Practical Data Analysis for Political Scientists

Brenton Kenkel

2017-03-14

Contents

Chapter 1

About This Book

This book contains the course notes for Brenton Kenkel's course Statistics for Political Research II (PSCI 8357 at Vanderbilt University). It covers the basics of statistical modeling and programming with linear models, along with applications in R.

This book is written in R Markdown and published via Bookdown on GitHub Pages. You can find the R Markdown source files at https://github.com/brentonk/pdaps.

Chapter 2

Principles of Programming

It may seem strange to begin a statistics class with two weeks on programming. It is strange. Here is why I have made this strange choice.

First, as a working social scientist, most of the time you spend on data analysis won't be on the *analysis* part. It'll be on obtaining and cleaning the data, to get it in a form that makes sense to analyze. Good programming skills will let you spend less time cleaning data and more time publishing papers.

Second, even if you don't want to develop good programming habits, journals are going to force you to. Every reputable political science journal requires that you provide replication scripts, and some of the best (e.g., *American Journal of Political Science*) have begun auditing the replication materials as a condition of publication. Better to learn The Right Way now when you have lots of time than to be forced to when you're writing a dissertation or on the market or teaching your own courses.

Third, while I feel embarrassed to invoke the cliché that is Big Data, that doesn't mean it's not a real thing. Political scientists have access to more data and more computing power than ever before. You can't collect, manage, clean, and analyze large quantities of data without understanding the basic principles of programming.

As Bowers (2011) puts it, "Data analysis is computer programming." By getting a PhD in political science,[1] by necessity you're going to become a computer programmer. The choice before you is whether to be a good one or a bad one.

Wilson et al. (2014) list eight "best practices for scientific computing." The first two encapsulate most of what you need to know:

1. Write programs for people, not computers.
2. Let the computer do the work.

[1] Or whatever other social science field.

2.1 Write Programs for People, Not Computers

The first two words here—*write programs*—are crucial. When you are doing analysis for a research project, you should be writing and running scripts, not typing commands into the R (or Stata) console. The console is ephemeral, but scripts are forever, at least if you save them.

Like the manuscripts you will write to describe your findings, your analysis scripts are a form of scientific communication. You wouldn't write a paper that is disorganized, riddled with grammatical errors, or incomprehensible to anyone besides yourself. Don't write your analysis scripts that way either.

Each script should be self-contained, ideally accomplishing one major task. Using an omnibus script that runs every bit of analysis is like writing a paper without paragraph breaks. A typical breakdown of scripts for a project of mine looks like:

- `0-download.r`: downloads the data
- `1-clean.r`: cleans the data
- `2-run.r`: runs the main analysis
- `3-figs.r`: generates figures

The exact structure varies depending on the nature of the project. Notice that the scripts are numbered in the order they should be run.

Within each script, write the code to make it as easy as possible for your reader to follow what you're doing. You should indent your code according to style conventions such as http://adv-r.had.co.nz/Style.html. Even better, use the `Code -> Reindent Lines` menu option in R Studio to automatically indent according to a sane style.

```
# Bad
my_results <- c(mean(variable),
quantile(variable,
probs = 0.25),
max(variable))

# Better
my_results <- c(mean(variable),
                quantile(variable,
                         probs = 0.25),
                max(variable))
```

Another way to make your code readable—one that, unfortunately, cannot be accomplished quite so algorithmically—is to add explanatory comments. The point of comments is not to document how the language works. The following comment is an extreme example of a useless comment.

```
# Take the square root of the errors and assign them to
# the output variable
```

```
output <- sqrt(errors)
```

A better use for the comment would be to explain *why* you're taking the square root of the errors, at least if your purpose in doing so would be unclear to a hypothetical reader of the code.

My basic heuristic for code readability is *If I got hit by a bus tomorrow, could one of my coauthors figure out what the hell I was doing and finish the paper?*

2.2 Let the Computer Do the Work

Computers are really good at structured, repetitive tasks. If you ever find yourself entering the same thing into the computer over and over again, you are Doing It Wrong. Your job as the human directing the computer is to figure out the structure that underlies the repeated task and to program the computer to do the repetition.

For example, imagine you have just run a large experiment and you want to estimate effects by subgroups.[2] Your respondents differ across four variables—party ID (R or D), gender (male or female), race (white or nonwhite), and education (college degree or not)—giving you 16 subgroups. You *could* copy and paste your code to estimate the treatment effect 16 times. But this is a bad idea for a few reasons.

- Copy-paste doesn't scale. Copy-paste is managable (albeit misguided) for 16 iterations, but probably not for 50 and definitely not for more than 100.
- Making changes becomes painful. Suppose you decide to change how you calculate the estimate. Now you have to go back and individually edit 16 chunks of code.
- Copy-paste is error-prone, and insidiously so. If you do the calculation wrong all 16 times, you'll probably notice. But what if you screwed up for just one or two cases? Are you *really* going to go through and check that you did everything right in each individual case?

We're going to look at the most basic ways to get the computer to repeat structured tasks—functions and control flow statements. To illustrate these, we will use a result you discussed in Stat I: the central limit theorem.

The central limit theorem concerns the *sampling distribution* of the sample mean,

$$\bar{X} = \frac{1}{N} \sum_{n=1}^{N} X_n,$$

[2]There could be statistical problems with this kind of analysis, at least if the subgroups were specified *post hoc*. See https://xkcd.com/882/ ("Significant"). We're going to leave this issue aside for now, but we'll return to it later when we discuss the statistical crisis in science.

where each X_n is independent and identically distributed with mean μ and variance σ^2. Loosely speaking, the CLT says that as N grows large, the sampling distribution of $\bar{X}$ becomes approximately normal with mean μ and variance σ^2/N.

Here's what we would need to do to see the CLT in practice. We'd want to take a bunch of samples, each of size N, and calculate the sample mean of each. Then we'd have a sample of sample means, and we could check to verify that they are approximately normally distributed with mean μ and variance σ^2/N. This is a structured, repetitive task—exactly the kind of thing that should be programmed. We'll try it out with a random variable from a Poisson distribution with $\lambda = 3$, which has mean $\mu = 3$ and variance $\sigma^2 = 3$.

First things first. We can use the `rpois` function to draw a random sample of N numbers from the Poisson distribution.

```
samp <- rpois(10, lambda = 3)
samp
```

```
##  [1] 2 3 8 3 5 4 3 4 2 2
```

To calculate the sample mean, we simply use the `mean` function.

```
mean(samp)
```

```
## [1] 3.6
```

We are interested in the distribution of the sample mean across many samples like this one. To begin, we will write a **function** that automates our core task—drawing a sample of N observations from Poisson(3) and calculating the sample mean. A function consists of a set of *arguments* (the inputs) and a *body* of code specifying which calculations to perform on the inputs to produce the output.

```
pois_mean <- function(n_obs) {
  samp <- rpois(n_obs, lambda = 3)
  ans <- mean(samp)
  return(ans)
}
```

This code creates a function called `pois_mean`. It has a single argument, called `n_obs`. It generates a random sample of `n_obs` draws from Poisson(3) and calculates its sample mean. It then `return`s the sample mean as the output.

Let's try calling this function a few times, each with a sample size of $N = 30$. Your output will differ slightly from what's printed here, since the function is generating random numbers.

```
pois_mean(n_obs = 30)
```

```
## [1] 3.0333
```

```
pois_mean(n_obs = 30)
```

```
## [1] 2.4667
```

```
pois_mean(n_obs = 30)
```

```
## [1] 2.9667
```

Remember that what we're interested in is the *sampling distribution* of the sample mean—the distribution of the sample mean across every possible sample of N observations. We can approximate this distribution by running `pois_mean` many times (e.g., 1000 or more). This would be infeasible via copy-paste. Instead, we will use a **for loop**.

```
# Set up a vector to store the output
n_replicates <- 1000
sampling_dist <- rep(NA, n_replicates)

for (i in 1:n_replicates) {
  sampling_dist[i] <- pois_mean(n_obs = 30)
}
```

Here's how the for loop works. We specified `i` as the name of the index variable, with values `1:n_replicates`. The for loop takes each value in the sequence, assigns it to the variable `i`, runs the given expression (in this case, assigning the output of `pois_mean` to the `i`'th element of `sampling_dist`), and then moves on to the next value in sequence, until it reaches the end.

Let's take a look at the results and compare them to our expectations.

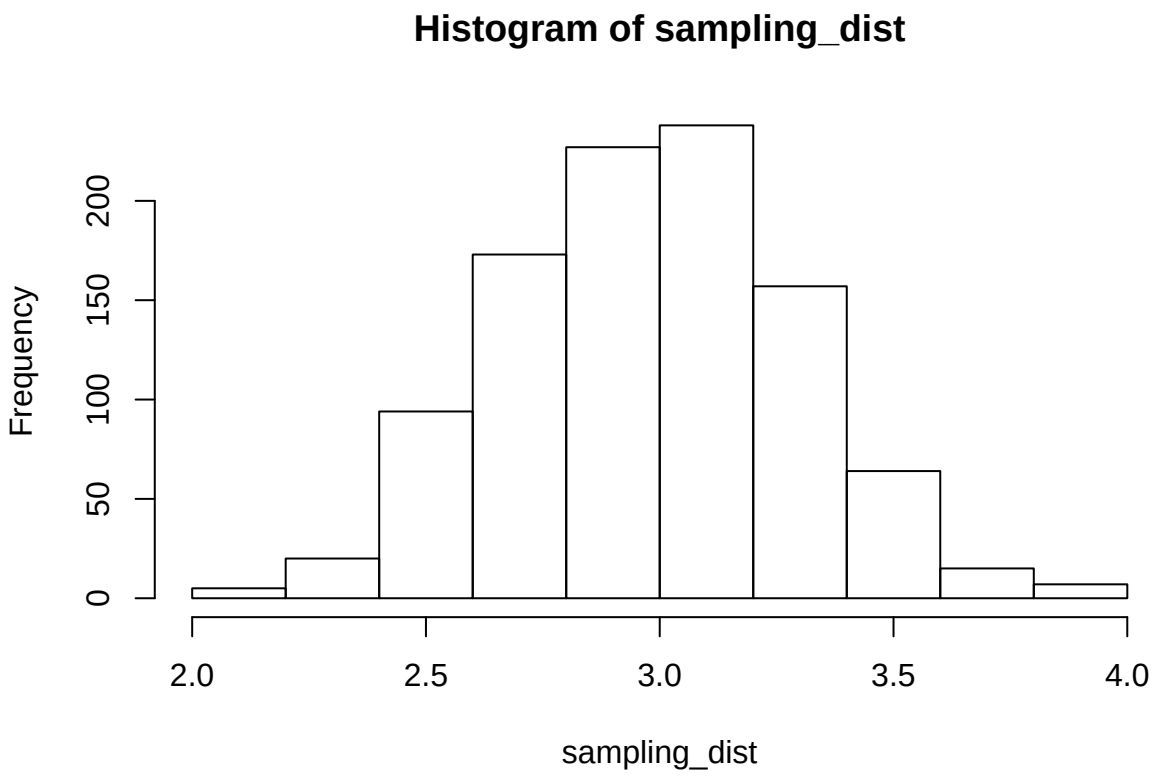
```
mean(sampling_dist)  # Expect 3
```

```
## [1] 2.9952
```

```
var(sampling_dist)  # Expect 1/10
```

```
## [1] 0.096944
```

```
hist(sampling_dist)  # Expect roughly normal
```



For loops are fun, but don't overuse them. Many simple operations are **vectorized** and don't require a loop. For example, suppose you want to take the square of a sequence of numbers. You could use a for loop …

```
input <- c(1, 3, 7, 29)
output <- rep(NA, length(input))

for (i in 1:length(input)) {
  output[i] <- input[i]^2
}

output
```

```
## [1]   1   9  49 841
```

… but it's faster (in terms of computational speed) and easier to just take advantage of vectorization:

```
input^2
```

```
## [1]   1   9  49 841
```

Another useful piece of control flow is **if/else statements**. These check a logical condition—an expression whose value is `TRUE` or `FALSE`—and run different code depending on the value of the expression. (You may want to catch up on the comparison operators: `==`, `>`, `>=`, `<`, `<=`, etc.)

Let's edit the `pois_mean` function to allow us to calculate the median instead of the mean. We'll add a second argument to the function, and implement the option using an if/else statement.

```
pois_mean <- function(n_obs, use_median = FALSE) {
  samp <- rpois(n_obs, lambda = 3)
  if (use_median) {
    ans <- median(samp)
  } else {
    ans <- mean(samp)
  }
  return(ans)
}
```

A couple of things to notice about the structure of the function. We use a comma to separate multiple function arguments. Also, we've specified `FALSE` as the *default* for the `use_median` argument. If we call the function without explicitly specifying a value for `use_median`, the function sets it to `FALSE`.

```
pois_mean(n_obs = 9)
```

```
## [1] 3.7778
```

```
pois_mean(n_obs = 9, use_median = TRUE)
```

```
## [1] 2
```

```
pois_mean(n_obs = 9, use_median = FALSE)
```

```
## [1] 2.6667
```

There is a vectorized version of if/else statements called, naturally, the `ifelse` function. This function takes three arguments, each a vector of the same length: (1) a logical condition, (2) an output value if the condition is `TRUE`, (3) an output value if the condition is `FALSE`.

```
x <- 1:10
big_x <- x * 100
small_x <- x * -100

ifelse(x > 5, big_x, small_x)
```

```
##  [1] -100 -200 -300 -400 -500  600  700  800  900 1000
```

Functions, for loops, and if/else statements are just a few of the useful tools for programming in R.[3] But even these simple tools are enough to allow you to do much more at scale than you could with a copy-paste philosophy.

