1a <- rep(0.2, 24)
pi_1b <- rbeta(24, 1, 4)

# generate deformed pups for each of 24 dams under Scenario 1a
count_1a <- rbinom(24, 10, pi_1a)
# generate deformed pups for each of 24 dams under Scenario 1b
count_1b <- rbinom(24, 10, pi_1b)
scenario_1 <- data.frame(pi_1a, count_1a, pi_1b, count_1b)
scenario_1

# Compare histograms of counts under both scenarios
hist_1a <- ggplot(data = scenario_1, aes(x = count_1a)) +
  geom_histogram(bins = 5) + coord_cartesian(xlim = c(0, 10))
hist_1b <- ggplot(data = scenario_1, aes(x = count_1b)) +
  geom_histogram(bins = 5) + coord_cartesian(xlim = c(0, 10))
grid.arrange(hist_1a, hist_1b, ncol=1)

# compare mean and variance for two scenarios
scenario_1 %>%
  summarise(mean_1a = mean(count_1a), sd_1a = sd(count_1a),
            mean_1b = mean(count_1b), sd_1b = sd(count_1b) )

scenario_1 <- scenario_1 %>%
  mutate(phat_1a = count_1a / 10, phat_1b = count_1b / 10)
```

```
# Model Scenario 1a data without overdispersion
fit_1a_binom <- glm(phat_1a ~ 1, family=binomial,
                    weight=rep(10,24), data = scenario_1)
summary(fit_1a_binom)
# estimated odds of deformity
exp(coef(fit_1a_binom))
# estimated prob of deformity
exp(coef(fit_1a_binom)) / (1+exp(coef(fit_1a_binom)))

ci.prof.noadj <- exp(confint(fit_1a_binom))
# CI for odds - profile likelihood
ci.prof.noadj
# CI for prob - profile likelihood
ci.prof.noadj / (1 + ci.prof.noadj)


# Model Scenario 1a data with overdispersion
fit_1a_quasi <- glm(phat_1a ~ 1, family=quasibinomial,
                    weight=rep(10,24), data = scenario_1)
summary(fit_1a_quasi)
# estimated odds of deformity
exp(coef(fit_1a_quasi))
# estimated prob of deformity
exp(coef(fit_1a_quasi)) / (1+exp(coef(fit_1a_quasi)))

ci.prof.adj <- exp(confint(fit_1a_quasi))
# CI for odds - profile likelihood
ci.prof.adj
# CI for prob - profile likelihood
ci.prof.adj / (1 + ci.prof.adj)


# Model Scenario 1b data without overdispersion
fit_1b_binom <- glm(phat_1b ~ 1, family=binomial,
                    weight=rep(10,24), data = scenario_1)
summary(fit_1b_binom)
# estimated odds of deformity
exp(coef(fit_1b_binom))
# estimated prob of deformity
exp(coef(fit_1b_binom)) / (1+exp(coef(fit_1b_binom)))

ci.prof.noadj <- exp(confint(fit_1b_binom))
# CI for odds - profile likelihood
ci.prof.noadj
```

```
# CI for prob - profile likelihood
ci.prof.noadj / (1 + ci.prof.noadj)

gof<-1-pchisq(fit_1b_binom$deviance,fit_1b_binom$df.residual)
gof                                  # test for goodness of fit


# Model Scenario 1b data with overdispersion
fit_1b_quasi = glm(phat_1b ~ 1, family=quasibinomial,
                   weight=rep(10,24), data=scenario_1)
summary(fit_1b_quasi)
# estimated odds of deformity
exp(coef(fit_1b_quasi))
# estimated prob of deformity
exp(coef(fit_1b_quasi)) / (1+exp(coef(fit_1b_quasi)))

# overdispersion parameter
phihat = sum(residuals(fit_1b_quasi, type="pearson")^2) /
  fit_1b_quasi$df.residual
phihat

ci.prof.adj <- exp(confint(fit_1b_quasi))
# CI for odds - profile likelihood
ci.prof.adj
# CI for prob - profile likelihood
ci.prof.adj / (1 + ci.prof.adj)
```

As expected, Scenarios 1a and 1b are very similar, with estimated overdispersion parameter $\hat{\phi} = 1.22$.

In Scenario 1b, the estimated overdispersion parameter is $\hat{\phi} = 2.18$, and there is significant lack of fit (p = .00011 based on residual deviance of 56.8 on 23 df) in a model without the overdispersion parameter.

In Scenario 2b, the CI for odds before adjusting for overdispersion = (.122, .248), while the CI for odds after adjusting for overdispersion = (.101, .289). Note that the true odds = .2 / .8 = .25. The CI after adjusting is wider, and it contains the true odds while the CI before adjusting does not.

Similarly, the CI for probability before adjusting for overdispersion = (.109, .199), while the CI for probability after adjusting for overdispersion = (.092, .224). Note that the true probability = .20. The CI after adjusting is wider, and it contains the true probability while the CI before adjusting does not.

c) The R code below will set up 3 doses with probabilities of .018, .119, and .500 and log odds given by $-4 + 2 * dose$. Scenario 2b probabilities at each

dose are then chosen according to a Beta distribution with $\beta = 0.5$ and α chosen so that the expected probability equals $\frac{\alpha}{\alpha+\beta}$.

```
set.seed(1133)

dose <- c(rep(0,8),rep(1,8),rep(2,8))

pi_2a <- exp(-4+2*dose)/(1+exp(-4+2*dose))
count_2a <- rbinom(24, 10, pi_2a)

b <- .5
a <- b*pi_2a / (1-pi_2a)
pi_2b <- rbeta(24, a, b)
count_2b <- rbinom(24, 10, pi_2b)
```

```
scenario_2 <- tibble(dose, pi_2a, count_2a, pi_2b, count_2b)
scenario2Tab <- scenario_2 %>%
          group_by(dose) %>%
          summarise(mean_2a_pi = round(mean(pi_2a),3),
                    sd_2a_pi = round(sd(pi_2a),3),
                    mean_2a_cnt = round(mean(count_2a),3),
                    sd_2a_cnt = round(sd(count_2a),3),
                    mean_2b_pi = round(mean(pi_2b),3),
                    sd_2b_pi = round(sd(pi_2b),3),
                    mean_2b_cnt = round(mean(count_2b),3),
                    sd_2b_cnt = round(sd(count_2b),3)) %>%
          as.data.frame()
colnames(scenario2Tab) <- c("Dosage","Mean p", "SD p",
    "Mean Count", "SD Count", "Mean p", "SD p",
    "Mean Count", "SD Count")
kable(scenario2Tab, booktabs = T,
    caption="Summary statistics of Scenario 2 by dose.") %>%
    add_header_above(c(" " = 1, "Scenario 2a" = 4,
                       "Scenario 2b" = 4)) %>%
    kable_styling(latex_options = "scale_down") %>%
    column_spec(c(4:5,8:9), width = "1cm")

scenario_2 <- scenario_2 %>%
  mutate(dose = dose,
         phat_2a = count_2a / 10, phat_2b = count_2b / 10,
    logit_2a = log ((count_2a + 0.5) / (10 - count_2a + 0.5)),
```

```
    logit_2b = log ((count_2b + 0.5) / (10 - count_2b + 0.5)) )

fit_2a_binom = glm(phat_2a ~ dose, family=binomial,
                   weight=rep(10,24), data=scenario_2)
summary(fit_2a_binom)
exp(coef(fit_2a_binom))   # not in textbook Rmd
exp(confint(fit_2a_binom))

gof<-1-pchisq(fit_2a_binom$deviance,fit_2a_binom$df.residual)
gof    # not in textbook Rmd

fit_2a_quasi = glm(phat_2a ~ dose, family=quasibinomial,
                   weight=rep(10,24), data=scenario_2)
summary(fit_2a_quasi)
exp(coef(fit_2a_quasi))   # not in textbook Rmd
exp(confint(fit_2a_quasi))

fit_2b_binom = glm(phat_2b ~ dose, family=binomial,
                   weight=rep(10,24), data=scenario_2)
summary(fit_2b_binom)
exp(coef(fit_2b_binom))   # not in textbook Rmd
exp(confint(fit_2b_binom))

gof<-1-pchisq(fit_2b_binom$deviance,fit_2b_binom$df.residual)
gof    # not in textbook Rmd

fit_2b_quasi = glm(phat_2b ~ dose, family=quasibinomial,
                   weight=rep(10,24), data=scenario_2)
summary(fit_2b_quasi)
exp(coef(fit_2b_quasi))   # not in textbook Rmd
exp(confint(fit_2b_quasi))
```

```
x <- 0:2
p_2 <- exp(-4+2*x)/(1+exp(-4+2*x))
p_2

theoretical_pi <- tibble(x = 1:50000,
  p1 = rbeta(x, shape1 = .5*p_2[1]/(1-p_2[1]), shape2 = .5),
  p2 = rbeta(x, shape1 = .5*p_2[2]/(1-p_2[2]), shape2 = .5),
  p3 = rbeta(x, shape1 = .5*p_2[3]/(1-p_2[3]), shape2 = .5))

hist1 <- ggplot() +
```

```
  geom_histogram(data = scenario_2[1:8,], bins = 5,
                 aes(x = pi_2b, y = ..density..),
                 color = "black", fill = "white") +
  coord_cartesian(xlim = c(0,1)) +
  geom_density(data = theoretical_pi, aes(x = p1),
               bw = 0.05, linetype = 3) +
  geom_vline(xintercept = p_2[1], color = "black", lwd = 2) +
  labs(title = "Dosage = 0 mg", x = "Probability of Deformity")

hist2 <- ggplot() +
  geom_histogram(data = scenario_2[9:16,],
                 aes(x = pi_2b, y = ..density..), bins = 5,
                 color = "black", fill = "white") +
  coord_cartesian(xlim = c(0,1)) +
  geom_density(data = theoretical_pi, aes(x = p2),
               bw = 0.05, linetype = 3) +
  geom_vline(xintercept = p_2[2], color = "black", lwd = 2) +
  labs(title = "Dosage = 1 mg", x = "Probability of Deformity")

hist3 <- ggplot() +
  geom_histogram(data = scenario_2[17:24,],
                 aes(x = pi_2b, y = ..density..), bins = 5,
                 color = "black", fill = "white") +
  coord_cartesian(xlim = c(0,1)) +
  geom_density(data = theoretical_pi, aes(x = p3),
               bw = 0.05, linetype = 3) +
  geom_vline(xintercept = p_2[3], color = "black", lwd = 2) +
  labs(title = "Dosage = 2 mg", x = "Probability of Deformity")

grid.arrange(hist1, hist2, hist3, ncol=1)
```

8

Introduction to Multilevel Models

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

8.1 Exercises

8.1.1 Conceptual Exercises

1. **Housing prices.** (a) L1 = house, L2 = neighborhood; (b) L1 = size, age, bedrooms, bathrooms, landscaping, etc.; L2 = quality of local schools, median housing price in neighborhood, median neighborhood income, commute time from neighborhood to central business district, etc.

2. The neighborhood factors apply fairly equally to all houses in a neighborhood. If we assumed all houses were independent, we would be overstating the effective sample size; two houses next door to each other contribute less than 2 independent pieces of information to our model, since certain effects are common to both houses.

3. The first model has 10 parameters to estimate:

- Level One:
$$Y_{ij} = a_i + b_i sqft_{ij} + c_i bedrooms_{ij} + \epsilon_{ij}$$

- Level Two:

$$a_i = \alpha_0 + u_i$$
$$b_i = \beta_0 + v_i$$
$$c_i = \gamma_0 + w_i$$

- Composite model:

$$Y_{ij} = \alpha_0 + \beta_0 sqft_{ij} + \gamma_0 bedrooms_{ij} + [\epsilon_{ij} + u_i + v_i sqft_{ij} + w_i bedrooms_{ij}]$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

$$\begin{bmatrix} u_i \\ v_i \\ w_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & & \\ \sigma_{uv} & \sigma_v^2 & \\ \sigma_{uw} & \sigma_{vw} & \sigma_w^2 \end{bmatrix} \right).$$

The second model has 5 parameters to estimate:

- Level One:

$$Y_{ij} = a_i + \epsilon_{ij}$$

- Level Two:

$$a_i = \alpha_0 + \alpha_1 income_i + \alpha_2 schools_i + u_i$$

- Composite model:

$$Y_{ij} = \alpha_0 + \alpha_1 income_i + \alpha_2 schools_i + [\epsilon_{ij} + u_i]$$

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

The third model has 19 parameters to estimate:

- Level One:

$$Y_{ij} = a_i + b_i sqft_{ij} + c_i bedrooms_{ij} + d_i age_{ij} + \epsilon_{ij}$$

- Level Two:

$$a_i = \alpha_0 + \alpha_1 medprice_i + u_i$$
$$b_i = \beta_0 + \beta_1 medprice_i + v_i$$
$$c_i = \gamma_0 + \gamma_1 medprice_i + w_i$$
$$d_i = \delta_0 + \delta_1 medprice_i + x_i$$

- Composite model:

$$\begin{aligned} Y_{ij} = {} & \alpha_0 + \alpha_1 medprice_i + \beta_0 sqft_{ij} + \beta_1 medprice_i sqft_{ij} + \gamma_0 bedrooms_{ij} + \\ & \gamma_1 medprice_i bedrooms_{ij} + \delta_0 age_{ij} + \delta_1 medprice_i age_{ij} + \\ & [\epsilon_{ij} + u_i + v_i sqft_{ij} + w_i bedrooms_{ij} + x_i age_{ij}] \end{aligned}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

$$\begin{bmatrix} u_i \\ v_i \\ w_i \\ x_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & & & \\ \sigma_{uv} & \sigma_v^2 & & \\ \sigma_{uw} & \sigma_{vw} & \sigma_w^2 & \\ \sigma_{ux} & \sigma_{vx} & \sigma_{wx} & \sigma_x^2 \end{bmatrix} \right).$$

The fourth model has 12 parameters to estimate:

- Level One:
$$Y_{ij} = a_i + b_i sqft_{ij} + \epsilon_{ij}$$
- Level Two:

$$a_i = \alpha_0 + \alpha_1 income_i + \alpha_2 schools_i + \alpha_3 medprice_i + u_i$$
$$b_i = \beta_0 + \beta_1 income_i + \beta_2 schools_i + \beta_3 medprice_i + v_i$$

- Composite model:

$$\begin{aligned} Y_{ij} = {} & \alpha_0 + \alpha_1 income_i + \alpha_2 schools_i + \alpha_3 medprice_i + \beta_0 sqft_{ij} + \\ & \beta_1 income_i sqft_{ij} + \beta_2 schools_i sqft_{ij} + \beta_3 medprice_i sqft_{ij} + \\ & [\epsilon_{ij} + u_i + v_i sqft_{ij}] \end{aligned}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

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

4. **Music performance anxiety.** The two plots can differ when the number of performances differ among musicians. Suppose, for instance, one vocalist who is very anxious had twice as many performances as anybody else. In that case, plot (a) would show a higher level of anxiety for vocalists compared to other instruments than plot (b). On the other hand, if a keyboardist had only one performance and she felt no anxiety in that one performance, that single performance would count just as much in plot (b) as a keyboardist with 20 performances averaged together, so plot (b) would show a lower anxiety level for keyboardists.

5. a_i denotes the true mean negative affect when Subject i is playing solos or small ensembles in the theoretical population of all Subject i performances, while $\widehat{a}_i$ denotes the predicted/estimated mean negative affect when Subject i is playing solos or small ensembles based on the data collected.

6. The tilt in plot (b) indicates a positive correlation between two error terms

– larger values of one are associated with larger values of the other. In plot (a), there is no correlation between error terms.

7. Independence does not account for the fact that observational units with the same Level Two unit carry overlapping information, making the effective sample size smaller. Thus, standard errors under independence divide by a larger n, making them smaller than under multilevel methods. LVCF simply reduces the sample size to the number of Level Two units.

8. Estimates of Level One and Level Two mean responses in Model A are not adjusted for any covariates at either level, giving us "unconditional means". Since every subject's intercept is a random variable centered at the same spot, Model A can also be referred to as a "random intercepts model"; at Level Two, there is only one random effect, and it is the subject effect on the intercept term. These two labels are entirely consistent.

9. The new model has 10 parameters to estimate (3 fixed effects and 7 variance components):

- Level One:
$$Y_{ij} = a_i + b_i LargeEns_{ij} + c_i SmallEns_{ij} + \epsilon_{ij}$$
- Level Two:
$$\begin{aligned} a_i &= \alpha_0 + u_i \\ b_i &= \beta_0 + v_i \\ c_i &= \gamma_0 + w_i \end{aligned}$$
- Composite model:
$$\begin{aligned} Y_{ij} = \alpha_0 + \beta_0 LargeEns_{ij} + \gamma_0 SmallEns_{ij} + \\ [\epsilon_{ij} + u_i + v_i LargeEns_{ij} + w_i SmallEns_{ij}] \end{aligned}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

$$\begin{bmatrix} u_i \\ v_i \\ w_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & & \\ \sigma_{uv} & \sigma_v^2 & \\ \sigma_{uw} & \sigma_{vw} & \sigma_w^2 \end{bmatrix} \right).$$

β_0 now represents the true mean difference in performance anxiety between large ensembles and solos.

10. In general it depends on whether or not Covariate A is involved in any interaction terms. If not, then its coefficient can be interpreted "holding all else constant". If it does appear in interaction terms, then the covariates that interact with Covariate A must be "set to 0" when interpreting the coefficient of Covariate A.

11. If LargeEns=1 and Orch=0, then

$$\hat{Y}_{ij} = \hat{\alpha}_0 + \hat{\alpha}_2 MPQnem_i + \hat{\beta}_0 + \hat{\beta}_2 MPQnem_i$$

On the other hand, if LargeEns=1 and Orch=1, then

$$\hat{Y}_{ij} = \hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\alpha}_2 MPQnem_i + \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 MPQnem_i$$

The difference between these two equations is: $\hat{\alpha}_1 + \hat{\beta}_1$, holding MPQNEM constant.

12.

a) $\hat{\alpha}_0$

A = mean performance anxiety across all subjects and all performances
B = same as A (all subjects) except only solos and small ensembles
C = same as B (solos and small ensembles) except only keyboardists and vocalists
D = same as C except only subjects with baseline negative emotionality of 0
E = same as D except only subjects with baseline negative emotionality of 31.63 (average)

b) $\hat{\beta}_0$

B = mean difference in performance anxiety across all subjects when playing in large ensembles rather than solos and small ensembles
C = same as B except only keyboardists and vocalists
D = same as C except only subjects with baseline negative emotionality of 0
E = same as D except only subjects with baseline negative emotionality of 31.63 (average)

c) $\hat{\alpha}_1$

C = difference in mean performance anxiety in solos and small ensembles for orchestral instrumentalists vs. keyboardists and vocalists
D = same as C except controlling for baseline negative emotionality
E = same as D

d) $\hat{\beta}_1$

C = mean difference in the effect of playing in large ensembles (as opposed to solos and small ensembles) for orchestral instrumentalists vs. keyboardists and vocalists
D = same as C except controlling for baseline negative emotionality
E = same as D

e) $\hat{\sigma}_u$

A = SD in between-subject deviations in performance anxiety
B = same as A except only solos and small ensembles
C = same as B except controlling for instrument played
D = same as C except controlling for baseline negative emotionality as well
E = same as D

f) $\hat{\sigma}_v$

B = SD in between-subject deviations in differences in performance anxiety levels between large ensembles and other performance types
C = same as B except controlling for instrument played
D = same as C except controlling for baseline negative emotionality as well
E = same as D

13. We haven't changed our Level One model, so $\hat{\sigma}_u^2$ should theoretically remain the same, but because all parameters are estimated simultaneously in a multilevel model (we don't estimate in stages like in the Two Stage model), estimates of performance anxiety and associated residuals can change and even get slightly worse.

14. **Model F**

$\hat{\alpha}_0$ = 8.37. The estimated mean performance anxiety for a musician's first diaried performance in a small or large ensemble played in front of an instructor for the population of keyboard players and vocalists with negative emotionality, positive emotionality, and absorption of 0 at baseline.

$\hat{\alpha}_2$ = 0.20. A one point increase in baseline absorption is associated with an estimated 0.20 mean increase in anxiety levels, after controlling for previous diary entries, audience, group size, positive emotionality, negative emotionality, and instrument.

$\hat{\alpha}_3$ = 1.53. The mean anxiety for an orchestral instrument player is an estimated 1.53 points higher than the mean anxiety of keyboard players and vocalists, after controlling for previous diary entries, audience, group size, positive emotionality, negative emotionality, and absorption.

$\hat{\beta}_0$ = -0.14. Each additional previous performance is associated with an estimated 0.14 mean decrease in performance anxiety, after controlling for audience, group size, positive emotionality, absorption, negative emotionality, and instrument.

$\hat{\gamma}_0$ = 3.61. The mean anxiety when playing in front of students is an estimated 3.61 points higher than the mean anxiety when playing in front of an instructor, after controlling for previous diary entries, group size, positive emotionality, absorption, negative emotionality, and instrument.

$\hat{\zeta}_0 = 0.51$. For musicians with a negative emotionality of 0, playing a solo is associated with an estimated mean increase in anxiety of 0.51 points compared to playing in an ensemble, after controlling for the effects of previous diary entries, audience, positive emotionality, absorption, and instrument.

$\hat{\rho}_{wx}$= 0.835. The estimated population correlation between increases in anxiety scores for performances before juries and before students is 0.835. Those musicians with higher anxiety scores (compared to the modeled scores) in front of juries tend to have higher scores (compared to the modeled scores) in front of students as well.

$\hat{\sigma} = 3.91$. The estimated population standard deviation in residuals for the individual regression models is 3.91 points.

$\hat{\sigma}_u = 3.80$. The estimated population standard deviation of anxiety levels for a musician's first diaried performance in an ensemble played in front of an instructor for the population of keyboard players and vocalists, after controlling for negative emotionality, positive emotionality, and absorption and instrument played.

$\hat{\sigma}_x = 4.28$. The estimated population standard deviation of differences in anxiety levels between performances before juries and before instructors is 4.28 points.

15. OLS regression would fit the pattern of the entire collection of points, minimizing the sum of squared residuals; this pattern is generally upward through the four clusters of points, since subjects with more previous performances also tend to have higher negative affect. Multilevel modeling essentially fits lines separately to each subject – each of which slopes downward – and then combines those 4 slopes into an overall model – the negatively sloping dashed black line.

16. While both OLS regression and multilevel modeling produce similar estimates of fixed effect coefficients, OLS regression tends to underestimate the standard errors for those coefficients, which would, in turn, overestimate the statistical significance of those fixed effects.

8.1.2 Guided Exercises

1. **Music performance joy.** [Sadler and Miller, 2010].

```
music <- read_csv("data/musicdata.csv")
music <- music %>%
  mutate(orch = ifelse(instrument=="orchestral instrument",1,0),
         large = ifelse(perform_type=="Large Ensemble",1,0),
         students = ifelse(audience=="Student(s)",1,0),
```

```
         juried = ifelse(audience=="Juried Recital",1,0),
         public = ifelse(audience=="Public Performance",1,0),
         instructor = ifelse(audience=="Instructor",1,0),
         solo = ifelse(perform_type=="Solo",1,0),
         memory1 = ifelse(memory=="Memory",1,0),
         female = ifelse(gender=="Female",1,0),
         vocal = ifelse(instrument=="voice",1,0),
         cmpqab = mpqab - mean(mpqab),
         male = ifelse(gender=="Male", 1, 0),
         c_years_study = years_study - mean(years_study))

box.perform <- ggplot(data=music,aes(factor(perform_type),pa)) +
  geom_boxplot() +
  coord_flip() +
  ylab("Positive affect") + xlab("")
box.memory <- ggplot(data=music,aes(factor(memory),pa)) +
  geom_boxplot() +
  coord_flip() +
  ylab("Positive affect") + xlab("")
box.audience <- ggplot(data=music,aes(factor(audience),pa)) +
  geom_boxplot() +
  coord_flip() +
  ylab("Positive affect") + xlab("")
scatter.previous <- ggplot(data=music, aes(x=previous,y=pa)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Positive affect") + xlab("Previous Performances")
mli.boxscatmat1 <- grid.arrange(box.perform,box.memory,
  box.audience,scatter.previous,ncol=2)

instr.all <- ggplot(data=music,aes(factor(instrument),pa)) +
  geom_boxplot() +
  coord_flip() +
  ylab("Positive affect") + xlab("") + ylim(10,50)
gender.all <- ggplot(data=music,aes(factor(gender),pa)) +
  geom_boxplot() + coord_flip() +
  ylab("Positive affect") + xlab("") + ylim(10,50)
mli.boxmat1 <- grid.arrange(instr.all, gender.all ,ncol=1)

scatter.age <- ggplot(data=music, aes(x=age,y=pa)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Positive affect") + xlab("Age")
scatter.years <- ggplot(data=music, aes(x=years_study,y=pa)) +
```

```
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Positive affect") + xlab("Years of Study")
scatter.pem <- ggplot(data=music, aes(x=mpqpem,y=pa)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Positive affect") + xlab("PEM")
scatter.nem <- ggplot(data=music, aes(x=mpqnem,y=pa)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Positive affect") + xlab("NEM")
scatter.con <- ggplot(data=music, aes(x=mpqcon,y=pa)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Positive affect") + xlab("Constraint")
scatter.abs <- ggplot(data=music, aes(x=mpqab,y=pa)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Positive affect") + xlab("Absorbtion")
mli.scatmat1 <- grid.arrange(scatter.age, scatter.years,
  scatter.pem, scatter.nem, scatter.con, scatter.abs, ncol=3)

hist(music$pa ,xlab="Positive affect",main="")

# Summary statistics by categorical covariates
by(music$pa,music$perform_type,summary)
by(music$pa,music$memory,summary)
by(music$pa,music$audience,summary)
by(music$pa,music$gender,summary)
by(music$pa,music$instrument,summary)

# Correlation coefficients by continuous covariates
select <- dplyr::select
main <- music %>%
  select(pa,previous,age,years_study,mpqpem,mpqnem,mpqcon,mpqab)
cor(main)

# Relationships among covariates
prop.table(table(music$perform_type,music$audience),2)
prop.table(table(music$perform_type,music$instrument),2)
prop.table(table(music$memory, music$perform_type),2)
```

1) Unlike negative affect, positive affect appears to be normally distributed across all subjects and performances. Higher levels of happiness seem to be

associated with performing in large ensembles (mean 35.4) rather than small ensembles (29.7) or solos (31.5), playing by memory (mean 34.2) instead of from a score (31.1), and playing in front of juries (mean 37.7) or the public (34.2) rather than instructors (30.1) or students (29.1). Males (mean 34.3) also express slightly higher happiness than females (31.3), and orchestral instrument players (mean 32.6) and vocal musicians (33.5) express higher happiness than keyboardists (28.2). There appears to be little trend in happiness across performances and at different ages, though. We do see that happiness tends to increase as years of study (correlation -.085) and negative emotionality (r=-.18) decrease, and as positive emotionality (r=.15) and absorption (r=.15) increase. Of course, many of these can be confounded; for instance, large ensembles may tend to feature orchestral instruments and public audiences and playing from memory.

2)

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

$\hat{\alpha}_0 = 32.56$, $\hat{\sigma}^2 = 41.70$, $\hat{\sigma}_u^2 = 23.72$

$\hat{\rho} = 23.72 \ / \ (23.72 + 41.70) = 0.363$

36.3% of the total variability in pre-performance happiness scores are attributable to differences among subjects.

3)

