Elementary Statistical Modeling for Applied Biostatistics

Copyright 2018 Jeffrey A. Walker

Draft: 2019-09-27

Contents

Preface

More cynically, one could also well ask "Why has medicine not adopted frequentist inference, even though everyone presents P-values and hypothesis tests?" My answer is: Because frequentist inference, like Bayesian inference, is not taught. Instead everyone gets taught a misleading pseudo-frequentism: a set of rituals and misinterpretations caricaturing frequentist inference, leading to all kinds of misunderstandings. – Sander Greenland

We use statistics to learn from data with uncertainty. Traditional introductory textbooks in biostatistics implicitly or explicitly train students and researchers to "discover by p-value" using hypothesis tests (Chapter 6). Over the course of many chapters, the student learns to use something like a look-up table or a dichotomous key to choose the correct "test" for the data at hand, compute a test statistic for their data, compute a p-value based on the test statistic, and compare the p-value to 0.05. Textbooks typically give very little guidance about what can be concluded if $p < 0.05$ or if $p > 0.05$, but many researchers conclude, incorrectly, they have "discovered" something or "shown" an effect if $p < 0.05$ but found "no effect" if $p > 0.05$.

Researchers learn nothing useful from a hypothesis test – that is, comparing p to 0.05 (or some other arbitrary value). A p-value is a measure of compatibility between the data and the null hypothesis and, consequently, a pretty good, but imperfect tool to dampen the frequency that we are fooled by randomness. But if we are investigating the effects of an increasingly acidified ocean on coral growth, $p = 0.002$ may be evidence of an effect of the experimental intervention, but, from everything we know about pH and cell biology, it would be absurd to conclude from any data that pH does not affect growth. To build useful models of how biological systems work, we want to know the magnitude of effects and our uncertainty in estimating these magnitudes. We can use a magnitude and uncertainty to make predictions about the future of coral reefs, under different scenarios of ocean acidification. We can use the estimated effects and uncertainty to model the consequences of the effects of acidification on coral growth on fish production or carbon cycling.

This book is an introduction to the estimation of effects of biological data, and measures of the uncertainty of theses estimates, using a statistical modeling approach. As an introduction, the focus will be linear models and extensions of the linear models including linear mixed models and generalized linear models. Linear models are the engine behind many hypothesis tests but the emphasis in statistical modeling is estimation and uncertainty instead of test statistics and p-values. All linear models, and their generalizations, are variations of

$$y_i \sim N(\mu_i, \theta) \tag{1}$$
$$\mathrm{E}(Y|X) = \mu \tag{2}$$
$$\mu_i = f(\beta_0 + \beta_1 x_i) \tag{3}$$

Chapter 1 explains the meaning of this **model specification** but the point to make here is that because all linear models and their generalizations are variations of this specification, a modeling strategy of learning or doing statistics is more coherent than the NHST strategy using look-up tables or dichotomous keys of hypothesis tests. Generalizations of the basic linear model include linear mixed models, generalized linear models, generalized additive models, causal graphical models, multivariate models, and machine learning.

This book is not a comprehensive source for any of these methods but, instead, *a path of the critical elements leading you to the doorway to the vast universe of each of these methods.*

NHST Blues – The "discovery by p-value" strategy, or Null-Hypothesis Significance Testing (NHST), has been criticized by statisticians for many, many decades. Nevertheless, introductory biostatistics textbooks written by both biologists and statisticians continue to organize textbooks around a collection of hypothesis tests, with much less emphasis on estimation and uncertainty. The NHST strategy of learning or doing statistics is easy in that it requires little understanding of the statistical model underneath the tests and its assumptions, limitations, and behavior. The NHST strategy in combination with point-and-click software enables "mindless statistics"[1] and encourages the belief that statistics is a tool like a word processor is a tool, afterall, a rigorous analysis of one's data requires little more than getting p-values and creating bar plots. Indeed, many PhD programs in the biosciences require no statistics coursework and the only training available to students is from the other graduate students and postdocs in the lab. As a consequence, the biological sciences literature is filled with error bars that imply data with negative values and p-values that have little relationship to the probability of the data under the null. More importantly for science, the reported statistics are often not doing for the study what the researchers and journal editors think they are doing.

0.1 Math

0.2 R and programming

---

[1]Gegenrezer

Part I: R fundamentals

Chapter 1

Organization – R Projects and R Notebooks

1.1 Importing Packages

The R scripts you write will include functions in packages that are not included in Base R. These packages need to be downloaded from an internet server to your computer. You only need to do this once. But, each time you start a new R session, you will need to load a package using the `library()` function. Now is a good time to import packages that we will use

1. Open R Studio and choose the menu item "Tools" > "Install Packages"
2. In the "packages" input box, insert "ggplot2, data.table, emmeans, lme4, reshape2". Make sure that "install dependencies" is clicked before you click "Install"

Again, once these are installed, you don't need to do this again. You simply need to use the `library()` function at the start of a script.

1.2 Create an R Studio Project for this Class

1. Create a project folder within the Documents folder (Mac OS) or My Documents folder (Windows OS). All files associated with this book will reside inside this folder. The name of the project folder should be something meaningful, such as "Applied_Biostatics" or the name of your class (for students in my Applied Biostatics class, this folder should be named "BIO_413").
2. Within the project folder, create new folders named
    1. "notebooks" – this is where your R notebooks are stored
    2. "R" – this is where R scripts are stored
    3. "data" – this is where data that we download from public archives are stored
    4. "output" – this is where you will store fake data generated in this class
    5. "images" – this is where image files are stored
3. Open R Studio and click the menu item File > New Project...
4. Choose "Existing Directory" and navigate to your project folder
5. Choose "Create Project"
6. Check that a ".Rproj" file is in your project folder

1.3   R Notebooks

A typical statistical modeling project will consist of:

1. importing data from Excel or text (.csv or .txt) files
2. cleaning data
3. analysis
4. generating plots
5. generating tables
6. writing text to describe the project, the methods, the analysis, and the interpretation of the results (plots and tables)

The best practice for reproducible research is to have all six of these steps in your R Notebook. Too many research projects are not reproducible because the data were cleaned in Excel, and then different parts of the data were separately imported into a GUI statistics software for analysis, and then output from the statistics software was transcribed to Excel to make a table. And other parts of the analysis are used to create a plot in some plotting software. And then the tables and plots are pasted into Microsoft Word to create a report. Any change at any step in this process will require the researcher to remember all the downstream parts that are dependent on the change and to re-do an analysis, or a table, or a plot, etc. etc.

The goal with an R Studio Notebook is to explicitly link all this so that changes in earlier steps automatically flow into the later steps. So, at the end of a project, a researcher can choose "run all" from the menu and the data are read, cleaned, analyzed, ploted, tabled, and put into a report with the text.

This means that you have to think of the organization of the R code that your write in a Notebook. Your cannot simply append new code to the end of a script if something earlier (or above) is dependent on it. You need to go back up and insert the new code at some earlier (and meaningful) point.

For example, an R chunk generates 100 random normal values and then plots these with a histogram. This was the chunk that I wrote

```
x <- rnorm(n)
qplot(x)
```

When I ran the chunk, I got the error "Error in rnorm(n) : object n not found". I was using the function `rnorm()` to generate values but I hadn't assigned any value to `n` yet, so I got the error. To get this to work properly, I could have just typed `n <- 100` in the console and then re-run the script but I want it to work properly on a fresh run of the chunk (after quitting and re-opening R Studio) so I instead inserted `n <- 100` at the start of the chunk, like this:

```
n <- 100
x <- rnorm(n)
qplot(x)
```

1.3.1   Create an R Notebook for this Chapter

1. The top-left icon in R Studio is a little plus sign within a green circle. Click this and choose "R Notebook" from the pull-down menu.
2. Change the title of the notebook to "Notebook_01-organization"
3. Delete the default R Markdown text starting with "This is an [R Markdown]…"

Now write some text documenting which packages you installed.

1.3.2 Create a "load-packages" chunk

1. Click on the "Insert" menu on the right hand side of the script (R Markdown) pane and choose "R". This will insert an R code chunk into your R markdown document.
2. The first R chunk of a notebook should be a setup chunk. Name the chunk "setup"
3. load the libraries ggplot2 and data.table and click the chunk's run button (the green triangle to the right of the chunk)

```r
library(ggplot2)
library(data.table)
```

I added the chunk option "message=FALSE". Run your chunk with and without this as an option.

1.3.3 Create a "simple plot" chunk

4. Create a new chunk and label it "simple plot"
5. insert the following R script and then click the chunk's run button. Do you get a plot?

```r
n <- 100
x <- rnorm(n)
qplot(x)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```
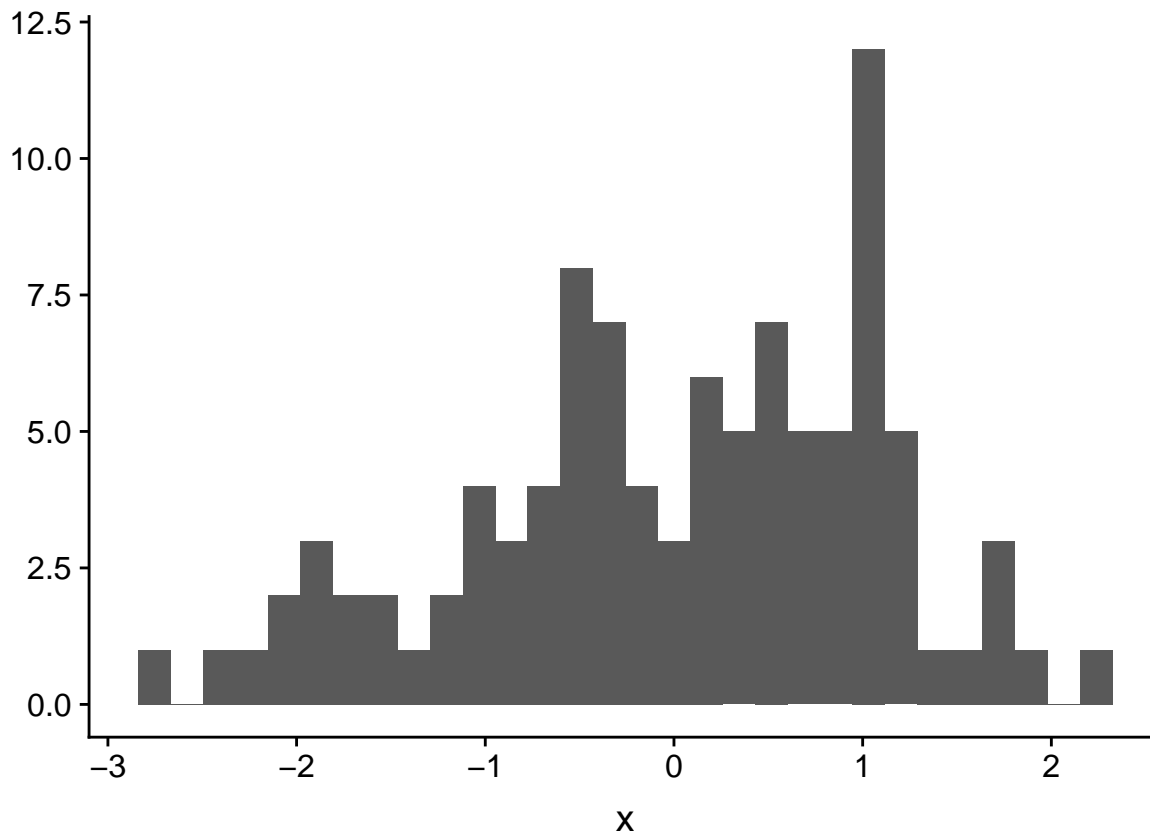

1.3.4 Create more R chunks and explore options and play with R code

Chapter 2

Data – Importing and Saving Data

Importing data into R can be a struggle for new R users and, unfortunately, most online sources give easy but superficial methods that don't follow best practices for increasing reproducibility or do not allow flexible organization of files within a project. Some examples:

(tl;dr – go to next section if you just want to import a file and don't want to understand this important background information!)

1. `df <- read.table(file="clipboard")` imports data copied to the clipboard, from an Excel/Sheets file or from an open text file. For this to be semi-reproducible, a comment specifying the filename, worksheet and range that was copied is necessary. More problematic (catastrophically so for reproducibility), is, how does a researcher know that they highlighted and copied the correct range in the Excel sheet?

2. `df <- read.csv(file.choose())` opens the familiar "open file" dialog box, which lets the user navigate to the file of choice. For this to be semi-reproducible, a comment specifying the filename to import is necessary. The catastrophic problem (for reproducibility) is, how does a researcher know which file was actually opened during an R session? The researcher might think they opened "walker_maine_bee_data_clean_androscoggin.csv" but mistakenly opened "walker_maine_bee_data_clean_aroostook.csv".

3. `df <- read.table(file="my_data.txt")` and `df <- read_excel(file="my_data.xlsx")` are reproducible because the filename is explicitly specified. But, this method requires that "my_data" is physically located in the same folder as the file containing the R script (the notebook .Rmd file in our case) and this violates the best practice of clean project organization with different folders for the different kinds of files (data, R scripts, images, manuscript text, etc.).

4. R Studio has an elegant import tool in the environment pane that opens a custom dialog box that allows the researcher to navigate to the file, and to specify what part of the file to import, such as the specific sheet and range for an Excel file. This has the same reproducibility issues as #1 and #2 but R Studio includes the equivalent script, which adds all relevant information for reproducility. One then simply copies and pastes this script into a code chunk and voila! The next time the script is run, the data can be imported from the script without using menus and dialog boxes. Except that..the script does not seem to take into account that the **working directory** of an R Markdown file is not the project folder but the folder containing the R Markdown file and so this two-step method fails. More personally, I'd prefer to run a chunk that quickly opens the data file instead of re-navigating through my file system and re-specifying the sheet and range every time I re-start the project in a new R session.

There are at least three solutions to the issues raised above, all requiring some understanding of **file paths** and directory structure in an operating system. A file such as "my_data.xlsx" has an **absolute** file path, which is the full address of the file (the filename is something like your house street number). The absolute file path of "my_data.xlsx" might be "/Users/jwalker/Documents/applied-biostatistics/data/my_data.xlsx". A

relative file path is the file path from the **working directory**. In an R Studio project, the working directory is the project directory, which is the directory containing the .Rproj file. This will be the working directory of the console. *Importantly*, the working directory of an R Markdown code chunk is the folder containing the saved R Markdown file. An R Studio Notebook is an R Markdown file so the working directory of a notebook code chunk is the folder containing the saved notebook file. If a notebook file is located within the notebooks folder, which is located within the project folder, then the relative file path to "my_file.xlsx" is "../data/my_file.xlsx". The ".." tells the file OS to move "up" into the parent directory (which is the project folder) and the "data" tells the file OS to move "down" into the data folder. These are put together into a single address using "/". The beauty of relative paths is that they remain the same – and so do not break one's script – if the project folder, and all of its contents including the data folder and the notebooks folder, is moved to another location on the hard drive (say into a new "Research" folder). By contrast, the absolute file path changes, which breaks any old script.

The three solutions are

1. Create a **relative path** to the file using something like `file_path <- "../data/my_data.xlsx"`. This should *always* work but it fails on some computers. For example, if the project folder is on a Windows OS (but not Mac OS) desktop, the assigned relative address doesn't seem to look in the folder containing the file.

2. Create a setup chunk that reroutes the working directory to the project folder using the script

```r
# use this in a chuck called "setup" to force the working directory to be
# at the level of the project file.
knitr::opts_knit$set(root.dir = rprojroot::find_rstudio_root_file())
```

For this to work, the chunk has to be named "setup", that is, the text inside the curly brackets at the top of the chunk should be "r setup". Then, with this chunk, the relative file path is `file_path <- "../data/my_data.xlsx"` if "my_data.xlsx" is immediately inside the data folder which is imediately inside the project folder. This should work on any machine, and should work even if a project folder is moved. However, it does not work when writing a book with Bookdown a Bookdown project has multiple .Rmd files and the chunk names cannot be repeated across files. There can only be one chunk called "setup".

3. Use the function here(). The most robust solution seems to be using the function here() from the here package. The function works something like this

```r
data_path <- "data" # path to data that are imported
file_name <- "my_data.xlsx"
file_path <- here(data_path, file_name) # paste together parts of the address
read_excel(file=file_path)
```

here::here() creates an absolute path, but one that is created on the fly, and will change (or should change) correctly if the project folder is moved on the same machine or to another machine. And no chunk named setup is required so this also works in every R markdown file when using Bookdown.

2.1   Create new notebook for this chapter

Create a new notebook and save it to the "notebooks" folder of your BIO_413 project. **Important:** In an R Studio project, the working directory is the project directory (the location of the .Rproj file) – this will be the working directory of the console. The working directory of an R Markdown code chunk is the folder containing the saved R Markdown file. An R Studio Notebook is an R Markdown file so the working directory of an notebook code chunk is the folder containing the saved notebook file. Importantly, the working directory of the code chunk isn't "moved" to the folder containing the saved file *until the file is saved* so import/export scripts will no work properly until the file is saved!

Annotate your notebook with notes! Update it as you learn more! We will use data.table for importing text files in tab-delimited or comma-separated formats and the readxl package for importing excel files.

Start each notebook with a chunk named "setup" (note that since this book is advocating the use of here() to set the file path, naming the setup chunk "setup" is not imperative).

```
knitr::opts_chunk$set(fig.width=6, fig.height=4)

# relative paths to project folders
data_path <- "data" # path to data that are imported
output_path <- "output" # path to data that are saved
```

Add a "load-packages" chunk.

```
library(ggplot2) # ggplot environment
library(ggpubr) # publication ready plots
library(data.table) # make data.frames data.tables
library(readxl) # import excel
library(emmeans) # get estimated marginal means and CIs, used for plot
library(janitor) # clean_names to clean col labels of imported data
library(here) # here::here() creates the absolute path to the file
```

2.2 Importing Data

Throughout this book, we will download data from the Dryad Digital Repository, which is a major resource for increasing reproducibility in science. My own view is that *all data* should be archived on some public server (exceptions include data that are proprietary or contain sensitive information – such as human health measures).

2.2.1 Excel File

The Excel dataset is from an experiment on the growth response of zebra finch chicks to an incubation call that presumably signals "hot environment" to the embryos (Mariette, M.M. and Buchanan, K.L., 2016. Prenatal acoustic communication programs offspring for high posthatching temperatures in a songbird. Science, 353(6301), pp.812-814). The source file is from the Dryad Repository here:

file name: "allDatasetsMarietteBuchanan2016.xls"

source: https://datadryad.org//handle/10255/dryad.122315

Steps

1. Copy the title of the Dryad page, which is "Data from: Prenatal acoustic communication programs offspring for high post-hatching temperatures in a songbird"
2. Create a new folder within "data" and paste in the copied title as the folder name
3. Remove the colon from the name, so the folder name is "Data from Prenatal acoustic communication programs offspring for high post-hatching temperatures in a songbird"
4. Download the .xls file into this folder

A .xls file is an old (pre 2007) Microsoft Excel file type. It is a binary file and can only be opened into a readable format with specialized software. The more modern Excel file type is .xlsx, which contains within it multiple xml components. An xml file is a text file, and so contains readable content, but the content is xml code to display something. In general, I am a big advocate of archiving stuff as text files (manuscripts, data, scripts, blog posts) because these will *always* be readable by future software. Microsoft Excel is not likely to die anytime soon and software that can read .xls and especially .xlsx files (again, .xlsx files are text files) is even less likely to disappear but we can feel even more confident if data are archived as text files. That said, a single microsoft excel file with multiple sheets is an efficient method for distributing data and the readxl package provides excellent tools for reading different sheets of a single .xls or .xlsx file.

The code below uses the function `read_excel()` from the package readxl. More about the amazing power of this package is the tidyverse page and chapter 11 in the *R for Data Science* book.

```
data_folder <- "Data from Prenatal acoustic communication programs offspring for high post-hatching temp
filename <- "allDatasetsMarietteBuchanan2016.xls"
file_path <- here(data_path, data_folder, filename)
chick <- read_excel(file_path, sheet="nestlingMass") # read the excel file
chick <- data.table(chick) # convert to data.table
```

In this book, we will consistently uses this protocol for storing and retrieving downloaded files. The first three lines in the script above creates the directory path to the file. This path includes

1. data_path – the relative path into the folder "data" (relative to the location of the project file)
2. data_folder – the name of the folder within "data" containing the file
3. filename – the name of the file to read

These are all put together into the absolute path using the function `here()` from the here package. Take a look at the value of file_path to confirm. The `read_excel(file_path, sheet="nestlingMass")` reads the nestlingMass sheet only. The next line:`chick <- data.table(chick)` converts the data.frame into a data.table.

2.2.1.1   Troubleshooting File Import

If you get an error that starts with "Error: `path` does not exist:" then R is not "seeing" your specified file given the path you've given it.

1. Make sure you loaded the package `here` in a "load-libraries" chunk and that you have run the chunk
2. Make sure you have assigned `data_path <- "data"` in the setup chunk and have run the chunk.
3. Make sure your "data" folder is *one level* inside your project folder. "one level" means it is not buried deeper inside other folders within the project folder.
4. Make sure your "Data from …" folder is one level inside your "data" folder
5. Make sure you have the name of the "Data from …" folder correct in your script. To do this, I do not type the name but instead go to the finder (on Mac OS) and highlight the "Data from …" folder and copy the name – the whole name – and then paste in between the quote mark in the line `data_folder <- ""`
6. Make sure your data file is one level inside the correct "Data from" folder.
7. Make sure the file name is correct in the script. As with the folder name, I go to the finder and copy the file name and paste it in place. In Windows use ctrl-a instead of ctrl-c to copy the full filename including the extension.

More generally, R is *very literal* when it comes to: Spelling. Humans are very good at understanding misspelled words but the R language (or any computer language) is very literal. "Data from Quantifying the effects of" and "Data from Quantifying the efects of" are different values. Capitalization. R is **case sensitive** (some programming languages are not). "Data from Quantifying", "Data from quantifying", and "data from quantifying" are all different values. Spelling AND capitalization have to be perfect, not simply close. Humans are very good at understanding misspelled and OdDLy capitalized words but the R language (or any computer language) is very literal.

2.2.1.2   Peak at the imported data.table to check that the file was imported correctly and to learn about the contents.

Insert the following after your import script and run:

```
View(chick) # check -- are there headers? are there the correct number of columns?
```

The line `View(chick)` script opens a new tab containing a spreadsheet-like display of the data frame `chick`. This is one way to check that the data were imported correctly, to examine the column labels to ensure they conform to best practices, and to simply get to know the contents of the imported file. From the view of chick, it may not be immediately apparent that the column names contain spaces. This can be seen more easily using

```
colnames(chick)
```

```
##  [1] "chick ID"                    "brood ID"
##  [3] "brood composition"           "sex"
##  [5] "rank in nest"                "playback treatment"
##  [7] "nest temperature above ambient" "max daily temp hatch day"
##  [9] "mean max temp hatch to day2"    "mean max temp hatch to day10"
## [11] "mean max temp hatch to day13"   "hatching mass"
## [13] "day1 mass"                   "day2 mass"
## [15] "day10 mass"                  "day13 mass"
## [17] "day13 tarsus"
```

In general, it is bad practice to include spaces, parentheses, and special characters such as $ or ^, in the column names of a data frame because these increase handling costs later on. The best practice is to replace a blank with an underscore, for example `brood_id`. Some coders separate words with a period (`brood.id`). Others mash words together into a single word like this `broodid` but this should generally be avoided because the result can be hard to read. Finally, some coders use Caps to designate new words like this `BroodId`. This is easier to read than simple concatenation but the underscore is the easiest to read.

The `janitor` package has a nice function to clean the column names of a data frame including replacing spaces with an underscore and stripping parentheses. The default clean includes changing any uppercase letter to lower case. Many coders like to work with all lowercase variable names to avoid having to hit the shift key.

```
chick <- clean_names(chick)
```

Importantly, **resist the temptation to change the column names in the data file**, which reduces reproducibility. Leave original data files original. Always increase reproducibility!

2.2.1.3   Best practices for creating data files

1. https://www.youtube.com/watch?time_continue=309&v=Ry2xjTBtNFE – An excellent video introduction to best practices for organizing data in a spreadsheet that will subsequently be analyzed by statistics software.

2. Broman, K. W., & Woo, K. H. (2017). Data organization in spreadsheets (No. e3183v1). https://doi.org/10.7287/peerj.preprints.3183v1 – An excelllent review of best practices for organizing data in a spreadsheet.

2.2.1.4   Explore with plots

Just for fun, let's plot the data and reproduce Fig. 2A and B. We are using the `qplot` function, which is from the ggplot2 package. Two plots are made and only a subset of the rows are plotted in each (in A, the subset in which playback_treatment=="treat" and, in B, the subset in which playback_treatment=="cont"). This book uses the ggplot2 package extensively.

```
qplot(x=nest_temperature_above_ambient, y=day13_mass, data=chick[playback_treatment=="treat"]) +
  geom_smooth(method="lm")
```
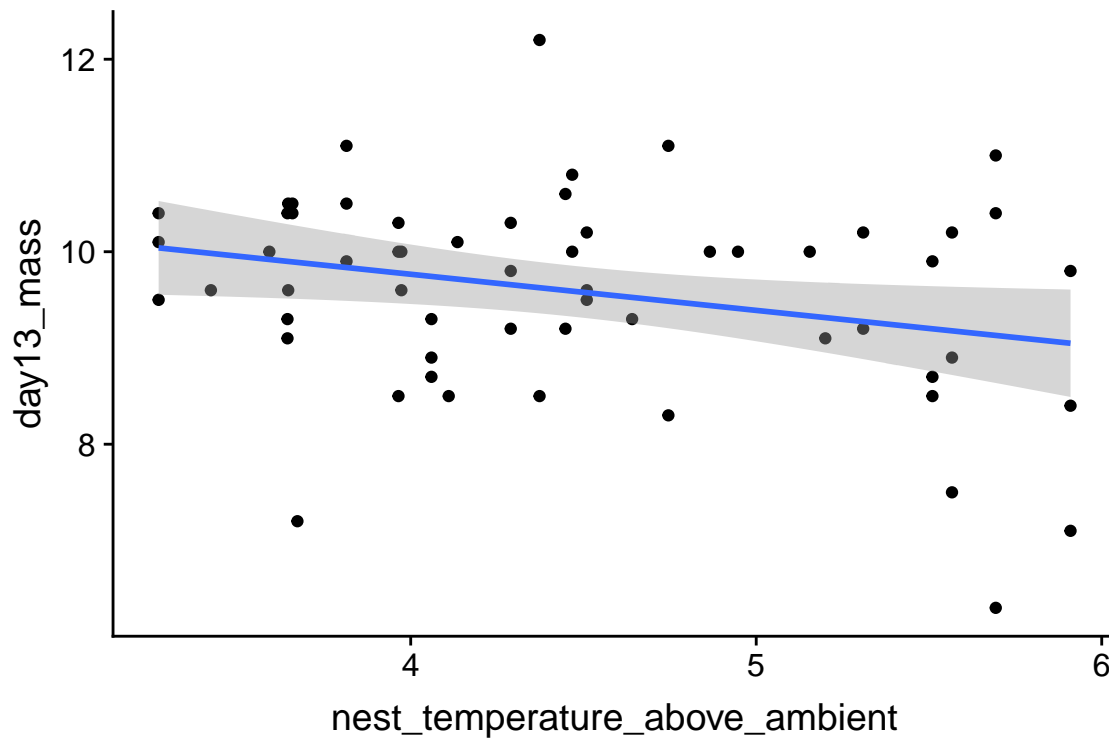
```
qplot(x=nest_temperature_above_ambient, y=day13_mass, data=chick[playback_treatment=="cont"]) +
  geom_smooth(method="lm")
```
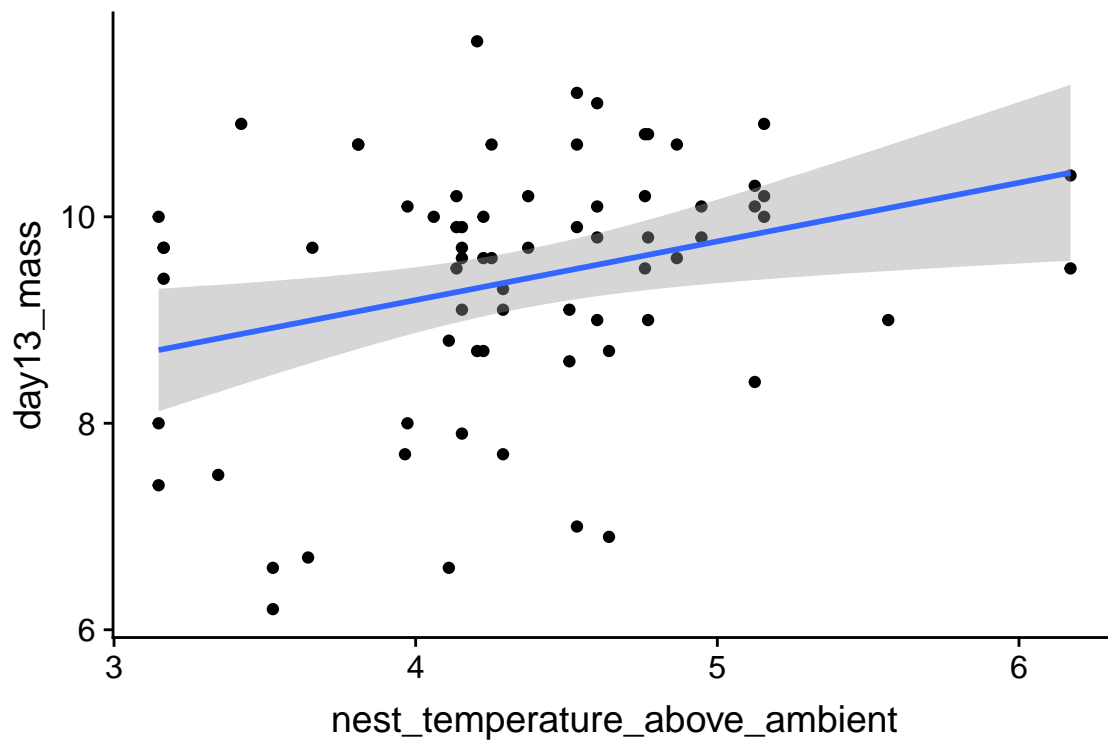
2.2.2 Text File

The example dataset comes from an experiment on the effect of neonicotinoid pesticides on bumble bee colony growth.

file name: "Whitehorn, O'Connor, Wackers, Goulson (2012) Data from 'Neonicotinoid pesticide reduces bumblebee colony growth and queen production'.csv.csv"

source: https://datadryad.org//resource/doi:10.5061/dryad.1805c973

Steps

1. Copy the title of the Dryad page, which is "Data from: Neonicotinoid pesticide reduces bumblebee colony growth and queen production"
2. Create a new folder within "data" and paste in the copied title as the folder name
3. Remove the colon from the name, so the folder name is "Data from Neonicotinoid pesticide reduces bumblebee colony growth and queen production"
4. Download the .csv file into this folder

A .csv file is a text file that is comma-delimted, which means that the entries of a row are separated by commas. A text file is readable by any text editor software and most other kinds of software. Datasets that are stored as text files are typically saved as either .csv (where the entries of a row are separated by commas) or .txt (where the entries are separated by tabs). The base R way to read a .csv file is using `read.csv`. The `read.table` function is more versatile, as the delimiter can be specified. The function `fread()` from the data.table package is fast, smart, and flexible. It is smart in the sense that it guesses what the delimter is. Unfortunately, because of spaces in the column labels for this file, fread guesses incorrectly (another reason why spaces in column labels should be avoided). To overcome this, the statement below specifies that the file contains a "header" (a line containing column labels)

```
data_folder <- "Data from Neonicotinoid pesticide reduces bumblebee colony growth and queen production"
filename <- "Whitehorn, O'Connor, Wackers, Goulson (2012) Data from 'Neonicotinoid pesticide reduces bu
file_path <- here(data_path, data_folder, filename)
bee <- fread(file_path, header=TRUE)
```

Here, as with the import of the Excel file, the first three lines create the directory path to the file. Peek at the file in the console. Again, there are spaces in the column names. **Here I'll leave it to you to change this**

Here is a reproduction of Fig 2 from the journal article.

```
bee[, treatment:=factor(treatment, c("Control", "Low", "High"))] # reorder factor levels
ggbarplot(data=bee, x="treatment", y="new_queens", add = "mean_se")
```
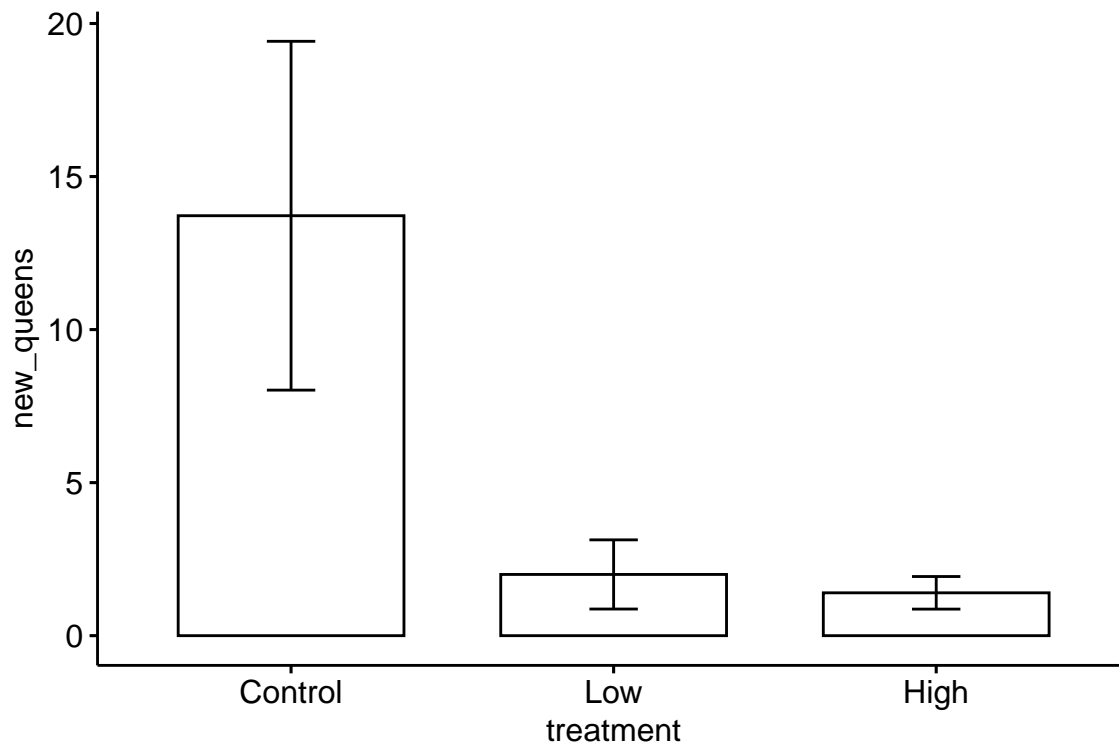
The plot suggests immediately some problems with the plot itself and the associated analysis. First, the y-axis is counts, which means that negative values are impossible. But the standard error bars look like they use standard errors computed from a model that allows infinetly large negative values, and the illustrated standard error bars imply that negative values exist. So these error bars are misleading. Second, it is good practice, especially if sample sizes are modest or small, to "show the data", which means, show the individual data points and not just a summary of the distribution.

Here are three alternative plots for exploratory purposes. The first simply "shows the data" but still uses the misleading standard error bars. The second uses a box plot. The last plots the means and 95% confidence intervals modeled with a GLM (generalized linear model) to account for the count data (the model used could be improved). Notice that the bar length above the mean is longer than the bar length below the mean (that is the interval is asymmetric about the mean). In order to stay focussed on importing data, I leave explanation of these plots and analysis to later chapters.

```
ggbarplot(data=bee, x="treatment", y="new_queens", add = c("mean_se", "point"))
```
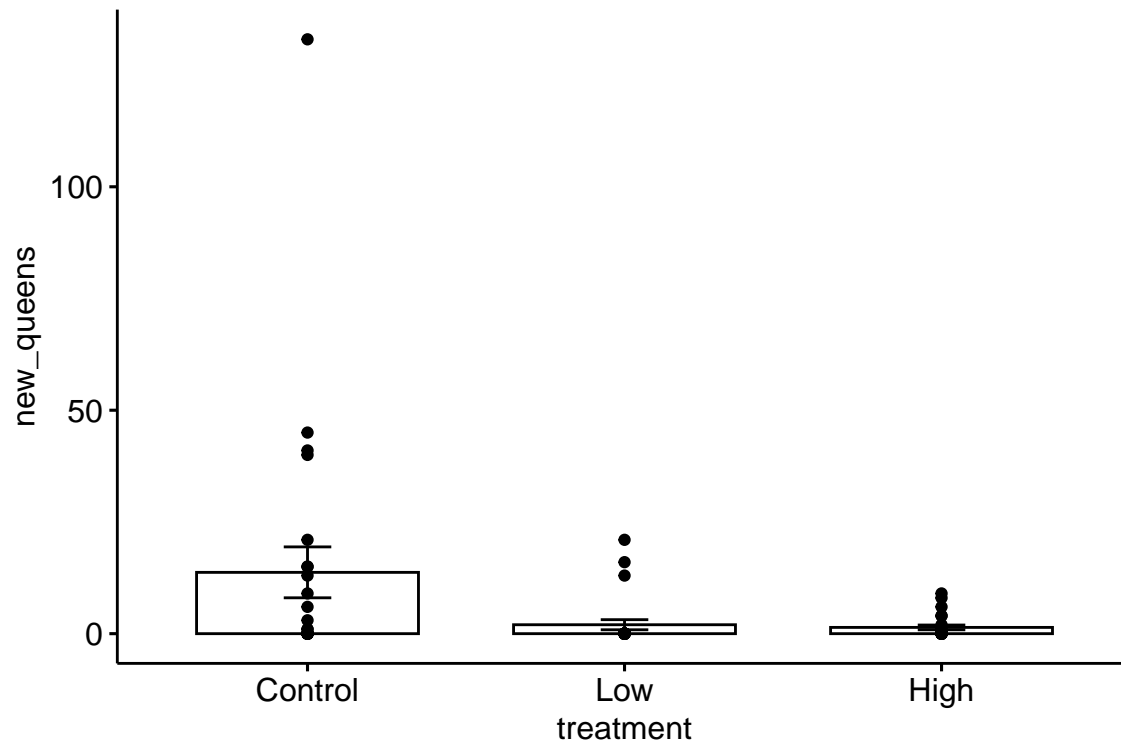
```
ggboxplot(data=bee, x="treatment", y="new_queens")
```
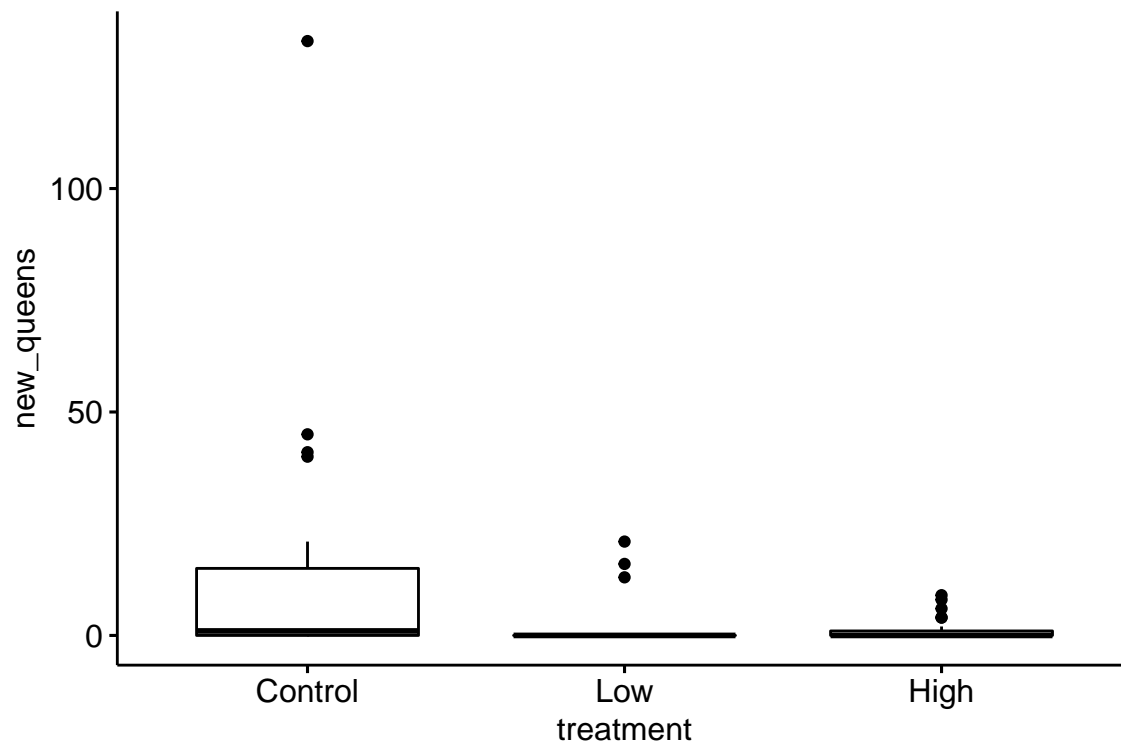


```
fit.glm <- glm(new_queens ~ treatment, data=bee, family=poisson())
means.glm <- emmeans(fit.glm, specs="treatment", type = "response")
gg <- ggplot(data=data.frame(means.glm), aes(x=treatment, y=rate)) +
  geom_col(fill="gray") +
  geom_errorbar(aes(x=treatment, ymin=asymp.LCL, ymax=asymp.UCL), width=0.3) +
```

```
  ylab("New queens") +
  NULL
gg
```
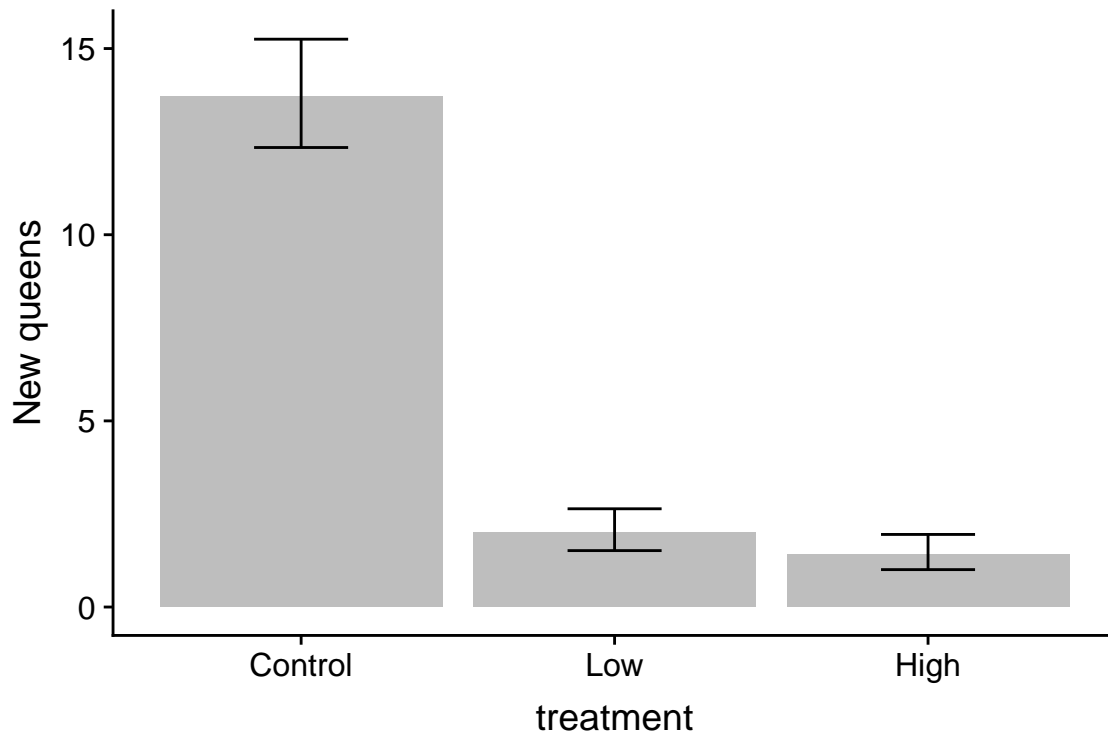

2.3   Saving Data

For many projects, it is uncommon to save data. I might save simulated data if it takes a long time (hours or days) to generate these and I simply want to work with the simulated data in the future (and not have to regenerate it). Or I might save processed data if it takes a long time to process it and I want to analyze the processed data in the future and not have to re-process it.

Here, we will create a small, simulated data set and save it to the "output" folder. You should have an "output" folder directly (one-level) inside the project folder. Create a folder within the output folder names "chapter_02". Make sure that you have the script assigning the path to the output folder code in the setup chunk.