[3] Others include the `replicate` function, the `apply` family of functions (`sapply`, `lapply`, `tapply`, `mapply`, ...), the **foreach** package, the **purrr** package, just to name a few of the most useful off the top of my head.

Chapter 3

Working with Data

Some material in this chapter is adapted from notes Matt DiLorenzo wrote for the Spring 2016 session of PSCI 8357.

Let me repeat something I said last week. In your careers as social scientists, starting with your dissertation research—if not earlier—you will probably spend more time collecting, merging, and cleaning data than you will on statistical analysis. So it's worth taking some time to learn how to do this well.

Best practices for data management can be summarized in a single sentence: *Record and document everything you do to the data.*

The first corollary of this principle is that raw data is sacrosanct. You should never edit raw data "in place". Once you download the raw data file, that file should never change.[1]

In almost any non-trivial analysis, the "final" data—the format you plug into your analysis—will differ significantly from the raw data. It may consist of information merged from multiple sources. The variables may have been transformed, aggregated, or otherwise altered. The unit of observation may even differ from the original source. You must document every one of these changes, so that another researcher working from the exact same raw data will end up with the exact same final data.

The most sensible way to achieve this level of reproducibility is to do all of your data merging and cleaning in a script. In other words, no going into Excel and mucking around manually. Like any other piece of your analysis, your pipeline from raw data to final data should follow the principles of programming that we discussed last week.

Luckily for you,[2] the **tidyverse** suite of R packages (including **dplyr**, **tidyr**, and others) makes it easy to script your "data pipeline". We'll begin by loading the package.

[1] Even if it's data you collected yourself, that data should still have a "canonical" representation that never gets overwritten. See Leek (2015) for more on distributing your own data.

[2] But not for me, because these tools didn't exist when I was a PhD student. Also, get off my lawn!

```
library("tidyverse")
```

3.1 Loading

The first step in working with data is to acquire some data. Depending on the nature of your research, you will be getting some or all of your data from sources available online. When you download data from online repositories, you should keep track of where you got it from. The best way to do so is—you guessed it—to script your data acquisition.

The R function `download.file()` is the easiest way to download files from URLs from within R. Just specify where you're getting the file from and where you want it to go. For the examples today, we'll use an "untidied" version of the World Development Indicators data from the World Bank that I've posted to my website.

```
download.file(url = "http://bkenkel.com/data/untidy-data.csv",
              destfile = "my-untidy-data.csv")
```

Once you've got the file stored locally, use the utilities from the **readr** package (part of **tidyverse**) to read it into R as a data frame.[3] We have a CSV file, so we will use `read_csv`. See `help(package = "readr")` for other possibilities.

```
untidy_data <- read_csv(file = "my-untidy-data.csv")
```

```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   gdp.2005 = col_double(),
##   gdp.2006 = col_double(),
##   gdp.2007 = col_double(),
##   gdp.2008 = col_double(),
##   pop.2005 = col_double(),
##   pop.2006 = col_double(),
##   pop.2007 = col_double(),
##   pop.2008 = col_double(),
##   unemp.2005 = col_double(),
##   unemp.2006 = col_double(),
##   unemp.2007 = col_double(),
##   unemp.2008 = col_double()
## )
```

Remember that each column of a data frame might be a different type, or more formally *class*, of object. `read_csv` and its ilk try to guess the type of data each column contains: character,

[3] More precisely, the **readr** functions produce output of class `"tbl_df"` (pronounced "tibble diff," I'm told), which are like data frames but better. See `help(package = "tibble")` for what can be done with `tbl_df`s.

integer, decimal number ("double" in programming-speak), or something else. The readout above tells you what guesses it made. If it gets something wrong—say, reading a column as numbers that ought to be characters—you can use the `col_types` argument to set it straight.

FYI, you could also run `read_csv()` directly on a URL, as in:

```
read_csv("http://bkenkel.com/data/untidy-data.csv")
```

However, in analyses intended for publication, it's usually preferable to download and save the raw data. What's stored at a URL might change or disappear, and you'll need to have a hard copy of the raw data for replication purposes.

Now let's take a look at the data we've just loaded in.

```
head(untidy_data)
```

```
## # A tibble: 6 × 13
##   country gdp.2005 gdp.2006 gdp.2007 gdp.2008  pop.2005  pop.2006
##     <chr>    <dbl>    <dbl>    <dbl>    <dbl>     <dbl>     <dbl>
## 1      AD   3.8423   4.0184   4.0216   3.6759  0.081223  0.083373
## 2      AE 253.9655 278.9489 287.8318 297.0189  4.481976  5.171255
## 3      AF   9.7630  10.3052  11.7212  12.1445 24.399948 25.183615
## 4      AG   1.1190   1.2687   1.3892   1.3902  0.082565  0.083467
## 5      AL   9.2684   9.7718  10.3483  11.1275  3.011487  2.992547
## 6      AM   7.6678   8.6797   9.8731  10.5544  3.014917  3.002161
## # ... with 6 more variables: pop.2007 <dbl>, pop.2008 <dbl>,
## #   unemp.2005 <dbl>, unemp.2006 <dbl>, unemp.2007 <dbl>, unemp.2008 <dbl>
```

We have a `country` variable giving country abbreviations. The other variables are numerical values: the country's GDP in 2005, 2006, 2007, and 2008; then the same for population and unemployment. Let's get this into a format we could use for analysis.

3.2 Tidying

Wickham (2014) outlines three qualities that make data "tidy":

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

For one thing, this means that whether a dataset is tidy or not depends—at least in part (some data collections are messy from any angle)—on the purpose it's being analyzed for.

Each row of `untidy_data` is a country. In observational studies in comparative politics and international relations, more commonly the unit of observation is the country-year.[4] How

[4]Insert lame joke about how Americanists haven't heard of other countries. But, seriously, if you're

can we take `untidy_data` and easily make it into country-year data?

We'll use the **tidyr** package (again, part of **tidyverse**) to clean up this data. The biggest problem right now is that each column, besides the country identifier, really encodes two pieces of information: the year of observation and the variable being observed. To deal with this, we'll have to first transform the data from one untidy format to another. We're going to use the `gather()` function to make each row a country-year-variable.

What `gather()` does is make a row for each entry from a set of columns. It's probably easiest to understand it by seeing it in practice:

```
long_data <- gather(untidy_data,
                    key = variable,
                    value = number,
                    gdp.2005:unemp.2008)
head(long_data)
```

```
## # A tibble: 6 × 3
##   country variable   number
##     <chr>    <chr>    <dbl>
## 1      AD gdp.2005   3.8423
## 2      AE gdp.2005 253.9655
## 3      AF gdp.2005   9.7630
## 4      AG gdp.2005   1.1190
## 5      AL gdp.2005   9.2684
## 6      AM gdp.2005   7.6678
```

With the first argument, we told `gather()` to use the `untidy_data` data frame. With the last argument, we told it the set of columns to "gather" together into a single column. The `key` column specifies the name of the variable to store the "key" (original column name) in, and the `value` column specifies the name of the variable to store the associated value. For example, the second row of `long_data` encodes what we previously saw as the `gdp.2005` column of `untidy_data`.

Now we have a new problem, which is that `variable` encodes two pieces of information: the variable and the year of its observation. **tidyr** provides the `separate()` function to solve that, splitting a single variable into two.

```
long_data <- separate(long_data,
                      col = variable,
                      into = c("var", "year"))
head(long_data)
```

```
## # A tibble: 6 × 4
##   country   var  year   number
##     <chr> <chr> <chr>    <dbl>
## 1      AD   gdp  2005   3.8423
```

confused because you haven't heard of other countries, just think of "state-years".

```
## 2      AE   gdp  2005 253.9655
## 3      AF   gdp  2005   9.7630
## 4      AG   gdp  2005   1.1190
## 5      AL   gdp  2005   9.2684
## 6      AM   gdp  2005   7.6678
```

So now we have country-year-variable data, with the year and variable conveniently stored in different columns. To turn this into country-year data, we can use the `spread()` function, which is like the inverse of `gather()`. `spread()` takes a key column and a value column, and turns each different key into a column of its own.

```
clean_data <- spread(long_data,
                     key = var,
                     value = number)
head(clean_data)
```

```
## # A tibble: 6 × 5
##   country  year      gdp      pop unemp
##     <chr> <chr>    <dbl>    <dbl> <dbl>
## 1      AD  2005   3.8423 0.081223    NA
## 2      AD  2006   4.0184 0.083373    NA
## 3      AD  2007   4.0216 0.084878    NA
## 4      AD  2008   3.6759 0.085616    NA
## 5      AE  2005 253.9655 4.481976   3.1
## 6      AE  2006 278.9489 5.171255   3.3
```

When using `spread()` on data that you didn't previously `gather()`, be sure to set the `fill` argument to tell it how to fill in empty cells. A simple example:

```
test_data
```

```
## # A tibble: 3 × 3
##        id     k     v
##     <chr> <chr> <dbl>
## 1 brenton     a    10
## 2 brenton     b    20
## 3 patrick     b     5
```

```
spread(test_data, key = k, value = v)
```

```
## # A tibble: 2 × 3
##        id     a     b
## *   <chr> <dbl> <dbl>
## 1 brenton    10    20
## 2 patrick    NA     5
```

```
spread(test_data, key = k, value = v, fill = 100)
```

```
## # A tibble: 2 × 3
```

```
##        id     a     b
## *   <chr> <dbl> <dbl>
## 1 brenton    10    20
## 2 patrick   100     5
```

One more important note on **tidyverse** semantics. It includes a fabulous feature called the *pipe*, `%>%`, which makes it easy to string together a truly mind-boggling number of commands.

In pipe syntax, `x %>% f()` is equivalent to `f(x)`. That seems like a wasteful and confusing way to write `f(x)`, and it is. But if you want to string together a bunch of commands, it's much easier to comprehend

```
x %>%
  f() %>%
  g() %>%
  h() %>%
  i()
```

than `i(h(g(f(x))))`.

You can pass function arguments using the pipe too. For example, `f(x, bear = "moose")` is equivalent to `x %>% f(bear = "moose")`.

The key thing about the **tidyverse** functions is that each of them takes a data frame as its first argument, and returns a data frame as its output. This makes them highly amenable to piping. For example, we can combine all three steps of our tidying above with a single command, thanks to the pipe:[5]

```
untidy_data %>%
  gather(key = variable,
         value = number,
         gdp.2005:unemp.2008) %>%
  separate(col = variable,
           into = c("var", "year")) %>%
  spread(key = var,
         value = number)
```

```
## # A tibble: 860 × 5
##   country  year      gdp      pop unemp
## *   <chr> <chr>    <dbl>    <dbl> <dbl>
## 1      AD  2005   3.8423 0.081223    NA
## 2      AD  2006   4.0184 0.083373    NA
## 3      AD  2007   4.0216 0.084878    NA
## 4      AD  2008   3.6759 0.085616    NA
## 5      AE  2005 253.9655 4.481976   3.1
```

[5] If you are reading the PDF copy of these notes (i.e., the ones I hand out in class), the line breaks are eliminated, making the piped commands rather hard to read. I am working on fixing this. For now, you may find the online notes at http://bkenkel.com/pdaps easier to follow.

```
## # ... with 855 more rows
```

Without the pipe, if we wanted to run all those commands together, we would have to write:

```
spread(separate(gather(untidy_data,
                       key = variable,
                       value = number,
                       gdp.2005:unemp.2008),
                col = variable,
                into = c("var", "year")),
       key = var,
       value = number)
```

Sad!

3.3 Transforming and Aggregating

Tidying the data usually isn't the end of the process. If you want to perform further calculations on the raw, that's where the tools in **dplyr** (part of, you guessed it, the **tidyverse**) come in.