```
#Model B (Add instructor and students as Level 1 covariates)
model.b <- lmer(pa ~ instructor + students + (instructor +
    students|id), REML=T, data=music)
summary(model.b)
```

$\hat{\alpha}_0 = 34.73$, $\hat{\beta}_0 = -4.19$, $\hat{\gamma}_0 = -4.45$, $\hat{\sigma}^2 = 36.39$, $\hat{\sigma}_u^2 = 20.34$, $\hat{\sigma}_v^2 = 11.61$, $\hat{\sigma}_w^2 = 12.08$

- $\hat{\alpha}_0 = 34.73$ = estimated mean happiness for juried recitals and public performances for the populations of musicians
- $\hat{\beta}_0 = -4.19$ = subjects have an estimated mean decrease in pre-performance happiness levels of 4.19 points when playing in front of instructors rather than juries or the general public

- $\hat{\sigma}_u = 4.51$ = the estimated population standard deviation of pre-performance happiness levels for juried recitals and public performances

$PseudoR^2_{L1}$ = (41.70 - 36.39) / 41.70 = .127

12.7% of the within-person variability in pre-performance happiness scores can be explained by audience (instructor vs. students vs. others).

4) Model C has 13 parameters to estimate:

- Level One:
$$Y_{ij} = a_i + b_i instructor_{ij} + c_i students_{ij} + \epsilon_{ij}$$
- Level Two:
$$\begin{aligned} a_i &= \alpha_0 + \alpha_1 cmpqab_i + u_i \\ b_i &= \beta_0 + \beta_1 cmpqab_i + v_i \\ c_i &= \gamma_0 + \gamma_1 cmpqab_i + w_i \end{aligned}$$
- Composite model:

$$\begin{aligned} Y_{ij} = {} & \alpha_0 + \alpha_1 cmpqab_i + \beta_0 instructor_{ij} + \beta_1 cmpqab_i instructor_{ij} + \\ & \gamma_0 students_{ij} + \gamma_1 cmpqab_i students_{ij} + \\ & [\epsilon_{ij} + u_i + v_i instructor_{ij} + w_i students_{ij}] \end{aligned}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

$$\begin{bmatrix} u_i \\ v_i \\ w_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_u & & \\ \sigma_{uv} & \sigma^2_v & \\ \sigma_{uw} & \sigma_{vw} & \sigma^2_w \end{bmatrix} \right).$$

5)

```
# Model C (Center absorption used as predictor for ints and slopes)
model.c <- lmer(pa ~ instructor + students + cmpqab +
    cmpqab:students + cmpqab:instructor +
    (instructor + students|id), REML=T, data=music)
summary(model.c)
```

$\hat{\alpha}_0 = 34.82$, $\hat{\beta}_0 = -4.25$, $\hat{\gamma}_0 = -4.65$, $\hat{\alpha}_1 = -0.02$, $\hat{\beta}_1 = 0.37$, $\hat{\gamma}_1 = 0.29$

$\hat{\sigma} = 6.04$, $\hat{\sigma}_u = 4.51$, $\hat{\sigma}_v = 2.84$, $\hat{\sigma}_w = 3.28$, $\hat{\rho}_{uv} = .095$, $\hat{\rho}_{vw} = .601$

- $\hat{\alpha}_0 = 34.82$ = estimated mean happiness for juried recitals and public performances for the populations of musicians with average levels of openness to absorbing sensory and imaginative experiences

- $\hat{\alpha}_1$ = -0.02 = for each 1 point increase in baseline absorption level, subjects have estimated mean happiness before performance in front of juries and the general public that is 0.02 points lower

- $\hat{\gamma}_0$ = -4.65 = subjects with average levels of absorption have an estimated mean decrease in pre-performance happiness levels of 4.65 points when playing in front of other students rather than juries or the general public

- $\hat{\beta}_1$ = 0.37 = for each 1 point increase in baseline absorption level, subjects performing before instructors have estimated mean happiness that is 0.35 points higher, compared to a 0.02 decrease for performances for juries and the general public. Thus, the effect of absorption for instructor performances is 0.37 points greater than the effect of absorption on happiness for juried recitals and public performances.

- $\hat{\sigma}_u$ = 4.51 = the estimated population standard deviation of pre-performance happiness levels for juried recitals and public performances, after controlling for absorption levels

- $\hat{\sigma}_v$ = 2.84 = the estimated population standard deviation of differences in pre-performance happiness levels between instructor audiences and juried recitals or public performances, after controlling for absorption levels

- $\hat{\rho}_{uv}$ = 0.095 = the estimated population correlation between errors in pre-performance happiness levels for juried recitals or public performances and errors in increases in happiness for instructor audiences, after controlling for absorption levels. There is little relationship between the two.

6)

```
# Model D (Add male at Level Two)
model.d <- lmer(pa ~ instructor + students + cmpqab + male +
    cmpqab:students + cmpqab:instructor + male:students +
    male:instructor + (instructor + students|id), REML=T,
    data=music)
summary(model.d)
```

$\hat{\alpha}_0$ = 33.92, $\hat{\beta}_0$ = -3.72, $\hat{\gamma}_0$ = -4.73, $\hat{\alpha}_1$ = -0.06, $\hat{\beta}_1$ = 0.40, $\hat{\gamma}_1$ = 0.29, $\hat{\alpha}_2$ = 3.05, $\hat{\beta}_2$ = -2.21, $\hat{\gamma}_2$ = -0.21

$\hat{\sigma}$ = 6.03, $\hat{\sigma}_u$ = 4.30, $\hat{\sigma}_v$ = 2.86, $\hat{\sigma}_w$ = 3.27, $\hat{\rho}_{uv}$ = .28, $\hat{\rho}_{uw}$ = -.75, $\hat{\rho}_{vw}$ = .43

- $\hat{\alpha}_0$ = 33.92 = estimated mean happiness for juried recitals and public

performances for the populations of **female** musicians with average levels of openness to absorbing sensory and imaginative experiences

- $\hat{\alpha}_1$ = -0.06 = for each 1 point increase in baseline absorption level, subjects have estimated mean happiness before performance in front of juries and the general public that is 0.06 points lower, **after controlling for gender**

- $\hat{\gamma}_0$ = -4.73 = **female** subjects with average levels of absorption have an estimated mean decrease in pre-performance happiness levels of 4.73 points when playing in front of other students rather than juries or the general public

- $\hat{\beta}_1$ = 0.40 = for each 1 point increase in baseline absorption level, subjects performing before instructors have estimated mean happiness that is 0.34 points higher **after controlling for gender**, compared to a 0.06 decrease for performances for juries and the general public. Thus, the effect of absorption for instructor performances is 0.40 points greater than the effect of absorption on happiness for juried recitals and public performances.

- $\hat{\sigma}_u$ = 4.30 = the standard deviation in happiness scores before juried recitals and public performances, after controlling for absorption **and gender**

- $\hat{\sigma}_v$ = 2.86 = the standard deviation of differences in pre-performance happiness levels between instructor audiences and juried recitals or public performances, after controlling for absorption **and gender**

- $\hat{\rho}_{uv}$ = 0.281 = the correlation between errors in happiness scores before juried recitals or public performances and errors in increases (or decreases) in happiness for instructor audiences, after controlling for absorption **and gender**

7)

- $\hat{\alpha}_2$ = 3.05 = male subjects have estimated mean happiness before performances in front of juries and the general public that is 3.05 points greater than female subjects, after controlling for baseline absorption

- $\hat{\beta}_2$ = -2.21 = after controlling for baseline absorption, male subjects have estimated mean happiness before performances in front of instructors which is 0.84 points higher than female subjects, compared to 3.05 points higher in front of juries and the general public

8)

```
# Model comparisons (it will refit models with ML)
anova(model.c,model.d)
```

We do not have significant evidence (deviance statistic = 5.79, p-value=.122 based on chi-square distribution with 3 df) that Model D is preferable to Model C – it does not pay off to add gender at Level Two.

$PseudoR^2_{L2_u}$ = (20.32-18.53)/20.32 = .088. We reduced musician-to-musician variability in the intercept (the mean happiness score for juried recitals and public performances) by 8.8% by adding gender as a predictor.

$PseudoR^2_{L2_v}$ = (8.06-8.19)/8.06 = -.016. We increased musician-to-musician variability in the instructor effect by 1.6% by adding gender as a predictor. This is likely due to the effect of simultaneous estimation, so it's better to say that gender explains little of the musician-to-musician variability in the instructor effect.

$PseudoR^2_{L2_w}$ = (10.7237-10.7192)/10.7237 = .0004. We reduced musician-to-musician variability in the student effect by 0.04% by adding gender as a predictor.

AIC-D = 3310.4 vs. AIC-C = 3310.2; BIC-D = 3377.7 vs. BIC-C = 3364.9. Both AIC and BIC indicate that Model C is better, so that adding gender as a predictor in all equations at Level Two does not improve the model.

8.1.3 Open-Ended Exercises

1. **Political ambiguity.** [Chapp et al., 2018].

```
ambiguity <- read_csv("data/ambiguity.csv")

ggplot(data = ambiguity, aes(x = demHeterogeneity,
                             y = ambiguity)) +
    geom_point() +
    geom_smooth(method = lm)

ambiguity %>%
  mutate(party = ifelse(democrat == 1,
                        "Democrat", "Republican")) %>%
  ggplot(aes(x = ideology, y = ambiguity, color = party)) +
    geom_point() +
    geom_smooth(method = lm)
```

```
amb.mod1 <- lmer(ambiguity ~ democrat + incumbent +
  demHeterogeneity + attHeterogeneity + mismatch + ideology +
  distLean + (1 |distID), data=ambiguity)
summary(amb.mod1)

amb.mod2 <- lmer(ambiguity ~ democrat + incumbent +
  demHeterogeneity + attHeterogeneity + ideology + distLean +
  (1 |distID), data=ambiguity)
summary(amb.mod2)

anova(amb.mod1, amb.mod2)

amb.mod3 <- lmer(ambiguity ~ democrat + incumbent +
  demHeterogeneity + attHeterogeneity + mismatch + ideology +
  distLean + democrat:ideology + (1 |distID), data=ambiguity)
summary(amb.mod3)
AIC(amb.mod3)
```

The best model with main effects only includes all predictors except `mismatch` and shows that the SD at Level One is more than 10 times greater than the SD at Level Two. We will, however, use a multilevel model with all main effects and the interaction between party and ideology in order to assess each of these hypotheses (in this model, Level Two variability does not really even register compared with Level One variability):

Hypothesis 1b - "when incumbents do hazard issue statements, these statements will be marked by a higher degree of clarity.". Yes - `incumbent` has significant positive association with clarity after accounting for other predictors (t = 3.84).

Hypothesis 2a - "ideological distance [from district residents] will be associated with greater ambiguity." Yes - `mismatch` has significant positive association with ambiguity (negative association with clarity) after accounting for other predictors (t = -3.72).

Hypothesis 2b - "controlling for ideological distance, ideological extremity [of the candidate] should correspond to less ambiguity." Yes - for Republicans, the `ideology` effect is significantly positive (t = 2.78), indicating more clarity for more extreme conservatives. And for Democrats, the `ideology` effect is significantly negative (coefficient estimate of -0.17 with interaction t = -6.44), indicating more clarity for more extreme liberals.

Hypothesis 3a - "more variance in attitudes [among district residents] will correspond to a higher degree of ambiguity in rhetoric". Yes - `attHeterogeneity` has significant negative association with clarity after accounting for other predictors (t = -2.54).

Hypothesis 3b - "a more heterogeneous mix of subgroups [among district residents] will also correspond to a higher degree of ambiguity in rhetoric". Yes - `demHeterogeneity` has significant negative association with clarity after accounting for other predictors (t = -5.66).

Further directions: full EDA; test terms with likelihood ratio test, or better yet, a parametric bootstrap; consider meaningful interactions (e.g., ideology by democrat).

2. **Airbnb in Chicago.** [Trinh and Ameri, 2018]

```
airbnb <- read_csv("data/airbnb.csv")
airbnb <- airbnb %>%
  mutate(HighBlack = ifelse(PctBlack > .6, 1, 0))

ggplot(airbnb, aes(price)) +
  geom_histogram(bins=150, fill = "red") +
  xlim(0,1250) + theme_minimal()+ xlab("Prices") +
  labs(title="Histogram of Listing Prices in Chicago
       (as of 8/21/2016)")

### incomplete look at Level One covariates

# room_type level one covariate
ggplot(airbnb, aes(room_type, price))+
  geom_boxplot() + coord_flip()+
  theme_minimal()+
  labs(x="Room Type", y="Price",
       title="Listing Prices by Room Type")
#bedrooms level one covariate
ggplot(airbnb, aes(bedrooms, price))+
  geom_point()+geom_smooth(method = lm)+
  theme_minimal()+
  labs(x="Number of Bedrooms", y="Price",
       title="Listing Prices by Number of Bedrooms Offered")
#overall_satisfaction level one covariate
ggplot(airbnb, aes(overall_satisfaction, price))+
  geom_point()+geom_smooth(method = lm)+
  theme_minimal()+
  labs(x="overall satisfaction", y="Price",
       title="Listing Prices by overall satisfaction")
cor(airbnb$accommodates, airbnb$bedrooms)

### incomplete look at Level Two covariates
```

```
#summary of listing prices by district
with(airbnb, by(price, district, summary))

#boxplots of prices by HighBlack
airbnb %>%
  mutate(Over60PctBlack = ifelse(HighBlack == 1, "yes",
                                 "no")) %>%
  ggplot(aes(Over60PctBlack, price,color=Over60PctBlack))+
    geom_boxplot()+theme_minimal() +
    coord_flip()+
    labs(title="Listing Prices by Over60PctBlack")

#walkscore, bikescore, transitscore -
#   denotes accessibility and mobility
airbnb %>% group_by(district) %>%
  summarise(listing = n(),
            walkscore = mean(WalkScore),
            bikescore = mean(BikeScore),
            transitscore = mean(TransitScore)) %>%
  arrange(desc(listing))

#ggplot with transit score
ggplot(airbnb, aes(TransitScore, price))+
  geom_point()+geom_smooth(method = lm)+
  labs(x="Transit Score", y="Price",
       title= "Listing Prices by Transit Scores")

# Potential model
airbnb <- airbnb %>%
  mutate(transit_cent = TransitScore - mean(TransitScore),
         bedroom_cent = bedrooms - mean(bedrooms),
         satisfy_cent = overall_satisfaction -
           mean(overall_satisfaction))
mean(airbnb$overall_satisfaction)
mean(airbnb$bedrooms)
mean(airbnb$TransitScore)

a.model <- lmer(price ~ HighBlack + TransitScore + bedrooms +
  HighBlack:bedrooms + TransitScore:bedrooms +
  overall_satisfaction + room_type + (1|neighborhood),
  REML=T, data=airbnb)
summary(a.model)

b.model <- lmer(price ~ HighBlack + transit_cent +
```

```
  bedroom_cent + HighBlack:bedroom_cent +
  transit_cent:bedroom_cent + satisfy_cent + room_type +
  (1|neighborhood), REML=T, data=airbnb)
summary(b.model)
```

Prices are most expensive and most varied for private rooms/apartments. We see that places with fewer bedrooms are less likely to be expensive; however, there are also many places with 2-4 bedrooms that are under $100. There are no listings under $250 that have 5 or more bedrooms, while very few places with zero bedrooms (possibly studio apartments) reach the $250 mark. There is a positive relationship between listing prices with the number of people each place can accommodate; however, the correlation between number of people accommodated and number of bedrooms is high (0.74).

We will discuss a model using price as the response, but one might also consider log(price) given the skewed distribution of prices. A possible model can be written as follows, denoting average nightly price of listing j in neighborhood i. All continuous variables (TransitScore, bedrooms, overall_satisfaction) are centered.

- Level One:

$$Y_{ij} = a_i + b_i bedrooms_{ij} + c_i OverallSatisfaction_{ij} + d_i PrivateRoom_{ij} + e_i SharedRoom_{ij} + \epsilon_{ij}$$

- Level Two:

$$\begin{aligned} a_i &= \alpha_0 + \alpha_1 HighBlack_i + \alpha_2 TransitScore_i + u_i \\ b_i &= \beta_0 + \beta_1 HighBlack_i + \beta_2 TransitScore_i \\ c_i &= \gamma_0 \\ d_i &= \delta_0 \\ e_i &= \varepsilon_0 \end{aligned}$$

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

Interpretations of fixed effects:

- $\hat{\alpha}_0 = 135.30$. The estimated mean price for an airbnb with the room type of entire home/apartment, with 1.4 bedrooms, an overall satisfaction score of 4.76, with a transit score of 71.35, in a neighborhood where less than 60% of the population is black.

- $\hat{\alpha}_1$ = -24.68. When more than 60% of the population in a neighborhood is black, there is an estimated $24.68 mean decrease in the price for an airbnb, for airbnbs with 1.4 bedrooms, after controlling for room_type, overall satisfaction, and transit score.

- $\hat{\alpha}_2$ = 2.39. A one unit increase in transit score is associated with an estimated $2.39 mean increase in the price for an airbnb, for airbnbs with 1.4 bedrooms, after controlling for room_type, overall satisfaction, and proportion of black residents in a neighborhood.

- $\hat{\beta}_0$ = 52.85. One additional bedroom is associated with an estimated $52.85 mean increase in price of the airbnb for airbnbs in neighborhoods with a transit score of 71.35 and percentage of black residents below 60%, after controlling for overall satisfaction and room type.

- $\hat{\beta}_1$ = -46.65. When more than 60% of the population in a neighborhood is black, the addition of one bedroom is associated with an estimated $6.20 mean increase in the price for an airbnb, compared to a $52.85 mean increase per bedroom in neighborhoods with less than 60% black, for airbnbs with a transit score of 71.35, after controlling for room_type and overall satisfaction.

- $\hat{\beta}_2$ = 1.42. The effect of a high transit score on price is greater for units with more bedrooms. For example, mean price increases by $2.39 for each 1 point increase in transit score for units with 1.4 bedrooms, but by $3.81 for each 1 point increase in transit score for units with 2.4 bedrooms, after controlling for overall satisfaction, percentage of black residents, and room type.

- $\hat{\gamma}_0$ = 25.70. A one unit increase in overall satisfaction is associated with an estimated $25.70 mean increase in the price for an airbnb, after controlling for room_type, number of bedrooms, transit score, and proportion of black residents in a neighborhood.

- $\hat{\delta}_0$ = -44.21. Private rooms are associated with average prices that are $44.21 lower than the entire apartment or home, after controlling for number of bedrooms, transit score, overall satisfaction, and proportion of black residents in a neighborhood.

- $\hat{\varepsilon}_0$ = -65.11. Shared rooms are associated with average prices that are $65.11 lower than the entire apartment or home, after controlling for number of bedrooms, transit score, overall satisfaction, and proportion of black residents in a neighborhood.

3. **Project 5183.**

```
rockies <- read_csv("data/FinalRockiesdata.csv")

rockies <- rockies %>%
  rename(K_9 = 'K/9') %>%
  mutate(PCL = ifelse(PCL == 'y', 1, 0),
         Coors = ifelse(Coors == 'y', 1, 0))

table(rockies$ID)
table(rockies$PCL)
table(rockies$Coors)
by(rockies$Pit, rockies$PCL, summary)

velocity <- ggplot(rockies, aes(x = vFA)) +
  geom_histogram(bins = 8, color="black", fill="white") +
  xlim(80, 95) + ylab("Frequency") +
  xlab("Average Fastball Velocity")
strikeout <- ggplot(rockies, aes(x =K_9)) +
  geom_histogram(bins = 8, color="black", fill="white") +
  ylab("Frequency") + xlab("Strikeouts per 9 innings")

avgs <- rockies %>%
  group_by(ID, Age) %>%
  summarise(meanK = mean(K_9),
            meanvel = mean(vFA),
            meanERA = mean(ERA),
            meanpct = mean(Pitpct))
avgs

avgvelocity <- ggplot(avgs, aes(x = meanvel)) +
  geom_histogram(bins = 8, color="black", fill="white") +
  xlim(80, 95) + ylab("Frequency") +
  xlab("Mean Avg Fastball Velocity")
avgstrikeout <- ggplot(avgs, aes(x =meanK)) +
  geom_histogram(binwidth = 1.75, color="black", fill="white") +
  xlim(0,15) + ylab("Frequency") +
  xlab("Mean Strikeouts per 9 innings")

ERA <- ggplot(rockies, aes(x =ERA )) +
  geom_histogram(bins = 12, color="black", fill="white") +
  xlim(0,20) + ylab("Frequency") + xlab("ERA")
pct <- ggplot(rockies, aes(x =Pitpct)) +
  geom_histogram(bins = 12, color="black", fill="white")  +
  ylab("Frequency") + xlab("Pct of strikes")
avgERA <- ggplot(avgs, aes(x = meanERA )) +
```

```
  geom_histogram(bins = 9, color="black", fill="white") +
  xlim(0,20) + ylab("Frequency") + xlab("Mean ERA")
avgpct <- ggplot(avgs, aes(x =meanpct)) +
  geom_histogram(binwidth = .02, color="black", fill="white") +
  xlim(0.5,0.7) + ylab("Frequency") + xlab("Mean Pct of strikes")

grid.arrange(velocity, strikeout, avgvelocity, avgstrikeout,
             ERA, pct, avgERA, avgpct, ncol=2)

ggplot(rockies,aes(x=vFA,y=factor(PCL))) +
  geom_dotplot(binaxis="y",stackdir="center", binwidth = 1/20) +
  xlim(85, 95) + facet_wrap(~ID,ncol=3) +
  theme(strip.text.x=element_blank())

ggplot(rockies,aes(x=K_9,y=factor(PCL))) +
  geom_dotplot(binaxis="y",stackdir="center", binwidth = 1/25) +
  xlim(0, 15) + facet_wrap(~ID,ncol=3) +
  theme(strip.text.x=element_blank())

ggplot(rockies,aes(x=ERA,y=factor(PCL))) +
  geom_dotplot(binaxis="y",stackdir="center", binwidth = 1/20) +
  xlim(0, 20) + facet_wrap(~ID,ncol=3) +
  theme(strip.text.x=element_blank())

ggplot(rockies,aes(x=Pitpct,y=factor(PCL))) +
  geom_dotplot(binaxis="y",stackdir="center", binwidth = 1/25) +
  xlim(0.5, 0.7) + facet_wrap(~ID,ncol=3) +
  theme(strip.text.x=element_blank())

ggplot(rockies,aes(x=vFA,y=factor(Coors))) +
  geom_dotplot(binaxis="y",stackdir="center", binwidth = 1/20) +
  xlim(85, 95) + facet_wrap(~ID,ncol=3) +
  theme(strip.text.x=element_blank())

ggplot(rockies,aes(x=K_9,y=factor(Coors))) +
  geom_dotplot(binaxis="y",stackdir="center", binwidth = 1/25) +
  xlim(0, 15) + facet_wrap(~ID,ncol=3) +
  theme(strip.text.x=element_blank())

ggplot(rockies,aes(x=ERA,y=factor(Coors))) +
  geom_dotplot(binaxis="y",stackdir="center", binwidth = 1/20) +
  xlim(0, 20) + facet_wrap(~ID,ncol=3) +
  theme(strip.text.x=element_blank())
```

```
ggplot(rockies,aes(x=Pitpct,y=factor(Coors))) +
  geom_dotplot(binaxis="y",stackdir="center", binwidth = 1/25) +
  xlim(0.5, 0.7) + facet_wrap(~ID,ncol=3) +
  theme(strip.text.x=element_blank())

rockies <- rockies %>%
  mutate(cage = Age - 27)

a1 <- ggplot(rockies, aes(x = cage, y = vFA)) +
  geom_point() +
  geom_smooth(method = "lm", color = 'black') +
  xlab("Centered Age") + labs(title="(a1)")
b1 <- ggplot(rockies, aes(x = cage, y = K_9)) +
  geom_point() +
  geom_smooth(method = "lm", color = 'black') +
  xlab("Centered Age") + labs(title="(b1)")
c1 <- ggplot(rockies, aes(x = cage, y = ERA)) +
  geom_point() +
  geom_smooth(method = "lm", color = 'black') +
  xlab("Centered Age") + labs(title="(c1)")
d1 <- ggplot(rockies, aes(x = cage, y = Pitpct)) +
  geom_point() +
  geom_smooth(method = "lm", color = 'black') +
  xlab("Centered Age") + labs(title="(d1)")

a2 <- ggplot(avgs, aes(x = Age, y = meanvel)) +
  geom_point() +
  geom_smooth(method = "lm", color = 'black') +
  ylab("Mean vFA") + xlab("Age") + labs(title="(a2)")
b2 <- ggplot(avgs, aes(x = Age, y = meanK)) +
  geom_point() +
  geom_smooth(method = "lm", color = 'black') +
  ylab("Mean K_9") + xlab("Age") + labs(title="(b2)")
c2 <- ggplot(avgs, aes(x = Age, y = meanERA)) +
  geom_point() +
  geom_smooth(method = "lm", color = 'black') +
  ylab("Mean ERA") + xlab("Age") + labs(title="(c2)")
d2 <- ggplot(avgs, aes(x = Age, y = meanpct)) +
  geom_point() +
  geom_smooth(method = "lm", color = 'black') +
  ylab("Mean Pitpct") + xlab("Age") + labs(title="(d2)")

grid.arrange(a1, b1, c1, d1, a2, b2, c2, d2, ncol=4)
```

```
model1a=lmer(vFA~cage+PCL+Coors+(PCL+Coors| ID), REML=T,
             data = rockies)
summary(model1a)

model1a2=lmer(vFA~PCL+Coors+(PCL+Coors| ID), REML=T,
             data = rockies)
summary(model1a2)

anova(model1a, model1a2)

model2a=lmer(Pitpct~cage+PCL+Coors+(PCL+Coors| ID),
             REML=T, data = rockies)
summary(model2a)

model3a=lmer(K_9~cage+PCL+Coors+(PCL+Coors|ID),REML=T,
             data = rockies)
summary(model3a)

model4a=lmer(ERA~cage+PCL+Coors+(PCL+Coors|ID),REML=T,
             data = rockies)
summary(model4a)


model1b <- lmer(vFA ~ cage + PCL + Coors + (1 | ID),
             data = rockies)
summary(model1b)

model2b <- lmer(Pitpct ~ cage + PCL + Coors + (1 | ID),
             data = rockies)
summary(model2b)

model3b <- lmer(K_9 ~ cage + PCL + Coors + (1 | ID),
             data = rockies)
summary(model3b)

model4b <- lmer(ERA ~ cage + PCL + Coors + (1 | ID),
             data = rockies)
summary(model4b)


model1c <- lmer(vFA ~ cage + PCL + Coors +
             cage:PCL + cage:Coors + PCL:Coors + (1 | ID),
             data = rockies)
```

```
summary(model1c)

model2c <- lmer(Pitpct ~ cage + PCL + Coors +
             cage:PCL + cage:Coors + PCL:Coors + (1 | ID),
             data = rockies)
summary(model2c)

model3c <- lmer(K_9 ~ cage + PCL + Coors +
             cage:PCL + cage:Coors + PCL:Coors + (1 | ID),
             data = rockies)
summary(model3c)

model4c <- lmer(ERA ~ cage + PCL + Coors +
             cage:PCL + cage:Coors + PCL:Coors + (1 | ID),
             data = rockies)
summary(model4c)


model1d <- lmer(vFA ~ cage + PCL +  cage:PCL + (1 | ID),
             data = rockies)
summary(model1d)

model2d <- lmer(Pitpct ~ Coors + (1 | ID),
             data = rockies)
summary(model2d)

model3d <- lmer(K_9 ~ cage + PCL +  cage:PCL + (1 | ID),
             data = rockies)
summary(model3d)

model4d <- lmer(ERA ~ cage + PCL + Coors + cage:Coors +
             (1 | ID), data = rockies)
summary(model4d)
```

Since we have such a limited data set documenting the effects of Project 5183 (118 games involving 7 pitchers who had at least one start before pitch count limits were enacted and one start afterwards), most analyses will show only non-statistically significant trends. One pitcher had only 2 starts under pitch count limits, while another pitcher only made 7 starts in all. In addition, the 75-pitch count limit was not strictly enforced; while the median pitch count dropped from 95 to 77, more than half of games pitched under a pitch count limit featured over 75 pitches.

Initial boxplots show some evidence that, under pitch count limits, fastball

velocity, strikeouts, and ERA all decrease, while the percentage of strikes increases. Lattice plots show that these trends are not consistent across all 7 starting pitchers. Initial scatterplots and boxplots also show that the effects of pitching in Coors Field and pitcher age will be important to control for, with ERA increasing during games at Coors Field and fastball velocity decreasing for older pitchers. Note that pitch count limit and Coors Field are both Level One variables, while age is a Level Two variable.

Multilevel models confirm the direction of Project 5183 effects noted in initial plots, even after accounting for the effect of Coors Field and an age-adjusted random effect for each starting pitcher, although few of the effects were statistically significant. Pitching under a pitch count limit was associated with a decrease in average fastball velocity of .27 mph (t = -1.35) for pitchers of average age, a decrease which grew as pitchers aged (for example, average velocity decrease by .92 mph (t = -2.16 for interaction term) for pitchers 5 years older than average). Pitching under a pitch count limit was also associated with no meaningful change in strike percentage after controlling for ballpark, and a decrease in earned runs average of 1.53 (t = -1.60) after controlling for ballpark and age. We also saw (non-significant) evidence that the negative association between age and strikeouts per nine innings only held with no pitch count limit, and that the higher ERA at Coors Field was even worse as pitchers aged.

In summary, there is some evidence from the Colorado Rockies in 2012 that a 75-pitch count limit can be effective in encouraging pitchers to pitch to contact and reduce the number of runs allowed, but results are very preliminary, based only on 7 pitchers from a single team.

4. **Replicate the Sadler and Miller paper.** [Sadler and Miller, 2010].