```
set.seed(1)
n <- 10
beta_0 <- 10.3
beta_1 <- 1.1
sigma <- 1.5
fake_data <- data.table(treatment=rep(c("Cn","Tr"), each=n),
                        weight=c(rnorm(n, mean=beta_0, sd=sigma),
                                 rnorm(n, mean=beta_0 + beta_1, sd=sigma)))
ggstripchart(data=fake_data, x="treatment", y="weight", add="mean_ci")
```
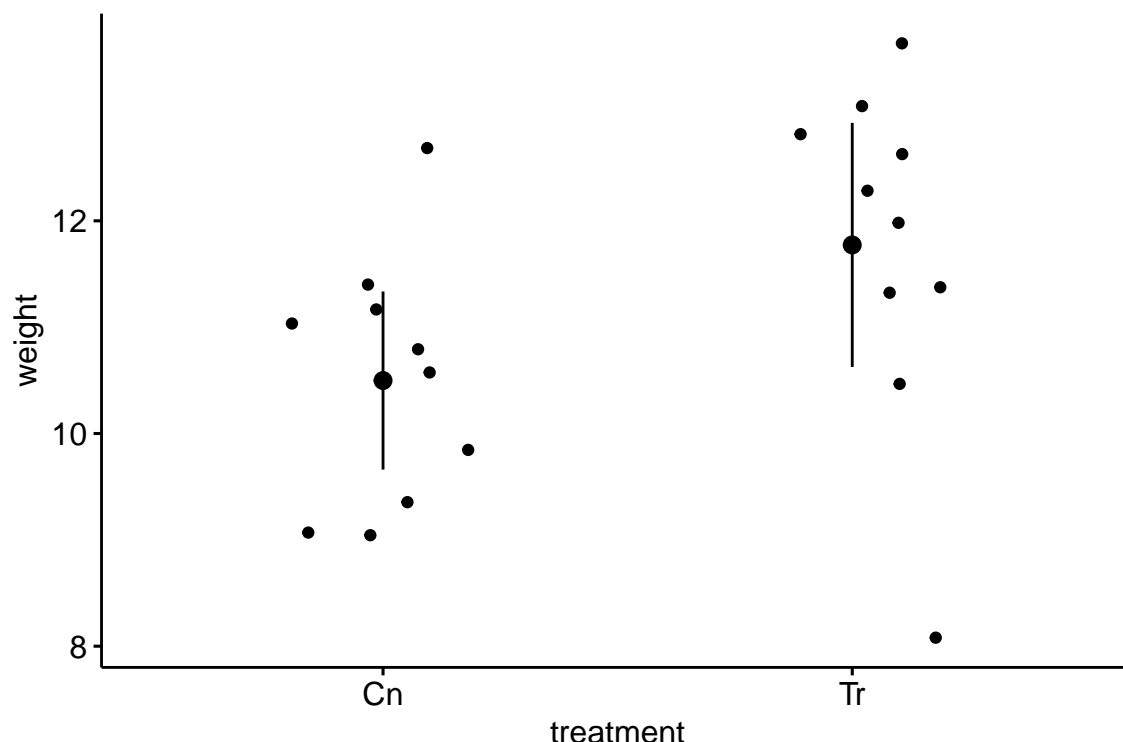
```
save_data_folder <- "chapter_02"
save_filename <- "my_first_fake_data.txt"
save_file_path <- here(output_path, save_data_folder, save_filename)
write.table(fake_data, save_file_path, sep="\t", quote=FALSE)
```

Look at your project directory to make sure the file is where it should be! We used `write.table()` to create a tab-delimited text file using `sep="\t"` to specify tabs to separate the row elements. "$\widehat{}$" is the standard character string for a tab. Check in your Fake_Data folder and open the file in a text editor.

2.4   Problems

1. Download the dataset "data-Lodjak.et.al-2016-FuncEcol.xlsx" from the Dryad repository at https://datadryad.org/resource/doi:10.5061/dryad.rd01s. The .xlsx file presents the data cleanly but the trade-off is that the 1) multiple header rows, and 2) spaces in the header labels, 3) parentheses in the header labels make it more complex to import in a usable way. Import the data and plot Body Mass against Age (that is make Body Mass the "Y" variable and Age the "X" variable) using the qplot function. You should recode the column labels to remove spaces and parentheses using the setnames function.

2. Download the dataset "Results2015.txt" from the Dryad repository at https://datadryad.org//resource/doi:10.5061/dryad.65vk4. Try to reproduce Fig. 1. It's not easy. I've inserted the figure below but also download the paper and look at Fig. 1.

3. (grad students only) Download and plot data from a Dryad Repository dataset of your choice.

4. (grad students only) Create fake experimental data with three treatment levels (control, lo_temp, high_temp). This will require three parameters: an intercept (beta_0), an effect of lo_temp (beta_1), and an effect of high_temp (beta_2). You should be able to plug and play from the script above even if you don't underdstand at this point what any of it is! Plot it as a strip chart, as above.
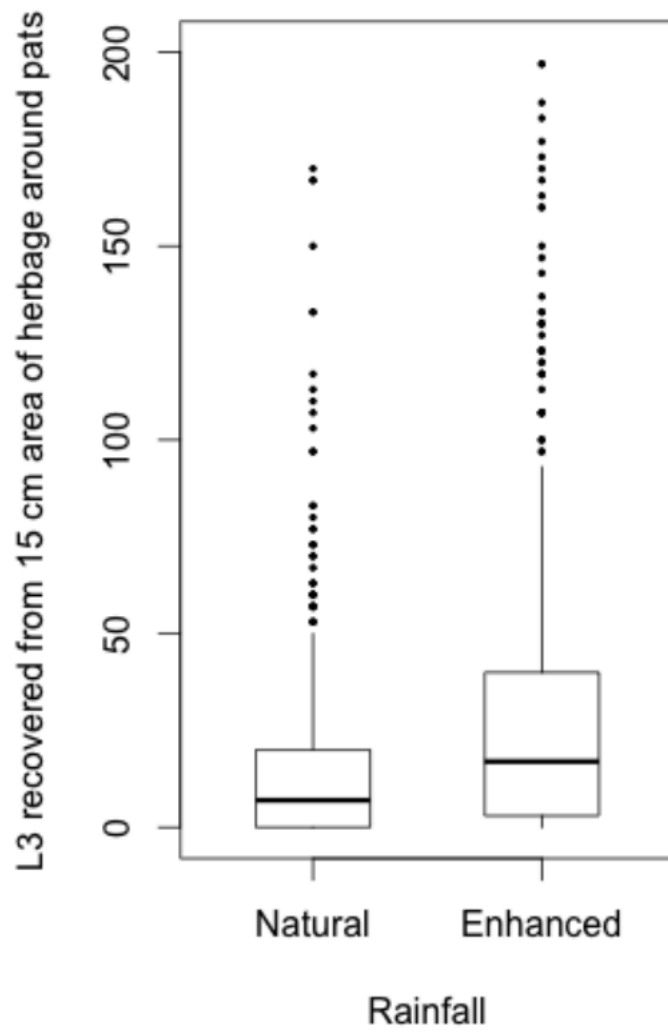
Fig. 1

Part II: Some Fundamentals of Statistical Modeling

Chapter 3

An Introduction to Statistical Modeling

This chapter introduces statistical modeling using the **linear model**. All students are familiar with the idea of a linear model from learning the equation of a line, which is

$$Y = mX + b \tag{3.1}$$

where m is the slope of the line and b is the Y-intercept. It is useful to think of equation (3.1) as a function that maps values of X to values of Y. Using this function, if we input some value of X, we always get the same value of Y as the output.

A linear model is a function, like that in equation (3.1), that is fit to a set of data, often to model a process that generated the data or something like the data. The line in Figure 3.1A is just that, a line, but the line in Figure 3.1B is a linear model fit to the data in Figure 3.1B.

3.1 Two specifications of a linear model

3.1.1 The "error draw" specification

A linear model is commonly specified using

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.2}$$

$$\tag{3.3}$$

This specification of a linear model has two parts: the **linear predictor** $Y = \beta_0 + \beta_1 X$ and the **error** ε. The linear predictor part looks like the equation for a line except that 1) β_0 is used for the intercept and β_1 for the slope and 2) the intercept term precedes the slope term. This re-labeling and re-arrangement make the notation for a linear model more flexible for more complicated linear models. For example $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ is a model where Y is a function of two X variables.

As with the equation for a line, the linear predictor part of a linear model is a function that maps a specific value of X to a value of Y. This mapped value is the **expected value** given a specific input value of X. This is often written as $E[Y|X]$, which is read as "the expected value of Y given X", where "given X" means
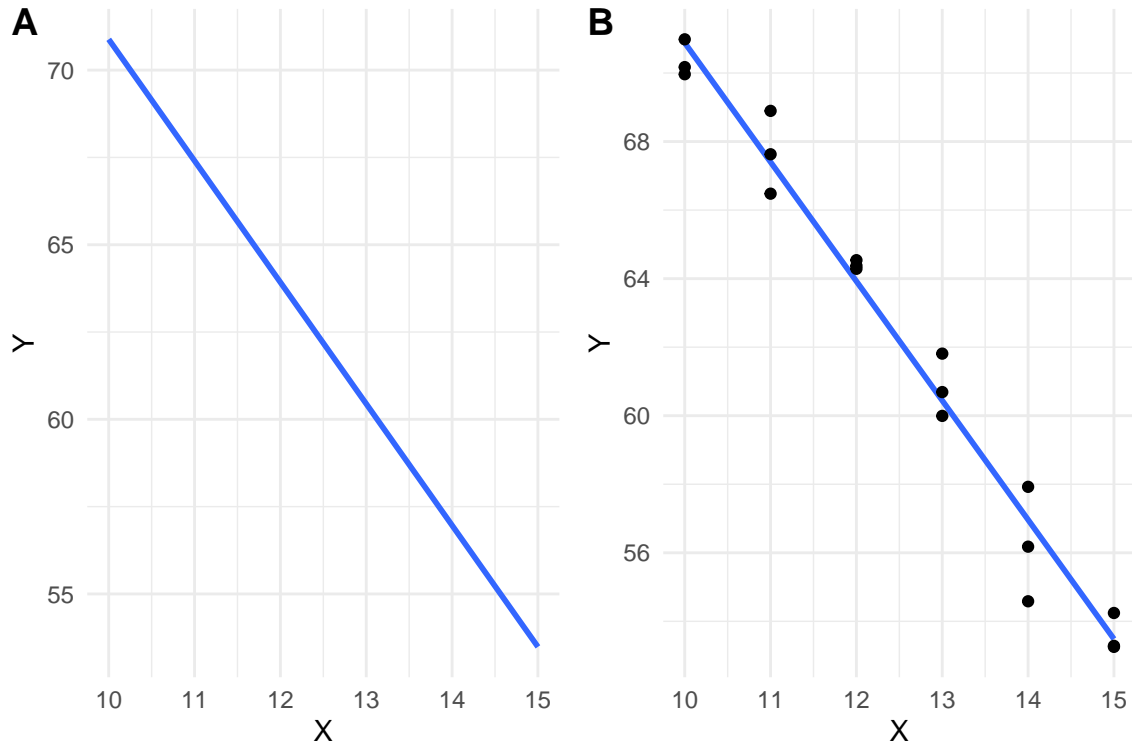
Figure 3.1: A line vs. a linear model. (A) the line $y = -3.48X + 105.7$ is drawn. (B) A linear model fit to the data. The model coefficients are numerically equal to the slope and intercept of the line in A.

a specific value of X. Importantly, $E[Y|X]$ is the **conditional mean**, which is the *modeled* value of Y for all observations in which X takes some specific value x.

Introductory textbooks almost always introduce linear models using equation (3.3) above. The key part of the model that is missing from the specification above is a second line

$$\varepsilon \sim N(0, \sigma^2) \tag{3.4}$$

which is read as "epsilon is distributed as Normal with mean zero and variance sigma squared". This line explicitly specifies the distribution of the error part. The error part of a linear model is a random "draw" from a normal distribution with mean zero and variance σ^2. Think of this as adding some random value to the expected value.

3.1.2   The "conditional draw" specification

A second way of specifying a linear model is

$$y_i \sim N(\mu_i, \sigma^2) \tag{3.5}$$
$$E(Y|X) = \mu \tag{3.6}$$
$$\mu_i = \beta_0 + \beta_1 x_i \tag{3.7}$$

The first line states that the response variable Y is a random variable independently drawn from a normal distribution with mean μ and variance σ^2. This first line is the **stochastic** part of the statistical model.

The second line simply states that μ (the greek letter "mu") from the first line is the conditional mean or conditional expectation. The third line states how μ_i is generated given that $X = x_i$. This is the linear predictor, which is the **systematic** (or deterministic) part of the statistical model. It is systematic because the same value of x_i will always generate the same μ_i.

3.1.3 Comparing the two ways of specifying the linear model

These two ways of specifying the model encourage slightly different ways of thinking about how the data (the response varible Y) were generated. The error draw specification "generates" data by randomly drawing some error ε_i from the specified distribution and adding this to x_i. The conditional draw specification "generates" data by constructing what y_i "should be" given x_i (the conditional expection), and then drawing a random variable from a distribution with this expectation. This random draw is y_i and not the "error". For the error draw generation, we need only one hat of random numbers, but for the conditional draw generation, we need a hat for each value of x_i.

The conditional draw specification explicitly defines all parameters, including the parameters of the linear predictor (β_0 and β_1), the conditional mean μ and the variance σ^2. The error draw specification only defines the parameters of the linear predictor, and often these are referrred to as "the parameters" in the sense that there are not other parameters. The error draw specification is most useful for thinking about model checking a fit linear model. The random draw specification is more generally useful in that it is easily generalized to more complex models, including hierarchical models, generalized linear models, and Bayesian models. In fact, *thinking about a model as a predictor plus error can lead to the misconception that in a generalized linear models, the error has the distribution (binomial, poisson, etc.) modeled.*

Although a linear model (or statistical model more generally) is a model of a data-generating process, linear models are not typically used to actually generate any data. Instead, when we use a linear model to understand something about a real dataset, we think of our data as one realization of a process that generates data like ours. A linear model is a model of that process. That said, it is incredibly useful to use linear models to create fake datasets for at least two reasons: to probe our understanding of statistical modeling generally and, more specifically, to check that a model actually creates data like that in the real dataset that we are analyzing.

3.2 What do we call the X and Y variables?

The inputs to a linear model (the X variables) have many names including "independent variables," "predictor variables,", "explanatory variables," "treatment variables," and "covariates". The output of a linear model (the Y variable or variables if the model is multivariate) is the "dependent variable," "response," or "outcome." The β in the linear model are model **parameters** and if a parameter is multiplied by an X variable then it is also a **coefficient** (for example, β_1 in model (3.3) is a coefficient). The coefficients of the X in a linear model (β_1 in model (3.3)) are often called "the effects" (so β_1 is the effect of X_1).

3.3 Statistical models are used for prediction, explanation, and description

Researchers typically use statistical models to understand relationships between one or more Y variables and one or more X variables. These relationships include

1. Descriptive modeling. Sometimes a researcher merely wants to describe the relationship between Y and a set of X variables, perhaps to discover patterns. For example, the arrival of a spring migrant bird (Y) as a function of sex (X_1) and age (X_2) might show that males and younger individuals arrive earlier. Importantly, if another X variable is added to the model (or one dropped), the coefficients,

and therefore, the precise description, will change. That is, the interpretation of a coefficient as a descriptor is *conditional* on the other covariates (X variables) in the model. In a descriptive model, there is no implication of causal effects and the goal is not prediction. Nevertheless, it is very hard for humans to discuss a descriptive model without using causal language, which probably means that it is hard for us to think of these models as *mere description*. Like natural history, descriptive models are useful as patterns in want of an explanation, using more explicit causal models including experiments.

2. Predictive modeling. Predictive modeling is very common in applied research. For example, fisheries researchers might model the relationship between population density and habitat variables to predict which subset of ponds in a region are most suitable for brook trout (*Salvelinus fontinalis*) reintroduction. The goal is to build a model with minimal prediction error, which is the error between predicted and actual values for a future sample. In predictive modeling, the X ("predictor") variables are largely instrumental – how these are related to Y is not a goal of the modeling, although sometimes an investigator may be interested in the relative importance among the X for predicting Y (for example, collecting the data may be time consuming, or expensive, or enviromentally destructive, so know which subset of X are most important for predicting Y is a useful strategy).

3. Explanatory (causal) modeling. Very often, researchers are explicitly interested in *how* the X variables are causally related to Y. The fisheries researchers that want to reintroduce trout may want to develop and manage a set of ponds to maintain healthy trout populations. This active management requires intervention to change habitat traits in a direction, and with a magnitude, to cause the desired response. This model is predictive – a specific change in X predicts a specific response in Y – because the coefficients of the model provide knowledge on how the system functions – how changes in the inputs *cause* change in the output. Causal interpretation of model coefficients requires a set of strong assumptions about the X variables in the model. These assumptions are typically met in **experimental designs** but not **observational designs**.

With observational designs, biologists are often not very explicit about which of these is the goal of the modeling and use a combination of descriptive, predictive, and causal language to describe and discuss results. Many papers read as if the researchers intend explanatory inference but because of norms within the biology community, mask this intention with "predictive" language. Here, I advocate embracing explicit, explanatory modeling by being very transparent about the model's goal and assumptions.

3.4 Modeling strategy

1. **choose a model**. Statistical modeling includes a diverse array of models, yet almost all methods used by researchers in biology, and all models in this book, are generalizations of the linear model specified in (16.7).

2. **fit the model**, in order to estimate the model parameters and the uncertainty in these estimates.

3. **check the model**, which means to use a series of diagnostic plots and computations of model output to check that the data reasonably approximate the chosen model.

4. **inference from the model**, which means to use the fit parameters to learn, with uncertainty, about the system, or to predict future observations, with uncertainty.

5. **plot the model**, which means to plot the estimated parameters (or other results dervived from the estimates) with their uncertainty.

In order to use a statistical model to describe, predict, or explain, we need to fit a model to data in order to estimate the parameters. A linear model fit to some data is

$$\hat{y}_i = b_0 + b_1 x_i + e_i \tag{3.8}$$

$$\tag{3.9}$$
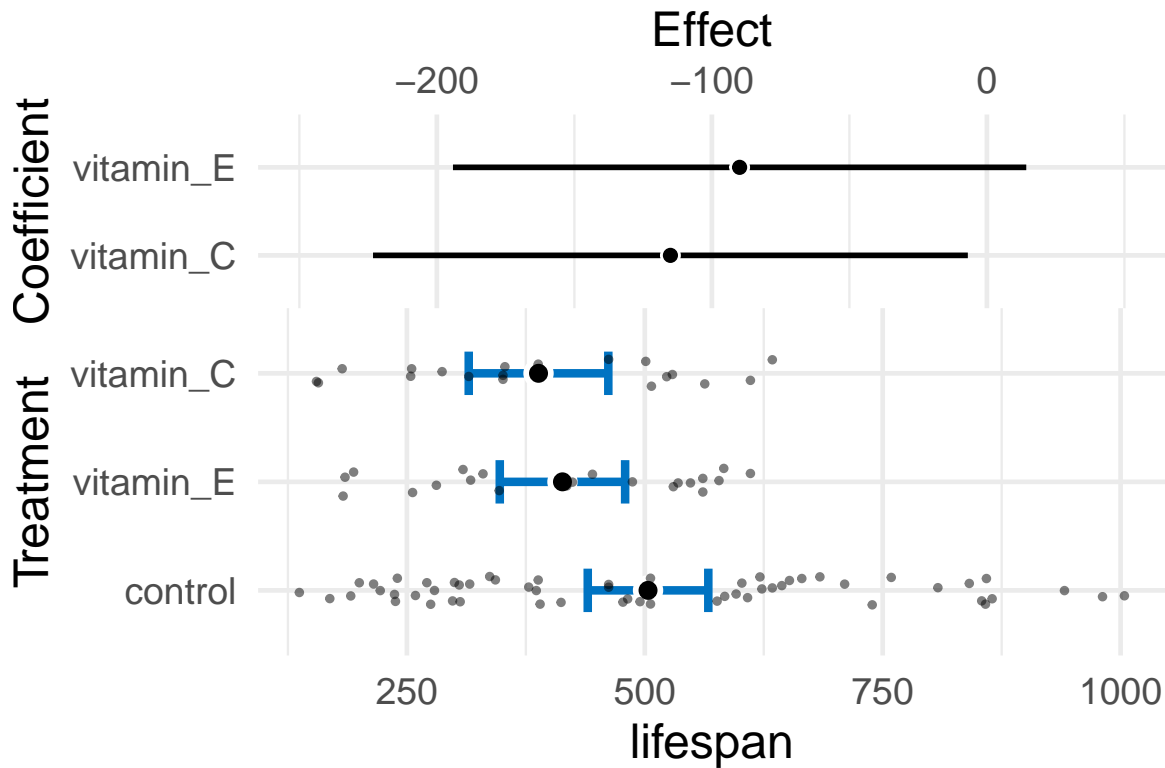
Figure 3.2: HarrellPlot of vole data.

$\hat{y}_i$ ("y hat") is the **predicted value** of individual i, b_0 and b_1 are the coefficients of the model fit (though technically b_0 is not a coefficient), and e_i is the residual. Sometimes $\hat{y}_i$ is simply called "the prediction".

If our goal is inference – to infer something about the "population" from the sample using the fit model, then $\hat{y}_i$ is the **point estimate** of the parameter μ_i, the coefficients b_0 and b_1 are point estimates of the parameters β_0 and β_1, and the standard deviation of the e_i is an estimate of σ. "Population" is in quotes because it is a very abstract concept. Throughout this book, Greek letters refer to a theoretical parameter and Roman letters refer to point estimates.

Throughout this text, I recommend reporting and interpreting **interval estimates** of the point estimate. A **confidence interval** is a type of interval estimate. A confidence interval of a parameter is a measure of the uncertainty in the estimate. A 95% confidence interval has a 95% probability (in the sense of long-run frequency) of containing the parameter This probability is a property of the population of intervals that could be computed using the same sampling and measuring procedure. It is not correct, without further assumptions, to state that there is a 95% probability that the parameter lies within the interval. Perhaps a more useful interpretation is that the interval is a **compatability interval** in that it contains the range of estimates that are compatible with the data, in the sense that a *t*-test would not reject the null hypothesis of a difference between the estimate and any value within the interval (this interpretation does not imply anything about the true value).

For the model fit to the data in Figure 3.1B, the coefficient of X is the slope of the line. Perhaps surprisingly, we can fit a model like equation (3.3) to data in which the X variable is categorical. A simple example is the experiment of antioxidants (vitamins C and E) on lifespan in Voles (Fig. 3.2). In this experiment, the X variable is categorical, with three **levels**: "Control", "Vitamin_E" and "Vitamin_C". Categorical X variables are often called **factors**. The trick to using a statistical model with categorical X is to recode the factor levels into numbers – how this is done is explained in Chapter xxx. When the X variable is categorical, the coefficients of the X are *differences in group means*. The linear model fit to the vole data has two coefficients, one for Vitamin E and one for vitamin C. The estimate and uncertainty of the these

two coefficients are shown in the top part of Figure 3.2. The bottom part shows the raw data, as well as the group (factor level) means and the uncertainty in the estimate of these means.

3.5   A mean is the simplest model

The simplest possible model that can be fit to the data is

$$E[Y] = b_0 \tag{3.10}$$

which is simply the mean of Y, or, more specifically, the **unconditional mean** of Y, since its value is not conditional on any value of X.

3.6   Assumptions for inference with a statistical model

Inference refers to using the fit model to generalize from the sample to the population, which assumes that the response is drawn from some specified probability distribution (Normal, or Poisson, or Bernouli, etc.). Throughout this text, I emphasize reporting and interpreting point estimates and confidence intervals. Another kind of inference is a **significance test**, which is the computation of the probability of "seeing the data" or something more extreme than the data, given a specified null hypothesis. A significance test results in a **p-value**, which can be reported with the point estimate and confidence interval. Somewhat related to a significance test is a hypothesis test, or what is now often perjoratively called a **Null-Hypothesis Signficance Test** (NHST), in which the p-value from a significance test is compared to a pre-specified error rate called α. NHST may be useful for some very limited kinds of science but, in general, is not useful for most biological research and, instead, leads to large misconceptions. A general rule of thumb is, do not compare a reported p-value to α.

1. The data were generated by a process that is "linear in the parameters", which means that the different components of the model are added together. This additive part of the model containing the parameters is the linear predictor in specifications (3.3) and (16.7) above. For example, a cubic polynomial model

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \tag{3.11}$$

is a linear model, even though the function is non-linear, because the different components are added. Because a linear predictor is additive, it can be compactly defined using matrix algebra

$$E(Y|X) = \mathbf{X}\boldsymbol{\beta} \tag{3.12}$$

where $mathbfX$ is the **model matrix** and $\boldsymbol{\beta}$ is the vector of parameters. We discuss these more in chapter xxx.

A **Generalized Linear Model** (GLM) has the form $g(\mu_i) = \eta_i$ where η (the Greek letter "eta") is the linear predictor

$$\eta = \mathbf{X}\boldsymbol{\beta} \tag{3.13}$$

GLMs are extensions of linear models. There are non-linear models that are not linear in the parameters, that is, the predictor is not a simple dot product of the model matrix and a vector of parameters. For example, the Michaelis-Menten model is a non-linear model

$$\text{E}(Y|X) = \frac{\beta_1 X}{\beta_2 + X} \tag{3.14}$$

that is non-linear in the parameters because the parts are not added together. This text covers linear models and generalized linear models, but not non-linear models that are also non-linear in the parameters.

2. The draws from the probability distribution are **independent**. Independence implies **uncorrelated** observations (the Y), that is, for any two observations with the same value of X, we cannot predict the value of one given the value of the other. For example, in the vole data above, uncorrelated implies that we cannot predict the lifespan of one vole within the Vitamin E treatment given the lifespan of another vole in the Vitamin E treatment. For linear models, this assumption is often stated as "independent errors" (the ε in model (3.3)) instead of independent observations.

There are lots of reasons that observations might be correlated. In the vole example, perhaps the voles were housed in batches of 5 individuals, and slight differences in the environment among the housing containers, caused all the voles in some containers to have shorter lifespans than expected given their treatment assigment and all voles in other containers to have longer lifespans than expected given their treatment assigment. More generally, if there are measures both within and among experimental units (field sites or humans or rats) then we'd expect the measures within the same unit to err from the model in the same direction. Multiple measures within experimental units (a site or individual) creates "clustered" observations. Lack of independence or clustered observations can be modeled using models with **random effects**. These models go by many names including linear mixed models (common in Ecology), hierarchical models, multilevel models, and random effects models. A linear mixed model is a variation of model (3.3). This text introduces linear mixed models in chapter xxx.

Measures that are taken from sites that are closer together or measures taken closer in time or measures from more closely related biological species will tend to have more similar values than measures taken from sites that are further apart or from times that are further apart or from species that are less closely related. Space and time and phylogeny create **spatial and temporal and phylogenetic autocorrelation**. Correlated error due to space or time or phylogeny can be modeled with **Generalized Least Squares** (GLS) models. A GLS model is a variation of model (3.3).

3.7 Specific assumptions for inference with a linear model

1. **Constant variance** or **homoskedasticity**. The most common way of thinking about this is the error term ε has constant variance, which is a short way of saying that random draws of ε in model (3.3) are all from the same (or **identical**) distribution. This is explicitly stated in the second line of model specification (3.3). If we were to think about this using model specification (16.7), then homoskedasticity means that σ in $N(\mu, \sigma)$ is constant for all observations (or that the *conditional* probability distributions are identical, where *conditional* would mean adjusted for μ)

Many biological processes generate data in which the error is a function of the mean. For example, measures of biological variables that grow, such as lengths of body parts or population size, have variances that "grow" with the mean. Or, measures of counts, such as the number of cells damaged by toxin, the number of eggs in a nest, or the number of mRNA transcripts per cell have variances that are a function of the mean. Heteroskedastic error can be modeled with **Generalized Least Squares**, a generalization of the linear model, and with **Generalized Linear Models** (GLM), which are "extensions" of the classical linear model.

2. Normal or **Gaussian** probability distribution. As above, the most common way of thinking about this is the error term ε is Normal. Using model specification (16.7), we'd say the conditional probablity distribution of the response is normal. A normal probability distribution implies that 1) the response is continuous and 2) the conditional probability is symmetric around mu_i. If the conditional probability distribution has a long left or right tail it is **skewed** left or right. Counts (number of cells, number

of eggs, number of mRNA transcripts) and binary responses (sucessful escape or sucessful infestation of host) are not continuous and often often have asymmetric probablity distributions that are skewed to the right and while sometimes both can be reasonably modeled using a linear model they are more often modeled using generalized linear models, which, again, is an extension of the linear model in equation (16.7).

A common misconception is that inference from a linear model assumes that the *response* (Y) is normally distributed. Both the "linear model" and "statistical model" ways of specifying the model show precisely why this conception is wrong. Model (3.3) states explicitly that it is the error that has the normal distribution – the distribution of Y is a mix of the distribution of X and the error. Model (16.7) states that the conditional outcome has a normal distribution, that is, the distribution after adjusting for variation in X.

3.8   "Statistical model" or "regression model"?

Statistical modeling terminology can be confusing. The X variables in a statistical model may be quantitative (continuous or integers) or categorical (names or qualitative amounts) or some mix of the two. Linear models with all quantitative independent variables are often called "regression models." Linear models with all categorical independent variables are often called "ANOVA models." Linear models with a mix of quantitative and categorical variables are often called "ANCOVA models" if the focus is on one of the categorical X or "regression models" if there tend to be many independent variables. Other patterns occur. For example "ANCOVA models" often include interaction effects but "regression models" rarely do. To avoid thinking of statistical analysis as "regression vs. ANOVA" (the type of thinking encouraged by many textbooks in biostatistics), I will most often use the term "statistical model" for general usage, and use a more specific term only to emphasize something about the model in that particluar context.

3.9   GLM vs. GLM vs. GLS

Linear models are sometimes called "general linear models" with the abbreviation GLM. This is unfortunate because the abbreviation GLM usually refers to **generalized linear models**. Regardless, don't confuse either version of GLM with GLS, which is the abbreviation of **generalized least squares**. GLS generalizes the linear model to allow for heteroskedastic and/or correlated error (using the "linear model" way of thinking about model specification)

Chapter 4

Variability and Uncertainty (Standard Deviations, Standard Errors, Confidence Intervals)

Uncertainty is the stuff of science. A major goal of statistics is measuring uncertainty. What do we mean by uncertainty? Uncertainty is the error in estimating a parameter, such as the mean of a sample, or the difference in means between two experimental treatments, or the predicted response given a certain change in conditions. Uncertainty is measured with a **variance** or its square root, which is a **standard deviation**. The standard deviation of a statistic is also (and more commonly) called a **standard error**.

Uncertainty emerges because of variability. In any introductory statistics class, students are introduced to two measures of variability, the "standard deviation" and the "standard error." These terms are absolutely fundamental to statistics – they are the start of everything else. Yet, many biology researchers confuse these terms and certainly, introductory students do too.

When a research biologist uses the term "standard deviation," they are probably referring to the sample standard deviation which is a measure of the variability of a sample. When a research biologist uses the term "standard error," they are probably referring to the standard error of a mean, but it could be the standard error of another statistics, such as a difference between means or a regression slope. An important point to remember and understand is that all standard errors *are* standard deviations. This will make more sense soon.

4.1 The sample standard deviation vs. the standard error of the mean

4.1.1 Sample standard deviation

The sample standard deviation is a measure of the variability of a sample. For example, were we to look at a histological section of skeletal muscle we would see that the diameter of the fibers (the muscle cells) is variable. We could use imaging software to measure the diameter of a sample of 100 cells and get a **distribution** like this
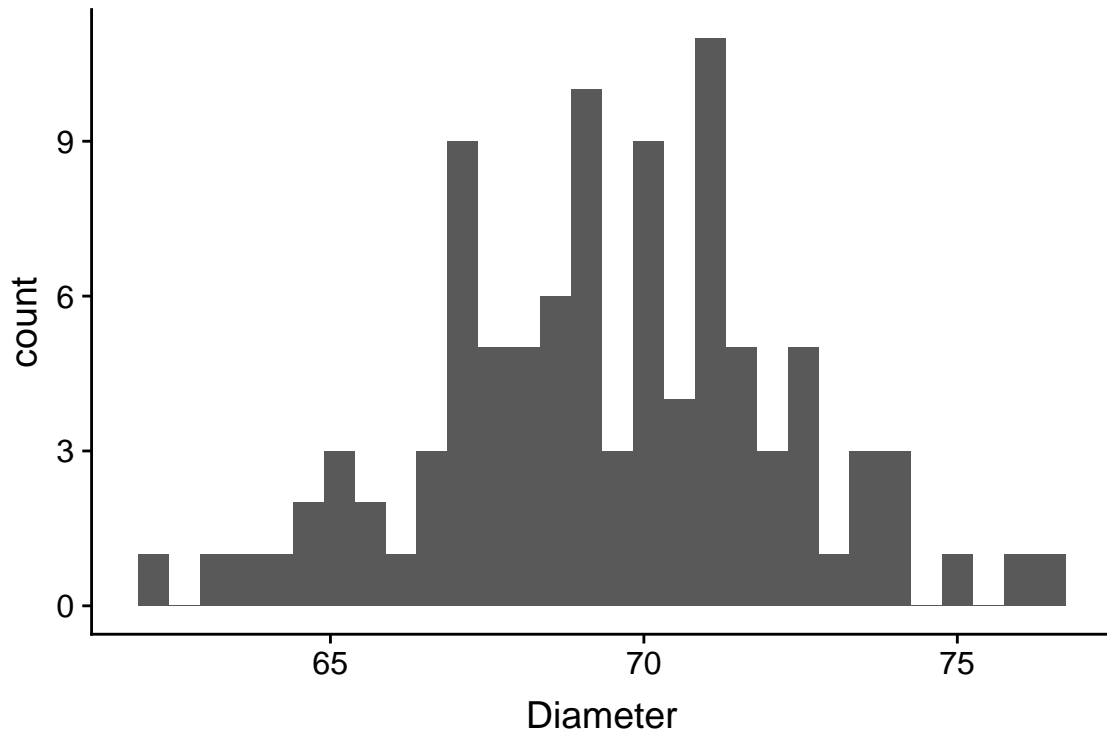
The mean of this sample is 69.4 and the standard deviation is 2.8. The standard deviation is the square root of the variance, and so computed by

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \overline{y})^2}{n - 1}} \tag{4.1}$$

Memorize this equation. To understand the logic of this measure of variability, note that $y_i - \overline{y}$ is the **deviation** of the ith value from the sample mean, so the numerator is the sum of squared deviations. The numerator is a sum over n items and the denominator is $n - 1$ so the variance is (almost!) an averaged squared deviation. More variable samples will have bigger deviations and, therefore, bigger average squared deviations. Since the standard deviation is the square root of the variance, a standard deviation is the square root of an average squared deviation. This makes it similar in value to the averaged deviation (or average of the absolute values of the deviations since the average deviation is, by definition of a mean, zero).

Notes on the variance and standard deviation

1. Variances are additive but standard deviations are not. This means that the variance of the sum of two independent (uncorrelated) random variables is simply the sum of the variances of each of the variables. This is important for many statistical analyses.
2. The units of variance are the square of the original units, which is awkward for interpretation. The units of a standard deviation is the same as that of the original variable, and so is much easier to interpet.
3. For variables that are approximately normally distributed, we can map the standard deviation to the quantiles of the distribution. For example, 68% of the values are within one standard deviation of the mean, 95% of the values are within two standard deviations, and 99% of the values are within three standard deviations.

4.1.2   Standard error of the mean

A standard error of a statistic is a measure of the precision of the statistic. The standard error of the mean is a measure of the precision of the estimate of the mean. The smaller the standard error, the more precise the estimate. The standard error of the mean (SEM) is computed as

$$SEM = \frac{s_y}{\sqrt{n}} \tag{4.2}$$

The SEM is often denoted $s_{\bar{y}}$ to indicate that it is a standard deviation of the mean $(\bar{y})$. In what sense is a standard error a measure of variability? This is kinda weird. If we sample 100 cells in the slide of muscle tissue and compute the mean diameter, how can the mean have a standard deviation? There is only one value! To understand how the SEM is a standard deviation, imagine 1) resampling 100 cells and 2) recomputing a mean from the re-sampled data, then repeating this resampling and recomputation an infinite number of times and each time, you write down the newly computed mean. The true standard error of the mean is the standard deviation of this infinitely long column of means. This means that a standard error of the mean, computed from a single sample using equation (4.2) is itself a sample statistic.

Notes on standard errors

1. The SEM is only one kind of standard error. A standard deviation can be computed for any statistic – these are all standard errors. For some statistics, such as the mean, the standard error can be computed directly using an equation, such as that for the SEM (equation (4.2). For other statistics, a computer intensive method such as the **bootstrap** is necessary to compute a standard error. We will return to the bootstrap at the end of this chapter.
2. The units of a standard error are the units of the measured variable.
3. A standard error is proportional to sample variability (the sample standard deviation, s_y) and inversely proportional to sample size (n). Sample variability is a function of both natural variation (there really is variation in diameter among fibers in the quadriceps muscle) and measurement error (imaging software with higher resolution can measure a diameter with less error). Since the SEM is a measure of the precision of estimating a mean, this means this precision will increase (or the SEM will decrease) if 1) an investigator uses methods that reduce measurement error and 2) an investigator computes the mean from a larger sample.
4. This last point (the SEM decreases with sample size) seems obvious when looking at equation (4.2), since n is in the denominator. Of course n is also in the denominator of equation (4.1) for the sample standard deviation but the standard deviation does not decrease as sample size increases. First this wouldn't make any sense – variability is variability. A sample of 10,000 cell diameters should be no more variable than a sample of 100 cell diameters (think about if you agree with this or not). Second, this should also be obvious from equation (4.1). The standard deviation is the square root of an average and averages don't increase with the number of things summed since both the the numerator (a sum) and denominator increase with n.

4.2   Using Google Sheets to generate fake data to explore the standard error

In statistics we are interested in estimated parameters of a **population** using measures from a **sample**. The goal in this section is to use Google Sheets (or Microsoft Excel) to use fake data to discover the behavior of sampling and to gain some intuition about uncertainty using standard errors.

4.2.1   Steps

1. Open Google Sheets

2. In cell A1 type "mu". mu is the greek letter μ and is very common notation for the poplation value (the TRUE value!) of the mean of some hypothetical measure. In cell B1, insert some number as the value of μ. Any number! It can be negative or positive.

3. In cell A2 type "sigma". sigma is the greek letter σ. σ^2 is very common (universal!) notation for the population (TRUE) variance of some measure or parameter. Notice that the true (population) values of the mean and variance are greek letters. This is pretty standard in statistics. In cell B2, insert some positive number (standard deviations are the positive square roots of the variance).

4. In cell A8 type the number 1

5. In cell A9 insert the equation "=A8 + 1". What is this equation doing? It is adding the number 1 to to the value in the cell above, so the resulting value should be 2.

6. In Cell B8, insert the equation "=normsinv(rand())*\$B\$2 + \$B\$1". The first part of the equation creates a random normal variable with mean 0 and standard deviation 1. multiplication and addition transform this to a random normal variable with mean μ and standard deviation σ (the values you set in cells B1 and B2).

7. copy cell B8 and paste into cell B9. Now Higlight cells A9:B9 and copy the equations down to row 107. You now have 100 random variables sampled from a infinite population with mean μ and standard deviation σ.

8. In cell A4 write "mean 10". In cell B4 insert the equation "=average(B8:B17)". The resulting value is the **sample mean** of the first 10 random variables you created. Is the mean close to μ?

9. In cell A5 write "sd 10". In cell B5 insert the equation "stdev(B8:B17)". The result is the **sample standard deviation** of the first 10 random variables. Is this close to σ?

10. In cell A6 write "mean 100". In cell B6 insert the equation "=average(B8:B107)". The resulting value is the **sample mean** of the all 100 random variables you created. Is this mean closer to μ than mean 10?

11. In cell A7 write "sd 100". In cell B7 insert the equation "=stdev(B8:B107)". The resulting value is the **sample standard deviation** of the all 100 random variables you created. Is this SD closer to σ than sd 10?

The sample standard deviation is a measure of the variability of the sample. The more spread out the sample (the further each value is from the mean), the bigger the sample standard deviation. The sample standard deviation is most often simply known as "The" standard deviation, which is a bit misleading since there are many kinds of standard deviations!

Remember that your computed mean and standard deviations are estimates computed from a sample. They are estimates of the true values μ and σ. Explore the behavior of the sample mean and standard deviation by re-calculating the spreadsheet. In Excel, a spreadsheet is re-calculated by simultaneously pressing the command and equal key. In Google, command-R recalculates but is painfully slow. Instead, if using Google Sheets, just type the number 1 into a blank cell, and the sheet recalculates quickly. Do it again. And again.

Each time you re-calculate, a new set of random numbers are generated and the new means and standard deviations are computed. Compare mean 10 and mean 100 each re-calculation. Notice that these estimates are variable. They change with each re-calculation. How variable is mean 10 compared to mean 100? The variability of the estimate of the mean is a measure of **uncertainty** in the estimate. Are we more uncertain with mean 10 or with mean 100? This variability is measured by a standard deviation. This **standard deviation of the mean** is also called the **standard error of the mean**. Many researchers are loose with terms and use "The" standard error to mean the standard error of the mean, even though there are many kinds of standard errors. In general, "standard error"" is abbreviated as "SE." Sometimes "standard error of the mean" is specifically abbreviated to "SEM."

The standard error of the mean is a measure of the precision in estimating the mean. The smaller the value the more precise the estimate. The standard error of the mean *is* a standard deviation of the mean. This is kinda weird. If we sample a population one time and compute a mean, how can the mean have a standard deviation? There is only one value! And we compute this value using the sample standard deviation: $SEM = \frac{SD}{\sqrt{N}}$. To understand how the SEM is a standard deviation, Imagine recalculating the spread sheet an infinite number of times and each time, you write down the newly computed mean. The standard error of the mean is the standard deviation of this infinitely long column of means.

4.3 Using R to generate fake data to explore the standard error

note that I use "standard deviation" to refer to the sample standard deviation and "standard error" to refer to the standard error of the mean (again, we can compute standard errors as a standard deviation of any kind of estimate)

4.3.1 part I

In the exercise above, you used Google Sheets to generate p columns of fake data. Each column had n elements, so the matrix of fake data was $n \times m$ (it is standard in most fields to specify a matrix as rows by columns). This is *much* easier to do in R and how much grows exponentially as the size of the matrix grows.

To start, we just generate a $n \times p$ matrix of normal random numbers.

```r
# R script to gain some intuition about standard deviation (sd) and standard error (se)
# you will probably need to install ggplot2 using library(ggplot2)
n <- 6 # sample size
p <- 100 # number of columns of fake data to generate
fake_data <- matrix(rnorm(n*p, mean=0, sd=1), nrow=n, ncol=p) # create a matrix
```

the 3rd line is the cool thing about R. In one line I'm creating a dataset with n rows and p columns. Each column is a sample of the standard normal distribution which by definition has mean zero and standard deviation of 1. But, and this is important, any sample from this distribution will not have exactly mean zero and standard deviation of 1, because it's a sample, the mean and standard deviation will have some small errror from the truth. The line has two parts to it: first I'm using the function "rnorm" (for random normal) to create a vector of n*m random, normal deviates (draws from the random normal distribution) and then I'm organizing these into a matrix (using the function "matrix")

To compute the vector of means, standard deviations, and standard errors for each column of `fake_data`, use the `apply()` function.

```r
means <- apply(fake_data,2,mean) # the apply function is super useful
sds <- apply(fake_data,2,sd)
sems <- sds/sqrt(n)
```

`apply()` is a workhorse in many R scripts and is often used in R scripts in place of a for-loop (see below) because it takes fewer lines of code.

The SEM is the standard deviation of the mean, so let's see if the standard deviation of the means is close to the true standard error. We sampled from a normal distribution with SD=1 so the true standard is

```r
1/sqrt(n)
```

```
## [1] 0.4082483
```

and the standard deviation of the p means is

```r
sd(means)
```

```
## [1] 0.3731974
```

Questions

1. how close is `sd(means)` to the true SE?
2. change p above to 1000. Now how close is sd(means) to the true SE?
3. change p above to 10,000. Now how close is sd(means) to the true SE?

4.3.2 part II - means

This is a visualization of the spread, or variability, of the sampled means

```
qplot(means)
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
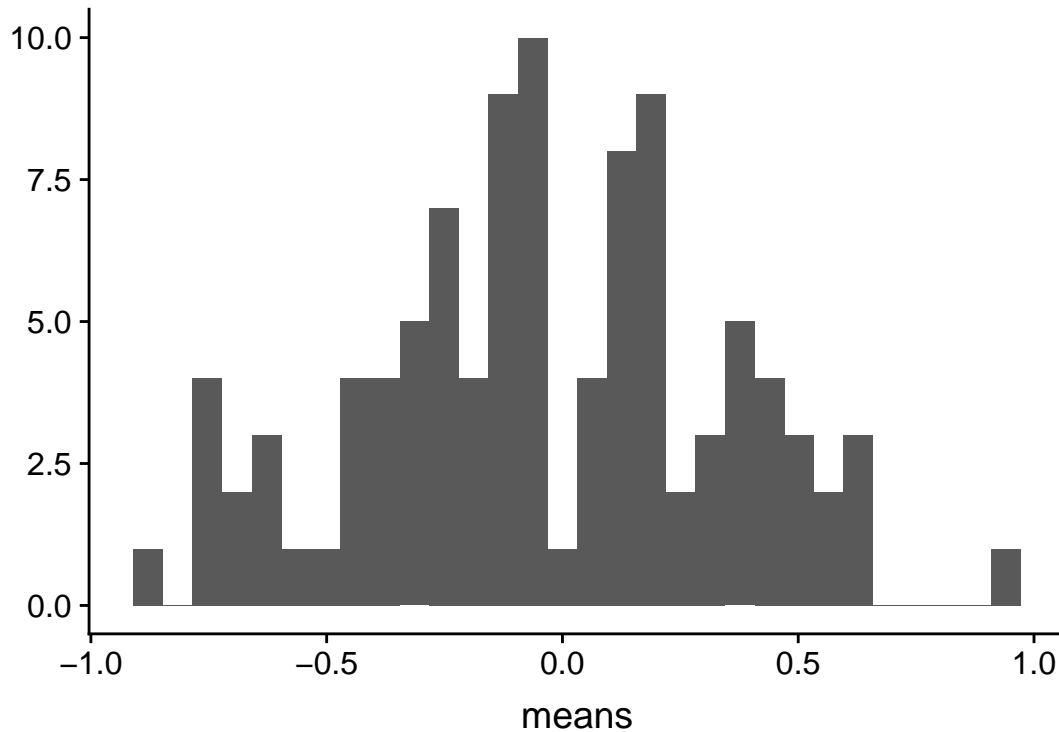
Compute the mean of the means

```
mean(means)
```

[1] -0.039961

Questions

1. Remember that the true mean is zero. How close, in general, are the sampled means to the true mean. How variable are the means? How is this quantified?
2. change n to 100, then replot. Are the means, in general, closer to the true mean? How variable are the means now?
3. Is the mean estimated with $n = 100$ closer to the truth, in general, then the mean estimated with $n = 6$?
4. Redo with $n = 10000$

4.3.3 part III - how do SD and SE change as sample size (n) increases?

```
mean(sds)
```

[1] 1.017144

Questions

1. what is the mean of the standard deviations when n=6 (set p=1000)

2. what is the mean of the standard deviations when n=100 (set p=1000)
3. when n = 1000? (set p=1000)
4. when n = 10000? (set p=1000)
5. how does the mean of the standard deviations change as n increases (does it get smaller? or stay about the same size)
6. repeat the above with SEM

```
mean(sems)
```

[1] 0.4152472

Congratulations, you have just done a Monte Carlo simulation!

4.3.4   Part IV – Generating fake data with for-loops

A **for-loop** is used to iterate a computation.

```
n <- 6 # sample size
n_iter <- 10^5 # number of iterations of loop (equivalent to p)
means <- numeric(n_iter)
sds <- numeric(n_iter)
sems <- numeric(n_iter)
for(i in 1:n_iter){
  y <- rnorm(n) # mean=0 and sd=1 are default so not necessary to specify
  means[i] <- mean(y)
  sds[i] <- sd(y)
  sems[i] <- sd(y)/sqrt(n)
}
sd(means)
```

[1] 0.4090702

```
mean(sems)
```

[1] 0.3883867

Questions

1. What do `sd(means)` and `mean(sems)` converge to as `n_iter` is increased from 100 to 1000 to 10,000?
2. Do they converge to the same number?
3. Should they?
4. What is the correct number?

Question number 4 is asking what is E(SEM), the "expected standard error of the mean". There is a very easy formula to compute this. What is it?

4.4   Bootstrapped standard errors

A standard error of the mean is the expected standard deviation of an infinite number of hypothetically re-sampled means. A bootstrap standard error of a statistic is the standard deviation of the statistic from a finite number of *resamples* of the data.

Let's download some data to explore this concept. The data are archived at Dryad Repository.

1. URL: https://datadryad.org//resource/doi:10.5061/dryad.31cc4
2. file: RSBL-2013-0432 vole data.xlsx
3. sheet: COLD VOLES LIFESPAN

The data are the measured lifespans of the short-tailed field vole (*Microtus agrestis*) under three different experimental treatments: vitamin E supplementation, vitamin C supplementation, and control (no vitamin supplementation). Vitamins C and E are antioxidants, which are thought to be protective of basic cell function since they bind to the cell-damaging reactive oxygen species that result from cell metabolism.

I've read in the file using read_excel and converted to a data.table named **vole**. I used **setnames** to rename the columns to lifespan, control, vitamin_E, and vitamin_C. The data are in a **wide format** – that is instead of a single "treatment" column, there are three columns ("control", "vitamin C", "vitamin E") with value = 1, if that row (or lifespan) was assigned the treatment of the column label and zero otherwise. In general, we want data.tables to be in long format.

Compute the standard error of the mean of the lifespan for the control group using equation (4.2). One simple way to do this for the control group is to extract the subset of the data satisfying the condition control = 1 (the value in the column "control" equals 1). In R, these conditional querries use ==.

```r
control_voles <- na.omit(vole[control==1, lifespan]) # subset of data satisfying condition and
# omitting missing data, if these exist
n <- length(control_voles) # the sample size
se <- sd(control_voles)/sqrt(n-1) # standard error of the mean of the control voles
```

Okay, the SEM using equation (4.2) is 31.9. Let's compare this with a bootstrap estimate of the SEM. There are several R packages that do a bootstrap but here I've coded one using a for-loop both to practice with for-loops and to help communicate what a bootstrap *is*.

The basic algorithm for a bootstrap is

1. **re-sample the data with replacement**. Here "the data" is the set of lifespans for the Control voles. "Resample with replacement" means to sample n times from the full set of values. If we were to do this manually, we would i) write down each value of Control lifespan on its own piece of paper and throw all pieces into a hat. ii) pick a paper from the hat, add its value to sample i, and return the paper to the hat. iii) repeat step ii n times, where n is the original sample size. The new sample contains some values multiple times (papers that were picked out of the hat more than once) and is missing some values (papers that were not picked out in any of the n picks).
2. Compute the statistic with the new sample.
3. Repeat steps 1 and 2 p times, which will result in a p-length vector of the statistic. Typically p is somewhere between 2000 and 10,000.
4. For the standard error of the statistic, compute the standard deviation of the re-sampled statistic.

A for-loop automates this algorithm.

```r
set.seed(1)
n_iter <- 2000 # number of bootstrap iterations, or p
means <- numeric(n_iter) # we will save the means each iteration to this
inc <- 1:n # the first sample is the actual sample
for(iter in 1:n_iter){ # this line sets up the number of iterations, p
  means[iter] <- mean(control_voles[inc]) # inc is the set of rows to include in the computation of the
  inc <- sample(1:n, replace=TRUE) # re-sample for the next iteration
}
se_boot <- sd(means)
getwd()
```

```
## [1] "/Users/jwalker/Documents/Github projects/Bookdown projects/applied-biostatistics"
```

The SEM using the bootstrap is 31.43. Again, the SEM using equation (4.2) is 31.9.

4.5 Confidence Interval

Here I introduce a **confidence interval** (abbreviated to CI) using the mean of a sample but the concept is easily generalized to any parameter of a statistical model. The mean of the Control voles is 503.4 and the SE of the mean is 31.9. The SE is used to construct the lower and upper boundary of a "1 - α" confidence interval using `lower <- mean(x) - qt(alpha/2, df = n)*se(x)` and `upper <- mean(x) + qt(alpha/2, df = n)*se(x)`.

$$upper = mean+ \tag{4.3}$$

A confidence interval of the mean is a measure of the uncertainty in the estimate of the mean. A 95% confidence interval has a 95% probability (in the sense of long-run frequency) of containing the true mean. This probability is a property of the population of intervals that could be computed using the same sampling and measuring procedure. It is not correct, without further assumptions, to state that "there is a 95% probability that the true mean lies within the interval". Perhaps a more useful interpretation is that the interval contains the range of means that are consistent with the data, in the sense that a *t*-test would not reject the null hypothesis of a difference between the estimate and any value within the interval (this interpretation does not imply anything about the true value).

Chapter 5

Covariance and Correlation

Variance is one of two major concepts in statistical modeling. The second is **covariance**, which arises when two variables measured on the same unit vary together. "Vary together" means that if we measured leg length and arm length on each individual in a sample of humans, we'd expect individuals with long arms to also have long legs while those with short arms to have short legs. "Measured on the same unit" means that we measure both leg length and arm length in each individual of a sample – we cannot compute a covariance if we measure legs in one sample and arms in a second. Covariance can be positive or negative. It is positive when the tendency is for both values to be large or both values to be small. It is negative when the tendency is for one value to be small when the other is large. Positive and negative covariance are easily visualized with a scatterplot 5.1.

1. Covariance

$$\mathrm{COV}[X, Y] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \tag{5.1}$$

Compare this to the equation for the variance. In (5.1), the numerator in the sum is the product of two different deviations (one for each variable) instead of the product of a deviation with itself.

2. Correlation

3. Regression
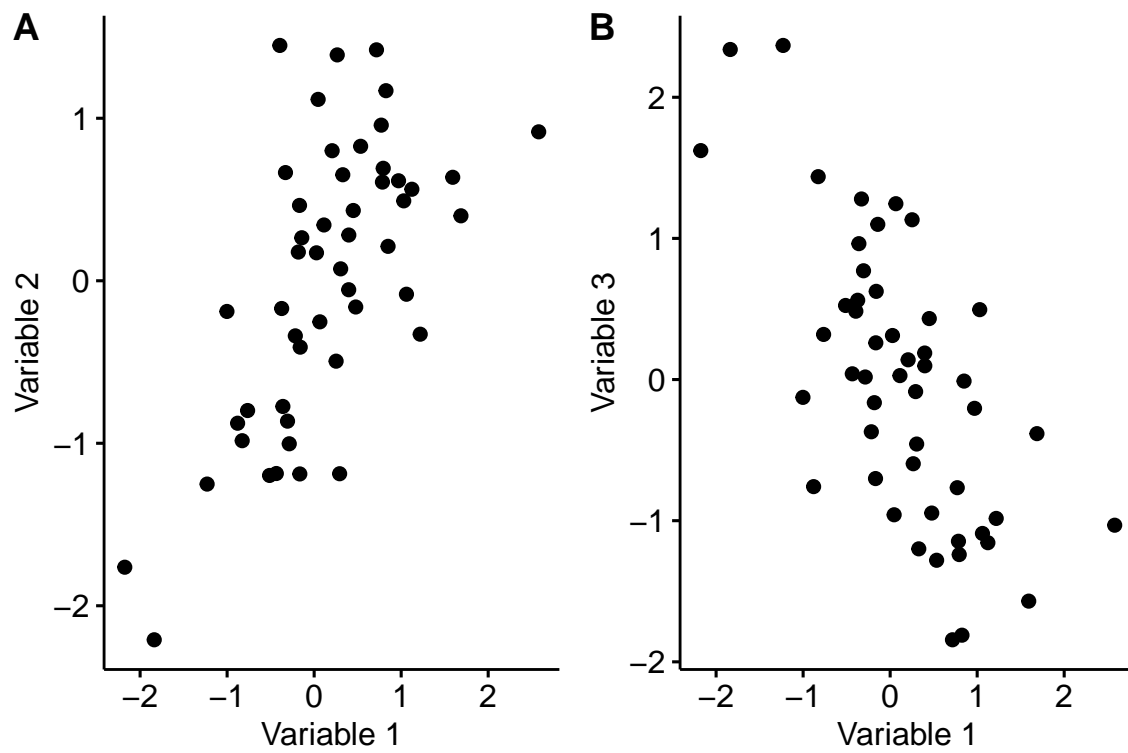
Figure 5.1: Scatterplot illustrating two variables with (A) positive covariance and (B) negative covariance

Chapter 6

P-values

6.1 p-values

Let's use the vole data to introduce the p-value. The vole data were archived in a format that requires some cleaning before it can be analyzed. The "lifespan (days)" column contains the lifespan for all three treatment levels, that is, the lifespan values are "stacked" or in "long" format. What is unusual is that there is no "treatment" column specifying the three treatment levels (control, vitamin E, vitamin C). Instead there are control, vitamin E, and vitamin C columns that contain a 1, if the lifespan value belongs to that treatment, and a NULL value if not. On import, the NULL value is replaced with NA, to indicate "missing". In order to analyze the data, the three treatment assignment columns need to be combined into a single "treatment" column.

```
folder <- "Data from Deleterious consequences of antioxidant supplementation on lifespan in a wild-deri
filename <- "RSBL-2013-0432 vole data.xlsx"
file_path <- here(data_path, folder, filename)
vole <- data.table(read_excel(file_path, sheet="COLD VOLES LIFESPAN", range="a2:d98"))
setnames(vole, old=colnames(vole), new=c("lifespan", "control", "vitamin_E", "vitamin_C"))
# not sure why ifelse fails so this is alternative
vole[control==1, treatment:="control"]
vole[vitamin_C==1, treatment:="vitamin_C"]
vole[vitamin_E==1, treatment:="vitamin_E"]
```

A typical analysis would compare the means of the two supplement treatment levels with the control treatment level and use a p-value as evidence of "an effect".

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
vitamin_C - control	-115.1	54	93	-223.2	-6.9	-2.1	0.037
vitamin_E - control	-89.9	52	93	-194.1	14.3	-1.7	0.090
vitamin_E - vitamin_C	25.2	65	93	-103.8	154.1	0.4	0.699

The table above gives the SE, t and p-value for each pairwise contrast (difference in means) among the three treatment levels. A typical report (one with several misconceptions) might read

"We found a significant effect of Vitamin C (t = -2.1, $p = 0.037$) on lifespan, but no effect of vitamin E (t = -1.7, $p = 0.09$) on lifespan."

A p value *is a continuous measure of evidence against the null.* As long as the data approximate the assumptions of the null hypothesis pretty well, a very small p-value, such as 0.002 or 0.0005, is pretty good evidence against the null hypothesis – but does not mean "an effect exists". To show an effect exists, we should have small p-values in multiple replicates *and* we should rigorously probe the hypothesis with different experiments that challenge the hypothesis in different ways. A small p is evidence for a research program to

move forward with replication and probing. A big p-value, say 0.22 or 0.76, is pretty weak evidence against the null, but does not mean "there is no effect." If an experiment is well designed, a big p could suggest abandoning any hypotheses that predict biologically consequential effects. Unfortunately, a big p could also reflect a weak experimental design. Between small and big p values, such as 0.009 or 0.011, problems arise. These intermediate p-values beg for replication. A major problem of inference using p values is that there is no sharp boundaries between these three regions. Instead biologists typically use the $p < 0.05$ as a sharp boundary to declare that an effect exists or not.

Okay. so what *is* a p-value? When we do a t-test, we get a p-value. There are several ways to think about this probability. The most compact way is $P(data|null)$, which is literally read as the probability of the data given the null (or "conditional" on the null), but is really short for *the probability of the data, or something more extreme than the data, given that the null hypothesis is true.* The "probability of the data" is kinda vague. More specifically, we mean the probability of some statistic about the data such as the difference in means between group A and group B or the t-value associated with this difference. So, a bit more formally, the probability returned in a t-test is $\text{prob}(t \geq t_{obs}|H_0)$. This is the long run frequency of observing a t-value as big or bigger than the observed t-value (the one you actually got with your data) if the null is true. Let's parse this into "long run frequency of observing a t-value as big or bigger than the observed t-value" and "null is true".

A thought experiment: You open a google sheet and insert 12 standard, normal random deviates (so the true mean is zero and the true variance is one) in Column A, rows 1-12. You arbitrarily assign the first six values (rows 1-6) to treatment A and the second six values (rows 7-12) to treatment B. You use the space immediately below these data to compute the mean of treatment A, the mean of treatment B, the difference in means (A - B), and a t-value. Unfortunately, google sheets doesn't have a t-value function so you'd have to compute this yourself. Or not, since this is a thought experiment. Now "fill right" or copy and paste these functions into 999 new columns. You now have 1000 t-tests. The expected value of the difference in means is zero (why?) but the actual values will form a normal distribution about zero. Most will be close to zero (either in the negative or positive direction) but some will be further from zero. The expected t-value will also be zero (why?) and the distribution of these 1000 t values will look normal but the tails are a little fuller. This row of t values is a null distribution, because in generating the data we used the exact same formula for the values assigned to A and the values assigned to B. Now think of a t-value in your head, say 0.72 (remember that t values will largely range from about -3 to +3 although the theoretical range is $-\infty$ to $+\infty$. What is the probability of observing a t of 0.72 *or bigger* if the null is true? Look at the row of t-values! Count the number of $t \geq 0.72$ and then divide by the total number of t-values in the row (1000) and you have a probability computed as a frequency. But remember the frequentist definition is the long run frequency, or the expected frequency at the limit (when you've generated not 1000 or even 1,000,000 but an infinite number of columns and t-values).

Some asides to the thought experiment: First, why "as big or bigger" and not just the probability of the value itself? The reason is that the probability of finding the exact t is 1/infinity, which doesn't do us much good. So instead we compute the probability of finding t as big, or bigger, than our observed t. Second, the t-test probability described above is a "one-tail probability". Because a difference can be both in the positive direction and the negative direction, we usually want to count all the $t \geq 0.72$ and the $t \leq -0.72$ and then add these two counts to compute the frequency of *as extreme or more extreme* values. This is called a "two-tailed probability" because we find extremes at both tails of the distribution. Third, we don't really count $t \geq 0.72$ but take advantage of the beautiful mathematical properties of the theoretical t distribution, which allows us to compute the frequentist probability (expected long range frequency) given the t-value and the degrees of freedom using the t-distribution.

Now what do I mean with the phrase "null is true"? Most people equate "null is true" with "no difference in means" but the phrase entails much more than this. Effectively, the phrase means that the p-value is based on modeling the real data with a theoretical sample in which all the points were randomly sampled from the same distribution and that the assignment of the individual points to treatment was random. This model means the theoretical sample has three properties: First, random assignment to treatment after sampling from the same distribution means that the expected means are the same, or put differently, the expected difference in means between the assigned groups is zero. Second, random assignment to treatment after

sampling from the same distribution *also* means that the expected variances of the two groups are equal. And third, random sampling means that the values of each point are independent – we cannot predict the value of one point knowing information about any other point. **Here is what is super important about this**: if we get a really low *p*-value, any one of these consequences may be untrue about our data, for example it could be that the true means of the two treatment groups really are different, or it could mean it is the variances that differ between the two groups, or it could mean that the data (or technically, the errors) are not independent of each other. This is why we need certain assumptions to make a *p*-value meaningful for empirical data. By assuming independent error and homogenous (equal) variances in our two samples, a low *p* value is evidence of unequal means.