Perhaps the simplest **dplyr** function (or "verb", as the R hipsters would say) is `rename()`, which lets you rename columns.

```
clean_data %>%
  rename(gross_domestic_product = gdp)
```

```
## # A tibble: 860 × 5
##   country  year gross_domestic_product      pop unemp
## *   <chr> <chr>                  <dbl>    <dbl> <dbl>
## 1      AD  2005                 3.8423 0.081223    NA
## 2      AD  2006                 4.0184 0.083373    NA
## 3      AD  2007                 4.0216 0.084878    NA
## 4      AD  2008                 3.6759 0.085616    NA
## 5      AE  2005               253.9655 4.481976   3.1
## # ... with 855 more rows
```

The **dplyr** functions, like the vast majority of R functions, do not modify their inputs. In other words, running `rename()` on `clean_data` will return a renamed copy of `clean_data`, but won't overwrite the original.

```
clean_data
```

```
## # A tibble: 860 × 5
##   country  year      gdp      pop unemp
## *   <chr> <chr>    <dbl>    <dbl> <dbl>
## 1      AD  2005   3.8423 0.081223    NA
```

```
## 2      AD  2006   4.0184 0.083373    NA
## 3      AD  2007   4.0216 0.084878    NA
## 4      AD  2008   3.6759 0.085616    NA
## 5      AE  2005 253.9655 4.481976   3.1
## # ... with 855 more rows
```

If you wanted to make the change stick, you would have to run:

```
clean_data <- clean_data %>%
  rename(gross_domestic_product = gdp)
```

`select()` lets you keep a couple of columns and drop all the others. Or vice versa if you use minus signs.

```
clean_data %>%
  select(country, gdp)
```

```
## # A tibble: 860 × 2
##   country      gdp
## *   <chr>    <dbl>
## 1      AD   3.8423
## 2      AD   4.0184
## 3      AD   4.0216
## 4      AD   3.6759
## 5      AE 253.9655
## # ... with 855 more rows
```

```
clean_data %>%
  select(-pop)
```

```
## # A tibble: 860 × 4
##   country  year      gdp unemp
## *   <chr> <chr>    <dbl> <dbl>
## 1      AD  2005   3.8423    NA
## 2      AD  2006   4.0184    NA
## 3      AD  2007   4.0216    NA
## 4      AD  2008   3.6759    NA
## 5      AE  2005 253.9655   3.1
## # ... with 855 more rows
```

`mutate()` lets you create new variables that are transformations of old ones.

```
clean_data %>%
  mutate(gdppc = gdp / pop,
         log_gdppc = log(gdppc))
```

```
## # A tibble: 860 × 7
##   country  year      gdp      pop unemp  gdppc log_gdppc
```

```
##     <chr> <chr>    <dbl>    <dbl> <dbl>  <dbl>     <dbl>
## 1      AD  2005   3.8423 0.081223    NA 47.305    3.8566
## 2      AD  2006   4.0184 0.083373    NA 48.198    3.8753
## 3      AD  2007   4.0216 0.084878    NA 47.381    3.8582
## 4      AD  2008   3.6759 0.085616    NA 42.935    3.7597
## 5      AE  2005 253.9655 4.481976   3.1 56.664    4.0371
## # ... with 855 more rows
```

`filter()` cuts down the data according to the logical condition(s) you specify.

```
clean_data %>%
  filter(year == 2006)
```

```
## # A tibble: 215 × 5
##   country  year      gdp       pop unemp
##     <chr> <chr>    <dbl>     <dbl> <dbl>
## 1      AD  2006   4.0184  0.083373    NA
## 2      AE  2006 278.9489  5.171255   3.3
## 3      AF  2006  10.3052 25.183615   8.8
## 4      AG  2006   1.2687  0.083467    NA
## 5      AL  2006   9.7718  2.992547  12.4
## # ... with 210 more rows
```

`summarise()` calculates summaries of the data. For example, let's find the maximum unemployment rate.

```
clean_data %>%
  summarise(max_unemp = max(unemp, na.rm = TRUE))
```

```
## # A tibble: 1 × 1
##   max_unemp
##       <dbl>
## 1      37.6
```

This seems sort of useless, until you combine it with the `group_by()` function. If you group the data before `summarise`-ing it, you'll calculate a separate summary for each group. For example, let's calculate the maximum unemployment rate for each year in the data.

```
clean_data %>%
  group_by(year) %>%
  summarise(max_unemp = max(unemp, na.rm = TRUE))
```

```
## # A tibble: 4 × 2
##    year max_unemp
##   <chr>     <dbl>
## 1  2005      37.3
## 2  2006      36.0
## 3  2007      34.9
```

```
## 4  2008      37.6
```

`summarise()` produces a "smaller" data frame than the input—one row per group. If you want to do something similar, but preserving the structure of the original data, use `mutate` in combination with `group_by`.

```
clean_data %>%
  group_by(year) %>%
  mutate(max_unemp = max(unemp, na.rm = TRUE),
         unemp_over_max = unemp / max_unemp) %>%
  select(country, year, contains("unemp"))
```

```
## Source: local data frame [860 x 5]
## Groups: year [4]
##
##    country  year unemp max_unemp unemp_over_max
##      <chr> <chr> <dbl>     <dbl>          <dbl>
## 1       AD  2005    NA      37.3             NA
## 2       AD  2006    NA      36.0             NA
## 3       AD  2007    NA      34.9             NA
## 4       AD  2008    NA      37.6             NA
## 5       AE  2005   3.1      37.3        0.08311
## # ... with 855 more rows
```

This gives us back the original data, but with a `max_unemp` variable recording the highest unemployment level that year. We can then calculate each individual country's unemployment as a percentage of the maximum. Whether grouped `mutate` or `summarise` is better depends, of course, on the purpose and structure of your analysis.

Notice how I selected all of the unemployment-related columns with `contains("unemp")`. See `?select_helpers` for a full list of helpful functions like this for `select`-ing variables.

3.4 Merging

Only rarely will you be lucky enough to draw all your data from a single source. More often, you'll be merging together data from multiple sources.

The key to merging data from separate tables is to have consistent identifiers across tables. For example, if you run an experiment, you might have demographic data on each subject in one table, and each subject's response to each treatment in another table. Naturally, you'll want to have a subject identifier that "links" the records across tables, as in the following hypothetical example.

```
subject_data
```

```
## # A tibble: 3 × 4
##      id gender loves_bernie does_yoga
```

```
##   <dbl>  <chr>        <chr>     <chr>
## 1  1001   male           yes        no
## 2  1002 female            no       yes
## 3  1003   male            no        no
```

```
subject_response_data
```

```
## # A tibble: 6 × 3
##      id treatment response
##   <dbl>     <chr>    <chr>
## 1  1001 read_book      sad
## 2  1001  watch_tv      sad
## 3  1002 read_book    happy
## 4  1002  watch_tv      sad
## 5  1003 read_book      sad
## 6  1003  watch_tv    happy
```

Let's practice merging data with our cleaned-up country-year data. We'll take two datasets from my website: a country-level dataset with latitudes and longitudes, and a country-year–level dataset with inflation over time.

```
latlong_data <- read_csv("http://bkenkel.com/data/latlong.csv")
latlong_data
```

```
## # A tibble: 245 × 3
##   country latitude longitude
##     <chr>    <dbl>     <dbl>
## 1      AD   42.546    1.6016
## 2      AE   23.424   53.8478
## 3      AF   33.939   67.7100
## 4      AG   17.061  -61.7964
## 5      AI   18.221  -63.0686
## # ... with 240 more rows
```

```
inflation_data <- read_csv("http://bkenkel.com/data/inflation.csv")
inflation_data
```

```
## # A tibble: 1,070 × 3
##   country  year inflation
##     <chr> <int>     <dbl>
## 1      AD  2004        NA
## 2      AD  2005        NA
## 3      AD  2006        NA
## 4      AD  2007        NA
## 5      AD  2008        NA
## # ... with 1,065 more rows
```

For your convenience, both of these datasets use the same two-letter country naming scheme

as the original data. Unfortunately, out in the real world, data from different sources often use incommensurate naming schemes. Converting from one naming scheme to another is part of the data cleaning process, and it requires careful attention.

dplyr contains various `_join()` functions for merging. Each of these take as arguments the two data frames to merge, plus the names of the identifier variables to merge them on. The one I use most often is `left_join()`, which keeps every row from the first ("left") data frame and merges in the columns from the second ("right") data frame.

For example, let's merge the latitude and longitude data for each country into `clean_data`.

```
left_join(clean_data,
          latlong_data,
          by = "country")
```

```
## # A tibble: 860 × 7
##   country  year      gdp      pop unemp latitude longitude
##     <chr> <chr>    <dbl>    <dbl> <dbl>    <dbl>     <dbl>
## 1      AD  2005   3.8423 0.081223    NA   42.546    1.6016
## 2      AD  2006   4.0184 0.083373    NA   42.546    1.6016
## 3      AD  2007   4.0216 0.084878    NA   42.546    1.6016
## 4      AD  2008   3.6759 0.085616    NA   42.546    1.6016
## 5      AE  2005 253.9655 4.481976   3.1   23.424   53.8478
## # ... with 855 more rows
```

Since `latlong_data` is country-level, the value is the same for each year. So the merged data contains redundant information. This is one reason to store data observed at different levels in different tables—with redundant observations, it is easier to make errors yet harder to catch them and fix them.

We can also merge data when the identifier is stored across multiple columns, as in the case of our country-year data. But first, a technical note.[6] You might notice that the `year` column of `clean_data` is labeled `<chr>`, as in character data. Yet the `year` column of `inflation_data` is labeled `<int>`, as in integer data. We can check that by running `class()` on each respective column.

```
class(clean_data$year)
```

```
## [1] "character"
```

```
class(inflation_data$year)
```

```
## [1] "integer"
```

From R's perspective, the character string `"1999"` is a very different thing than the integer number `1999`. Therefore, if we try to merge `clean_data` and `inflation_data` on the `year` variable, it will throw an error.

[6] This won't be the case if you got `clean_data` by loading it in directly from `clean-data.csv` on my website, since `read_csv()` will have correctly encoded `year` as an integer.

```
left_join(clean_data,
          inflation_data,
          by = c("country", "year"))
```

```
## Error in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): Can't join on 'year' x 'yea
```

To fix this, let's use `mutate()` to convert the `year` column of `clean_data` to an integer. We probably should have done this in the first place—after all, having the year encoded as a character string would have thrown off plotting functions, statistical functions, or anything else where it would be more natural to treat the year like a number.

```
clean_data <- mutate(clean_data,
                     year = as.integer(year))
clean_data
```

```
## # A tibble: 860 × 5
##    country  year      gdp      pop unemp
##      <chr> <int>    <dbl>    <dbl> <dbl>
## 1       AD  2005   3.8423 0.081223    NA
## 2       AD  2006   4.0184 0.083373    NA
## 3       AD  2007   4.0216 0.084878    NA
## 4       AD  2008   3.6759 0.085616    NA
## 5       AE  2005 253.9655 4.481976   3.1
## # ... with 855 more rows
```

Looks the same as before, except with an important difference: `year` is now labeled `<int>`.

Now we can merge the two datasets together without issue. Notice how we use a vector in the `by` argument to specify multiple columns to merge on.

```
left_join(clean_data,
          inflation_data,
          by = c("country", "year"))
```

```
## # A tibble: 860 × 6
##    country  year      gdp      pop unemp inflation
##      <chr> <int>    <dbl>    <dbl> <dbl>     <dbl>
## 1       AD  2005   3.8423 0.081223    NA        NA
## 2       AD  2006   4.0184 0.083373    NA        NA
## 3       AD  2007   4.0216 0.084878    NA        NA
## 4       AD  2008   3.6759 0.085616    NA        NA
## 5       AE  2005 253.9655 4.481976   3.1        NA
## # ... with 855 more rows
```

You might remember that `inflation_data` contained some country-years not included in the original data (namely, observations from 2004). If we want the merged data to use the observations from `inflation_data` rather than `clean_data`, we can use the `right_join()` function.

```
right_join(clean_data,
           inflation_data,
           by = c("country", "year"))
```

```
## # A tibble: 1,070 × 6
##   country  year    gdp      pop unemp inflation
##     <chr> <int>  <dbl>    <dbl> <dbl>     <dbl>
## 1      AD  2004     NA       NA    NA        NA
## 2      AD  2005 3.8423 0.081223    NA        NA
## 3      AD  2006 4.0184 0.083373    NA        NA
## 4      AD  2007 4.0216 0.084878    NA        NA
## 5      AD  2008 3.6759 0.085616    NA        NA
## # ... with 1,065 more rows
```

One last common issue in merging is that the identifier variables have different names in the two datasets. If it's inconvenient or infeasible to correct this by renaming the columns in one or the other, you can specify the `by` argument as in the following example.

```
inflation_data <- rename(inflation_data,
                         the_country = country,
                         the_year = year)
inflation_data
```

```
## # A tibble: 1,070 × 3
##   the_country the_year inflation
##         <chr>    <int>     <dbl>
## 1          AD     2004        NA
## 2          AD     2005        NA
## 3          AD     2006        NA
## 4          AD     2007        NA
## 5          AD     2008        NA
## # ... with 1,065 more rows
```

```
left_join(clean_data,
          inflation_data,
          by = c("country" = "the_country", "year" = "the_year"))
```

```
## # A tibble: 860 × 6
##   country  year      gdp      pop unemp inflation
##     <chr> <int>    <dbl>    <dbl> <dbl>     <dbl>
## 1      AD  2005   3.8423 0.081223    NA        NA
## 2      AD  2006   4.0184 0.083373    NA        NA
## 3      AD  2007   4.0216 0.084878    NA        NA
## 4      AD  2008   3.6759 0.085616    NA        NA
## 5      AE  2005 253.9655 4.481976   3.1        NA
## # ... with 855 more rows
```

3.5 Appendix: Creating the Example Data

I used the same tools this chapter introduces to create the untidy data. I may as well include the code to do it, in case it helps further illustrate how to use the **tidyverse** tools (and, as a bonus, the **WDI** package for downloading World Development Indicators data).