```
music <- read_csv("data/musicdata.csv")

# Add new indicators to music data set
music <- mutate(music, c_years_study = years_study - 8,
  piano = ifelse(instrument=="keyboard (piano or organ)", 1, 0),
  orchestra = ifelse(instrument=="orchestral instrument", 1, 0),
  solo = ifelse(perform_type=="Solo", 1, 0),
  small = ifelse(perform_type=="Small Ensemble", 1, 0),
  memory1 = ifelse(memory=="Memory", 1, 0),
  students = ifelse(audience=="Student(s)", 1, 0),
  juried = ifelse(audience=="Juried Recital", 1, 0),
  public = ifelse(audience=="Public Performance", 1, 0))

# Models from Table 2
model1 <- lmer(na ~ gender + c_years_study + orchestra + piano +
```

```
  solo + small + students + juried + public + memory1 + diary +
  (1|id), REML=T, data=music)
summary(model1)
AIC(model1)

model2 <- lmer(na ~ mpqnem + c_years_study + orchestra + solo +
  students + juried + public + diary + mpqnem:c_years_study +
  (1|id), REML=T, data=music)
summary(model2)
AIC(model2)

# Calculate variance explained at each level by comparing to
#   unconditional means model
model0 <- lmer(na ~ 1 + (1|id), REML=T, data=music)
summary(model0)
AIC(model0)
```

For the most part, model coefficients and SEs agree very closely between SAS output in the paper and our own R models. The two exceptions are NEM in Model 2 (R coefficient and SE are half as large) and Orch in Model 2 (R coefficient and SE are twice as large). AICs in R are a bit lower (2952.8 for Model 1 and 2953.2 for Model 2) and actually paint a different picture about preferences. Variance explained is similar at both levels for Model 2 (13.5% for L1 and 45.1% for L2) but similar only at L1 for Model 1 (13.9% for L1 and -17.3% for L2).

Issues with Model 2 as presented on page 284:

- audience and performance type should be broken into several indicators, leading to more equations at Level Two. Instead, the authors use audience and performance indicators as L2 variables.
- piano and small ensemble are not included in Model 2

9

Two-Level Longitudinal Data

```
# Packages required for Chapter 9
library(GGally)
library(data.table)
library(Hmisc)
library(mice)
library(lattice)
library(nlme)
library(reshape2)
library(MASS)
library(mnormt)
library(lme4)
library(gridExtra)
library(knitr)
library(kableExtra)
library(broom)
library(tidyverse)
```

9.1 Exercises

9.1.1 Conceptual Exercises

1. **Parenting and gang activity.** [Walker-Barnes and Mason, 2001] (a) L1 = time, L2 = subjects; (b) L1 = linear or quadratic effects of time, hours per week working or in extracurricular activities (if it varies over the year), weather, etc.; L2 = ethnicity, parental behavior and peer gang activity (since these were collected only at baseline), neighborhood, GPA, socioeconomics, etc.

2. The wide format would have one row per subject (300 rows) and separate columns for each of the 8 assessments for subject gang activity. The long

format would have one row per assessment per subject (about 2400 rows) and just a single column for subject gang activity.

3. A lattice plot would make it easier to tell if, for instance, subjects' gang activity increased or decreased in a linear fashion, or if a quadratic or spline model would be more appropriate. Side-by-side spaghetti plots would make it easier to tell if, for instance, initial gang activity or change over time differed (on average) between members of different ethnic groups.

4. Let *time* = assessment number (where the initial assessment is at time 0); thus, an increase of 1 in time is about a one month timespan. Then the first model has 6 parameters to estimate (2 fixed effects and 4 variance components):

- Level One:
$$Y_{ij} = a_i + b_i time_{ij} + \epsilon_{ij}$$
- Level Two:
$$a_i = \alpha_0 + u_i$$
$$b_i = \beta_0 + v_i$$
- Composite model:

$$Y_{ij} = \alpha_0 + \beta_0 time_{ij} + [\epsilon_{ij} + u_i + v_i time_{ij}]$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

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

The second model has 8 parameters to estimate (4 fixed effects and 4 variance components):

- Level One:
$$Y_{ij} = a_i + b_i time_{ij} + \epsilon_{ij}$$
- Level Two:
$$a_i = \alpha_0 + \alpha_1 peergang_i + u_i$$
$$b_i = \beta_0 + \beta_1 peergang_i + v_i$$
- Composite model:

$$\begin{aligned} Y_{ij} = \alpha_0 + \alpha_1 peergang_i + \beta_0 time_{ij} + \\ \beta_1 peergang_i time_{ij} + [\epsilon_{ij} + u_i + v_i time_{ij}] \end{aligned}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

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

At this point, we'll assume ethnicity is a single indicator variable (e.g. Hispanic = 1 and Other = 0). Then the third model has 14 parameters to estimate (10 fixed effects and 4 variance components):

- Level One:
$$Y_{ij} = a_i + b_i time_{ij} + \epsilon_{ij}$$
- Level Two:
$$\begin{aligned} a_i &= \alpha_0 + \alpha_1 peergang_i + \alpha_2 ethnicity_i + \alpha_3 parenting_i + \\ &\quad \alpha_4 parenting_i ethnicity_i + u_i \\ b_i &= \beta_0 + \beta_1 peergang_i + \beta_2 ethnicity_i + \beta_3 parenting_i + \\ &\quad \beta_4 parenting_i ethnicity_i + v_i \end{aligned}$$
- Composite model:

$$\begin{aligned} Y_{ij} &= \alpha_0 + \alpha_1 peergang_i + \alpha_2 ethnicity_i + \alpha_3 parenting_i + \\ &\quad \alpha_4 parenting_i ethnicity_i + \beta_0 time_{ij} + \beta_1 peergang_i time_{ij} + \\ &\quad \beta_2 ethnicity_i time_{ij} + \beta_3 parenting_i time_{ij} + \beta_4 parenting_i ethnicity_i time_{ij} + \\ &\quad [\epsilon_{ij} + u_i + v_i time_{ij}] \end{aligned}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

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

5. Note that Walker-Barnes and Mason's model in Table 2 is slightly different than our model from (4). They treated ethnicity as a categorical variable with 3 levels and thus represented it with two indicator variables (one for Black and one for White/Other, leaving Hispanic as the reference level). This resulted in two additional fixed effects – one in the Level Two model for intercept, and one in the Level Two model for slope.

- Intercept coefficient for Peer behavior = .252. The estimated mean gang involvement at the beginning of a student's ninth grade year increases by .252 for each 1 unit increase in the gang involvement of peers, after controlling for ethnicity and parental psychological control.

- Intercept coefficient for Black ethnicity = .671. The estimated mean gang involvement at the beginning of a student's ninth grade year is .671 greater for Black students than Hispanic students whose parents exhibit no psychological control, after controlling for peer gang involvement.

- Intercept coefficient for Black ethnicity X Parenting = -.161. The estimated difference in the effect of an increase of 1 unit in parental psychological control on initial gang involvement between Black and Hispanic students, after controlling for peer gang involvement. In particular, an increase of 1 unit in parental psychological control is associated with a mean increase in gang involvement of .076 at the beginning of the ninth grade year for Hispanic students, but that same increase in parental psychological control is associated with a mean decrease in gang involvement of .085 for Black students, holding peer gang involvement constant.

- Slope coefficient for Peer behavior = -.011. The estimated mean gang involvement decreases by .011 per month for each 1 unit increase in the gang involvement of peers, after controlling for ethnicity and parental psychological control.

- Slope coefficient for Black ethnicity = -.132. The estimated mean monthly decrease in gang involvement is .132 greater for Black students than Hispanic students whose parents exhibit no psychological control, after controlling for peer gang involvement.

- Slope coefficient for Parenting = -.015. The estimated mean gang involvement decreases by .015 per month for each 1 unit increase in parental psychological control for Hispanic students, after controlling for peer gang involvement.

- Slope coefficient for Black ethnicity X Parenting = .048. The estimated difference in the effect of an increase of 1 unit in parental psychological control on monthly change in gang involvement between Black and Hispanic students, after controlling for peer gang involvement. In particular, an increase of 1 unit in parental psychological control is associated with a mean decrease in gang involvement of .015 per month for Hispanic students, but that same increase in parental psychological control is associated with a mean increase in gang involvement of .033 per month for Black students, holding peer gang involvement constant (note: there is a typo in Table 9.5 and this coefficient should be positive).

6. **Charter schools.** An indicator variable for charter schools should be included in Level Two equations for both intercept and slope (the coefficient for time at Level One).

7. Percent free and reduced lunch should be included in the Level Two equation

for intercept, but probably not slope. The boxplot in (a) illustrates that there is some association between charter schools and percent free and reduced lunch (charter schools tend to have more students in poverty), but probably not strong enough to cause multicollinearity issues. It will be important to control for free and reduced lunch before comparing charter and non-charter schools.

8. Figure 9.15 allows us to assess any potential interactions between charter schools and free and reduced lunch. It appears like the difference in initial math scores between charter and non-charter schools is greater with higher percentages of students in poverty, while the differences in yearly change in math scores between charter and non-charter schools is fairly similar for low and high percent free and reduced lunch students. Thus, an interaction term might be more valuable in the Level Two model for intercept.

9. There would be two sets of boxplots – one for low percent non-white (e.g. below the median) and one for high percent non-white. Then each plot would features boxplots for low percent special education (e.g. below the median) and high percent special education on the same set of axes. Possibly math scores for low percent special education would be higher in one set of boxplots, while math scores for high percent special education would be higher in the other. In that case, we should consider a model with an interaction term in the Level Two equation for intercept.

10. Within-school deviations could be caused by Level One covariates that change over time (e.g. time trends, student-teacher ratio, state funding per student), while between-school deviations could be caused by Level Two covariates that are essentially constant over time (e.g. percent of students in poverty, percent non-white, urban or rural).

11. These models are essentially the same. "Random slopes and intercepts" implies a single covariate at Level One (time in this case) and no covariates in the two Level Two equations, while "unconditional growth" implies the same thing – we model growth with a time covariate at Level One, but then include no Level Two covariates (to make the model unconditional).

12. Model A has just a single Level Two equation, while Model B has two Level Two equations, and the two dependent variables at Level Two are interpreted differently from the single dependent variable in Model A, so there is no "apples-to-apples" comparison at Level Two between Models A and B.

13. First, we'd have to decide the number of pieces and the range of years for each piece. Let's assume we wish to fit two pieces – one from 2001-2005 and one from 2006-2010. Then we could fit the following model, which produces two line segments with different slopes connected in 2005: $\widehat{y} = \widehat{\beta}_1 x_1 + \widehat{\beta}_2 (x_1 - 2005) x_2$, where $x_2 = 1$ if $x_1 > 2005$ and 0 otherwise.

14. Yes. As a Level Two covariate, we are using a school's percent free and reduced lunch compared to other schools as a predictor of initial math scores

and change in math scores. As a Level One covariate, we are testing if the year-to-year changes in a school's percent free and reduced lunch students are predictive of that school's yearly math scores. Using a school's average or 2008 percent free and reduced lunch instead of 2010 values as a Level Two predictor would not change models or interpretations much, assuming that relative comparisons between schools stay similar.

15. A 1% increase in percent free and reduced lunch is associated with a very small change in mean math scores, so a 10% increase is used to demonstrate when a change in free and reduced lunch would be associated with a notable change in average math scores.

16. The first Level Two equation (for intercept) would include a term for the charter by free and reduced lunch interaction.

17. The error structure at Level Two is modeling the pattern of residuals from trying to model schools' 2008 math scores and their yearly change in math scores based on characteristics of those schools. Each residual has underlying variability and can be correlated with the others. On the other hand, the within-school error structure refers to the 3 math scores over time recorded for each school—how much school-to-school variability exists for each year, how closely related are 2008 and 2009 scores from the same school, etc.

18. If it were compound symmetry, all variances and all covariances would be the same. If it were autoregressive, covariances of scores the same number of years apart (e.g. 36.46 and 39.84) would be equal and the correlation 2 years apart would be the square of the correlation 1 year apart. If it were Toeplitz, covariances of scores the same number of years apart (e.g. 36.46 and 39.84) would be equal. If it were our standard two-level model, variances would be quadratically related to time and variance and covariances could be calculated from estimated variance components using formulas in Section 9.7.1.

9.1.2 Guided Exercises

1. **Teen alcohol use.** Article: [Curran et al., 1997]; data source: [Singer and Willett, 2003].

a) Level One = age; Level Two = coa, male, peer

b)

```
alcohol <- read_csv("data/alcohol.csv")

alcohol <- alcohol %>%
  mutate(male2 = ifelse(male==1, "male", "female"),
         coa2 = ifelse(coa==1, "alcoholic parent", "none"),
```

```
        peer01 = ifelse(peer > 0, "peer use above 0",
                        "peer use equals 0"),
        peersum = round(peer^2),
        alcsum = round(alcuse^2),
        age14 = age - 14)
```

```
# Exploratory data analyses #

# Quick look at covariates
# - individually and relationship with response
#   Note: we only have Level Two covariates in this study
#     (other than time at Level One)
#   Also, these plots contains all 246 (dependent) observations

ggplot(alcohol, aes(x = alcuse)) +
  geom_histogram(bins = 8, color="black", fill="white") +
  ylab("Frequency") + xlab("Alcohol use")

plot1 <- ggplot(alcohol, aes(x = coa2, y = alcuse)) +
  geom_boxplot() +
  ylab("Alcohol use") + xlab("Child of an alcoholic")
plot2 <- ggplot(alcohol, aes(x = male2, y = alcuse)) +
  geom_boxplot() +
  ylab("Alcohol use") + xlab("Male")
plot3 <- ggplot(alcohol, aes(x = as.factor(peersum),
                             y = alcuse)) +
  geom_boxplot() +
  ylab("Alcohol use") + xlab("Peer alcohol use")
plot4 <- ggplot(alcohol, aes(x = peersum, y = alcuse)) +
  geom_jitter(width = .25) +
  geom_smooth(method = lm, se = FALSE) +
  ylab("Alcohol use") + xlab("Peer alcohol use")
mli.plotmat1 <- grid.arrange(plot1, plot2, plot3, plot4,ncol=2)

alcohol %>% group_by(coa2) %>%
  summarise(means = mean(alcuse), sds = sd(alcuse),
            meds = median(alcuse),
            q1s = quantile(alcuse, 0.25),
            q3s = quantile(alcuse, 0.75),
            mins = min(alcuse), maxs = max(alcuse), ns = n())
alcohol %>% group_by(male2) %>%
  summarise(means = mean(alcuse), sds = sd(alcuse),
```

```
            meds = median(alcuse),
            q1s = quantile(alcuse, 0.25),
            q3s = quantile(alcuse, 0.75),
            mins = min(alcuse), maxs = max(alcuse), ns = n())
cor(alcohol$alcuse, alcohol$peer)
```

Alcohol use is right skewed, mostly because of the large number of observations from non-drinkers. We will analyze this variable without transformation (since logging or other options won't help in this case), but we will consider other options later in the course.

Alcohol use is more prevalent among children of alcoholics (mean of 1.3) than children of non-alcoholics (0.6), and there is a fairly strong positive correlation between peer alcohol use and individual alcohol use (r=0.42). We can examine a boxplot of alcohol use vs. peer use, since most subjects report peer use of 0, 1, 2, 3, or 4. There is little difference between genders in alcohol use (male mean = 0.97, female mean = 0.87).

c)

```
ggplot(alcohol, aes(x = age, y = alcuse)) +
  geom_point() + geom_line() + facet_wrap(~id, ncol = 10) +
  theme(strip.text.x=element_blank()) + ylab("Alcohol use")
```

In general, alcohol use tends to increase between the ages of 14 and 16, although there is large variability in both intercepts (alcohol use at age 14) and slope (rate of increase between 14 and 16). Note that there are 25 subjects with no alcohol use at any time point.

d)

```
spaghetti1 <- ggplot(alcohol, aes(x = age, y = alcuse)) +
  geom_line(aes(group = id), color = "dark grey") +
  facet_grid(~ coa2) +
  geom_smooth(aes(group = 1), color = "black", size = 1)

spaghetti2 <- ggplot(alcohol, aes(x = age, y = alcuse)) +
  geom_line(aes(group = id), color = "dark grey") +
  facet_grid(~ male2) +
  geom_smooth(aes(group = 1), color = "black", size = 1)

spaghetti3 <- ggplot(alcohol, aes(x = age, y = alcuse)) +
```

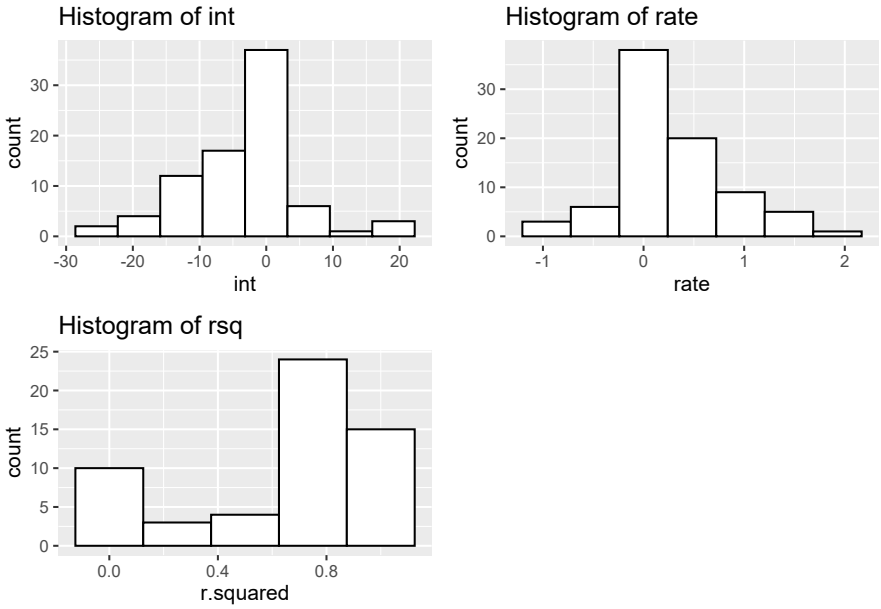
```
  geom_line(aes(group = id), color = "dark grey") +
  facet_grid(~ peer01) +
  geom_smooth(aes(group = 1), color = "black", size = 1)

lon.spaghetti <- grid.arrange(spaghetti1, spaghetti2,
                              spaghetti3, ncol=2)
```

Subjects who are children of alcoholics have more alcohol use at age 14, but their rate of growth is about the same as children of non-alcoholics. Boys and girls start at about the same level, but boys tend to increase their alcohol use at a higher rate, especially between 15 and 16. Subjects with peer usage above 1 start at a much higher level, and their rate of increase is slightly higher than subjects whose peers have low usage.

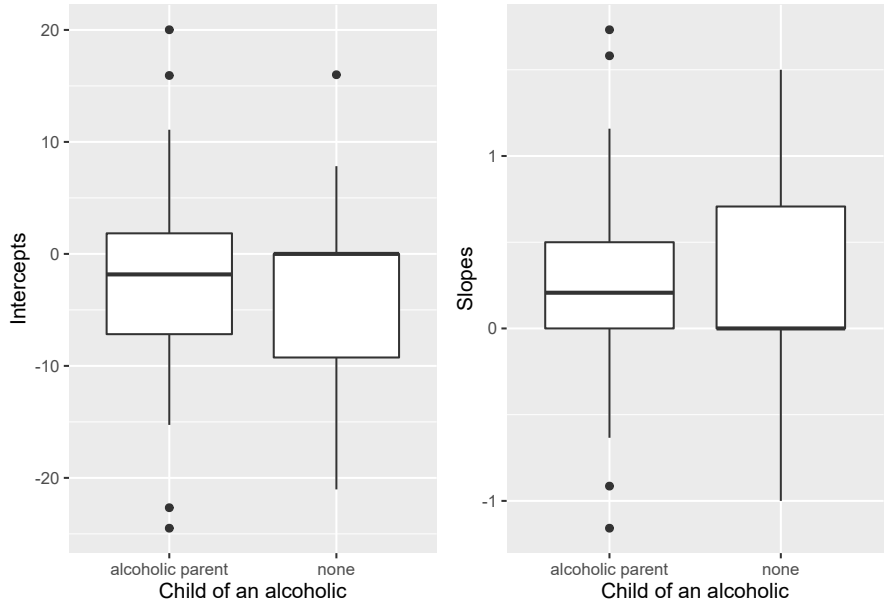
e)

```
select = dplyr::select

regressions <- alcohol %>%
  group_by(id) %>%
  do(fit = lm(alcuse ~ age, data=.))

lm_info1 <- regressions %>%
  tidy(fit) %>%
  ungroup() %>%
  select(id, term, estimate) %>%
  spread(key = term, value = estimate) %>%
  rename(rate = age, int = `(Intercept)`)

lm_info2 <- regressions %>%
  tidy(fit) %>%
  ungroup() %>%
  select(id, term, std.error) %>%
  spread(key = term, value = std.error) %>%
  rename(se_rate = age, se_int = `(Intercept)`)

lm_info <- regressions %>%
  glance(fit) %>%
  ungroup() %>%
  select(id, r.squared, df.residual) %>%
  inner_join(lm_info1, by = "id") %>%
  inner_join(lm_info2, by = "id") %>%
  mutate(tstar = qt(.975, df.residual),
```

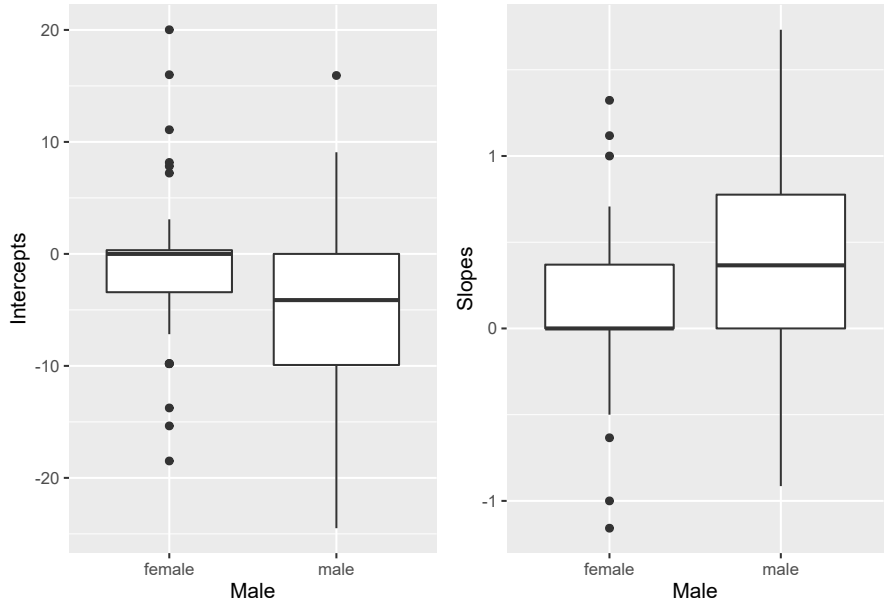
```
         intlb = int - tstar * se_int,
         intub = int + tstar * se_int,
         ratelb = rate - tstar * se_rate,
         rateub = rate + tstar * se_rate)
#head(data.frame(lm_info))

alc.wide <- alcohol %>%
  select(-X1, -alcsum, -peersum, -age14) %>%
  spread(key = age, value = alcuse)
alc.wide <- lm_info %>%
  select(id, int, rate , r.squared) %>%
  right_join(alc.wide, by = "id")

int.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=int), bins = 8, color="black",
                 fill="white") +
  ggtitle("Histogram of int")
rate.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=rate), bins=7, color="black",
                 fill="white") +
  ggtitle("Histogram of rate")
rsq.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=r.squared), bins = 5, color="black",
                 fill="white") +
  ggtitle("Histogram of rsq")
lon.histmat1 <- grid.arrange(int.hist1, rate.hist1,
                             rsq.hist1, ncol=2)
```



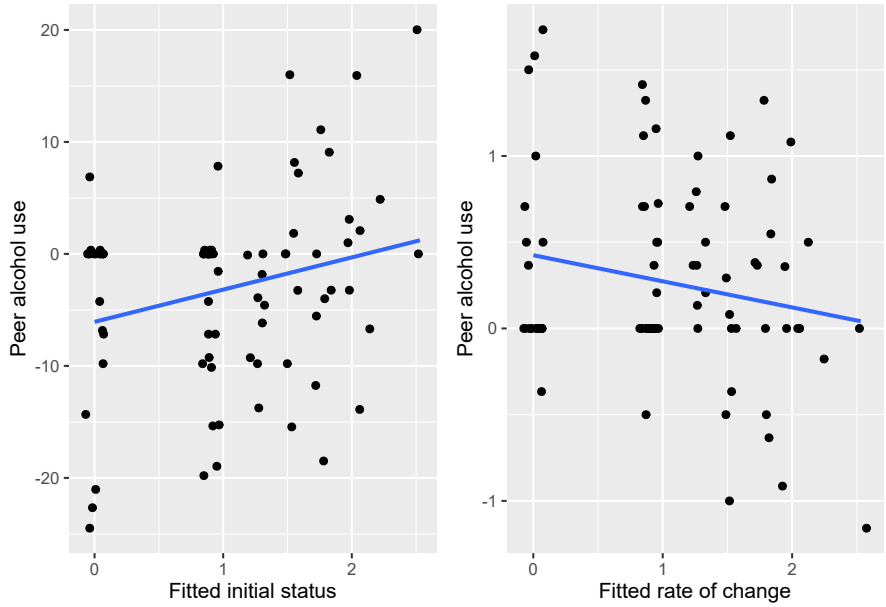

```
int.box <- ggplot(alc.wide) +
  geom_boxplot(aes(x = coa2, y = int)) +
  ylab("Intercepts") + xlab("Child of an alcoholic")
rate.box <- ggplot(alc.wide) +
  geom_boxplot(aes(x = coa2, y = rate)) +
  ylab("Slopes") + xlab("Child of an alcoholic")
lon.box <- grid.arrange(int.box, rate.box, ncol=2)
```




```
int.box <- ggplot(alc.wide) +
  geom_boxplot(aes(x = male2, y = int)) +
  xlab("Male") + ylab("Intercepts")
rate.box <- ggplot(alc.wide) +
  geom_boxplot(aes(x = male2, y = rate)) +
  xlab("Male") + ylab("Slopes")
lon.box <- grid.arrange(int.box, rate.box, ncol=2)
```




```
int.scat <- ggplot(alc.wide, aes(x = peer, y = int)) +
  geom_jitter() +
  geom_smooth(method = lm, se = FALSE) +
  ylab("Peer alcohol use") + xlab("Fitted initial status")
rate.scat <- ggplot(alc.wide, aes(x = peer, y = rate)) +
  geom_jitter() +
  geom_smooth(method = lm, se = FALSE) +
  ylab("Peer alcohol use") + xlab("Fitted rate of change")
lon.scat <- grid.arrange(int.scat, rate.scat, ncol=2)
```



f)

```
regressions <- alcohol %>%
  group_by(id) %>%
  do(fit = lm(alcuse ~ age14, data=.))

lm_info1 <- regressions %>%
  tidy(fit) %>%
  ungroup() %>%
  select(id, term, estimate) %>%
  spread(key = term, value = estimate) %>%
  rename(rate = age14, int = `(Intercept)`)

lm_info2 <- regressions %>%
  tidy(fit) %>%
  ungroup() %>%
  select(id, term, std.error) %>%
  spread(key = term, value = std.error) %>%
  rename(se_rate = age14, se_int = `(Intercept)`)

lm_info <- regressions %>%
  glance(fit) %>%
  ungroup() %>%
  select(id, r.squared, df.residual) %>%
```