6.2 Creating a null distribution.

Let's repeat: A pretty good definition of a *p*-value is: the long-run frequency of observing a test-statistic as large or larger than the observed statistic, if the null were true. A more succinct way to state this is

$$p = \text{prob}(t \geq t_o | H_o) \tag{6.1}$$

where t is a hypothetically sampled *t*-value from a null distribution, t_o is the observed *t*-value, and H_o is the null hypothesis. Part of the null hypothesis is the expected value of the parameter estimated is usually (but not always) zero – this can be called the nil null. For example, if there is no vitamin E effect on lifespan, then the expected difference between the means of the control and vitamin E treatment levels is zero. Or,

$$\text{E}(\bar{vitamin}_E - \bar{control} | H_o) = 0.0 \tag{6.2}$$

let's plot the data and look at the group means. Below is a strip chart of the vole data with superimposed treatment level means, using the function `ggstripchart` from the ggpubr package (can you make this?). I'm going to refer to this kind of chart as a "dot plot", which is what most biology researchers call this type of chart.
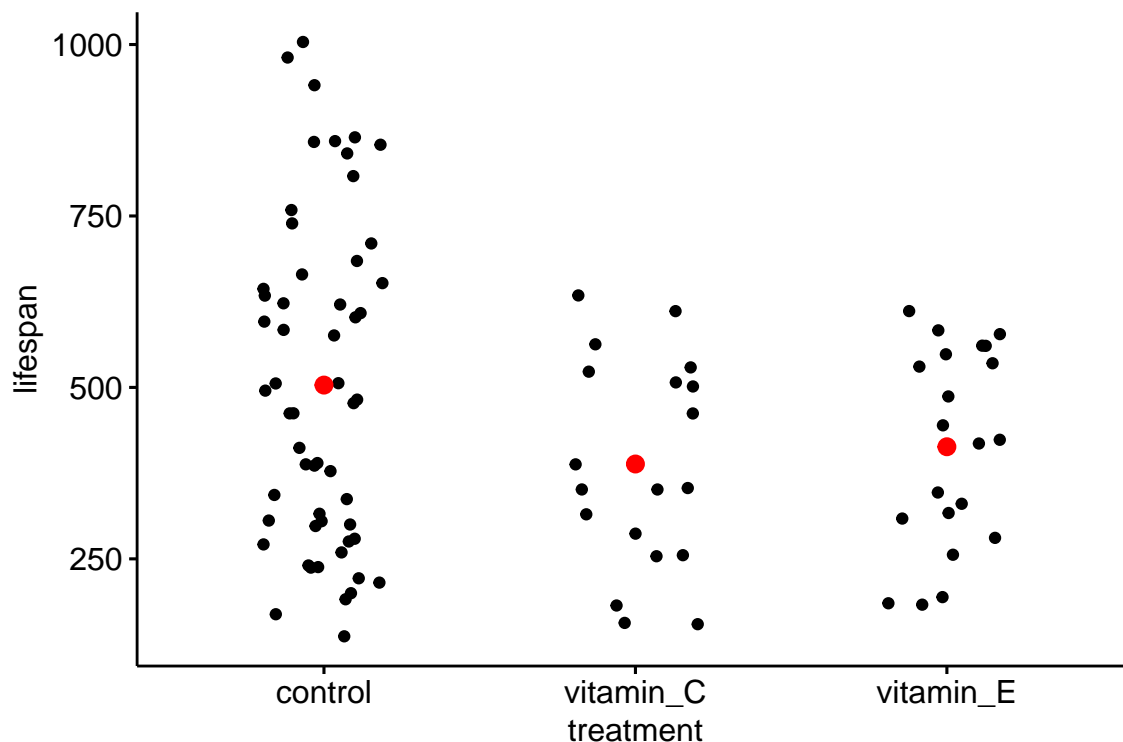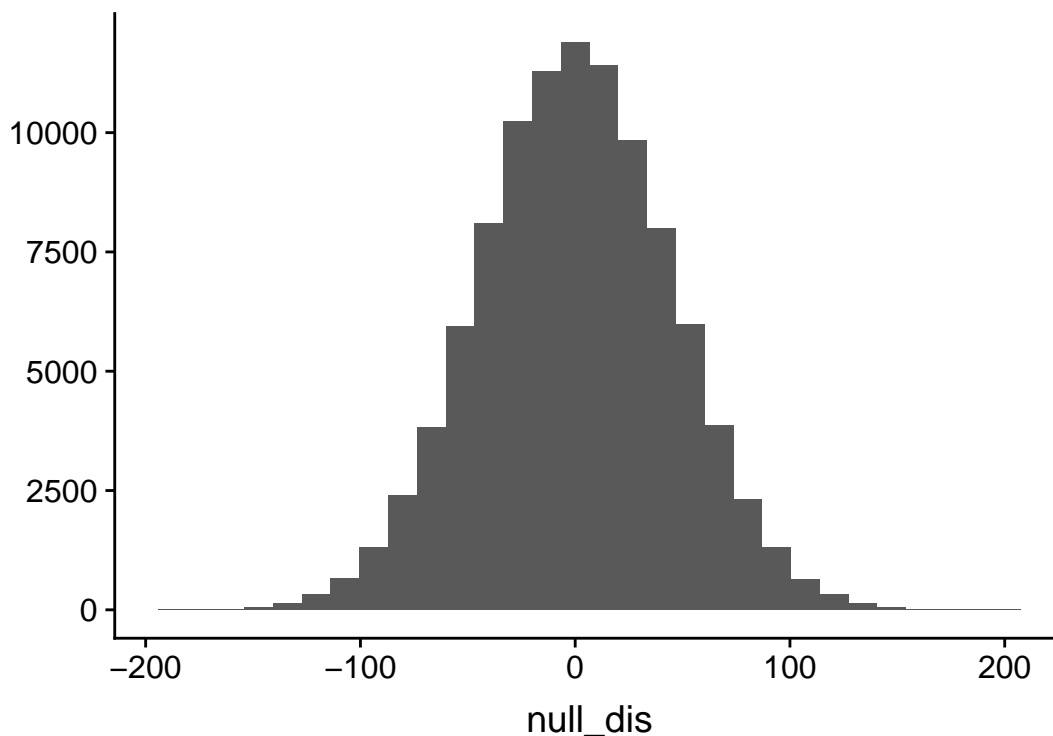
Figure 6.1: Null distribution for an infinitely large data set that looks curiously like the lifespans of the cold-rear voles from the control treatment.

6.2.1   the Null Distribution

The mean lifespan in the vitamin_E treatment is -89.9 days shorter than the mean lifespan in the control treatment. And, the mean lifespan in the vitamin_E treatment is -115.1 days shorter than the mean lifespan in the control treatment. These are the measured effects, or the **observed differences in means**. How confident are we in these effects? Certainly, if the researchers did the experiment with *two* control treatment groups, they would measure some difference in their means simply because of finite sampling (more specifically, the many, many random effects that contribute to lifespan will differ between the two control groups). So let's reframe the question: are the observed differences unusually large compared to a distribution of differences that would occur if there were no effect? That is, if the "null were true". To answer this, we compare our observed difference to this **null distribution**. This comparison gives the probability (a long-run frequency) of "sampling" a random difference from the null distribution of differences that is as large, or larger, than the observed difference.

What is a null distribution? It is the distribution of a statistic (such as a difference in means, or better, a *t*-value) if the null were true. Here, I am generating a null distribution that is relevant to the cold vole data. See if you can understand the script before reading the explanation below.

```
seed <- 1
n_iter <- 10^5 # number of iterations
mu <- mean(vole[treatment=='control', lifespan])
sigma <- sd(vole[treatment=='control', lifespan])
n <- nrow((vole[treatment=='control',]))
sample1 <- matrix(rnorm(n*n_iter, mean=mu, sd=sigma), nrow=n) # 100,000 samples (each size n)
sample2 <- matrix(rnorm(n*n_iter, mean=mu, sd=sigma), nrow=n) # 100,000 samples
null_dis <- apply(sample2, 2, mean) - apply(sample1, 2, mean)
qplot(null_dis)
```

What have we done above? We've simulated an infinitely large population of voles that have a distribution of lifespans similar to that of the cold-reared voles assigned to the control group. The mean μ and standard deviation σ of the simulated lifespan are equal to the observed mean and standard deviation of the lifespans of the control voles. Then, the script:

1. randomly sample 56 values from this population of simulated lifespans and assign to sample1. We sample 56 values because that is the sample size of our control in the experiment.
2. randomly sample 56 values from this population of simulated lifespans and assign to sample2.
3. compute the difference $\bar{Y}_{sample2} - \bar{Y}_{sample1}$.
4. repeat 1-3 100,000 times, each time saving the difference in means.
5. plot the distribution of the 100,000 differences using a histogram

The distribution of the differences is a null distribution. Notice that the mode of the null distribution is at zero, and the mean (-0.11584) is close to zero (if we had set n to infinity, the mean would be precisely zero). *The expected difference between the means of two random samples from the same population is, of course, zero.* Don't gloss over this statement if that is not obvious. The tails extend out to a little more than +100 and -100. What this means is that it would be rare to randomly sample two sets of data from the same population with mean μ and standard deviation σ and find a difference of, say, -257. In fact, in the 100,000 runs, there were no difference as large as |-257| (the absolute value of -257). The minimum and maximum differences sampled over the 100,000 iterations was -187 days and 201 days.

How do our observed differences compare? Let's focus on vitamin E. The vitamin_E effect is -89.9 days. There are 2110 sampled differences less than the observed value and 2126 greater than the absolute value of the observed value. Together this is 4236 so the frequency of differences from the simulated null distribution that as larger or larger than the observed difference is 0.042 (this compuation includes the observed value in both the numerator and denominator).

6.2.2 *t*-tests

A *t*-test is a test of differences between two values. These could be

1. the difference between the means of two samples (a "two-sample" *t*-test)
2. the difference between a mean of a sample and some pre-specified value (a "one-sample" *t*-test)
3. the difference between a coefficient from a linear model and a value (often zero)

A *t*-test compares an observed *t*-value to a *t*-distribution. The null distribution introduced above was a distribution of mean differences. This isn't generally useful, since the distribution of expected mean differences is a function of sample variability (standard deviations) in addition to sample size and, therefore, a mean-difference null distribution will be unique to every study. A *t*-distribution is a distribution of *t*-values under the null (statistical jargon for "given the null is true"), where a *t*-value is a difference standardized by its standard error. Standardizing by a standard deviation (remember that a standard error is an estimate of the statistic's standard deviation) removes the effect of sample variability on the distribution. A *t*-distribution, then, is only a function of sample size (or "degrees of freedom"). As n increases a t distribution converges on the standard, normal distribution.

The difference between the mean of the vitamin_E treatment and the control treatment is -89.9. A two-sample *t*-test of this difference is

```
## 
##  Two Sample t-test
## 
## data:  vole[treatment == "vitamin_E", lifespan] and vole[treatment == "control", lifespan]
## t = -1.6275, df = 75, p-value = 0.1078
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -199.97362   20.14029
## sample estimates:
```

```
## mean of x mean of y
##   413.4762  503.3929
```

The *p*-value comes from comparing the observed *t* to a null *t* distribution and "counting" the values that are bigger than the observed *t*. These are counted in both tails, because *p* is the probability of a *t* more extreme than the observed value, and *t* can be more extreme in the negative direction and in the positive direction. We can simulate this with a finite, instead of infinite, null distribution using the t-distribution instead of the distribution of mean differences, as above. I show the script, but don't just cut and paste the code. Spend time thinking about what the each line does. Explore it by copying parts and pasting into console.

```r
set.seed(1)
n_iter <- 10^4 # number of iterations
mu <- mean(vole[treatment=='control', lifespan])
sigma <- sd(vole[treatment=='control', lifespan])
n <- nrow((vole[treatment=='control',]))
sample1 <- matrix(rnorm(n*n_iter, mean=mu, sd=sigma), nrow=n) # 100,000 samples
sample2 <- matrix(rnorm(n*n_iter, mean=mu, sd=sigma), nrow=n) # 100,000 samples

#way no. 1 - compute the t-tests manually
mean_diffs <- apply(sample2, 2, mean) - apply(sample1, 2, mean) # what is the apply function returning?
se_mean_diffs <- sqrt(apply(sample2, 2, sd)^2/n + apply(sample1, 2, sd)^2/n)
t_dis <- mean_diffs/se_mean_diffs

#way no.2 - compute the t-tests using the base R function t.test
# rbind stacks sample1 under sample2. I used this order to be consistent with
# t.test function
t_dis2 <- apply(rbind(sample2, sample1), 2, function(x)
  {t.test(x[1:n], x[(n+1):(2*n)], var.equal=TRUE)$statistic})
# confirm that t_dis = t_dis2

# plot the null distribution of t-values
qplot(t_dis)
```

```r
# what is the p-value?
count <- sum(abs(t_dis) >= abs(ttest.E$statistic)) + 1
p_vitamin_E <- count/(n_iter + 1)
```

Hey that looks pretty good! A *p* value can be computed by counting the number of simulated *t*-values, *including the observed value*, that are more extreme (in either the positive or negative direction) than the observed *t*. Including the observed *t*, there are 1015 values that are more extreme than that observed. An approximate measure of *p* is this count divided by 100,001 (why is 1 added to the denominator?), which is 0.1014899. This simulation-based *p*-value is very (very!) close to that computed from the observed *t*-test.

6.2.3   P-values from the perspective of permutation

A very intuitive way to think about *p*-values is with a permutation test. Consider two of the treatment levels in the vole data, say vitamin E and the vitamin C (I'm bored with the control!). Think about the structure of the dataset: there are two columns, "Treatment", which contains the assigned treatment, and "Lifespan". The values in the Treatment column were randomly assigned prior to the start of the experiment. If there is an effect of treatment on lifespan, then assginment matters – the values in the lifespan column for the vitamin E rows will be more or less, on average, than the values in the lifepan column for the vitamin C rows. Or, the lifespan values are what they are *because* of the values in the treatment column.

Now let's leave the values in the treatment column be, and just randomly re-arrange or permute the lifespan values. What is the new expected diference in lifespan between the two treatments? Zero, of course! That
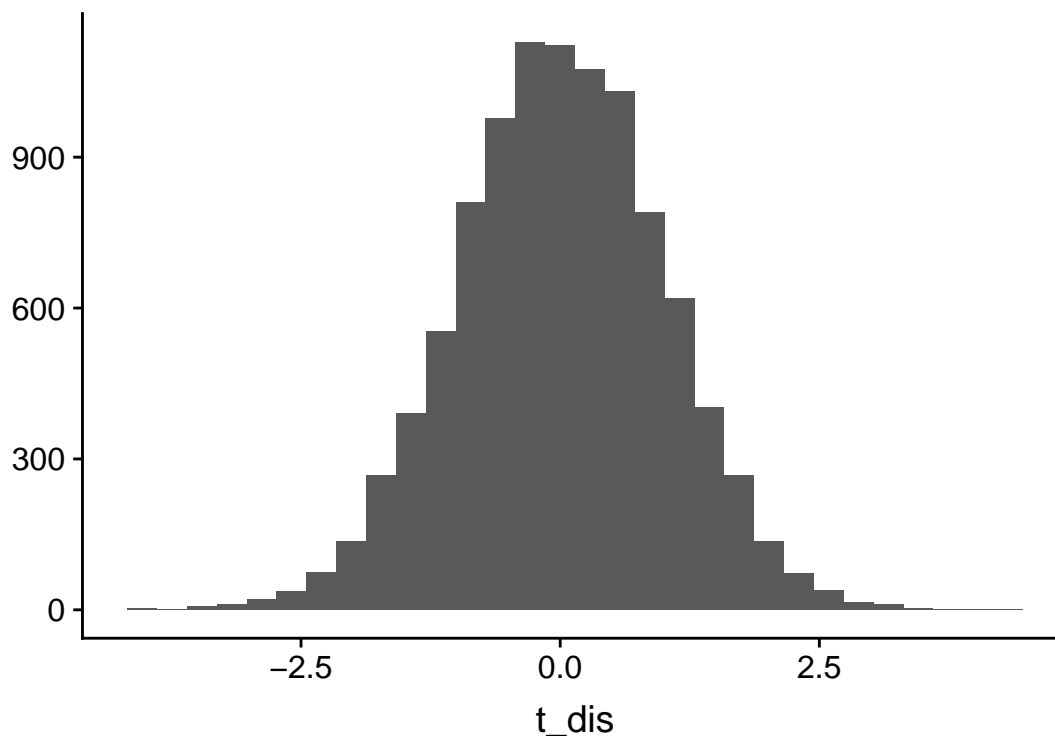
Figure 6.2: Null distribution of t-values. The simulation generated 10,000 t-tests with a true null.

is, because the lifespans were randomly re-arranged, they cannot be caused by treatment assignment!

A permutation is a random re-arrangement of values in a column. Consider the many thousands of permutations of the values in the lifespan column. A difference in means can be computed from each of these permuations and a distribution of differences can be generated. Is the observed difference extreme relative to the other values in this distribution? This is a permutation test – it compares an observed statistic to a distributin of the statistic computed over many thousands of permutations.

6.3 Statistical modeling instead of hypothesis testing

This chapter is an introduction to a *p*-value by way of *t*-tests. I advocate that you analyze *t*-test like questions using statistical modeling instead of null hypothesis significance testing. The reason is that we learn much more from an estimate of the effect and a CI than from a *t* and *p*-value. But, it is also good to know that a *t* test is a special case of a linear model, and you can get that *t* and *p* using a statistical modeling approach should your boss want them (and you cannot convince them otherwise). Let's explore this.

1. Using the emmeans package, compute the effects (differences in means) of vitamin E and vitamin C on lifespan, relative to the control, with their 95% CI and the *t* and *p* values for the cold-reared vole data.
2. Compute a separate *t*-test of vitamin-E vs. control and vitamin C vs. control.

Are the *t* and *p* values the same? No! The reason is that the statistical model had three groups and the SE of the difference was computed from the sample standard deviation of all three groups. Each t-test computes the SE of the difference from only the two groups being compared. In general, the SE computed from all three groups is better because it uses more information. This is one reason to prefer the linear model instead of the separate t-tests.

3. To convince yourself that a *t*-test is a special case as of a linear model, compute the effects of the

vitamin E treatment (relative to control) **but exclude the vitamin C data from the model fit**. Now compare the t and p values with the t-test. These should be the same.

4. Now use the default t.test function by deleting "var.equal=TRUE" from the function. Are t and p still equal to those from the statistical model? No! the reason is because the default t.test function uses a modification of the t-test called "Welsch's t-test". This test allows for heterogenity of variances. Several sources argue that one should always uses Welsch's test since it simplifies to the classical t-test when the sample variances are equal. This is true, but only relevant if you're into t-tests. And, we can model heterogenous variances using a statistical model. We'll do this in a later chapter.

5. Use the function `pairwise.t.test` to compute all pairwise t.tests among the three treatment levels. Is the p-value for the vitamin_E - control contrast the same as that if using t.test (with var.equal=TRUE) or the statistical model with vitamin_C data excluded? No! The reason is that pairwise.t.test adjusts the p-values for multiple testing as a default.

Pro tip: Before you use a new R function like t.test or pairwise.t.test, it is really advisable to read the help page and look at the defaults for the parameters! Researchers publish errors because they failed to look closely at what the R function was doing and they think the function is doing something else. Ooops!

6.4   frequentist probability and the interpretation of p-values

6.4.1   Background

There are at least three different meanings of **probability**.

1. **subjective probability** is the probability that an individual assigns to an event based on prior knowledge and the kinds of information considered reliable evidence. For example, if I asked a sample of students, what is the probability that a 30c homeopathic medicine could clear a *Streptococcus* infection from your respiratory system, their answers would differ because of variation in their knowledge of basic science, including chemistry and physics, their knowledge of what homeopathic medicines are, and how they weight different kinds of evidence.

2. **classical probability** is simply one divided by the number of possible unique events. For example, with a six-sided die, there are six possible unique events. The probability of rolling a 2 is $\frac{1}{6}$ and the probability of rolling an odd number is $\frac{1}{2}$.

3. **frequentist probability** is based on the concept of *long run frequency*. If I roll a die 10 times, the frequency of rolling a 2 will be approximately $\frac{1}{6}$. If I roll the die 100 times, the frequency of rolling a two will be closer, but to $\frac{1}{6}$. If I roll the die 1000 times, the frequency of rolling the die will be even closer to $\frac{1}{6}$. So the frequentist definition is the expected frequency given an infinite number of rolls. For events with continous outcomes, a frequentist probability is the long run frquency of *observing the outcome or one more extreme*.

6.4.2   This book covers frequentist approaches to statistical modeling and when a probability arises, such as the *p*-value of a test statistic, this will be a frequentist probability.

When we do a t-test, we get a p-value. There are several ways to think about this probability. The most compact way is $P(data|null)$, which is literally read as the probability of the data given the null (or "conditional" on the null), but is really short for *the probability of the data, or something more extreme than the data, given that the null hypothesis is true.* The "probability of the data" is kinda vague. More specifically, we mean the probability of some statistic about the data such as the difference in means between group A and group B or the t-value associated with this difference. So, a bit more formally, the probability returned in a t-test is $\text{prob}(t \geq t_{obs}|H_0)$. This is the long run frequency of observing a t-value as big or bigger than

the observed t-value (the one you actually got with your data) if the null is true. Let's parse this into "long run frequency of observing a t-value as big or bigger than the observed t-value" and "null is true".

A thought experiment: You open a google sheet and insert 12 standard, normal random deviates (so the true mean is zero and the true variance is one) in Column A, rows 1-12. You arbitrarily assign the first six values (rows 1-6) to treatment A and the second six values (rows 7-12) to treatment B. You use the space immediately below these data to compute the mean of treatment A, the mean of treatment B, the difference in means (A - B), and a t-value. Unfortunately, google sheets doesn't have a t-value function so you'd have to compute this yourself. Or not, since this is a thought experiment. Now "fill right" or copy and paste these functions into 999 new columns. You now have 1000 t tests. The expected value of the difference in means is zero (why?) but the actual values will form a normal distribution about zero. Most will be close to zero (either in the negative or positive direction) but some will be further from zero. The expected t-value will also be zero (why?) and the distribution of these 1000 t values will look normal but the tails are a little fuller. This row of t values is a null distribution, because in generating the data we used the exact same formula for the values assigned to A and the values assigned to B. Now think of a t-value in your head, say 0.72 (remember that t values will largely range from about -3 to +3 although the theoretical range is $-\infty$ to $+\infty$. What is the probability of observing a t of 0.72 *or bigger* if the null is true? Look at the row of t-values! Count the number of $t \geq 0.72$ and then divide by the total number of t-values in the row (1000) and you have a probability computed as a frequency. But remember the frequentist definition is the long run frequency, or the expected frequency at the limit (when you've generated not 1000 or even 1,000,000 but an infinite number of columns and t-values).

Some asides to the thought experiment: First, why "as big or bigger" and not just the probability of the value itself? The reason is that the probability of finding the exact t is 1/infinity, which doesn't do us much good. So instead we compute the probability of finding t as big, or bigger, than our observed t. Second, the t-test probability described above is a "one-tail probability". Because a difference can be both in the positive direction and the negative direction, we usually want to count all the $t \geq 0.72$ and the $t \leq -0.72$ and then add these two counts to compute the frequency of *as extreme or more extreme* values. This is called a "two-tailed probability" because we find extremes at both tails of the distribution. Third, we don't really count $t \geq 0.72$ but take advantage of the beautiful mathematical properties of the theoretical t distribution, which allows us to compute the frequentist probability (expected long range frequency) given the t-value and the degrees of freedom using the t-distribution.

Now what do I mean with the phrase "null is true"? Most people equate "null is true" with "no difference in means" but the phrase entails much more than this. Effectively, the phrase means that the p-value is based on modeling the real data with a theoretical sample in which all the points were randomly sampled from the same distribution and that the assignment of the individual points to treatment was random. This model means the theoretical sample has three properties: First, random assignment to treatment after sampling from the same distribution means that the expected means are the same, or put differently, the expected difference in means between the assigned groups is zero. Second, random assignment to treatment after sampling from the same distribution *also* means that the expected variances of the two groups are equal. And third, random sampling means that the values of each point are independent – we cannot predict the value of one point knowing information about any other point. **Here is what is super important about this**: if we get a really low p-value, any one of these consequences may be untrue about our data, for example it could be that the true means of the two treatment groups really are different, or it could mean it is the variances that differ between the two groups, or it could mean that the data (or technically, the errors) are not independent of each other. This is why we need certain assumptions to make a p-value meaningful for empirical data. By assuming independent error and homogenous (equal) variances in our two samples, a low p value is evidence of unequal means.

6.4.3 Two interpretations of the p-value

Since we want to be working scientists who want to use p-values as a tool, we need to know how to interpret (or use) the p-value to make reasonable inferences and how to avoid mis-interpreting the p-value and making unreasonable or even incorrect inferences. Ronald Fisher, the inventor of frequentist statistics, developed

an interpretation of the *p*-value that is probably most useful for academic and applied research programs. Neyman and Pearson (Neyman-Pearson) gave the *p*-value a different interpretation, one that is probably most useful for industrial quality control. Today's biology researchers use an interpretation that is an odd hybrid of the two, which often leads to silly inference. Regardless, understanding the distinction between Fisher and Neyman-Pearson will inform how we write up our results in a manuscript. I'll describe these in the context of the two-sample *t*-test.

6.4.3.1   Fisher's interpretation

Fisher was working in the context of an agricultural field station, the goal of which is to discover better agricultural practices. Does this new fertilizer work better than our old fertilizer? This is the context of much of modern biosciences and clinical medicine. Fisher thought of *p* as evidence against the null; the smaller the *p* the better the evidence that the means differ, which, in an experimental context, implies a treatment effect. If an experiment results in a large *p*-value, we can move on and test other fertilizers. If an experiment results in a small *p*-value, we want to pursue this new fertilizer more. Do more experiments! Fisher never thought of a single experiment as definitive. The decision to move on or pursue is only partly informed by the *p*-value and Fisher offered no rule about what *p*-value lies on the threshold of this decision. When pressed, Fisher might say that $p = 0.05$ is a reasonable threshold.

6.4.3.2   Neyman-Pearson interpretation

Neyman-Pearson thought of *p* as the necessary and sufficient information to make a decision between accepting the null (or at least not rejecting the null) or rejecting the null and accepting an alternative hypothesis. This decision balances two sorts of errors: Type I (false positives), which they called α, and Type II (false negatives), which they called β. A false positive means the null was rejected but there really is no effect. A false negative means that the null was not rejected but there actually is an effect. α is set by the experimenter and is the long-term frequency (or "rate") of false positives **when the null is true** that the experimenters are willing to accept. This is easily understood in the context of manufacturing. I've just made a batch of beer that I now need to ship. I sample 10 cans and test the quality against a norm. If $p < \alpha$, we reject the null in favor of the alternative – something may be wrong with the batch, it differs from the norm. We throw the beer away. If $p > \alpha$, we do not reject the null, nor the beer! We ship it.

After setting α, the experimenter designs the experiment to achieve an acceptable rate of β. Since β is the false negative rate then $1 - \beta$ is the rate of not making a false negative error, that is, the rate of rejecting the null when there really is an effect. This is called the **power** of the experiment. An experiment with high power will have a low probability of a Type II error. An experiment with low power will have a high probability of a Type II error. Power is partly determined by sample size, the bigger the sample the smaller the *p*-value, all other things equal (think about why in the context of the formula for the *t*-value). Power is also a function of α. If we set a low α (say, $\alpha = 0.01$), the test is conservative. We are more likely to fail to reject the null even if the null is false. This is the balance. We want to make sure that we test our batch of beer using enough cans to find a bad batch if it exists, but we don't want to test too many cans because this is a waste of money. An experimenter sets α, computes the sample size needed to achieve a certain level of power $(1 - \beta)$, and then does the experiment.

In Fisher's interpretation, there is no α, no β, no alternative hypothesis, and no sharp decision rule. Instead, in Fisher, *p* is a continuous measure of evidence against the null and its value is interpreted subjectively by an informed and knowledgeable expert using additional information to make decisions. Neyman-Pearson rejected Fisher's conception of *p* as evidence against the null and used *p* as a tool to make a decision that maintains long-term type I error rates at alpha given a certain power. In Neyman-Pearson, *p* is compared to a threshold, α and this alone makes the decision. In Neyman-Pearson, *p* is not treated as continuous information. $p = 0.00000001$ is no more evidence to use to reject the null than $p = 0.049$.

6.4.4 NHST

Modern researchers interpret p using a combination of Fisher and Neyman-Pearson concepts in what has become known as Null Hypothesis Significance Testing (NHST). Similar to Neyman-Pearson, a p-value is compared to α but similar to Fisher, many researchers, and many textbooks and statistics software (including base R) trichtomize a statistically significant p into "significance levels" (three asterisks for $p < 0.001$, two asterisks for $0.001 < p < 0.01$, and one asterisk for $0.01 < p < 0.05$) but many researchers also casuallly partition non-significant p values into "marginally signifiant" (or similar) and "not significant".

6.4.5 Some major misconceptions of the p-value

Setting the type I error rate α to 0.05 is so pervasive that I'm going to simply use "0.05" instead of "alpha" in discussing misconceptions.

6.4.5.1 Misconception: p is the probability that the null is true *and* $1 - p$ is probability that the alternative is true

Many researchers believe that if $p > 0.05$ then "there is no effect." A frequentist hypothesis test cannot show that an effect doesn't exist, only that the null has a low probablity of producing a test statistic as extreme or more extreme than the observed effect.

Many researchers believe that if $p < 0.05$ then "there is an effect." Again, a frequentist hypothesis test cannot show that an effect exists, only that the null has a low probablity of producing a test statistic as extreme or more extreme than the observed effect.

1. The statement "There is no effect of predators on feeding behavior" is not a valid conclusion of a frequentist hypothesis test.
2. The statement "We found no effect of predators on feeding behavior" is misleading because a frequentist hypothesis test can neither find an effect nor find no effect.

The two errors above are gross misconceptions that are pervasive in the biology literature. A more subtle issue is the belief that a low p-value shows that the researcher's explanatory hypothesis is correct. For example, researchers believe the result "the prey fish fed 14.2 (95% CI: 9.2, 19.2) minutes shorter in the presence of the predator fish" confirms their hypothesis that prey modulate feeding duration as a function of their ability to assess the risk of predation. Some alternative explanations:

1. The predator fish also competes with the prey fish for the prey fish's food and with less food the prey fish spends less time feeding because it gives up when food density drops below some amount.
2. The predator fish is introduced to the prey tank by hand and odorant molecules from the researcher's hands are detected by the prey and the prey reduces feeding duration because of these odorants.

Importantly, no single experiment confirms an explanatory hypothesis. Instead, alternative explanations require multiple experiments with different controls to "rigrously probe" the preferred hypothesis.

6.4.5.2 Misconception: a p-value is repeatable

Many researchers believe that a p-value is a precise measure – that if the experiment were replicated, a similar p would result. This belief requires at least two misconceptions. First, if the null were true, then *any* p-value is equally likely. $p = 0.00137$ is just as likely as $p = 0.492$. In other words, if the null were true, the p-value is not replicable at all! Second, the p value is highly dependent on the sample, and can be highly variable among replications, but there is no true p-value, so there can be no estimate or standard error. Let's explore these.

What is the distribution of p-values under the null? I often ask students, "if the null were true, what is the most likely p-value?" or "if null were true, what kind of p-values would we expect, that is what is

the expected distribution". A common answer is $p = 0.5$ is the most likely value and something like a normal curve, except the tails abruptly stop at 0 and 1, is the expected distribution.
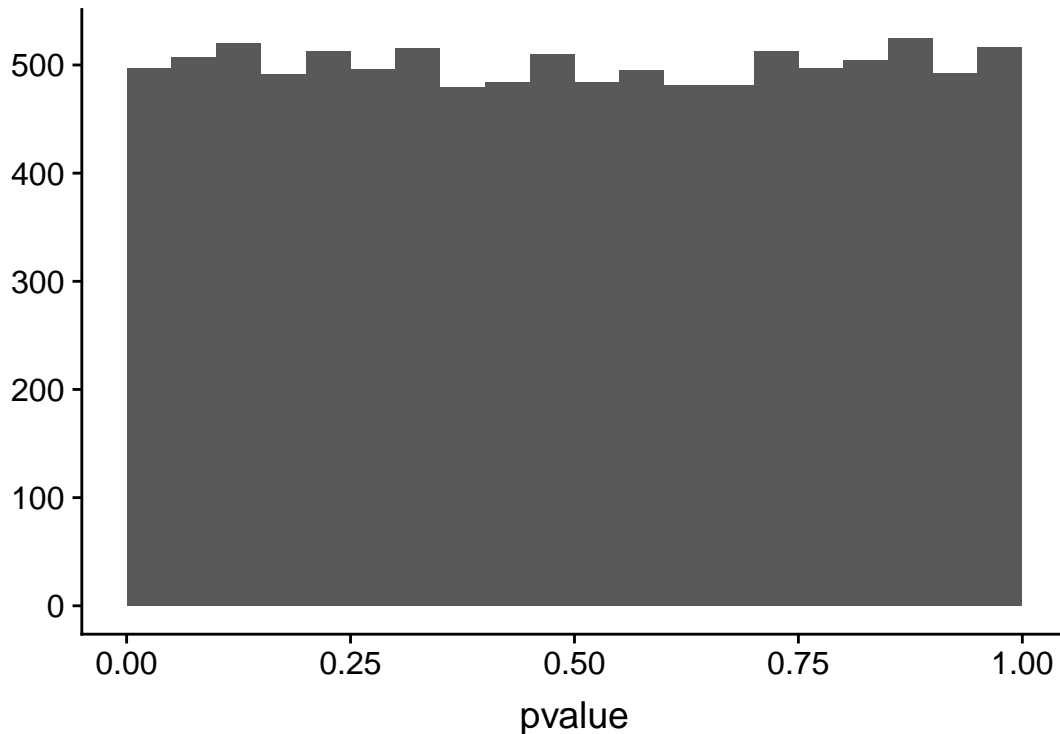
The incredible inconsistency of the p-value

How replicable is the conclusion of an experiment if the p-value for a t-test is 0.03? If our conclusion is based on $p < 0.05$, then the conclusion is not very replicable. The simulation below shows the results of 15 replicates of an experiment with true power of 40%. There are five "significant" results (one less than expected) but several replicates have very high p-values.

6.4.5.3   Misconception: 0.05 is the lifetime rate of false discoveries

An important and widespread misconception is that if a researcher consistently uses $\alpha = 0.05$, then the frequency of incorrectly concluding an effect exists, or "discovering" an effect, over the lifetime of the researcher, will be 5%. This is incorrect. α is the rate of false positive if the null hypothesis is true, so our lifetime "false discovery" rate could only be 5% if everything we ever tested has no true effect! More generally, the **false discovery rate** is the frequency of false positives divided by the frequency of positives (the sum of false and true positives). This differs from the the Type I error rate, which is the frequency of false positives divided by the frequency of tests *in which the null is true.*

Imagine we test

1. 1000 null hypotheses over a lifetime
2. 60% are true nulls, this means there are 600 true nulls and 400 true effects
3. alpha is 5%. This means we expect to find $p \leq 0.05$ 30 times (0.05×600) when the null is true
4. power is 25%. This means we expect to find $p \leq 0.05$ 100 times (0.25×400) when the null is false
5. We have made $30 + 100 = 130$ "discoveries" (all experiments with $p \leq 0.05$), but
6. 30 of the 130, or 23%, are "false discoveries". This is the false discovery rate.

Think about this. If the null is never true, you cannot have a false discovery–every $p \leq 0.05$ is a true discovery (the false discovery rate is 0%). And if the null is always true, every $p < 0.05$ is a false discovery (the false discovery rate is 100%).
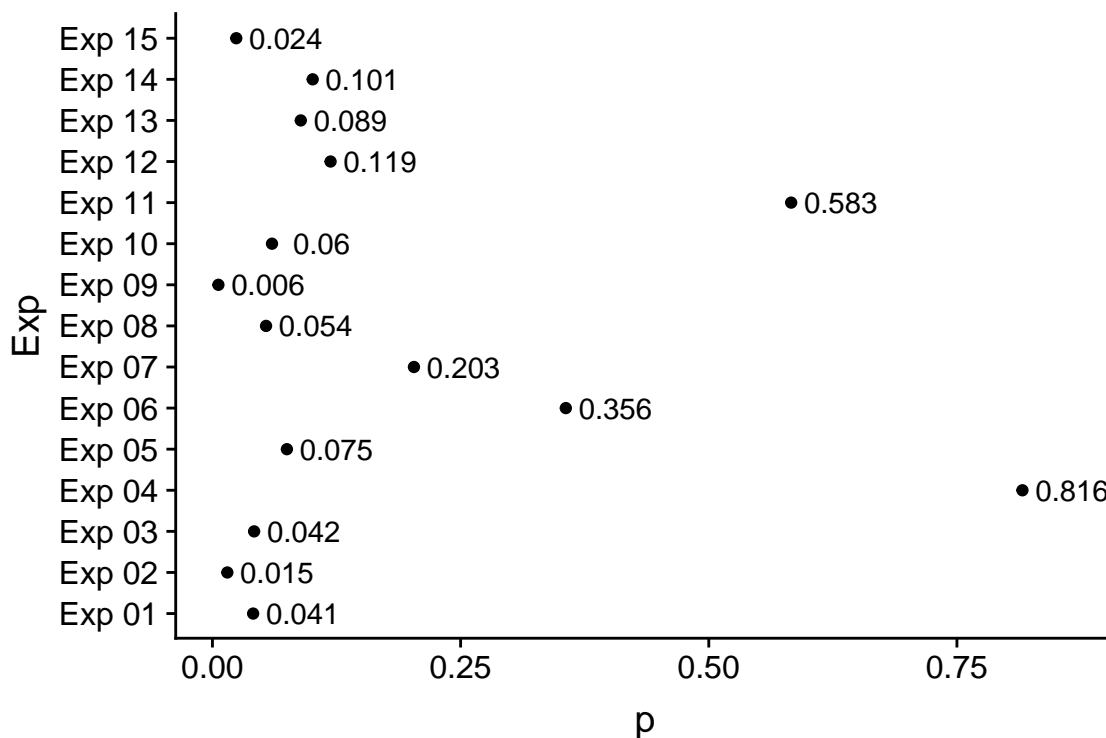
Figure 6.3: Variability of *p*-values when the power is 0.4

6.4.5.4   Misconception: a low *p*-value indicates an important effect

Many researchers write results as if they believe that a small *p*-value means the effect is big or important. This may misconception may arise because of the ubiquitous use of "significant" to indicate a small p-value and "very" or "extremely" or "wicked" significant to indicate a really small p-value. Regardless, this is a misconception. A small p-value will usually result when there is high power (but can occur even if power is low) and power is a function of effect size, variability (the standard deviation), and sample size. A small *p* could result from a large effect size but can also result with a small effect size if the sample size is big enough.

This is easy to simulate (see script below). Let's model the effect of the genotype of a gene on height

```
set.seed(1)
rho <- 0.5
n <- 10^4
genotype <- c("+/+", "+/-", "-/-")
Sigma <- diag(2)
Sigma[1,2] <- Sigma[2,1] <- rho
X <- rmvnorm(n, mean=c(0,0), sigma=Sigma)
colnames(X) <- c("X1", "X2")
beta <- c(0.05, 0.05)
y <- X%*%beta + rnorm(n)
fit <- lm(y ~ X)
coefficients(summary(fit))
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.007472959 0.01007946 0.7414046 4.584656e-01
## XX1         0.044304824 0.01154709 3.8368830 1.253725e-04
```

```
## XX2          0.048228101 0.01170855 4.1190490 3.835033e-05
```

6.4.5.5 Misconception: a low p-value indicates high model fit or high predictive capacity

On page 606, of Lock et al "Statistics: Unlocking the Power of Data", the authors state in item D "The p-value from the ANOVA table is 0.000 so the model as a whole is effective at predicting grade point averages." This is incorrect. A p-value is not a measure of the predictive capacity of a model because the p-value is a function of the signal, noise (unmodeled error), and *sample size* while predictive capacity is a function of just the signal:noise ratio. If the signal:noise ratio is tiny, the predictive capacity is small but the p-value can be tiny if the sample size is large. This is easy to simulate (see script below). The whole-model p-value is exceptionally small (0.00001002) but the relative predictive ability, measured by the R^2, is near zero (0.002).

```r
set.seed(1)
rho <- 0.5
n <- 10^4
Sigma <- diag(2)
Sigma[1,2] <- Sigma[2,1] <- rho
X <- rmvnorm(n, mean=c(0,0), sigma=Sigma)
colnames(X) <- c("X1", "X2")
beta <- c(0.05, -0.05)
y <- X%*%beta + rnorm(n)
fit <- lm(y ~ X)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6449 -0.6857  0.0148  0.6756  3.6510
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.007473   0.010079   0.741 0.458466
## XX1          0.044305   0.011547   3.837 0.000125 ***
## XX2         -0.051772   0.011709  -4.422  9.9e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 9997 degrees of freedom
## Multiple R-squared:  0.0023, Adjusted R-squared:  0.002101
## F-statistic: 11.52 on 2 and 9997 DF,  p-value: 1.002e-05
```

6.4.5.5.1 What the p-value does not mean

1. p is not the probability of the null being true. More formally, this probability is $P(null|data)$ but our p-value is $P(data|null)$. These are not the same. $P(null|data)$ is the probability of the null being true given the data. $P(data|null)$ is the probability of our data, or something more extreme than our data, conditional on a true null.
2. $1 - p$ is not the probability of the alternative
3. p is not a measure of effect size.
4. p in one experiment is not the same level of evidence against the null as in another experiment
5. p is not a great indicator of which is more likely, H0 or H1.

6. If one treatment level has $p < 0.05$ and another treatment level has $p > 0.05$, this is not evidence that the treatment levels have different effects on the outcome.

6.4.6 Recommendations

If you are working on basic science research simply report the exact p-value, along with a CI. If $p < 0.05$ (or some other α) do not report this as "significant" – in fact, avoid the word "significant". In the english language, "significant" implies big or important. Small p-values can result even with trivially small effects if n is big or sample variation is small. If p is smaller than say 0.001, then this is pretty good evidence that the data is not a fluke of sampling. But if p is closer to 0.01 or 0.05, this is only weak evidence of a fluke because of the sampling variability of p.

If you are working on quality control then a p value is a useful tool, but is only relevant compared to a decision rule with well-reasoned values of α and β – exact values of p are not very meaningful.

6.5 Problems

Problem 1 – simulate the distribution of p under the null. There are many ways to do this but a straightforard approach is to

1. Create a $2n \times m$ matrix of random normal deviates with mean 0 and sd 1
2. Do a t-test on each column, with the first n values assigned to one group and the remaining n values assigned to the second group. Save the p-value from each.
3. Plot a histogram of the p-values.
4. What is the distribution? What is the most likely value of p?

Problem 2 – simulate power. Again, many ways to do this but following up on Problem 1. 1. Create a $2n \times m$ matrix of random normal deviates with mean 0 and sd 1 2. Add an effect to the first n values of each column. Things to think about a. what is a good effect size to add? The effect/sd ratio, known as Cohen's d, is a relative (or standardized) measure of effect size. Cohen suggest 0.2, 0.5, and 0.8 as small, medium, and large standardized effects. b. should the same effect be added to each individual? Yes! It is the random component that captures the individual variation in the response. 3. Do a t-test on each column of the matrix, using the first n values in group 1 and the remaining n values in group 2. Save the p-values for each. 4. Compute the power, the relative frequency $p \leq 0.05$. 5. Repeat with different values of n, effect size, and sd, but only vary one at a time. How does power vary with these three parameters?

Problem 3 – write a script for a permutation test of the vitamin E and vitamin C levels of the vole data. Compare this to the t-test.

Problem 4 – grad students only. Simulate the false discovery rate. Explore the parameters: the frequency of true nulls and the power.

Chapter 7

Creating Fake Data

Fake data are generated by sampling from one of R's random sampling functions. These functions sample from different distributions including

1. uniform – function `runif(n, min=0, max=1)`, which samples n continous values between `min` and `max`.
2. normal (Gaussian) – function `rnorm(n, mean=0, sd=1)`, which samples n continous values from a distribution with the specified mean and standard deviation. The default is the "standard" normal distribution.
3. poisson – function `rpois(n, lambda)`, which samples `n` counts from a distribution with mean and variance equal to `lambda`.
4. negative binomial – `rnegbin(n, mu=n, theta)`, which samples `n` counts with mean `mu` and variance `mu + mu^2/theta`.

7.0.1 Continuous X (fake observational data)

A very simple simulation of observational design (the X are not at "controlled" levels)

```
n <- 25
# the paramters
beta_0 <- 25 # the true intercept
beta_1 <- 3.4 # the true slope
sigma <- 2 # the true standard deviation

x <- rnorm(n)
y <- beta_0 + beta_1*x + rnorm(n, sd=sigma)
qplot(x, y)
```
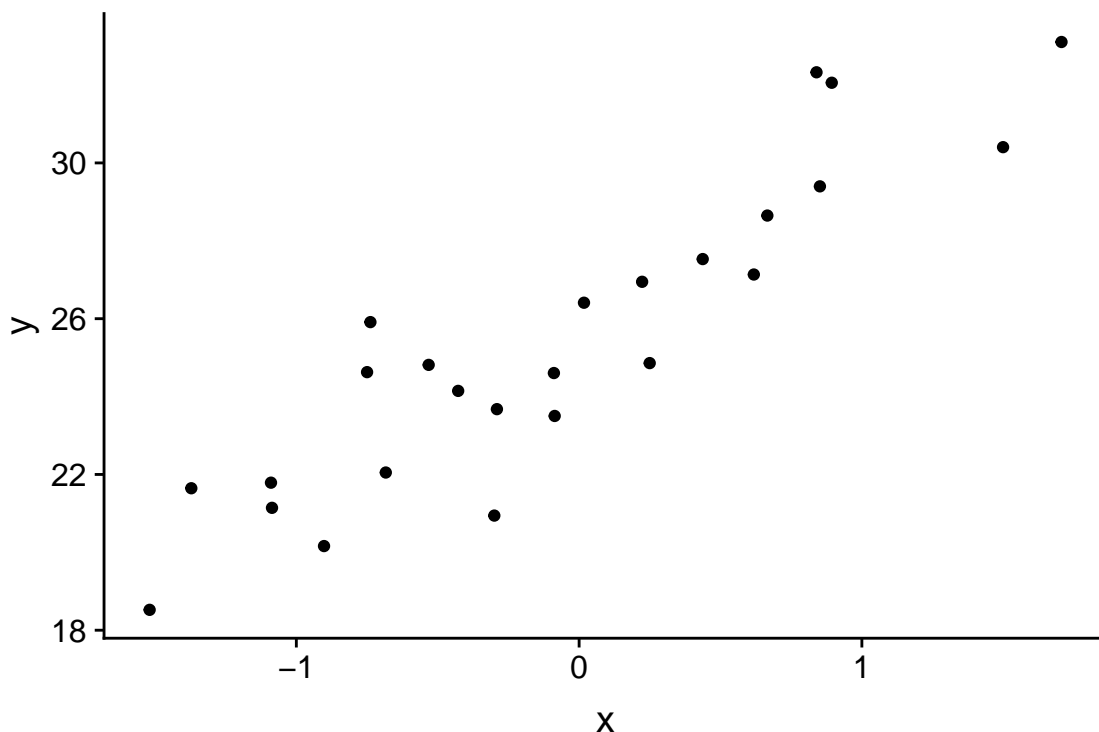
How well does a model fit to the data recover the true parameters?

```r
fit <- lm(y ~ x)
knitr::kable(coefficients(summary(fit)), digits=c(1, 2, 1, 4))
```

	Estimate	Std. Error	t value	Pr($>$|t|)
(Intercept)	25.8	0.34	75.8	0
x	4.2	0.40	10.4	0

The coefficient of x is the "Estimate". How close is the estimate? Run the simulation several times to look at the variation in the estimate – this will give you a sense of the uncertainty. Increase n and explore this uncertainty. Increase all the way up to $n = 10^5$. Commenting out the qplot line will make this exploration easier.

7.0.2 Categorical X (fake experimental data)

Similar to above but the X are at controlled levels and so this simulates an experimental design

```r
n <- 5 # the sample size per treatment level

fake_data <- data.table(Treatment=rep(c("control", "treated"), each=n))
beta_0 <- 10.5 # mean of untreated
beta_1 <- 2.1 # difference in means (treated – untreated)
sigma <- 3 # the error standard deviation
# the Y variable ("Response") is a function of treatment. We use some matrix
# algebra to get this done.
# Turn the Treatment assignment into a model matrix. Take a peak at X!
X <- model.matrix(~ Treatment, fake_data)
# to make the math easier the coefficients are collected into a vector
beta <- c(beta_0, beta_1)
# you will see the formula Y=Xb many times. Here it is coded in R
fake_data[, Response:=X%*%beta + rnorm(n, sd=sigma)]
```

```
# plot it with a strip chart (often called a "dot plot")
ggstripchart(data=fake_data, x="Treatment", y="Response", add = c("mean_se"))
```
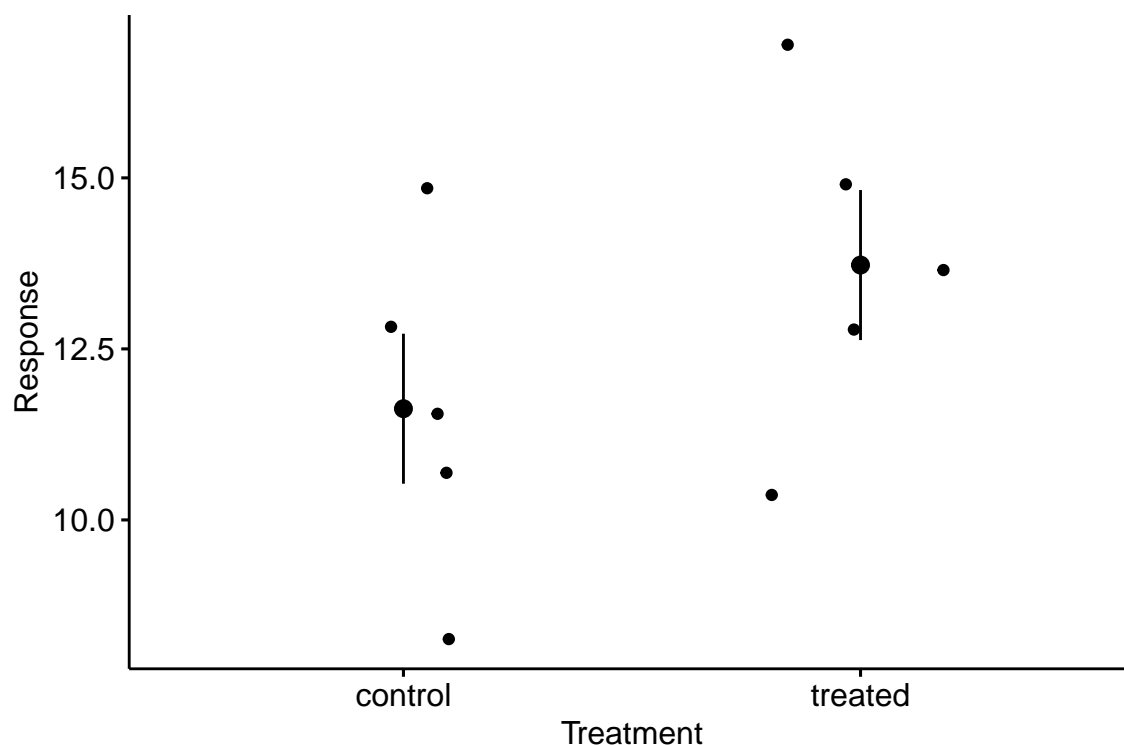


```
# fit using base R linear model function
fit <- lm(Response ~ Treatment, data=fake_data)
# display a pretty table of the coefficients
knitr::kable(coefficients(summary(fit)), digits=3)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.626	1.097	10.601	0.000
Treatmenttreated	2.100	1.551	1.354	0.213

Check that the intercept is close to beta_0 and the coefficient for Treatment is close to beta_1. This coefficient is the difference in means between the treatment levels. It is the simulated effect. Again, change n. Good values are $n = 20$ and $n = 100$. Again, comment out the plot line to make exploration more efficient.

7.0.3 Correlated X (fake observational data)

7.0.3.1 Generating correlated X variables

It's useful to think about how correlated data are generated because often we want to generate fake_data with an expected correlation. Let's say we want to generate two X variables that have an expected correlation of 0.6. To generate this, we take advantage of the fact that two variables, X1 and X2, are correlated if they share a "common cause" – a variable Z that effects (or "causes") both X1 and X2. If the expected variances of X1, X2, and Z are all 1, then the expected correlation between X1 and X2 is the product of the causal effect from Z to each X. The easiest way to implement this is to simply make the effect from Z to both X equal to $\sqrt{(0.6)}$.