First I load the necessary packages.

```
library("tidyverse")
library("WDI")
library("countrycode")
library("stringr")
```

Next, I download the relevant WDI data. I used the `WDIsearch()` function to locate the appropriate indicator names.

```
dat_raw <- WDI(country = "all",
               indicator = c("NY.GDP.MKTP.KD",  # GDP in 2000 USD
                             "SP.POP.TOTL",     # Total population
                             "SL.UEM.TOTL.ZS"), # Unemployment rate
               start = 2005,
               end = 2008)

head(dat_raw)
```

```
##   iso2c    country year NY.GDP.MKTP.KD SP.POP.TOTL SL.UEM.TOTL.ZS
## 1    1A Arab World 2005     1.6428e+12   313430911        12.1402
## 2    1A Arab World 2006     1.7629e+12   320906736        11.3296
## 3    1A Arab World 2007     1.8625e+12   328766559        10.8961
## 4    1A Arab World 2008     1.9799e+12   336886468        10.5060
## 5    1W      World 2005     5.7703e+13  6513959904         6.1593
## 6    1W      World 2006     6.0229e+13  6594722462         5.9000
```

I want to get rid of the aggregates, like the "Arab World" and "World" we see here. As a rough tack at that, I'm going to exclude those so-called countries whose ISO codes don't appear in the **countrycode** package data.[7]

```
dat_countries <- dat_raw %>%
  filter(iso2c %in% countrycode_data$iso2c)
```

Let's check on which countries are left. (I cut it down to max six characters per country name for printing purposes.)

```
dat_countries$country %>%
  unique() %>%
  str_sub(start = 1, end = 6)
```

[7] **countrycode** is a very useful, albeit imperfect, package for converting between different country naming/coding schemes.

```
##   [1] "Andorr" "United" "Afghan" "Antigu" "Albani" "Armeni" "Angola"
##   [8] "Argent" "Americ" "Austri" "Austra" "Aruba"  "Azerba" "Bosnia"
##  [15] "Barbad" "Bangla" "Belgiu" "Burkin" "Bulgar" "Bahrai" "Burund"
##  [22] "Benin"  "Bermud" "Brunei" "Bolivi" "Brazil" "Bahama" "Bhutan"
##  [29] "Botswa" "Belaru" "Belize" "Canada" "Congo," "Centra" "Congo,"
##  [36] "Switze" "Cote d" "Chile"  "Camero" "China"  "Colomb" "Costa "
##  [43] "Cuba"   "Cabo V" "Curaca" "Cyprus" "Czech " "German" "Djibou"
##  [50] "Denmar" "Domini" "Domini" "Algeri" "Ecuado" "Estoni" "Egypt,"
##  [57] "Eritre" "Spain"  "Ethiop" "Finlan" "Fiji"   "Micron" "Faroe "
##  [64] "France" "Gabon"  "United" "Grenad" "Georgi" "Ghana"  "Gibral"
##  [71] "Greenl" "Gambia" "Guinea" "Equato" "Greece" "Guatem" "Guam"
##  [78] "Guinea" "Guyana" "Hong K" "Hondur" "Croati" "Haiti"  "Hungar"
##  [85] "Indone" "Irelan" "Israel" "Isle o" "India"  "Iraq"   "Iran, "
##  [92] "Icelan" "Italy"  "Jamaic" "Jordan" "Japan"  "Kenya"  "Kyrgyz"
##  [99] "Cambod" "Kiriba" "Comoro" "St. Ki" "Korea," "Korea," "Kuwait"
## [106] "Cayman" "Kazakh" "Lao PD" "Lebano" "St. Lu" "Liecht" "Sri La"
## [113] "Liberi" "Lesoth" "Lithua" "Luxemb" "Latvia" "Libya"  "Morocc"
## [120] "Monaco" "Moldov" "Monten" "St. Ma" "Madaga" "Marsha" "Macedo"
## [127] "Mali"   "Myanma" "Mongol" "Macao " "Northe" "Maurit" "Malta"
## [134] "Maurit" "Maldiv" "Malawi" "Mexico" "Malays" "Mozamb" "Namibi"
## [141] "New Ca" "Niger"  "Nigeri" "Nicara" "Nether" "Norway" "Nepal"
## [148] "Nauru"  "New Ze" "Oman"   "Panama" "Peru"   "French" "Papua "
## [155] "Philip" "Pakist" "Poland" "Puerto" "West B" "Portug" "Palau"
## [162] "Paragu" "Qatar"  "Romani" "Serbia" "Russia" "Rwanda" "Saudi "
## [169] "Solomo" "Seyche" "Sudan"  "Sweden" "Singap" "Sloven" "Slovak"
## [176] "Sierra" "San Ma" "Senega" "Somali" "Surina" "South " "Sao To"
## [183] "El Sal" "Sint M" "Syrian" "Swazil" "Turks " "Chad"   "Togo"
## [190] "Thaila" "Tajiki" "Timor-" "Turkme" "Tunisi" "Tonga"  "Turkey"
## [197] "Trinid" "Tuvalu" "Tanzan" "Ukrain" "Uganda" "United" "Urugua"
## [204] "Uzbeki" "St. Vi" "Venezu" "Britis" "Virgin" "Vietna" "Vanuat"
## [211] "Samoa"  "Yemen," "South " "Zambia" "Zimbab"
```

With that out of the way, there's still some cleaning up to do. The magnitudes of GDP and population are too large, and the variable names are impenetrable. Also, the `country` variable, while helpful, is redundant now that we're satisfied with the list of countries remaining.

```
dat_countries <- dat_countries %>%
  select(-country) %>%
  rename(gdp = NY.GDP.MKTP.KD,
         pop = SP.POP.TOTL,
         unemp = SL.UEM.TOTL.ZS,
         country = iso2c) %>%
  mutate(gdp = gdp / 1e9,
         pop = pop / 1e6)
```

```
head(dat_countries)
```

```
##   country year      gdp      pop unemp
## 1      AD 2005   3.8423 0.081223    NA
## 2      AD 2006   4.0184 0.083373    NA
## 3      AD 2007   4.0216 0.084878    NA
## 4      AD 2008   3.6759 0.085616    NA
## 5      AE 2005 253.9655 4.481976   3.1
## 6      AE 2006 278.9489 5.171255   3.3
```

Now I convert the data to "long" format.

```
dat_countries_long <- dat_countries %>%
  gather(key = variable,
         value = value,
         gdp:unemp)

head(dat_countries_long)
```

```
##   country year variable    value
## 1      AD 2005      gdp   3.8423
## 2      AD 2006      gdp   4.0184
## 3      AD 2007      gdp   4.0216
## 4      AD 2008      gdp   3.6759
## 5      AE 2005      gdp 253.9655
## 6      AE 2006      gdp 278.9489
```

I then smush `variable` and `year` into a single column, and drop the individual components.

```
dat_countries_long <- dat_countries_long %>%
  mutate(var_year = paste(variable, year, sep = ".")) %>%
  select(-variable, -year)

head(dat_countries_long)
```

```
##   country    value var_year
## 1      AD   3.8423 gdp.2005
## 2      AD   4.0184 gdp.2006
## 3      AD   4.0216 gdp.2007
## 4      AD   3.6759 gdp.2008
## 5      AE 253.9655 gdp.2005
## 6      AE 278.9489 gdp.2006
```

Finally, I "widen" the data, so that each `var_year` is a column of its own.

```
dat_countries_wide <- dat_countries_long %>%
  spread(key = var_year, value = value)
```

```
head(dat_countries_wide)
```

```
##   country gdp.2005 gdp.2006 gdp.2007 gdp.2008  pop.2005  pop.2006
## 1      AD   3.8423   4.0184   4.0216   3.6759  0.081223  0.083373
## 2      AE 253.9655 278.9489 287.8318 297.0189  4.481976  5.171255
## 3      AF   9.7630  10.3052  11.7212  12.1445 24.399948 25.183615
## 4      AG   1.1190   1.2687   1.3892   1.3902  0.082565  0.083467
## 5      AL   9.2684   9.7718  10.3483  11.1275  3.011487  2.992547
## 6      AM   7.6678   8.6797   9.8731  10.5544  3.014917  3.002161
##    pop.2007  pop.2008 unemp.2005 unemp.2006 unemp.2007 unemp.2008
## 1  0.084878  0.085616         NA         NA         NA         NA
## 2  6.010100  6.900142        3.1        3.3        3.4        4.0
## 3 25.877544 26.528741        8.5        8.8        8.4        8.9
## 4  0.084397  0.085350         NA         NA         NA         NA
## 5  2.970017  2.947314       12.5       12.4       13.5       13.0
## 6  2.988117  2.975029       27.8       28.6       28.4       16.4
```

Now we have some ugly data. I save the output to upload to my website.

```
write_csv(dat_countries_wide, path = "untidy-data.csv")
```

And here's how I made the second country-year dataset used in the merging section. The country dataset with latitudes and longitudes is from https://developers.google.com/public-data/docs/canonical/countries_csv.

```
dat_2 <-
  WDI(country = "all",
      indicator = "FP.CPI.TOTL.ZG",
      start = 2004,
      end = 2008) %>%
  as_data_frame() %>%
  select(country = iso2c,
         year,
         inflation = FP.CPI.TOTL.ZG) %>%
  mutate(year = as.integer(year)) %>%
  filter(country %in% clean_data$country) %>%
  arrange(country, year)

write_csv(dat_2, path = "inflation.csv")
```

Chapter 4

Data Visualization

Visualization is most important at the very beginning and the very end of the data analysis process. In the beginning, when you've just gotten your data together, visualization is perhaps the easiest tool to explore each variable and learn about the relationships among them. And when your analysis is almost complete, you will (usually) use visualizations to communicate your findings to your audience.

We only have time to scratch the surface of data visualization. This chapter will cover the plotting techniques I find most useful for exploratory and descriptive data analysis. We will talk about graphical techniques for presenting the results of regression analyses later in the class—once we've, you know, learned something about regression.

4.1 Basic Plots

We will use the **ggplot2** package, which is part of—I'm as tired of it as you are—the **tidyverse**.

```
library("tidyverse")
```

For the examples today, we'll be using a dataset with statistics about the fifty U.S. states in 1977,[1] which is posted on my website.

```
state_data <- read_csv("http://bkenkel.com/data/state-data.csv")
state_data
```

```
## # A tibble: 50 × 12
##        State Abbrev Region Population Income Illiteracy LifeExp Murder
##        <chr>  <chr>  <chr>      <dbl>  <dbl>      <dbl>   <dbl>  <dbl>
## 1    Alabama     AL  South       3615   3624        2.1   69.05   15.1
## 2     Alaska     AK   West        365   6315        1.5   69.31   11.3
```

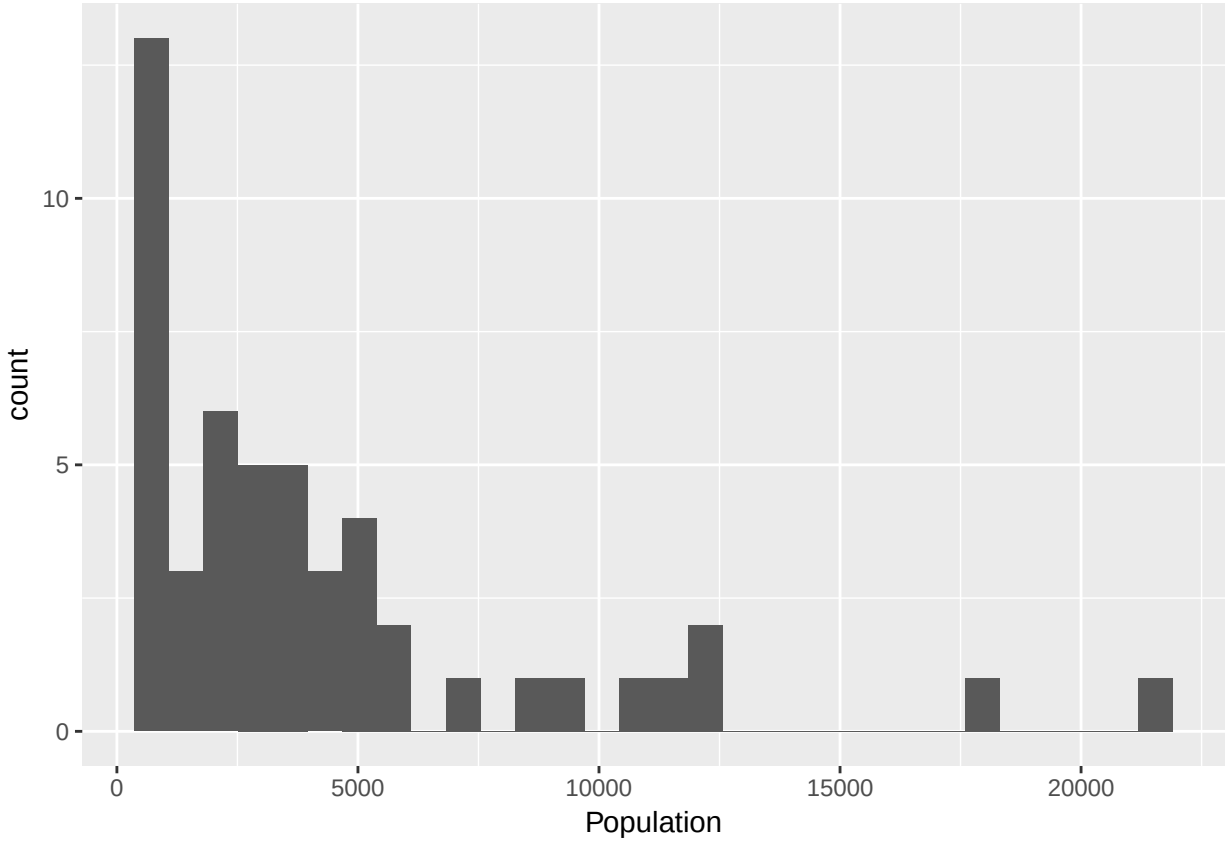
[1] Why 1977? Because it was easily available. See the appendix to this chapter.

```
## 3    Arizona     AZ   West       2212   4530        1.8   70.55    7.8
## 4   Arkansas     AR  South       2110   3378        1.9   70.66   10.1
## 5 California     CA   West      21198   5114        1.1   71.71   10.3
## # ... with 45 more rows, and 4 more variables: HSGrad <dbl>, Frost <dbl>,
## #   Area <dbl>, IncomeGroup <chr>
```