```
  inner_join(lm_info1, by = "id") %>%
  inner_join(lm_info2, by = "id") %>%
  mutate(tstar = qt(.975, df.residual),
         intlb = int - tstar * se_int,
         intub = int + tstar * se_int,
         ratelb = rate - tstar * se_rate,
         rateub = rate + tstar * se_rate)
#head(data.frame(lm_info))

alc.wide <- alcohol %>%
  select(-X1, -alcsum, -peersum, -age14) %>%
  spread(key = age, value = alcuse)
alc.wide <- lm_info %>%
  select(id, int, rate , r.squared) %>%
  right_join(alc.wide, by = "id")

int.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=int), bins=8, color="black",
                 fill="white") +
  ggtitle("Histogram of int")
rate.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=rate), bins=8, color="black",
                 fill="white") +
  ggtitle("Histogram of rate")
rsq.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=r.squared), bins=8, color="black",
                 fill="white") +
  ggtitle("Histogram of rsq")
lon.histmat1 <- grid.arrange(int.hist1, rate.hist1,
                             rsq.hist1, ncol=2)

int.box <- ggplot(alc.wide) +
  geom_boxplot(aes(x = coa2, y = int)) +
  ylab("Intercepts") + xlab("Child of an alcoholic")
rate.box <- ggplot(alc.wide) +
  geom_boxplot(aes(x = coa2, y = rate))  +
  ylab("Slopes") + xlab("Child of an alcoholic")
lon.box <- grid.arrange(int.box, rate.box, ncol=2)

int.box <- ggplot(alc.wide) +
  geom_boxplot(aes(x = male2, y = int)) +
  xlab("Male") + ylab("Intercepts")
rate.box <- ggplot(alc.wide) +
  geom_boxplot(aes(x = male2, y = rate)) +
```

```
  xlab("Male") + ylab("Slopes")
lon.box <- grid.arrange(int.box, rate.box, ncol=2)

int.scat <- ggplot(alc.wide, aes(x = peer, y = int)) +
  geom_jitter() +
  geom_smooth(method = lm, se = FALSE) +
  ylab("Peer alcohol use") + xlab("Fitted initial status")
rate.scat <- ggplot(alc.wide, aes(x = peer, y = rate)) +
  geom_jitter() +
  geom_smooth(method = lm, se = FALSE) +
  ylab("Peer alcohol use") + xlab("Fitted rate of change")
lon.scat <- grid.arrange(int.scat, rate.scat, ncol=2)
```

Children of alcoholics have higher levels of alcohol use at age 14 but similar rates of increased usage between 14 and 16. Males have slightly lower alcohol use at age 14 than females but a slightly faster rate of increased usage over the next two years. Alcohol use at age 14 is positively associated with peer use, but subjects whose peers have lower use experience faster growth between 14 and 16.

The linear relationship between time and alcohol use does not change, but the intercept becomes more meaningful – the new intercept reflects alcohol use at age 14 rather than age 0. Most of the old intercepts were negative, which is outside the range of the alcohol use scale.

g) The individual slopes and R-square values remain exactly the same, only the intercepts change. That is, the linear relationship between time and alcohol use does not change, but the intercept becomes more meaningful – the new intercept reflects alcohol use at age 14 rather than age 0. Most of the old intercepts were negative, which is outside the range of the alcohol use scale.

h)

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

$\hat{\rho} = .573 \ / \ (.573 + .562) = .505$
50.5% of the total variability in alcohol use is attributable to differences among subjects.

i)

```
#Model B (Unconditional growth)
model.b <- lmer(alcuse ~ age14 + (age14|id), REML=T, data=alcohol)
summary(model.b)
```

$\hat{\alpha}_0 = 0.65$, $\hat{\beta}_0 = 0.27$
$PseudoR^2_{L1} = (.562 - .337)/.562 = .400$

40.0% of the within-person variability in alcohol use can be explained by linear trends over time.

0.65 = estimated mean alcohol use scale score for 14-year-olds

0.27 = subjects have an estimated mean yearly increase in alcohol use scale score of 0.27 points

j)

```
#Model C (Add coa and peer at Level Two)
model.c <- lmer(alcuse ~ coa + peer + age14 + coa:age14 +
                  peer:age14 + (age14|id), REML=T, data=alcohol)
summary(model.c)
```

$\hat{\alpha}_0 = -0.32$, $\hat{\alpha}_1 = 0.58$, $\hat{\alpha}_2 = 0.69$, $\hat{\beta}_0 = 0.43$, $\hat{\beta}_1 = -0.014$, $\hat{\beta}_2 = -0.150$

- $\hat{\alpha}_0 = -0.32$ = estimated mean alcohol use scale score for 14-year-olds who are not children of alcoholics and whose peers do not use alcohol
- $\hat{\alpha}_1 = 0.58$ = children of alcoholics have mean alcohol use 0.58 points greater than children of non-alcoholics at age 14, after controlling for peer alcohol use
- $\hat{\alpha}_2 = 0.69$ = each one point increase in peer alcohol use is associated with a mean increase in a 14-year-old's alcohol use scale score of 0.69, after controlling for parental alcoholism
- $\hat{\beta}_0 = 0.43$ = teens who are not children of alcoholics and whose peers do not use alcohol increase their alcohol use by 0.43 points per year, on average, between the ages of 14 and 16
- $\hat{\beta}_1 = -0.014$ = teens who are children of alcoholics have a mean yearly increase in alcohol use that is .014 points smaller than children of non-alcoholics, after controlling for peer alcohol use. For example, teens who are children of alcoholics have a mean yearly increase in alcohol of 0.415

points, if their peers are non-drinkers, compared to a mean yearly increase of 0.429 in teens who are not children of alcoholics.

- $\hat{\beta}_2 = -0.150 =$ for each 1 unit increase on the peer alcohol use scale, teens have a mean yearly increase in alcohol use that is .150 points smaller, after controlling for parent alcoholism. For example, teens whose parents are not alcoholics with a score of 1 on the peer alcohol use scale have a mean yearly increase in alcohol use of 0.28, while teens with a score of 2 on the peer alcohol use scale have a mean yearly increase in alcohol use of 0.13.

k)

```
#Model D (Remove coa:age14)
model.d <- lmer(alcuse ~ coa + peer + age14 + peer:age14 +
                  (age14|id), REML=T, data=alcohol)
summary(model.d)

# Model comparisons using REML=F
model.c <- lmer(alcuse ~ coa + peer + age14 + coa:age14 +
                  peer:age14 + (age14|id), REML=F, data=alcohol)
model.d <- lmer(alcuse ~ coa + peer + age14 + peer:age14 +
                  (age14|id), REML=F, data=alcohol)
anova(model.c,model.d)
```

Model D has 9 parameters to estimate:

- Level One:
$$Y_{ij} = a_i + b_i age14_{ij} + \epsilon_{ij}$$
- Level Two:
$$a_i = \alpha_0 + \alpha_1 coa_i + \alpha_2 peer_i + u_i$$
$$b_i = \beta_0 + \beta_1 peer_i + v_i$$
- Composite model:

$$Y_{ij} = \alpha_0 + \alpha_1 coa_i + \alpha_2 peer_i + \beta_0 age14_{ij} + \beta_1 peer_i age14_{ij} + [\epsilon_{ij} + u_i + v_i age14_{ij}]$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

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

We do not have significant evidence (deviance statistic = 0.0126, p-value=.9105 based on chi-square distribution with 1 df) that Model C is preferable to Model D—it does not pay off to add COA as a predictor of slope at Level Two.

2. **Ambulance diversions**. [Fisher et al., 2019]

```
# Changes made when creating amb3 from raw data:
# - stations is all values less than 90... makes it a
#     more normal distribution
# - admit is less than 25,000 to get rid of influential
#     points that are skewing the distribution
# - admit1 is all values of admit/1000 to change the
#     scale
# - totalvisits1 is all values of totalvisits/1000 to
#     change the scale

amb3 <- read_csv("data/ambulance3.csv") %>%
  mutate(high_stations = ifelse(stations > 23.5, "Yes", "No"))

as_tibble(amb3) %>%
  arrange(id, year2013) %>%
  print(width = Inf)

# quick peak at divert hours variable (response)
ggplot(amb3, aes(diverthours))+
  geom_histogram(bins=30)
ggplot(amb3, aes(logdivert))+
  geom_histogram(bins=30)

# quick peak at admit1 variable
ggplot(amb3, aes(admit1))+
  geom_histogram(bins=30)

# quick peak at stations variable (level 2)
ggplot(amb3, aes(stations))+
  geom_histogram(bins=30)

# quick peak at total visits variable (level 1)
ggplot(amb3, aes(totalvisits))+
  geom_histogram(bins=30)

# EDA
```

```
amb3_eda <- filter(amb3, id <= 106190110 )
ggplot(amb3_eda, aes(x = year2013, y = diverthours)) +
  geom_point() +
  geom_line() +
  facet_wrap(~as.factor(id), ncol = 6) +
  scale_x_continuous(limits=c(0,2), breaks=c(0,1,2)) +
  theme(strip.text.x=element_blank()) +
  labs(x = "Years since 2013", y = "Diversion hours")

ggplot(amb3_eda, aes(x = totalvisits1, y = diverthours)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Total visits / 1000", y = "Diversion hours")

ggplot(amb3, aes(x = year2013, y = diverthours)) +
  geom_line(aes(group = as.factor(id)), color = "dark grey") +
  facet_grid(~ems_basic) +
  geom_smooth(aes(group = 1), color = "black", size = 1) +
  labs(x = "Years since 2013", y = "Diversion hours")

ggplot(amb3, aes(x = year2013, y = diverthours)) +
  geom_line(aes(group = as.factor(id)), color = "dark grey") +
  facet_grid(~high_stations) +
  geom_smooth(aes(group = 1), color = "black", size = 1) +
  labs(x = "Years since 2013", y = "Diversion hours")


regressions <- amb3 %>%
  group_by(id) %>%
  do(fit = lm(diverthours ~ year2013, data=.))

sd_filter <- amb3 %>%
  group_by(id) %>%
  summarise(sds = sd(diverthours))

regressions <- regressions %>%
  right_join(sd_filter, by="id") %>%
  filter(!is.na(sds))

lm_info1 <- regressions %>%
  tidy(fit) %>%
  ungroup() %>%
  select(id, term, estimate) %>%
```

```
  spread(key = term, value = estimate) %>%
  rename(rate = year2013, int = `(Intercept)`)

lm_info2 <- regressions %>%
  tidy(fit) %>%
  ungroup() %>%
  select(id, term, std.error) %>%
  spread(key = term, value = std.error) %>%
  rename(se_rate = year2013, se_int = `(Intercept)`)

lm_info <- regressions %>%
  glance(fit) %>%
  ungroup() %>%
  select(id, r.squared, df.residual) %>%
  inner_join(lm_info1, by = "id") %>%
  inner_join(lm_info2, by = "id") %>%
  mutate(tstar = qt(.975, df.residual),
         intlb = int - tstar * se_int,
         intub = int + tstar * se_int,
         ratelb = rate - tstar * se_rate,
         rateub = rate + tstar * se_rate)
head(data.frame(lm_info))

# summary stats for intercepts
summary(lm_info$int)
sd(lm_info$int)

# summary stats for fitted rate of change
summary(lm_info$rate)
sd(lm_info$rate,na.rm=T)

# summary stats for R sq
summary(lm_info$r.squared)

# histograms for ints, rates of change, and Rsq values
int.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=int), binwidth=4, color="black",
                 fill="white") +
  labs(x="Intercepts", y="Frequency", title="(a)")
rate.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=rate), binwidth=2, color="black",
                 fill="white") +
  labs(x="Slopes", y="Frequency", title="(b)")
rsq.hist1 <- ggplot(lm_info) +
```

```
  geom_histogram(aes(x=r.squared), binwidth=0.2,
                 color="black", fill="white") +
  labs(x="Rsquared values", y="Frequency", title="(c)")
lon.histmat1 <- grid.arrange(int.hist1, rate.hist1,
                             rsq.hist1, ncol=2)

# correlation between slopes and intercepts for subjects
#   with slope
with(lm_info, cor(int, rate, use="complete.obs"))
ggplot(data = lm_info, aes(x = int, y = rate)) +
  geom_point(color = "dark grey") +
  geom_smooth(se=FALSE, method="lm", color="black") +
  xlab("Fitted Intercepts") + ylab("Fitted Slopes")


amb3_wide <- amb3 %>%
  group_by(id) %>%
  filter(row_number() == 1)

# Boxplots to evaluate L2 variables
amb3_wide <- lm_info %>%
  select(id, int, rate , r.squared) %>%
  right_join(amb3_wide, by = "id")

int.box1 <- ggplot(amb3_wide) +
  geom_boxplot(aes(x = factor(ems_basic), y = int)) +
  coord_flip() +
  labs(x = "EMS basic", y="Fitted Intercepts", title="(a)")
rate.box1 <- ggplot(amb3_wide) +
  geom_boxplot(aes(x = factor(ems_basic), y = rate)) +
  coord_flip() +
  labs(x = "EMS basic", y="Fitted Slopes", title="(b)")
lon.box2 <- grid.arrange(int.box1, rate.box1, nrow=2)

int.box1 <- ggplot(amb3_wide) +
  geom_boxplot(aes(x = high_stations, y = int)) +
  coord_flip() +
  labs(x = "Stations", y="Fitted Intercepts", title="(a)")
rate.box1 <- ggplot(amb3_wide) +
  geom_boxplot(aes(x = high_stations, y = rate)) +
  coord_flip() +
  labs(x = "Stations", y="Fitted Slopes", title="(b)")
lon.box2 <- grid.arrange(int.box1, rate.box1, nrow=2)
```

```
##Correlation structure
hgtm.lm <- lm(diverthours ~ year2013, data = amb3)
amb3 <- amb3 %>%
  mutate(lmres = resid(hgtm.lm))

hgtwm <- amb3 %>%
  select(id, lmres, year2013) %>%
  mutate(name = rep("lmres", n())) %>%
  unite(newcol, name, year2013, sep = ".") %>%
  spread(key = newcol, value = lmres)

ggpairs(hgtwm[,c(2:4)], upper = list(),
        lower = list(continuous = "smooth"),
  diag = list(continuous = "bar", discrete = "bar"),
  axisLabels = "show")

# Two Level without logging

# Model A - unconditional means
mod.A = lmer(diverthours ~ 1 + (1|id), REML=T, data=amb3)
summary(mod.A)

# Model B - unconditional growth
mod.B = lmer(diverthours ~ year2013 + (year2013|id),
  REML=T, data=amb3)
summary(mod.B)

amb3 <- mutate(amb3, year2015 = year2013 - 2)
mod.B1 = lmer(diverthours ~ year2015 + (year2015|id),
  REML=T, data=amb3)
summary(mod.B1)  # corr changes

mod.C = lmer(diverthours ~ year2013 + totalvisits1 + (1|id),
  REML=T, data=amb3)
summary(mod.C)

mod.D = lmer(diverthours ~ year2013 + ems_basic +
  (year2013|id), REML=T, data=amb3)
summary(mod.D)

mod.D0 = lmer(diverthours ~ year2013 + ems_basic + (1|id),
  REML=T, data=amb3)
summary(mod.D0)
```

```
anova(mod.D, mod.D0, test = "Chisq")

mod.E = lmer(diverthours ~ year2013 + ems_basic +
  ems_basic:year2013 + (year2013|id), REML=T, data=amb3)
summary(mod.E)

mod.F = lmer(diverthours ~ year2013 + totalvisits1 +
  totalvisits1:year2013 + ems_basic + ems_basic:year2013 +
  (year2013|id), REML=T, data=amb3)
summary(mod.F)

mod.G = lmer(diverthours ~ year2013 + totalvisits1 +
  ems_basic + stations + ems_basic:year2013 +
  stations:year2013 + (year2013|id), REML=T, data=amb3)
summary(mod.G)

# Parametric bootstrap code for lme4-models
#   from Fabian Scheipl on stack exchange

#m0 is the lmer model under the null hypothesis (smaller model)
#mA is the lmer model under the alternative

bootstrapAnova <- function(mA, m0, B=1000){
  oneBootstrap <- function(m0, mA){
    d <- drop(simulate(m0))
    m2 <-refit(mA, newresp=d)
    m1 <-refit(m0, newresp=d)
    return(anova(m2,m1)$Chisq[2])
  }
  nulldist <- replicate(B, oneBootstrap(m0, mA))
  ret <- anova(mA, m0)
  ret$"Pr(>Chisq)"[2] <- mean(ret$Chisq[2] < nulldist)
  names(ret)[8] <- "Pr_boot(>Chisq)"
  attr(ret, "heading") <- c(attr(ret, "heading")[1],
    paste("Parametric bootstrap with", B,"samples."),
    attr(ret, "heading")[-1])
  attr(ret, "nulldist") <- nulldist
  return(ret)
}

#use like this (increase B for stronger results):
# bRLRT <- bootstrapAnova(mA=<BIG MODEL>, m0=<SMALLER MODEL>)

# also: will likely run faster to save models under ML
```

```
#   instead of REML, since anova() will refit models
#   under ML before calculating LRT statistic

bRLRT = bootstrapAnova(mA = mod.D, m0 = mod.D0, B=100)
bRLRT
nullLRT = attr(bRLRT,"nulldist")
x=seq(0,max(nullLRT),length=100)
y=dchisq(x,2)
nullLRT.1 <- as.data.frame(cbind(nullLRT=nullLRT,x=x,y=y))
ggplot(nullLRT.1) +
  geom_histogram(aes(x=nullLRT,y=..density..),binwidth=1,
                 color="black",fill="white") +
  geom_vline(xintercept=44.2,size=1) +
  geom_line(aes(x=x,y=y)) +
  labs(x="Likelihood Ratio Test Statistics from Null Dist",
       y="Density")
sum(nullLRT>=44.2)/100
# parametric bootstrap p-value - where you must fill in
#   LRT from original data and number of bootstraps

# Confidence intervals for model parameters
#  - can reduce nsim to run faster
confint(mod.D, method = c("boot"), boot.type = c("perc"))
confint(mod.D, method = c("Wald"))
confint(mod.D, method = c("profile"))
```

a) Observational units = year for specific hospital (L1); hospital (L2). Explanatory variables = year2013, totalvisits1 (L1); ems_basic, stations (L2).

b) In order of promise: `ems_basic`, `ems_basic:year2013`, `station:year2013`, `station`. The spaghetti plot shows generally higher numbers of diversion hours when ems_basic is 0 (can handle higher severity levels) and a higher slope for ems_basic of 0 (especially in the first year) and for higher stations (above the median).

c) An unconditional growth model:

- Level One:
$$Y_{ij} = a_i + b_i time_{ij} + \epsilon_{ij}$$
- Level Two:
$$a_i = \alpha_0 + u_i$$
$$b_i = \beta_0 + v_i$$
- Composite model:

$$Y_{ij} = \alpha_0 + \beta_0 time_{ij} + [\epsilon_{ij} + u_i + v_i time_{ij}]$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

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

a_i is the estimated diversion hours for hospital i in 2013.

v_i is how the yearly change in diversion hours for hospital i differs from the average yearly change across all hospitals (β_0).

d)

$\hat{\beta}_1 = -266.1$. This is the difference in yearly increases in diversion hours between hospitals with basic services or more advanced. While we estimate an increase of 392.7 diversion hours per year for hospitals that can handle more severe cases, we only expect an increase of 126.6 diversion hours per year for hospitals with basic services.

The true null distribution of the t-statistic is not known since exact degrees of freedom are unknown with correlated data.

Since confidence intervals using bootstrapping and the Wald method for β_1 both contain 0, we can conclude that there is *not* statistically significant evidence that yearly changes in diversion hours differ by EMS level.

e) We must estimate 11 parameters in Model G: 7 fixed effects and 4 variance components.

- Level One:
$$Y_{ij} = a_i + b_i year2013_{ij} + c_i totalvisits1_{ij} + \epsilon_{ij}$$

- Level Two:
$$\begin{aligned} a_i &= \alpha_0 + \alpha_1 emsbasic_i + \alpha_2 stations_i + u_i \\ b_i &= \beta_0 + \beta_1 emsbasic_i + \beta_2 stations_i + v_i \\ c_i &= \gamma_0 \end{aligned}$$

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

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

f) $\hat{\beta}_0 = 275.5$. The number of diversion hours increases by 275.5 per year for hospitals where the EMS level is beyond basic and there are no stations, after controlling for total emergency visits.

$\hat{\alpha}_2 = -2.8$. In 2013, each additional station was associated with a 2.8 hour decline in diversion hours, after accounting for total visits and EMS level.

g)

- $H_O : \sigma_v^2 = 0$ and $\rho_{uv} = 0$.
- We have statistically significant evidence (LRT = 44.157, p<.001) that the full model (D) performs better—that it's beneficial to allow hospital-to-hospital variability in yearly changes and correlation among errors.
- The above statement is confirmed by a parametric bootstrap which showed the p-value is below 1 in 100.
- Code is shown above.
- The likelihood ratio test typically does not perform well for variance components, especially when testing at boundary values.
- To produce a bootstrapped value for Y_{11} under the null model:

$a_1 = \alpha_0 + \alpha_1 emsbasic_1 + u_1 = 1184.7 - 881.6 * 1 + u_1$ where u_1 is drawn from a normal distribution with mean 0 and SD 601.5. $b_1 = \beta_0 = 146.0$. Then $Y_{11} = a_1 + b_1 year2013_{11} + \epsilon_{11} = a_1 + b_1 * 0 + \epsilon_{11}$ where ϵ_{11} is drawn from a normal distribution with mean 0 and SD 433.3.

So let's say $u_1 = 1000$ and $\epsilon_{11} = -200$. Then $a_1 = 1184.7 - 881.6 + 1000 = 1303.1$ and $Y_{11} = 1303.1 - 200 = 1103.1$.

9.1.3 Open-Ended Exercises

1. **UCLA nurse blood pressure study.** Article: [Goldstein and Shapiro, 2000]; data source: [Weiss, 2005].

```
nurse <- read_csv("data/nursebp.csv")

head(nurse)
summary(as.numeric(table(nurse$SNUM)))
table(nurse$POSTURE)
selected <- nurse %>%
  dplyr::select(SYS, DIA,HRT, MNACT5, STR, HAP, TIR, time)
cor(selected, use="pairwise.complete.obs")

nursecomplete <- na.omit(nurse)
```

```
means <- nursecomplete %>%
  group_by(SNUM) %>%
  summarise(meanSYS = mean(SYS),
            meanMNACT5 = mean(MNACT5),
            meanHAP = mean(HAP),
            meanSTR = mean(STR),
            meanTIR = mean(TIR))

sys <- ggplot(nurse, aes(x = SYS)) +
  geom_histogram(bins = 65, color="black", fill="white") +
  xlim(80, 200) + ylab("Frequency") +
  xlab("Systolic Blood Pressure")
meansys <- ggplot(means, aes(x = meanSYS)) +
  geom_histogram(bins = 65, color="black", fill="white") +
  xlim(80, 200) + ylab("Frequency") +
  xlab("Mean Systolic Blood Pressure")
grid.arrange(sys, meansys, ncol=1)

meanMNACT5  <- ggplot(means, aes(x = meanMNACT5)) +
  geom_histogram(bins = 30, color="black", fill="white")  +
  ylab("Frequency") + xlab("meanMNACT5")
meanSTR  <- ggplot(means, aes(x = meanSTR)) +
  geom_histogram(bins = 30, color="black", fill="white")  +
  ylab("Frequency") + xlab("meanSTR")
meanHAP  <- ggplot(means, aes(x = meanHAP)) +
  geom_histogram(bins = 30, color="black", fill="white")  +
  ylab("Frequency") + xlab("meanHAP")
meanTIR  <- ggplot(means, aes(x = meanTIR)) +
  geom_histogram(bins = 30, color="black", fill="white")  +
  ylab("Frequency") + xlab("meanTIR")
grid.arrange(meanMNACT5, meanSTR, meanHAP, meanTIR, ncol = 2)

ggplot(data=nurse,aes(factor(POSTURE),SYS)) +
  geom_boxplot() +
  coord_flip()

spag1 <- ggplot(nurse, aes(x = time, y = SYS)) +
  geom_line(aes(group = SNUM), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess)
spag2 <- ggplot(nurse, aes(x = MNACT5, y = SYS)) +
  geom_line(aes(group = SNUM), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess)
```

```
spag3 <- ggplot(nurse, aes(x = STR, y = SYS)) +
  geom_line(aes(group = SNUM), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess)
spag4 <- ggplot(nurse, aes(x = HAP, y = SYS)) +
  geom_line(aes(group = SNUM), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess)
spag5 <- ggplot(nurse, aes(x = TIR, y = SYS)) +
  geom_line(aes(group = SNUM), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess)
grid.arrange(spag1, spag2, spag3, spag4, spag5, ncol = 3)

means1 <- nurse %>%
  select(DAY, FH123, PHASE, SYS, SNUM) %>%
  group_by(SNUM) %>%
  mutate(meanSYS = mean(SYS))

box1 <- ggplot(data=means1,aes(factor(DAY),meanSYS)) +
  geom_boxplot() +
  coord_flip()
box2 <- ggplot(data=means1,aes(factor(FH123),meanSYS)) +
  geom_boxplot() +
  coord_flip()
box3 <- ggplot(data=means1,aes(factor(PHASE),meanSYS)) +
  geom_boxplot() +
  coord_flip()
grid.arrange(box1, box2, box3, ncol = 2)

spag6 <- ggplot(nurse, aes(x = time, y = SYS)) +
  geom_line(aes(group = SNUM), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess) + facet_wrap(~DAY)
spag7 <- ggplot(nurse, aes(x = time, y = SYS)) +
  geom_line(aes(group = SNUM), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess) + facet_wrap(~FH123)
spag8 <- ggplot(nurse, aes(x = time, y = SYS)) +
  geom_line(aes(group = SNUM), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess) + facet_wrap(~PHASE)
grid.arrange(spag6, spag7, spag8, ncol = 2)
```

```
#Model A
model.a <- lmer(SYS~1 + (1|SNUM), REML=T, data=nurse)
summary(model.a)

#Model B
model.b <- lmer(SYS~time + (time|SNUM), data=nurse)
summary(model.b)

#Model C
model.c <- lmer(SYS~timepass + (1|SNUM), data=nurse)
summary(model.c)

# Moving toward a final model
model.z <- lmer(SYS~timepass + STR + TIR + MNACT5 + DAY +
  FH123 + (1 |SNUM), data=nurse)
summary(model.z)
```

With 203 nurses having between 28 and 60 assessments (median 48) over the course of a day, we have a rich data set for investigating factors that may impact systolic blood pressure. Missing data is not a big issue, either, with complete data for 7896 of the 9573 observations – only ratings of emotion and activity were missing in any noticeable amount.

Initial lattice plots and spaghetti plots show substantial variability in systolic blood pressure throughout the day for the same individual and between individuals at the beginning of the day. There appears to be a tendency for blood pressure to fall slightly throughout the day. Histograms show that self-ratings of stress levels and tiredness tend to be low (1 or 2), with relatively few ratings of 3, 4, or 5. Nevertheless, lattice plots and spaghetti plots show that blood pressure levels are a bit higher when individuals are feeling stressed and a bit lower when they are tired (and higher when the nurse has been more active in the past five minutes). Thus, among Level One covariates, time of the day, stress and tiredness self-ratings, and recent activity all appear important to consider in final models.

Among Level Two covariates, initial boxplots show that whether it's a work day and whether the nurse has two parents with history of hypertension both are associated with increased systolic blood pressure. Spaghetti plots of blood pressure over time, separated by categories of Level Two covariates, show little effect on trends over time, other than possibly the two parents with hypertension group.

Initial multilevel models show that 29% of total variability in systolic blood pressure is based on differences between nurses, while 71% is based on differences over the course of the day for the same nurse. An unconditional growth

model shows that blood pressure significantly falls over the course of a day (an estimated 0.116 points per hour, t = -3.8).

In our final multilevel model, however, this negative trend over time was no longer significant after controlling for stress, tiredness, activity level, work day, and family history of hypertension. Holding all else constant, blood pressure is higher when nurses feel more stressed (t = 3.00) and when nurses have been more active (t = 18.16), but lower when nurses feel more tired (t = -2.99). Since over the course of the day nurses become less active and more tired, accounting for these covariates explains much of the negative trend over time. After adjusting for emotional state and recent activity at the time that blood pressure is taken, average blood pressure is 2.51 points higher on work days (t = 2.07) and 7.15 points higher for nurses with two parents having histories of hypertension.

2. **Completion rates at U.S. colleges.** [National Center for Education Statistics, 2018]