```
n <- 10^3
z <- rnorm(n) # the common cause, with sigma = 1
rho <- 0.6 # the true correlation between X1 and X2
beta_z <- sqrt(rho) # the easiest way to get effects of z on X1 and X2 that generates rho
sigma_x <- sqrt(1 - rho) # we will make the variance of X1 and X2 = 1, so the "explained" variance in X
x1 <- beta_z*z + rnorm(n, sd=sigma_x)
x2 <- beta_z*z + rnorm(n, sd=sigma_x)
# check
cov(data.frame(X1=x1, X2=x2)) # is the diagonal close to 1, 1?
```

```
##             X1        X2
## X1 1.0523598 0.5978408
## X2 0.5978408 0.9713292
```

```
cor(x1, x2) # is the value close to rho?
```

```
## [1] 0.5913168
```

```
qplot(x1, x2)
```
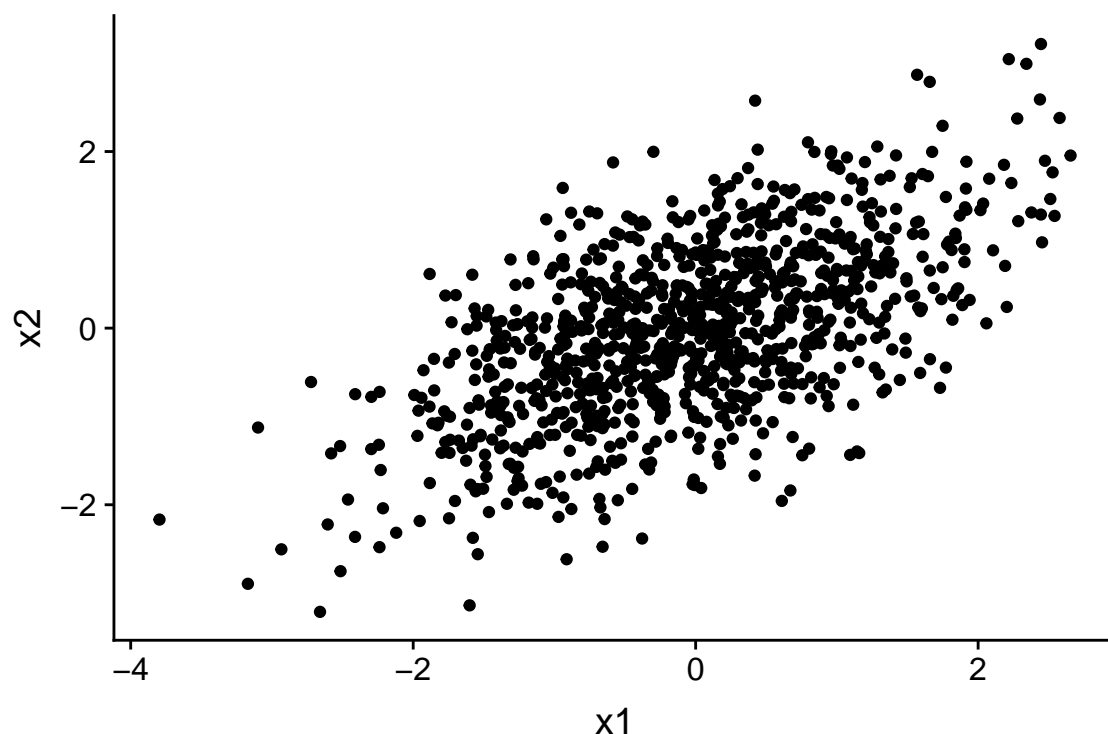

Now create fake Y that is a function of both X_1 and X_2. Create "standardized" fake data, where $\sigma_Y = 1$.

```
beta_0 <- 3.2
beta_1 <- 0.7
beta_2 <- -0.3
explained_sigma <- beta_1^2 + beta_2^2 + 2*beta_1*beta_2*rho # Wright's rules! Compare to Trig!
sigma_Y.X <- sqrt(1 - explained_sigma) # sqrt unexplained variance
y <- beta_0 + beta_1*x1 + beta_2*x2 + rnorm(n, sd=sigma_Y.X)

# check
var(y) # should be close to 1 as n gets bigger
```

```
## [1] 0.9791502
```

```r
# check parameters
coef(summary(lm(y ~ x1 + x2))) # should be near parameters as n gets bigger
```

```
##               Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)  3.1887712 0.02536483 125.716255 0.000000e+00
## x1           0.6885565 0.03064195  22.471043 8.422461e-91
## x2          -0.3037457 0.03189446  -9.523465 1.223898e-20
```

Note that the variance of Y is the variance of the explained part due to X1 and X2 and the unexplained part and if the expected variance of $Y = 1$ then this sets an upper limit for the explained part. This means that

$$\beta_1^2 + \beta_2^2 + 2\beta_1\beta_2\rho < 1 \tag{7.1}$$

which means the magnitude of β_1 and β_2 should generally be less than 1.

7.0.3.2 Creating mulitple X variables using the package mvtnorm

The package mvtnorm provides a function to generate multivariate data (multiple columns) with a specified vector of means (the means of each column) and covariance matrix among the means.

```r
rcov1 <- function(p){
  # p is the number of columns or number of variables
  pp <- p*(p-1)/2 # number of elements in lower tri
  max_r <- 0.7
  r <- rexp(pp)
  r <- r*max_r/max(r)

  # create correlation matrix
  R <- matrix(1, nrow=p, ncol=p)
  R[lower.tri(R, diag=FALSE)] <- r
  R <- t(R)
  R[lower.tri(R, diag=FALSE)] <- r

  # convert to covariance matrix
  L <- diag(sqrt(rexp(p))) # standard deviations
  S <- L%*%R%*%L

  # check -- these should be the same
  # R
  # cov2cor(S)
  return(S)
}
```

Now let's use mvtnorm to generate fake correlated X

```r
p <- 5 # number of X variables
S <- rcov1(p)

# make the fake X
n <- 10^5
mu <- runif(p, min=10, max=100) # vecctor of p means
X <- rmvnorm(n, mean=mu, sigma=S)

# how close? (check the cor as this is easier to scan)
cov2cor(S)
```

```
##           [,1]       [,2]       [,3]       [,4]       [,5]
## [1,] 1.0000000 0.15511020 0.11863165 0.48788949 0.14545246
## [2,] 0.1551102 1.00000000 0.70000000 0.03514487 0.04345391
## [3,] 0.1186317 0.70000000 1.00000000 0.05346376 0.54083807
## [4,] 0.4878895 0.03514487 0.05346376 1.00000000 0.44504979
## [5,] 0.1454525 0.04345391 0.54083807 0.44504979 1.00000000
```

```
cor(X)
```

```
##           [,1]       [,2]       [,3]       [,4]       [,5]
## [1,] 1.0000000 0.15618693 0.12146387 0.48987841 0.14569745
## [2,] 0.1561869 1.00000000 0.69709529 0.03879653 0.03961509
## [3,] 0.1214639 0.69709529 1.00000000 0.05598629 0.54138395
## [4,] 0.4898784 0.03879653 0.05598629 1.00000000 0.44214983
## [5,] 0.1456974 0.03961509 0.54138395 0.44214983 1.00000000
```

7.0.3.3   The rcov1 algorithm is naive

A problem with generating a fake covariance matrix as above is that it is likely to be **singular** as p get's bigger. A singular covariance matrix is one where there are fewer orthogonal axes of variation then there are variables. Imagine a multidimensional scatterplot of a data set with the fake covariance matrix. If we zoom around this multidimensional space, we will come across a "view" in which all the points are compressed along a single line – that is there is no variation on the axis orthogonal to this line of points. This is bad, as it means we cannot fit a linear model using least squares (because the inverse of the covariance matrix doesn't exist).

Let's explore this. If the a covariance matrix is singular, then at least one eigenvalue of the matrix is negative (eigenvalues is a multivariate term beyond the scope of this text but, effectively, these are the variances of the orthogonal axes referred to above). Here I compute the frequency of covariance matrices with at least one negative eigenvalue as p increases

```
niter <- 1000
p_vec <- 3:10
counts <- numeric(length(p_vec))
i <- 0
for(p in 3:10){
  i <- i+1
  for(iter in 1:niter){
    counts[i] <- ifelse(eigen(rcov1(p))$values[p] < 0, counts[i]+1, counts[i])
  }
}
data.table(p=p_vec, freq=counts/niter)
```

```
##     p  freq
## 1:  3 0.000
## 2:  4 0.025
## 3:  5 0.066
## 4:  6 0.135
## 5:  7 0.190
## 6:  8 0.225
## 7:  9 0.275
## 8: 10 0.368
```

7.0.3.4 Generating multiple columns of X variables with a non-singular covariance matrix

This section uses some ideas from matrix algebra. The goal is to create a $n \times p$ matrix of X variables that have some random covariance structure that is full-rank (not singular, or no negative eigenvalues). The algorithm starts with a $p \times p$ random eigenvector matrix $\mathbf{E}$ and a $p \times p$ random eigenvalue matrix $\mathbf{L}$ and then computes the random covariance matrix using $\mathbf{ELE}^\top$

1. Generate a random $p \times p$ random eigenvector matrix from a covariance matrix of $p \times p$ matrix of random normal variables.

```r
fake.eigenvectors <- function(p){
    a <- matrix(rnorm(p*p), p, p) # only orthogonal if p is infinity so need to orthogonalize it
    a <- t(a)%*%a # this is the sum-of-squares-and-cross-product-matrix
    E <- eigen(a)$vectors # decompose to truly orthogonal columns
    return(E)
}
```

2. Generate p random eigenvalues in descending order and that sum to 1. There are several ways to create this sequence. Here are two:

```r
# force the eigenvalues to descend at a constant rate
fake.eigenvalues <- function(p, m=p, start=1, rate=1){
  # m is the number of positive eigenvalues
  # start and rate control the decline in the eigenvalue
  s <- start/seq(1:m)^rate
  s <- c(s, rep(0, p-m)) # add zero eigenvalues
  L <- diag(s/sum(s)*m) # rescale so that sum(s)=m and put into matrix,
  # which would occur if all the traits are variance standardized
  return(L)
}
```

```r
# random descent
fake.eigenvalues.exp <- function(p, m=p, rate=1){
  # exponential distribution of eigenvalues
  # m is the number of positive eigenvalues
  # start and rate control the decline in the eigenvalue
  s <- rexp(m, rate)
  s <- s[order(s, decreasing=TRUE)] # re-order into descending order
  s <- c(s, rep(0, p-m)) # add zero eigenvalues
  L <- diag(s/sum(s)*m) # rescale so that sum(s)=m and put into matrix,
  # which would occur if all the traits are variance standardized
  return(L)
}
```

3. Generate the random covariance matrix

```r
fake.cov.matrix <- function(p){
    # p is the size of the matrix (number of cols and rows)
    E <- fake.eigenvectors(p)
    L <- diag(fake.eigenvalues(p))
    S <- E%*%L%*%t(E)
    return(S)
}
```

4. Generate the random X variables using $\mathbf{X} = \mathbf{X}'(\mathbf{E}\sqrt{\mathbf{L}})^\top$

```r
# two functions to compute the random data
fake.X <- function(n,p,E,L){
  # n is number of observations
  # p is number of variables
  X <- matrix(rnorm(n*p),nrow=n,ncol=p) %*% t(E%*%sqrt(L))
    return(X)
}
```

An example

```r
n <- 10^5
p <- 5
E <- fake.eigenvectors(p)
L <- fake.eigenvalues(p, start=1, rate=1)
X <- fake.X(n, p, E, L)
colnames(X) <- paste0("X", 1:p)
E
```

```
##              [,1]         [,2]        [,3]        [,4]        [,5]
## [1,]   0.4762049  0.76679336  0.02189702 -0.37886958 -0.20306443
## [2,] -0.3226816  0.26514622 -0.21147219  0.50680583 -0.72387940
## [3,]   0.6343127 -0.07414265  0.40381446  0.65434666  0.03024185
## [4,]   0.2450305 -0.53523178  0.22900376 -0.40906173 -0.65856872
## [5,] -0.4546568  0.22305882  0.85982044 -0.06406746 -0.01166667
```

```r
E <- eigen(cov(X))$vectors
scores <- X%*%E
colnames(scores) <- paste0("pc", 1:p)
cor(cbind(X, scores))[1:p, (p+1):(p*2)]
```

```
##            pc1         pc2         pc3        pc4         pc5
## X1   0.6381822  0.71536319  0.009665104  0.2581645 -0.119317350
## X2 -0.5701649  0.32632936 -0.213505507 -0.4289443 -0.582102493
## X3   0.8417769 -0.06005797  0.312645510 -0.4358132  0.011149679
## X4   0.4130194 -0.64271492  0.232581875  0.3517727 -0.488358607
## X5 -0.6579685  0.23893819  0.712352851  0.0501983 -0.004418134
```

Part III: Introduction to Linear Models

Chapter 8

A linear model with a single, continuous X

8.1 A linear model with a single, continuous X is classical "regression"

To introduce modeling with a single continuous X variable, I'll use data from

1. Source: Dryad Digital Repository. https://doi.org/10.5061/dryad.b3h4q
2. File: "FCM data dryad.csv"

The data are from **?**, who showed that North American red squirrel (*Tamiasciurus hudsonicus*) mothers from Yukon, Alaska produce faster growing pups in years with increased squirrel density. Remarkably, they even showed that perceived (but not actual) density results in faster growing pups. To begin to investigate how pregnant mothers control the future growth rate of pups, the researchers measured the relationship between local squirrel density and the amount of fecal cortisol metabolites from pregnant mothers. Cortisol is a hormone that is secreted as part of stress response. The researchers were interested in cortisol because it had previously been shownt that, in mammals, blood cortisol levels in pregnant mothers have numerous effects on offspring long past birth. If increased squirrel density causes increased blood cortisol levels then we would expect to find a positive relationship between *Density* and

Figure 8.1 is a **scatterplot** of the data with the amount of cortisol metabolites in the feces on the Y axis and local squirrel density on the X axis. The line through the data is a graphical representation of a linear model fit to the data and the gray cloud around the line is a graphical representation of the uncertainty in the model. The researchers wanted to model the "effect" of squirrel density on the amount of cortisol metabolites in the feces of the pregnant mothers. Graphically, this effect is the slope of the line in Figure 8.1.

The model fit to the data is

$$FCM_i = \beta_0 + \beta_1 Density_i + \varepsilon_i \tag{8.1}$$

which contains both the linear predictor and the error. For inference, for example, computing standard errors of the coefficients, We need to model the error. Here, we use the simplest model of error which is "IID $N(0, \sigma)$". This means, the modeled error is

1. Independent – individual error values are independent of other values.
2. Identical – individual error can be thought of as a sample from a single **random distribution** (the same for each individual value). For this model, this distribution is
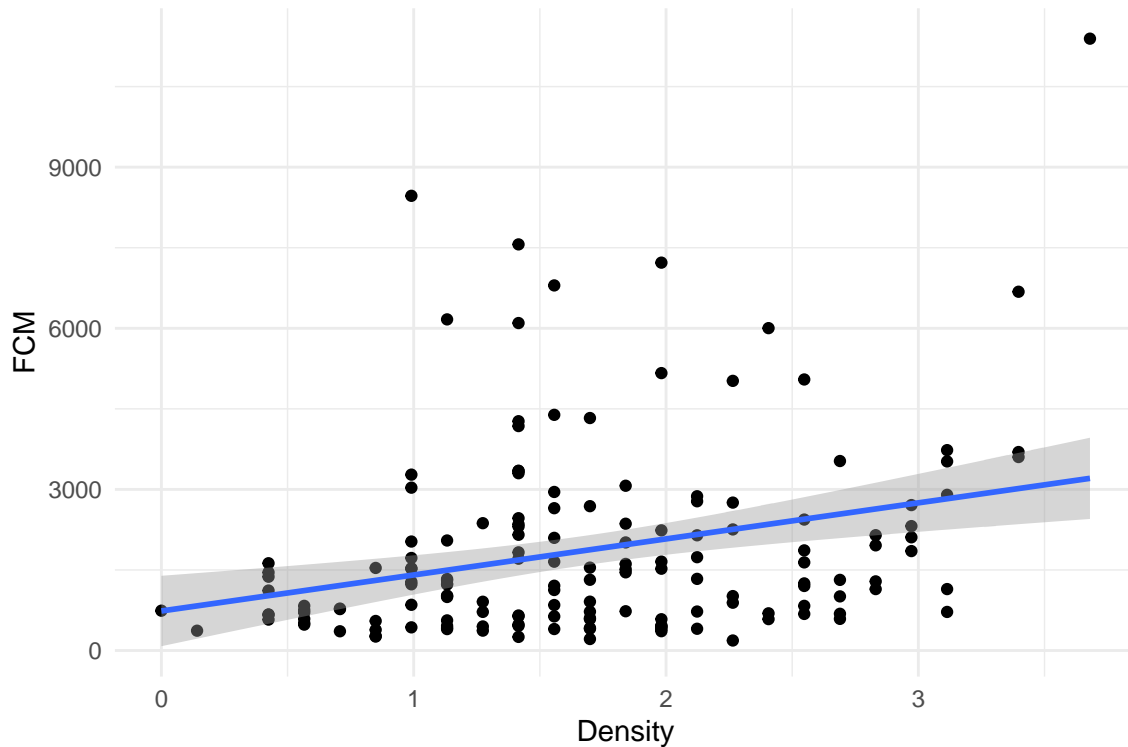
Figure 8.1: A scatterplot of Fecal cortisol matabolites and squirrel density.

3. $N(0, \sigma)$ – the modeled distribution is "Normal" or "Gaussian", with a mean of zero and a standard deviation of σ.

The predictor part of the model is

$$\mathrm{E}[FCM|Density] = \beta_0 + \beta_1 Density \tag{8.2}$$

In words, model (8.2) reads "the expected value of FCM conditional on density is beta-knot plus beta-one times density". An **expected value** is a long run average – if we were to sample lots and lots of red squirrel populations with $Density = x$ (where x is a specific value), we'd expect the average FCM across these samples to be $\beta_0 + \beta_1 x$.

Let's unpack this. $\mathrm{E}[Y]$ is the **expectation** or **expected value** of Y. An expection is the long-run average of Y if we were to run an experiment or re-sample a population many times. The sample mean of Y is an estimate of $\mathrm{E}[Y]$. $\mathrm{E}[Y|X]$ is a conditional expectation of Y – it is the expectation given additional conditions. Using the red squirrel example, these conditions are a specific value of $Density$. If FCM is linearly related to $Density$ (the right-hand side of equation (8.2)) then the expected value of FCM given a local density of 2.8 squirrels differs from the expected value of FCM given a local density of 1.4 squirrels (the units of $Density$ are squirrels per 150 meter radius of the individual female's midden).

In model (8.2), there is a single X variable (FCM). While the X variables are often called the "dependent" variables, in this model FCM does not "depend" on the independent variable $Density$ in any causal sense – meaning if I were to intervene and set $Density$ to some value x, I would expect FCM to equal $\beta_0 + \beta_1 x$. Rather, FCM only "depends" on $Density$ in a probablistic sense – if $Density = x$ then the most probable value of FCM is $\beta_0 + \beta_1 x$. With some strong assumptions model (8.2) can be turned into a model of causal dependency, which is the focus of chapter xxx.

β_0 and β_1 are the **parameters** of model (8.2). Specifically β_0 is the model **intercept** and β_1 is the modeled **effect** of *Density*. Again, the effect (β_1) has a probabilistic, and not causal, interpretation. This interpretation is

$$\beta_1 = \mathrm{E}[FCM|Density = x + 1] - \mathrm{E}[FCM|Density = x] \tag{8.3}$$

Or, in words, "beta-1 is the expected value of FCM when density equals x + 1 minus the expected value of FCM when the density equals x." β_1 is simply the difference in expected values given a one unit difference in *Density*. A very short way to state this is "β_1 is a difference in conditional means".

8.1.1 Using a linear model to estimate explanatory effects

The goal of the statistical model here is to estimate β_1 – the probabalistic effect of *Density* on *FCM*. This estimate, and a measure of the uncertainty of this estimate, are in the table of coefficients of the fit model

	Estimate	Std. Error	t value	Pr($>$|t|)
(Intercept)	736.0	331.9	2.2	0.0281
Density	671.1	178.9	3.8	0.0002

where the entries in the column "Estimate" are estimates of the parameters β_0 and β_1 in model (8.2). The entries in the column "Std. Error" are the standard errors (SE) of the estimates, which are measures of the uncertainty of the estimates.

The parameter estimates in the table above are the coefficients of the fitted model

$$FCM_i = b_0 + b_1 Density_i + e_i \tag{8.4}$$

where the subscript i refers to the ith individual. The coefficients b_0 and b_1 are the y-intercept and the slope of the line in Figure 8.1. The coefficient for *Density* (b_1) is 671.1, and (given the definition of the parameter β_1 in equation (8.3)) we expect squirrel mothers with a local density of 2 squirrels within a 150 m radius of her midden to average 671.1 more units of FCM (ng of fecal cortical metabolites per gram dry food) than mother squirrels with a local density of only 1 squirrel within a 150 m radius of her midden.

8.1.1.1 Probabilistic vs. causal conditioning

Remember that this coefficient is estimating a probabilistic parameter. Consequently, the coefficient b_1 is simply a descriptor of a pattern of relationship between local density and fecal cortisol metabolites - no causal effect is implied. With the strong assumptions explained in chapter xxx, however, b_1 can estimate a causal effect.

8.1.2 Using a linear model for prediction

Model (8.4) gives the measured value of *FCM* for each squirrel. The equation includes the linear predictor ($b_0 + b_1 Density_i$) and the **residual** from the predictor (e_i). The predictor part is called "predictor" because it is the equation for predicting the value of an individual's *FCM* given that individual's value of *Density*:

$$\widehat{FCM} = b_0 + b_1 Density \tag{8.5}$$

where $\widehat{FCM}$ is read as "FCM hat" and is the **predicted value** or simply "prediction". Very often, we use the predictor part (equation (8.5)) to predict unknown or future values given different modeled inputs (the X).

8.1.3   Reporting results

The authors of the squirrel fcm data published a figure and table similar to fig. xxx and table above but used a slightly more complex linear model. Here is how the author's reported the results:

> Across 6 years (2006 to 2011), we found a positive relationship between local density and concentrations of fecal cortisol metabolites [FCM; $t_1 55 = 3.63$, P = 0.0002 (table S4 and Fig. 3A)].

I would advocate reporting the estimate and a confidence interval instead of t and p. For example "Across 6 years (2006 to 2011), the probabilistic effect of local density on fecal cortisol metabolites is 671.1 (95% CI: 317.7, 1024.5). If a p-value is report *in addition* to the effect and CI, always report the exact p-value, which emphasizes the continuous nature of evidence against the null, and not something like"$p < 0.05$", which artificially dichotomizes the evidence against the null.

8.2   Working in R

8.2.1   Exploring the bivariate relationship between *Y* and *X*

Questions

1. Import the "FCM data dryad.csv" data from the Dryad repository as the data.table `fcm`
2. How are different words in the column labels demarcated? Is this good practice?

Here we want to fit a model of `FCM.ng.g.dry` as a function of `Raw.Squirrel.Density`. The authors used prior knowledge to expect a positive relationship between these two variables. Use qplot to generate a scatterplot of *FCM* against *Density*

Questions

3. Is there a trend? If so, does the trend look linear or non-linear?
4. Does the residual variation (the deviation from the trend on the *Y* axis) look homogenous along the *X*-axis?
5. Are there any obvious outliers?

8.2.2   Fitting the linear model

We will fit a linear model to the data using the `lm` function, which is very general and will be our workhorse throughout the class. The minimal input to the function is a model formula and the name of the data.frame (remember, a data.table is a data.frame). A formula is of the form `Y ~ X`. All of the output we assign to the object `fit`.

Let's fit the linear model to the data using density as the predictor

```
fit <- lm(FCM.ng.g.dry ~ Raw.Squirrel.Density, data=fcm)
```

R will look for the specified *Y* and *X* variables in the column names of `fcm`. If these are not found, R will return an error, for example

```
fit <- lm(FCM_ng_g_dry ~ Raw_Squirrel_Density, data=fcm)
```

will return the error "Error in eval(predvars, data, env) : object 'FCM_ng_g_dry' not found". This means your spelling and capitalization have to be exact!

8.2.3 Getting to know the linear model: the `summary` function

The `lm` function returns an `lm` object, which we've assigned to the name `fit`. `fit` contains lots of information about our fit of the linear model to the data. Most of the information that we want for most purposes can be retrieved with the `summary` function, which is a general-purpose R command the works with many R objects.

```
summary(fit)
```

```
##
## Call:
## lm(formula = FCM.ng.g.dry ~ Raw.Squirrel.Density, data = fcm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2107.5 -1108.3  -434.9   511.8  8186.8
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              736.0      331.9   2.217 0.028078 *
## Raw.Squirrel.Density     671.1      178.9   3.752 0.000248 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1732 on 154 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.08374,    Adjusted R-squared:  0.07779
## F-statistic: 14.07 on 1 and 154 DF,  p-value: 0.0002484
```

What is here:

Call. This is the model that was fit

Residuals. This is a summary of the distribution of the residuals. From this one can get a sense of the distribution (for inference, the model assumes a normal distribution with mean zero). More useful ways to examine this distribution will be introduced later in this chapter.

Coefficients table. This contains the linear model coefficients and their standard error and associated t and p values.

1. The column of values under "Estimate" are the coefficients of the fitted model (equation (8.4)). Here, 735.9604344 is the intercept (b_0) and 671.1379749 is the effect of *Density* (b_1).
2. The column of values under "Std. Error" are the standard errors of the coefficients.
3. The column of values under "t value" are the *t-statistics* for each coefficient. A t-value is a **signal to noise ratio**. The coefficient b_1 is the "signal" and the SE is the noise. Get used to thinking about this ratio. Any t less than 2 is indicative of too much noise to say much about the signal. A t between 2 and 3 means the noise is large enough to suggest an effect. A t greater than 3 is pretty good evidence of an effect.
4. The column of values under "Pr(>|t|)" is the p-value, which is the exact probability associated with a particular t. What is the p-value a test of? The p-value tests the hypothesis "how probable are the data if the coefficient is zero?". Formally $P = \text{freq}(t' \geq t | H_o)$, where t' is the hypothetical t-value, t is the observed t-value, and H_o is the null hypothesis. We will return to p-values in Chapter xxx.

Signif. codes. I am surprised that base R returns this. These are useless because the concept of "levels of significance" is muddled, as will be discussed in Chapter xxx.

Beneath the Signif. codes are some model statistics which are useful

Residual standard error This is $\sqrt{\sum e_i^2/(n-2)}$, where e_i are the residuals in the fitted model. "degrees of freedom" is the number of e_i that are "allowed to vary" after fitting the parameters, so is the total sample size (n) minus the number of parameters fit. The fit model has two fit parameters (b_0 and b_1 so the df is $n-2$. Note that this is the denominator in the residual standard error equation.

Multiple R-squared. This is an important but imperfect summary measure of the whole model that effectively measures how much of the total variance in the response variable "is explained by" the model. Its value lies between zero and 1. **It's a good measure to report in a manuscript**.

F-statistic and p-value. These are statistics for the whole model (not the individual coefficients) and I just don't find these very useful.

Note that the p-value for the coefficient for Raw.Squirrel.Density is very small and we could conclude that the data are not consistant with a model of no slope. But did we need a formal hypothesis test for this? We haven't learned much if we have only learned that the slope is "not likely to be exactly zero". What we want to know is not *if* there is a relationship between FCM and $Density$, which is imperfectly answered with a p-value, but *the sign and magnitude* of the relationship and the uncertainty in this estimate. For this, we don't need the p-value. Instead, we want to interpret the coefficient to its SE directly (for a quick-and-dirty interpretation) or the confidence interval of the effect (for a more formal interpretation). Please read this paragraph again. We will come back to it over and over.

8.2.4 display: An alternative to summary

Much of what we want to know about a model fit is returned by the `display` function from the `arm` package.

```
display(fit)
```

```
## lm(formula = FCM.ng.g.dry ~ Raw.Squirrel.Density, data = fcm)
##                      coef.est coef.se
## (Intercept)           735.96   331.94
## Raw.Squirrel.Density  671.14   178.90
## ---
## n = 156, k = 2
## residual sd = 1732.02, R-Squared = 0.08
```

The `display` function does not give a t-value or a p-value of the coefficients because the authors of the arm package do not think p-values are very valuable. We don't need a t because one can mentally compute the approximate ratio of the coefficient to its SE and get a sense of the signal to noise, and that's all the authors of the display function think we need.

8.2.5 Confidence intervals

Confidence intervals for the coefficients of the model are obtained by

```
confint(fit)
```

```
##                          2.5 %    97.5 %
## (Intercept)           80.21785 1391.703
## Raw.Squirrel.Density 317.73057 1024.545
```

`confint` returns by default the 95% confidence interval (CI) of all parameters. The most useful way of thinking about the meaning of a CI is

A confidence interval contains the range of parameter values that are consistent with the data, in the sense that a t-test would not reject the null hypothesis of a difference between the estimate and any value within the interval

A more textbook way of defining a CI is: A 95% CI of a parameter has a 95% probability of including the true value of the parameter. It does not mean that there is a 95% probability that the true value lies in the interval. This is a subtle but important difference. Here is a way of thinking about the proper meaning of the textbook definition: we don't know the true value of β_1 but we can 1) repeat the experiment or sampling, 2) re-estimate β_1, and 3) re-compute a 95% CI. If we do 1-3 many times, 95% of the CIs will include β_1 within the interval.

Confidence intervals are often interpreted like p-values. That is, the researcher looks to see if the CI overlaps with zero and if it does, concludes there is "no effect". First, this conclusion is not correct – **the inability to find sufficient evidence for an effect does not mean there is no effect, it simply means there is insufficient evidence to conclude there is an effect**!

Second, what we want to use the CI for is to guide us about how big or small the effect might reasonably be, given the data. Again, A CI is a measure of parameter values that are "consistent" with the data. If our biological interpretations at the small-end and at the big-end of the interval's range radically differ, then we don't have enough *precision* in our analysis to reach an unambiguous conclusion. Remember this.

8.2.6 How good is our model?

How well does variation in *Density* "explain" variation in *FCM*? The answer to this is in the R^2 value, which is given in `display(fit)` and in `summary(fit)` and accessed directly with

```r
summary(fit)$r.squared
```

```
## [1] 0.08373756
```

R^2 is the fraction of the total variance of Y explained by the model, or more specifically, the linear predictor. It will vary from zero (the model explains nothing) to one (the model explains everything). If $R^2 = 0$ the response is completely unpredictable by the predictors. We can think of the values of the response as white noise or all error. This doesn't mean that the values are "not caused" or "random" or not predicted by some other variable. It only means the values are random with respect to the X variable(s) in the model. If $R^2 = 1$ we can *exactly* predict the response from the X variables in the model. So the bigger the R^2, the better the model in the sense that the response is more predicatable. **Super importantly**, "explains" is in a probabilistic and not causal sense. We will explore this concept much more in future worksheets.

8.2.7 Model checking

`plot` is a very useful base R function for "model checking" or "model diagnostics" to see if our model fit is acceptable.

```r
plot(fit)
```

Residuals vs Fitted

lm(FCM.ng.g.dry ~ Raw.Squirrel.Density)

Scale–Location

lm(FCM.ng.g.dry ~ Raw.Squirrel.Density)
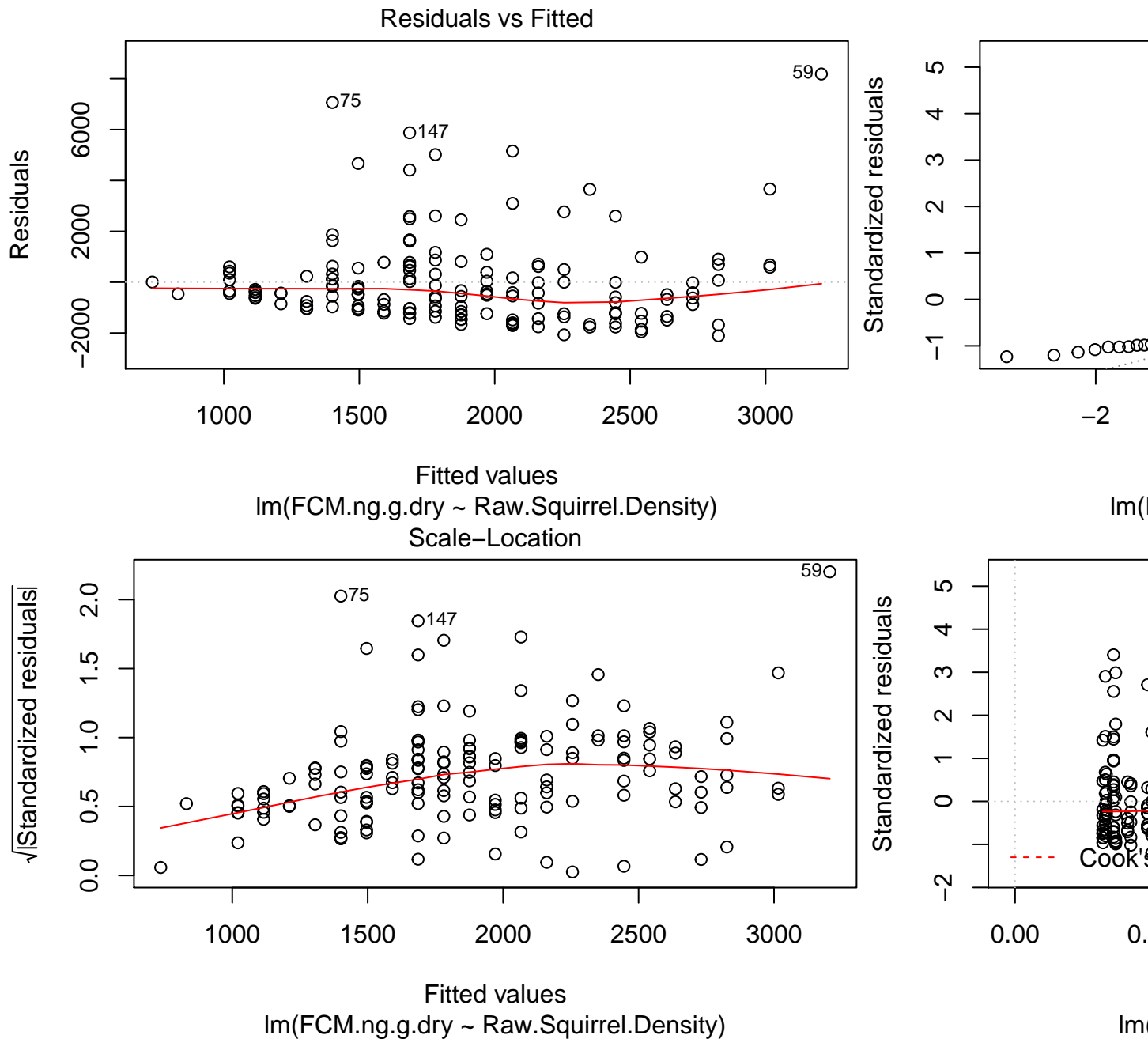
Compare the four diagnostic plots using the guidelines from here http://data.library.virginia.edu/diagnostic-plots/

Questions

6. Look at the plots you just made. What is a residual? What is a fitted value?

8.2.8   exploring a lm object

`fit` contains much information but simply typing `fit` into the console gives us only the model and the coefficients. `names` is a super important R function. It gives us the names of all the parts of some R object. `fit` is an lm object. `names(fit)` gives us all the parts contained in an lm object.

```
names(fit)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "na.action"     "xlevels"       "call"          "terms"
## [13] "model"
```

You can see any of these parts using the dollar sign

Questions

7. What does `fit$residuals` return? Answer using equation (8.4)

8. What does `fit$fitted.values` return? Answer using equation @ref(eq:fcmi

You can use qplot to make a plot similar to the first plot of `plot(fit)`

```
qplot(fit$fitted.values, fit$residuals, geom=c('point', 'smooth'))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```
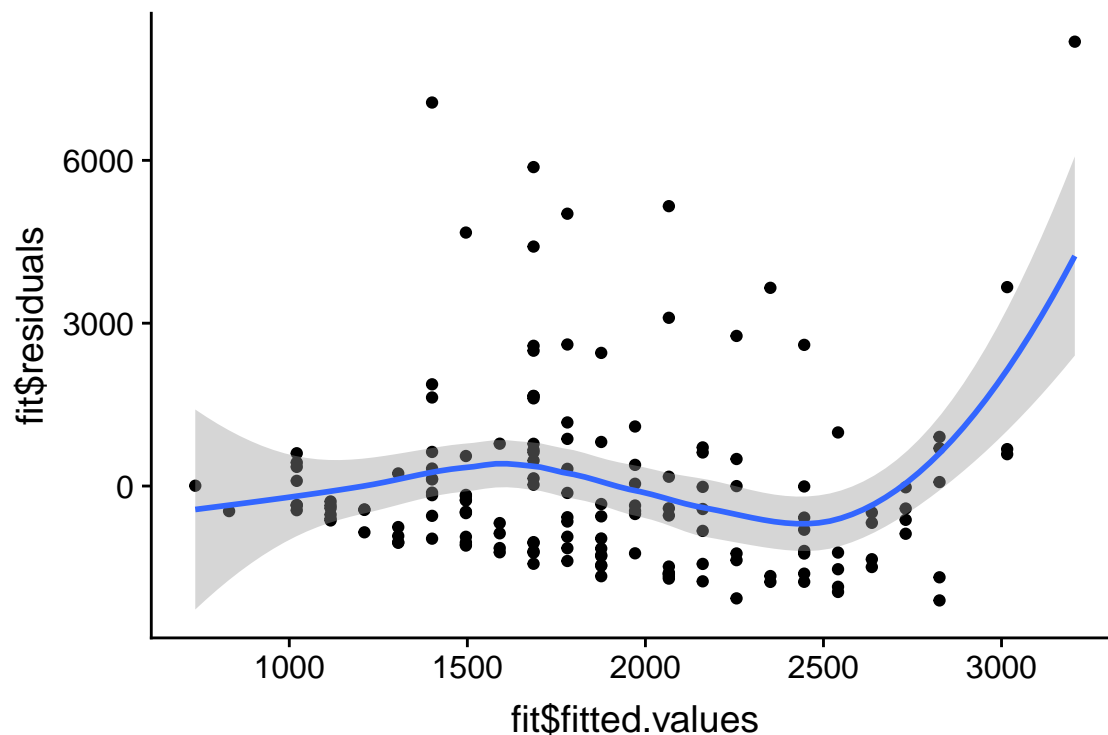

8.3 Problems

1. Using the chick data from Chapter 3. Compare the effects of nest_temperature_above_ambient on day13_mass by fitting two separate linear models 1) one using only the control group and one using the treated group. The grouping variable is playback_treatment. These models were plotted in Chapter 3 so `lm` will return the linear model behind these plots.

Report the results using the two effect estimates and a 95% confidence interval (we will learn in a later chapter a more sophisticated way of comparing the effects between the groups)

file name: "allDatasetsMarietteBuchanan2016.xls"

source: https://datadryad.org//handle/10255/dryad.122315

2. (Grad students only) – find a dataset using Dryad that has data that can be fit by a simple linear model with a single continuous X (its okay if the authors fit the data with a more complex model). Fit the data and report the results with a plot and text.

Chapter 9

A linear model with a single, categorical X

9.1 A linear model with a single, categorical X is the engine behind a single factor (one-way) ANOVA and a t-test is a special case of this model.

To introduce modeling with a single, categorical X variable, I'll use the vole data from

1. Source: Dryad Digital Repository. https://doi.org/10.5061/dryad.31cc4/1
2. File: "RSBL-2013-0432 vole data.xlsx"
3. Sheet: "COLD VOLES LIFESPAN"

Normal cellular metabolism creates reactive oxygen species (ROS) that can disrupt cell function and potentially cause cell damage. Anti-oxidants are molecules that bind ROS, inhibiting their ability to disrupt cell activity. A working hypothesis for many years is that supplemental anti-oxidants should improve cell function and, scaling up, whole-animal function (such as lifespan). The vole data explores this with supplemental Vitamins C and E, which are anti-oxidants, in the diet of the short-tailed field vole (*Microtus agrestis*).

The goal of the study is to measure the effect of anti-oxidants on lifespan. The researchers randomly assigned the voles to one of thre treatment levels: "control", "vitamin E", and "vitamin C". The variable *treatment*, is a single, categorical X variable. Categorical variables are often called **factors** and the treatment levels are often called **factor levels**. There are no units to a categorical X variable (even though a certain amount of each anti-oxidant was supplemented). The response (Y) is *lifespan* measured in days.

The linear model with a categorical X variable with three levels is not immediately obvious, and so I don't present the model until after showing the table of model coefficients

9.1.1 Table of model coefficients

Here is the table of coefficients from the linear model fit

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	503.4	27.4	18.4	0.000
treatmentvitamin_E	-89.9	52.5	-1.7	0.090
treatmentvitamin_C	-115.1	54.5	-2.1	0.037

The table has estimates for three parameters. The first estimate (the intercept) is the mean response in the reference level. Here the reference level is the "control" group. The additional estimates are the differences
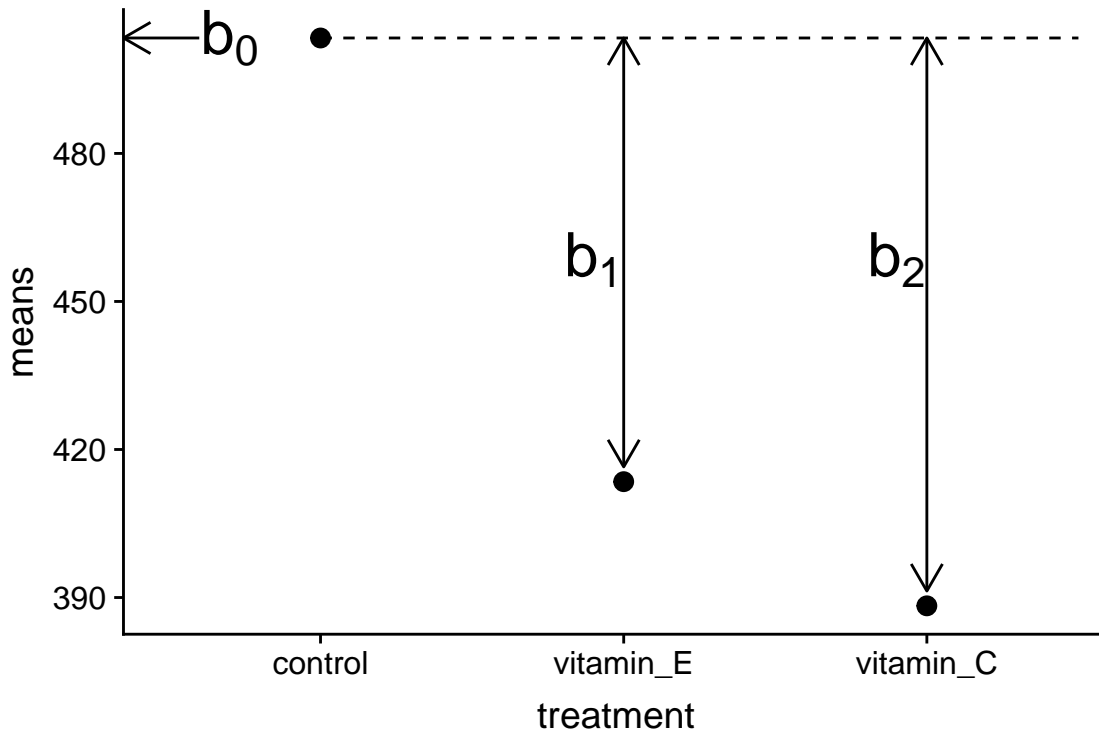
Figure 9.1: What the coefficients of a linear model with a single categorical X mean. The means of the three treatment levels for the vole data are shown with the filled circles. The length of the double-headed arrows are differences in means. The intercept (b_0) is the mean of the reference treatment level. The coefficients (b_1 and b_2) are the differences between the treatment level's mean and the reference mean. As with a linear model with a continuous X, the coefficients are effects.

in the mean between each of the other treatment levels and the reference level. These are the "effects" in the model. So typically with categorical X, when we speak of an *effect* we mean a difference in means. These estimates and their meaning are illustrated in Figure 9.1.

(note. The default in R is to set the level that is first alphabetically as the reference level. In the vole data, "control" comes before "vitamin_E" and "vitamin_C" alphabetically, and so by default, it is the reference level. This makes sense for these data – we want to compare the lifespan of the vitamins E and C groups to that of the control group. The reference level can be changed of course.)

9.1.2   The linear model

We can see an immediate difference between the coefficient table for a linear model fit to a single, categorical X and that for a single, continuous X. For the latter, there is a single coefficient for X. For the former, there is a coefficient for each level of the categorical X *except* the "reference" level.

The linear model for a single, continuous X with three factor levels is

$$lifespan = \beta_0 + \beta_1 vitamin\_E + \beta_2 vitamin\_C + \varepsilon \qquad (9.1)$$

and the estimates in the coefficient table are the coefficients of the fit model

$$lifespan_i = b_0 + b_1 vitamin\_E + b_2 vitamin\_C + e_i \qquad (9.2)$$

Remember, b_0 is the mean of the control group, b_1 is the difference in means between the vitamin E and control groups, and b_2 is the difference in means between the vitamin C and control groups (Figure 9.1).

In this model, *vitamin_E* and *vitamin_C* are **dummy variables** that contain a one, if the data is from that treatment level, and zero otherwise. This is called dummy coding or treatment coding. The `lm` function creates these dummy variables under the table, in something called the **model matrix**, which we'll cover in the next chapter. You won't see these columns in your data. But if you did, it would look something like this

lifespan	treatment	vitamin_E	vitamin_C
621	control	0	0
865	control	0	0
583	vitamin_E	1	0
561	vitamin_E	1	0
315	vitamin_C	0	1
157	vitamin_C	0	1

There are alternative coding methods. Dummy coding is the default in R. Note that the method of coding can make a difference in an ANOVA table, and many published papers using R have published incorrect interpretations of ANOVA table outputs. This is both getting ahead of ourselves and somewhat moot, because I don't advocate publishing ANOVA tables.

9.1.2.1 Some math to convince you that the intercept of a linear model with a categorical X is the mean of the reference group *and* the intercept of a line. And some math to convince you that the coefficient of a dummy variable in a linear model with a categorial X is a difference in means *and* a slope.

The interecept of a model is the value of the model when all X-variables are set to zero. The X variables in the model (Equation (9.2)) are the dummy variables *vitamin_E* and *vitamin_C*. If we set vitamin_E\$ and *vitamin_C* in Equation (9.2) to zero, the modeled (or expected) value reduces to

$$\text{E}(lifespan|X_1 = 0, X_2 = 0) = b_0 \tag{9.3}$$

% Since both dummy variables are set to zero, we have modeled the expected value or mean of the control group.

The slope of a model is the difference in the modeled value given a one unit increase in X. If we increase the dummy variable *vitamin_E* from zero to one (that is, if we are modeling the expected value of the vitamin E group), we get

$$\text{E}(lifespan|X_1 = 1, X_2 = 0) = b_0 + b_1 \tag{9.4}$$

which can be re-arranged to

$$b_1 = \text{E}(lifespan|X_1 = 1, X_2 = 0) - b_0 \tag{9.5}$$

and since $\text{E}(lifespan|X_1 = 0, X_2 = 0) = b_0$ then

$$b_1 = \text{E}(lifespan|X_1 = 1, X_2 = 0) - \text{E}(lifespan|X_1 = 0, X_2 = 0) \tag{9.6}$$

or, the coefficient of vitamin E is the difference in means between the vitamin E and control groups, which is also a slope since this is the expected difference given a one unit increase in *vitamin_E*.
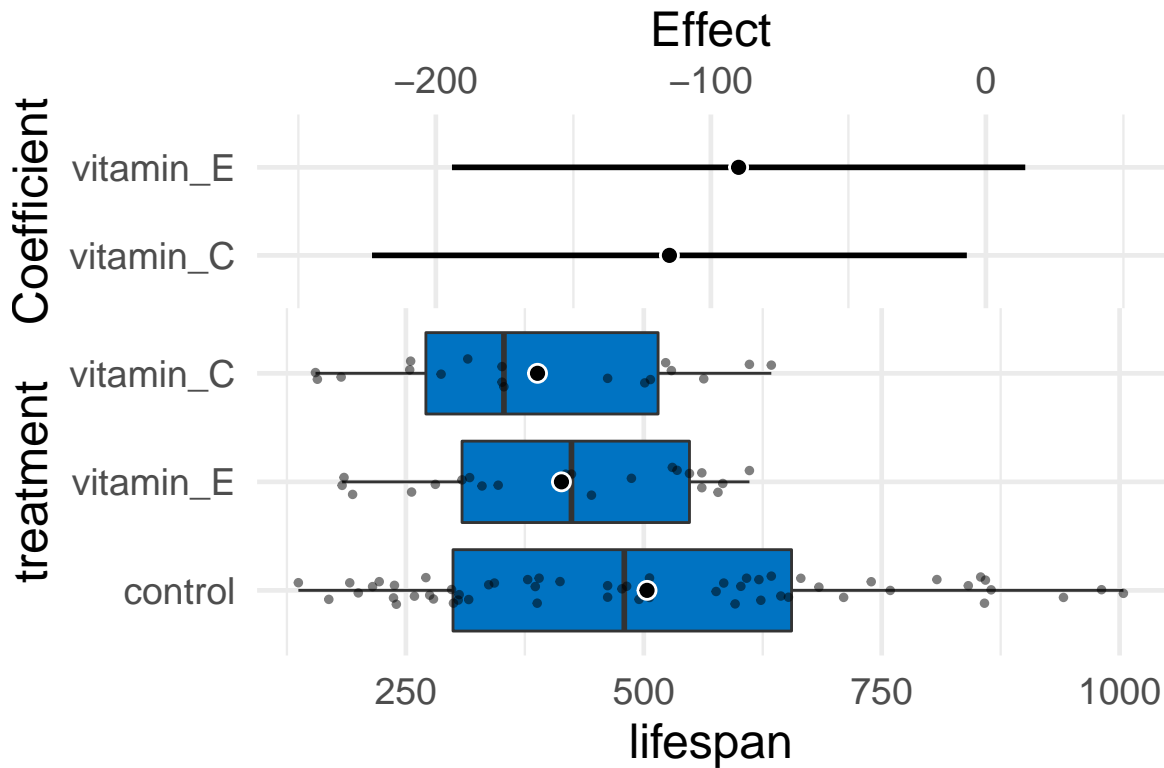
Figure 9.2: HarrellPlot of the raw data, distribution, and effects of the vole lifespan data.

9.1.3 Reporting results

What should be reported for the analyis of effects of anti-oxidant supplements on vole lifespan? Best practice includes reporting the raw data with a summary distribution and treatment effects with CIs. "Raw data" means the individual lifespans as a function of treatment level.

9.1.3.1 Harrell Plot of the data

The raw data, the distributions within treatment level, and the effects (difference in means) of treatment can be combined into a single plot that I call a Harrell plot (Figure **??**). Notice that the x-axis and y axes are flipped so that *lifespan* is on the x-axis. It is still the "response" or "Y" variable! The Harrell plot contains two parts

1. The bottom contains a **strip chart** (often called a "dot plot") of the raw response measures grouped by factor level. Superimposed over the strip chart is a **box plot** summarizing the distribution of each factor level. The line in the center of a box is the median *lifespan* for that group, the left and right edges of the box are the 25% and 75% quantiles of *lifespan* for that grop, and the lines extending to the left and right of the box are the "whiskers", which are the smallest and largest value within $1.5IQR$ (inter-quartile range, which is the interval bounded by box).
2. The top is a **forest plot** of the effects and the 95% CI of the effects. For categorical X, the effects could be model coefficients or treatment **contrasts**, which are differences in means between treatment levels. Model coefficients are a subset of possible treatment contrasts.

The Harrell plot above shows the effects as model coefficients, which (again!) are differences between the mean of the response in a specific treatment level and the mean of the response in a reference level. Here the reference level is the control group.

9.1.3.2 In-text reporting

"The mean lifespan of cold-reared voles supplmented with vitamin E was -89.9 days shorter than the mean lifespan for the control group (95% CI: -194.1, 14.3). The mean lifespan of cold-reared voles supplmented with vitamin C was -115.1 days shorter than the mean lifespan for the control group (95% CI: -223.2, -6.9).

9.1.3.3 Correct interpretation of the Confidence Interval is key

Remember, that the CI contains the range of parameter values that are consistent with the data (in the sense that a t-test wouldn't reject the hypothesis test). This means that a true value at the low end or the high end of the CI is consistent with the data. Your technical report/manuscript should discuss the consequences of this. For example, A small, increase in lifespan is consistant with the Vitamin E but not Vitamin C supplementation, if we use the 95% CI as a pretty good range for inferring "consistent with". Both a 223 day and a 7 day decrease in lifespan are consistant with the Vitamin C effect. 223 days seems like a huge effect, especially for a short lived vole. 7 days is certainly a much smaller effect, but this doesn't mean that it doesn't have important ecological, behavioral, or fitness consequences.

9.2 Working in R

9.2.1 Exploring the relationship between Y and X

Questions

1. Import the vole data from the Dryad repository as the data.table `vole_wide`. Replace the spaces in the column labels with the underscore "_".

The data are in "wide" format. A pretty good script for for converting these to long format is

```
vole_long <- melt(vole_wide, measure.vars=c("control", "vitamin_E", "vitamin_C"), variable.name="treatm
vole <- na.omit(vole_long)
# melt is from reshape2 package and is a workhorse in R
# the resulting data frame has 3 stacked copies of the same rows
# na.omit removes the superfluous two extra sets of rows created by melt
# the more compact way to do this is combine the steps:
# vole <- na.omit(melt(vole_wide, measure.vars=c("control", "vitamin_E", "vitamin_C"))
# but I suggest two steps so you can see what melt does. This isn't the best example of using melt.
```

Use the ggpubr package to create a box plot, grouped by treatment, with superimposed strip chart ("dots")

Questions

2. Do the response as a function of group look fairly normally distributed or are there red flags such as skewness, outliers, bimodal clusters, etc.
3. Is the direction of the effect consistent with the expected direction?

9.2.2 Fitting the model

As with a single, continuous X, we fit the model using the `lm` function and with the model formula of the form `y ~ x`. Note that the R formula can use the single categorical variable *treatment*. The code underneath lm will note that *treatment* is a factor with three levels and will automatically create the two dummy variables noted above in the linear model.

```
fit <- lm(lifespan ~ treatment, data=vole)
```

All of the same scripts to access the information in `fit` that we used with the continuous X analysis are the same. For example, the base R `summary` function gives the same information as in the continuous X example.