When I obtain data, I start by looking at the univariate distribution of each variable via a histogram. The following code creates a histogram in ggplot.

```
ggplot(state_data, aes(x = Population)) +
  geom_histogram()
```

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



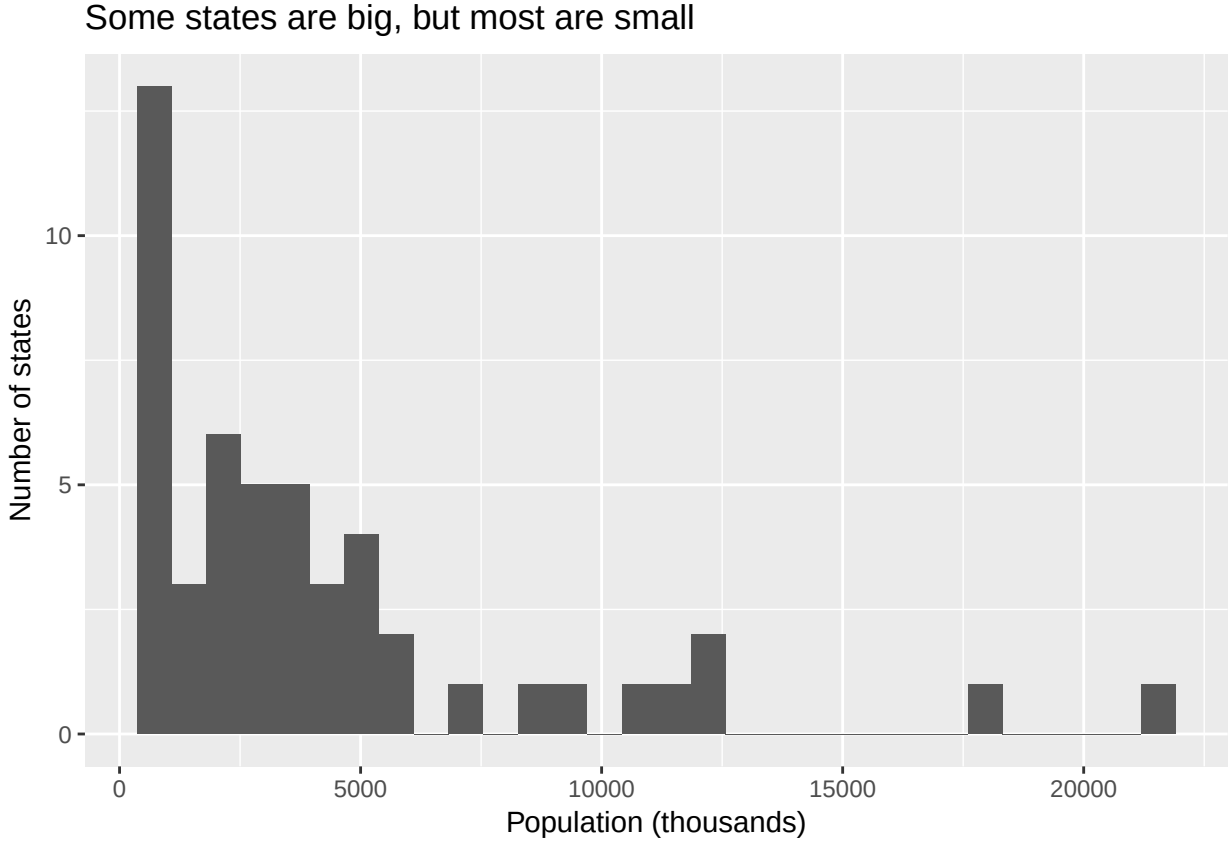
Let's walk through the syntax there. In the first line, we call `ggplot()`, specifying the data frame to draw from, then in the `aes()` command (which stands for "aesthetic") we specify the variable to plot. If this were a bivariate analysis, here we would have also specified a `y` variable to put on the y-axis. If we had just stopped there, we would have a sad, empty plot. The `+` symbol indicates that we'll be adding something to the plot. `geom_histogram()` is the command to overlay a histogram.

We'll only be looking at a few of the ggplot commands today. I recommend taking a look at the online package documentation at http://docs.ggplot2.org to see all of the many features available.

When you're just making graphs for yourself to explore the data, you don't need to worry about things like axis labels as long as you can comprehend what's going on. But when you prepare graphs for others to read (including those of us grading your problem sets!) you need to include an informative title and axis labels. To that end, use the `xlab()`, `ylab()`, and `ggtitle()` commands.

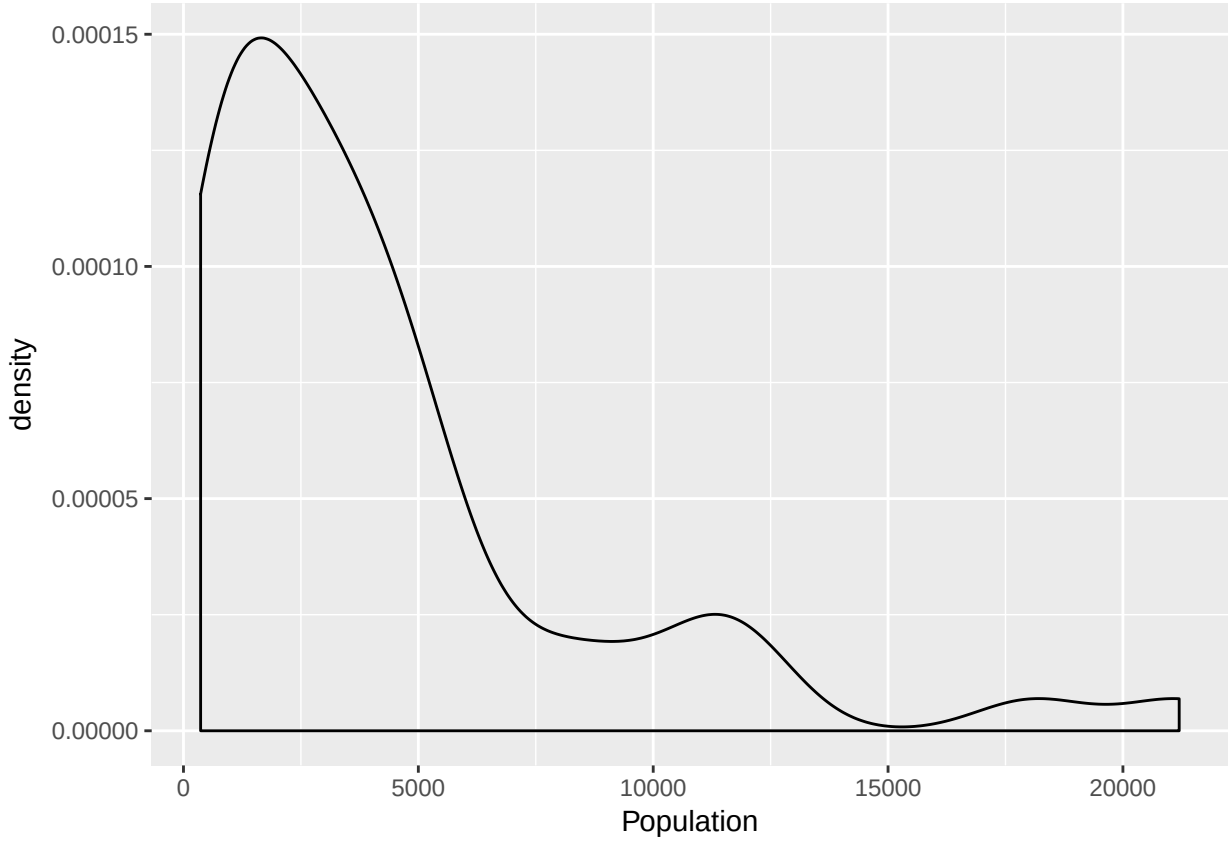
```
ggplot(state_data, aes(x = Population)) +
  geom_histogram() +
  xlab("Population (thousands)") +
  ylab("Number of states") +
  ggtitle("Some states are big, but most are small")
```

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



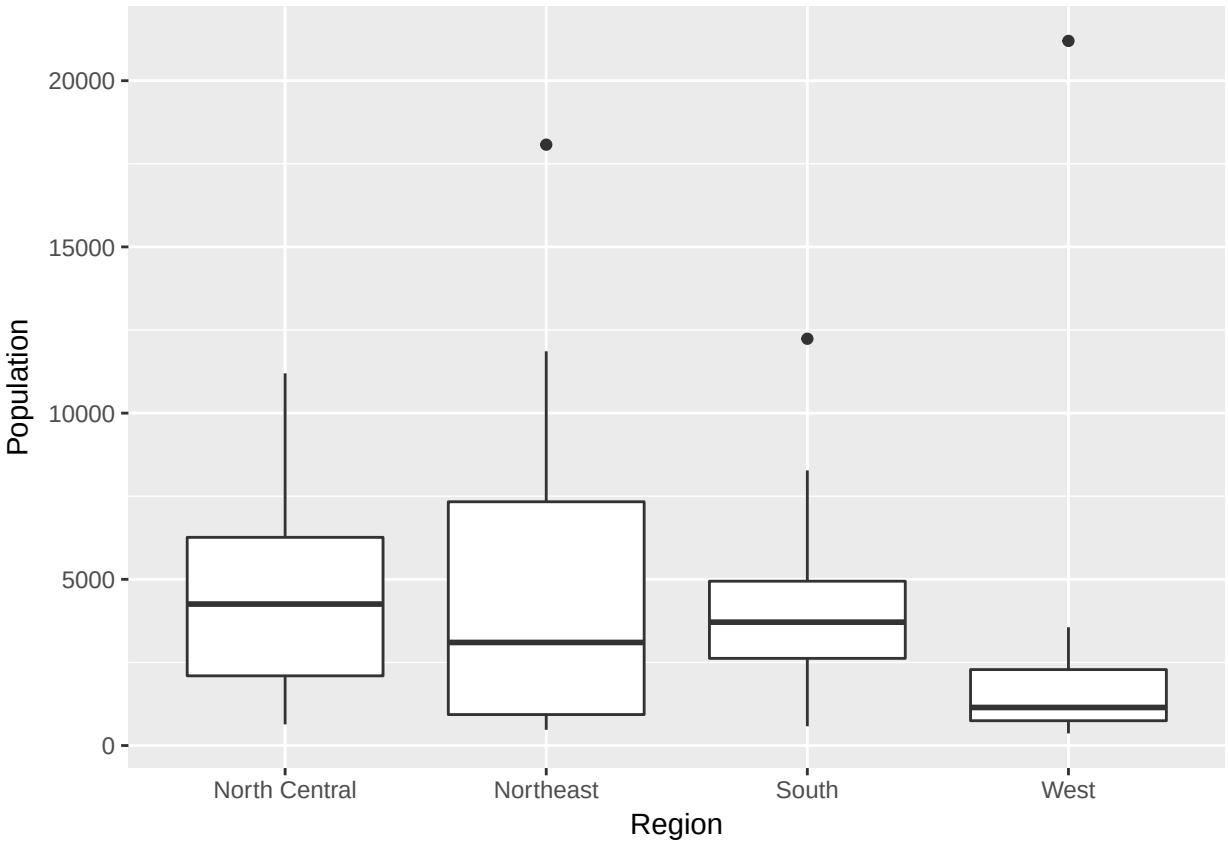
The density plot is a close relative of the histogram. It provides a smooth estimate of the probability density function of the data. Accordingly, the area under the density plot integrates to one. Depending on your purposes, this can make the y-axis of a density plot easier or (usually) harder to interpret than the count given by a histogram.

```
ggplot(state_data, aes(x = Population)) +
  geom_density()
```



The box plot is a common way to look at the distribution of a continuous variable across different levels of a categorical variable.

```
ggplot(state_data, aes(x = Region, y = Population)) +
  geom_boxplot()
```