```
colleges <- read_csv("data/colleges.csv")

table(as.numeric(table(colleges$id)))
cor(data.frame(colleges$rate, colleges$year,
               colleges$instpct, colleges$instamt),
    use = "pairwise.complete.obs")

aid.lev2 <- colleges %>%
  group_by(id) %>%
  summarise(meanrate = mean(rate),
            meaninstpct = mean(instpct),
            meaninstamt = mean(instamt))
cor(data.frame(aid.lev2$meanrate,aid.lev2$meaninstpct,
  aid.lev2$meaninstamt),use="pairwise.complete.obs")

rate <- ggplot(colleges, aes(x = rate)) +
  geom_histogram(bins = 40, color="black", fill="white") +
  ylab("Frequency")
instpct <- ggplot(colleges, aes(x = instpct)) +
  geom_histogram(bins = 40, color="black", fill="white") +
  ylab("Frequency")
instamt <- ggplot(colleges, aes(x = instamt)) +
  geom_histogram(bins = 40, color="black", fill="white") +
  ylab("Frequency")
grid.arrange(rate, instpct, instamt, ncol = 2)

meanrate <- ggplot(aid.lev2, aes(x = meanrate)) +
```

```
  geom_histogram(bins = 65, color="black", fill="white") +
  ylab("Frequency")
meaninstpct <- ggplot(aid.lev2, aes(x = meaninstpct)) +
  geom_histogram(bins = 65, color="black", fill="white") +
  ylab("Frequency")
meaninstamt <- ggplot(aid.lev2, aes(x = meaninstamt)) +
  geom_histogram(bins = 65, color="black", fill="white") +
  ylab("Frequency")
grid.arrange(meanrate, meaninstpct, meaninstamt, ncol=2)

faculty <- ggplot(colleges, aes(x = faculty)) +
  geom_histogram(bins = 40, color="black", fill="white") +
  ylab("Frequency")
tuition <- ggplot(colleges, aes(x = tuition)) +
  geom_histogram(bins = 40, color="black", fill="white") +
  ylab("Frequency")
grid.arrange(faculty, tuition, ncol = 2)

spag9 <- ggplot(colleges, aes(x = year, y = rate)) +
  geom_line(aes(group = id), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess)
spag10 <- ggplot(colleges, aes(x = year, y = instpct)) +
  geom_line(aes(group = id), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess)
spag11 <- ggplot(colleges, aes(x = year, y = instamt)) +
  geom_line(aes(group = id), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess)
grid.arrange(spag9, spag10, spag11, ncol = 2)

rate <- colleges %>%
  group_by(id) %>%
  mutate(meanrate = mean(rate))
fac <- ggplot(rate, aes(faculty, meanrate)) +
  geom_point() +
  geom_smooth(method = "lm")
tui <- ggplot(rate, aes(tuition, meanrate)) +
  geom_point() +
  geom_smooth(method = "lm")
grid.arrange(fac, tui, ncol = 2)

# remove 3 schools with rates above 100
```

```
#   and median center others (except year at first year)
`%!in%` = Negate(`%in%`)
colleges <- colleges %>%
  filter(id %!in% c(198109,146755,233912)) %>%
  mutate(year.c = year - 2002,
         instamt.c = instamt - 5.88,
         instpct.c = instpct - 74.0,
         tuition.c = tuition - 15.11,
         faculty.c = faculty - 5.36)

model.a1 <- lmer(rate~ 1 + (1|id), data=colleges)
summary(model.a1)

model.b1 <- lmer(rate ~ year.c + (year.c|id), data=colleges)
summary(model.b1)

model.g1 <- lmer(rate~ year.c + instamt.c + instpct.c +
  faculty.c + year.c:instamt.c +
  (year.c + instamt.c + instpct.c + year.c:instamt.c|id),
  data=colleges)
summary(model.g1)
```

Initial histograms show that our primary response variable, completion rate, is reasonably normal with just a few high outliers (we removed observations from 3 institutions which indicated over 100 degrees per 100 students enrolled). In addition, `instpct` has a strong left skew, with the bulk of observations near 100, while `instamt` has a strong right skew. Finally, `tuition` appears bimodal—possibly public and private institutions.

Initial multilevel models show that 85% of total variability in completion rates occurs at Level Two (between colleges), while only 15% is based on changes between 2002 and 2008 for the same college. An unconditional growth model shows that completion rates are significantly rising over time (an estimated 0.13 degrees per 100 students per year, t = 5.17).

In our final multilevel model, however, this positive trend over time was no longer significant after controlling for percentage and amount of institutional grants and faculty-student ratio. With a fixed percentage of students receiving an institutional grant, an additional $1000 in the typical grant size in 2002 was associated with an increase in completion rate of 0.22 degrees per 100 students (t = 6.17 after adjusting for faculty-student ratio too). However, the impact of grant size is significantly decreasing over time (t = -2.77 for the year-by-grant size interaction); in 2008, for instance, the same $1000 increase in typical grant size was associated with an increase in completion rate of 0.14 degrees per 100 students. With a fixed typical grant amount, an increase of

4% in the percentage of students receiving institutional grants is associated with an increase in completion rate of 0.10 degrees per 100 students (t = 6.12 after adjusting for faculty-student ratio too); this effect is constant from year-to-year. Finally, there is some evidence (t = 1.68) that colleges with higher faculty-student ratios have higher completion rates, after adjusting for institutional grant activity and time trends. This particular model centered year on 2002 and centered continuous covariates on their median in order to provide more stable and interpretable parameter estimates.

Future analyses might target sources of variability between colleges, such as size, quality, or public/private status. Tuition is related to some of these factors, but it did not prove to be significantly associated with completion rates after adjusting for institutional grant activity and faculty-student ratio. One reason for this might be the high correlation between tuition and mean grant size at the college level. Nevertheless, analysis of this data has the potential for providing insights to colleges about the best way to spend their institutional grant money—provide more money per student or provide money to more students.

3. **Beating the Blues.** Article: [2003]; data source: [Everitt and Hothorn, 2006].

```
# Read in data in wide format (one row per person)
BtheB <- read_csv("data/BtheB.csv")

#Alternative way to read in data after installing HSAUR package
# data("BtheB", package = "HSAUR")

# Investigate patterns of missingness
md.pattern(BtheB)

# Remove 3 subjects with no post-baseline data
BtheB <- BtheB %>%
  filter(is.na(bdi.2m) + is.na(bdi.4m) + is.na(bdi.6m) +
           is.na(bdi.8m) < 4) %>%
  rowid_to_column("subject")

# Create data frame in LONG form (one obs per subject-visit)
blues <- BtheB %>%
  pivot_longer(cols = bdi.2m:bdi.8m,
               names_to = "time",
               values_to = "bdi") %>%
  mutate(time = 10 * parse_number(time))
```

```
# Exploratory data analyses #

#Quick look at covariates - individually and relationship with
#  response.  We only have Level Two covariates in this study.
table(BtheB$length)
table(BtheB$drug)
table(BtheB$treatment)

ggplot(data = BtheB, aes(x = bdi.pre)) +
  geom_histogram(bins = 15) +
  xlab("Baseline BDI")

ggplot(data = blues, aes(x = length, y = bdi)) +
  geom_boxplot() +
  coord_flip() +
  xlab("Length of current episode") + ylab("BDI")
ggplot(data = blues, aes(x = drug, y = bdi)) +
  geom_boxplot() +
  coord_flip() +
  xlab("Concomitant drug use") + ylab("BDI")
ggplot(data = blues, aes(x = treatment, y = bdi)) +
  geom_boxplot() +
  coord_flip() +
  xlab("Treatment group") + ylab("BDI")

ggplot(data = blues, aes(x = bdi.pre, y = bdi)) +
  geom_point() +
  geom_smooth() +
  xlab("Baseline BDI") + ylab("BDI")
ggplot(data = blues, aes(x = time, y = bdi)) +
  geom_point() +
  geom_smooth() +
  xlab("Months since treatment ended") + ylab("BDI")

# Plot time trend by individual - just connect points
ggplot(blues, aes(x = time, y = bdi)) +
  geom_point() +
  geom_line() +
  facet_wrap(~as.factor(subject), ncol = 6) +
  theme(strip.text.x=element_blank()) +
  labs(x = "Months since treatment ended", y = "BDI")
```

```
# spaghetti plot
ggplot(blues, aes(x = time, y = bdi)) +
  geom_line(aes(group = subject), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess)

# spaghetti plot by treatment
ggplot(blues, aes(x = time, y = bdi)) +
  geom_line(aes(group = subject), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess) +
  facet_wrap(~ treatment)

# spaghetti plot by baseline BDI with loess overall trend
ggplot(blues, aes(x = time, y = bdi)) +
  geom_line(aes(group = subject), color = "dark grey") +
  geom_smooth(aes(group = 1), color = "black", size = 1,
              method = loess) +
  facet_wrap(~ cut_number(bdi.pre, 4), ncol = 4)

# Find slope and intercept of all subjects
regressions <- blues %>%
  group_by(subject) %>%
  do(fit = lm(bdi ~ time, data=.))

sd_filter <- blues %>%
  group_by(subject) %>%
  summarise(sds = sd(bdi, na.rm = TRUE))

regressions <- regressions %>%
  right_join(sd_filter, by="subject") %>%
  filter(!is.na(sds))

lm_info1 <- regressions %>%
  tidy(fit) %>%
  ungroup() %>%
  select(subject, term, estimate) %>%
  spread(key = term, value = estimate) %>%
  rename(rate = time, int = `(Intercept)`)

lm_info2 <- regressions %>%
  tidy(fit) %>%
  ungroup() %>%
  select(subject, term, std.error) %>%
```

```
  spread(key = term, value = std.error) %>%
  rename(se_rate = time, se_int = `(Intercept)`)

lm_info <- regressions %>%
  glance(fit) %>%
  ungroup() %>%
  select(subject, r.squared, df.residual) %>%
  inner_join(lm_info1, by = "subject") %>%
  inner_join(lm_info2, by = "subject") %>%
  mutate(tstar = qt(.975, df.residual),
         intlb = int - tstar * se_int,
         intub = int + tstar * se_int,
         ratelb = rate - tstar * se_rate,
         rateub = rate + tstar * se_rate)
head(data.frame(lm_info))

# summary stats for intercepts
summary(lm_info$int)
sd(lm_info$int)

# summary stats for fitted rate of change
summary(lm_info$rate)
sd(lm_info$rate,na.rm=T)

# summary stats for R sq
summary(lm_info$r.squared)

# histograms for ints, rates of change, and Rsq values
int.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=int), binwidth=4, color="black",
                 fill="white") +
  labs(x="Intercepts", y="Frequency", title="(a)")

rate.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=rate), binwidth=2, color="black",
                 fill="white") +
  labs(x="Slopes", y="Frequency", title="(b)")

rsq.hist1 <- ggplot(lm_info) +
  geom_histogram(aes(x=r.squared), binwidth=0.2, color="black",
                 fill="white") +
  labs(x="Rsquared values", y="Frequency", title="(c)")

grid.arrange(int.hist1, rate.hist1, rsq.hist1, ncol=2)
```

```
# correlation between slopes and intercepts for
#   subjects with slope
with(lm_info, cor(int, rate, use="complete.obs"))

BtheB <- lm_info %>%
  select(subject, int, rate , r.squared) %>%
  right_join(BtheB, by = "subject")

# Boxplots to compare treatment groups
int.box1 <- ggplot(BtheB) +
  geom_boxplot(aes(x = treatment, y = int)) +
  coord_flip() +
  labs(x="Treatment", y="Fitted Intercepts", title="(a)")
rate.box1 <- ggplot(BtheB) +
  geom_boxplot(aes(x = treatment, y = rate)) +
  coord_flip() +
  labs(x="Treatment", y="Fitted Slopes", title="(b)")
grid.arrange(int.box1, rate.box1, nrow=2)

# Boxplots to compare current episode length
int.box1 <- ggplot(BtheB) +
  geom_boxplot(aes(x = length, y = int)) +
  coord_flip() +
  labs(x="Length of Current Episode", y="Fitted Intercepts",
       title="(a)")
rate.box1 <- ggplot(BtheB) +
  geom_boxplot(aes(x = length, y = rate)) +
  coord_flip() +
  labs(x="Length of Current Episode", y="Fitted Slopes",
       title="(b)")
grid.arrange(int.box1, rate.box1, nrow=2)

# Boxplots to compare concomitant drug
int.box1 <- ggplot(BtheB) +
  geom_boxplot(aes(x = drug, y = int)) +
  coord_flip() +
  labs(x="Concomitant Drug", y="Fitted Intercepts",
       title="(a)")
rate.box1 <- ggplot(BtheB) +
  geom_boxplot(aes(x = drug, y = rate)) +
  coord_flip() +
  labs(x="Concomitant Drug", y="Fitted Slopes", title="(b)")
grid.arrange(int.box1, rate.box1, nrow=2)
```

```
# Scatterplots of baseline BDI vs intercept and slope
int.scat1 <- ggplot(BtheB, aes(x=bdi.pre, y=int)) +
  geom_point() +
  geom_smooth(method="lm",color="black",size=.75) +
  labs(x="Baseline BDI",
       y="Fitted Intercepts",title="(a)")
rate.scat1 <- ggplot(BtheB, aes(x=bdi.pre, y=rate)) +
  geom_point() +
  geom_smooth(method="lm",color="black",size=.75) +
  labs(x="Baseline BDI",
       y="Fitted Slopes",title="(b)")
grid.arrange(int.scat1, rate.scat1, ncol=2)

# Investigate current episode by treatment interaction
int.box <- ggplot(BtheB) +
  geom_boxplot(aes(x=treatment, y=int)) +
  coord_flip() +
  labs(x="Treatment",y="Fitted Intercepts",title="(a)") +
  facet_grid(length~.)
rate.box <- ggplot(BtheB) +
  geom_boxplot(aes(x=treatment, y=rate)) +
  coord_flip() +
  labs(x="Treatment",y="Fitted Slopes",title="(b)") +
  facet_grid(length~.)
grid.arrange(int.box,rate.box,ncol=2)

# Examine correlation structure
score.nonm <- blues %>%
  filter(is.na(bdi) == FALSE)
hgtm.lm <- lm(bdi~time, data=score.nonm)
score.nonm <- score.nonm %>%
  mutate(lmres = resid(hgtm.lm))

hgtwm <- score.nonm %>%
  select(subject, lmres, time) %>%
  mutate(name = rep("lmres", n())) %>%
  unite(newcol, name, time, sep = ".") %>%
  spread(key = newcol, value = lmres)

ggpairs(hgtwm[,c(2:5)],upper=list(),
        lower=list(continuous="smooth"),
  diag=list(continuous="bar", discrete="bar"),
  axisLabels="show")
```

```
#########################################################


#Model A (Unconditional means model)
model.a <- lmer(bdi~ 1 + (1|subject), data=blues)
summary(model.a)

#Model B (Unconditional growth)
model.b <- lmer(bdi~ time + (time|subject), data=blues)
summary(model.b)

blues <- blues %>%
  mutate(btheb = ifelse(treatment=="BtheB",1,0),
         cbdi.pre = bdi.pre - mean(bdi.pre),  # mean is 23.15
         longepi = ifelse(length==">6m",1,0),
         druguse = ifelse(drug=="Yes",1,0))

#Model C (Uncontrolled effects of treatment on intercept
#  and slope)
model.c <- lmer(bdi~ btheb*time + (time|subject), data=blues)
summary(model.c)

#Model D (Centered baseline BDI)
model.d <- lmer(bdi~ btheb*time + cbdi.pre*time +
                (time|subject), data=blues)
summary(model.d)

# Model F (all 4 covariates on intercept; just trt on slope)
#   - Final Model?
model.f0 <- lmer(bdi~ btheb*time + cbdi.pre + druguse +
                longepi + (time|subject), data=blues)
summary(model.f0)
AIC(model.f0)

model.f <- lmer(bdi~ time + btheb + cbdi.pre + druguse +
  longepi + btheb:longepi + btheb:time + (time|subject),
  data=blues)
summary(model.f)
AIC(model.f)
```

Among the 100 subjects, 3 had no post-baseline data and were removed; another 24 had no BDI data after 2 months, so we can use them to help evaluate

factors that affect the intercept (end of active treatment) but not factors that affect BDI once active treatment ends. 52 subjects had complete data at all 4 time points.

Exploratory data analysis shows that BDI levels continue to drop throughout the study, from an average near 20 at two months to an average near 10 at 8 months (but we also drop from 97 to 52 patients over that time). Beating the Blues therapy produces notably lower BDI values (and less variability) at the end of active treatment compared to Treatment As Usual, but the rates of change during the 6-month post-treatment phase are similar in the two treatment groups. Subject with higher baseline BDI had higher intercepts (BDI at the end of active treatment) but slightly steeper declines post-treatment. Subjects with longer current episodes had higher intercepts but similar slopes to those with shorter current episodes. Subjects with concomitant drug use had slightly lower intercepts but slightly higher slopes than those without. Boxplots showed a potential interaction with treatment for current episode length (intercept).

A potential final model evaluates the effect of treatment on intercept (performance during active treatment phase) and slope (change in BDI for 6 months after active treatment ended), while controlling for all covariates for intercept and examining the interaction between treatment and episode length on intercept. Based on this model, we see that Beating the Blues really only outperforms Treatment as Usual during active treatment for subjects with longer current episodes, after accounting for baseline BDI and concomitant drug use. The difference between Beating the Blues and Treatment as Usual in mean BDI at the end of active treatment is not significant for subjects with shorter current episodes (difference of .81 points; t = -0.30), but treatment effect is significantly larger for subjects with longer current episodes (t = -1.97 for treatment-by-length interaction). For subjects with longer episodes, Beating the Blues has mean BDI which is 7.14 points lower that Treatment as Usual, after accounting for baseline BDI and concomitant drug use. A simpler model without the treatment by episode length interaction showed significantly lower BDI scores for Beating the Blues patients overall, after accounting for baseline BDI, concomitant drug use, and episode length (difference of 4.21 points, t = -2.08), but we've seen that most of that difference is due to subjects with longer current episodes. Finally, as expected, mean BDI after active treatment was lower with lower baseline BDI and concomitant drug use, but the size of the treatment effect did not vary with baseline BDI levels or concomitant drug use.

After accounting for episode length, concomitant drug use, and baseline BDI, there is a slight trend for slower declines in BDI post-treatment with Beating the Blues compared to Treatment as Usual. Subjects receiving Treatment as Usual had a average drop of 0.95 BDI points per month post-treatment, while those receiving Beating the Blues therapy had an average drop of 0.45

BDI points (interaction t = 1.65). Change per month post-treatment did not depend on any other covariate.

10

Multilevel Data 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)
```

10.1 Exercises

10.1.1 Conceptual Exercises

1. **Seed germination.** Excluding plants with no height data may produce misleading conclusions. For example, if twice as many sterilized plants germinated compared to non-sterilized plants, but if those plants which sterilized had the same initial growth and growth rate, then our analysis would find no effect of sterilization. We could replace all missing values with 0's, indicating lack of germination, but then we'd likely have non-normal data, with a significant percentage of values right at 0.

2. A Level Two covariate could potentially differ from plant to plant within a pot. For instance, seed size, planting depth, distance from edge of pot, etc.

3. Mean heights by pot in Figure 10.2 would be reasonable, especially since

treatments were applied at the pot level. The story would almost invariably be the same; the only difference would be that each pot would be weighted the same regardless of the number of germinated plants in each pot. One line per pot would also be acceptable in Figure 10.3, although it seems a bit more natural to have one line per plant and to be able to see which plants germinated late or died early.

4. The standard deviation of intercepts and slopes from a single pot gives a measure of variability between plants from the same pot; we'd get one SD per pot, and we can average them together to get an overall measure of plant-to-plant variability. If we are interested instead in variability between pots, we'd get an average intercept or slope from the plants in each pot, and then measure how different the pots' intercepts and slopes are using SD.

5. You could subtract the mean number of days to center time, but centering just refers to any shifting of the times to make 0 (and the scale in general) more sensible. By subtracting 13, centered time will be at 0 for a plant's initial height measurement (Day 13). Otherwise, the initial heights would have a negative time, representing the number of days before the average time that all heights were taken.

6. There is evidence that heights further apart in time are less correlated, a gap of 10 days produces about the same correlation no matter when the start date was, and that those correlations seem to be decreasing in a systematic way.

7. A multilevel model can still be very helpful. Time (at Level One) is probably the biggest factor, and once time trends have been accounted for, a larger percentage of unexplained variability may come from the Level Three experimental factors. In some ways, using 72 mean heights by pot is appealing, since they'd represent 72 independent observations, and we could use traditional regression or anova analyses. However, there is so much information that is thrown away in that process. Multilevel models help us model growth rates, control for differences among plants and pots, and then evaluate the effects of experimental factors on initial heights, growth rates, etc.

8. A likelihood ratio test can be used when comparing nested models (especially models nested in fixed effects). In this case, Model B is a reduced version of Model C, with the same variance components but with all fixed effects other than the intercept and slope set to 0.

9. No—they have different interpretations; they are not estimating the same thing. In Model A, $\hat{\sigma}_u^2$ estimates the plant-to-plant variability in average height (across all time points), while in Model B $\hat{\sigma}_u^2$ estimates plant-to-plant variability in initial (day 13) heights.

10. Boundary constraints come into play when "legal" values for estimates of model parameters have limits (e.g., variances can't be negative, correlations must be between -1 and 1). In multilevel models we use likelihoods to

simultaneously estimate values for all model parameters, including fixed effects and variance components. It's possible that our likelihood function could be maximized at an "illegal" value of a parameter, so we might have to adjust the estimate of that parameter and accept a slightly smaller value of the likelihood.

11. We could have removed the error term on Time at Level Two, which would effectively remove the correlation term with the intercept error term as well. We could also remove insignificant fixed effects to see if that changes estimates of variance components, we could decide to use a single random effect at each level (on the intercept equation), or we could reparameterize covariates (although Time is really the only candidate, and its parameterization seems reasonable).

12. The bootstrapped test statistics (in the histogram) are shifted much closer to 0 than the chi-square curve. Thus, the area under the chi-square curve above the observed test statistic (vertical line) is much greater than the proportion of bootstrapped values from the histogram above the observed test statistic.

13. (a) We'll have 1 L1 equation, 3 L2 equations, and 6 L3 equations. Each Level Three equation has 4 fixed effects coefficients for a total of 24 fixed effects. At Level One there is a single variance term, at Level Two there are 3 variance terms and 3 covariance terms, and at Level Three there are 6 variance terms and 15 covariances, for a total of 1+6+21=28 variance components. (b) A much simplified model would include only random intercepts (thus removing all covariances and leaving just one variance term per level) and remove all interactions (so covariates would only be included in the intercept equation at each level). This would leave 7 fixed effects and 3 variance components = 10 parameters. (c) In part (b) we assumed no interaction (so the effect of each covariate is constant across all levels of all other covariates). We also accounted for variability among days, plants, and pots holding covariates at each level constant, but we assumed the effect of each covariate is constant across all days, plants, or pots.

14. We generate bootstrapped data according to estimated parameters from Model F:

Level Three:

$$\begin{aligned} a_1 &= \hat{\alpha}_0 + \tilde{u}_1 = 1.529 + N(0, .221) = 1.529 + .280 = 1.809 \\ b_1 &= \hat{\beta}_0 + \hat{\beta}_1 strl_1 + \hat{\beta}_2 rem_1 + \hat{\beta}_3 strl_1 rem_1 \\ &= .091 + .060(1) - .016(0) - .039(1)(0) = .151 \end{aligned}$$

Level Two:

$$\begin{aligned} a_{11} &= a_1 + u_{11} = 1.809 - .402 = 1.407 \\ b_{11} &= b_1 + v_{11} = .151 - .030 = .121 \end{aligned}$$

Note: $[u_{11}, v_{11}]$ are selected simultaneously from a bivariate normal distribution where u_{11} has SD .542, v_{11} has SD .038, and they have correlation .155.

Level One: Since $Y_{ijk} = a_{ij} + b_{ij}time_{ijk} + \epsilon_{ijk}$, we have:

$$Y_{111} = 1.407 + .121(0) + N(0, .286) = 1.407 + .121(0) - .114 = 1.293$$
$$Y_{112} = 1.407 + .121(5) + N(0, .286) = 1.407 + .121(5) - .066 = 1.946$$
$$Y_{113} = 1.407 + .121(10) + N(0, .286) = 1.407 + .121(10) + .311 = 2.928$$
$$Y_{114} = 1.407 + .121(15) + N(0, .286) = 1.407 + .121(15) + .050 = 3.272$$

Then to carry out the parametric bootstrap test:

a) Generate 413 observations from 107 plants and 32 pots according to the process above.
b) Fit both Models D and F to the data generated in (a). Compute the likelihood ratio test statistic comparing the two models.
c) Repeat (a) and (b) 1000 times.
d) Plot the 1000 likelihood ratio test statistics, and find a p-value by counting the number of bootstrapped test statistics above the observed test statistic and dividing by 1000.

15. At Day 13 (time13=0), there is no difference in the effect of sterilization for remnant and non-remnant soil. However, by Day 28 (time13=15), sterilization increases average height by .900 mm in non-remnant soil (.060 x 15) but only by .315 mm in remnant soil (.060 x 15 - .039 x 15).

16. **Collective efficacy and violent crime.** L1 = item on survey, L2 = resident, L3 = neighborhood.

17.

- After adjusting for individual demographics and neighborhood characteristics, the collective efficacy score of homeowners was, on average, 0.122 points higher than that of non-homeowners. Thus, homeowners report more collective efficacy than non-homeowners.

- After adjusting for individual demographics and neighborhood characteristics, the average collective efficacy score increased by .021 points for each additional 10 years in age. Thus, older residents report more collective efficacy than younger residents.

- After adjusting for individual demographics and neighborhood characteristics, the average collective efficacy score increased by .035 points for each additional point in socio-economic status. Thus, higher SES is associated with more collective efficacy.

- After adjusting for individual demographics and neighborhood characteristics of concentrated disadvantage and residential stability, the average collective efficacy score decreased by .037 points for each additional point in immigrant concentration. Thus, higher immigrant concentration is associated with less collective efficacy.

- After adjusting for individual demographics and neighborhood characteristics of concentrated disadvantage and immigrant concentration, the average collective efficacy score increased by .074 points for each additional point in residential stability. Thus, higher residential stability is associated with more collective efficacy.

18.

- Level One:

$$Y_{ijk} = a_{ij} + \epsilon_{ijk}$$

- Level Two:

$$\begin{aligned} a_{ij} = a_i &+ b_i female_{ij} + c_i married_{ij} + d_i divorced_{ij} + f_i single_{ij} \\ &+ k_i homeowner_{ij} + l_i Latino_{ij} + m_i Black_{ij} + n_i mobility_{ij} + o_i age_{ij} \\ &+ q_i YearsNbd_{ij} + r_i SES_{ij} + u_{ij} \end{aligned}$$

- Level Three:

$$\begin{aligned} a_i &= \alpha_0 + \alpha_1 ConcDisadv_i + \alpha_2 ImmigConc_i + \alpha_3 ResidStabil_i + \tilde{u}_i \\ b_i &= \beta_0 \\ c_i &= \gamma_0 \\ d_i &= \delta_0 \\ f_i &= \phi_0 \\ k_i &= \kappa_0 \\ l_i &= \lambda_0 \\ m_i &= \mu_0 \\ n_i &= \nu_0 \\ o_i &= \omega_0 \\ q_i &= \xi_0 \\ r_i &= \rho_0 \end{aligned}$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$, $u_{ij} \sim N(0, \sigma_u^2)$, and $\tilde{u_i} \sim N(0, \sigma_{\tilde{u}}^2)$.

In composite form, we have:

$$\begin{aligned} Y_{ijk} = [&\alpha_0 + \beta_0 female_{ij} + \gamma_0 married_{ij} + \delta_0 divorced_{ij} + \phi_0 single_{ij} \\ &+ \kappa_0 homeowner_{ij} + \lambda_0 Latino_{ij} + \mu_0 Black_{ij} + \nu_0 mobility_{ij} \\ &+ \omega_0 age_{ij} + \chi_0 YearsNbd_{ij} + \rho_0 SES_{ij} + \alpha_1 ConcDisadv_i \\ &+ \alpha_2 ImmigConc_i + \alpha_3 ResidStabil_i] + [\tilde{u}_i + u_{ij} + \epsilon_{ijk}] \end{aligned}$$

19. Side-by-side spaghetti plots of each L2 and L3 covariate – for example, separate plots for males and females, where the x-axis is survey item number and the y-axis is collective efficacy score. Boxplots of L2 and L3 categorical covariates, where each resident's average collective efficacy score is calculated; similarly can show scatterplots of L2 and L3 continuous covariates vs. average score. Lattice plot showing spaghetti plot for each neighborhood.