Questions

4. Review different output in the `summary` function and list which are useful and which are not so useful and why.

Other useful functions on the lm object ("fit") are `coefficients(summary())`, `coef()` or `coefficients()` and `confint`. Assigning the output of these functions to an R object allows you to increase reproducibility. For example, if I assign the coefficients to `b`

```
b <- coef(fit)
```

then I can report these in R markdown text by embedding r code directly in the text. For example, if I embed "r round(b["treatmentvitamin_E"], 1)" between a pair of single accent characters, then r markdown inserts -89.9 into the rendered text.

9.2.3   An introduction to contrasts

We often want to compare more than just the non-reference levels to the reference level. For example, we might want to compare the effects of the vitamin E supplementation to vitamin C supplementation. Or, we might want to combine (or "pool") vitamin C and vitamin E levels effects into a single "anti-oxidant" level and compare to the control. These comparisons of means are called linear **contrasts**. The emmeans package is a good package for obtaining contrasts for both simple linear models computed with `lm` and for more complicated statistical models. If you haven't already, download the emmeans package.

```
fit.em <- emmeans(fit, spec="treatment")
fit.em
```

```
##  treatment emmean    SE df lower.CL upper.CL
##  control      503 27.4 93      449      558
##  vitamin_E    413 44.8 93      325      502
##  vitamin_C    388 47.1 93      295      482
##
## Confidence level used: 0.95
```

The `emmeans()` function returns various estimated means, depending on what is specified with the `spec=` parameter. Here the grouping variable "treatment" is specified, so the means returned are estimates of $E(lifespan|treatment)$, the modeled means for each level of treatment. For this simple analysis, the modeled means are simply the group means. Note that the default value returned is a table with the standard error and 95% confidence limits of the estimates.

Let's use the emmeans object to get the contrasts for all combinations of treatment levels.

```
summary(contrast(fit.em, method="revpairwise", adjust="none"), infer=c(TRUE, TRUE))
```

```
##  contrast               estimate   SE df lower.CL upper.CL t.ratio p.value
##  vitamin_E - control       -89.9 52.5 93     -194    14.31  -1.713  0.0900
##  vitamin_C - control      -115.1 54.5 93     -223    -6.93  -2.113  0.0373
##  vitamin_C - vitamin_E     -25.2 64.9 93     -154   103.81  -0.387  0.6993
##
## Confidence level used: 0.95
```

I've sent to parameters to the contrast function and one to the summary function

1. method="revpairwise". `contrast` can create different combinations of differences between means. Here I've specified all pairwise differences (the "rev" reverses the order of the subtraction). Notice that the

statistics (estimate, SE, etc) are equal to the same statistics for b_1 and b_2 of the linear model. I said earlier that these coefficients are contrasts!

2. adjust="none". In classical frequentist hypothesis testing, the p-value of a contrast in what are called "post-hoc tests" is adjusted to reflect "multiple testing" (more than one p-value is being computed). This adjustment is almost standard in biology, but the practice is hugely controversial. The concept of multiple testing is important, and we will return to this in a future chapter, but here I have chosen to show the unadjusted p-value. The reason is that I want the unadjusted confidence interval and the adjustment would adjust these as well. If deleted `adjust="none"` from the script, the contrast function would default to the **Tukey HSD** (Honestly Significant Difference) test. There are literally dozens and dozens of post-hoc tests, which largely reflects the misplaced emphasis on "better" p-values rather than parameter estimates and their uncertainty.

3. infer=c(TRUE, TRUE). This parameter controls what kind of inference to put in the table. The first value specifies the inclusion of the CI (emmeans uses "CL" for confidence limit), the second value specifies the inclusion of t and p-values.

9.2.4 Harrell plot

9.2.4.1 Installing the harrellplot package

The harrellplot package is available on github but not a cran repository and, therefore, takes a little more work to install. To install a package from a github repository, 1. load library(devtools) – this may need to be installed first using the R Studio Tools > Install Packages... tool 2. install harrellplot from github. In the console, type

```
install_github("middleprofessor/harrellplot")
```

3. load the harrellplot package
4. harrellplot requires other packages including broom, Hmisc, car, lme4, and lmerTest. If you haven't installed these do. load these with the library() function at the start of your notebook.

9.2.4.2 Using harrellplot to make a nice, publishable plot of treatment effects

In the console type '?harrellplot to see the many parameters. Unlike ggplot2, variable names need to be specified with quotes in the harrellplot function. The harrellplot function is a list with several elements.

Here is the default plot

```
vole.harrellplot <- harrellplot(x="treatment", y="lifespan", data=vole)
vole.harrellplot$gg # gg is the plot object
```
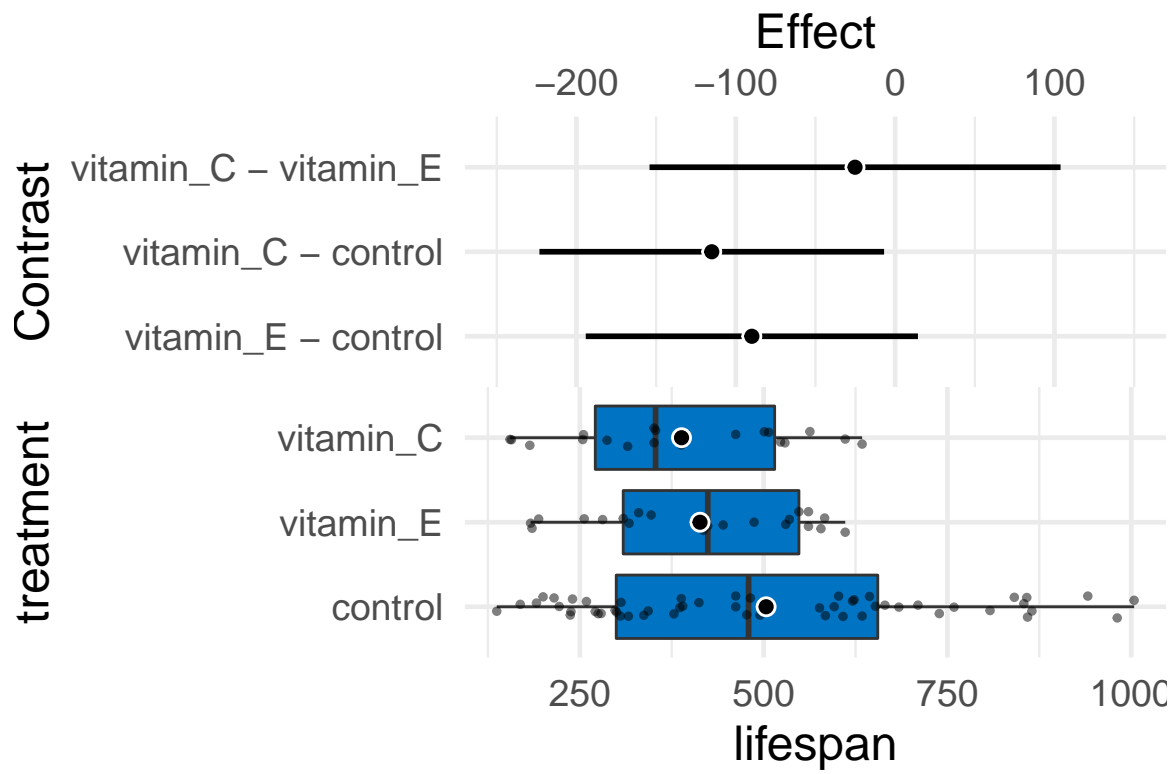
Chapter 10

Model Checking

```
# a function to transform a vector into quantiles
# not if the data are 1:n then the output is "rankits"
quantilize_1 <- function(x){
  # this is the ppoints(x) function
  m <- length(x)
  s <- trunc(rank(x))
  a <- ifelse(m <= 10, 3/8, 1/2)
  q <- (s-a)/(m + (1-a) - a)
  return(q)
}
```

10.1 Do coefficients make numeric sense?

10.2 All statistical analyses should be followed by model checking

We us a linear model (or statistical model more generally) to infer effects or predict future outcomes. Our inference is uncertain. Given some modeling assumptions, we can quantify this uncertainty with standard errors, and from these standard errors we can compute confidence intervals and p-values. It is good practice to use a series of **diagnostic plots**, diagnostic statistics, and simulation to check how well the data approximate the fit model and model assumptions. **Model checking** is used to both check our subjective confidence in the modeled estimates and uncertainty and to provide empirical evidence for subjective decision making in the analysis workflow.

NHST blues – Researchers are often encouraged by textbooks, colleagues, or the literature to test the assumptions of a t-test or ANOVA with formal hypothesis tests of distributions such as a Shaprio-Wilks test of normality or a Levine test of homogeneity. In this strategy, an alternative to the t-test/ANOVA is used if the distribution test's p-value is less than some cut-off (such as 0.05). Common alternatives include 1) transformations of the response to either make it more normal or the variances more homogenous, 2) implementation of alternative tests such as a Mann-Whitney-Wilcoxon (MWW) test for non-normal data or a Welch t-test/ANOVA for heterogenous variances. The logic of a test of normality or homogeneity before a t-test/ANOVA isn't consistent with frequentist thinking because the failure to reject a null hypothesis does not mean the null hypothesis is true. We shouldn't conclude that a sample is "normal" or that the variances are "homogenous" because a distributional test's p-value > 0.05. But, maybe we should of the distributional pre-test as an "objective" model check? The logic of this objective decision rule suffers from several issues. **First**, the subsequent p-value of the $t$test/ANOVA test is not valid because this p-value is

the long-run frequency of a test-statistic as large or larger than the observed statistic conditional on the null – not conditional on the subset of nulls with $p > 0.05$ in the distribution test. **Second**, real data are only approximately normal; with small n, it will be hard to reject the null of a normal distribution because of low power, but, as n increses, a normality test will reject any real dataset. **Third**, and most importantly, our analysis should follow the logic of our goals. If our goal is the estimation of effects, we cannot get meaningful estimates from a non-parametric test (with a few exceptions) or a transformed response, as these methods are entirely about computing a "correct" p-value. Good alternatives to classic non-parametric tests and transformations are bootstrap estimates of confidence limits, permutation tests, and generalized linear models.

10.3   Linear model assumptions

Assumptions of a linear model concern the distribution of the "random draw" in the underlying statistical model. Again, in the random error specification of a linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{10.1}$$
$$\varepsilon \sim N(0, \sigma) \tag{10.2}$$

the random draw (the "error") is from a normal distribution with mean zero and standard deviation σ. In the random conditional response specification

$$y_i \sim N(\mu_i, \sigma) \tag{10.3}$$
$$\mathrm{E}(Y|X) = \mu \tag{10.4}$$
$$\mu_i = \beta_0 + \beta_1 x_i \tag{10.5}$$

the random draw is a value drawn from a normal distribution with mean $\mu_i = \beta_0 + \beta_1 x_i$ and variance σ^2. Any inference about the parameter β_1 (such as confidence intervals or hypothesis tests) assumes that the these distributions are IID Normal where IID is **independent and identically distributed** and Normal refers to the Normal (or Gaussian) distribution.

1. Independent means that the error for one case cannot be predicted from the error of any other case. This lack of independence creates *correlated error*. There are lots or reasons that errors might be correlated. If individuals are measured both within and among cages, or tanks, or plots, or field sites, then we'd expect the measures within the same unit (cage, tank, plot, site) to err from the model in the same direction because of environmental features shared by individuals within the unit but not by individuals in other units. Multiple measures within experimental units create "clusters" of error. Lack of independence or clustered error can be modeled using models with **random effects**. These models go by many names including linear mixed models (common in Ecology), hierarchical models, multilevel models, and random effects models. A linear mixed model is a variation of model (10.2).

tl;dr – Measures taken within the same individual over time (*repeated measures*) are correlated and are common in all areas of biology. In ecology and evolutionary studies, measures that are taken from sites that are closer together or measures taken closer in time or measures from more closely related biological species will tend to have more similar error than measures taken from sites that are further apart or from times that are further apart or from species that are less closely related. Space and time and phylogeny create **spatial, temporal, and phylogenetic autocorrelation**. Correlated error due to space or time or phylogeny can be modeled with **Generalized Least Squares** (GLS) models. A GLS model is a variation of model (10.2).

2. Identical means that the errors are "drawn" from the same distribution. Since the model is a linear model, this distribution is a Normal distribution with mean zero and variance σ^2. A consequence of

"identical" is that the error variance is **homoskedastic**, or constant, or independent of X. If the error variance differs among the X then the errors are **heteroskedastic**. Many biological processes generate data in which the error is a function of the mean. For example, measures of biological variables that grow, such as lengths of body parts or population size, have variances that "grow" with the mean. Or, measures of counts, such as the number of cells damaged by toxin, the number of eggs in a nest, or the number of mRNA transcripts per cell have variances that are a function of the mean. Both growth and count measures can sometimes be reasonably modeled using a linear model but more often, they are better modeled using a **generalized linear model** (GLM), which is an extension of a linear model. Heteroskedasitc error arising for other reasons, both biological and experimental, can be modeled with Generalized Least Squares (GLS) or with linear mixed models..

3. Normal (Gaussian) error means that 1) the response is continuous and 2) the probability of sampling an individual measuring 0.5 units below the population mean is the same as the probability of sampling an individual measuring 0.5 units above the population mean. Counts (number of cells, number of eggs, number of mRNA transcripts) and binary responses (sucessful escape or sucessful infestation of host) are not continous and often often have asymmetric probablity distributions that are skewed to the right and while sometimes both can be reasonably modeled using a linear model they are more often modeled using a **generalized linear model** (GLM), which, again, is an extension of the linear model in equation (10.2).

10.4   Diagnostic plots use the residuals from the model fit

10.4.1   Residuals

A residual of a statistical model is $y_i - \hat{y}_i$. Remember that $\hat{y}_i$ is the predicted value of Y when X has the value x_i (compaactly written as $X = x_i$). And remember that $\hat{y}_i$ is the estimate of μ_i. For linear models (but not generalized linear models), the residuals of the fit model are estimates of the ε in equation (10.2). This *is not* true for generalized linear models because GLMs are not specified using (10.2).

Alert A common misconception is that inference from a linear model assumes that the *response* (the measured Y) is IID Normal. This is wrong. Either specification of the linear model shows precisely why this conception is wrong. Model (10.2) explicitly shows that it is the error that has the normal distribution – the distribution of Y is a mix of the distribution of X and that of the error. A more general way of thinking about the assumed distribution uses the specification in model (10.5), which shows that it is the *conditional* response that is assumed to be IID normal. Remember, a conditional response (y_i) is a random draw from the infinite set of responses at a given value of X.

Let's look at the distribution of residuals versus the response for a hypothetical experiment with a single, categorical X variable (the experimental factor) with two levels ("Cn" for control and "Tr" for treatment). The true parameters are $\beta_0 = 10$ (the true mean for the control group, or μ_0), $\beta_1 = 4$ (the difference between the true mean for the treatment minus the true mean for the control, or $\mu_1 - \mu_0$), and $\sigma = 2$ (the error standard deviation).

The plot above shows a histogram of the response (A) and residuals (B). In the plot of the response, the mode (the highest bar, or bin with the most cases) includes true mean for each group. And, as expected given $\beta_1 = 4$, the modes of the two groups are 4 units apart. It should be easy to see from this plot that the response does not have a normal distribution. Instead, it is distincly bimodal. But the distribution of the response *within* each level looks like these are drawn from a normal distribution – and it should. In the plot of the residuals, the values of both groups are shifted so that the mean of each group is at zero. The consequence of the shift is that the combined set of residuals does look like it is drawn from a Normal distribution.

The two plots suggest two different approaches for model checking. First, we could examine the responses within each level of the experimental factor. Or, second, we could examine the residuals of the fit model,

model-check-residuals1-1.bb
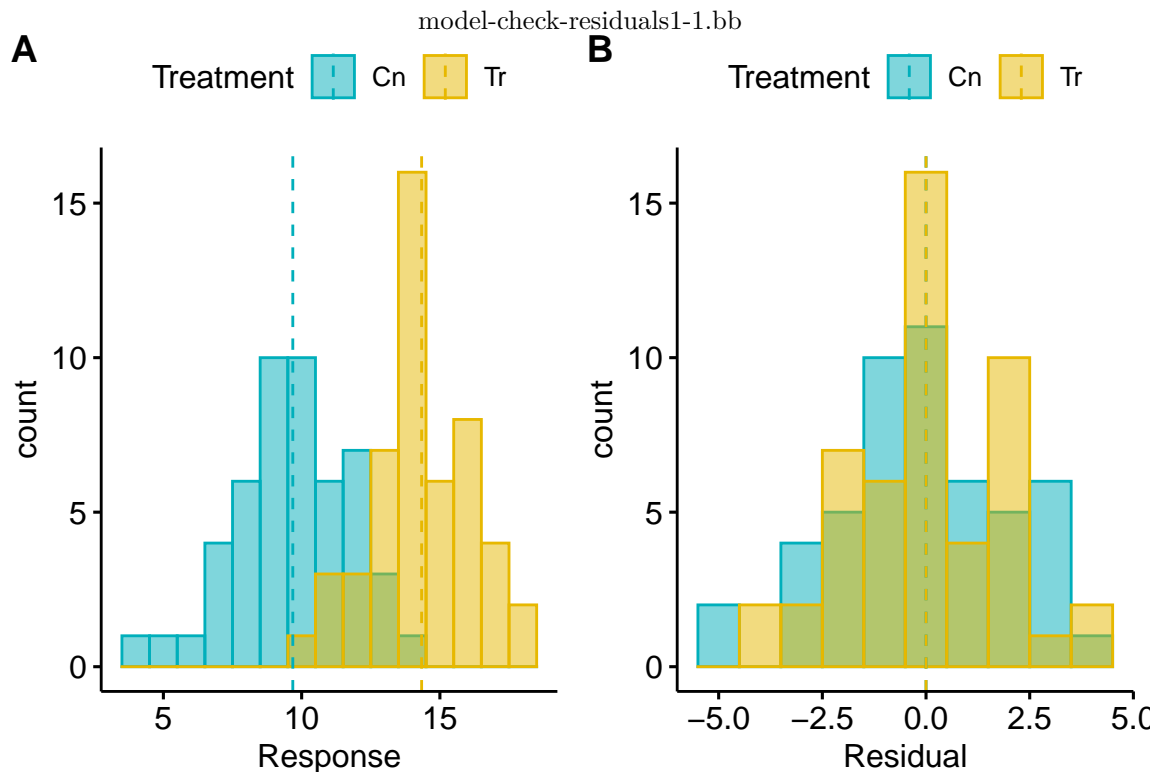
Figure 10.1: (#fig:model-check-histogram, model-check-residuals1)Histogram of the (A) response, showing with modes near the true means of each group and (B) residuals, with a mode for both groups at zero.

ignoring that the residuals come from multiple groups. The first is inefficient because it requires as many checks as there are levels in the factor. The second requires a single check.

Alert Some textbooks that recommend formal hypothesis tests of normality recommend the inefficient, multiple testing on each group separately. This isn't wrong, it's just more work than it needs to be and also suffers from "multiple testing".

10.4.2   A Normal Q-Q plot is used to check normality

A Normal Q-Q plot of the residuals can be used to check how closely the residuals approximate a normal distribution. A Normal Q-Q plot is a scatterplot of

1. **sample quantiles** on the y axis. The sample quantiles is the vector of N residuals in rank order, from smallest (most negative) to largest (most positive). Sometimes this vector is standardized (doing this makes not difference to the interpretation of the Q-Q plot).
2. **standard normal quantiles** on the x axis. This is the vector of standard normal quantiles given N elements in the vector.

Stats 101 A quantile is the value of a distribution that is greater than p percent of the values in the distribution. The 2.5% quantile of a uniform distribution from 0 to 1 is 0.025. The 2.5% quantile of a standard normal distribution is -1.96 (remember that 95% of the values in a standard normal distribution are between -1.96 and 1.96). The 50% quantile of a uniform distribution is 0.5 and the 50% quantile of a standard normal distribution is 0.0 (this is the median of the distribtion – 50% of the values are smaller and 50% of the values are larger).

Stats 201 A Q-Q plot more generally is a scatter plot of two vectors of quantiles either of which can come

from a sample or a theoretical distribution. In the GLM chapter, the text will introduce Q-Q plots of residual quantiles transformed to have an expected uniform distribution. These are plotted against theoretical uniform quantiles from 0 to 1.

If the sampled distribution closely approximates a normal distribution, the scatter should fall along a line from the bottom, left to the top, right of the plot. The interpretation of a normal Q-Q plot is enhanced with a line of "expected values" of the sample quantiles if the sample residuals are drawn from a normal distribution. The closer the sample quantiles are to the line, the more closely the residuals approximate a normal distribution. Because of sampling, the sampled values always deviate from the line, especially at the ends. If the sample was drawn from a normal distribution, these deviations should be small if the sample size is big, but can be more pronounced with a small sample size. This makes it hard to have much confidence in the "normality" of a small sample.

Let's have a look at a Normal Q-Q plot of the residuals of the fake data generated above.
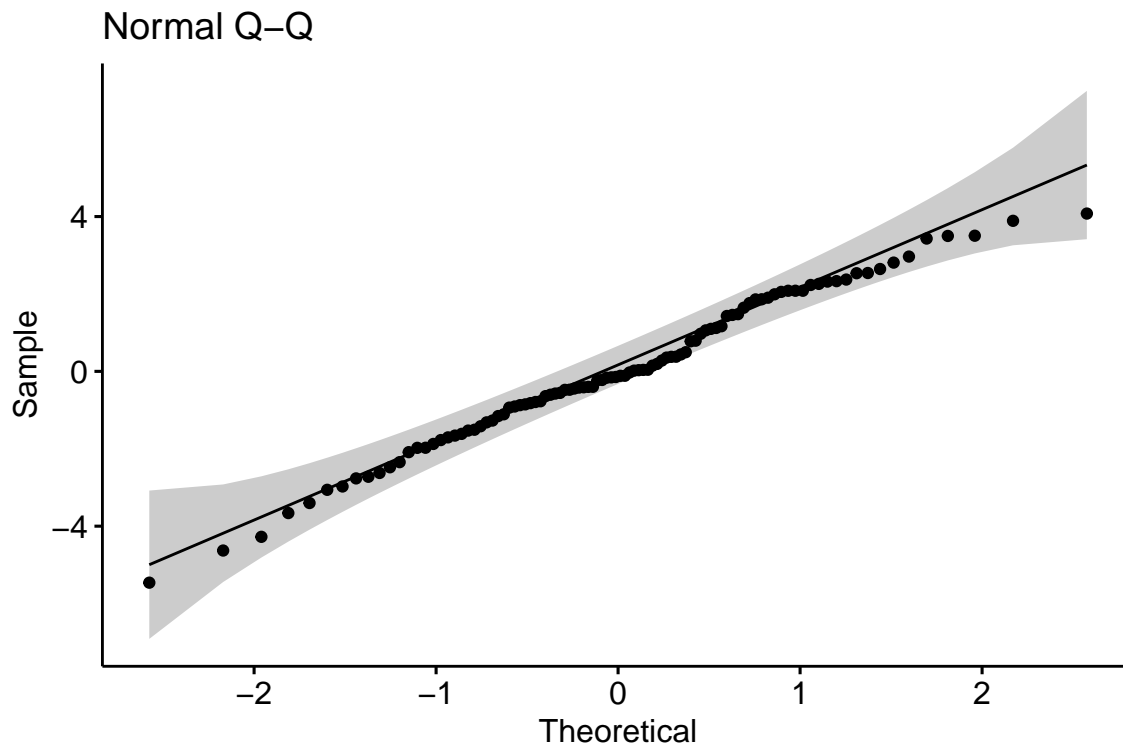
Rules of a Q-Q plot

At the small end of the distribution (bottom-left), the sample values are a bit more negative than expected, which means the left tail is a bit extended. At the large end (upper-right), the sample values are, a bit less positive than expected, which means the right tail is a bit shortened. Should we fit a different model given these deviations? To answer this, we look look at the shaded area, which represents the range of expected deviations from expectation (the line) given the sample size. Clearly the deviations are within this range.

Now let's look at simulated samples drawn from non-normal distributions to identify their characteristic deviations.

10.4.2.1 Right skewed

Many biological measures are from a distribution with long, right tails (right skewed). Examples include many count variables (number of eggs in a clutch, number of cells colonized by a parasite), and measures of time, weight, or length. What is common to all of these is unconstrained upper bounday but a constrained

lower boundary at or above zero (A nest might have zero but eggs. The weight of a fat depot must be greater than zero but the weight of a specific species of fish in a trawl catch might be zero).

A long right tail of conditional responses creates a characteristic positive deviation of the largest quantiles in a Normal Q-Q plot of the residuals from a linear model. Positive deviations at the upper end indicate larger values than expected given a normal distribution. This is the signature of the residuals of a linear model fit to right skewed data.
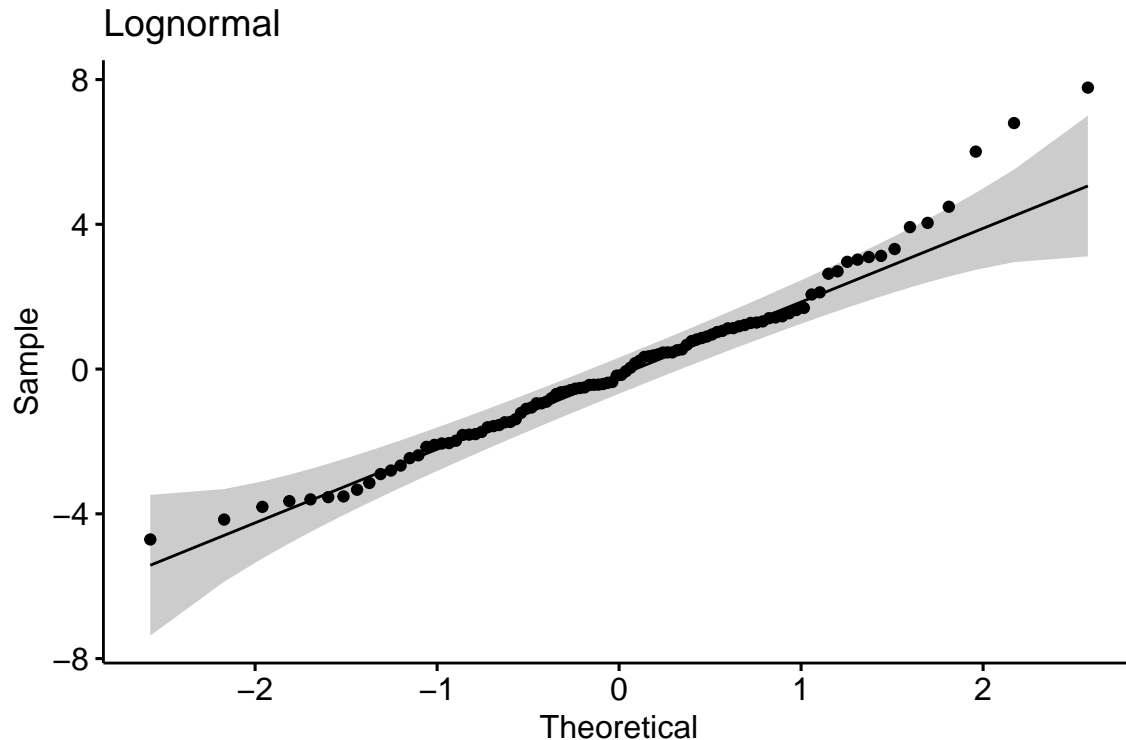
A continous response with a right skewed distribution can be modeled with a generalized linear model using a lognormal or gamma distribution. A count response can be modeled with a generalized linear model using a Poisson, quasi-Poisson, or negative binomial distribution (Chapter xxx).

10.4.2.2 Excess zeroes

Count data often have an excess of zeroes (for example a lake with no nests or a host with no parasites), resulting in positive deviations (closer to zero) at the lower end of the quantile range.

```
# zero inflated
mu_pois <- 2.7
beta_pois <- 2
p_zero <- 0.3

fd.zipois <- data.table(
  Treatment=rep(c("Cn", "Tr"), each=n),
  Response=c(SpiecEasi::rzipois(n, lambda = mu_pois, pstr0 = p_zero), SpiecEasi::rzipois(n, lambda = (mu
)
m1 <- lm(Response ~ Treatment, data=fd.zipois)
fd.zipois[, Residual:=residuals(m1)]
gg1.zipois <- gghistogram(data=fd.zipois, x = "Response",
                          color="Treatment", fill="Treatment",
                          add = "mean", rug = FALSE,
```

```
                          bins=9,
                          palette = c("#00AFBB", "#E7B800")
)

gg3.zipois <- ggqqplot(data=fd.zipois, x = "Residual", title="zero-inflated poisson")

gg3.zipois
```
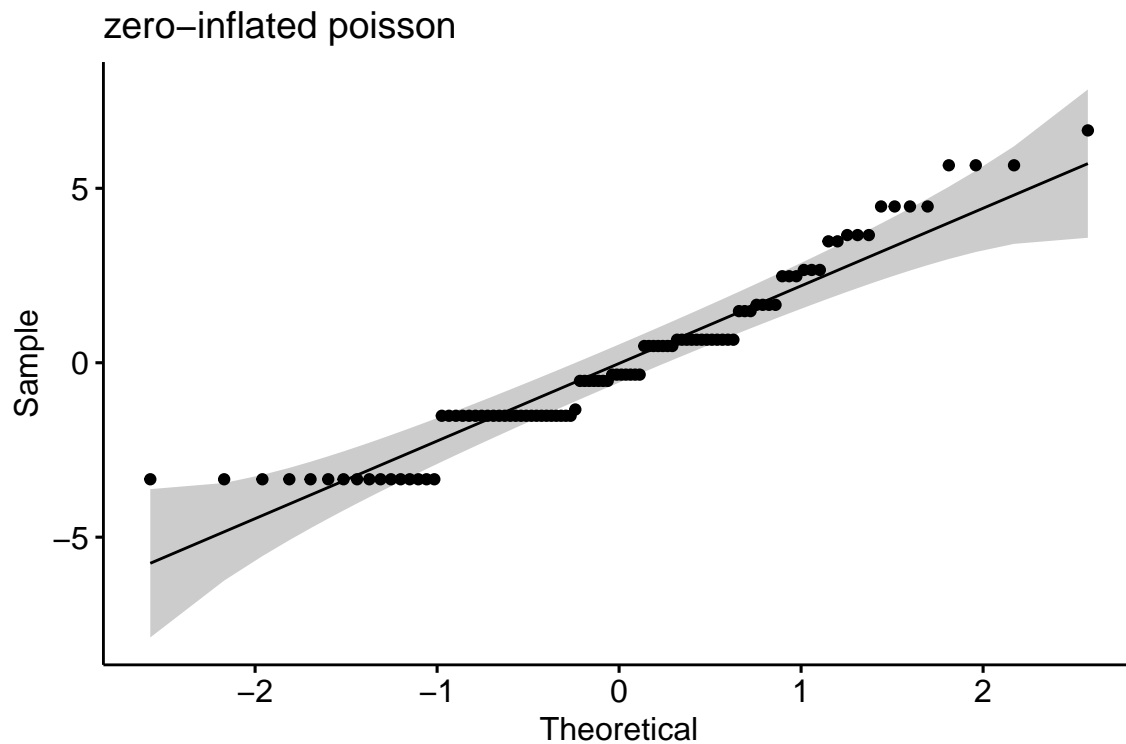

10.4.2.3   Constrained lower and upper bounds

Proportions are constrained to values between zero and one. A proportion can have a distribution that approximates a normal distribution if the mean is near 0.5 and the standard deviation is small. But, more generally, proportions can have distributions with diverse shapes.
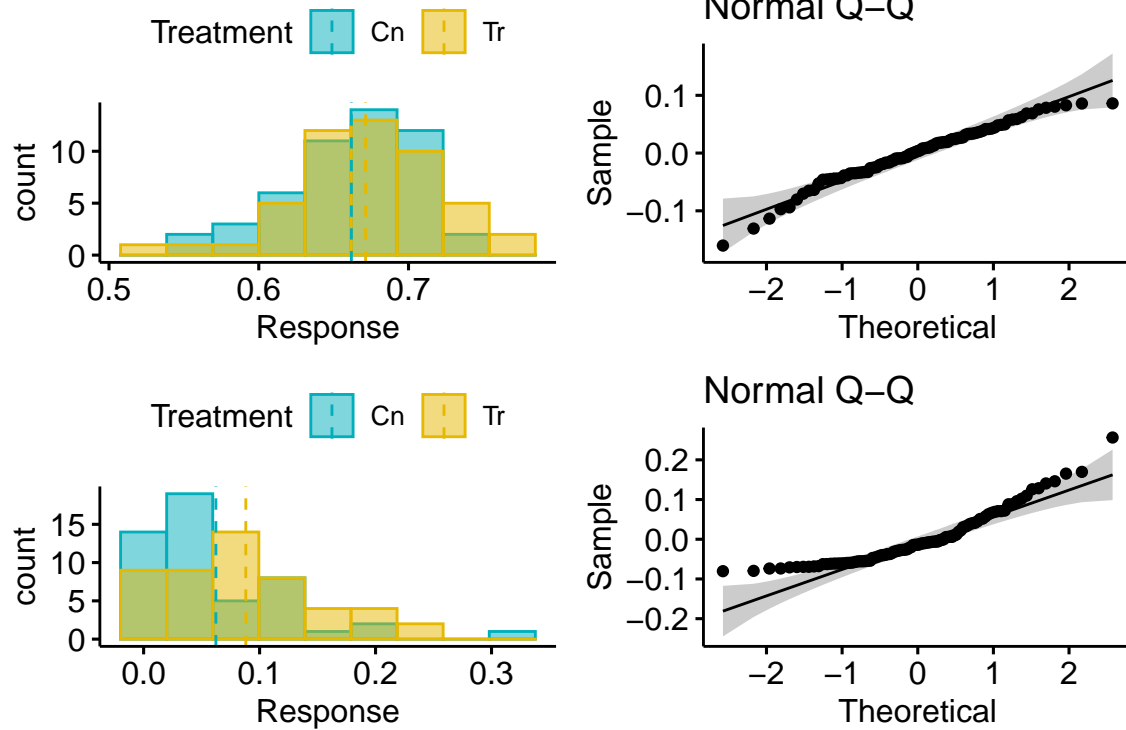
10.4.2.4   Binary responses

10.4.3   Outliers - an outlier is a point that is highly unexpected given the modeled distribution.

10.5   Model checking homoskedasticity

10.6   Model checking independence - hapiness adverse example.

10.7   Using R

Chapter 11

Model Fitting and Model Fit (OLS)

11.1 Least Squares Estimation and the Decomposition of Variance

The linear models in the last chapter and for much of this book are fit to data using a method called "ordinary least squares" (OLS). This chapter explores the meaning of OLS and related statistics, including R^2, as well as some alternative methods for bivariate regression.

11.2 OLS regression
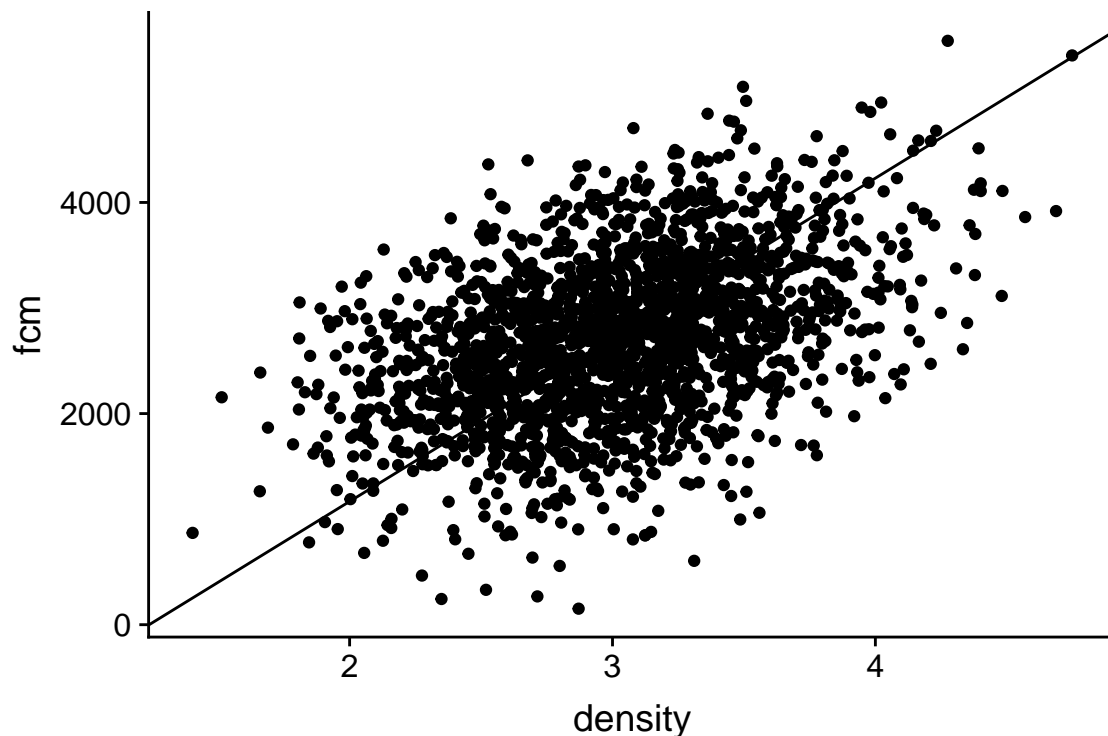
The fake data illustrated in the scatterplot above (Figure **??**) were modeled to look something like the squirrel fecal cortisol metabolite data in the previous chapter. If a typical student is asked to draw a regression line through the scatter, they typically draw a line similar to that in Figure **??**. This line is not the OLS regression line but the major axis of an elipse that encloses the scatter of points–that students invariably draw this line

suggests that the brain interprets the major axis of an elliptical scatter of points as a trend (This major axis line is an alternative method for estimating a slope and is known as standard major-axis regression. More about this at the end of this chapter.)
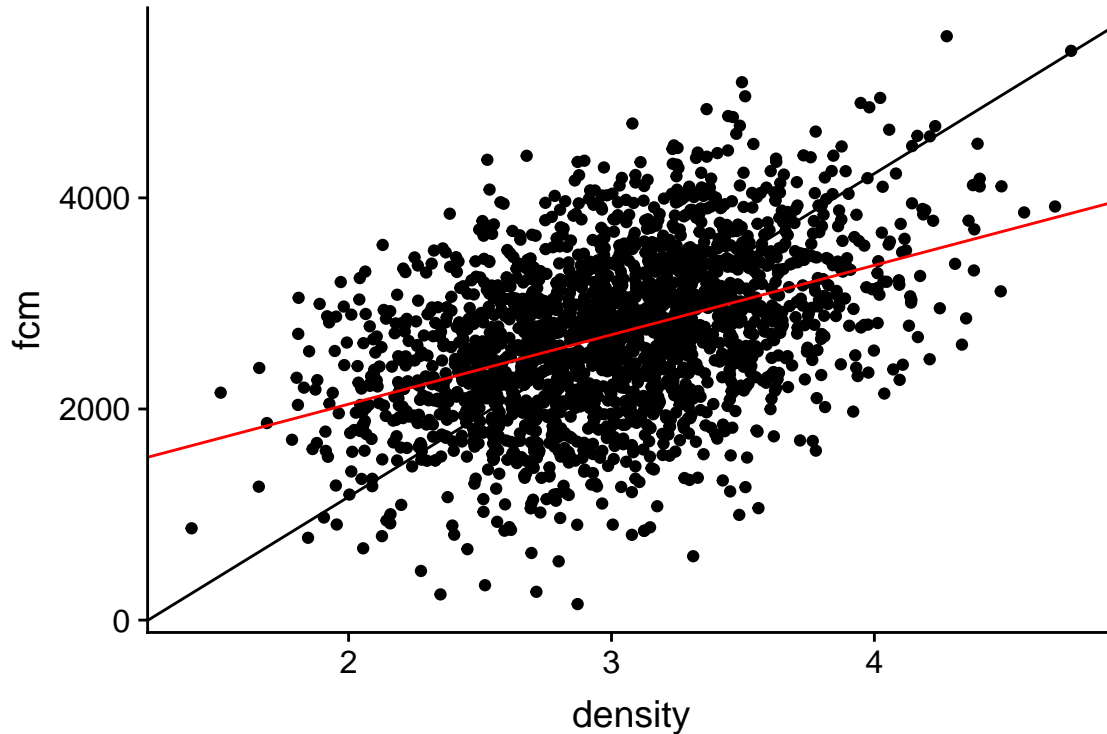
The OLS regression line is the red line in Figure **??** – the standard major axis line is left for comparison). The OLS regression line

1. passes through the bivariate mean $(\bar{x}, \bar{y})$ of the scatter, and
2. minimizes the sum of the squared deviations from each point to it's modeled value $\sum (y_i - \hat{y}_i)^2$

There are an infinite number of lines that pass through the bivariate mean (think of anchoring a line at the bivariate mean and spinning it). The OLS line is the line that minimizes the squared (vertical) deviations ("least squares").

For a bivariate regression, the slope (coefficient b_1 of X) of the OLS model fit is computed by

$$b_1 = \frac{\text{COV}(X, Y)}{\text{VAR}(X)} \tag{11.1}$$

This equation is worth memorizing. We will generalize this into a more flexible equation in a few chapters.

11.3   How well does the model fit the data? R^2 and "variance explained"

Let's switch to real data.

1. Source: Dryad Digital Repository. https://doi.org/10.5061/dryad.056r5
2. File: "Diet-shift data.xls"

Fish require arachidonic acid (ARA) and other highyly unsaturated fatty acids in their diet and embryo and yolk-stage larvae obtain these from yolk. Fuiman and Faulk (xxx) designed an experiment to investigate

if red drum (*Sciaenops ocellatus*) mothers provision the yolk with ARA from recent dietary intake or from stored sources in somatic tissues. The data below are from experiment 8. The *x*-axis is the days since a diet shift to more and less ARA (*days*) and the *y*-axis is the ARA content of the eggs (*ARA*).
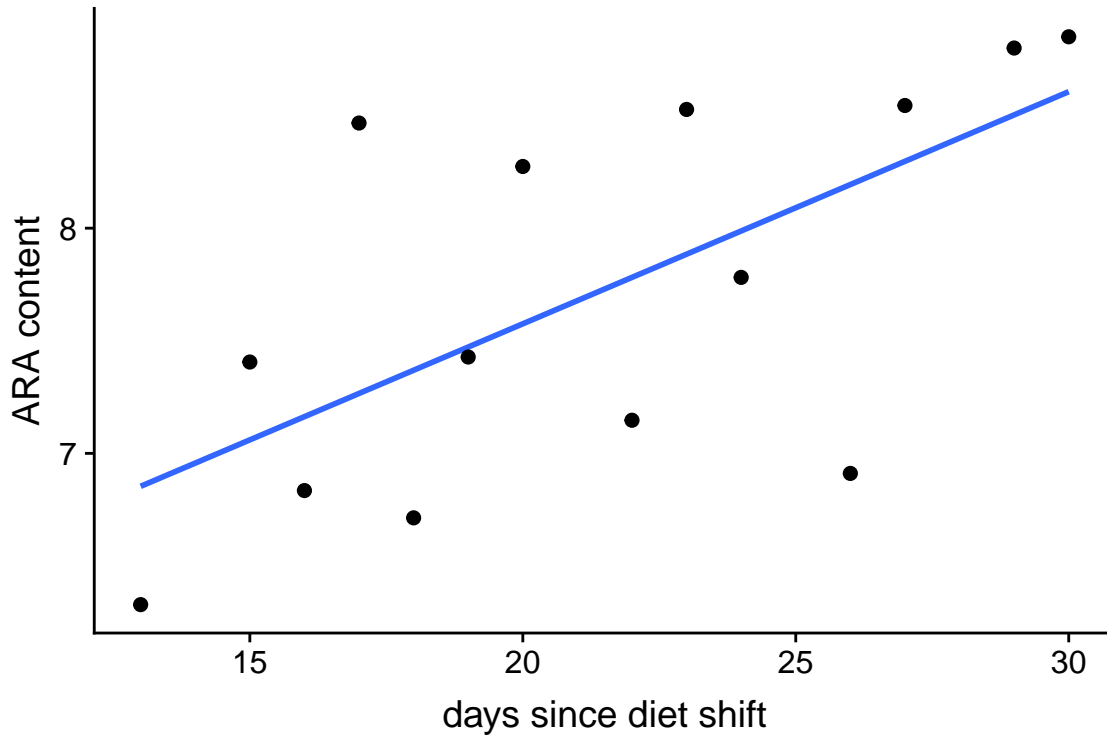
The statistic R^2 is a measure of the fit of a model to data. The R^2 for the fit of the egg data is 0.42. R^2 is the fraction of two variances $\frac{\text{VAR}(Model)}{\text{VAR}(Y)}$, or, the fraction of the variance of Y "explained by the model." The value of R^2 ranges from zero (the fit cannot be any worse) to one (the fit is "pefect").

To understand R^2, and its computation, a bit more, let's look at three kinds of deviations.

Figure 11.1A shows the deviations from the measured values to the mean value (dashed line). These are the deviations in the numerator of the equation to compute the variance of ARA_EGG_MG. Figure 11.1B shows the deviations of the measured values from the modeled values. The sum of these deviations squared is what is minimized by the OLS fit. The bigger these deviations are, the worse the model fit. Figure 11.1C shows the deviations of the modeled values to the mean value. The bigger these deviations are, the better the model fit.

The sums of the squares of these deviations (or "sums of squares") have names:

$$\text{SS(total)} = \sum (y_i - \bar{y})^2 \tag{11.2}$$

$$\text{SS(error)} = \sum (y_i - \hat{y}_i)^2 \tag{11.3}$$

$$\text{SS(model)} = \sum (\hat{y}_i - \bar{y})^2 \tag{11.4}$$

Again, SS(total) is the numerator of the equation for the sample variance. It is called "s-s-total" because SS(total) = SS(model) + SS(error). That is, the total sums of squares can be **decomposed** into two **components**: the modeled sums of squares and the error sums of squares. Given these components, it's easy to understand R^2
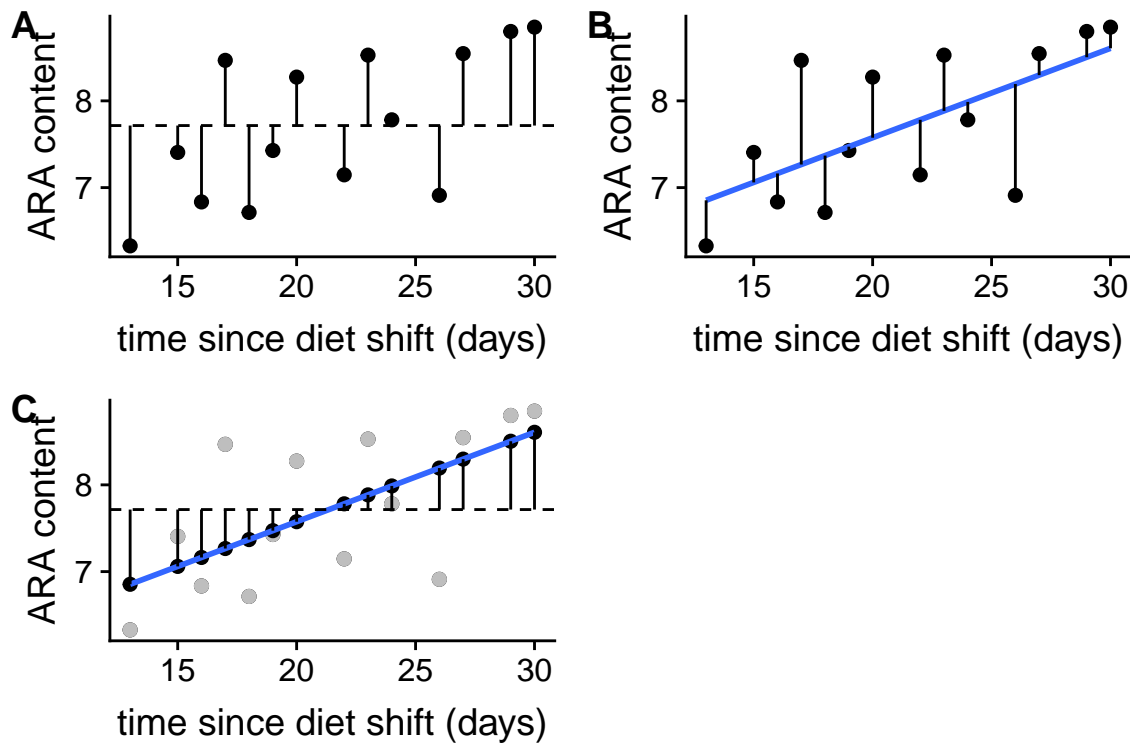
Figure 11.1: Three kinds of deviations from a fit model. A. Deviations of the measured values from the mean. These are in the numerator of the equation of the sample variance. The dashed line is the mean ARA content. B. Deviations of the measured values from the modeled values. The sum of these deviations squared is what is minimized in an OLS fit. C. Deviations of the modeled values from the mean ARA content. The measured values are in gray, the modeled values in black

$$R^2 = \frac{SS(model)}{SS(total)} \tag{11.5}$$

R^2 is the fraction of the total sums of squares that is due to (or "explained by") the model sums of squares. Above I said that R^2 is the fraction of *variance* explained by the model. Equation xxx is a ratio of variance, but the $(n-1)^{-1}$ in both the numerator and the denominator cancel out. Finally, many sources give the equation for R^2 as

$$R^2 = 1 - \frac{SS(error)}{SS(total)} \tag{11.6}$$

which is an obvious alternative given the decomposition. I prefer the former equation because it emphasizes the model fit instead of model ill-fit.

Part IV: More than one X – Multivariable Models

Chapter 12

Adding covariates to a linear model

In its most general sense, **Covariates** are simply the X variables in a statistical model. With data from experiments, "covariates" more typically refers to X variables that are added to a model to increase precision of the treatment effects. In observational designs, covariates might be added to a model to 1) increase predictive ability, 2) because the researcher is interested in specific conditional effects, or 3) to eliminate confounding. These are discussed in later chapters.

12.1 Adding covariates can increases the precision of the effect of interest

I use fake data to introduce the concept of **statistical elimination** of a **covariate** in a statistical model. Here I am modeling the effect of a new drug on blood LDL-C levels. LDL is a kind of lipoprotein, which are particles in the blood that transport fats and cholesterol to and from different tissues. LDL-C is cholesterol associated with LDL particles. LDL-C is considered "bad cholesterol" because LDL is believed to transport cholesterol and other lipids to arterial walls, which is the basis for atherosclerosis.

Twenty applied biostats students are recruited and are randomly assigned to either the new or old drug. The response is blood LDL-C level and the two treatment levels are "old" and "new". For the dummy variable coding, I'll make "old" the reference level.

Let's model this with

$$ldlc = \beta_0 + \beta_1 drug + \varepsilon \tag{12.1}$$

where $drug$ is the dummy variable with $old = 0$ and $new = 1$.

```
##                  Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 161.640772    3.974506 40.6694043 3.620049e-19
## drugnew       1.785414    5.620800  0.3176442 7.544101e-01
```

The plot shows large overlap in LDL-C and no obvious effect of the drug. There "is no effect of the drug ($p = 0.754$)" is of course an incorrect interpretation of the hypothesis test of the estimate of β_1. A correct interpretation is, there is too much noise to say much about any effect.

In addition to assigning treatment level randomly, I also had the 20 students count calories from fat over the course of the experiment. Here is a plot of LDL-C vs. percent calories from fat, with treatment assignment color coded. Remember, these are the exact same values of LDL-C as in the first figure.