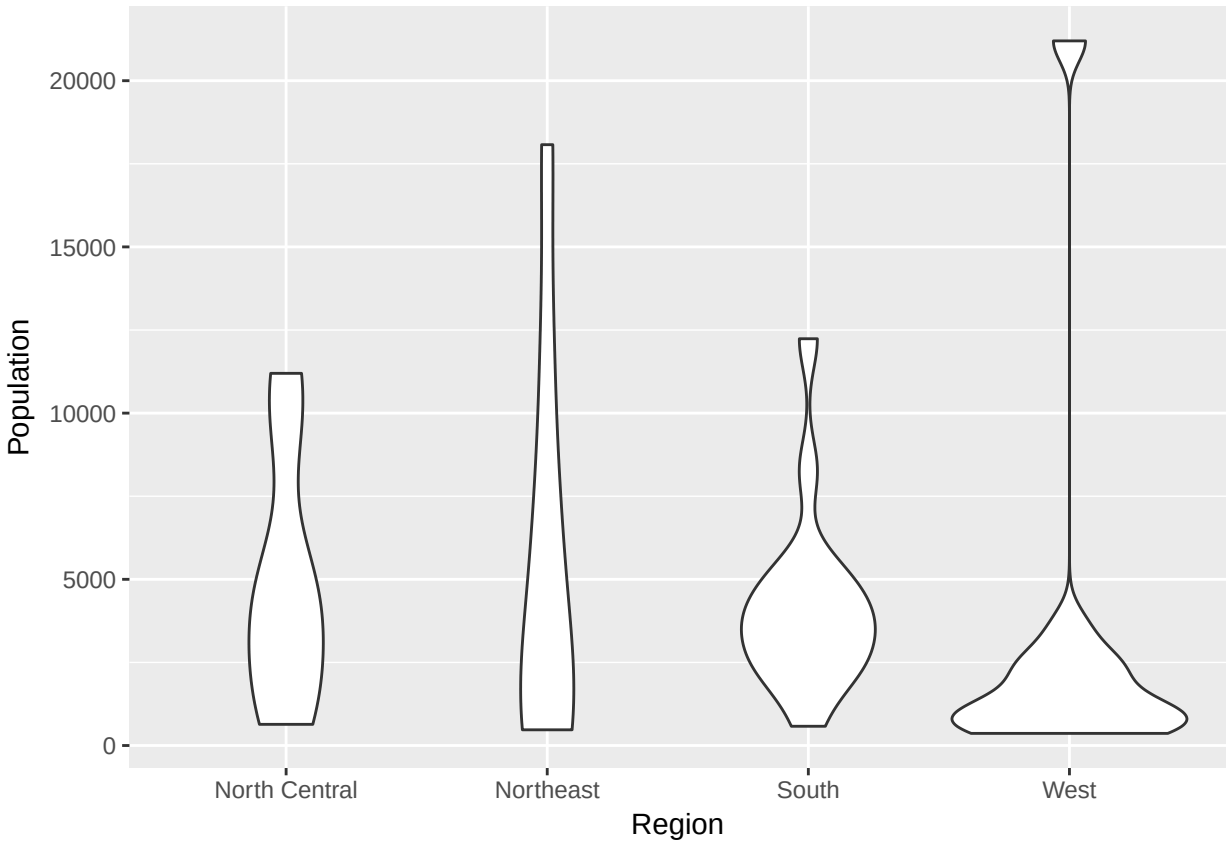


A box plot consists of the following components:

- Center line: median of the data
- Bottom of box: 25th percentile
- Top of box: 75th percentile
- Lower "whisker": range of observations no more than 1.5 IQR (height of box) below the 25th percentile
- Upper "whisker": range of observations no more than 1.5 IQR above the 75th percentile
- Plotted points: any data lying outside the whiskers

If you want to skip the summary and plot the full distribution of a variable across categories, you can use a violin plot.

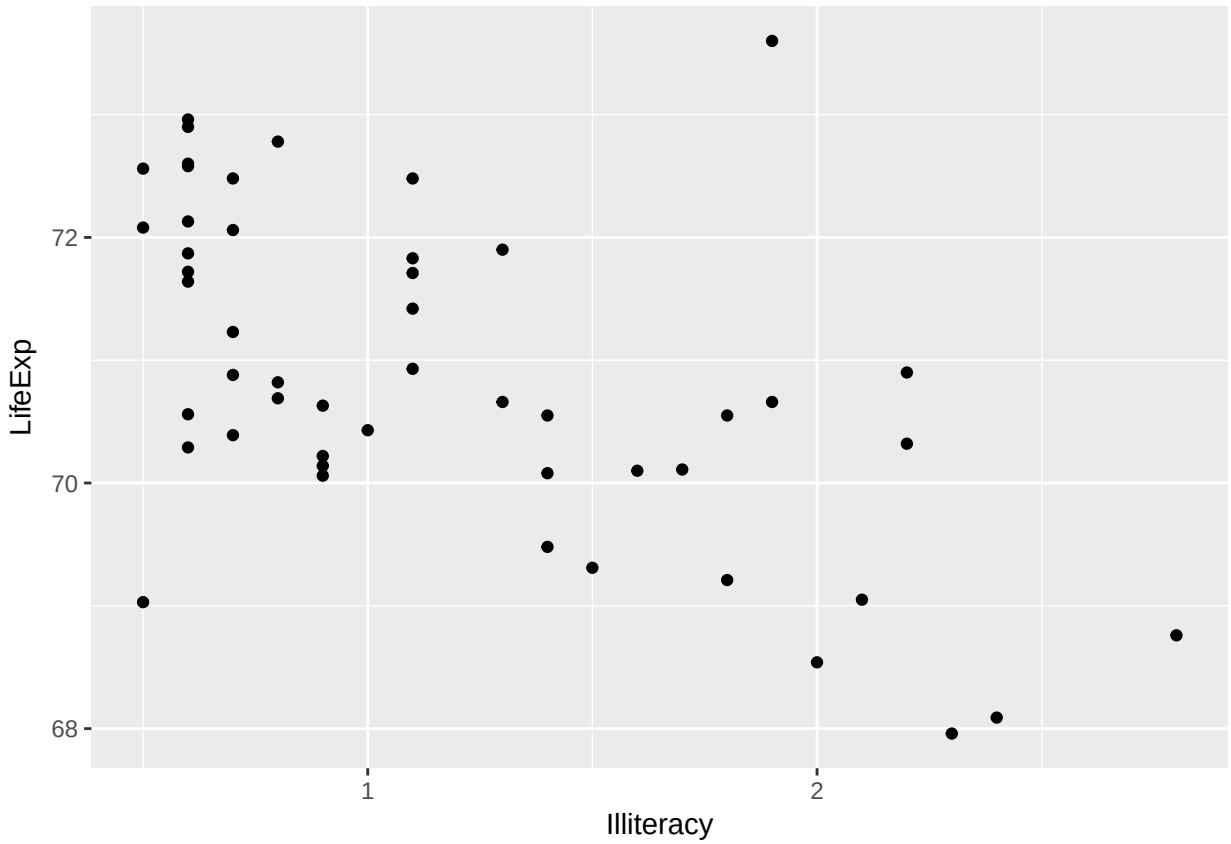
```
ggplot(state_data, aes(x = Region, y = Population)) +
  geom_violin()
```



Technically, violin plots convey more information than box plots since they show the full distribution. However, readers aren't as likely to be familiar with a violin plot. It's harder to spot immediately where the median is (though you could add that to the plot if you wanted). Plus, violin plots look goofy with outliers—see the "West" column above—whereas box plots handle them easily.

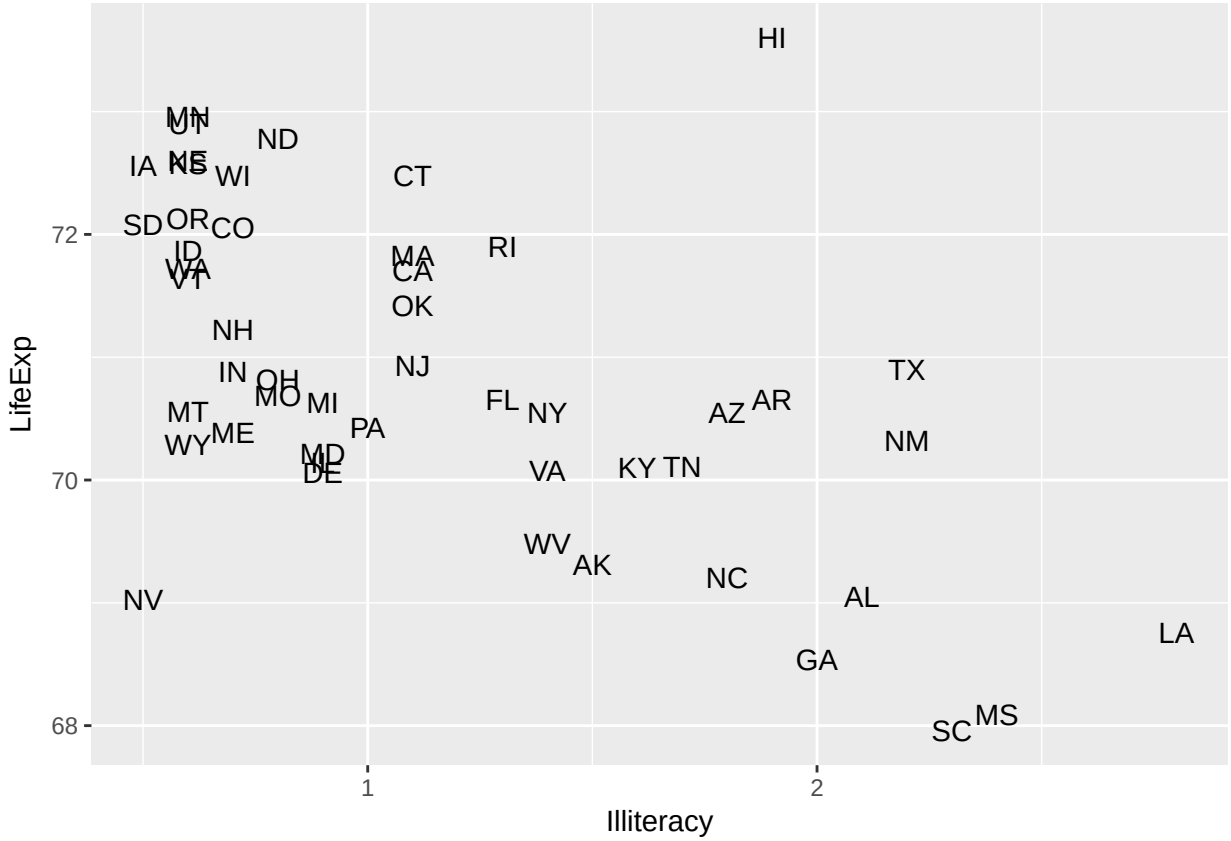
For visualizing relationships between continuous variables, nothing beats the scatterplot.

```
ggplot(state_data, aes(x = Illiteracy, y = LifeExp)) +
  geom_point()
```



When you're plotting states or countries, a hip thing to do is plot abbreviated names instead of points. To do that, you can use `geom_text()` instead of `geom_point()`, supplying an additional aesthetic argument telling ggplot where to draw the labels from.

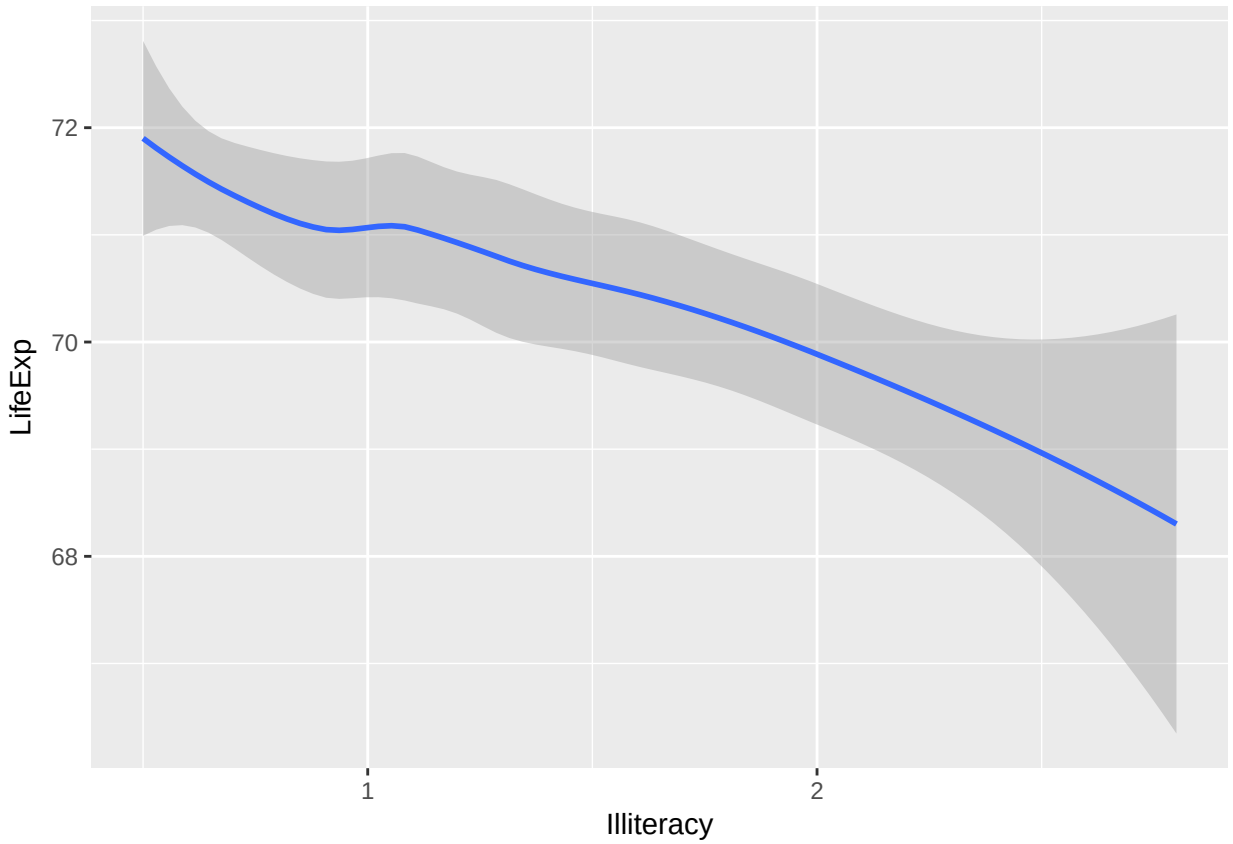
```
ggplot(state_data, aes(x = Illiteracy, y = LifeExp)) +
  geom_text(aes(label = Abbrev))
```