20. There would be 95 parameters to estimate: 15 fixed effects, 1 variance at L1, 1 variance at L2, and 12 variances and 66 covariances at L3. With the 3 neighborhood covariates in each L3 equation, that would add an extra 33 parameters to estimate (but no new variance components).

21. Those are likely pseudo R2 values. Adding 11 individual demographics explained 3.2% of the person-to-person variability in collective efficacy scores within a neighborhood, compared to the random intercepts model. Adding 3 neighborhood characteristics explained 70.3% of the nbd-to-nbd variability in collective efficacy scores, compared to the random intercepts model.

22. Model 1 is similar to Table 3, with 11 L2 covariates and 3 L3 covariates, except perceived neighborhood violence is the response. Model 2 then just adds collective efficacy summarized at the neighborhood level as a 4th L3 covariate. The primary results is that higher levels of collective efficacy in a neighborhood are associated with lower levels of perceived violence, after controlling for individual demographics and other neighborhood characteristics like concentrated disadvantage, residential stability, and immigrant concentration.

10.1.2 Guided Exercises

1. **Tree tubes.** [Eisinger et al., 2011].

```
treeTubes <- read_csv("data/treetube.csv")

#transect/height
ggplot(treeTubes) +
```

```
  geom_density(aes(x = HEIGHT)) +
  facet_wrap(~TRANSECT)
ggplot(treeTubes) +
  geom_boxplot(aes(x = as.factor(TRANSECT), y = HEIGHT)) +
  coord_flip()

#tubex/height
ggplot(treeTubes) +
  geom_density(aes(x = HEIGHT, color = as.factor(TUBEX)))
ggplot(treeTubes) +
  geom_boxplot(aes(x = as.factor(TUBEX), y = HEIGHT)) +
  coord_flip()

#year/height
ggplot(treeTubes) +
  geom_density(aes(x = HEIGHT, color = as.factor(YEAR)))

#species/height
ggplot(treeTubes) +
  geom_density(aes(x = HEIGHT)) +
  facet_wrap(~SPECIES)

# Repeat all with log HEIGHT
ggplot(treeTubes) +
  geom_density(aes(x = log(HEIGHT))) +
  facet_wrap(~TRANSECT)
ggplot(treeTubes) +
  geom_boxplot(aes(x = as.factor(TRANSECT), y=log(HEIGHT))) +
  coord_flip()

ggplot(treeTubes) +
  geom_density(aes(x = log(HEIGHT), color = as.factor(TUBEX)))
ggplot(treeTubes) +
  geom_boxplot(aes(x = as.factor(TUBEX), y = log(HEIGHT))) +
  coord_flip()

ggplot(treeTubes) +
  geom_density(aes(x = log(HEIGHT), color = as.factor(YEAR)))

ggplot(treeTubes) +
  geom_density(aes(x = log(HEIGHT))) +
  facet_wrap(~SPECIES)
ggplot(treeTubes) +
  geom_density(aes(x = log(HEIGHT), color = SPECIES))
```

a) Tree heights are decidedly right skewed, so relationships with covariates are easier to see after logging heights. There appears to be quite a bit of variation in heights within transect, but overall heights are pretty similar except for transect #15, which tends to have taller trees. Trees with tubes tend to have slightly larger heights, although most trees had no tube (n = 4100 vs. 543). As time goes by, typical heights increase in an orderly fashion, although distributions of log heights for later years have more left skewness. Finally, typical heights and variability vary somewhat by species, with black walnut and ironwood having the largest concentrations of short trees.

```
table(treeTubes$TRANSECT, treeTubes$YEAR)
table(treeTubes$TRANSECT, treeTubes$TUBEX)

ggplot(treeTubes) +
  geom_bar(aes(x = as.factor(YEAR), fill = as.factor(TUBEX))) +
  facet_wrap(~TRANSECT)
```

b) We find that transects #15 and #17 have no measurements from 1990, and few (or none) in 1992. Transect #18 was not measured in 1995. More importantly, all three of these transects with missing years are those which had tubes. If we were to fit a model with an interaction for `TUBEX` and `YEAR`, the coefficient for `TUBEX` would represent the difference in mean tree heights between trees with and without tubes in 1990. This could be problematic.

c) At Level Three, data would be grouped by transect with `TUBEX` as a Level Three variable. At Level Two, data would be grouped by trees with `SPECIES` as a Level Two variable. Finally, at Level One, data would be grouped by `YEAR` with `YEAR` as a Level One variable.

```
ggplot(treeTubes, aes(x = YEAR, y = HEIGHT)) +
  geom_line(aes(group = ID), color = "dark grey") +
  facet_wrap(~as.factor(TUBEX), ncol=2) +
  geom_smooth(se = FALSE, color = "black")
```

d) We find that trees with tubes tend to grow at slower rates than trees without tubes, as indicated by the greater average rate of increase among trees without tubes. But there's more variability among trees without tubes.

```
treeModA <- lmer(HEIGHT ~ 1 + (1|ID) + (1|TRANSECT),
                 REML=T, data=treeTubes)
summary(treeModA)
```

e) At level one: $Y_{ijk} = a_{ij} + \epsilon_{ijk}$ where $\epsilon_{ijk} \sim \mathrm{N}(0, 2.46^2)$.
At level two: $a_{ij} = a_i + u_{ij}$ where $u_{ij} \sim \mathrm{N}(0, 0.57^2)$.
At level three: $a_i = 2.23 + \tilde{u}_i$ where $\tilde{u}_i \sim \mathrm{N}(0, 0.49^2)$.
This gives the composite model: $Y_{ijk} = 2.23 + \tilde{u}_i + u_{ij} + \epsilon_{ijk}$

```
treeTubes <- mutate(treeTubes, TIME = YEAR - 1990)
treeModB <- lmer(HEIGHT ~ TIME + (TIME|ID) + (TIME|TRANSECT),
                 REML = T, data = treeTubes)
summary(treeModB)
```

f) At level one: $Y_{ijk} = a_{ij} + b_{ij}TIME + \epsilon_{ijk}$ where $\epsilon_{ijk} \sim \mathrm{N}(0, 0.61^2)$.

At level two:

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

where

$$\begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33^2 & \\ -0.99(0.33)(0.16) & 0.16^2 \end{bmatrix} \right)$$

At level three:

$$a_i = -0.16 + \tilde{u}_i$$
$$b_i = 0.30 + \tilde{v}_i$$

where

$$\begin{bmatrix} \tilde{u}_i \\ \tilde{v}_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.20^2 & \\ -0.91(0.20)(0.06) & 0.06^2 \end{bmatrix} \right)$$

There is a correlation of random effects at Level Two of -0.99, implying the model was fit on the boundary.

```
treeModC <- lmer(HEIGHT ~ TIME + (1|ID) + (0+TIME|ID) +
  (1|TRANSECT) + (0+TIME|TRANSECT), REML = T, data = treeTubes)
summary(treeModC)
```

g) At level two:

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

where

$$u_{ij} \sim N(0, 0.26^2)$$
$$v_{ij} \sim N(0, 0.15^2)$$

At level three:

$$a_i = -0.16 + \tilde{u}_i$$
$$b_i = 0.30 + \tilde{v}_i$$

where

$$\tilde{u}_i \sim N(0, 0.20^2)$$
$$\tilde{v}_i \sim N(0, 0.07^2)$$

h) As Section 10.6 details, using a chi-square distribution to compare models that differ in their variance components would likely yield a p-value that is artificially too large.

```
bootstrapAnova.par <- function(mA, m0, B=1000, cores = 1){
  library(parallel)
  oneBootstrap <- function(m0, mA){
    d <- drop(simulate(m0))
    m2 <-refit(mA, newresp=d)
    m1 <-refit(m0, newresp=d)
    return(anova(m2,m1)$Chisq[2])
  }
  if(cores>1){
    cl <- makeCluster(cores)
    clusterEvalQ(cl, library(lme4))
    clusterExport(cl,c("mA", "m0"), envir = environment())
    nulldist <- parSapply(cl, 1:B, function(i, ...)
      {oneBootstrap(m0, mA)})
    stopCluster(cl)
  } else {
    nulldist <- replicate(B, oneBootstrap(m0, mA))
  }
  ret <- anova(mA, m0)
  ret$"Pr(>Chisq)"[2] <- mean(ret$Chisq[2] < nulldist)
```

```
  names(ret)[8] <- "Pr_boot(>Chisq)"
  attr(ret, "heading") <- c(attr(ret, "heading")[1],
      paste("Parametric bootstrap with", B,"samples."),
      attr(ret, "heading")[-1])
  attr(ret, "nulldist") <- nulldist
  return(ret)
}

# Takes about 5 minutes to run
treeBVsTreeC <- bootstrapAnova.par(mA=treeModB, m0=treeModC,
                                   B=500, cores=4)
treeBVsTreeC
```

i) The test favors Model B (the alternative model) which allows for nonzero correlation of random effects at Levels Two and Three. We have statistically significant evidence (LRT = 423, parametric bootstrap p < .002) that Model B provides a better fit than Model C, and the correlations of random effects at Levels Two and Three are nonzero.

```
treeModD <- lmer(HEIGHT ~ TIME + TIME:TUBEX + (1|ID) +
  (0+TIME|ID) + (1|TRANSECT) + (0+TIME|TRANSECT),
  REML = T, data = treeTubes)
summary(treeModD)
```

j) The average yearly increase in height is 0.056 meters per year lower among trees with tubes than trees without tubes (although it is not statistically significant, Z = -1.057). Trees without tubes have an average yearly increase in height of .316 meters, compared to an average yearly increase of .260 meters for trees with tubes.

```
anova(treeModD, treeModC)
```

k) We do not have statistically significant evidence (LRT = 5.64 on 2 df, p = .060) that tube presence affects initial height or yearly increase, although this could be considered marginally significant evidence. We could also test, for example, a model with `TIME` vs. a model with `TIME` and `TIME:TUBEX` (assuming no tube effect on initial heights) or a model with `TIME` and `TUBEX` vs. a model that adds the interaction term.

2. **Kentucky math scores.** [Bickel, 2007].

```
kentucky <- read_csv("data/kentucky.csv")
kentucky <- kentucky %>%
  mutate(female2 = ifelse(female == 1, "Female", "Male"),
    nonwhite2 = ifelse(nonwhite == 1, "Non-white", "White"))

kentucky %>% group_by(female2) %>%
  filter(!is.na(female2) & !is.na(mathn)) %>%
  summarise(means = mean(mathn),
            sds = sd(mathn),
            meds = median(mathn),
            q1s = quantile(mathn, 0.25),
            q3s = quantile(mathn, 0.75),
            mins = min(mathn),
            maxs = max(mathn),
            ns = n())
kentucky %>% group_by(nonwhite2) %>%
  filter(!is.na(nonwhite2) & !is.na(mathn)) %>%
  summarise(means = mean(mathn),
            sds = sd(mathn),
            meds = median(mathn),
            q1s = quantile(mathn, 0.25),
            q3s = quantile(mathn, 0.75),
            mins = min(mathn),
            maxs = max(mathn),
            ns = n())

with(kentucky, cor(cbind(mathn,readn,sch_size,sch_ses,
        dis_size,dis_ses), use="pairwise.complete.obs"))

sampdata <- kentucky[sample(1:48168,size=2000),]
ggplot(sampdata, aes(x = female2, y = mathn)) +
  geom_boxplot() + coord_flip()
ggplot(sampdata, aes(x = nonwhite2, y = mathn)) +
  geom_boxplot() + coord_flip()
ggplot(sampdata, aes(x = sch_ses, y = mathn)) +
  geom_point() + geom_smooth()
ggplot(sampdata, aes(x = dis_size, y = mathn)) +
  geom_point() + geom_smooth()
```

a) There are moderately strong negative correlations between math scores and SES at the school level (r=-0.25) and at the district level (r=-0.19), but little correlation between math scores and school or district size. Nonwhite students have a noticeably lower mean math score than white students (40.8

vs. 51.1), while female students have a slightly higher mean score than males (50.3 vs. 49.4).

```
#Model A
modela <- lmer(mathn ~ 1 + (1|sch_id) + (1|dis_id),
               REML=T, data=kentucky)
summary(modela)
```

b) 91% of variability in math scores is found at Level One (student), 7.0% at Level Two (school), and 1.6% at Level Three (district).

```
#Model B
modelb <- lmer(mathn~female + nonwhite + female:nonwhite +
  (female+nonwhite+female:nonwhite|sch_id) +
  (female+nonwhite+female:nonwhite|dis_id), REML=T,
  data=kentucky)
summary(modelb)
```

c)

- Level One:

$$Y_{ijk} = a_{ij} + b_{ij}female_{ijk} + c_{ij}nonwhite_{ijk} + d_{ij}female:nonwhite_{ijk} + \epsilon_{ijk}$$

- Level Two:

$$\begin{aligned} a_{ij} &= a_i + u_{ij} \\ b_{ij} &= b_i + v_{ij} \\ c_{ij} &= c_i + w_{ij} \\ d_{ij} &= d_i + z_{ij} \end{aligned}$$

- Level Three:

$$\begin{aligned} a_i &= \alpha_0 + \tilde{u}_i \\ b_i &= \beta_0 + \tilde{v}_i \\ c_i &= \gamma_0 + \tilde{w}_i \\ d_i &= \delta_0 + \tilde{z}_i \end{aligned}$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

$$\begin{bmatrix} u_{ij} \\ v_{ij} \\ w_{ij} \\ z_{ij} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & & & \\ \sigma_{uv} & \sigma_v^2 & & \\ \sigma_{uw} & \sigma_{vw} & \sigma_w^2 & \\ \sigma_{uz} & \sigma_{vz} & \sigma_{wz} & \sigma_z^2 \end{bmatrix} \right).$$

and then at Level Three:

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

We must estimate 25 parameters in Model B: 4 fixed effects and 21 variance components, with 1 variance term at Level One, 4 variance terms and 6 correlations at Level Two, and 4 variance terms and 6 correlations at Level Three.

Based on a pseudo R-squared value ((416.219-399.568) / 416.219), 4.0% of the student-to-student variability in math scores is explained by gender, race, and their interaction.

There is no significant evidence (t=-0.68) of an interaction between female and nonwhite. In layman's terms, the gender gap is similar among whites and nonwhites, and the gap between whites and nonwhites is similar among boys and girls.

```
modelc <- lmer(mathn~female + nonwhite + sch_ses +
  female:sch_ses + nonwhite:sch_ses +
  (female+nonwhite|sch_id) + (female + nonwhite + sch_ses +
  female:sch_ses + nonwhite:sch_ses | dis_id),
  REML=T, data=kentucky)
summary(modelc)
```

d) There are 6+1+6+21 = 34 parameters to estimate (fixed + varcomp at L1 + varcomp at L2 + varcomp at L3).

```
modeld <- lmer(mathn~female + sch_ses + dis_size +
  female:sch_ses + female:dis_size + sch_ses:dis_size +
  female:sch_ses:dis_size + (female|sch_id) +
  (female+sch_ses+female:sch_ses|dis_id), REML=T,
  data=kentucky)
summary(modeld)
```

```
modeld0 <- lmer(mathn~female + sch_ses + dis_size +
  female:sch_ses + female:dis_size + sch_ses:dis_size +
  female:sch_ses:dis_size + (female|sch_id) +
  (1|dis_id), REML=T, data=kentucky)
summary(modeld0)

anova(modeld0,modeld)

# Investigate 3-way interaction
summary(kentucky$sch_ses)
summary(kentucky$dis_size)
# Use parameter estimates from Model D0
mathn_hat <- function(female, schses, dissize) {
  46.5 - .042*female - .214*schses - 1.94*dissize -
    .0434*female*schses + .844*female*dissize -
    .076*schses*dissize + .0175*female*schses*dissize
}
mathn_hat(1, -15.64, 1.088)
mathn_hat(0, -15.64, 1.088)
mathn_hat(1, -29.38, .12)
mathn_hat(0, -29.38, .12)
mathn_hat(1, -29.38, 1.58)
mathn_hat(0, -29.38, 1.58)
mathn_hat(1, -4.6, .12)
mathn_hat(0, -4.6, .12)
mathn_hat(1, -4.6, 1.58)
mathn_hat(0, -4.6, 1.58)
```

e) There are 8+1+3+10 = 22 parameters to estimate.

- Level One:
$$Y_{ijk} = a_{ij} + b_{ij} female_{ijk} + \epsilon_{ijk}$$
- Level Two:
$$\begin{aligned} a_{ij} &= a_i + c_i schses_{ij} + u_{ij} \\ b_{ij} &= b_i + d_i schses_{ij} + v_{ij} \end{aligned}$$
- Level Three:
$$\begin{aligned} a_i &= \alpha_0 + \alpha_1 dissize_i + \tilde{u}_i \\ b_i &= \beta_0 + \beta_1 dissize_i + \tilde{v}_i \\ c_i &= \gamma_0 + \gamma_1 dissize_i + \tilde{w}_i \\ d_i &= \delta_0 + \delta_1 dissize_i + \tilde{z}_i \end{aligned}$$
- Composite model:

$$\begin{aligned} Y_{ijk} &= [\alpha_0 + \alpha_1 dissize_i + \beta_0 female_{ijk} + \beta_1 dissize_i female_{ijk} + \\ &\quad \gamma_0 schses_{ij} + \gamma_1 dissize_i schses_{ij} + \delta_0 schses_{ij} female_{ijk} + \\ &\quad \delta_1 dissize_i schses_{ij} female_{ijk}] + [\tilde{u}_i + \tilde{v}_i female_{ijk} + \\ &\quad \tilde{w}_i schses_{ij} + \tilde{z}_i schses_{ij} female_{ijk} + u_{ij} + v_{ij} female_{ijk} + \epsilon_{ijk}] \end{aligned}$$

where $\epsilon_{ijk} \sim N(0, \sigma^2)$, and we assume the following variance-covariance structure at Level Two:

$$\begin{bmatrix} u_{ij} \\ v_{ij} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix} \right).$$

and then at Level Three:

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

For Model D0, we assume all variance and covariances at Level 3 are 0 except for $\sigma_{\tilde{u}}^2$. That is, after adjusting for district size, the effects from Level 2 do not vary by district; there is just a single overall district effect that is a random effect.

Estimates of fixed effects are pretty similar, although SEs are lower with `(1|dis_id)`, making fixed effects appear more significant.

I would choose the simpler model (D0). The smaller model has lower AIC and BIC, and some of the correlation coefficients in the fuller model are suspiciously close to 1 or -1. The likelihood ratio test is not significant (LRT=4.64 on 9 df, p=.865), although a parametric bootstrap test would be more reliable. It's not surprising that fixed effects in the fuller model are less significant—multilevel models often provide a more accurate assessment of the lack of independence, which increases SEs—but if a model has too many terms, sometimes the estimates can be suspect.

The gender gap (with females outperforming males on average) is fairly consistent except in schools with high SES from smaller districts. Both genders perform much better, on average, in schools with lower SES. In addition, both genders perform slightly better in districts with larger sizes if schools have lower SES, but in districts with smaller sizes if schools have higher SES.

```
kentucky %>% mutate(dis_size2 = ifelse(dis_size > 0.76,
                                       "large", "small")) %>%
```

```
  ggplot(aes(x = sch_ses, y = mathn, color = female2)) +
    geom_point(size = .25) +
    geom_smooth(method = "lm", se = FALSE) +
    facet_wrap(~ dis_size2, ncol = 2)
```

Faceted scatterplots of `sch_ses` vs. `mathn` illustrate the three-way interaction contained in Model D. Districts with size above the median are shown on the left, with the relationship between `sch_ses` and `mathn` fitted separately for males and females. Smaller districts are shown on the right. In smaller districts, males outperform females for higher levels of `sch_ses`, but females outperform males for lower levels of `sch_ses`. In larger districts, females outperform males by a consistent amount regardless of `sch_ses` level. Of course, these plots do not control for other variables.

10.1.3 Open-Ended Exercises

1. **Seed germination: coneflowers.**

```
seedwd <- read_csv("data/seeds2.csv")
dim(seedwd)
seedwd[135:146,2:11]   # illustrate wide data

# Remove plants that did not germinate
seedwd <- seedwd %>%
  mutate(nope = is.na(hgt13) + is.na(hgt18) + is.na(hgt23) +
           is.na(hgt28)) %>%
  filter(nope < 4)

# Create data frame in LONG form (one obs per plant
#   measurement time)
seedlg <- seedwd %>%
  gather(key = time, value = hgt, hgt13:hgt28) %>%
  mutate(time = as.integer(str_sub(time, -2)),
         time13 = time - 13 )

# Variables in seedlg:
# 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.  Should be Y for
#   all observations in seedlg (true except for plant 281 -
#   probably an error)
# time = number of days after planting when height measured
# time13 = centered time, so that first day of measurement
#   is Day 0
# hgt = height of plant (in mm).

# create indicator variables for later analyses
seedlg <- seedlg %>%
  mutate(cult=ifelse(soil=="CULT",1,0),
         rem=ifelse(soil=="REM",1,0),
         stp=ifelse(soil=="STP",1,0),
         lead=ifelse(species=="L",1,0),
         cone=ifelse(species=="C",1,0),
         strl=ifelse(sterile=="Y",1,0),
         nostrl=ifelse(sterile=="N",1,0) )

# create separate data sets of leadplants and coneflowers
conedata <- seedlg %>%
  filter(cone==1)
leaddata <- seedlg %>%
  filter(lead==1)

## Exploratory data analysis ##

# Add average across all time points for each plant for EDA plots
meanplant <- seedlg %>% group_by(plant) %>%
  summarise(meanplant = mean(hgt, na.rm = TRUE))
seedwd <- seedwd %>%
  left_join(meanplant, by = "plant")
conewd <- seedwd %>%
  filter(species=="C")
leadwd <- seedwd %>%
  filter(species=="L")

# One obs per plant - assume plants relatively independent
#   within pots
ggplot(conedata,aes(x=soil,y=hgt)) +
  geom_boxplot() +
  labs(x="Soil type",y="Plant Height (mm)",
       title="Coneflowers (a)")
```

```
ggplot(conedata,aes(x=sterile,y=hgt)) +
  geom_boxplot() +
  labs(x="Sterilized",y="Plant Height (mm)",
       title="Coneflowers (b)")

seedlg <- seedlg %>%
  mutate(speciesname = ifelse(species=="C","Coneflowers",
                              "Leadplants") )

ggplot(conedata, aes(x=time, y=hgt)) +
  geom_line(aes(group=plant), color="dark grey") +
  facet_wrap(~pot, ncol=7) +
  geom_smooth(se=FALSE, color="black") +
  labs(x="Days since seeds planted",y="Plant height (mm)")

conedata <- conedata %>%
  mutate(soilname = recode(soil, STP="Reconstructed",
                           CULT="Cultivated", REM="Remnant") )
ggplot(conedata,aes(x=time,y=hgt)) +
  geom_line(aes(group=plant),color="dark grey") +
  facet_wrap(~soilname,ncol=3) +
  geom_smooth(se=FALSE,color="black") +
  labs(x="Days since seeds planted",y="Plant height (mm)")

conedata <- conedata %>%
  mutate(sterilename = recode(sterile, Y="Sterilized",
                              N="Not Sterilized") )
ggplot(conedata,aes(x=time,y=hgt)) +
  geom_line(aes(group=plant),color="dark grey") +
  facet_wrap(~sterilename,ncol=2) +
  geom_smooth(se=FALSE,color="black") +
  labs(x="Days since seeds planted",y="Plant height (mm)")

# Summary stats for linear models by plant - use centered time
hgt.list=lmList(hgt~time13 | plant, data=conedata,
                na.action=na.exclude)
int = as.matrix(coef(hgt.list))[,1]
summary(int)
rate = as.matrix(coef(hgt.list))[,2]
summary(rate)
rsq <- by(conedata, conedata$plant, function(data)
  summary(lm(hgt ~ time13,
             data = data,na.action=na.exclude))$r.squared)
summary(rsq)
```

```
sum(rsq[!is.na(rsq)]>=.8)/length(rsq[!is.na(rsq)])

int.rate.rsq <- as.data.frame(cbind(int,rate,rsq))
int.hist <- ggplot(int.rate.rsq,aes(x=int)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Intercepts",y="Frequency",title="(a)")
rate.hist <- ggplot(int.rate.rsq,aes(x=rate)) +
  geom_histogram(binwidth=0.05,color="black",fill="white") +
  labs(x="Slopes",y="Frequency",title="(b)")
rsq.hist <- ggplot(int.rate.rsq,aes(x=rsq)) +
  geom_histogram(binwidth=0.1,color="black",fill="white") +
  labs(x="R-squared values",y="Frequency",title="(c)")
grid.arrange(int.hist,rate.hist,rsq.hist,ncol=2)

# Descriptive statistics of the estimates obtained by
#   fitting the linear model by plant.
mean(int)
sd(int)
mean(rate, na.rm=T)
sd(rate, na.rm=T)
cor(int, rate, use="complete.obs")

# Summary stats for linear models by pot
hgt2.list=lmList(hgt~time13 | pot, data=conedata,
                 na.action=na.exclude)
int2 = as.matrix(coef(hgt2.list))[,1]
summary(int2)   # summary statistics for 32 intercepts
rate2 = as.matrix(coef(hgt2.list))[,2]
summary(rate2)
rsq2 <- by(conedata, conedata$pot, function(data)
  summary(lm(hgt ~ time13,
             data = data,na.action=na.exclude))$r.squared)
summary(rsq2)

int.rate.rsq2 <- as.data.frame(cbind(int2,rate2,rsq2))
int2.hist <- ggplot(int.rate.rsq2,aes(x=int2)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Intercepts",y="Frequency",title="(a)")
rate2.hist <- ggplot(int.rate.rsq2,aes(x=rate2)) +
  geom_histogram(binwidth=0.05,color="black",fill="white") +
  labs(x="Slopes",y="Frequency",title="(b)")
rsq2.hist <- ggplot(int.rate.rsq2,aes(x=rsq2)) +
  geom_histogram(binwidth=0.2,color="black",fill="white") +
  labs(x="R-squared values",y="Frequency",title="(c)")
```

```
grid.arrange(int2.hist,rate2.hist,rsq2.hist,ncol=2)

# Descriptive statistics of the estimates obtained by
#   fitting the linear model by pot.
mean(int2)
sd(int2)
mean(rate2, na.rm=T)
sd(rate2, na.rm=T)
cor(int2, rate2, use="complete.obs")

# Try to compare variability between plants and between pots
#   Plus ANOVA-type boxplot of within pot vs within
#   plant variability
seedwdplus <- seedwd %>%
  filter(species=="C") %>%
  mutate(int = int, rate = rate)
bypot <- seedwdplus %>%
  group_by(pot) %>%
  summarise(sd_int = unname(sd(int)),
            mean_int = unname(mean(int)),
            sd_rate = unname(sd(rate)),
            mean_rate = unname(mean(rate)) )

# intercept variability between plants = .677
mean(bypot$sd_int,na.rm=T)
# rate variability between plants = .058
mean(bypot$sd_rate,na.rm=T)
# intercept variability between pots = .640
sd(bypot$mean_int,na.rm=T)
# rate variability between pots = .062
sd(bypot$mean_rate,na.rm=T)

# Boxplots to compare ints and rates by factor levels
hgt.listl=lmList(hgt~time13 | plant, data=leaddata,
                 na.action=na.exclude)
intl = as.matrix(coef(hgt.listl))[,1]
ratel = as.matrix(coef(hgt.listl))[,2]
allfits <- tibble(int = c(int, intl),
                  rate = c(rate, ratel),
                  species = c(rep("Coneflowers", 176),
                              rep("Leadplants", 107)) )

int.byspecies <- ggplot(allfits,aes(x=species,y=int)) +
  geom_boxplot(aes(group=species)) +
```

```
  labs(x="Species",y="Intercepts",title="(a)")
rate.byspecies <- ggplot(allfits, aes(x=species,y=rate)) +
  geom_boxplot(aes(group=species)) +
  labs(x="Species",y="Slopes",title="(b)")
grid.arrange(int.byspecies,rate.byspecies,ncol=2)

seedwdplus <- seedwdplus %>%
  mutate(soilname = recode(soil, STP="Reconstructed",
                           CULT="Cultivated",
                           REM="Remnant") )
int.bysoil <- ggplot(seedwdplus,aes(x=soilname,y=int)) +
  geom_boxplot() +
  labs(x="Soil type",y="Intercepts",title="(a)")
rate.bysoil <- ggplot(seedwdplus,aes(x=soilname,y=rate)) +
  geom_boxplot() +
  labs(x="Soil type",y="Slopes",title="(b)")
grid.arrange(int.bysoil,rate.bysoil,ncol=2)

seedwdplus <- seedwdplus %>%
  mutate(sterilename = recode(sterile, Y="Sterilized",
                              N="Not Sterilized") )
int.bysterile <- ggplot(seedwdplus,aes(x=sterilename,y=int)) +
  geom_boxplot() +
  labs(x="Sterile",y="Intercepts",title="(a)")
rate.bysterile <- ggplot(seedwdplus,aes(x=sterilename,y=rate)) +
  geom_boxplot() +
  labs(x="Sterile",y="Slopes",title="(b)")
grid.arrange(int.bysterile,rate.bysterile,ncol=2)

# Examine correlation structure
seed.nona <- conedata %>%
  filter(!is.na(hgt))
hgt.lm = lm(hgt~time13, data=seed.nona)
seed.nona <- seed.nona %>%
  mutate(lmres = resid(hgt.lm))
hgtw <- seed.nona %>%
  dplyr::select(plant, time13, lmres) %>%
  spread(key = time13, value = lmres, sep = "=")
hgtw.1 <- na.omit(hgtw)
ggpairs(hgtw.1[,2:5], lower=list(continuous="smooth"),
  upper=list(), diag=list(continuous="bar", discrete="bar"),
  axisLabels="show")
```

```
### Model fitting - coneplants ###

# Model A - unconditional means
modelac = lmer(hgt ~ 1 + (1|plant) + (1|pot), REML=T,
               data=conedata)
summary(modelac)

# Model B - unconditional growth
modelbc = lmer(hgt ~ time13 + (time13|plant) + (time13|pot),
  REML=T, data=conedata)
summary(modelbc)

# Model C - add covariates at pot level
modelcc = lmer(hgt ~ time13 + strl + cult + rem + time13:strl +
  time13:cult + time13:rem + (time13|plant) + (time13|pot),
  REML=T, data=conedata)
summary(modelcc)

# go with Model C; remember that anova() assumes REML=F
anova(modelbc,modelcc)

# Model D - add interactions to Model C
modeldc = lmer(hgt ~ time13 + strl + cult + rem + time13:strl +
  time13:cult + time13:rem + strl:cult + strl:rem +
  time13:strl:cult + time13:strl:rem + (time13|plant) +
  (time13|pot), REML=T, data=conedata)
summary(modeldc)

# Model E - remove insignificant terms from Model D
modelec = lmer(hgt ~ time13 + strl + rem + strl:rem +
  time13:strl + time13:rem + time13:strl:rem + (time13|plant) +
  (time13|pot), REML=T, data=conedata)
summary(modelec)

anova(modeldc,modelec)        # go with Model E
```

After investigating the effects of soil type and sterilization on leadplants, we now turn to the 36 pots with coneflowers. More coneflower plants (176) germinated than leadplants (107), so we have greater power for statistical comparisons among germinated plants. Side-by-side spaghetti plots of coneflowers and leadplants show similar starting heights and similar total growth over 15 days, although coneflowers tend to see faster growth earlier compared to leadplants. There is also greater variability in initial heights and growth rates among coneflower plants than we had observed among leadplants.