The line is the bivariate regression fit to the data ignoring treatment level so is the model
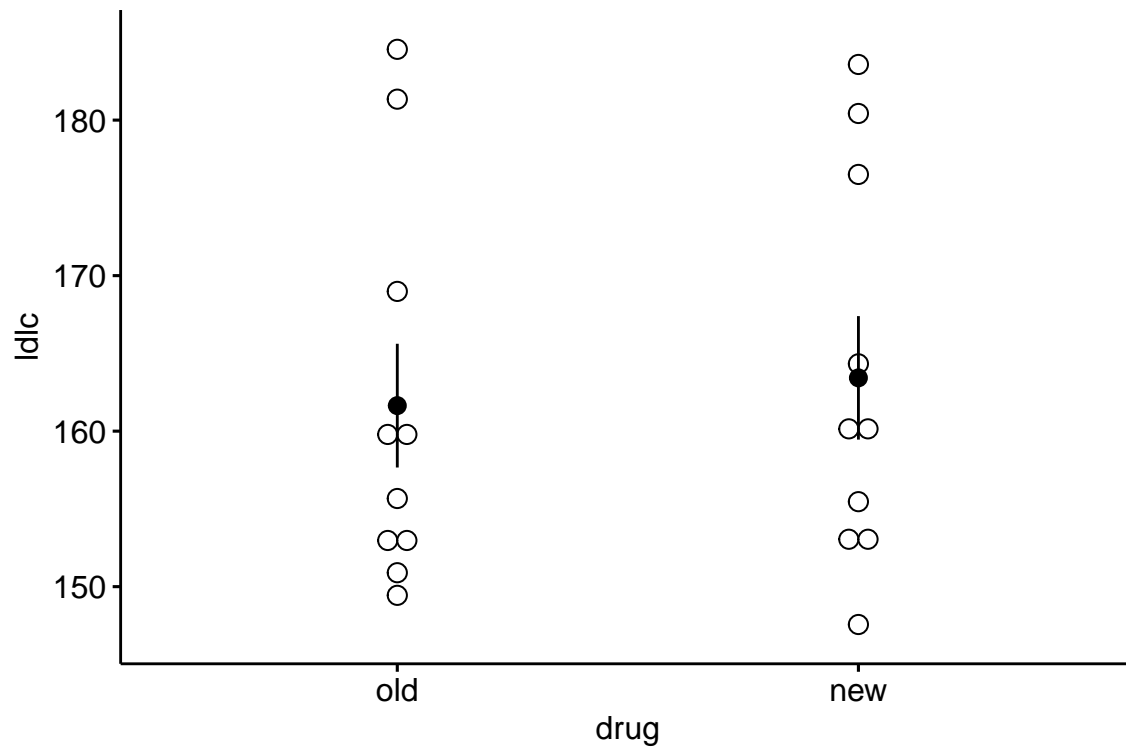
Figure 12.1: The fake LDL-C experiment.
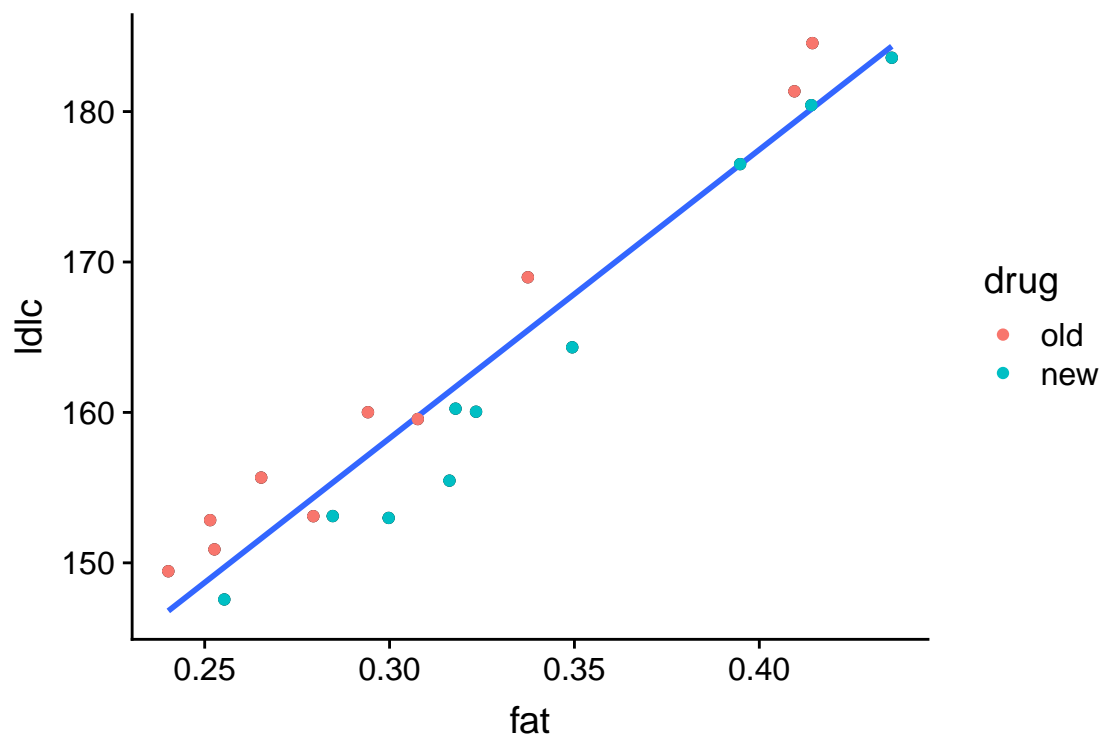
Figure 12.2: Linear regression of *ldlc* on dietary *fat* fit to the fake LDL-C data. The points are color coded by treatment.

$$ldlc = \beta_0 + \beta_1 fat + \varepsilon \tag{12.2}$$

I've color coded the points by treatment level but *drug* is not in the model. It is clear that most of the "old" data points are above the line, or have positive residuals from the model, while the "new" data points are below the line, or have negative residuals from the model. A better way to think about this pattern is that **at any specific level of fat, the LDL-C for old is higher than the LDL-C for new**.

What is happening? Dietary fat is contributing to the variance of LDL-C and this added noise makes it harder to measure the effect of the new drug relative to the old drug. If we could somehow measure the effect of drug at a specific level of dietary fat, then we could get a more precise estimate of the effect. But how to do this?

1. We could just analyze a subset of the data, that is, only the cases in which the value of dietary fat is nearly equal. This throws away perfectly good data and, consequently, greatly reduces the sample size and thus precision to estimate the effect.

2. We could use the residuals of the fitted model (12.2) to estimate the effect of drug treatment (this is what we did by eye in figure 12.2). Here is the new model

$$ldlc.r = \beta_0 + \beta_1 drug + \varepsilon \tag{12.3}$$

where *ldlc.r* are the residuals. The effect of the new drug on these residuals is

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```
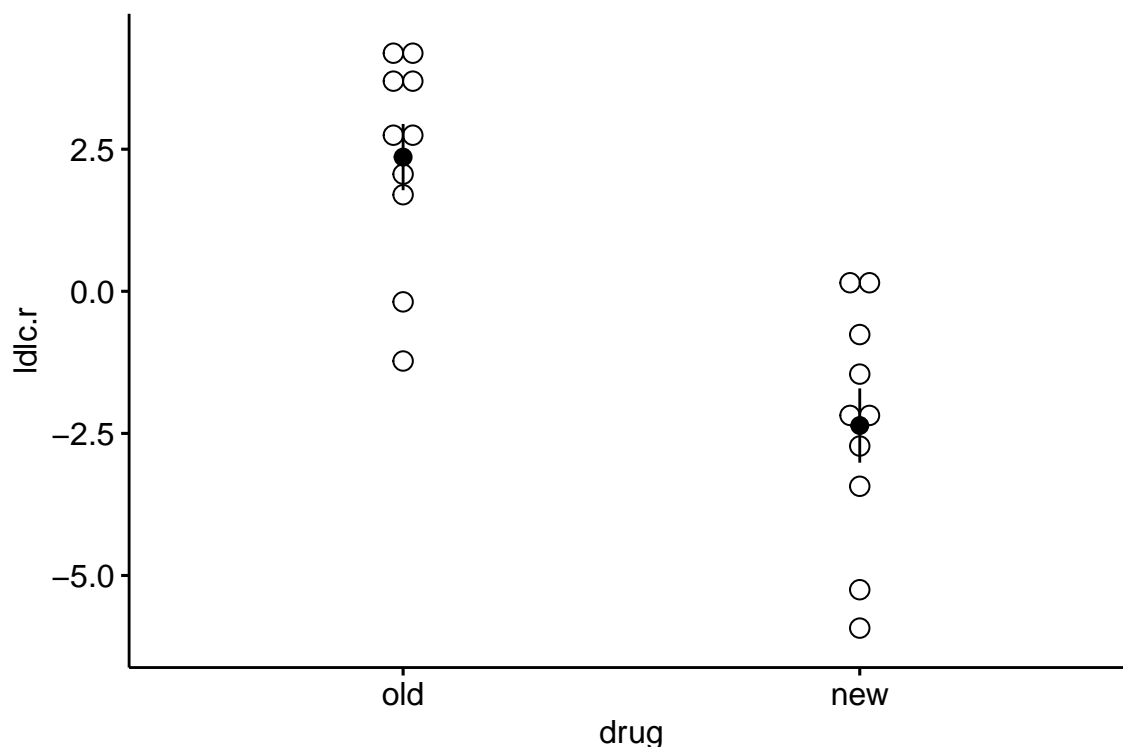


```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  2.361321  0.6190269  3.814570 1.269495e-03
## drugnew     -4.722642  0.8754362 -5.394616 3.984159e-05
```

In this two-stage analysis (stage 1: fit ldlc ~ fat to get residuals, stage 2: fit residuals ~ drug), we have *eliminated the effect of dietary fat* on the variance of the response and, as a consequence, the estimate of the effect of the drug is much more precise. Now the estimate of the effect is -4.7 mg/dL blood and the SE is

only .9 (compare this to the values in the original analysis). While the SE of the diference is correct, any confidence interval or t-value is not because the df is wrong. In the two stage analysis we fit two parameters – the slope (coefficient b_1) in stage 1 and the difference in means (coefficient b_1) in stage 2 – but the t in the table assumes we only fit one parameter (that from stage 2). Effectively, the stage 2 test is ignorant that the data ($ldlc.r$) are the result of a previous model fit. We could manually modify the computation of t, but the more proper method is to simply…

3. Add dietary fat into the original linear model.

$$ldlc = \beta_0 + \beta_1 fat + \beta_2 drug + \varepsilon \tag{12.4}$$

```
##                Estimate Std. Error  t value      Pr(>|t|)
## (Intercept)  99.449257  2.2812187 43.59479  6.880431e-19
## fat         203.773799  7.2192309 28.22652  1.009563e-15
## drugnew      -5.128034  0.8711147 -5.88675  1.794002e-05
```

Here, the estimate is -5.1 and the SE is 0.9. Look back at the script generating the fake data; the true effect (β_1) of the new drug was set to -5.0 so this estimate is quite good.

12.1.1   Interaction effects with covariates

12.1.2   Add only covariates that were measured before peaking at the data

12.2   Regression to the mean

It is common to measure the outcome variable (Y) both before and after the experimental treatments are applied and then compare the pre-post *change* in Y in response to the treatment using a t-test or ANOVA. Don't do this.

Instead, add the pre-treatment measure into the model as a covariate.

$$Y_{post} = \beta_0 + \beta_1 Y_{pre} + \beta_2 Treatment + \varepsilon \tag{12.5}$$

where $Treatment$ is a dummy variable for a two-level factor. A pre-treatment measure (Y_{pre}) is often called the *baseline* measure. The change in Y ($\Delta Y = Ypost - Y_{pre}$) is sometimes called a change score or gain score. ΔY can be modeled as in equation (12.5) and the p-value will be precisely the same (the estimate and SE will differ of course because the response variable is different).

$$\Delta Y = \beta_0 + \beta_1 Y_{pre} + \beta_2 Treatment + \varepsilon \tag{12.6}$$

The reason why a researcher should not model ΔY as a function of $Treatment$ without Y_{pre} as a covariate is that the **regression to the mean**. To explain regression to the mean, I use fake data simulated to model the results from an important study on gut microbiomes. In this study, the authors (Turnbaugh et al. xxx) showed that mice with feces from obese (genotype ob/ob) donors had higher weight gain than mice with feces from lean (genotype $+/+$) donors, presumably because of the differences in microbial communities between the donor types (shown elsewhere in their paper). To support the inference of a large difference in weight change, they illustrated the percent change in each treatment level in their Fig 3C, which is replicated here using simulated data generated to match the original summary statistics (Figure 12.3).

That looks like a big difference, with the mice from the obese-donor treatment level gaining much more fat than the mice from the lean-donor treatment level. Turnbaugh et al. used a simple t-test of this percent change to test the effect of the ob/ob treatment. The linear model underneath this t-test is
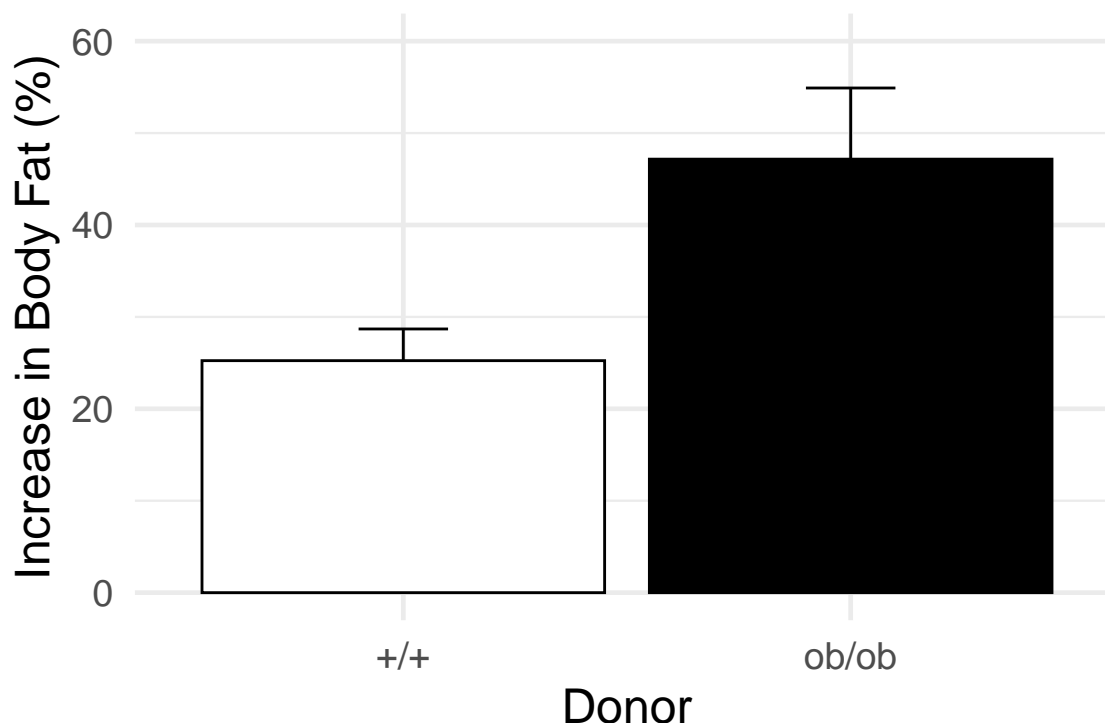
Figure 12.3: Figure 3c of Turnbaugh *et al* 2006. This figure was generated with simulated data matching the summary statistics given in Turnbaugh *et al* 2006

$$fat.gain = \beta_0 + \beta_1 obese + \varepsilon \tag{12.7}$$

where *fat.gain* is the percent gain in fat from baseline and *obese* is a dummy variable with *ob/ob* = 1. The model coefficients are

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)  25.24015   5.627515 4.485134 0.0003259533
## treatmentob/ob 21.92156   8.176589 2.681016 0.0157879742

##                 2.5 %   97.5 %
## (Intercept)   13.367137 37.11317
## treatmentob/ob  4.670468 39.17266
```

Or, the increase in fat in the obese-treated mice was 21.9% (95%CI: 4.7, 39.2%, $p = 0.016$) greater than the increase in lean-treated mice. This result, if generally verified with replication and rigorous probing, would have spectacular implications for human health.

One might reasonably expect that if mice are randomized into two groups, then the expected difference in percent change from baseline is zero. This is unconditionally true but not conditionally true. That is, if we ignore initial fat weight, the expected difference is zero. But, the expected difference is also conditional on the initial difference in fat weights. More specifically, the expected difference is opposite in sign but proportional in magnitude to the initial difference. This conditional expectation is a consequence of regression to the mean. If the first measure of a random variable is extreme, the second measure will tend to be less extreme. And, if a second measure is extreme, the first measure will tend to be less extreme.

Despite random treatment assignment, the mean initial fat weight of the *ob/ob* group was 1.2SD less than the mean initial weight of the *+/+* group. By contrast, the mean final weight of the *ob/ob* group was 0.06SD larger than the mean final weight of the *+/+* group. This first difference is an extreme measure. The second

Figure 12.4: Effect of initial difference in weight on the difference in change score. Increased initial difference in weight results in an increased differences in change score between treatment and control. Four different values of *rho* (the correlation between initial and final weights) were simulated. Only when *rho*=1 is there no influence of initial difference, because whatever differences occur at baseline will be perfectly preserved in the final measure. The X gives the values in the original Turnbaugh data

is extremely close to the expectation if there is no treatment effect. Because the initial difference in weight is unusually negative, the expected difference in percent change will be unusually positive.

This dependency between difference in percent change from baseline and difference in initial weight is easy to simulate. Simply

1. randomly sample a normal distribution as the "initial weight" and randomly assign to treatment class
2. let the final weight have some correlation (ρ) with the initial weight. Some correlation should make sense – we expect a mouse that has more fat than average at the start of the experiment to also have more fat than average at the end of the experiment. Run the experiment at different values of this correlation to see how it effects regression to the mean.
3. Do not add a treatment effect. We want to explore the behavior of the nill null hypothesis.

What's happening in Figure 12.4? Each point is a result for a single, simulated experiment. In total, there are 1000 simulated experiments for each of four values of ρ. The *x*-axis is the difference between the means of the two treatment levels at baseline (*Initial difference*). The *y*-axis is the difference in mean change score between the two treatment levels – that is the difference in the means of ΔY from equation (12.6). This difference in ΔY is the effect of the treatment the researchers are interested in. The *unconditional* expectation of this difference is zero

$$\mathrm{E}(\Delta Y_{ob/ob} - \Delta Y_{+/+}) = 0 \qquad (12.8)$$

but the change conditional on baseline is not zero

$$\mathrm{E}(\Delta Y_{ob/ob} - \Delta Y_{+/+}) \neq 0 \tag{12.9}$$

Instead, the conditional expectation is a function of the difference at baseline. If the initial difference in weight happens to be unusually large and negative, the expected difference in change score is unusually positive. This non-zero expectation means that the estimate of the treatment effect is **conditionally biased** for any model that does not include the baseline fat weight as a covariate. And, from a frequentist perspective, the Type I error for a test of a difference in ΔY is strongly dependent on the initial difference in weight.

The big X in the plot indicates the difference at baseline and difference in ΔY for the original mice study. The difference in $DeltaY$ is unusually positive (about .6% of the $|\delta Y|$ are larger) but very close to the expected value given the unusually large, negative difference at baseline. In other words, the probability of the data, or more extreme than the data, is not 0.006 but something larger and perhaps, much larger (the computed value depends on the observed ρ. From, the plot, the X is very unusual if $\rho = 1$, pretty unusual if $\rho = 0.66$, but pretty common if $\rho = 0.33$ or if $\rho = 0$).

12.2.1 Do not use percent change, believing that percents account for effects of initial weights

Some researchers mistakenly believe that a t-test of percent change automatically adjusts for effects in initial weight, since this initial weight is in the denominator of the percent. This is wrong. The dependency of the difference in change between treatments on the initial difference between treatments is more severe if change is measured as a percent, because the numerator (the change score) is expected to be larger if the denominator is smaller (initial measure). Using the simulated data from above, here is this dependency.

12.2.2 Do not "test for balance" of baseline measures

Contrary to some advice and maybe to intuition, it makes no sense to "test for balance" at baseline with a t-test of the difference in initial measures of Y. And, it makes no sense to use this test as a decision rule for how to proceed: if $p > 0.05$ then use a simple t test of the change scores, if $p < 0.05$ then use ANCOVA with baseline measures in the model. First, a null-hypothesis significance test cannot tell you that there is "no difference" – this is not what null-hypothesis tests do. Second, any p-value after the initial test isn't strictly valid as it does not take into account this decision step, but this is minor. Third, it doesn't matter; there will always be some difference in the actual means of the initial measures and, consequently, the conditional expectation of the final measures, or change in measures, or percent change will be dependent on this initial difference. So, if one has initial measures, one should use an linear model that adjusts for baseline measures to estimate the treatment effect in pre-post designs. And, if one isn't planning on taking an initial measure, then maybe you should, because the initial measure used in a linear model allows a better estimate of the treatment effect!
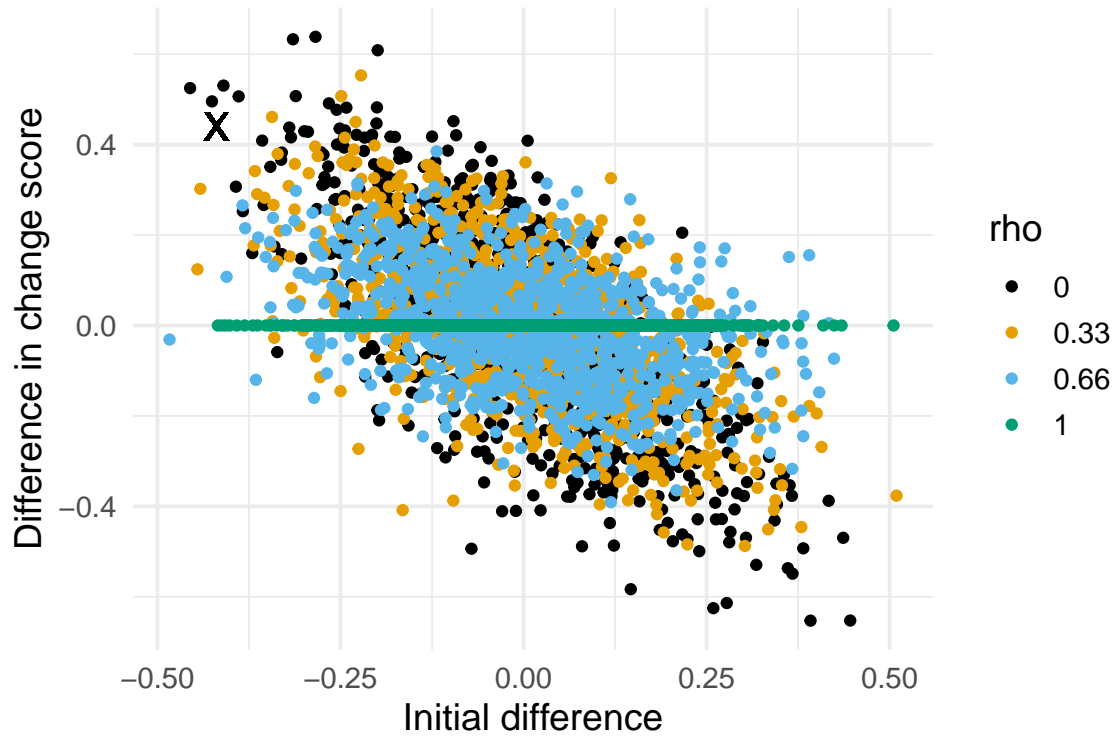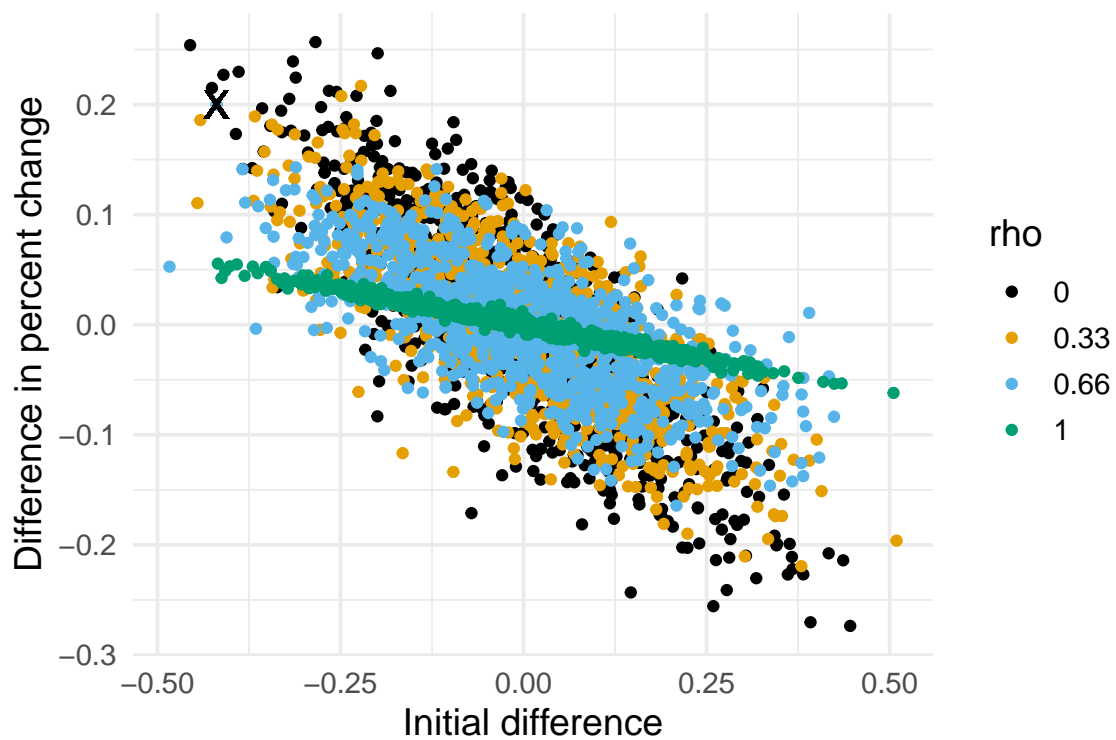
Figure 12.5: Effect of initial difference in weight on the difference in percent change. Increased initial difference in weight results in an increased differences in Percent change between treatment and control. Four different values of *rho* (the correlation between initial and final weights) were simulated. Note there is no value of *rho* where the difference in percent change is independent of the initial difference. The X gives the values in the original Turnbaugh data.

Chapter 13

Two (or more) Categorical X – Factorial designs

13.1 Factorial experiments

A factorial experiment is one in which there are two or more categorical X that are **crossed**, so the groups contain different combinations of levels of each factor. Factorial experiments are used to estimate the **interaction** between factors. Let's explore this using data from xxx

The urchin data are from an experiment measuring the effect of $Temp$ and $CO2$ on larval sea urchin metabolic rate ($Resp$) (there are other outcome measures in the study too). The units of metabolic rate are pmol O2/hr/larva. There are two $Temp$ levels (13C and 18C) and two $CO2$ levels (400 µAtm and 1100) and the factors are crossed, so this is a 2 × 2 (crossed or factorial) design. There are $n = 6$ replicates for each combination of the levels. A good way to visualize the treatment combinations in a crossed design is with a $m \times p$ table showing all combinations of the m levels of factor 1 ($Temp$) against the p levels of factor 2 ($CO2$)

The upper left cell contains the 13C level with the Temperature factor and the 400 µAtm level within the CO2 factor. The replicates in this cell were grown with no added treatments, so this cell is the "control" for Temp and the control for CO2, which we will use as the "reference" group for the linear model. The replicates in the lower left cell were grown with an added temperature treatment (in this case, a 5 C higher temperature). The replicates in the upper right cell were grown an added CO2 treatment (a 700 µATM higher CO2). And finally, the replicates in the bottom right cell were grown with both the added temperature (+5 C) and added CO2 (+700 µATM). Here, I use a "+" or "-" to designate the addition (or not) of the treatment, so our 2 × 2 treatment levels are Temp-/CO2-, Temp+/CO2-, Temp-/CO2+ and Temp+/CO2+.

Figure 13.1: The urchin experimental design as a 2 x 2 table

13.1.1 Model coefficients: an interaction effect is what is leftover after adding the treatment effects to the control

A factorial design allows a researcher to estimate the interaction between two factors. To clarify this, let's fit the factorial model and look at the coefficient table. The fit model is

$$Resp_i = b_0 + b_1 Temp^+ + b_2 CO2^+ + b_3 Temp^+ CO2^+ + e_i \qquad (13.1)$$

% Again, $Temp^+$ and $CO2^+$ are dummy variables. The model also includes $Temp^+ CO2^+$, which is a dummy variable for the interaction between Temp and CO2. The value of this interaction dummy variable is literally the product of the two main factor dummy variables ($Temp^+$ and $CO2^+$), which can be verified with the model matrix (which here, is computed from the subset of the data that includes only the first two rows of each treatment combination)

(Intercept)	Temp+	CO2+	Temp+:CO2+
1	0	0	0
1	0	0	0
1	1	0	0
1	1	0	0
1	0	1	0
1	0	1	0
1	1	1	1
1	1	1	1

The coefficient table is

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.23	0.73	11.3	0.000
Temp+	4.51	1.03	4.4	0.000
CO2+	-0.32	1.03	-0.3	0.761
Temp+:CO2+	-2.68	1.45	-1.9	0.079

The Intercept (b_0) is the mean (8.23) of the reference (Temp-/CO2-) group, and so the mean of the upper left cell in Table 1). The Temp+ coefficient (b_1) is the estimate of the added temperature effect relative to the reference, and so is the mean of the lower left cell minus the mean of the upper left cell ($b_1 = \bar{Y}_{Temp+} - \bar{Y}_{-/-}$). The CO2+ coefficient (b_2) is the estimate of the added CO2 effect relative to the reference, and so is the mean of the upper right cell minus the mean of the upper left cell ($b_2 = \bar{Y}_{CO2+} - \bar{Y}_{-/-}$). The Temp+:CO2+ coefficient (b_3) is the estimate of the **interaction effect**, which is the effect in addition to the $Temp^+$ and $CO2^+$ effects. If you added b_1 and b_2 to b_0, you would get the mean of the $Temp^+/CO2^+$ group *if the effects were purely additive*. So the interaction effect is the difference between the mean of the bottom right cell and the sum of the coefficients of the other three cells ($b_3 = \bar{Y}_{Temp+CO2+} - (b0 + b1 + b2)$). An interaction is a **non-additive effect**. Think about this. Adding 5 C increases respiration by 4.51 units. Adding 700 μATM CO2 decreases respiration by .32 units. If these effects were purely additive, then adding both 5 C and 700 μATM should result in a mean of 8.23 + 4.51 - .32 = 12.42 units for the $Temp^+/CO2^+$ group. What is the mean of this group?

9.74! So the difference between the "additive expectation" and the actual mean is $9.74 - 12.42 = -2.68$, which is the interaction effect (coefficient).

13.1.2 What is the biological meaning of an interaction effect?

I can dead lift 150 pounds and my friend Jake can deadlift 175 pounds. Working together, we should be able to lift 325 pounds. What if together, we could actually lift 400 pounds? If this were the case, this would be an interaction with an effect equal to 75 pounds. Is this biologically plausible? If so, what is the mechanism? Here is a possible mechanism (although I am highly skeptical of it having a magnitude of 75 pounds): when

lifting an object as part of a group, the central nervous system allows increased motor unit recruitment, and so each person can lift more weight than they could if lifting alone. A positive interaction like this is called *synergistic*. Always think about the biological meaning of an interaction effect.

13.1.3 What about models with more than two factors?

A factorial model can have more than two factors, for example, a model with three factors (A, B, and C), each with two levels (which I'll designate with a "+"), is

$$Y = \beta_0 + \beta_1 A^+ + \beta_1 B^+ + \beta_3 C^+ + \beta_4 A^+ B^+ + \beta_5 A^+ C^+ + \beta_6 B^+ C^+ + \beta_7 A^+ B^+ C^+ + \varepsilon \qquad (13.2)$$

It is easy enough to get an ANOVA table with p-values for this model but I don't recommend it because

1. If space and/or time and/or materials are limited then it typically makes more sense to prioritize the power to estimate standard errors by choosing one of the two-factor models and increasing sample size
2. Interaction effects in 2-factor models are hard enough to interpret. A 3-way interaction is very, very tough to interpret. If all we did was table up F-ratios and p-values, this wouldn't matter. But it does matter.

13.1.4 The additive model

Sometimes researchers fit reduced models. For example, the p-value for the interaction effect in the full model is 0.079, which is not statististically significant. This would lead some researchers to explicitly conclude (mistakenly), that there is no interaction between CO2 and Temp and that the effects of CO2 and Temp on metabolism are **additive** (see below). This would lead others (controversially) to refit the model excluding the interaction term:

$$Resp_i = b_0 + b_1 Temp^+ + b_2 CO2^+ + e_i \qquad (13.3)$$

% This particular reduced model is often referred to as the **additive model**, since it excludes the the interaction term, which is a *product* of other terms. Indeed, the title of the urchin paper is "Temperature and CO2 additively regulate physiology, morphology and genomic responses of larval sea urchins, *Strongylocentrotus purpuratus*." It is important to recognize that unless the interaction effect is precisely zero, the coefficients of Temp+ and CO2+ will change. Here are the coefficients of the additive model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.90	0.66	13.4	0.000
Temp+	3.17	0.77	4.1	0.000
CO2+	-1.66	0.77	-2.2	0.042

For several reasons, one should not drop a term because $p > 0.05$.

1. The p-value is an arbitrary dichotomization of a continuous variable. Would it make sense to behave differently if $p = 0.051$ vs. $p = 0.049$, given that these two p-values are effectively identical?
2. A p-value is not evidence that an effect is zero, or "doesn't exist", or event that an effect is "trivially small". This is because p-values are a function of measurement error, sampling error, and sample size, in addition to effect size.

13.1.5 Contrasts – simple vs. main effects

In a $m \times p$ design, there are mp cell means and, therefore, $mp(mp-1)/2$ pairwise contrasts. For the urchin data, these are

1. Temp+/CO2- - Temp-/CO2- (bottom left - top left)
2. Temp+/CO2+ - Temp-/CO2+ (bottom right - top right)
3. Temp-/CO2+ - Temp-/CO2- (top right - top left)
4. Temp+/CO2+ - Temp+/CO2- (bottom right - bottom left)
5. Temp-/CO2+ - Temp+/CO2- (top right - bottom left)
6. Temp+/CO2+ - Temp-/CO2- (bottom right - top left)

(1) and (2) are the **simple effects** of Temp at each level of CO2. (3) and (4) are the simple effects of CO2 at each level of Temp. (5) is the response of the CO2 treatment at the reference level of Temp minus the response of the Temp treatment at the reference level of CO2. This probably is not a very interesting contrast. (6) is the combined response of Temp and CO2, including the interaction, minus the control. This also is unlikely to be of much interest.

contrast	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
Temp+,CO2- - Temp-,CO2-	4.51	1.03	20	2.4	6.7	4.4	0.000
Temp-,CO2+ - Temp-,CO2-	-0.32	1.03	20	-2.5	1.8	-0.3	0.761
Temp-,CO2+ - Temp+,CO2-	-4.83	1.03	20	-7.0	-2.7	-4.7	0.000
Temp+,CO2+ - Temp-,CO2-	1.51	1.03	20	-0.6	3.6	1.5	0.157
Temp+,CO2+ - Temp+,CO2-	-3.00	1.03	20	-5.1	-0.9	-2.9	0.008
Temp+,CO2+ - Temp-,CO2+	1.82	1.03	20	-0.3	4.0	1.8	0.091

Main effects are contrasts between

13.2 Reporting results

13.2.1 Text results

The effect of the increased temperature at the control CO2 level was 4.5 (95% CI: 2.4, 6.7; $p < 0.001$) pmol O2/hr/larva. The effect of increased CO2 at the control temperature was -0.3 (95% CI: -2.4, 1.8; $p = .76$) pmol O2/hr/larva. The interaction effect was -2.7 (95% CI: -5.7, 0.3; $p = 0.079$) pmol O2/hr/larva. Because of the relatively large interaction, the effect of temperature at the high level of CO2 was less than half the effect at the low level of CO2 (estimate: 1.82; 95% CI: -0.3, 4.0; $p = 0.091$) and the effect of CO2 at the high level of Temp was 10 times greater than that at the low level of Temp (estimate: -3.0; 95% CI: -5.1, -.9; $p = 0.0084$).

The CI on the interaction includes both large negative values and trivially small values, including zero, and, consequently, our data is compatible with both scientific models (that is, we can neither support nor reject the predictions of the scientific model using these results).

13.2.2 Harrellplot

If the focus is on the interaction, then a Harrell Plot of coefficients is a good choice for communicating this.
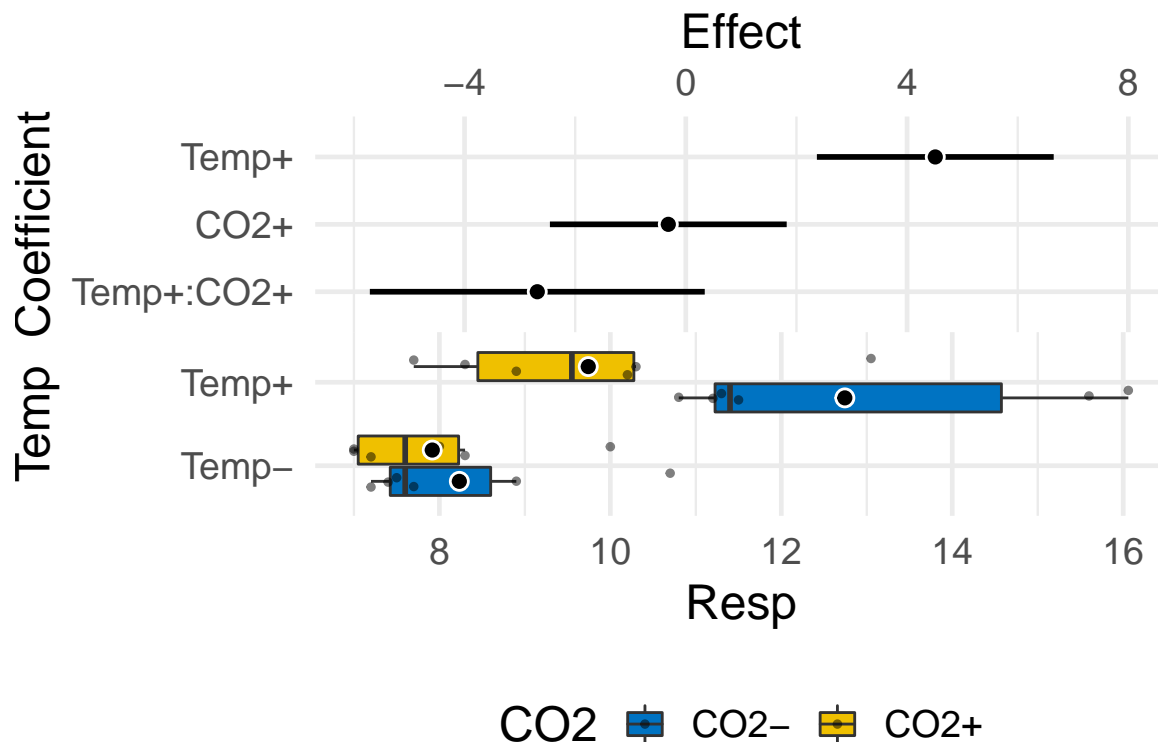
The plot shows that a relatively big and positive Temp effect, a relatively small CO2 effect (and one with too much noise to have any confidence of the sign) and a intermediate interaction effect. In the original paper, the researchers were testing a scientific (not statistical!) model that predicted no interaction, and the researchers argued that these data support this model because of the small p-value for this effect (see the table above). Well, the data do support that model (one end of the 95% CI for the interaction includes zero, and other very small values) but they also support a model of a large interaction (the other end of the 95% CI includes large negative effects). The data are too course (or the signal:noise ratio is to small) to have much confidence in the size of the interaction effect. Some comments:

1. It is not valid to claim that an effect "doesn't" exist because p > 0.05 (see chapter xxx)
2. Researchers need to draw conclusions based on *both* ends of the CI, as well as the mean!
3. A sample size of 6 will generally be very low power to get a precise estimate of the interaction unless the sampling variability is very small.

If we are more interested in differences between the different treatment combinations, then a Harrell plot with the contrasts is a good way to communicate this.
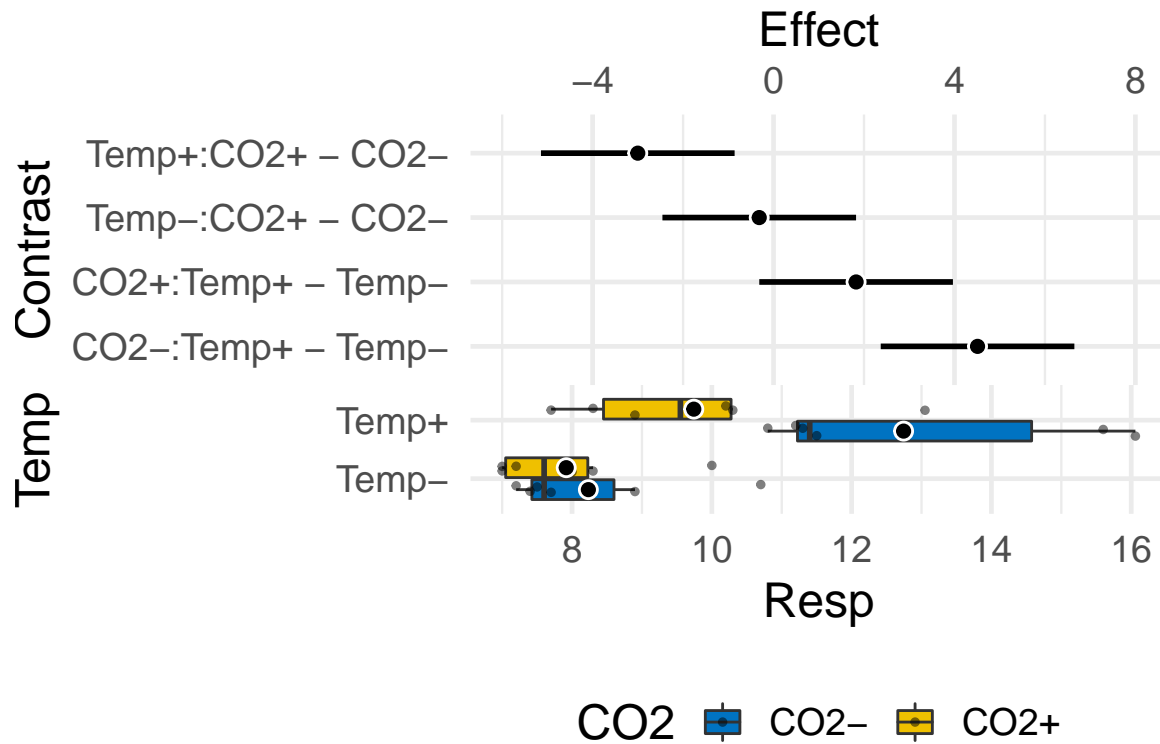
The four contrasts are differences between the means of pairs of cells in the 2×2 table:

1. "Temp+: CO2+ - CO2-" – lower right - lower left
2. "Temp-: CO2+ - CO2-" – upper right - upper left
3. "CO2+: Temp+ - Temp-" – lower right - upper right
4. "CO2-: Temp+ - Temp-" – lower left - upper left

13.2.3 Interaction plots

A common way to plot the results of the model is with an **interaction plot**.
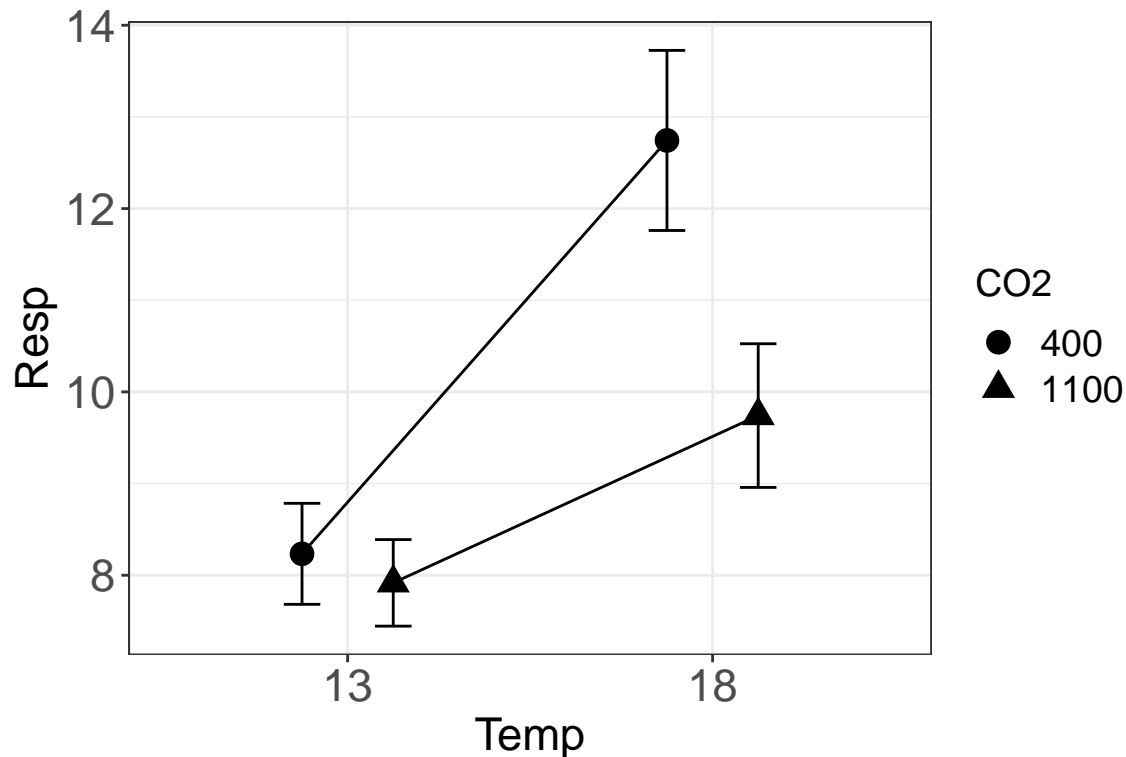
In this interaction plot, the X-axis contains the two $Temp$ treatment levels and the Y-axis is the outcome ($Resp$). The plot shows the four cell means indicated by the circles (low CO2 levels) or triangles (high CO2 levels). The solid lines connect the cell means *across Temp levels within CO2 levels*.

1. The slope of a line is the effect of $Temp$ on $Resp$
2. The relative *elevation* of the lines is the effect of $CO2$ on $Resp$
3. The difference in slope *or* the relative elevation at each level of $Temp$ is the interaction effect

Let's deconstruct this. The top (CO2-) line is the effect of $Temp$ at the control (400 μATM) value of $CO2$. The slope of the bottom (CO2+) line is the effect of $Temp$ at the high (1100 μATM) value of $CO2$. *These lines have different slopes*, or the slope *is conditional on* the level of CO2. This means that the effect of $Temp$ on respiration is *conditional on the value of CO2*. Think about this. This is what an interaction implies–conditional effects.

At the reference temperature (13 C), the CO2+ line is barely below the CO2- line. But at the high temperature (18 C), the CO2+ line is far below the CO2- line. That is, the relative elevation (the $CO2$ effect) is conditional on the level of $Temp$. It will always be the case that if the effect of Factor A is conditional on the levels of Factor B, then the effect of Factor B will be conditional on the levels of Factor A.

An interaction plot is an okay plot. It doesn't show the data, only a descriptive summary (means and standard errors). If we are interested in the interaction effect, it doesn't give us a very good sense of the error in this effect. And *that* is a problem because with real data, two lines are never precisely parallel. Our interpretation of the similarity of the slopes would probably mostly reflect our pre-conceived scientific model.

13.3   Recommendations

1. Should an interaction term be dropped from a model? In biological data, there will almost certainly be an interaction effect with everything. The goal, then, should be to model the effect and use the estimate and the uncertainty in the effect, *in combination with the goals of the analysis*, to make any decision about dropping a term.

13.4 Working in R

A **full-factorial** model with two factors is specified in the model formula as `y ~ A*B` where A is the first factor, and B is the second factor. R expands this formula to `y ~ A + B + A:B` where the colon indicates an interaction (multiplicative) effect.

```r
# re-label levels and convert to factors.
urchin1 <- copy(urchin)
urchin1[, CO2:=factor(ifelse(CO2==400,"-", "+"))]
urchin1[, Temp:=factor(ifelse(Temp==13, "-", "+"))]
fit <- lm(Resp ~ Temp*CO2, data=urchin1) # use urchin1 data with relabeled levels
coef(summary(fit))
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  8.2333333  0.7255577 11.3475922 3.626935e-10
## Temp+        4.5100000  1.0260936  4.3953106 2.792573e-04
## CO2+        -0.3166667  1.0260936 -0.3086138 7.608069e-01
## Temp+:CO2+  -2.6850000  1.4511155 -1.8503007 7.910035e-02
```

The first thing this script does is to re-label the levels of Temp and CO2. The relabeling is more aesthically pleasing in the tabled output. If we didn't relable, it would be necessary to convert CO2 and Temp to factors or else R would treat these as numeric and not factors.

The additive model is specified by the formula `y ~ A + B`

```r
fit <- lm(Resp ~ Temp + CO2, data=urchin1) # use urchin1 data with relabeled levels
coef(summary(fit))
```

```
##              Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  8.904583  0.6636207 13.418183 9.038657e-12
## Temp+        3.167500  0.7662831  4.133590 4.721000e-04
## CO2+        -1.659167  0.7662831 -2.165214 4.203445e-02
```

13.5 Problems

1. Draw four 2×2 tables and label the row and column headers using the levels of the urchin treatment. In the first table, insert the cell means. In the 2nd table, insert the equation for the coefficient. In the third table, solve the equations. And in the fourth column, insert the estimates from the table above. Are tables 3 and 4 the same? If not, you've goofed somewhere.

2. Frew et al. (2017) showed that increasing atmospheric CO2 increases grub activity in the soil which in turn increases root damage to sugarcane. They used a 2 x 2 experiment to then show that silicon added to the soild decreased the damage to the roots by the grubs (silicon minerals are very hard and plants uptake silicon from the soil to mineralize tissues to protect against insect damage). There are lots of analyses in the paper, but I want you to reproduce Fig. 4b, but using gg_interaction_plot and fit the model using lm.

(The treatment assignments are in a different file than the experimental results. Use the `merge` function to glue the two tables together, keying on the common column "plant")

1. **file name**: "canegrub_feedingtrial.csv"
2. **file name**: "treatments.csv"
3. **source**: https://datadryad.org/resource/doi:10.5061/dryad.r3s16

3. Kardol et al investigated the effect of moss growth in response to rainfall and community structure. Analyze the effect of these two factors on biomass gain and generate a Harrell plot alternative to their bar plot in Fig. 3 (see below). What is striking about your plot compared to theirs?

Figure 4. (a) Effects of silicon treatments on the relative growth rate [=mass gained (g)/time(days)] and (b)

relative consumption of roots [=food ingested (mg change in dry mass)/ initial body mass (mg fresh mass)]

under aCO_2 and eCO_2. Levels of significance are shown for effects of silicon and CO_2 treatments. Values are

means ± SE. Degrees of significance are indicated as follows: ns = not significant, . P<0.1 , * P<0.05, ** P<0.01,

*** P<0.001.

Figure 13.2: Frew et al. Figure 4b

Filename "Data file for Dryad.xlsx" **sheet** "Data" **Source*:** https://datadryad.org/resource/doi:10.5061/dryad.66d5f

4. (Grad students only) Generate a fake experiment! The experiment should have two factors each with two levels. Experiment with power by varying sample size and effect size.
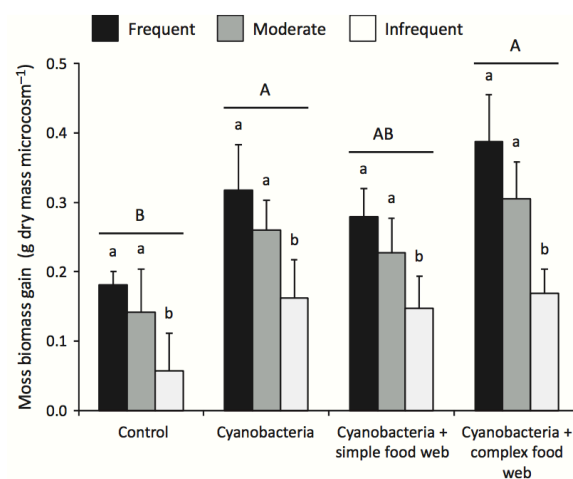
Figure 3 Responses of moss biomass gain to food-web structure (control consisting of sterilised moss with no cyanobacteria added, cyanobacteria, cyanobacteria + simple food web, cyanobacteria + complex food web) and precipitation regime (frequent, moderate, infrequent). Uppercase letters above groups of bars denote significant differences (*post hoc* comparisons based on estimated marginal means, $P < 0.05$) among moss food-web treatments. Lowercase letters above bars indicate significant differences among precipitation regimes *across* food-web treatments. Data are means $\pm$ SE ($N = 16$). ANOVA results are shown in Table 2.

Figure 13.3: Kardol et al. Figure 3

Chapter 14

ANOVA Tables

Treatment effects are most often analyzed using ANOVA, which is short for "Analysis of Variance". This is somewhat of an odd name for a method to test for treatments effects - what do differences in means have to do with an analyis of variance? The name makes sense in light of the decomposition of the total variance into a model variance and the residual variance (chapter xxx). If there are differences among the means, then the total variance is increased because of variation among groups.

The engine underneath modern ANOVA is a linear model. If the model has a single categorical factor, the ANOVA is **one-way**. If the model has two categorical factors it is a two-way ANOVA. If the model has a single categorical factor and one continuous factor it is an ANCOVA, short for **analysis of covariance** (next chapter). More complex experimental designs classically analyzed with ANOVA are nested, split-plot, latin-square, and many others.

14.1 Summary of usage

If you choose to report an ANOVA, also report the effects and their uncertainty in some way, either the model coefficients or contrasts.

1. ANOVA generates a table with one row for each term in the linear model. A term is a factor or a covariate or an interaction. For a two-way factorial ANOVA, these terms are the two main effects and the interaction effect.
2. The ANOVA generates an F and p-value for the whole model and for each term in the ANOVA table.
3. The p-value of an interaction term is often used as a decision rule to interpret the main effects. If $p \leq 0.05$ then do not interpret the main effects but instead examine the condition ("simple") effects. If $p > 0.05$, then interpret the main effects. Regardless, this sort of decision rule is itself controversial, and for good reason.
4. If the main effects are to be interpreted, some statisticians advocate re-fitting the model without the interaction effect, others advocate interpreting the main effects with the interaction term in the model. This only matters if the design is unbalanced (see below).
5. Regardles of any decision, always plot the data using a Harrell plot or interaction plot to understand and communicate the magnitude and pattern of interaction.
6. For factors with more than two levels, the p-value is often used as a decision rule to dissect the factor with post-hoc tests, such as Tukey HSD.
7. A design is balanced if all the cells have the same number of replicates. A design is unbalanced if one or more of the cells has a different number of replicates. Unbalanced designs make it necessary to make decisions, none of which are perfect, and all of which are controversial. Some statisticians have even advocated randomly excluding data until the design is back in balance. Don't do this.

8. There are multiple ways to decompose the sum of squares. I highlight the major three: Type I (sequential), Type II (partial sequential), and Type III. Most statistics software and introductory statistics books default to Type III and, consequently, many researchers are unaware that Types I and II exist. R's default is Type I, and this can make a difference if the design is unbalanced. This is *not* a rare error in publications.

9. Because R defaults to Type I sum of squares, the *p*-value of a factor depends on the order of the factors in the model if the design is unbalanced. This is a feature, not a bug.

10. ANOVA based on type II sum of squares do not depend on factor order if the design is unbalanced, but it does assume that the interaction is zero.

11. ANOVA based on type III sum of squares do not depend on order if the design is unbalanced and does not assume the interaction is zero.

12. If the design is balanced, Type I, II, and III sum of squares generate the same ANOVA table. And the ANOVA table of just the main effects is the same as the ANOVA table that includes the interaction term. None of this is true when the design is unbalanced, However, the decision to use type II or type III is very controversial.

14.2 Example: a one-way ANOVA using the vole data

The vole data has a single factor ("treatment") with three levels ("control", "vitamin_E", "vitamin_C"). In statistics textbooks that emphasize hypothesis testing, the "Which test should I use" flowchart would guide a researcher given this design to a single classification, or one-way ANOVA, since a t-test can only compare two levels but an ANOVA can compare more than two levels. There are better ways to think about what ANOVA is doing, but okay.

Here is an ANOVA table of the vole data:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatment	2	248446	124223.0	2.95	0.057
Residuals	93	3912751	42072.6		

I'll explain all the parts of the ANOVA table later, but for now, focus on the *p*-value, which is that most researchers want out of the table. What null hypothesis does this *p*-value test? The p-value gives the probability of the observed *F* or larger *F*, if the null were true. The null hypothesis models the data as if they were sampled from a single, normal distribution and randomly assigned to different groups. Thus the null hypotheis includes the equality of the means among factor levels. In the vole data, the single treatment factor has three levels and a small *p*-value could occur because of a difference in means between the vitamin_E treatment and control, or between the vitamin_C treatment and control, or between the two vitamin treatments. The *p*-value or ANOVA table doesn't indicate what is different, only that the observed *F* is unexpectedly large if the null were true. As a consequence, researchers typically interpret a low *p*-value in an ANOVA table as evidence of "an effect" of the term but have to use additional tools to dissect this effect. The typical additional tools are either **planned comparisons**, which are contrasts among a subset of a priori identified treatment levels (or groups of levels) or unplanned comparisons ("post-hoc" tests) among all pairs of levels.

The *p*-value in the ANOVA table acts as a decision rule: if $p < 0.05$ then it is okay to further dissect the factor with planned comparisons or post-hoc tests because the significant *p* "protects" the type I error of further comparisons. I'm not fond of using *p*-values for these sorts of decision rules.

14.3 Example: a two-way ANOVA using the urchin data

Let's use the urchin data from the previous chapter xxx to explore the ANOVA table, which is what is typically reported. The experiment has two factors ($Temp$ and $CO2$), each with two levels. Here is the linear model

$$Resp = \beta_0 + \beta_1 Temp + \beta_2 CO2 + \beta_3 TempCO2 + \varepsilon \tag{14.1}$$

In order to understand factorial ANOVA (or any ANOVA with multiple factors), it is useful to know the difference between **conditional means** and **marginal means**