Maybe it's overwhelming to look at all that raw data and you just want a summary. For example, maybe you want an estimate of expected `LifeExp` for each value of `Illiteracy`. This is called the *conditional expectation* and will be the subject of much of the rest of the course. For now, just now that you can calculate a smoothed conditional expectation via `geom_smooth()`.

```
ggplot(state_data, aes(x = Illiteracy, y = LifeExp)) +
  geom_smooth()
```

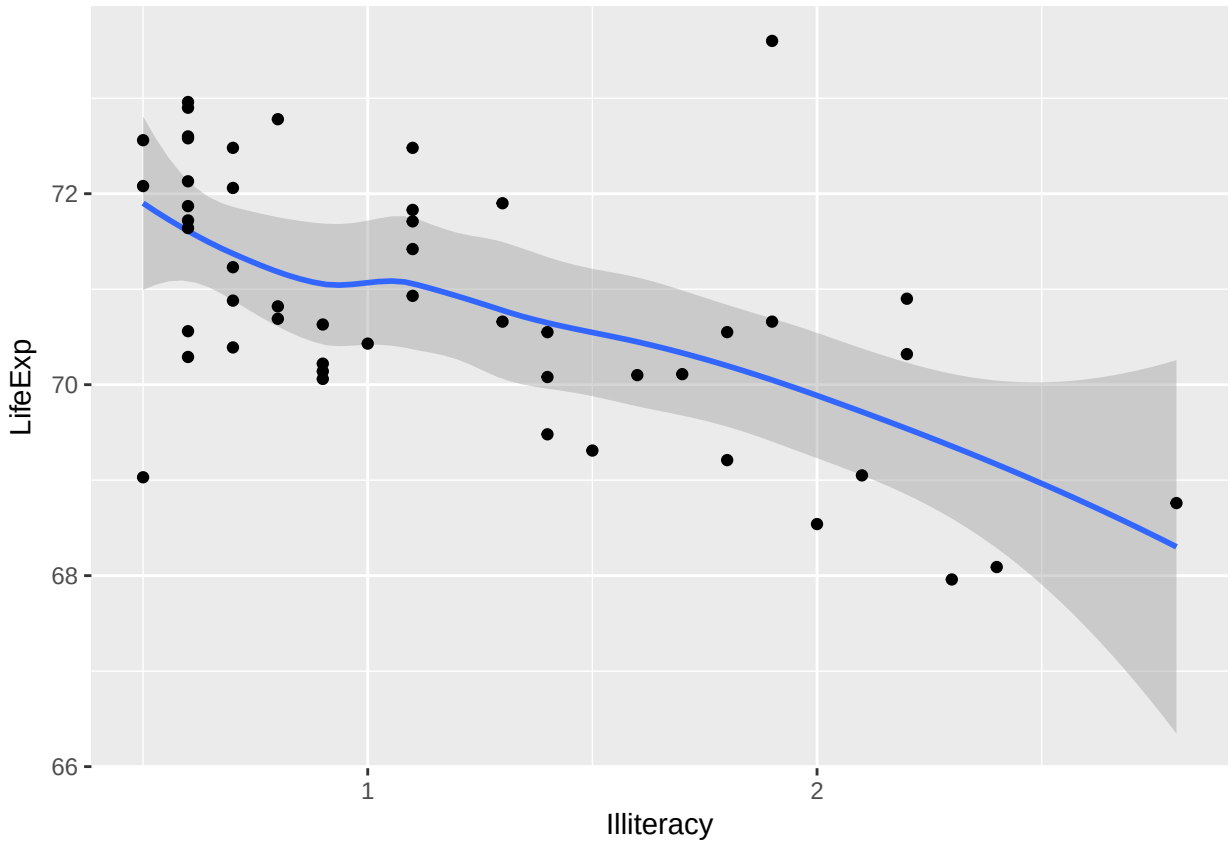
```
## `geom_smooth()` using method = 'loess'
```



And if you're the kind of overachiever who likes to have the raw data *and* the summary, you can do it. Just add them both to the `ggplot()` call.

```
ggplot(state_data, aes(x = Illiteracy, y = LifeExp)) +
  geom_smooth() +
  geom_point()
```

```
## `geom_smooth()` using method = 'loess'
```



4.2 Saving Plots

When you're writing in R Markdown, the plots go straight into your document without much fuss. Odds are, your dissertation will contain plots but won't be written in R Markdown, which means you'll need to learn how to save them.

It's pretty simple:

1. Assign your `ggplot()` call to a variable.
2. Pass that variable to the `ggsave()` function.

```
pop_hist <- ggplot(state_data, aes(x = Population)) +
  geom_histogram()

ggsave(filename = "pop-hist.pdf",
       plot = pop_hist,
       width = 6,
       height = 3)
```

If you want plot types other than PDF, just set a different extension. See `?ggsave` for the possibilities.

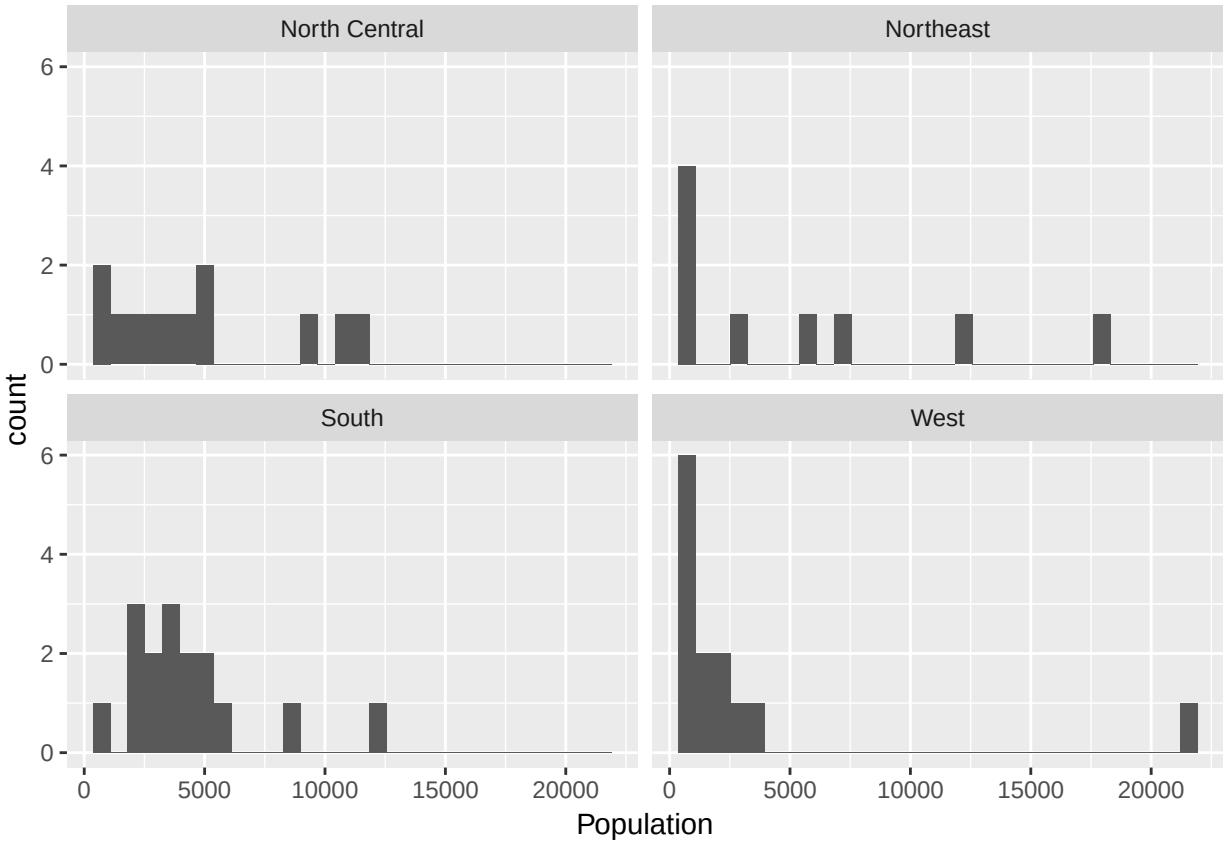
4.3 Faceting

Suppose you want to split the data into subgroups, as defined by some variable in the data (e.g., the region states are in), and make the same plot for each subgroup. ggplot's *faceting* functions, `facet_wrap()` and `facet_grid()`, make this easy.

To split up plots according to a single grouping variable, use `facet_wrap()`. This uses R's *formula* syntax, defined by the tilde `~`, which you'll become well acquainted with once we start running regressions.

```
ggplot(state_data, aes(x = Population)) +
  geom_histogram() +
  facet_wrap(~ Region)
```

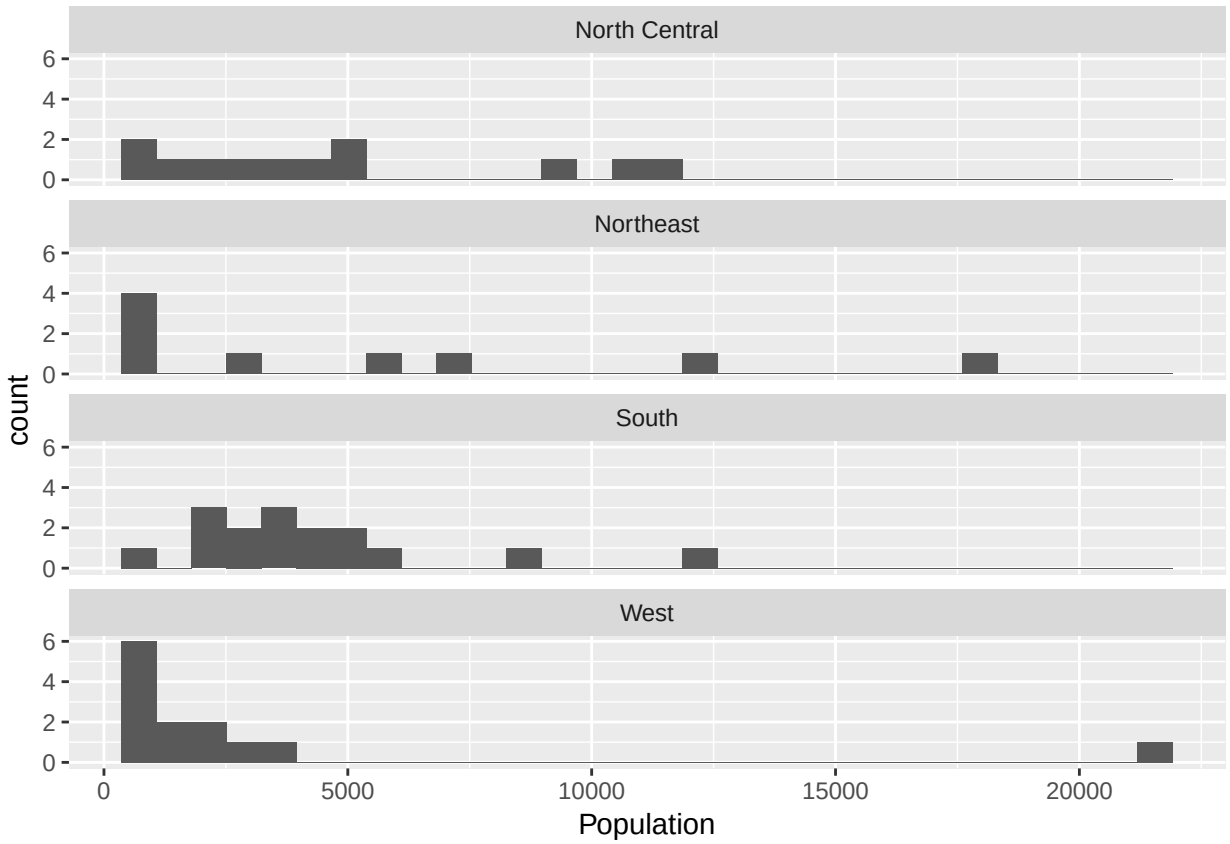
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



If you don't like the default arrangement, use the `ncol` argument.

```
ggplot(state_data, aes(x = Population)) +
  geom_histogram() +
  facet_wrap(~ Region, ncol = 1)
```