Side-by-side spaghetti plots and boxplots of fitted slopes and intercepts show that soil from remnant prairie is associated with the tallest initial heights but the slowest growth rates. In addition, sterilization is associated with both taller plants initially and faster growth rates than non-sterilized plants. Heights over time for the same plant are highly correlated, with stronger correlations for time points that are closer together and later in the study period (for example, Day 23 and 28 heights have a correlation of .93).

Initial multilevel models show that 46% of total variability in plant height is found within individual plants, while 26% occurs between plants in the same pot, and 29% occurs between different pots. Accounting for time explains 84% of variability in heights within plants, and then our final model explains 72% of the pot-to-pot variability in initial heights and 85% of the pot-to-pot variability in growth rate.

In our final multilevel model, the largest effects on initial height and growth rate were associated with sterilization and soil from remnant prairie. The effect of sterilization on initial heights was significantly different for remnant and non-remnant soil (t=5.511); at Day 13, average initial height in sterilized non-remnant soil was 0.81 mm greater than the same height in unsterilized non-remnant soil, but average initial height in sterilized remnant soil was -0.89 mm less than the same height in unsterilized remnant soil. A similar effect with a greater difference is seen at Day 23: sterilization produced an additional 1.76 mm of height in non-remnant soil, but 0.48 mm less height in remnant soil.

The effect of sterilization on growth rate is significantly different for remnant and non-remnant soil (t=-2.103). In non-remnant soil, sterilization increases the average growth rate from .079 mm/day to .175 mm.day, while in remnant soil, sterilization increases the average growth rate from .058 mm/day to .099 mm/day. So, while remnant soil has higher initial heights and slower growth than other soil types, sterilization has a negative effect on both initial height and growth rate in remnant soil compared to other soil types. Since this was a randomized experiment, we can conclude that differences in initial height and growth rate of coneflowers were caused by sterilization and soil type, although we must be careful about generalizing results which occurred in carefully controlled greenhouse conditions with the effects of sterilization and soil type that would be seen in nature.

2. **Mudamalai leaf growth.** [Pray, 2009].

```
# mudamalai - 3 level analysis with leaf data from India.
#   Note slight imbalance with 1960 observations - a perfect
#   data set would have 45 trees (3x3x5) with 9 branches per
#   tree (3 br x 3 strata) and 5 leaves per branch = 2025 obs
```

```
leaves <- read.csv("data/mudamalai.csv")
dim(leaves)
head(leaves)

# Investigate patterns of missingness
md.pattern(leaves)

## Exploratory data analysis ##

# Rough look at key Level Three and Two covariates
table(leaves$Species)
table(leaves$Zone)
table(leaves$Strata)

# Histograms of Level One variables
Area.hist <- ggplot(leaves,aes(x=Area)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Area",y="Frequency",title="(a)")
Width.hist <- ggplot(leaves,aes(x=Width)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Width",y="Frequency",title="(b)")
Petl.hist <- ggplot(leaves,aes(x=Petiole.length)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Petiole length",y="Frequency",title="(c)")
Petw.hist <- ggplot(leaves,aes(x=Petiole.width)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Petiole width",y="Frequency",title="(d)")
Herb.hist <- ggplot(leaves,aes(x=Herbivory)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Herbivory",y="Frequency",title="(e)")
grid.arrange(Area.hist,Width.hist,Petl.hist,Petw.hist,
             Herb.hist,ncol=3)

# Histograms of Level Two and Three variables
Height.hist <- ggplot(leaves,aes(x=Tree.height)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Tree height",y="Frequency",title="(a)")
Girth.hist <- ggplot(leaves,aes(x=Tree.girth)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Tree girth",y="Frequency",title="(b)")
Length.hist <- ggplot(leaves,aes(x=Branch.length)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Branch length",y="Frequency",title="(c)")
```

```
Stomata.hist <- ggplot(leaves,aes(x=Stomata)) +
  geom_histogram(binwidth=0.5,color="black",fill="white") +
  labs(x="Stomata",y="Frequency",title="(d)")
grid.arrange(Height.hist,Girth.hist,Length.hist,
             Stomata.hist,ncol=2)

# Extract first row from each branch to create Level 2 data set
temp.lev2 = leaves[order(leaves$Tree,leaves$Branchnum),]
leaf.lev2 = temp.lev2[c(TRUE, temp.lev2$Branchnum[-1] !=
  temp.lev2$Branchnum[-length(temp.lev2$Branchnum)]), ]
dim(leaf.lev2)   # 405 x 17
head(leaf.lev2)

# Add average across all leaves for each branch for EDA plots
meanAreabyBrch = by(leaves$Area,leaves$Branchnum,mean,na.rm=T)
leaf.lev2 = data.frame(leaf.lev2,
    meanAreabyBrch=as.numeric(meanAreabyBrch))

# Extract first row from each tree to create Level 3 data set
leaf.lev3 = temp.lev2[c(TRUE, temp.lev2$Tree[-1] !=
  temp.lev2$Tree[-length(temp.lev2$Tree)]), ]
dim(leaf.lev3)   # 45 x 17
head(leaf.lev3)

# Add average across all leaves and branches for each tree for
#   EDA plots
meanAreabyTree = by(leaves$Area,leaves$Tree,mean,na.rm=T)
leaf.lev3 = data.frame(leaf.lev3,
  meanAreabyTree=as.numeric(meanAreabyTree))

# One obs per branch - mean Area vs. L2 covariates
L2.1 <- ggplot(leaf.lev2,aes(x=Strata,y=meanAreabyBrch)) +
  geom_boxplot() +
  labs(x="Strata",y="Mean Area",title="(a)")
L2.2 <- ggplot(data=leaf.lev2,aes(x=Branch.length,
                                  y=meanAreabyBrch)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Mean Area") + xlab("Branch length") + labs(title="(b)")
L2.3 <- ggplot(data=leaf.lev2,aes(x=Stomata,y=meanAreabyBrch)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Mean Area") + xlab("Stomata") + labs(title="(c)")
grid.arrange(L2.1,L2.2,L2.3,ncol=2)
```

```
# One obs per tree - mean Area vs. L3 covariates
L3.1 <- ggplot(leaf.lev3,aes(x=Species,y=meanAreabyTree)) +
  geom_boxplot() +
  labs(x="Species",y="Mean Area",title="(a)")
L3.2 <- ggplot(leaf.lev3,aes(x=Zone,y=meanAreabyTree)) +
  geom_boxplot() +
  labs(x="Zone",y="Mean Area",title="(b)")
L3.3 <- ggplot(data=leaf.lev3,aes(x=Tree.height,
                                  y=meanAreabyTree)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Mean Area") + xlab("Tree height") + labs(title="(c)")
L3.4 <- ggplot(data=leaf.lev3,aes(x=Tree.girth,
                                  y=meanAreabyTree)) +
  geom_point() +
  geom_smooth(method="lm",color="black") +
  ylab("Mean Area") + xlab("Tree height") + labs(title="(d)")
grid.arrange(L3.1,L3.2,L3.3,L3.4,ncol=2)

# No spaghetti plots or lattice plots since leaves and
#   branches randomly selected at levels one and two

##############################################################

### Model fitting ###
model1 <- lmer(Area~ 1 + (1|Branchnum) + (1|Tree),
               REML=T, data=leaves)
summary(model1)

model2 <- lmer(Area~ Strata + Species + Zone +
  (1|Branchnum) + (Strata|Tree), REML=T, data=leaves)
summary(model2)

model2a <- lmer(Area~ Strata + Species + Zone +
  (1|Branchnum) + (1|Tree), REML=T, data=leaves)
summary(model2a)

model3 <- lmer(Area~ Strata + Species + Zone + Species:Zone +
  (1|Branchnum) + (Strata|Tree), REML=T, data=leaves)
summary(model3)

model3a <- lmer(Area~ Strata + Species + Zone + Strata:Zone +
  (1|Branchnum) + (Strata|Tree), REML=T, data=leaves)
```

```
summary(model3a)

# Final Model
model4 <- lmer(Area~ Strata + Species + Zone + Species:Zone +
  Strata:Zone + (1|Branchnum) + (Strata|Tree),
  REML=T, data=leaves)
summary(model4)

anova(model3,model4)
anova(model3a,model4)

# Compare model2 vs model2a using parametric bootstrap
anova(model2a,model2)   # fill in observed LRT below (7.716)

# Parametric bootstrap code for lme4-models
#   from Fabian Scheipl on stack exchange

#m0 is the lmer model under the null hypothesis (smaller model)
#mA is the lmer model under the alternative

bootstrapAnova <- function(mA, m0, B=1000){
  oneBootstrap <- function(m0, mA){
    d <- drop(simulate(m0))
    m2 <-refit(mA, newresp=d)
    m1 <-refit(m0, newresp=d)
    return(anova(m2,m1)$Chisq[2])
  }
  nulldist <- replicate(B, oneBootstrap(m0, mA))
  ret <- anova(mA, m0)
  ret$"Pr(>Chisq)"[2] <- mean(ret$Chisq[2] < nulldist)
  names(ret)[8] <- "Pr_boot(>Chisq)"
  attr(ret, "heading") <- c(attr(ret, "heading")[1],
    paste("Parametric bootstrap with", B,"samples."),
    attr(ret, "heading")[-1])
  attr(ret, "nulldist") <- nulldist
  return(ret)
}

# run bootstrapAnova function first
bRLRT = bootstrapAnova(mA=model2, m0=model2a, B=100)
bRLRT
nullLRT = attr(bRLRT,"nulldist")
x=seq(0,max(nullLRT),length=100)
y=dchisq(x,5)
```

```
nullLRT.1 <- as.data.frame(cbind(nullLRT=nullLRT,x=x,y=y))
ggplot(nullLRT.1) +
  geom_histogram(aes(x=nullLRT,y=..density..),binwidth=1,
                 color="black",fill="white") +
  geom_vline(xintercept=7.716,size=1) +
  geom_line(aes(x=x,y=y)) +
  labs(
x="Likelihood Ratio Test Statistics from Null Distribution",
y="Density")
sum(nullLRT>=7.716)/100


# Other potential directions for analysis (very rough analyses)

# careful - not exactly 15 observations per strata
stomatarows=seq(1,1960,by=15)
stomata=leaves[stomatarows,]
dim(stomata)
head(stomata)

models1=lmer(Stomata~Strata*Species*Zone+Tree.height+(1|Tree),
             REML=F, data=stomata)
summary(models1)

models2=lmer(Stomata~Strata*Zone+Species+Tree.height+(1|Tree),
             REML=F, data=stomata)
summary(models2)

models3=lmer(Stomata~Strata+Zone+Species+(1|Tree),
             REML=F, data=stomata)
summary(models3)

stomataC=stomata[stomata$Species=="Cassia fistula",]
modelsc1=lmer(Stomata~Strata*Zone+Tree.height+(1|Tree),
              REML=F, data=stomataC)
summary(modelsc1)

modelsc2=lmer(Stomata~Strata+Zone+Tree.height+(1|Tree),
              REML=F, data=stomataC)
summary(modelsc2)

library(pscl)
herb1=zeroinfl(Herbivory~Strata+Species+Zone+Tree.height,
```

```
                data=stomata)
summary(herb1)
```

An initial analysis of missing data shows that branch length is the only problematic variable in this rich data set spanning 1960 leaves from 45 trees (5 trees representing each combination of species and zone). Initial histograms show that our primary response variable, estimated leaf surface area, is slightly right skewed, but not to the extent it should be problematic to fit normal models to error terms.

Initial multilevel models show that 72% of total variability in surface areas occurs at Level Three (between trees), while only 10% occurs between different branches from the same tree, and 18% occurs between different leaves from the same branch.

Our final multilevel model focused on the effects on leaf surface area of tree species, climate zone in which the tree was located, and branch height (upper, middle, or lower strata of the tree). Drop-in-deviance tests showed the significance of both the species-by-zone interaction (chi-square statistic = 12.527, df=4, p=.014) and the strata-by-zone interaction (chi-square statistic = 11.897, df=4, p=.018). A parametric bootstrap showed marginal significance for the inclusion of error terms and their associated covariance terms on the Level Three equations for strata effects (approximate p-value .07); we noted that the likelihood ratio test based on the chi-square distribution provided a conservative test in this case (test statistic = 7.716, df=5, p=.173).

According to our final model, golden shower leaves are significantly larger than leaves from the other two species; this difference is especially notable in the dry deciduous zone, although still present in the other two climate zones. Surface areas in the dry thorn climate tend to be smaller than in other climates, except that the difference is mitigated with mountain ebonies. In the dry deciduous zone, leaves from the middle strata of branches are smaller than leaves from the lower strata, but in the other two climate zones leaves from the middle and upper strata of branches are larger than leaves from the lower strata.

Future analyses could control for characteristics of trees or branches, or, more likely, they could focus on alternative response variables such as petiole length, petiole width, or herbivory (which would likely need a zero-inflated model). Like coneflowers and leadplants, we could also analyze tree species separately to determine how individual species experience different growth in different climate zones.

11

Multilevel Generalized Linear Models

```
# Packages required for Chapter 11
library(gridExtra)
library(lme4)
library(pander)
library(ggmosaic)
library(knitr)
library(kableExtra)
library(tidyverse)
```

11.1 Exercises

11.1.1 Conceptual Exercises

1. Response = number of drinks during past weekend. Level one = student; level two = college. Research question: are individual student characteristics or attributes of their college more predictive of drinking behavior?

2. **College basketball referees.** In plot (a), if you draw a vertical line at a certain foul differential (say +1), then the length of that line in the dark region represents the empirical probability that a foul called when the foul differential is +1 is on the home team; the length of the vertical line in the lighter region then represents the empirical probability that a foul in the same situation is on the visitors. In plot (d), each point represents the empirical log-odds that a foul is called on the home team at a certain foul differential (say +1) – i.e. the log of the probability of a home team foul divided by the probability of a foul on the visitors. We see as the foul differential increases, the probability of a home foul decreases, and the log-odds of a home foul decreases in a linear fashion.

3. Yes. First, we must test for statistical significance. Second, we must adjust

for potential confounding variables, such as game score, type of foul, teams involved, etc.

4. A multilevel model allows all parameters to be estimated simultaneously, so that effects of individual games can borrow strength from information in other games (so they don't end up so extreme), and we can model correlation structure within individual games or teams.

5. If two random effects are crossed, levels of one random effect can occur at different levels of the second random effect. But if two random effects are nested, a certain level of one random effect can only occur in one specific level of the second random effect.

6. Assume that we collect the performances of students from both charter and public non-charter schools across all their classes, and we would like to include a random effect for teacher. Then student is nested in school, but student and teacher would be crossed random effects, since each student takes classes from multiple teachers and each teacher has multiple classrooms full of students.

7. (a) z_i is the effect of Game i on the coefficient for foul differential (e.g. is the effect of foul differential greater or smaller than average in Game i?). r_h is the effect of Home Team h on the coefficient for foul differential (e.g. is the effect of foul differential greater or smaller than average for Home Team h?). s_g is the effect of Visiting Team g on the coefficient for foul differential (e.g. is the effect of foul differential greater or smaller than average for Visiting Team g?). (b) We'd need 3 new parameters—all variance terms: σ_z^2, σ_r^2, σ_s^2.

8. No, we cannot use a likelihood ratio test to compare two models that are not nested.

9. Here are the basic steps for running a parametric bootstrap procedure to compare Model A1 with Model A3:

Parametric bootstrap process:

- Set fixed effects ($\widehat{\alpha}_0 = -.189$, $\widehat{\beta}_0 = -.268$) and variance components ($\widehat{\sigma}_u = .522$) under the null model.
- Sample $u_i \sim N(0, .522)$ for each i.
- Solve for $\widehat{p}_{ij}$ in $\log(\frac{\widehat{p}_{ij}}{1-\widehat{p}_{ij}}) = -.189 - .268\text{foul.diff}_{ij} - u_i$.
- Generate a 0 or 1 for Y_{ij} by randomly sampling from a Bernoulli distribution with probability $\widehat{p}_{ij}$ (i.e., flip a weighted coin, where $\widehat{p}_{ij}$ is the weight).
- Follow the above steps to generate a new set of 4972 binary observations.
- Fit both the null model (A1) and the alternative model (A3) to the data.

- Calculate a likelihood ratio test statistic = 2 x [max log likelihood under larger model – max log likelihood under null model].
- Repeat the above steps 1000 times.
- Plot the 1000 simulated LRT statistics.
- P-value = proportion of simulated LRT statistics greater than the observed LRT statistic (16.074). Note: for a confidence interval for, say, α_0 based on the null model, save the estimates of α_0 from each of your 1000 bootstrapped data sets, and form a confidence interval (could take middle 95% of estimates, or could use estimates to assess standard error, or there are many other possible approaches)

Specific parametric bootstrap example:

- $\hat{\alpha}_0 = -.189$, $\hat{\beta}_0 = -.268$, and $\hat{\sigma}_u = .522$
- $u_1 = -.314$ (based on random draw from a normal distribution with mean 0 and SD .522) and $\text{foul.diff}_{11} = 0$
- $\log(\frac{p_{11}}{1-p_{11}}) = -.189 - .268(0) - .314 = -.503$ so that $p_{11} = (\frac{e^{-.503}}{1+e^{-.503}}) = .377$
- select Y_{11} from a Bernoulli distribution with $p = .377$
- since all fixed and random effects are constant within a game, all that changes from foul to foul is foul.diff_{ij}. Thus, for the second foul from Game 1, we have $\log(\frac{p_{12}}{1-p_{12}}) = -.189 - .268(-1) - .314 = -.235$ and $p_{12} = (\frac{e^{-.235}}{1+e^{-.235}}) = .442$. So Y_{12} will be randomly generated from a Bernoulli distribution with $p = .442$
- continue in this manner for all 4972 observations

10. The chi-square distribution will produce p-values that are too large.

11. That would have required 3 additional variance components—correlations between the random effects for intercept and the effect of foul differential associated with games, home teams, and visiting teams.

12. We allow the log-odds of a foul on the home team to vary by game, by home team, and by visiting team, after controlling for foul differential, score differential, time remaining, and type of foul. But we assume the effects of these covariates are fixed across game, home team, and visiting team.

13. Because foul differential interacts with type of foul and time remaining, the coefficient for foul differential only applies when type of foul is at its reference

level (shooting fouls) and time=0 (end of the half). Otherwise, we'd have to factor in the interaction terms into the effect of foul differential.

14. The non-linear bump comes from the addition of an indicator variable for whether the home team has the lead (and its interaction with time). If the score is tied (score.diff=0 and lead.home=0) and there are 10 minutes left in the half (time=10), then an extra point scored by the home team (score.diff increases by 1 and lead.home changes from 0 to 1) is associated with an increase in log odds of a foul on the home team of .034-.150+.026(10)=.144. If, instead, the home team is ahead by 2 (score.diff=2 and lead.home=1) with 10 minutes left, then an extra point for the home team (score.diff increases by 1 but lead.home remains at 1) is associated with an increase in log odds of a foul on the home team of .034.

15. For a shooting foul, the effect of an extra foul on the visiting team (foul.diff decreases by 1) is: 1/exp(-.172-.0087x10) = 1.296. For an offensive foul, the effect of an extra foul on the visiting team (foul.diff decreases by 1) is: 1/exp(-.172-.0087x10-.103) = 1.436. Note that 1.436/1.296 = 1.109.

16. With a foul differential of 2, an increase of 1 minute in time remaining is associated with a decrease of 2.6% in the odds of a home team foul, while with a foul differential of -2, an increase of 1 minute in time remaining is associated with an increase of 0.9% in the odds of a home team foul. Note that exp(-.00871-.00869x2) = .974 and exp(-.00871-.00869x(-2)) = 1.009.

17.

- $\exp(\widehat{\phi}_0)$ = exp(-.00871) = .9913 and 1/.9913 = 1.00875. As the time remaining increases by 1 minute, the odds of a home foul increases by 0.875% (or the odds increase by 9.1% (1/exp(-.0871) = 1.091) as time remaining increases by 10 minutes). This interpretation applies to situations where the home team is not leading and fouls are equal between the home and visitors, after controlling for score differential and type of foul.

- $\exp(\widehat{\kappa}_0)$ = exp(-.080943) = .922 and 1/.922 = 1.084. The odds of a home foul are 8.4% higher for shooting fouls compared to offensive fouls at the point where an equal number of fouls have been called on the two teams, after controlling for score differential, time remaining, and whether the home team has the lead.

- $\exp(\widehat{\xi}_0)$ = exp(.02595x10) = 1.296. The effect of the home team gaining the lead on the odds of a home team foul increases by 29.6% for each extra 10 minutes in time remaining, after controlling for foul differential and type of foul. For example, with 5 minutes remaining, if the home team scores 2 points to take the lead (lead.home changes from 0 to 1 and score.diff increases by 2), then the odds of a home team foul increases by 4.8%, while the same basket to take a lead with 15 minutes remaining increases the odds

of a home team foul by 35.8%. Note that exp(-.150+.0335x2+.02595x5) = 1.048 and exp(-.150+.0335x2+.02595x15) = 1.358.

18. The baseline odds for DePaul is a random effect; we assume it was chosen from a normal distribution with mean 0 and standard deviation $\hat{\sigma}_v = 0.28$ (an estimated model parameter).

19. **Heart attacks in Aboriginal Australians.** [Randall et al., 2014]. Level One = subgroup within SLA; Level Two = SLA.

20. Level One:

$$\begin{aligned}\log(\lambda_{ij}) = a_i &+ b_i \text{aborig}_{ij} + c_i \text{age35}-44_{ij} + d_i \text{age45}-54_{ij} \\ &+ f_i \text{age55}-64_{ij} + k_i \text{age65}-74_{ij} + l_i \text{age75}-84_{ij} \\ &+ m_i \text{female}_{ij} + n_i \text{year03}_{ij} + o_i \text{year04}_{ij} \\ &+ q_i \text{year05}_{ij} + r_i \text{year06}_{ij} + s_i \text{year07}_{ij} + log(\text{population}_{ij})\end{aligned}$$

- Level Two:

$$\begin{aligned}a_i &= \alpha_0 + u_i \\ b_i &= \beta_0 \\ c_i &= \gamma_0 \\ d_i &= \delta_0 \\ f_i &= \phi_0 \\ k_i &= \kappa_0 \\ l_i &= \lambda_0 \\ m_i &= \mu_0 \\ n_i &= \nu_0 \\ o_i &= \omega_0 \\ q_i &= \xi_0 \\ r_i &= \psi_0 \\ s_i &= \chi_0,\end{aligned}$$

where error terms at Level Two can be assumed to follow a normal distribution: $u_i \sim N(0, \sigma_u^2)$.

There are 14 parameters to estimate – 13 fixed effects and 1 variance component. Note that the Level One equation has no separate error term, but it does have an offset term for the population of each subgroup within each SLA.

21.

- The rate of AMI events is 2.10 times higher for Aboriginal Australians compared to non-Aboriginal Australians, after adjusting for age, sex, and year of event.

- We are 95% confident that the rate of AMI events is between 1.98 and 2.23 times higher for Aboriginal Australians than non-Aboriginal Australians, after adjusting for age, sex, and year of event.

- There is significant evidence (p<.01) that the rate of AMI events differs among age groups, after adjusting for sex, year, and Aboriginal status.

- We are 95% confident that the rate of AMI events is between 5.44 and 6.64 times higher among those aged 35 to 44 compared to those aged 25 to 34, after adjusting for sex, year, and Aboriginal status.

- The rate of AMI events in females is 45% that of males (i.e., 55% lower), after adjusting for age, year, and Aboriginal status.

- We are 95% confident that the rate of AMI events in 2007 is between 86% and 91% the rate in 2002 (i.e., between 9% and 14% lower), after adjusting for age, sex, and Aboriginal status.

22. $\exp(\hat{\beta}_0) = 2.10 \Rightarrow \hat{\beta}_0 = \log(2.10) = 0.74$. Similarly, (log(1.98), log(2.23)) = (.68, .80) = $(\hat{\beta}_0 \pm 1.96 * SE) \Rightarrow$ SE = .03.

23. Since two nested models would have been compared, a likelihood ratio test or parametric bootstrap test might have been used.

24. 11 interaction terms would have to have been added to the Level One equation—5 interacting Aboriginal status with the 5 indicator variables for different age groups, 1 interacting Aboriginal status with the indicator variable for females, and 5 interacting Aboriginal status with the 5 indicator variables for different years.

25. Two new models would have been created. First, 3 indicator variables would have been added to the equation for a_i at Level Two, representing 3 levels of remoteness (other than major city). Second, after removing the 3 indicator variables for remoteness, 4 indicator variables would have been added to the equation for a_i at Level Two, representing 4 levels of SES.

26.

- There is significant evidence (p<.01) that the rate of AMI events differs among remoteness levels, after adjusting for age, sex, year, and Aboriginal status.

- The rate of AMI events is 22% higher for residents of remote and very remote areas compared to residents of major cities, after adjusting for age, sex, year, and Aboriginal status.

- We are 95% confident that the rate of AMI events is between 52% and 91% higher for residents of the most economically disadvantaged areas compared to residents of the least disadvantaged areas, after adjusting for age, sex, year, and Aboriginal status.

27. The difference between Aboriginal Australians and non-Aboriginals is greater before we account for variability among SLAs.

28. They are referring to estimating the effects of SLAs using random effect estimates, which tend to be partway between the "crude" empirical estimate of the SLA and the average from similar subgroups from all SLAs.