```
##             CO2-  CO2+ Temp-mm
## Temp-      8.233 7.917   8.075
## Temp+     12.743 9.742  11.243
## CO2-mm    10.488 8.829   9.659
```

In the table above, the upper, left 2×2 grid of cells are the conditional means, which are the means of each group, where a group is a specific combination of factor levels. The first two values of the third row are the marginal means for CO2. The first (10.488) is the mean of the two means when CO2=CO2-. This can be written as $E(Resp|CO2-)$. The second (8.829) is the mean of the two means when CO2=CO2+, or $E(Resp|CO2+)$. The first two elements of the third column are the marginal means for Temp. These are $E(Resp|Temp-)$ and $E(Resp|Temp+)$. The bottom right value (9.659) is the grand mean.

A **conditional effect** is a difference between conditional means. For example the conditional effect of *Temp conditional on* CO2=CO2- is $12.743 - 8.233$. A **marginal effect** is a difference in marginal means within a factor, for example the marginal effect of *Temp* is $11.243 - 8.075$.

Here is the ANOVA table of the urchin data

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temp	1	60.2	60.2	19.1	0.0003
CO2	1	16.5	16.5	5.2	0.0332
Temp:CO2	1	10.8	10.8	3.4	0.0791
Residuals	20	63.2	3.2		

This ANOVA table uses what are called Type 3 sum of squares, which is *NOT* the default in R but is the default in many other statistics software and is, therefore, the *only* type of ANOVA that many researchers know (and, many researchers are unaware that there are multiple types of ANOVA table). Understanding these differences is important, at least if one is reporting ANOVA tables. I'll return to the importance of this later.

14.3.1 How to read an ANOVA table

An ANOVA table has a row for each term in the underlying linear model – each of these adds a component of variance to the total, and a row for the residual variance (this residual variance row is frequently excluded from the published table). The urchin model has three terms (one level of *Temp*, one level of *CO2*, and one interaction). The statistics for each term are

1. **Degrees of Freedom** (df) – If the term is a factor, the df will equal the number of levels (k) for the factor minus 1. Think of it this way: the contribution of the variance due to a factor is a function of the variability of the k level means around the grand mean. How many degrees of independent variation do these level means have, given that we know the grand mean? The answer is $k - 1$ – once the values of $k - 1$ level means are written down, the kth level mean has no freedom to vary; its value has to be $k\bar{\bar{Y}} - \sum_i^{k-1} Y_i$. For an interaction term, the df is the product of the df of each of the factors in the interaction.
2. **Sum of Squares** – the sum of squared differences between the modeled value and the grand mean. In addition to a sum of squares for each term, a **residual mean square** is computed as the sum of squared differences between the measured and modeled values.
3. **Mean Square** – The sum of squares divided by the df (this is a "mean" with df acting as the number of things that were summed).
4. **F-ratio** – the Mean Square of the term dived by the residual mean square.

5. **p-value** – the p-value for the F-ratio. F is compared to an F-distribution, which is a distribution of F-ratios under the null.

14.3.1.1   Each row in the table tests a null hypothesis

The row for each term in an ANOVA table tests a null hypothesis. In order to understand the null hypotheses, I need to define a few more terms

For the ANOVA table above, which uses Type 3 sum of squares, the probabilities are

1. Temp – $p = \text{prob}(F \geq F_o|CO2, Temp : CO2)$. The null is no difference in means conditional on the level of CO2 and Temp:CO2. This is equivalent to no difference between the grand mean and the marginal mean of Temp+, or

$$b_1 = \overline{\overline{Resp}} - \text{E}(Resp|Temp^+) \tag{14.2}$$

2. CO2– $p = \text{prob}(F \geq F_o|Temp, Temp : CO2)$. The null is no difference in means conditional on the level of Temp and Temp:CO2. This is equivalent to no difference between the grand mean and the marginal mean of CO2+, or

$$b_2 = \overline{\overline{Resp}} - \text{E}(Resp|CO2^+) \tag{14.3}$$

3. Temp:CO2 – $p = \text{prob}(F \geq F_o|Temp, CO2)$. The null is no difference in means conditional on the level of Temp and CO2. This is equivalent to the difference between the conditional mean of Temp+/CO2+ and the expected conditional mean of Temp+/CO2+ if there were no interaction.

$$b_3 = \text{E}(Resp|Temp^+, CO2^+) - (\overline{\overline{Resp}} - b_1 - b_2) \tag{14.4}$$

As noted in the equations, these three differences are the coefficients of the linear model behind the ANOVA. Here is the coefficient table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.66	0.36	26.6	0.00000
Temp1	-1.58	0.36	-4.4	0.00030
CO21	0.83	0.36	2.3	0.03325
Temp1:CO21	-0.67	0.36	-1.9	0.07910

In ANOVA with type 3 sum of squares, the dummy variables are coded using effect coding, which differs from the dummy coding introduced in chapter xxx. The consequence is that the **grand mean** (the mean of *Resp* across all values) is now the "reference" value. The intercept in this table, then, is the grand mean. The coefficients are *differences from the grand mean*, as described above.

Use the table of conditional and marginal effects above to check that the coefficients equal the differences in the equations above. Also not that the *p*-values for the effects in the coefficient table equals the *p*-values in the ANOVA table.

It is important to note that this table differs from the coefficient table with dummy coding because that reference is the mean of Temp-/CO2- and not the grand mean.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.23	0.73	11.3	0.00000
TempTemp+	4.51	1.03	4.4	0.00028
CO2CO2+	-0.32	1.03	-0.3	0.76081
TempTemp+:CO2CO2+	-2.68	1.45	-1.9	0.07910

Importantly, note that *p*-values for b_1 (the Temp effect) and b_2 differ between the two tables. This is because the *t*-value tests different hypotheses! In the coefficient table with effect coding (that behind the ANOVA

with type 3 sums of squares), the p-value tests marginal effects and so is a function of both marginal means within a factor. By contrast, in the coefficient table with dummy coding, the p-value tests conditional effects, and so is only a function of the conditional means when the other factor is at its reference level (right? The coefficient b_1 in the dummy coded coefficient table is the effect of only increasing $Temp - CO2$ is left at its reference level). For the interaction effect, the coefficient differs between the effects coded model and the dummy coded model (because different reference means) but the p-value ultimately tests the same hypothesis (non-additive effects of the factors) and so the t and p values are the same.

14.3.1.2 What to do after ANOVA?

Researchers frequently report ANOVA statistics (F and p values) for factorial models in a way that suggests that they misunderstand the hypotheses tested. It probably doesn't help that there is a long-standing debate among statisticians about the most sensible strategy for interpreting factorial ANOVA results. And it doesn't help that the default ANOVA table in R can suggest a very different interpretation than the default ANOVA table in some other software packages.

Here are three strategies for interpreting a factorial ANOVA table that uses Type III sum of squares. All strategies use p-values to make a series of decision rules. In the first strategy, which is a type of model simplification or model selection, a researcher starts with the interactions at the bottom of the ANOVA table and works up, eliminating terms with $p > 0.05$ and re-fitting the reduced model before interpreting main effects. In the second strategy, the researcher uses the original ANOVA table that includes all terms to interpret main effects.

Strategy 1

1. is interaction $p < 0.05$?
    a. if yes, then do NOT test main effects. Show a graph to show pattern of conditional effects. Test conditional effects if this is of interest.
    b. if no, then refit model without the interaction and test main effects – This now is equivalent to ANOVA using Type II sum of squares.
        2. is main effect $p < 0.05$$?
            a. if yes, then keep in model
            b. if no, then refit model without that main effect

Strategy 2

2. is interaction $p < 0.05$?
    a. if yes, then do NOT test main effects. Show a graph to show pattern of conditional effects. Test conditional effects if this is of interest.
    b. if no, then use the same table as the test of the main effects. This is interpreting the main effects with the interaction term in the model. This is the logic of ANOVA using type III sum of squares.

Strategy 3

3. is interaction $p < 0.05$?
    a. if yes, then look at interaction plot to determine if it makes sense test main effects. For example, if CO2+ had obviously lower *Resp* at both levels of *Temp*, even if one was much lower (ie. interactaction), then some people would say that the test of the main effect is meaningful. Test conditional effects if this is of interest.
    b. if no, then use the same table as the test of the main effects

In general, statisticians advise against strategy 3 (interpreting main effects in the presence of interaction) – its not wrong, its just that a main effect has an awkward interpretation if there is an interaction. Of course this is true if there is *any* interaction term in the model, not just a statistically significant term. The controversy is more, if the interaction p is not significant, then do we implement stategy 1 (refit model excluding interaction to test main effects) or strategy 2 (use full factorial anova table to test main effects).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temp	1	45.2	45.2	14.5	0.0011
CO2	1	4.1	4.1	1.3	0.2630
Temp:CO2	1	14.8	14.8	4.8	0.0413

then one shouldn't report the ANOVA results using something like "Temperature had a significant effect on metabolism ($F_{1,20} = 14.5$, $p = 0.001$). There was no effect of CO2 on metabolism ($F_{1,20} = 4.1$, $p = 0.26$)". There was a significant interaction effect between Temperature and CO2 on metabolism ($F_{1,20} = 14.8$, $p = 0.04$)". If one accepts that the small interaction p-value is evidence of an interaction effect then this interpretation of the main factors makes no sense, as the first two results imply that the interaction effect is zero (or, that there is a constant effect of $Temp$ or $CO2$ across both levels of the other factor), which is then contradicted by the third result.

More specifically, if one is using a p-value to guide decision making, then a significant interaction p indicates that there is no single "main" effect of a factor. Instead, the effect of $Temp$ is conditional on the level of $CO2$, and the effect of $CO2$ is conditional on the level of $Temp$. This is easily seen in the interaction plot, where the effect of $Temp$ is large when $CO2$ is high but much smaller when $CO2$ is low. Indeed, the effect of $Temp$ at the low CO2 is 0.16.

Instead of interpreting the factors as constant effects, A better strategy is to compare the **conditional effects**, that is, the effects of $Temp$ within each level of $CO2$ and the effects of $CO2$ within each level of $Temp$ (conditional effects are sometimes called the "simple effects").

The controversy arises in what to do after an ANOVA if the interaction effect has a non-significant p-value. At this point, I am punting instead of explaining the basis for the controversy, because ultimately I think the major problem with both strategies is the use of null hypothesis significance testing to make analysis decisions.

In fact, the entire reason that I use the urchin data as the example for factorial ANOVA is because it beautifully illustrates the absurdity of the interaction p-value decision rule. Why should we interpret the results of the ANOVA when the interaction p is 0.079 differently than when the interaction p is 0.04? Remember, the p-value is a "sample" statistic (in the sense that it is entirely a function of the sampled data) and in conditions of low power (which is likely, but not necessarily, true for the urchin data given n=6), a p-value is highly variable.

There are several problems with this approach. 1) a p-value is not evidence of "no effect", 2) the power to test interaction effects is small relative to that for the main effects (this is a general rule, not something specific to these data), 3) the interaction SS accounts for about 7.2% of the total SS, which doesn't seem inconsequential, and 4) the interaction p-value is small enough to raise a red flag, and, most importantly, 5) the confidence interval of the interaction effect indicates that the large, negative values of the interaction are *as consistent with the data* as trivially small values (or a value of zero). But the CI is not in an ANOVA table and many researchers fail to report it. These five points suggest that this experiment be replicated, with a larger sample size, to get a better estimate of the interaction effect. The problem of course is that experiments are rarely replicated, except in biomedical research.

The absurdity of the p-value decision rule strategy for interpretation of effects after an ANOVA is highlighted by comparing the forest plot of model coefficients of the real and fake urchin data. It would be absurd to use an ANOVA table to interpret these patterns as radically different (one without an interaction and constant main effects, the other with an interactioni and conditional effects).

14.3.2 How to read ANOVA results reported in the text

ANOVA results are often reported in the text of a results section, using something like "Temperature had a significant effect on metabolism ($F_{1,20} = 14.5$, $p = 0.001$). There was no effect of CO2 on metabolism ($F_{1,20} = 4.1$, $p = 0.26$)". The subscripts of the F statistic are the numerator and denominator degrees of freedom (df) of the F-value (These df are a column in the ANOVA table. The denomintor df may not appear in the table if it is the residual df and the row for the residual term was not reported). Sometimes I find the
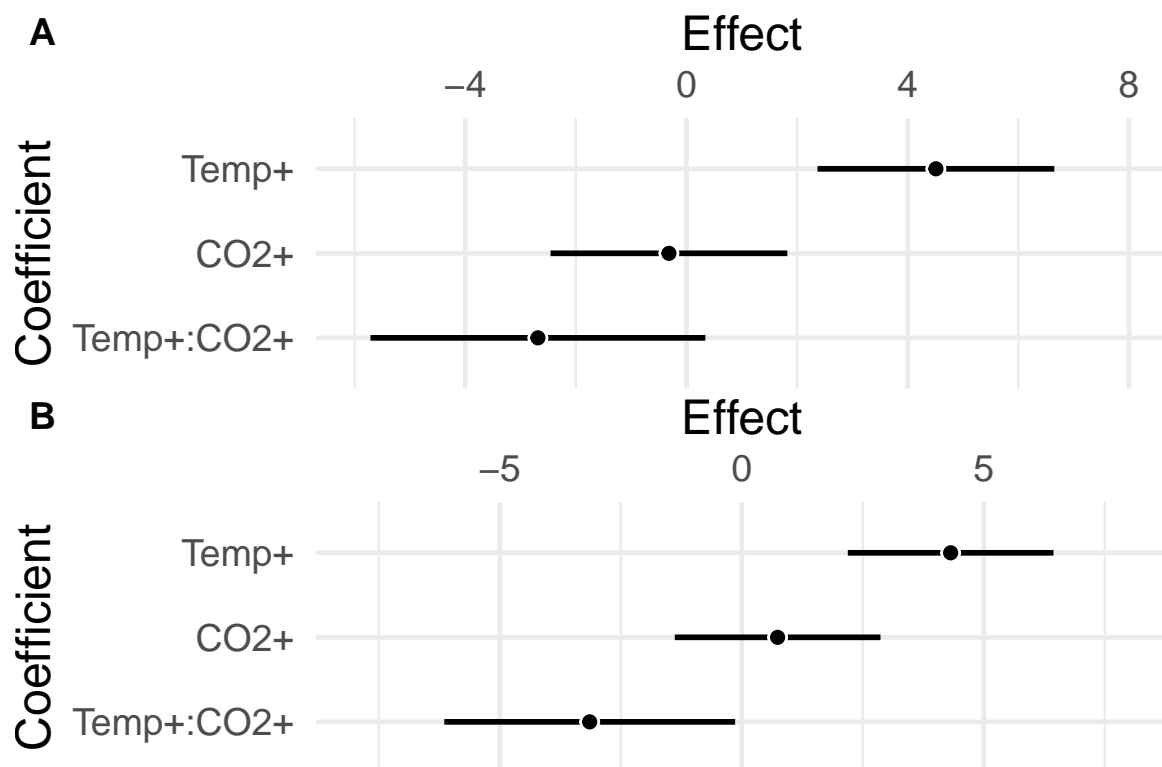
Figure 14.1: Forest plots (the upper part of a Harrell plot) of the actual and fake urchin data. A) Real urchin data. The interaction effect is not significant ($p = 0.079$). B) Fake urchin data. The interaction effect is significant ($p = 0.04$).

reported df are not consistent with the description of the design and analysis, which means the data were not analyzed as stated.

14.3.3   Better practice – estimates and their uncertainty

As emphasized in the previous chapter, the decision to include or exclude an interaction effect in the model should not be based on a p-value but on the goals of the model.

1. If the goal is the interaction (because a scientific model predicts one, or because this is biology and everything is conditional), then estimate the interaction effect (as a coefficient of the model!) and its uncertainty, including a CI and p-value. There is no controversy on how to estimate this effect and its uncertainty. The coefficient will be different between dummy and effect coded models but this is okay because they have different specific interpretations but the same general interpretation. Use a Harrel plot with the coefficients (including the interaction coefficient) to show this estimate and uncertainty.

2. If the goal is to estimate constant main effects, then exclude the interaction effect from the model and report the main effects (again, as coefficients from the model or contrasts if other pairwise effects are desired) with their uncertainty. Use an interaction plot (or bottom part of the harrell plot) to justify forcing the interaction to zero (for example the interaction effect adds little to the total sum of squares or the interpretation of a single main effect or two (or more) conditional effects would be the same. Use a Harrel plot that excludes the interaction term to show these main effects and uncertainty.

3. And if a researcher is interested in the effects of the factors but there is strong evidence for a non-trivial interaction, then report the conditional effects (as contrasts) with their uncertainty. Use a Harrel plot that includes the interaction term to show these conditional effects and uncertainty. If there is an obvious interaction, it probably doesn't make sense to interpret the main effects, contrary to what some people argue. If there is a positive effect of factor A across all levels of factor B, we don't really need a p-value to test that the average of these positive effects is significant. This doesn't add value to the plot and any conditional effects that are reported.

Notice that an ANOVA table has no role in this recommendation.

14.4   Unbalanced designs

My recommendation above is to not bother with ANOVA, but to simply compute the contrasts of interest using the linear model. But if you really want to use ANOVA, you should be aware that **if the design is unbalanced, factor order matters in the default R anova function** and that I routinely find published ANOVA tables that report statistics (F and p values) that are not what the authors think they are.

An **unbalanced** design is one in which the number of replicates differs among the cell. The urchin data is balanced because there are six replicates in each cell. If the respirometer broke before taking the respiratory measures of the final tank, the design would be unbalanced, one of the cells would have only five replicates.

Let's look at the effect of row order on the statistics of the urchin data using R's default anova function.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temp	1	60.20	60.20	19.06	0.00030
CO2	1	16.52	16.52	5.23	0.03325
Temp:CO2	1	10.81	10.81	3.42	0.07910

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CO2	1	16.52	16.52	5.23	0.03325
Temp	1	60.20	60.20	19.06	0.00030
CO2:Temp	1	10.81	10.81	3.42	0.07910

Now let's unbalance the data, by removing three random replicates (these may be both in one cell or spread across cells. First, here is the number of replicates in each cell:

```
##
##         CO2- CO2+
##   Temp-   6    4
##   Temp+   6    5
```

And here are the two tables with the order of Temp and CO2 reversed in the model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temp	1	62.25	62.25	18.44	0.00049
CO2	1	21.49	21.49	6.36	0.02190
Temp:CO2	1	6.38	6.38	1.89	0.18720

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CO2	1	17.59	17.59	5.21	0.03561
Temp	1	66.14	66.14	19.59	0.00037
CO2:Temp	1	6.38	6.38	1.89	0.18720

Several observations are important.

1. the statistics for the last row, which is the interaction, does not change.
2. if these data were analyzed in the software package JMP, or SAS, or SSPS then **order wouldn't matter**. Here is what the tables would look like

	Sum Sq	Df	F value	Pr(>F)
Temp	58.77	1	17.41	0.00064
CO2	19.93	1	5.90	0.02648
Temp:CO2	6.38	1	1.89	0.18720

	Sum Sq	Df	F value	Pr(>F)
CO2	19.93	1	5.90	0.02648
Temp	58.77	1	17.41	0.00064
CO2:Temp	6.38	1	1.89	0.18720

3. Order does not change the statistics in the coefficient table, even for unbalanced data:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.407	23.367	0.0000
Temp1	-1.70	0.407	-4.172	0.0006
CO21	0.99	0.407	2.430	0.0265
Temp1:CO21	-0.56	0.407	-1.374	0.1872

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.50	0.407	23.367	0.0000
CO21	0.99	0.407	2.430	0.0265
Temp1	-1.70	0.407	-4.172	0.0006
CO21:Temp1	-0.56	0.407	-1.374	0.1872

14.4.1 What is going on in unbalanced ANOVA? – Type I, II, III sum of squares

Type I sum of squares. Here is the (default) ANOVA table using Type I sum of squares for the urchin data with the three missing rows.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temp	1	62.248	62.248	18.4	0.0005
CO2	1	21.488	21.488	6.4	0.0219
Temp:CO2	1	6.377	6.377	1.9	0.1872
Residuals	17	57.399	3.376		

The default coding of dummy variables in R's `lm` function is dummy coding, which is the coding used for Type I or **Sequential Sum of Squares**. The hypothesis tested by each row in the ANOVA table using Type I sum of squares is the effect of that row's term conditional on all terms before it in the model (or above it in the table) and ignoring all terms after it in the model (or below it in the table).

1. The hypothesis tested by the p-value for $Temp$ is the same as if $Temp$ were the only term in the model (other than the intercept). That is, the means are estimated for each level of $Temp$ ignoring the fact that half the replicates within each level of $Temp$ experienced low CO_2 and half experienced high CO_2

2. The hypothesis tested by the p-value for CO_2 is conditional on $Temp$. That is, the difference in metabolism between CO_2+ and CO_2- when $Temp$ is "held constant" (or for all cases where $Temp$ takes the same value). This is equivalent to the hypothesis that the difference in the marginal means of CO2 is zero.

3. The hypothesis tested by the p-value for the interaction is conditional on all other terms and nothing is ignored.

Type II sum of squares. Here is the ANOVA table using Type II sum of squares for the urchin data with missing values. The interaction term is excluded from the linear model, because type II sum of squares are used to estimate main effects ignoring the interaction (so this would make sense only if a plot of the effects suggested a small interaction relative to the main effects). The sum of squares for the main effects would be the same if the interaction were included but the residual df, and thus the F and P-values would differ.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temp	1	66.145	66.145	18.7	0.0004
CO2	1	21.488	21.488	6.1	0.0241
Residuals	18	63.776	3.543		

The hypothesis tested by each row in the ANOVA table using Type II sum of squares is the effect of that row's term conditional on all terms *at the same level or below* but ignoring all terms at a higher level in the model (or below it in the table). For example, the hypothesis test for a factor is conditioned on other factors but ignores interaction terms among the factors. Consequently, these hypotheses tested are

1. The hypothesis tested by the p-value for $Temp$ is conditional on CO_2. This is the same hypothesis that would occur using Type I sum of squares but placing $Temp$ second in the model, after CO_2 (and it is in fact how I computed it for the table).

2. The hypothesis tested by the p-value for CO_2 is conditional on $Temp$. This is exactly the hypothesis for CO_2 using the Type I sum of squares above.

Type III sum of squares. Here is the ANOVA table using Type III sum of squares for the urchin data for missing data. The interaction term is excluded from the linear model, and advocates of using Type III sum of squares explicitly want this in the model.

	Sum Sq	Df	F value	Pr(>F)
Temp	58.770	1	17.406	0.0006
CO2	19.935	1	5.904	0.0265
Temp:CO2	6.377	1	1.889	0.1872
Residuals	57.399	17		

The hypothesis tested by each row in the ANOVA table using Type III sum of squares is the effect of that row's term conditional on all terms in the model.

1. The hypothesis tested by the p-value for $Temp$ is conditional on CO_2 and $Temp : CO_2$.

2. The hypothesis tested by the *p*-value for $CO2$ is conditional on $Temp$ and $Temp : CO2$.

3. The hypothesis tested by the *p*-value for $Temp : CO2$ is conditional on $Temp$ and $CO2$. This is the same for Type I sum of squares (and Type II, if the interaction term were included)

14.4.2 Back to interpretation of main effects

14.4.3 The anova tables for Type I, II, and III sum of squares are the same if the design is balanced.

14.5 Working in R

14.5.1 Type I sum of squares in R

The base R function `anova()` computes the ANOVA table using Type I sum of squares for any fit model object, such as that returned by `lm`. Here is a script for the urchin data. I first create unbalanced data by deleting the first row that is the control row.

```
cn_rows <- which(urchin[, Temp]=="Temp-" & urchin[, CO2]=="CO2-") # gives the rows of the controls
urchin_unbalanced <- urchin[-cn_rows[1],] # deletes the row that is in first element of cn_rows
urchin.t1 <- lm(Resp ~ Temp*CO2, data=urchin_unbalanced)
anova(urchin.t1)
```

```
## Analysis of Variance Table
##
## Response: Resp
##           Df Sum Sq Mean Sq F value     Pr(>F)
## Temp       1 55.696  55.696 16.9244 0.0005907 ***
## CO2        1 18.411  18.411  5.5946 0.0288072 *
## Temp:CO2   1  9.204   9.204  2.7970 0.1108298
## Residuals 19 62.527   3.291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

14.5.2 Type II and III Sum of Squares

Type II sum of squares can be computed manually simply by fitting the model twice, once with the factors ordered one way and then with the factors ordered the opposite way. The car package has the function `Anova` that specifically outputs Type II and Type III ANOVA tables.

Type II sum of squares can be fit with the interaction in the model, and this generates the Type II sum of squares for the main terms but the residual is wrong for the F-ratio because it is the residual from the full model and Type II assumes the interaction effect is zero. So, if one wants an ANOVA table with a F and p that reflect this, then the interaction should be dropped from the model.

```
urchin.t2 <- lm(Resp ~ Temp*CO2, data=urchin_unbalanced)
Anova(urchin.t2, type="2")
```

```
## Anova Table (Type II tests)
##
## Response: Resp
##        Sum Sq Df F value    Pr(>F)
## Temp   52.711  1 16.0173 0.0007624 ***
```

```
## CO2       18.411  1  5.5946 0.0288072 *
## Temp:CO2   9.204  1  2.7970 0.1108298
## Residuals 62.527 19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
urchin.t2 <- lm(Resp ~ Temp + CO2, data=urchin_unbalanced)
Anova(urchin.t2, type="2")
```

```
## Anova Table (Type II tests)
##
## Response: Resp
##          Sum Sq Df F value   Pr(>F)
## Temp     52.711  1 14.6968 0.001038 **
## CO2      18.411  1  5.1333 0.034725 *
## Residuals 71.731 20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To get type III sum of squares, we need to specify effects coding for the model matrix. The safest way to do this is something like this

```
con3 <- list(Temp=contr.sum, CO2=contr.sum) # change the contrasts coding for the model matrix
urchin.t3 <- lm(Resp ~ Temp*CO2, data=urchin_unbalanced, contrasts=con3)
Anova(urchin.t3, type="3")
```

```
## Anova Table (Type III tests)
##
## Response: Resp
##              Sum Sq Df  F value    Pr(>F)
## (Intercept) 2148.60  1 652.8939 3.559e-16 ***
## Temp          54.71  1  16.6241 0.0006422 ***
## CO2           17.15  1   5.2119 0.0341221 *
## Temp:CO2       9.20  1   2.7970 0.1108298
## Residuals     62.53 19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chapter 15

Predictive Models

This chapter focusses on modeling **observational data** with multiple X variables, both continous and categorical. The classical analysis of multiple X variables is **multiple regression**, sometimes called **multivariable regression** and occassionally, but incorrectly, called **multivariate regression** – "multivariate" refers to multiple Y variables.

The models in this chapter have the structure

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...\beta_p X_p + \varepsilon \tag{15.1}$$

% where p is the number of X variables or **predictors** in the model. This equation is easily generalized to both generalized linear models, linear mixed models, and generalized linear mixed models.

15.1 Overfitting

When a model is fit to data, the model coefficients are estimates of the parameters that "generated the data". The value of an estimate is partly a function of the signal (the parameter) and partly a function of the noise, which is unique to the sample. At a low signal to noise ratio a model is mostly fitting the noise. A measure of how well the model "fits" the data is R^2, which is

$$R^2 < -1 - \frac{SS_{residual}}{SS_{total}} \tag{15.2}$$

As X variables are added to a model, the R^2 necessarily increases. Part of this increase is due to added signal, but part is due to added noise. If the added noise is more than the added signal, then the model fit – that is the parameter estimates – increasingly reflects the noise unique to the sample rather the signal common to every sample. This is the basis of **overfitting**.

To demonstrate overfitting, I fit completely random X variables to the lifespans for the control voles.

Think about it this way: if I create fake data in there are ten X variables that are correlewhich Y is a simple column of random, normal variables that are not a function of

15.2   Model building vs. Variable selection vs. Model selection

15.2.1   Stepwise regression

15.2.2   Cross-validation

15.2.3   Penalization

15.2.3.1   AIC

15.2.3.2   LASSO

15.3   Shrinkage

Part V: Expanding the Linear Model – Generalized Linear Models and Multilevel (Linear Mixed) Models

Chapter 16

Generalized linear models I: Count data

Biologists frequently count stuff, and design experiments to estimate the effects of different factors on these counts. For example, the effects of environmental mercury on clutch size in a bird, the effects of warming on parasite load in a fish, or the effect of exercise on RNA expression.

Count data differ from data with normal error in many ways, including 1) counts are discrete, and can be zero or positive integers only, 2) counts tend to bunch up on the small side of the range, creating a distribution with a positive skew, 3) a sample of counts can have an abundance of zeros, and 4) the variance of counts increases with the mean (see Figure 16.1 for some of these properties). Some count data can be approximated by a normal distribution and reasonably modeled with a linear model but more often, count data are modeled with **Poisson distribution** or **negative binomial distribution** using a **generalized linear model** (GLM). Poisson and negative binomial distributions are **discrete probability distributions** with two important properties: 1) the distribution contains only zero and positive integers and 2) the variance is a function of the mean. Back before modern computing and fast processors, count data were often analyzed by either **transforming** the response or by **non-parametric hypothesis tests**. One reason to prefer a statistical modeling approach with a GLM is that we can get interpretable parameter estimates. By contrast, both the analysis of transformed data and non-parametric hypothesis tests are really tools for computing "correct" p-values.

16.1   The generalized linear model

As outlined in section [Assumptions for inference with statistical models] in Chapter 1, a common way that biological researchers think about a response variable is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{16.1}$$
$$\varepsilon \sim N(0, \sigma) \tag{16.2}$$

That is, we can think of a response as the sum of some systematic part and "random errror", which is a random draw from a normal distribution with mean zero and variance σ^2. This way of thinking about the generation of the response is useful for linear models, and model checking linear models, but is not useful for generalized linear models or model checking generalized liner models. For example, if we want to model the number of parasites that have infected a fish using a Poisson distribution, the following is the **wrong** way to think about the statistical model
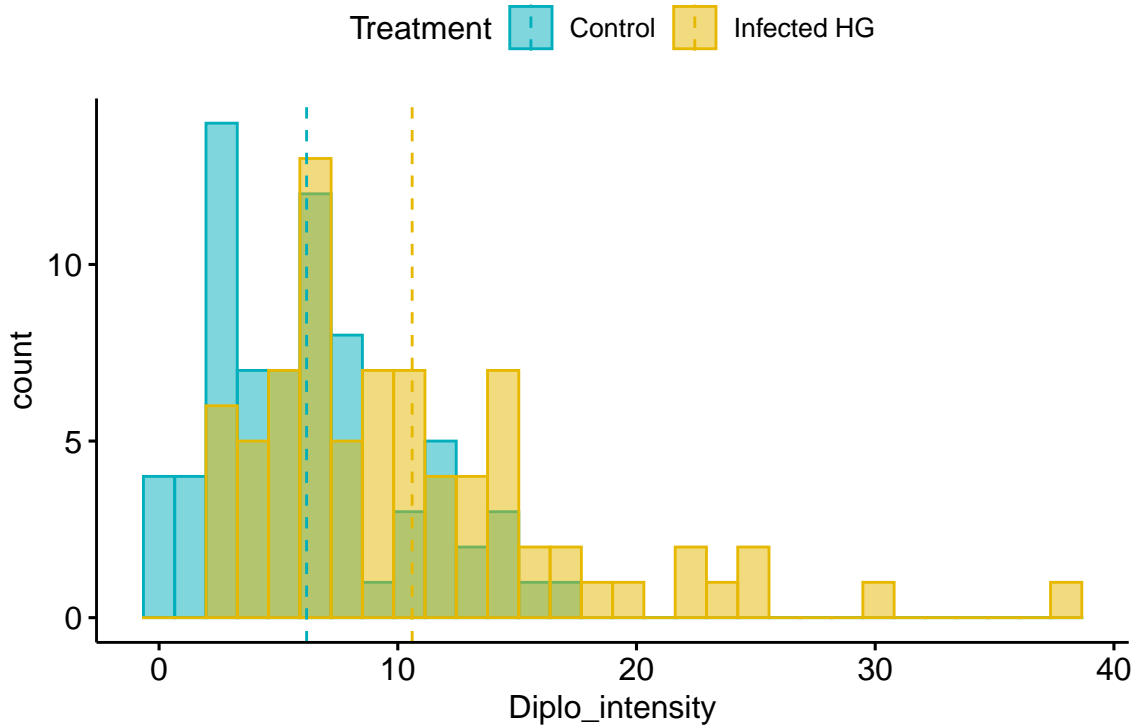
Figure 16.1: Histogram of the count of a trematode parasite larvae in Control vs. Infected fish. Fish in the Infected treatment are infected with a tapeworm.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{16.3}$$
$$\varepsilon \sim Poisson(\lambda) \tag{16.4}$$

That is, we should not think of a count as the sum of a systematic part and a random draw from a Poisson distribution. Why? Because it is the counts (or the counts conditional on X) that are poisson distributed, not the residuals from the fit model.

Thinking about the distribution of count data using model (16.4) leads to absurd consequences. For example, if we set the mean of the Poisson "error" to zero (like with a normal distribution), then the error term for every observation would *have to* be zero (because the only way to get a mean of zero with non-negative integers is if every value is zero). Or, if the study is modeling the effect of a treatment on the counts (that is, the X are dummy variables) then β_0 is the expected mean count of the control (or reference) group. But if we add non-zero Poisson error to this, then the mean of the control group would be larger than β_0. This doesn't make sense. And finally, equation (16.4) generates a continuous response, instead of an integer, because β_0 and β_1 are continuous.

A better way to think about the data generation for a linear model, because this naturally leads to the *correct* way to think about data generation for a generalized linear model, is

$$y_i \sim N(\mu_i, \sigma) \tag{16.5}$$
$$E(Y|X) = \mu \tag{16.6}$$
$$\mu_i = \beta_0 + \beta_1 x_i \tag{16.7}$$

That is, a response is a random draw from a normal distribution with mean mu (not zero!) and variance σ^2. Line 1 is the stochastic part of this specification. Line 3 is the systematic part.

The specification of a generalized linear model has both stochastic and systematic parts but adds a third part, which is a **link function** connecting the stochastic and systematic parts.

1. **The stochastic part**, which is a probability distribution from the exponential family (this is sometimes called the "random part")

$$y_i \sim \text{Prob}(\mu_i) \tag{16.8}$$

2. **the systematic part**, which is a linear predictor (I like to think about this as the deterministic part)

$$\eta = \mathbf{X}\boldsymbol{\beta} \tag{16.9}$$

3. a **link function** connecting the two parts

$$\eta_i = g(\mu_i) \tag{16.10}$$

μ (the Greek symbol mu) is the conditional mean (or expectation $\text{E}(Y|X)$) of the response on the **response scale** and η (the Greek symbol eta) is the conditional mean of the response on the **link scale**. A GLM models the response with a distribution specified in the stochastic part. The probability distributions introduced in this chapter are the Poisson and Negative Binomial. The natural link function for the Poisson and Negative Binomial is the "log link", $\eta = log(\mu)$. More generally, while each distribution has a natural (or, "canonical") link function, one can use alternatives. Given this definition of a generalized linear model, a linear model is a GLM with a normal distribution and an Identity link ($\eta = \mu$).

When modeling counts using the Poisson or negative binomial distributions with a log link, the link scale is linear, and so the effects are additive on the link scale, while the response scale is nonlinear (it is the exponent of the link scale), and so the effects are multiplicative on the response scale. If this doesn't make sense now, an example is worked out below. The inverse of the link function backtransforms the parameters from the link scale back to the response scale. So, for example, a prediction on the response sale is $\exp(\hat{\eta})$ and a coefficient on the response scale is $\exp(b_j)$.

16.2 Count data example – number of trematode worm larvae in eyes of threespine stickleback fish

The example is an experiment measuring the effect of the parasitic tapeworm *Schistocephalus solidus* infection on the susceptibility of infection from a second parasite, the trematode *Diplostomum pseudospathaceum*, in the threespine stickleback fish *Gasterosteus aculeatus*[1]. The treatment levels are "Control" (unexposed to the tapeworm), "Uninfected" " (exposed to the tapeworm but uninfected), "Infected LG" (exposed and infected with the low growth population of the tapeworm), and "Infected HG" (exposed and infected with the high growth population of tapeworm). The response is the number of trematode larvae counted in the eyes (right and left combined) of the fish. A histogram of the counts is shown in Figure 16.1 for the control and Infected HG treatment levels.

16.2.1 Modeling strategy

NHST blues – Students are often encouraged by textbooks, colleagues, or the literature to start the analysis by first "testing" assumptions with hypothesis tests – for example using a Shaprio-Wilks test of normality as a decision rule to decide if to use a parametric test such as a t-test or ANOVA if the null hypothesis of normality is not rejected, or a non-parametric test such as a Mann-Whitney U test if the null hypothesis of normality is rejected. I advise against this, because 1) this pre-test filtering automatically invalidates the p-value of the hypothesis test as it does not adjust for the filtering procedure, 2) real data are only

---

[1]Benesh, D. P., & Kalbe, M. (2016). Experimental parasite community ecology: intraspecific variation in a large tapeworm affects community assembly. Journal of Animal Ecology, 85(4), 1004-1013
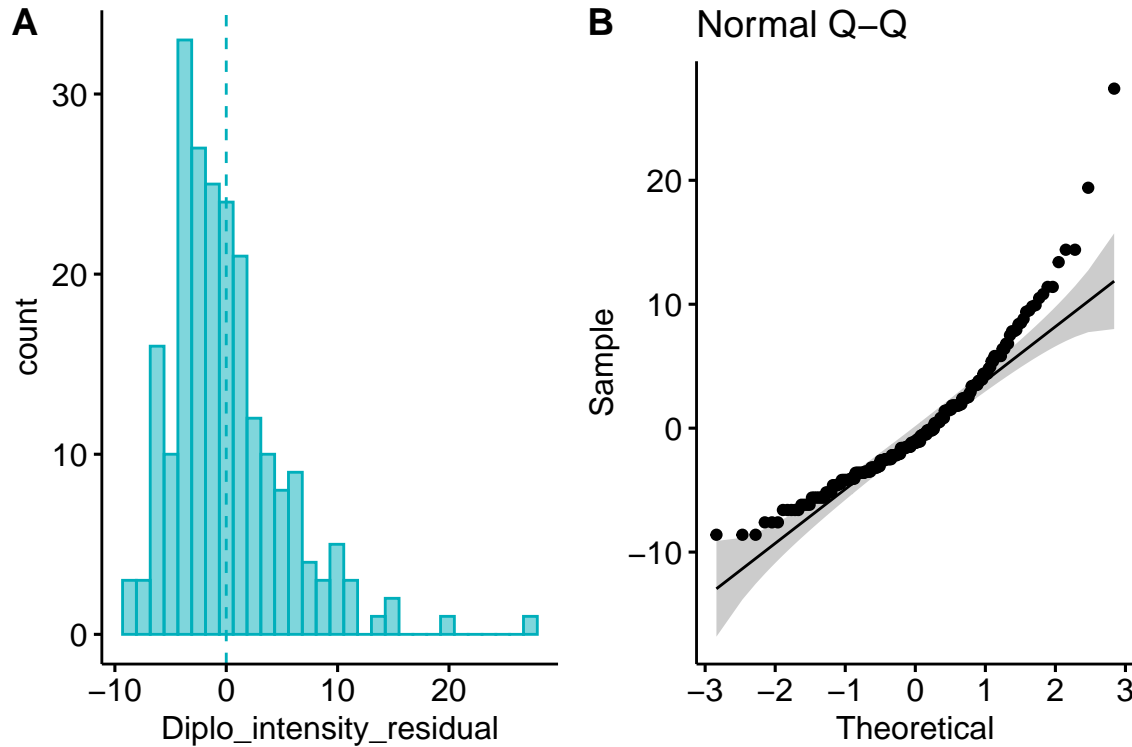
Figure 16.2: Diagnostic plots of stickleback parasite data. A) Distribution of the residuals of the fit linear model. B) Normal Q-Q plot of the residuals of the fit linear model.

approximately normal and as n increses, a normality test will reject any real dataset, and 3) hypothesis tests are pretty robust to non-normality anyway.

Instead of testing assumptions of a model using formal hypothesis tests *before* fitting the model, a better strategy is to 1) fit a model, and then do 2) **model checking** using **diagnostic plots**, diagnostic statistics, and simulation.

With these data, a researcher would typically fit a GLM with a Poisson or negative binomial distribution and log link. Here, I start with a linear model to illustrate the interpretation of diagnostic plots with non-normal data. I use the "linear model" specification (equation (16.2)) because the diganostic plots for model checking a linear model use the residuals of the fit model.

$$Diplo\_intensity_i = \beta_0 + \beta_1 Uninfected_i + \beta_2 Infected\_LG_i + \beta_3 Infected\_HG_i + \varepsilon_i \qquad (16.11)$$
$$\varepsilon \sim N(0, \sigma) \qquad (16.12)$$

16.2.2   Checking the model I – a Normal Q-Q plot

Figure 16.2A shows a histogram of the residuals from the fit linear model. The plot shows that the residuals are clumped at the negative end of the range, which suggests that a model with a normally distributed conditional outcome (or normal error) is not well approximated.

A better way to investigate this is with the **Normal Q-Q** plot in Figure 16.2B, which plots the sample quantiles for a variable against their theoretical quantiles. If the conditional outcome approximates a normal distribution, the points should roughly follow the line. Instead, for the worm data, the points are above the line at both ends. At the left (negative) end, this means that we aren't seeing the most negative values that
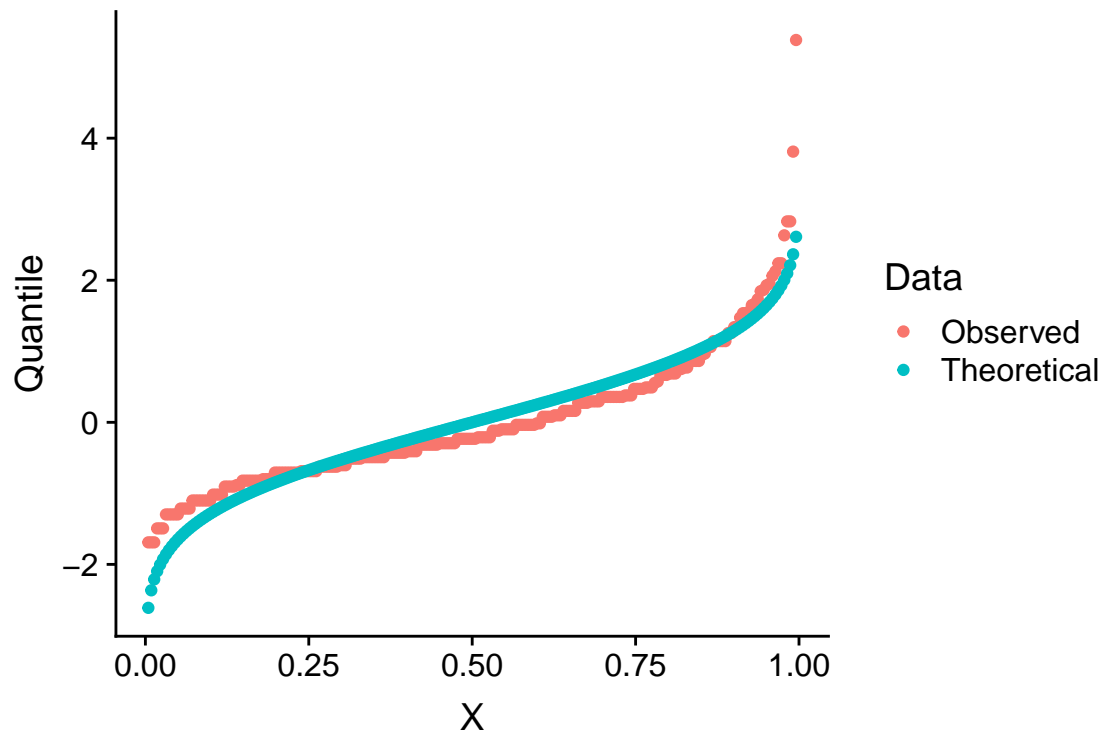
would be expected (the observed values are more positive than the theoretical values). Remembering that this plot is of residuals, if we think about this as counts, this means that our smallest counts are not as small as we would expect given the mean and a normal distribution. This shouldn't be surprising – the counts range down to zero and counts cannot be below zero. At the positive end, the sample values are again more positive than the theoretical values. Thinking about this as counts, this means that are largest counts are larger than expected given the mean and a normal distribution. This pattern is exactly what we'd expect of count data, or at least count data that borders zero.

Intuition Pump – Let's construct a Normal Q-Q plot. A **quantile** (or percentile) of a vector of numbers is the value of the point at a specified percentage rank. The median is the 50% quantile. The 95% confidence intervals are at the 2.5% and 97.5% quantiles. In a Normal Q-Q plot, we want to plot the quantiles of the residuals against a set of theoretical quantiles.
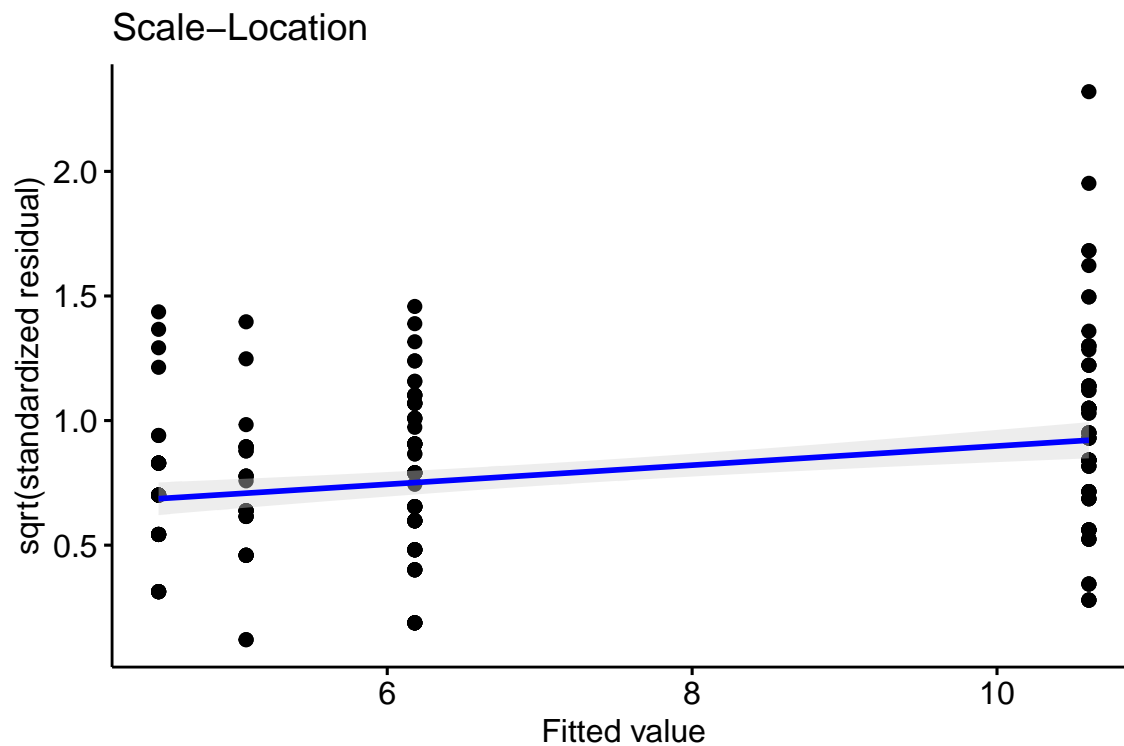
1. To get the observed quantiles, rank the residuals of the fit linear model from most negative to most positive – these are your quantiles! For example, if you have $n = 145$ residuals, then the 73rd point is the 50% quantile.
2. A theoretical quantile from the normal distribution can be constructed using the `qnorm` function which returns the normal quantiles for a specified vector of percents. Alternatively, one could randomly sample n points using `rnorm`. These of course will be sampled quantiles so will only approximate the expected theoretical quantiles, but I add this here because we use this method below.

Now simply plot the observed against theoretical quantiles. Often, the **standardized** quantiles are plotted. A standardized variable has a mean of zero and a standard deviation of one and is computed by 1) centering the vector at zero by subtracting the mean from every value, and 2) dividing each value by the standard deviation of the vector. Recognize that because a standard deviation is a function of deviations from the mean, it doesn't matter which of these operations is done first. A standardized theoretical quantile is specified by `qnorm(p, mean = 0, sd = 1)`, which is the default.

Below, I've plotted the standardized observed and theoretical quantiles against the vector of percents (from 0 to 100%). This plot also nicely shows how the residuals of the worm data deviate from that expected if these had a normal distribution. The plot nicely shows that the most negative observed quintiles are not as negative as expected given a normal distribution, which again makes sense because this would imply negative counts since the mean is close to zero. And it nicely shows that the most positive observed quantiles are more positive than expected given a normal distribution, again this makes sense in right skewed count data. Finally, the plot nicely shows that the median is less positive than that expected given a normal distribution, which is at the mean (a right skew tends to pull the mean to the right of the median).

16.2.3   Checking the model II – scale-location plot for checking homoskedasticity

A linear model also assumes the error has constant variance (that is, the error variance is not a function of the value of X), or homoskedasticity. The fit model can be checked for homoskedasticity using a scale-location

plot, which is a scatterplot of the positive square-root of the standardized residuals against the fitted values[2]. If the residuals approximate a normal distribution, then a regression line through the scatter should be close to horizontal. The regression line in the scale-location plot of the fit of the linear model to the worm data shows a distinct increase in the "scale" (the square root of the standardized residuals) with increased fitted value, which is expected of data that are lognormally, Poisson, or negative binomially distributed.

16.2.4   Two distributions for count data – Poisson and Negative Binomial

The pattern in the normal Q-Q plot in Figure 16.2B should discourage one from modeling the data with a normal distribution and instead model the data with an alternative distribution using a Generalized Linear Model. There is no unique mapping between how data are generated and a specific distribution, so this decision is not as easy as thinking about the data generation mechanism and then simply choosing the "correct" distribution. Section 4.5 in Bolker (xxx) is an excellent summary of how to think about the generating processes for different distributions in the context of ecological data. Since the response in the worm data are counts, we need to choose a distribution that generates integer values, such as the Poisson or the negative binomial.

1. Poisson – A Poisson distribution is the probability distribution of the number of occurrences of some thing (an egg, a parasite, or a specific mRNA transcript) generated by a process that generates the thing at a constant rate per unit effort (duration or space). This constant rate is λ, which is the expectation, so $E(Y) = \mu = \lambda$. Because the rate per effort is constant, *the variance of a Poisson variable equals the mean*, $\sigma^2 = \mu = \lambda$. Figure **??** shows three samples from a Poisson distribution with λ set to 1, 5, and 10. The plots show that, as the mean count (λ) moves away from zero, a Poisson distribution 1) becomes less skewed and more closely approximates a normal distribution and 2) has an increasingly low probability of including zero (less than 1% zeros when the mean is 5).

A Poisson distribution, then, is useful for count data in which the conditional variance is close to the conditional mean. Very often, biological count data are not well approximated by a Poisson distribution because the variance is either less than the mean, an example of **underdispersion**[3], or greater than the mean, an example of **overdispersion**[4]. A useful distribution for count data with overdispersion is the negative binomial.

---

[2]fitted values are the predicted values, $\hat{Y}$

[3]the variance is less than that expected by the probability model

[4]the variance is greater than that expected by the probability model

2. Negative Binomial – The negative binomial distribution is a discrete probability distribution of the number of successes that occur before a specified number of failures k given a probability p of success. This isn't a very useful way of thinking about modeling count data in biology. What is useful is that the Negative Binomial distribution can be used simply as way of modeling an "overdispersed" Poisson process. The mean of a negative binomial variable is $\mu = k\frac{p}{1-p}$ and the variance is $\sigma^2 = \mu + \mu^2/k$. As a method for modeling an overdispersed Poisson variable, k functions as a **dispersion parameter* controlling the amount of overdispersion and can be any real, positive value (not simply a positive integer), including values less than 1.

16.2.5   Fitting a GLM with a Poisson distribution to the worm data

Let's fit a GLM with a Poisson distribution to the worm data. The model is

$$Diplo\_intensity_i \sim Poisson(\mu_i) \tag{16.13}$$
$$\mathrm{E}(Diplo\_intensity|Treatment) = \mu \tag{16.14}$$
$$\mu_i = \exp(\eta_i) \tag{16.15}$$
$$\eta_i = \beta_0 + \beta_1 Uninfected_i + \beta_2 Infected\_LG_i + \beta_3 Infected\_HG_i \tag{16.16}$$

1. The first line of the model is the stochastic part stating the response is modeled as a random Poisson variable with mean and variance μ (the rate parameter λ of the Poisson distribution).
2. The second line states the μ is the conditional mean or conditional expectation
3. The third line connects the conditional mean on the link scale (η) with the conditional mean on the response scale (μ)
4. The fourth line is the linear predictor, and includes three dummy variables.

Remember that the conditional mean is the expected/predicted/fitted/modeled value when $X = x_i$.
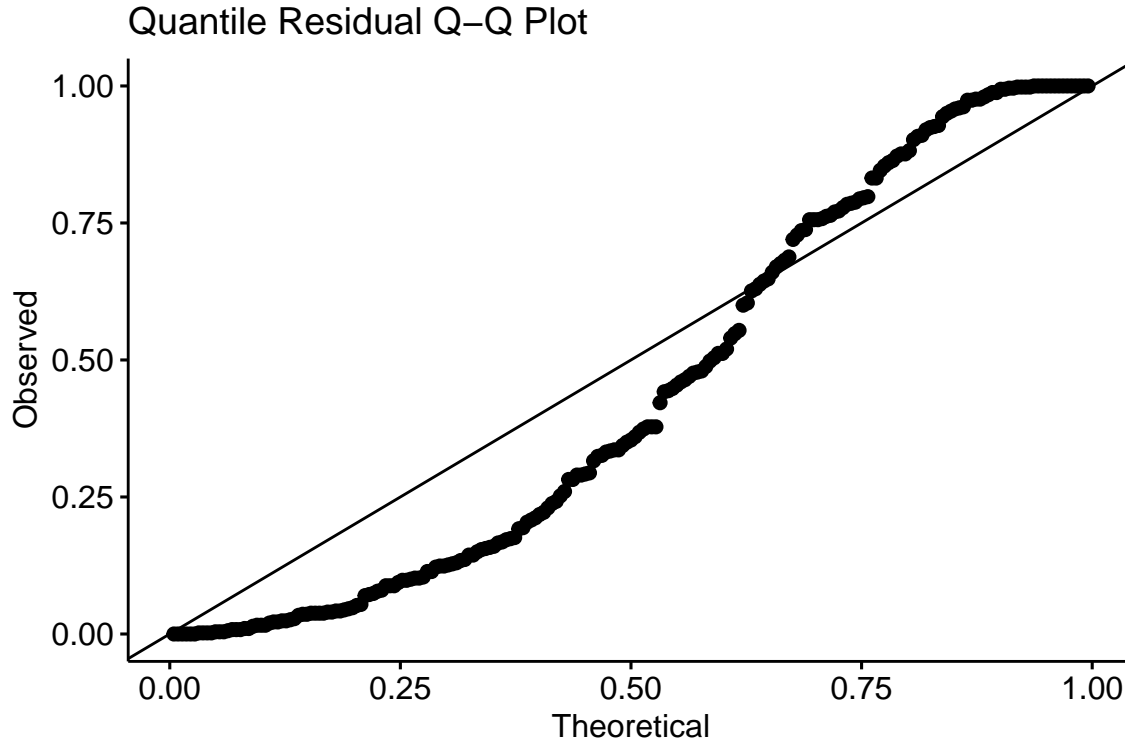
16.2.6   Model checking fits to count data

we use the fit model to check 1. the overall similarity of observed and theoretical distributions 2. if the observed distribution is over or under dispersed 3. if there more zeros than expected by the theoretical distribution. If so, the observed distribution is **zero-inflated**

16.2.6.1   Model checking a GLM I – the quantile residual Q-Q plot

A quantile-quantile (Q-Q) plot is used to check overall similarity of the observed distribution with the distribution that would be expected under the model. An alternative to a Normal Q-Q plot for a GLM fit is a quantile residual Q-Q plot of observed vs. expected **quantile residuals**. The basic algorithm for this is

1. Use the model parameters to simulate p fake values of the response for each row of the data. This will be a $n \times p$ matrix of fake data where each column is a new, random sample of a population with parameters equal to that estimated for the observed data. For the Poisson, the parameter for each observation will be $\hat{\mu}_i$, the modeled value of observation i. For the negative binomial, the parameters will be $\hat{\mu}_i$ and the dispersion parameter k, which is the same for all observations.
2. For each observation (each row of the matrix of fake data), compute the fraction of simulated values smaller than the observed value of the response variable for that row. This fraction is the observed **quantile residual**, which ranges in value from 0 to 1. If the true data are distribitued as that specified by the model, then quantile residuals will have a uniform distribution.
3. Sort the observed quantile residuals from smallest to largest and plot against theoretical quantile residuals from a uniform distribution. One could transform the quantile residuals to standard, normal residuals and then plot using a traditional Normal Q-Q plot but this step isn't necessary (if reported, a Normal Q-Q plot of transformed quantile residuals might confuse readers who failed to read the fine print).

Misconceivable – A common misconception is that if the distribution of the response approximates a Poisson distribution, then the residuals of a GLM fit with a Poisson distribution should be normally distributed, which could then be checked with a Normal Q-Q plot, and homoskedastic, which could be checked with a scale-location plot. Neither of these is true because a GLM does not transform the data and, in fact, the model definition does not specify anything about the distribution of an "error" term – there is no ε in the model defintion above! This is why thinking about the definition of a linear model by specifying an error term with a normal distribution can be confusing and lead to misconceptions when learning GLMs.

Quantile Residual Q–Q Plot

The Q-Q plot using quantile residuals with a Poisson distribution indicates that the counts of *Diplostomum* larvae in the eyes of the threespine stickleback are not well approximated by a Poisson distribution – there are too many observed values near the ends of the expected tails, indicating the expected values are not spread out enough. This pattern emerges because the observed counts are overdispersed compared to a Poisson distribution.

16.2.6.2 Model checking a GLM II – a dispersion plot

If observed counts are Poisson distributed, then the **Pearson residuals** (r_i) and the residual degrees of freedom of the fit model (df) can be used to compute a dispersion statistic

$$\frac{\sum r_i}{df} \tag{16.17}$$

that has an expected value of 1. Instead of a formal hypothesis test of this statistic, I use a simulation approach and ask, "if the observed counts are Poisson distributed, what is the expected frequency distribution of this dispersion statistic?" and then use simulation to generate this expected distribution. The algorithm for this is

1. For each observation i, generate a random Poisson count using $\hat{\mu}$ as the parameter.
2. Fit the model and compute the dispersion statistic.
3. Repeat 1 and 2 N_{iter} times.

The plot below shows a histogram of the dispersion statistic computed for 1000 simulations of the worm data. The observed dispersion statistic is 3.4. The expected value is 1.0. The mean of the simulated values is 1.

16.2.7   Fitting a GLM with a Negative Binomial distribution to the worm data

The model is

Observed vs. expected dispersion statistic

Figure 16.3: Observed vs. expected dispersion statistic. The observed statistic marked by the red line. The histogram of expected statistics are from 1000 simulations of the observed data.

$$Diplo\_intensity \sim NB(\mu, k) \tag{16.18}$$
$$\mathrm{E}(Diplo\_intensity|Treatment) = \mu \tag{16.19}$$
$$\mu = \exp(\eta) \tag{16.20}$$
$$\eta = \beta_0 + \beta_1 Uninfected + \beta_2 Infected\_LG + \beta_3 Infected\_HG \tag{16.21}$$

This model specifies a negative binomial distribution but otherwise is just like that above specifying a Poisson distribution.

16.2.7.1   Model checking

Quantile Residual Q–Q Plot

A quantile residual Q-Q plot of the GLM model fit with negative binomial distribution is illustrated above. This looks pretty good.

16.2.7.2   Model means and coefficients

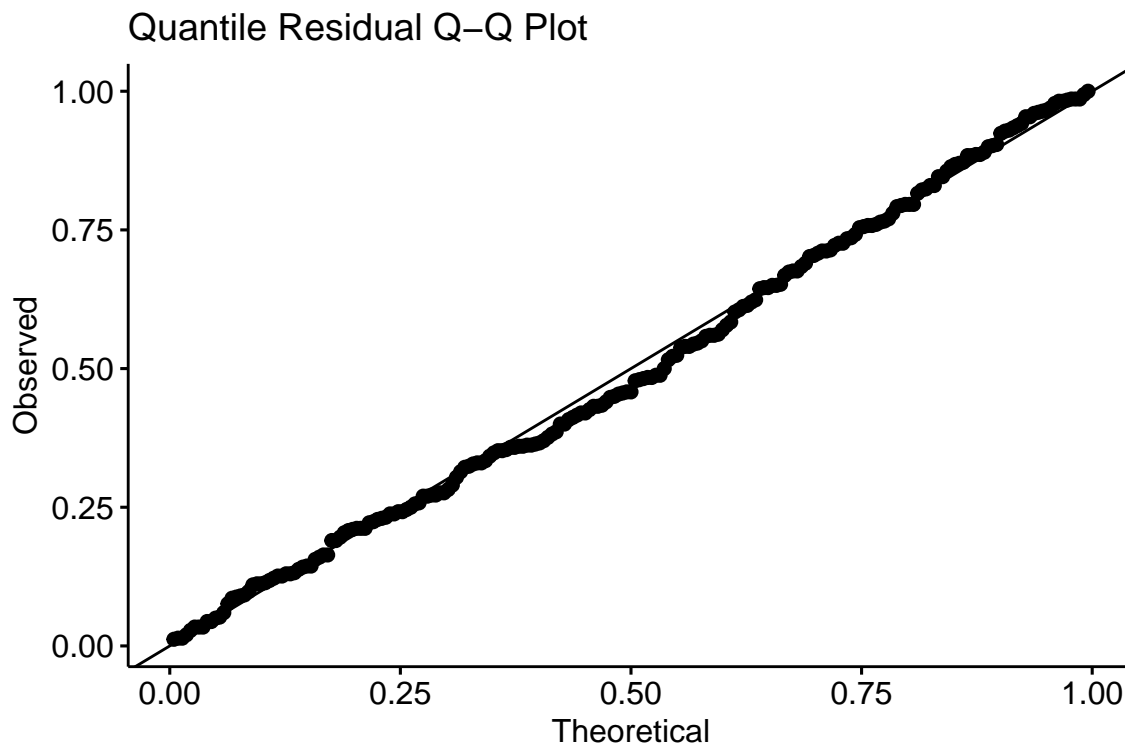In a Generalized Linear Model of counts using either a Poisson or negative binomial distribution, modeled means, coefficients, and contrasts can be reported either on the link or response scale. Remember, the response scale is a count, while the link scale is a log(count).

The modeled means on the link scale are

```
## Treatment    emmean      SE  df asymp.LCL asymp.UCL
## Control       1.82 0.0804 Inf      1.66      1.98
## Uninfected    1.50 0.1093 Inf      1.29      1.72
## Infected LG   1.62 0.1362 Inf      1.36      1.89
## Infected HG   2.36 0.0714 Inf      2.22      2.50
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
```

While the means on response scale are

```
## Treatment    response     SE  df asymp.LCL asymp.UCL
## Control          6.18 0.497 Inf      5.28      7.24
## Uninfected       4.50 0.492 Inf      3.63      5.58
## Infected LG      5.07 0.691 Inf      3.89      6.63
## Infected HG     10.60 0.757 Inf      9.22     12.20
##
## Confidence level used: 0.95
```

```
## Intervals are back-transformed from the log scale
```

1. A mean on the response scale is simply the exponent of the mean on the link scale. For example, the mean of the Control treatment level on the response scale is $\exp(1.821408) = 6.180555$.

2. The CIs on the link scale are symmetric around the mean but those on the response scale are not. This is a feature, not a bug. Remember that counts are right skewed which means a CI will have a wider right than left interval. Check this!

3. If a plot includes a 1 SE error bar on the response scale, this is technically correct but it encourages the practice of computing CIs using the 2*SE rule of thumb. This rule breaks down for count data with right skewed distributions.

4. Plotting the response scale CIs is both technically correct and makes the 2*SE rule of thumb unnecessary.

The model coefficients on the link scale are

```
##  contrast              estimate    SE  df asymp.LCL asymp.UCL z.ratio
##  Uninfected - Control    -0.317 0.136 Inf    -0.583   -0.0514  -2.339
##  Infected LG - Control   -0.197 0.158 Inf    -0.507    0.1126  -1.248
##  Infected HG - Control    0.540 0.108 Inf     0.329    0.7504   5.019
##  p.value
##  0.0193
##  0.2122
##  <.0001
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
```

Backtransforming the coefficients (but not the intercept) to the response scale (using $\exp b_j$) results in a **response ratio**.

```
##  contrast               ratio     SE  df asymp.LCL asymp.UCL z.ratio
##  Uninfected / Control   0.728 0.0988 Inf     0.558      0.95  -2.339
##  Infected LG / Control  0.821 0.1298 Inf     0.602      1.12  -1.248
##  Infected HG / Control  1.715 0.1845 Inf     1.389      2.12   5.019
##  p.value
##  0.0193
##  0.2122
##  <.0001
##
## Confidence level used: 0.95
## Intervals are back-transformed from the log scale
## Tests are performed on the log scale
```

1. Note how the emmeans package reports the name of the term as the ratio of the coefficient term to the intercept term (the reference treatment level). Why are the coefficients tranformed to ratios on the response scale? Remember that a coefficient is a difference in conditional means and that $\exp(B - A) = \frac{\exp(B)}{\exp(A)}$. For a dummy variable as here (say "Infected HG"), the response ratio is

$$RR_{Infected\_HG} = \frac{\overline{Infected\_HG}}{\overline{Control}} \tag{16.22}$$

which give us the **relative effect** of Infected_HG compared to the Control. Relative effects could be reported as a response ratio in a table, or in the text it could be reported as a percent "Infected HG fish had 71.5% (95%CI: 38.9% - 111.8%) more *Diplostomum* larvae than Control fish." Where do these percents come from?

The percent effect is $100(RR_j - 1)$ larger than the reference mean if the $RR_j > 1$ or $100(1 - RR_j)$ smaller than the reference mean if the $RR_j < 1$.

2. Backtransforming the intercept does not generate a ratio since the intercept on the link scale is not a difference. For the worm analysis, the intercept on the link scale is the mean count of the control group on the link scale and the backtransformed intercept is the mean count of the control group on the response scale.

3. Effects on the response scale are not additive but multiplicative! So, for example, the mean of the Infected HG treatment level on the response scale is $\overline{Control} * RR_{Infected\_HG}$ (remember that with a linear model the mean would be $b_{Control} + b_{Infected\_HG}$). Check and see if this works.

16.3   Working in R

Source publication: Benesh, D. P., & Kalbe, M. (2016). Experimental parasite community ecology: intraspecific variation in a large tapeworm affects community assembly. Journal of Animal Ecology, 85(4), 1004-1013.

Source data URL: https://datadryad.org/resource/doi:10.5061/dryad.bq8j8

Source file: "Lab_exp.csv"

Poisson fit. A quantile residual Q-Q plot can be generated using the package DHARMa

```
fit.pois <- glm(Diplo_intensity ~ Treatment, family="poisson", data=worm)

# from the DHARMa package
  n_sim <- 250
  simulationOutput <- simulateResiduals(fittedModel = fit.pois, n = n_sim)
  plot(simulationOutput, asFactor = F)
```


DHARMa scaled residual plots

A plot of the dispersion statistic can be generated using the object returned by the `SimulateOutput` function but with `refit = TRUE`, which refits a model each iteration. This refitting isn't necessary if only the quantile

residuals are needed. The Dharma package does not divide the sum of squared Pearson residuals by the residual degrees of freedom and so the expected value of the statistic is df.

```r
# from the DHARMa package
  n_sim <- 250
  simulationOutput <- simulateResiduals(fittedModel=fit.pois, n=n_sim, refit=TRUE)
  testDispersion(simulationOutput)
```

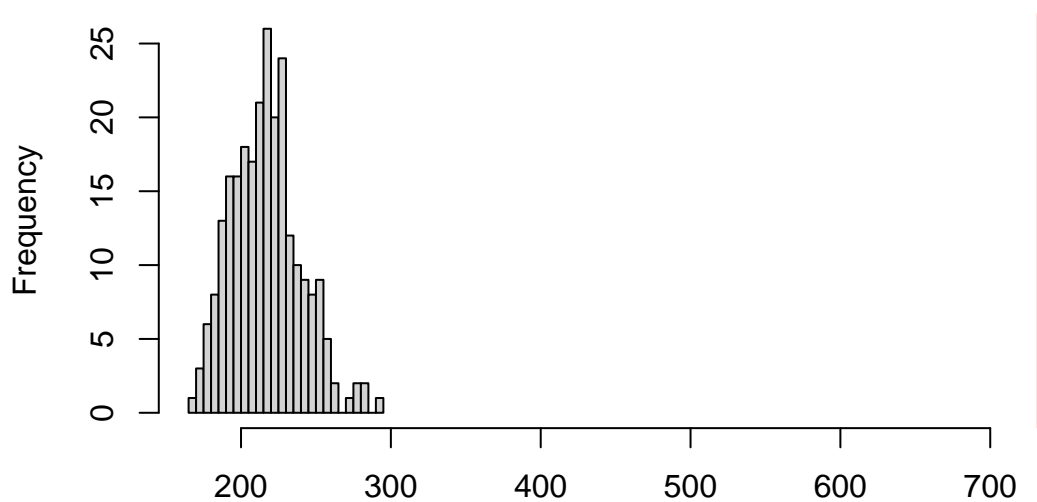## DHARMa nonparametric dispersion test via mean deviance residual fitted vs. simulated−refitted

Simulated values, red line = fitted model. p−value (two.sided) = 0

```
##
##  DHARMa nonparametric dispersion test via mean deviance residual
##  fitted vs. simulated-refitted
##
## data:  simulationOutput
## dispersion = 3.3788, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

Negative binomial fit.

```r
fit.nb <- glm.nb(Diplo_intensity ~ Treatment, data=worm)
# from the DHARMa package
  simulationOutput <- simulateResiduals(fittedModel = fit.nb, n = n_sim)
  plot(simulationOutput, asFactor = F)
```

DHARMa scaled residual plots

QQ plot residuals

**Residual vs. predicted
lines should match**

KS test: p= 0.789
Deviation n.s.

Outlier test: p= 1
Deviation n.s.

Observed

Standardized residual

Expected

Predicted values (rank transformed)

```r
# link scale
emm <- emmeans(fit.nb, specs="Treatment")
emm
```

```
## Treatment     emmean      SE  df asymp.LCL asymp.UCL
## Control         1.82 0.0804 Inf      1.66      1.98
## Uninfected      1.50 0.1093 Inf      1.29      1.72
## Infected LG     1.62 0.1362 Inf      1.36      1.89
## Infected HG     2.36 0.0714 Inf      2.22      2.50
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
```

```r
summary(contrast(emm, method="trt.vs.ctrl", adjust="none"), infer=c(TRUE, TRUE))
```

```
## contrast               estimate     SE  df asymp.LCL asymp.UCL z.ratio
## Uninfected - Control     -0.317 0.136 Inf    -0.583   -0.0514 -2.339
## Infected LG - Control    -0.197 0.158 Inf    -0.507    0.1126 -1.248
## Infected HG - Control     0.540 0.108 Inf     0.329    0.7504  5.019
## p.value
## 0.0193
## 0.2122
## <.0001
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
```

```r
# response scale
emm.response <- emmeans(fit.nb, specs="Treatment", type="response")
summary(contrast(emm, method="trt.vs.ctrl", adjust="none", type="response"), infer=c(TRUE, TRUE))
```

```
## contrast              ratio      SE  df asymp.LCL asymp.UCL z.ratio
## Uninfected / Control  0.728 0.0988 Inf     0.558      0.95 -2.339
## Infected LG / Control 0.821 0.1298 Inf     0.602      1.12 -1.248
```

```
##  Infected HG / Control 1.715 0.1845 Inf     1.389     2.12  5.019
##  p.value
##  0.0193
##  0.2122
##  <.0001
##
## Confidence level used: 0.95
## Intervals are back-transformed from the log scale
## Tests are performed on the log scale
```

16.4 Problems

Analyze the data that went into Fig 6B of Tena, A., Pekas, A., Cano, D., Wäckers, F. L., & Urbaneja, A. (2015). Sugar provisioning maximizes the biocontrol service of parasitoids. Journal of Applied Ecology, 52(3), 795-804.

1. Compute contrasts and CIs among all pairs of all three treatment levels
2. Make a better plot like 6b including 1) use the modeled mean instead of the simple group mean and 2) use the modeled CI of the mean instead of the SE computed within each group independently.

source URL: https://datadryad.org/resource/doi:10.5061/dryad.bj001

source file: "4_Parastism_Fig_6.csv"

Chapter 17

Linear mixed models

17.1 Random effects

Researchers often collect data in batches, for example

1. Researchers interested in the effects of insectivorous birds on tree seedling performance in a forest stake out ten 1 m^2 plots and use a wire-mesh cage to cover half of each plot [1]. The cage allows insect herbivores into the seedlings inside but excludes insectivorous birds that eat the insects from the seedlings. In every plot, five seedlings are planted within the exclosure and five outside of the exclosure. At the end of the experiment, the total leaf mass is measured on all seedlings. Small, uncontrolled, environmental factors (including soil factors and density of insectivorous birds) will differ between plots but will be common to all seedlings within a plot and we would expect a common response to this uncontrolled variation on top of the differential response to each treatment. As a consequence, the ten measures of leaf mass within a plot are not independent.
2. A nutrition researcher wants to compare the effect of glucose vs. fructose on glucose metabolism in humans. Ten individuals are recruited. Each individual has blood insulin measured 60 minutes after a noon meal over six successive days. The meal alternates between high glucose and high fructose on each day. Each individual has three measures under high glucose treatment and three measures under high fructose treatment. Small, uncontrolled, environmental factors (including metabolic variation, other meals, activity levels) will differ between the individuals but be common within an individual and we would expect a common response to this uncontrolled variation on top of the differential response to each treatment. As a consequence, the six measures of insulin within an individual are not independent.
3. A researcher is using a mouse model to compare growth of a wildtype and engineered mutant strain of *Staphylococcus*. A small spot on both right and left forelimbs of ten mice is shaved and abraded. The two strains are randomly assigned to a side (so each mouse is infected with each strain). Small, uncontrolled, environmental factors (including immune responses) will differ between the mice but be common between the two limbs within a mouse and we would expect a common response to this uncontrolled variation on top of the differential response to each treatment. As a consequence, the two measures of growth within a mouse are not independent.
4. An ecologist wants to measure the effect of an invasive plant on the reproduction of a native plant. They stake-out ten 2 m^2 plots in a forest and divide each plot into four quadrants, with each quadrant assigned a different treatment: control, activated carbon (a procedural control), extract from the invasive plant's leaves, and both activated carbon and extract from the invasive plant's leaves. The response is seedling count. Small, uncontrolled, environmental factors (including soil, drainage, and light) will differ between plots but will will be common to all four quadrants within a plot and we would expect a common response to this uncontrolled variation on top of the differential response to

---

[1]Giffard, B., Corcket, E., Barbaro, L., & Jactel, H. (2012). Bird predation enhances tree seedling resistance to insect herbivores in contrasting forest habitats. Oecologia, 168(2), 415-424

each treatment. As a consequence, the four sets of counts within a plot are not independent.

5. An ecologist wants to measure the effect of habitat on chick growth in a bird. Five individuals nest in artifical nest boxes built on the boundary between the forest and a large, agricultural field. Five other individuals nest in boxes built deep in the interior of the forest. Chicks in each nest are weighed 13 days after hatching. Small, uncontrolled, environmental factors (including parenting, food availablity, temperature, etc.) will differ between the nests but be common within the nests and we would expect a common response to this uncontrolled variation on top of the differential response to each treatment. As a consequence, the measures of weight within a nest are not independent.

6. A physiologists has skeletal muscle cells growing in 5 control cultures, and 5 treated cultures. The *Y* variable is cell diameter, which is measured in 10 cells per culture. Small, uncontrolled, environmental factors (including chemical) will differ between cultures but will be common to all cells within a culture and we would expect a common response to this uncontrolled variation on top of the differential response to each treatment. As a consequence, the ten measures of diameter within a culture are not independent.

7. A behavioral biologist wants to measure the effect of a predator fish on the preferred feeding location (open water or vegetation) of a prey fish. Ten tanks are set up with equal amounts of vegetated and unvegetated area. One-third of each tank is screened off to house a predator fish, which are added to five of the tanks. Ten prey fish are added to each tank. The response is minutes spent foraging in the open water as a fraction of total time foraging, which is measured in each fish in each tank. Small, uncontrolled, environmental factors (including temperature, water chemistry, light, and fish behavior) will differ between the tanks but be common within tanks and we would expect a common response to this uncontrolled variation on top of the differential response to each treatment. As a consequence, the ten measures of foraging of each fish within a tank are not independent.

The batches – plots in experiment 1, individuals in experiment 2, mice in experiment 3, plots in experiment 4, nests in experiment 5, cultures in experiment 6, and tanks in experiment 7 – are often referred to as **blocks** or **clusters**. I'll generally use the term "block" in this book. The blocks are the experimental units, meaning that it is at this level that the experimenter is controlling the conditions. There is systematic variation at two levels: among treatments due to treatment effects and among blocks due to **block effects**. This among-block variation is the **random effect**. An assumption of modeling random effects is that the blocks (plots/individuals/mice/nests/cultures/tanks) are a random sample of the blocks that could have been sampled. This is often not strictly true as blocks are often **convenience samples**.

The multiple measures within a block are often called **repeated measures**, especially if the block is an experimental animal such as a mouse or human. If multiple measures within a treatment level within a block (that is, within a *block × treatment* combination) are taken over time, the data are **longitudinal**. Sometimes in cell biology, the subsamples within a treatment within a block are called "replicates", as they are replicates of this *block × treatment* combination, but this can be confusing because the treatments are replicated at the level of the block and not at the level of the subsamples within a treatment by block combination. The blocks are the independent experimental units. Instead the multiple measures of the response within a *block × treatment* combination are **subsamples**.

Experiments 1 and 2 are examples of a **complete randomized block with subsampling** design. "Complete" means that each block has all treatment levels or combinations of levels if there is more than one factor. Experiments 3 and 4 are examples of a **complete randomized block** design. The blocks are complete but there is only one measure of the response per treatment. Experiments 5, 6, and 7 are examples of an **incomplete randomized blocks** design. The blocks are incomplete because they do not contain less than all treatment levels and combinations. In these examples, each block contains only one treatment level.

17.2 Random effects in statistical models

In all of the above examples, the researcher is interested in the treatment effect but not the variation due to differences among the blocks. The blocks are nuissance factors that add additional variance to the response, with the consequence that estimates of treatment effects are less precise, unless the varriance due to the

Table 17.1: The linear mixed model specified above estimates a fixed intercept and fixed slope (treatment effect) that are common to all observations and a random intercept and random slope for each block, each of which is common among observations within a block but differ among observations in different blocks.

block	b_0	b_{0j}	b_1	b_{1j}
1	b_0	$b_{0,j=1}$	b_1	$b_{1,j=1}$
2	b_0	$b_{0,j=2}$	b_1	$b_{1,j=2}$
3	b_0	$b_{0,j=3}$	b_1	$b_{1,j=3}$
4	b_0	$b_{0,j=4}$	b_1	$b_{1,j=4}$
5	b_0	$b_{0,j=5}$	b_1	$b_{1,j=5}$
6	b_0	$b_{0,j=6}$	b_1	$b_{1,j=6}$

blocks is explicitly modeled. Including block structure in the design and in the statistical model is known as **blocking**. A natural way to think about the block factor is as a **random effect**, meaning that plots in experiment 1 or the mice in experiment 3 are simply random samples from a population of plots or mice. Modeling this using the residual-error specification looks like

$$y_{ij} = (\beta_0 + \beta_{0j}) + (\beta_1 + \beta_{1j})x_i + \varepsilon_i \tag{17.1}$$

where i indexes the observation and j indexes the block (culture, plot, mouse, etc). The intercept parameter β_{0j} is a **random intercept** and the slope parameter β_{1j} is a **random slope**. The random intercept has a **fixed** component (β_0) that is common to all observations and a random component (β_{0j}) that is common within a block but differs among blocks (see table below). In the above equation, I've used parentheses to show how these combine into the random intercept that is unique for each block. Similarly, the random slope (treatment effect) has a fixed part (β_1) that is common to all observations and a random component (β_{1j}) that is common within a block but differs among blocks (see table below). Again, these are collected within a pair of parentheses in the equation above.

boundary (singular) fit: see ?isSingular

Linear mixed models are called "mixed models" because they are a mix of fixed and random components. Another useful way to specify this model is to think about it hierarchically, using

$$y_{ij} = \beta_{0j} + \beta_{1j}x_i + \varepsilon_i \tag{17.2}$$
$$\varepsilon_i \sim N(0, \sigma) \tag{17.3}$$
$$\beta_{0j} = \beta_0 + N(0, \sigma_0) \tag{17.4}$$
$$\beta_{1j} = \beta_1 + N(0, \sigma_1) \tag{17.5}$$

The first line states that the response is a function of a block-specific intercept and a block specific slope plus some error that is unique to each observation. The third and fourth lines state that these block-specific effects are themselves a function of a common effect and a random term that is unique to each block. That is, we have a hierarchical or multi-level structure to the model. Line 1 is the top level and the effects that are specified in line 1 are a function of effects at a second, lower level, which are specified in lines 3 and 4. Because of this structure, linear mixed models are sometimes called hierarchical or multi-level models.

Finally, it's useful to think how to specify a linear mixed model using the random-draw specification, as this leads naturally to generalized linear mixed models, or GLMMs.

$$y_{ij} \sim N(\mu_{ij}, \sigma) \tag{17.6}$$
$$\mu_{ij} = \beta_{0j} + \beta_{1j} x_i \tag{17.7}$$
$$\beta_{0j} \sim N(\beta_0, \sigma_0) \tag{17.8}$$
$$\beta_{1j} \sim N(\beta_1, \sigma_1) \tag{17.9}$$

17.3 Linear mixed models are flexible

The linear mixed model in Equation (17.1) specifies both a random intercept and a random slope but a researcher might limit the random effect to the intercept only, or less commonly, the slope only. Excluding the random slope from Equation (17.1) results in the model

$$y_{ij} = (\beta_0 + \beta_{0j}) + \beta_1 x_i + \varepsilon_i \tag{17.10}$$

We might use a random-intercept-only model if we think that features of the block would effect the mean response among blocks but not effect the difference in treatment level (or treatment effect) among blocks. For example, differences in the immune systems among the individual mice in experiment 3 might effect growth in both the wild-type and engineered strains of staph but won't effect the difference in growth between wild-type and engineered strains from one mouse to another.

Not more than you should know – For more complex mixed models, matrix algebra makes the specification of the model much more manageable than the scalar algebra in **??**.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{17.11}$$

where $\mathbf{y}$ is the vector of the response, $\mathbf{X}\boldsymbol{\beta}$ is the linear predictor of fixed effects and $\mathbf{Z}\mathbf{u}$ is the linear predictor of random effects. $\mathbf{X}$ is the model matrix for the fixed effects and $\boldsymbol{\beta}$ is the vector of fixed-effect terms (the fixed part of the intercept (β_0), including the fixed-effect coefficients for each of the

17.4 Visualizing block effects

To visualize random effects due to block, Let's create fake data that look something like experiments 1 or 2, with a single factor with two treatment levels, $k = 10$ blocks, and $n = 3$ measures for each treatment level within each block. This is a randomized complete block design with subsampling and has a total of $N = 2 \times k \times n$ measures of Y (and rows of the data.table).

Figure 17.1A shows the response as a function of treatment. The responses are nicely symmetric around the treatment means (the blue and yellow lines). A linear model (and generalized linear models, more generally) assumes that a response, conditional on the X, are independent. Figure 17.1B shows how this assumption is violated for the simulated data. That pattern of residuals within a block around the treatment means does not look at all random. Instead, there is a distinct pattern within a block for the points to cluster either below the treatment means or above the treatment means. In blocks a, b, c, g, and h, all or most of the responses are below their treatment mean (for example in block a, all the blue points are below the blue line and 2 of 3 yellow points are below the yellow line). In blocks, e, f, i, and j, all or most of the responses are above their treatment mean (for example, in block i, all three yellow points are above the yellow line and 2 of 3 blue points are above the blue line). In other words, the responses within a block covary together. For a linear model, this is known as **correlated error**.
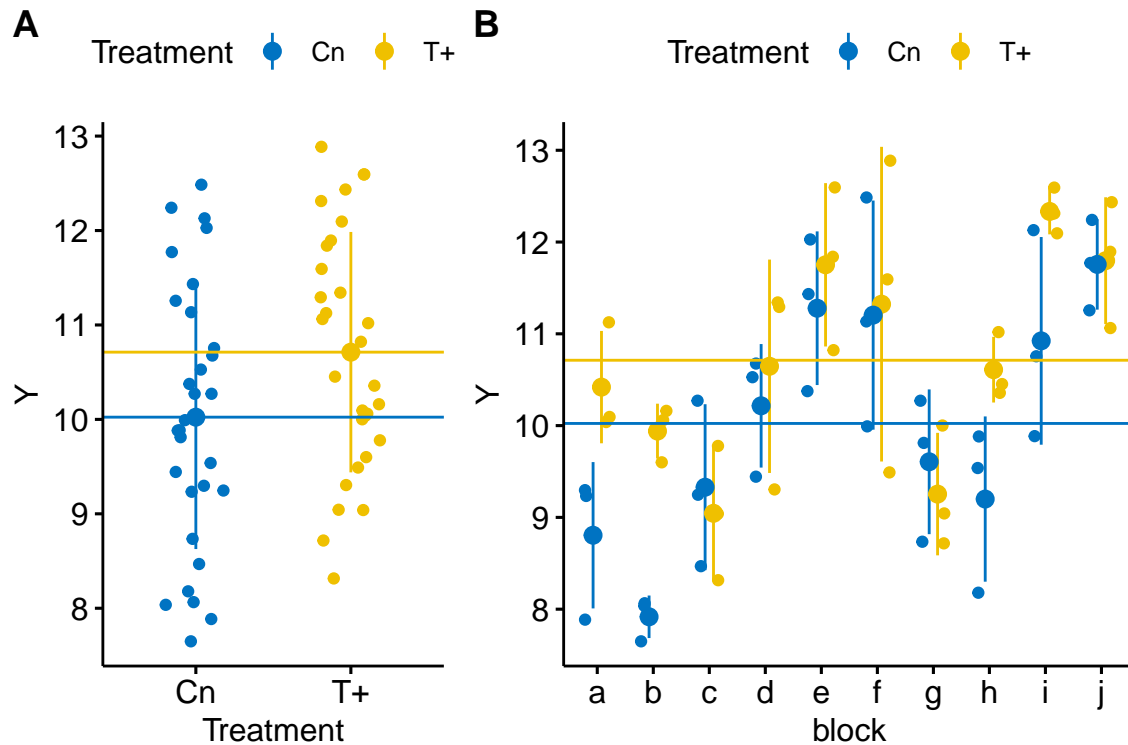
Figure 17.1: Visualizing random effects. A) The response in the two treatment levels. B) The same data but separated by block. The blue line is at the control mean and the yellow line is at the treated mean. The black dots are the mean response within a block.

17.5 Linear mixed models can increase precision of point estimates

Block effects are differences in expected mean response among blocks due to unmeasured factors that are shared within blocks but not among blocks. A classical linear model fails to model this component of the total variance in the response, and as a consquence, this block-specific variance becomes part of the error variance. One way to visualize this is by moving the random intercept and random slope components of equation (17.1) to the right and combining it with the observation-specific error

$$y_{ij} = \beta_0 + \beta_1 x_i + (\beta_{0j} + \beta_{1j} + \varepsilon_i) \tag{17.12}$$

% which shows that the random effects β_{0j} and β_{1j} are modeled as error in a linear model. As a consequence, the residual variance is larger and, therefore, the standard errors of point estimates, including means, coefficients of the model, and contrasts from the model, are larger. Here is the table of model coefficients of the data in Figure 17.1 fitted using a classical linear model

```
##               Estimate Std. Error  t value      Pr(>|t|)
## (Intercept) 10.0235749  0.2439544 41.08791 1.393185e-44
## TreatmentT+  0.6884787  0.3450036  1.99557 5.068466e-02
```

and a linear mixed model

```
##               Estimate Std. Error       df  t value      Pr(>|t|)
## (Intercept) 10.0235749  0.3954736 8.999737 25.345750 1.114804e-09
## TreatmentT+  0.6884787  0.2700797 9.000052  2.549169 3.123718e-02
```

The linear mixed model has increased precision (a smaller SE of the estimates) because it estimates the

value of β_{0j} and β_{1j} for each block. The linear model does not estimate these parameters and the variance in these parameters is swept into the residual variance.

NHST blues – A **paired t-test** is equivalent to the special case of a linear mixed model with a single factor with two treatment levels, k blocks, and a single measure of each treatment level within each block. A good example is the wild type vs. engineered staph count in mice in experiment 3 above. A linear mixed model is much more flexible than a paired t-test because it allows a researcher to add treatment levels, additional factors, and covariates to the model. In addition, a linear mixed model can handle missing data.

Here is fake data similar in design to experiment 3 with a single factor with two treatment levels and both levels applied to the same experimental unit.

```
set.seed(2)
n <- 10 # number of mice (blocks)
x <- rep(c("WT","T+"), each=n) # treatments
id <- rep(letters[1:n], 2) # block id
y <- c(rnorm(n, mean=10), rnorm(n, mean=11))
fake_data <- data.table(Y=y, X=x, ID=id)
```

The t-test p-value is

```
t.test(Y~X, data=fake_data, paired=TRUE)$p.value
```

```
## [1] 0.05336815
```

and the coefficient table of the fixed effect in the linear mixed model is

```
coef(summary(lme(Y~X, random = ~1|ID, correlation=corCompSymm(form=~1|ID), data=fake_data)))
```

```
##                 Value Std.Error DF   t-value       p-value
## (Intercept) 11.1797704 0.3438775  9 32.510914 1.212113e-10
## XWT         -0.9686188 0.4358740  9 -2.222245 5.336815e-02
```

17.6   Linear mixed models are used to avoid pseudoreplication

17.7   Linear mixed models shrink coefficients by partial pooling

In experiment 1 above, there are 10 sites (maybe different woodlots). In each plot, five seedlings are planted inside a cage and five outside the cage. The cage excludes insectivorous birds but not herbivorous insects. The researchers are investigating how birds affect plant growth indirectly – by eating insects that feed on the plants. The response is total leaf area in each seedling.

Let's say we want to know the treatment effect in each of these sites. There are several ways of estimating this.

1. Fit k separate models, one for each site. The intercept (control mean) and slope (treatment effect) parameters for each site are estimated independently from all other sites. Conequently, the model parameters are computed using **no pooling**. For the estimation of the β terms, this is equivalent to a single, factorial linear model with *Site* modeled as a **fixed effect** (this is not true for the estimate of the standard errors of these terms since these are computed from the residual sum of squares of the model. For balanced data, all of the "intercept" or "slope" terms will have the same SE in the factorial analysis but differ among the k independent analyses).

2. Fit a linear model to all the data combined as if these were from a single site, and assign the intercept and treatment effect paramters to all sites. The model parameters are computed using **complete pooling**.
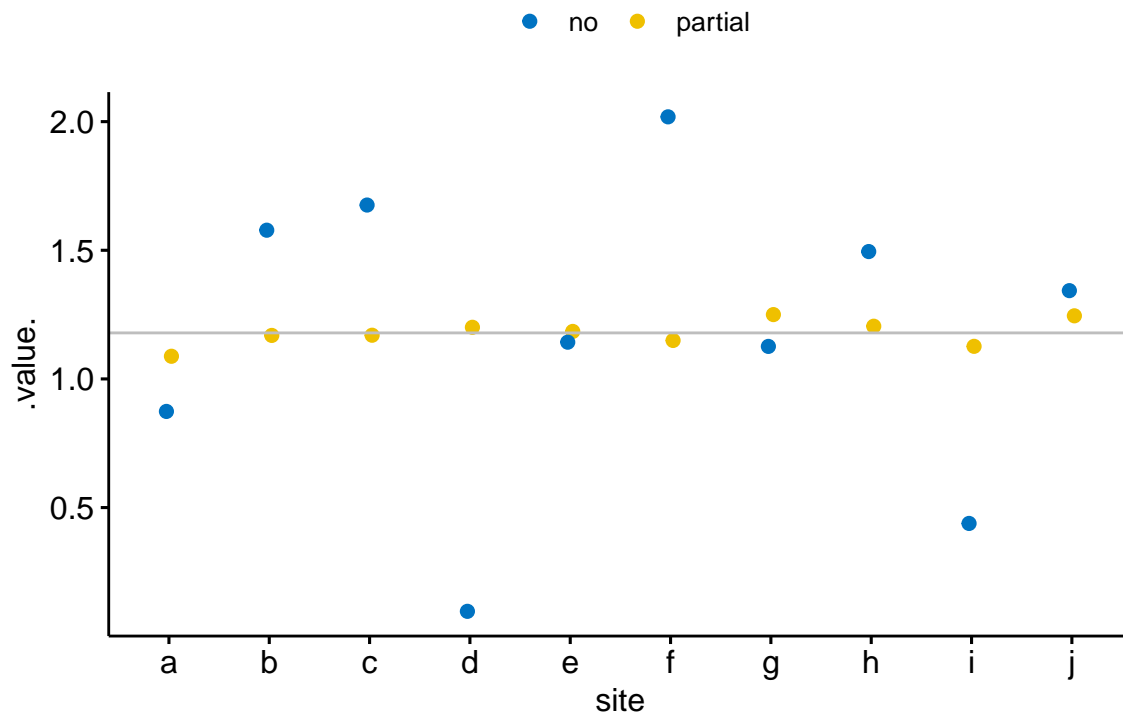
Figure 17.2: Shrinkage estimates of the treatment effect in a linear mixed model. The grey line is the estimate using complete pooling (so there is only one estimate which is assigned to each site). In general, the partial-pooling (linear mixed model) estimates (yellow) are generally closer to the complete pooling estimate than the no-pooling (separate linear models) estimates (blue). More specifically, if the no-pooling estimate is far from the complete pooling estimate, the partial pooling estimate is *much* closer to the complete pooling estimate. The consequence of partial pooling is that the differences among the estimates are shrunk toward zero.

3. Fit a linear mixed model to all the data, using site as a random factor to estimate both random intercepts and slopes. Similar to the no-pooling analysis, there will be different intercept (control mean) and slope (treatment effect) estimates for each site, but unlike the no-pooling analysis, these estimates are computed by combining information from the other sites. The information used to estimate parameters in a linear mixed model is somewhere in between no pooling and complete pooling and is sometimes called **partial pooling**.

The consequence of partial pooling in a linear mixed model is that site intercepts (control means) are pulled toward the single intercept in the complete-pooling analysis and the site slopes (treatment effects) are pulled toward the single slope in the complete-pooling analysis. This has the consequence that the **differences** in parameter estimates among sites are shrunk toward zero. A consequence of this shrinkage is that the variance of the intercept estimates or of the slope estimates is smaller than that in the no-pooling analysis. Figure 17.2 shows this shrinkage effect using fake data simulating the seedling experiment.

The linear mixed model estimates of the treatment effects for each site are a type of **shrinkage estimate** and a linear mixed model is one kind of **shrinkage estimator**. Shrinkage estimates have fascinating properties:

1. the variance of shrinkage estimates is less than that of ordinary least squares estimates (no-pooling, or using the block as a fixed factor)
2. shrinkage estimates are **biased** but the OLS estimates are not. This means that the expected value of a coefficient from the linear mixed model *does not equal* the true (parameter) value! Or, more formally, $E(b_j) \neq \beta_j$.
3. the **mean square error** of shrinkage estimates will be smaller than that for OLS estimates.

Table 17.2: Effect of new expression vector on protein production in ten strains of *E. coli* using a fixed effect factorial model and linear mixed model.

Strain	β_{1j}	fixed b_{1j}	random b_{1j}
a	0.91	1.07	0.98
b	0.87	0.94	0.85
c	0.90	-1.03	0.30
d	0.81	0.64	0.63
e	1.09	1.00	1.07
f	0.62	0.91	1.14
g	1.33	2.26	1.36
h	1.27	1.48	0.96
i	1.61	0.57	1.13
j	0.89	1.50	0.93

The first property was discussed above and shown in Figure 17.2. The second property raises the question, if we want to estimate the treatment effects within each site, why would we ever want to use *Site* as a random instead of fixed effect? The answer is the third property, which can be summarized as, "if we were to replicate the experiment many times, the shrinkage estimates will be, on average, less wrong (or closer to the true value) than the OLS estimates, where"wrong" is the absolute deviation from the true value."

When shrinkage estimators were first discovered, the third property surprised stasticians. The third property has profound consequences. Consider a scenario where researchers want to compare the performance of a new expression vector to that of an existing expression vector on protein production using *E. coli*. The researchers have ten different *E. coli* strains and are interested in strain-specific effects because they will choose the three strains with the largest effects for further testing. The researchers measure the response of each strain five times.

The table above shows the true strain-specific effect and both the fixed (OLS) and random (LMM) effect estimates. The largest OLS estimate is 70% larger than the true effect and the strain with the largest true effect is not among the top three biggest OLS estimates (its ranking is 9/10). By contrast, the LMM estimates are closer to the true effects and the top strain is among the three largest LMM estimates.

These results are specific to these fake data but more generally, 1) the largest OLS estimates are inflated (larger error from the true effect), relative to the largest LMM estimates 2) overall, the LMM estimates will be closer than the OLS estimates to the true effects

To understand this, rank order the treatment effects for each strain. An individual strain's position in this rank is the sum of the true effect for that strain and some random error. Because OLS, relative to shrinkage estimates, have greater variance in the estimate (that is, the random error component is bigger), the biggest effects estimated by OLS are more likely to be big because of the error component, compared to shrinkage estimates.

Not more than you want to know – Shrinkage estimators are not only useful when we are interested in block-specific effects but are also useful for estimating effects when there are **multiple responses**. For example, consider a researcher interested in measuring the effects of some exercise treatment on gene expression in adipose cells. The researcher measures expression levels in 10,000 genes. Given the typical content in undergraduate biostatics courses, a researcher would probably model these responses using 10,000 *t*-tests, or equivalently, 10,000 separate linear models. If the tests were ranked by *p*-value or absolute effect size, many of the genes with largest absolute effect would be there because of a large error component and many of the largest effects would be massively overinflated. Re-imagining the design as a single, linear mixed model with each gene modeled as a block would lead to a rank order in which the biggest measured effects more closely approximate the true effects.

17.8 Working in R

The major function for working with linear mixed models is `lmer()` from the lme4 package. An older, and still sometimes used and useful function is `lme()` from the nlme package. The authors of the lme4 package argue that the df in a linear mixed model are too approximate for a useful *p*-value and, consequently, the `lme` function does not return a *p*-value. Many biological researchers want a *p*-value and typically use the `lmerTest` package to get this.

17.8.1 coral data

Source Zill, J. A., Gil, M. A., & Osenberg, C. W. (2017). When environmental factors become stressors: interactive effects of vermetid gastropods and sedimentation on corals. Biology letters, 13(3), 20160957.

Dryad source https://datadryad.org/resource/doi:10.5061/dryad.p59n8

file name "VermetidSedimentData_ZillGilOsenberg_DRYAD.xlsx"

```r
folder <- "Data from When environmental factors become stressors- interactive effects of vermetid gastro
fn <- "VermetidSedimentData_ZillGilOsenberg_DRYAD.xlsx"
sheet_i <- "Coral Growth Rate Data"
file_path <- paste(data_path, folder, fn, sep="/")
coral <- data.table(read_excel(file_path, sheet=sheet_i))
setnames(coral, old=colnames(coral), new=clean_label(colnames(coral)))
coral[, Vermetids:=factor(Vermetids)]
coral[, Sediment:=factor(Sediment)]
```

`lmer` adds the random component to the formula. `lme` adds the random component as a separate argument

```r
# to reproduce the results
# observation 2 should be excluded from the analysis
inc <- c(1, 3:nrow(coral))

# specification using lmer
# random intercept only
fit.lmer1 <- lmer(GrowthRate ~ Vermetids*Sediment + (1|Block), data=coral[inc])
# random intercept and slope
fit.lmer2 <- lmer(GrowthRate ~ Vermetids*Sediment + (Vermetids|Block) + (Sediment|Block), data=coral[inc
```

```
## boundary (singular) fit: see ?isSingular
```

```r
# random intercept and slope
fit.lmer3 <- lmer(GrowthRate ~ Vermetids*Sediment + (Vermetids|Block) + (Sediment|Block), data=coral[inc
```

```
## boundary (singular) fit: see ?isSingular
```

```r
# to include the interaction as a random effect we'd need subsampling within each factorial treatment c

# specification using lme
fit.lme <- lme(GrowthRate ~ Vermetids*Sediment, random= ~1|Block, data=coral[inc])

# results using lmer fit
coefficients(fit.lmer1)
```

```
## $Block
##   (Intercept) Vermetids1 Sediment1 Vermetids1:Sediment1
## 1    1.205018 0.004655556 0.2852023           -0.7735467
## 2    1.336068 0.004655556 0.2852023           -0.7735467
```

```
## 3      1.213966 0.004655556 0.2852023          -0.7735467
## 4      1.262805 0.004655556 0.2852023          -0.7735467
## 6      1.320787 0.004655556 0.2852023          -0.7735467
## 7      1.201242 0.004655556 0.2852023          -0.7735467
## 8      1.314441 0.004655556 0.2852023          -0.7735467
## 9      1.199831 0.004655556 0.2852023          -0.7735467
## 10     1.361541 0.004655556 0.2852023          -0.7735467
##
## attr(,"class")
## [1] "coef.mer"
```

```r
coefficients(summary(fit.lmer1))
```

```
##                         Estimate Std. Error       df     t value
## (Intercept)            1.268411111  0.1541680 30.42768  8.22745788
## Vermetids1             0.004655556  0.2091398 22.94243  0.02226049
## Sediment1              0.285202305  0.2160126 23.53130  1.32030428
## Vermetids1:Sediment1  -0.773546750  0.3006674 23.24638 -2.57276556
##                           Pr(>|t|)
## (Intercept)            3.129900e-09
## Vermetids1             9.824326e-01
## Sediment1              1.994327e-01
## Vermetids1:Sediment1   1.693404e-02
```

```r
fit.emm <- emmeans(fit.lmer1, specs=c("Vermetids", "Sediment"))
summary(contrast(fit.emm, method="revpairwise", adjust="none"), infer=c(TRUE, TRUE))
```

```
##   contrast   estimate    SE   df lower.CL upper.CL t.ratio p.value
##   1,0 - 0,0  0.00466 0.209 23.0   -0.428   0.4373   0.022  0.9824
##   0,1 - 0,0  0.28520 0.217 23.6   -0.163   0.7330   1.316  0.2009
##   0,1 - 1,0  0.28055 0.217 23.6   -0.167   0.7283   1.294  0.2081
##   1,1 - 0,0 -0.48369 0.209 23.0   -0.916  -0.0511  -2.313  0.0300
##   1,1 - 1,0 -0.48834 0.209 23.0   -0.921  -0.0557  -2.335  0.0286
##   1,1 - 0,1 -0.76889 0.217 23.6   -1.217  -0.3211  -3.547  0.0017
##
## Confidence level used: 0.95
```

```r
fit.lmer1.ml <- lmer(GrowthRate ~ Vermetids*Sediment + (1|Block), data=coral[inc], REML=FALSE)
# random intercept and slope
fit.lmer3.ml <- lmer(GrowthRate ~ Vermetids*Sediment + (Vermetids|Block) + (Sediment|Block), data=coral
```

```
## boundary (singular) fit: see ?isSingular
```

```r
AIC(fit.lmer1.ml)
```

```
## [1] 52.80349
```

```r
AIC(fit.lmer3.ml)
```

```
## [1] 61.04197
```

The formula for `lmer`

Appendix 1: Getting Started with R

17.9   Get your computer ready

17.9.1   Install R

R is the core software

Download R for your OS

17.9.2   Install R Studio

R Studio is a slick (very slick) GUI interface for developing R projects

Download R Studio Desktop

17.9.3   Resources for installing R and R Studio

On Windows

On a Mac

17.9.4   Install LaTeX

LaTeX ("la-tek") is necessary to use the pdf output of R Markdown.

On Windows

On a Mac

17.10   Start learning

17.10.1   Start with Data Camp Introduction to R

Data Camp: Introduction to R (free online course)

17.10.2   Then Move to Introduction to R Studio

R Studio Essentials, Programming Part 1 (Writing code in RStudio)

17.10.3   Develop your project with an R Studio Notebook

Getting Started with R Markdown

Introducing Notebooks with R Markdown

17.11   Getting Data into R

Getting your data into R

17.12   Additional R learning resources

Getting used to R, RStudio, and R Markdown

Link to list of R Studio webinars

Link to set of R package cheat sheets (amazing!)

Bookdown online books

17.13   Packages used extensively in this text

1. ggplot2
2. data.table
3. mvtnorm
4. lme4
5. nlme
6. emmeans
7. readxl
8. reshape2

Data Visualisation chapter from *R for Data Science*

Graphics for communication chapter from *R for Data Science*

Youtube: An Introduction to The data.table Package

Coursera: The data.table Package

Appendix 2: Online Resources for Getting Started with Statistical Modeling in R

Roughly, in order from most elementary to most advanced

Learning Statistics with R by Danielle Navarro and adapted to Bookdown (for web viewing) by Emily Kothe.

Statististical Thinking for the 21st Century by Russell A. Poldrack

Regression Models for Data Science in R by Brian Caffo

Broadening Your Statistical Horizons: Generalized Linear Models and Multilevel Models by J. Legler and P. Roback

Modern Statistics for Modern Biology

The Art of Data Science by Roger D. Peng and Elizabeth Matsui

Bibliography