11.1.2 Open-Ended Exercises

1. **Airbnb in Chicago.** [Trinh and Ameri, 2018]

```
airbnb <- read_csv("data/airbnb.csv")

airbnb <- airbnb %>%
  mutate(satisfaction = ifelse(overall_satisfaction==5, 1, 0),
    satisfact = ifelse(satisfaction==1, "5", "Below 5"),
    logprice = log(price),
    HighBlack = ifelse(PctBlack > .60, 1, 0),
    cWalkScore = WalkScore - mean(WalkScore),
    cTransit = TransitScore - mean(TransitScore),
    cBikeScore = BikeScore - mean(BikeScore),
    EntireUnit = ifelse(room_type == "Entire home/apt", 1, 0),
    PrivateRoom = ifelse(room_type == "Private room", 1, 0),
    SharedRoom = ifelse(room_type == "Shared room", 1, 0))

# EDA
summary(airbnb)

ggplot(data = airbnb, aes(x = logprice)) +
  geom_density(aes(fill = satisfact), position = "fill",
               adjust = 2, alpha = 0.5)
ggplot(data = airbnb, aes(x = bedrooms)) +
  geom_density(aes(fill = satisfact), position = "fill",
               adjust = 2, alpha = 0.5)
```

```
ggplot(data = airbnb, aes(x = reviews)) +
  geom_density(aes(fill = satisfact), position = "fill",
               adjust = 2, alpha = 0.5)
ggplot(data = airbnb, aes(x = accommodates)) +
  geom_density(aes(fill = satisfact), position = "fill",
               adjust = 2, alpha = 0.5)
ggplot(data = airbnb, aes(x = minstay)) +
  geom_density(aes(fill = satisfact), position = "fill",
               adjust = 2, alpha = 0.5)
ggplot(data = airbnb, aes(x = PctBlack)) +
  geom_density(aes(fill = satisfact), position = "fill",
               adjust = 2, alpha = 0.5)
ggplot(data = airbnb, aes(x = WalkScore)) +
  geom_density(aes(fill = satisfact), position = "fill",
               adjust = 2, alpha = 0.5)
ggplot(data = airbnb, aes(x = BikeScore)) +
  geom_density(aes(fill = satisfact), position = "fill",
               adjust = 2, alpha = 0.5)
ggplot(data = airbnb, aes(x = TransitScore)) +
  geom_density(aes(fill = satisfact), position = "fill",
               adjust = 2, alpha = 0.5)

ggplot(data = airbnb, aes(x = room_type)) +
  geom_bar(aes(fill = satisfact), position = "fill")
ggplot(data = airbnb, aes(x = as.factor(HighBlack))) +
  geom_bar(aes(fill = satisfact), position = "fill")
ggplot(data = airbnb, aes(x = as.factor(minstay>1))) +
  geom_bar(aes(fill = satisfact), position = "fill")

# Initial models
model0 <- glmer(satisfaction~1+(1|district), family=binomial,
                data=airbnb)
summary(model0)

model0a <- glmer(satisfaction~1+(1|neighborhood),
                 family=binomial, data=airbnb)
summary(model0a)

# Potential model for binary satisfaction response (Ch 11)
gfinal1 <- glmer(satisfaction ~ PrivateRoom + logprice +
  bedrooms + HighBlack + (1 | neighborhood),
  family = binomial, data = airbnb)
summary(gfinal1)
exp(coef(summary(gfinal1))[,1])
```

```
gfinal2 <- glmer(satisfaction ~ PrivateRoom + logprice +
  bedrooms + HighBlack + PrivateRoom:logprice +
  logprice:bedrooms + (1 | neighborhood),
  family = binomial, data = airbnb)
summary(gfinal2)
```

Our final model revealed 3 significant covariates at Level One (individual unit)—room type, bedrooms, and logged price—and 1 at Level Two (neighborhood)—whether proportion of black residents is above 60%. The odds of highest satisfaction (5 out of 5) are 2.18 times higher in private rooms than in shared rooms or the entire unit, after controlling for price, bedrooms, and proportion of black residents. A doubling of price is associated with a 77.3% increase in the odds of highest satisfaction, after controlling for room type, bedrooms, and the proportion of black residents. Each additional bedroom is associated with a 24.4% decrease in the odds of highest satisfaction, after controlling for room type, price, and the proportion of black residents. Finally, units in neighborhoods with over 60% black residents have 50.6% lower odds of highest satisfaction, after controlling for room type, bedrooms, and price.

There is also evidence of interactions between room type and price, and between bedrooms and price. The effect of private rooms is greater at higher prices, and the effect of bedrooms is less negative at higher prices.

2. **Seed germination.** [Angell, 2010]

```
seedwd <- read_csv("data/seeds2.csv")
dim(seedwd)
seedwd[135:146,2:11]   # illustrate wide data

# create indicator variables for later analyses
seedwd <- seedwd %>%
  mutate(cult=ifelse(soil=="CULT",1,0),
         rem=ifelse(soil=="REM",1,0),
         stp=ifelse(soil=="STP",1,0),
         lead=ifelse(species=="L",1,0),
         cone=ifelse(species=="C",1,0),
         strl=ifelse(sterile=="Y",1,0),
         nostrl=ifelse(sterile=="N",1,0),
         germ=ifelse(germin=="Y",1,0))

# create separate data sets of leadplants and coneflowers
conedata <- filter(seedwd, cone==1)
```

```
leaddata <- filter(seedwd, lead==1)

## Exploratory data analysis ##

# Rough look at Level Two covariates
table(seedwd$species)
table(seedwd$soil)
table(seedwd$sterile)
table(seedwd$germ)
prop.table(table(seedwd$species))
prop.table(table(seedwd$soil))
prop.table(table(seedwd$sterile))
prop.table(table(seedwd$germ))

ggplot(data = seedwd, aes(x = species)) +
  geom_bar(aes(fill = germin), position = "fill")

# From here on, do everything separately for
#   leadplants and coneflowers

ggplot(data = conedata, aes(x = soil)) +
  geom_bar(aes(fill = germin), position = "fill")
ggplot(data = conedata, aes(x = sterile)) +
  geom_bar(aes(fill = germin), position = "fill")
ggplot(data = conedata, aes(x = soil)) +
  geom_bar(aes(fill = germin), position = "fill") +
  facet_wrap(~ sterile)

ggplot(data = leaddata, aes(x = soil)) +
  geom_bar(aes(fill = germin), position = "fill")
ggplot(data = leaddata, aes(x = sterile)) +
  geom_bar(aes(fill = germin), position = "fill")
ggplot(data = leaddata, aes(x = soil)) +
  geom_bar(aes(fill = germin), position = "fill") +
  facet_wrap(~ sterile)

############################################################

# Model A - unconditional means
modelal = glmer(germ ~ 1 + (1|pot), family=binomial,
  data=leaddata)
summary(modelal)

modelac = glmer(germ ~ 1 + (1|pot), family=binomial,
```

```
  data=conedata)
summary(modelac)

# Add covariates at pot level
modelcl = glmer(germ ~ strl + cult + rem + cult:strl +
  rem:strl + (1|pot), family=binomial, data=leaddata)
summary(modelcl)

modelcl0 = glmer(germ ~ strl + cult + rem +
  (1|pot), family=binomial, data=leaddata)
summary(modelcl0)   # Final leadplant model

modelcc = glmer(germ ~ strl + cult + rem + cult:strl +
  rem:strl + (1|pot), family=binomial, data=conedata)
summary(modelcc)  # Final coneflower model

modelcc0 = glmer(germ ~ strl + cult + rem +
  (1|pot), family=binomial, data=conedata)
summary(modelcc0)

anova(modelcl0,modelcl)   # go with Model CL (with interactions)
# could also try parametric bootstrap
```

Initial exploratory analyses show that overall germination rates are higher for coneflowers and higher for restored prairie soil (STP) for both species. Sterilization appears to have no effect on the germination of leadplants, but coneflowers have a higher germination rate in non-sterilized soil (87.6% vs 71.3%). Finally, there is graphical evidence of differing interactions between soil type and sterilization in coneflowers and leadplants—in coneflowers, remnant soil produces lower germination rates in sterilized soil, while sterilization has relatively little impact in the other two soil types; in leadplants, germination rates in cultivated soil are higher if the soil is sterilized, while germination rates in remnant soil are lower if the soil is sterilized.

Our final multilevel model for coneflowers contains only main effects for sterilization and soil type. Odds of germination are 3.74 times higher in non-sterilized soil (p=.049), holding soil type constant. Odds of germination in restored soil are 66.4 times higher than in cultivated soil (p<.001) and 34.1 times higher than in remnant soil (p=.004), holding sterilization constant. Note that parameter estimates are a bit extreme, likely due to the high germination rate in restored soil.

Our final multilevel model for leadplants contains interaction terms for sterilization by soil type. The effect of soil type is significantly different depending on whether the soil is sterilized (likelihood ratio test for interaction: LRT

= 14.5, p<.001). Odds of germination in restored soil are 2.42 times higher than in cultivated soil, if the soil was sterilized, and 35.9 times higher than in cultivated soil that has not been sterilized. Similarly, odds of germination in restored soil are 10.0 times higher than in remnant soil, if the soil was sterilized, and 5.34 times higher than in remnant soil that has not been sterilized.

Since this was a randomized experiment, we can conclude that differences in germination rate of coneflowers and leadplants were caused by sterilization and soil type, although we must be careful about generalizing results which occurred in carefully controlled greenhouse conditions with the effects of sterilization and soil type that would be seen in nature.

3. **Book banning.** [Fast and Hegland, 2011].

```
noTex.df <- read_csv("data/bookbanningNoTex.csv")
noTex.df

length(unique(noTex.df$author))
length(unique(noTex.df$book))
length(unique(noTex.df$state))
sort(table(noTex.df$state),decr=T)
sort(table(noTex.df$author),decr=T)[1:20]
sort(table(noTex.df$booktitle),decr=T)[1:20]
table(noTex.df$sexexp)
table(noTex.df$antifamily)
table(noTex.df$violence)
table(noTex.df$occult)
table(noTex.df$language)
table(noTex.df$homosexuality)
table(noTex.df$freqchal)
table(noTex.df$removed)

hist(noTex.df$pvi2); summary(noTex.df$pvi2)
hist(noTex.df$cperhs); summary(noTex.df$cperhs)
hist(noTex.df$cmedin); summary(noTex.df$cmedin)
hist(noTex.df$cperba); summary(noTex.df$cperba)

# Proportions of successful challenges
by(noTex.df$removed,noTex.df$sexexp,mean)
by(noTex.df$removed,noTex.df$homosexuality,mean)
by(noTex.df$removed,noTex.df$antifamily,mean)
by(noTex.df$removed,noTex.df$language,mean)
by(noTex.df$removed,noTex.df$occult,mean)
by(noTex.df$removed,noTex.df$violence,mean)
```

```
prop.table(table(removed=noTex.df$removed,
  freqchal=noTex.df$freqchal),2)
ggplot(data = noTex.df, aes(x = as.factor(freqchal))) +
  geom_bar(aes(fill = as.factor(removed)), position = "fill")

prop.table(table(removed=noTex.df$removed,
  obama=noTex.df$obama),2)
ggplot(data = noTex.df, aes(x = as.factor(obama))) +
  geom_bar(aes(fill = as.factor(removed)), position = "fill")

prop.table(table(removed=noTex.df$removed,
  sexexp=noTex.df$sexexp),2)
ggplot(data = noTex.df, aes(x = as.factor(sexexp))) +
  geom_bar(aes(fill = as.factor(removed)), position = "fill")

prop.table(table(removed=noTex.df$removed,
  sexexp=noTex.df$sexexp,obama=noTex.df$obama),c(2,3))
ggplot(data = noTex.df, aes(x = as.factor(sexexp))) +
  geom_bar(aes(fill = as.factor(removed)), position = "fill") +
  facet_wrap(~ as.factor(obama))

ggplot(data = noTex.df, aes(x = pvi2)) +
  geom_density(aes(fill = as.factor(removed)),
               position = "fill", adjust = 2, alpha = 0.5)
ggplot(data = noTex.df, aes(x = cperhs)) +
  geom_density(aes(fill = as.factor(removed)),
               position = "fill", adjust = 2, alpha = 0.5)

ggplot(data = noTex.df, aes(x = pvi2)) +
  geom_density(aes(fill = as.factor(removed)),
               position = "fill", adjust = 2, alpha = 0.5) +
  facet_wrap(~ as.factor(freqchal))

# Models

nottex1 <- glmer(removed ~ 1 +
  (1|book) + (1|state), data=noTex.df, family=binomial)
summary(nottex1)

nottex2 <- glmer(removed~freqchal + pvi2 + sexexp +
  obama + cperhs + sexexp:obama + freqchal:pvi2 +
  (1|book) + (1|state), data=noTex.df, family=binomial)
summary(nottex2)
```

```
nottex3 <- glmer(removed~freqchal + pvi2 + obama +
  cmedin + cperba + cperhs + violence + occult + sexexp +
  antifamily + language + homosexuality + (1|book) + (1|state),
  data=noTex.df, family=binomial)
summary(nottex3)

nottex4 <- glmer(removed~freqchal + pvi2 + obama +
  cperhs + antifamily + (1|book) + (1|state),
  data=noTex.df, family=binomial)
summary(nottex4)

nottex4a <- glmer(removed~freqchal + pvi2 + obama +
  cperhs + antifamily + (1|book) + (obama|state),
  data=noTex.df, family=binomial)
summary(nottex4a)

# A "final" model
nottex5 <- glmer(removed~freqchal + pvi2 + obama +
  cperhs + freqchal:pvi2 +
  sexexp + sexexp:obama + (1|book) + (1|state),
  data=noTex.df, family=binomial)
summary(nottex5)

nottex5a <- glmer(removed~freqchal + pvi2 + obama +
  cperhs + freqchal:pvi2 +
  sexexp + sexexp:obama + (1|book) + (1|state) + (1|author),
  data=noTex.df, family=binomial)
summary(nottex5a)
```

After setting aside challenges from Texas, we have 931 challenges from 47 states related to 694 unique book titles and 571 unique authors. Of the 931 challenges, 217 were successful (or 23.3%). The most common reasons given for challenges include sexually explicit material (299 challenges), inappropriate language (237), and violent material (134).

Exploratory analysis examined relationships between potential predictors and the probability of a challenge being successful (resulting in removal of a book). For example, the reasons for challenges that produced the highest success rates included: violence (32.8%), inappropriate language (31.6%), and sexually explicit material (28.4%). Whether a book was frequently challenged and whether the challenge occurred before the Obama Presidency were both also associated with higher success rates. Among state-level variables, there is evidence in conditional density plots that higher levels of PVI and proportion with high school degrees were both associated with lower success rates. Fi-

nally, there is visual evidence of a couple of potential interactions. The effect of sexually explicit challenges is much more dramatic during the Obama Presidency, and the effect of PVI is much more dramatic with frequently challenged authors.

Our final multilevel model (accounting for the crossed random effects of book and state) suggested that characteristics of the context in which a challenge was made were more predictive of the success of a challenge than the specific reason for the challenge. In general, challenges were more successful during the Bush Presidencies (at least when challenges were not for sexually explicit material), when frequently challenged authors were involved, and in states that leaned Republican (especially among frequently challenged authors) and had lower percentages of high school graduates.

More precisely, for every extra 1 percent of high school graduates in a state, the odds of a successful challenge decreases by 7.8%, after controlling for PVI, whether an author is frequently challenged, whether the challenge was during the Obama Presidency, and whether the challenge was over sexually explicit material. For a state leaning Democrat (say PVI of +10), the odds of a successful challenge are 47.4% lower if a book is frequently challenged, while for a state leaning Republican (say PVI of -10), the odds of a successful challenge are 5.35 times greater if a book is frequently challenged, after controlling for the percentage of high school graduates in a state, whether the challenge was during the Obama Presidency, and whether the challenge was over sexually explicit material. Finally, challenges for non-sexually explicit material had odds of success that were 70.5% lower if made during the Obama Presidency rather than the Bush Presidency, while challenges for sexually explicit material had odds of success that were 20.4% greater during the Obama era vs. the Bush era, after controlling for PVI, whether an author is frequently challenged, and the percentage of high school graduates in a state.

4. **Yelp restaurant reviews.** [Janusz and Mohr, 2018]

```
# Data wrangling
yelp <- read_csv("data/yelp.csv")
yelp

yelp <- yelp %>%
  mutate(usefulYN = ifelse(useful > 0, "Yes", "No"),
    stars5 = ifelse(stars_given == 5, "5 stars", "1-4 stars"),
    useful01 = ifelse(useful > 0, 1, 0),
    stars01 = ifelse(stars_given == 5, 1, 0),
    years_on_yelp = days_on_yelp / 365.25,
    length100 = length / 100,
    revs_biz100 = rev_count_biz / 100,
```

```
    revs_user100 = rev_count_user / 100)

temp <- yelp %>% count(name) %>% count(n)
print(temp, n=Inf)
temp <- yelp %>% count(user_id) %>% count(n)
print(temp, n=Inf)

yelp2 <- yelp %>% group_by(user_id) %>% filter(n()>1)
yelp3 <- yelp %>% sample_n(5000)

#EDA
prop.table(table(stars5=yelp3$stars5,
  ParkingLot=yelp3$BusinessParking_lot),2)
ggplot(data = yelp3, aes(x = BusinessParking_lot)) +
  geom_bar(aes(fill = stars5), position = "fill")
prop.table(table(stars5=yelp3$stars5,
  ParkingLot=yelp3$BusinessParking_validated),2)
ggplot(data = yelp3, aes(x = BusinessParking_validated)) +
  geom_bar(aes(fill = stars5), position = "fill")
prop.table(table(stars5=yelp3$stars5,
  ParkingLot=yelp3$BusinessParking_street),2)
ggplot(data = yelp3, aes(x = BusinessParking_street)) +
  geom_bar(aes(fill = stars5), position = "fill")
prop.table(table(stars5=yelp3$stars5,
  ParkingLot=yelp3$BusinessParking_valet),2)
ggplot(data = yelp3, aes(x = BusinessParking_valet)) +
  geom_bar(aes(fill = stars5), position = "fill")

ggplot(data = yelp3, aes(x = length100)) +
  geom_density(aes(fill = stars5), position = "fill",
               adjust = 5, alpha = 0.5)
ggplot(data = yelp3, aes(x = useful)) +
  geom_density(aes(fill = stars5), position = "fill",
               adjust = 25, alpha = 0.5)
ggplot(data = yelp3, aes(x = funny)) +
  geom_density(aes(fill = stars5), position = "fill",
               adjust = 25, alpha = 0.5)
ggplot(data = yelp3, aes(x = cool)) +
  geom_density(aes(fill = stars5), position = "fill",
               adjust = 25, alpha = 0.5)
ggplot(data = yelp3, aes(x = revs_user100)) +
  geom_density(aes(fill = stars5), position = "fill",
               adjust = 20, alpha = 0.5)
ggplot(data = yelp3, aes(x = avg_user_stars)) +
```

```
  geom_density(aes(fill = stars5), position = "fill",
               adjust = 2, alpha = 0.5)
ggplot(data = yelp3, aes(x = years_on_yelp)) +
  geom_density(aes(fill = stars5), position = "fill",
               adjust = 2, alpha = 0.5)
ggplot(data = yelp3, aes(x = revs_biz100)) +
  geom_density(aes(fill = stars5), position = "fill",
               adjust = 20, alpha = 0.5)
ggplot(data = yelp3, aes(x = avg_biz_stars)) +
  geom_density(aes(fill = stars5), position = "fill",
               adjust = 4, alpha = 0.5)

#Models
mod1 <- glmer(stars01 ~ 1 + (1|user_id) + (1|name),
  data=yelp, family=binomial)
summary(mod1)

mod2 <- glmer(stars01 ~ 1 +
  length100 + useful + funny + cool +
  revs_user100 + avg_user_stars + years_on_yelp +
  revs_biz100 + avg_biz_stars + BusinessParking_valet +
    BusinessParking_validated + BusinessParking_street +
    BusinessParking_lot +
  (1|user_id) + (1|name), data=yelp3, family=binomial)
summary(mod2)

mod3 <- glmer(stars01 ~ 1 +
  length100 + useful + funny + cool +
  revs_user100 + avg_user_stars +
  revs_biz100 + avg_biz_stars +
  (1|user_id) + (1|name), data=yelp3, family=binomial)
summary(mod3)
exp(coef(summary(mod3))[,1])
```

Exploratory data analysis shows that attributes of the restaurant, the reviewer, and the review itself are all predictive of whether or not a review is 5 stars or not. Longer reviews were less likely to award 5 stars, as were reviews that more people rated as useful, funny, or cool (except for a handful of reviews where a lot of people rated them with those attributes). More 5 star reviews were associated with reviewers with fewer previous reviews and those with higher average ratings, while years on yelp appeared unrelated. Finally, more 5 star reviews were strongly positively associated with average restaurant rating, slightly positively associated with number of reviews for a restaurant, and not strongly associated with parking lot type.

One potential final model accounts for crossed random effects of reviewer and restaurant, while including covariates associated with the restaurant, the reviewer, and the review itself. The odds of a 5-star review decrease by 4.2% for each 100 extra characters in the review, after controlling for how many readers found a review useful or funny or cool, the average rating and number of previous reviews by the reviewer, and the average rating and number of previous reviews for the restaurant. (All other interpretations are based on controlling for the same set of covariates.) The odds of a 5-star review decrease by 7.4% for each extra reader who rates a review as useful, decrease by 14.5% for each extra reader who rates a review as funny, and increase by 38.1% for each extra reader who rates a review as cool. As expected, the odds of a 5-star review are 3.57 times greater for each 1 point increase in the average user rating, and 3.40 times greater for each 1 point increase in the average restaurant rating. However, the odds of a 5-star review decrease by 17.5% for each extra 100 previous reviews by the reviewer, and by 3.5% for each extra 100 previous reviews of the restaurant.

Further modeling could investigate potential interactions between covariates when modeling 5-star reviews, or the type of restaurant as determined in the `categories` field. We could also focus on a different response like whether at least one person felt a review was useful or cool.

Bibliography

Diane Angell. Effects of soil type and sterilization on the growth of coneflowers and leadplants. St. Olaf College. Class data for Biology 261, 2010.

Annika Awad, Evan Lebo, and Anna Linden. Intercontinental comparative analysis of Airbnb booking factors. St. Olaf College. Statistics 316 Project, 2017.

Robert Bickel. *Multilevel Analysis for Applied Research: It's Just Regression!* Guilford Publications, New York, 2007.

J. A. Bishop. An experimental study of the cline of industrial melanism in biston betularia (l.) (lepidoptera) between urban liverpool and rural north wales. *Journal of Animal Ecology*, 41(1):209–243, 1972. doi: 10.2307/3513.

Margaret Blakeman, Tim Renier, and Rami Shandaq. Modeling Donald Trump's voters in the 2016 Election. St. Olaf College. Statistics 316 Project, 2018.

H. Jane Brockmann. Satellite male groups in horseshoe crabs, limulus polyphemus. *Ethology*, 102(1):1–21, 1996. URL http://dx.doi.org/doi:10.1111/j.1439-0310.1996.tb01099.x.

A.C. Cameron and P.K. Trivedi. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, 1:29–53, 1986.

Christopher Chapp, Paul Roback, Kendra Jo Johnson-Tesch, Adrian Rossing, and Jack Werner. Going vague: Ambiguity and avoidance in online political messaging. *Social Science Computer Review*, Aug 2018. URL https://doi.org/10.1177/0894439318791168.

Patrick J. Curran, Eric Stice, and Laurie Chassin. The relation between adolescent alcohol use and peer alcohol use: A longitudinal random coefficients model. *Journal of Consulting and Clinical Psychology*, 65(1):130–140, 1997. URL http://dx.doi.org/10.1037/0022-006X.65.1.130.

Samantha Dahlquist and Jin Dong. The effects of credit cards on tipping. St. Olaf College. Statistics 272 Project, 2011.

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.

Brian S. Everitt and Torsten Hothorn. *A Handbook of Statistical Analyses using R*. Chapman & Hall/ CRC, Boca Raton, FL, 2006.

Anthony Farrar and Thomas H. Bruggink. A new test of the moneyball hypothesis. *The Sport Journal*, May 2011. URL http://thesportjournal.org/article/a-new-test-of-the-moneyball-hypothesis/.

Shannon Fast and Thomas Hegland. Book challenges: A statistical examination. St. Olaf College. Statistics 316 Project, 2011.

Lisa Fisher, Katie Murney, and Tyler Radtke. Emergency department overcrowding and factors that contribute to ambulance diversion. St. Olaf College. Statistics 316 Project, 2019.

I. B. Goldstein and D. Shapiro. Ambulatory blood pressure in women: Family history of hypertension and personality. *Psychology, Health & Medicine*, 5 (3):227–240, 2000. URL https://doi.org/10.1080/713690197.

Walt Hickey. The dollar-and-cents case against Hollywood's exclusion of women. FiveThirtyEight, Apr 2014. URL https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/.

P. A. Holst, D. Kromhout, and R. Brand. For debate: pet birds as an independent risk factor for lung cancer. *British Medical Journal*, 297(6659): 1319–1321, Nov 1988. doi: 10.1136/bmj.297.6659.1319.

Brooke Janusz and Michael Mohr. Predicting user Yelp star ratings based on restaurant attributes. St. Olaf College. Statistics 316 Project, 2018.

Kaggle. House sales in King County, USA, 2018a. URL https://www.kaggle.com/harlfoxem/housesalesprediction/home.

Kaggle. NBA enhanced box scores and standings, 2018b. URL https://www.kaggle.com/pablote/nba-enhanced-stats.

National Center for Education Statistics. The Integrated Postsecondary Education Data System, 2018. URL https://nces.ed.gov/ipeds/.

Joyce H. Poole. Mate guarding, reproductive success and female choice in African elephants. *Animal Behaviour*, 37:842–849, 1989. URL http://www.sciencedirect.com/science/article/pii/0003347289900687.

Ian Pray. Effects of rainfall and sun exposure on leaf characteristics. St. Olaf College. Bio in South India Project, 2009.

J Proudfoot, D Goldberg, A Mann, B Everitt, I Marks, and J A Gray. Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice. *Psychological Medicine*, 33(2):217–27, Feb 2003. doi: 10.1017/s0033291702007225.

Fred Ramsey and Daniel Schafer. *The Statistical Sleuth: A course in methods of data analysis*. Brooks/Cole Cengage, Boston, Massachusetts, 2nd edition, 2002.

D.A. Randall, L.R. Jorm, S. Lujic, S.J. Eades, T.R. Churches, A.J. O'Loughlin, and A.H. Leyland. Exploring disparities in acute myocardial infarction events between Aboriginal and non-Aboriginal Australians: Roles of age, gender, geography and area-level disadvantage. *Health & Place*, 28: 58–66, 2014. ISSN 1353-8292. doi: https://doi.org/10.1016/j.healthplace.2014.03.009.

Marieke Roskes, Daniel Sligte, Shaul Shalvi, and Carsten K. W. De Dreu. The right side? Under time pressure, approach motivation leads to right-oriented bias. *Psychology Science*, 22(11):1403–1407, 2011. URL https://doi.org/10.1177/0956797611418677.

Michael E. Sadler and Christopher J. Miller. Performance anxiety: A longitudinal study of the roles of personality and experience in musicians. *Social Psychological and Personality Science*, 1(3):280–287, 2010. URL http://dx.doi.org/10.1177/1948550610370492.

Prabha Siddarth, Alison C. Burggren, Harris A. Eyre, Gary W. Small, and David A. Merrill. Sedentary behavior associated with reduced medial temporal lobe thickness in middle-aged and older adults. *PLOS ONE*, 13(4): 1–13, Apr 2018. doi: 10.1371/journal.pone.0195549.

Judith D. Singer and John B. Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, Inc., New York, 1st edition, 2003.

Ly Trinh and Pony Ameri. Airbnb price determinants: A multilevel modeling approach. St. Olaf College. Statistics 316 Project, 2018.

UCLA Statistical Consulting Group. Zero-inflated negative binomial regression: R data analysis examples, 2018. URL https://stats.idre.ucla.edu/r/dae/zinb/.

Chanequa J. Walker-Barnes and Craig A. Mason. Ethnic differences in the effect of parenting on gang involvement and gang delinquency: A longitudinal, hierarchical linear modeling perspective. *Child Development*, 72(6): 1814–1831, 2001. URL http://dx.doi.org/10.1111/1467-8624.00380.

Robert E. Weiss. *Modeling Longitudinal Data*. Springer-Verlag, New York, 2005.

Wikipedia contributors. Kentucky Derby. In *Wikipedia*, 2018. URL https://en.wikipedia.org/wiki/Kentucky_Derby.