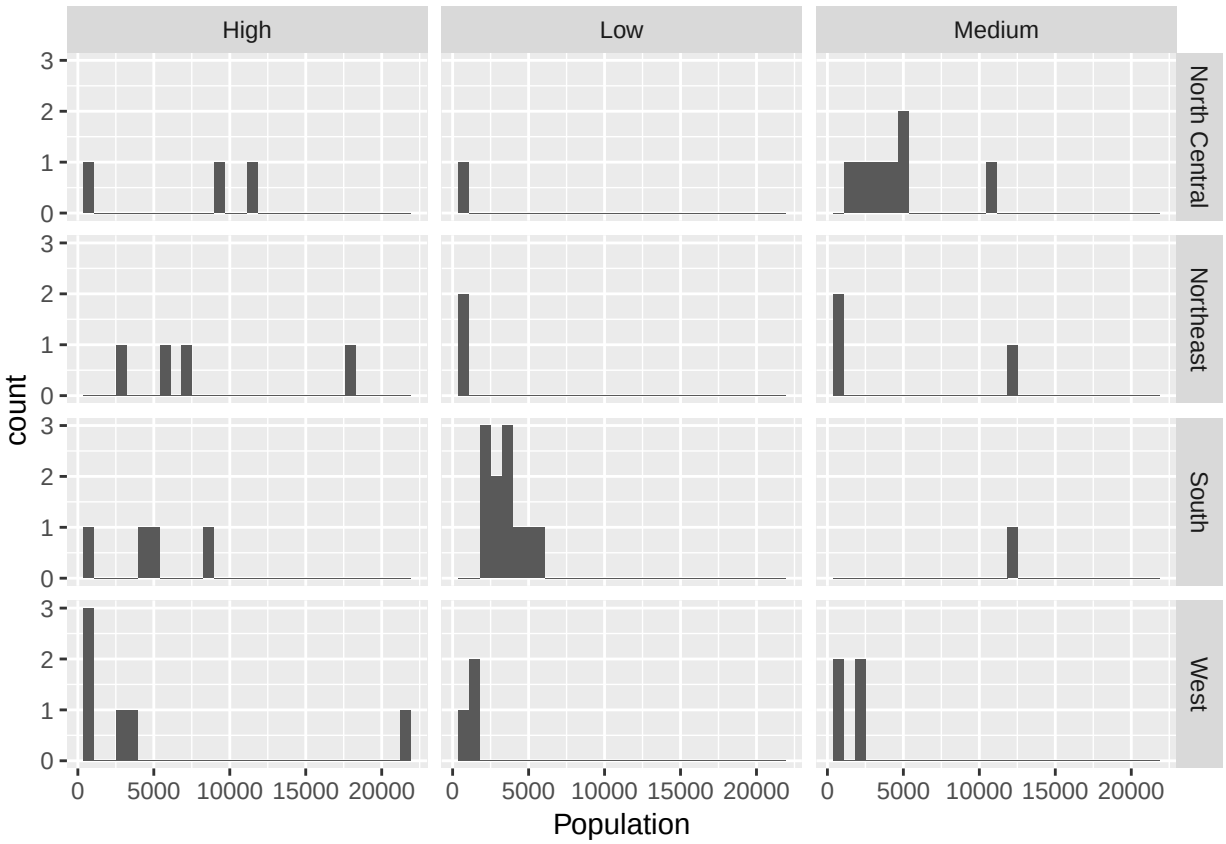
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



For two grouping variables, use `facet_grid()`, putting variables on both sides of the formula.

```
ggplot(state_data, aes(x = Population)) +
  geom_histogram() +
  facet_grid(Region ~ IncomeGroup)
```

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

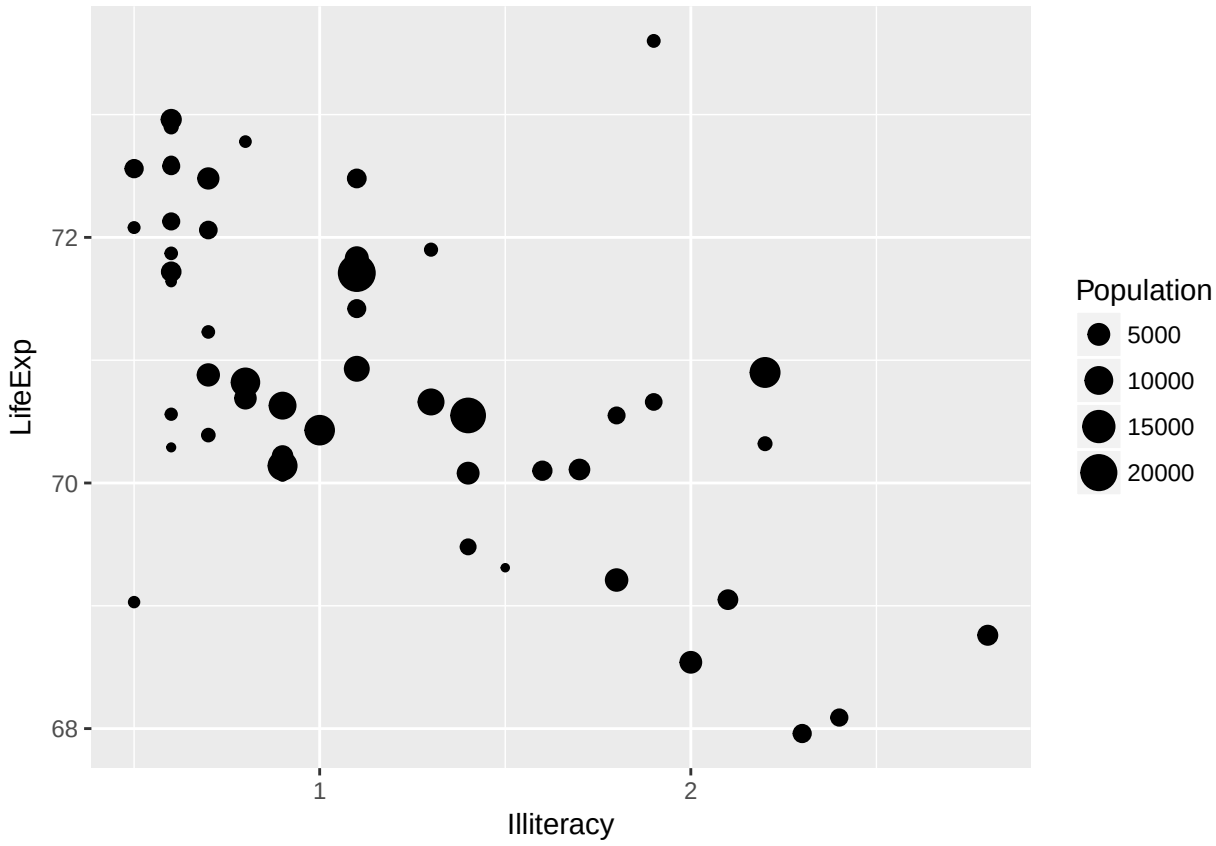


4.4 Aesthetics

Faceting is one way to incorporate information about additional variables into what would otherwise be a plot of just one or two variables. Aesthetics—which alter the appearance of particular plot features depending on the value of a variable—provide another way to do that.

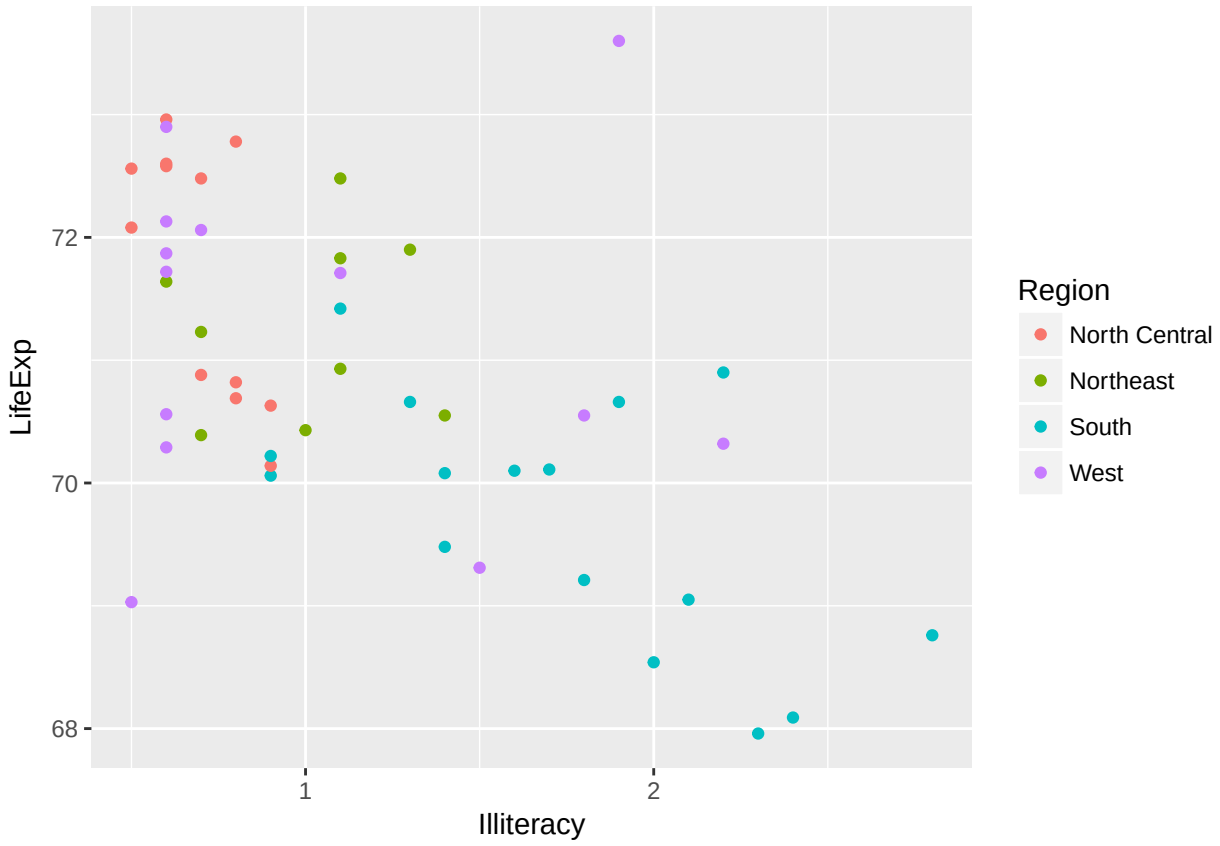
For example, when visualizing the relationship between statewide illiteracy and life expectancy, you might want larger states to get more visual weight. You can set the `size` aesthetic of the `point` geometry to vary according to the state's population.

```
ggplot(state_data, aes(x = Illiteracy, y = LifeExp)) +
  geom_point(aes(size = Population))
```



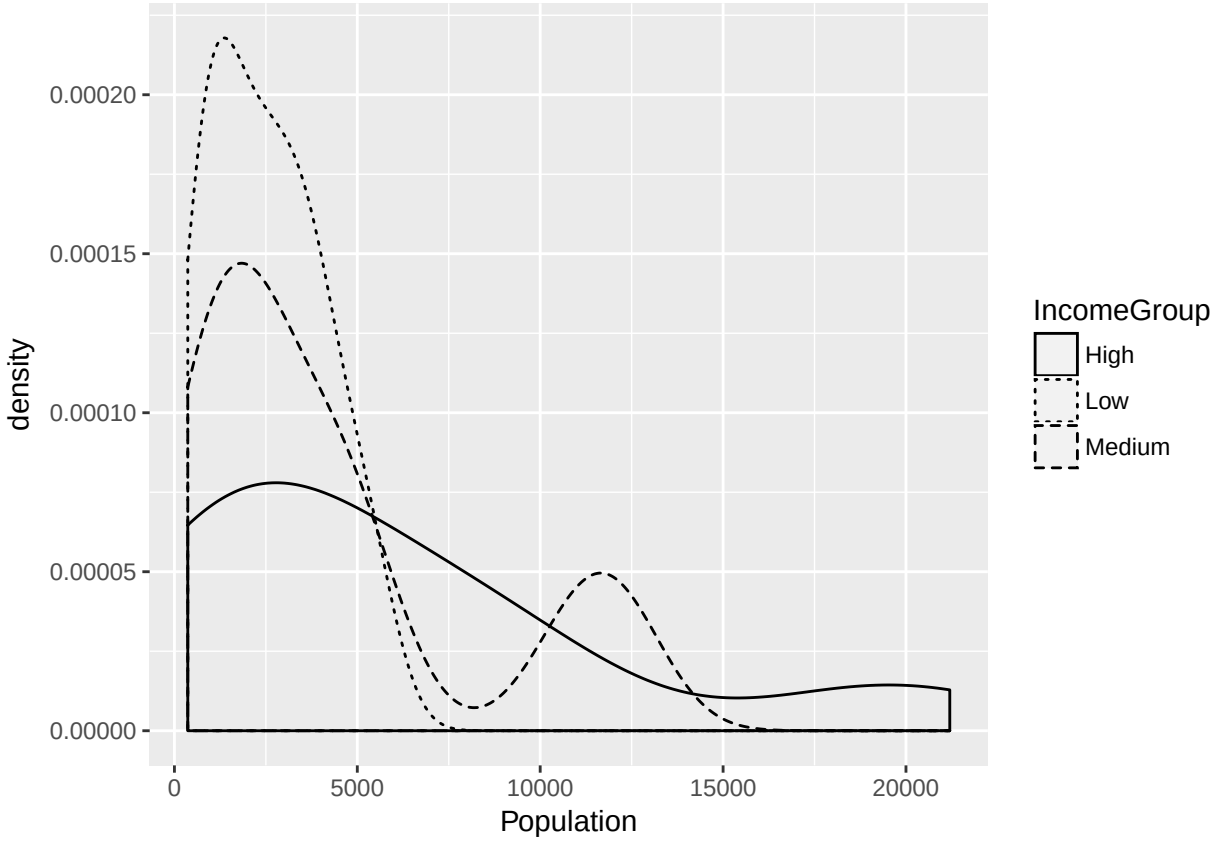
The **ggplot2** documentation lists the available aesthetics for each function. Another popular one is `colour`, which is great for on-screen display but not so much for the printed page. (And terrible for the colorblind!)

```
ggplot(state_data, aes(x = Illiteracy, y = LifeExp)) +
  geom_point(aes(colour = Region))
```



For line graphs or density plots, you can set the `linetype` to vary by category.

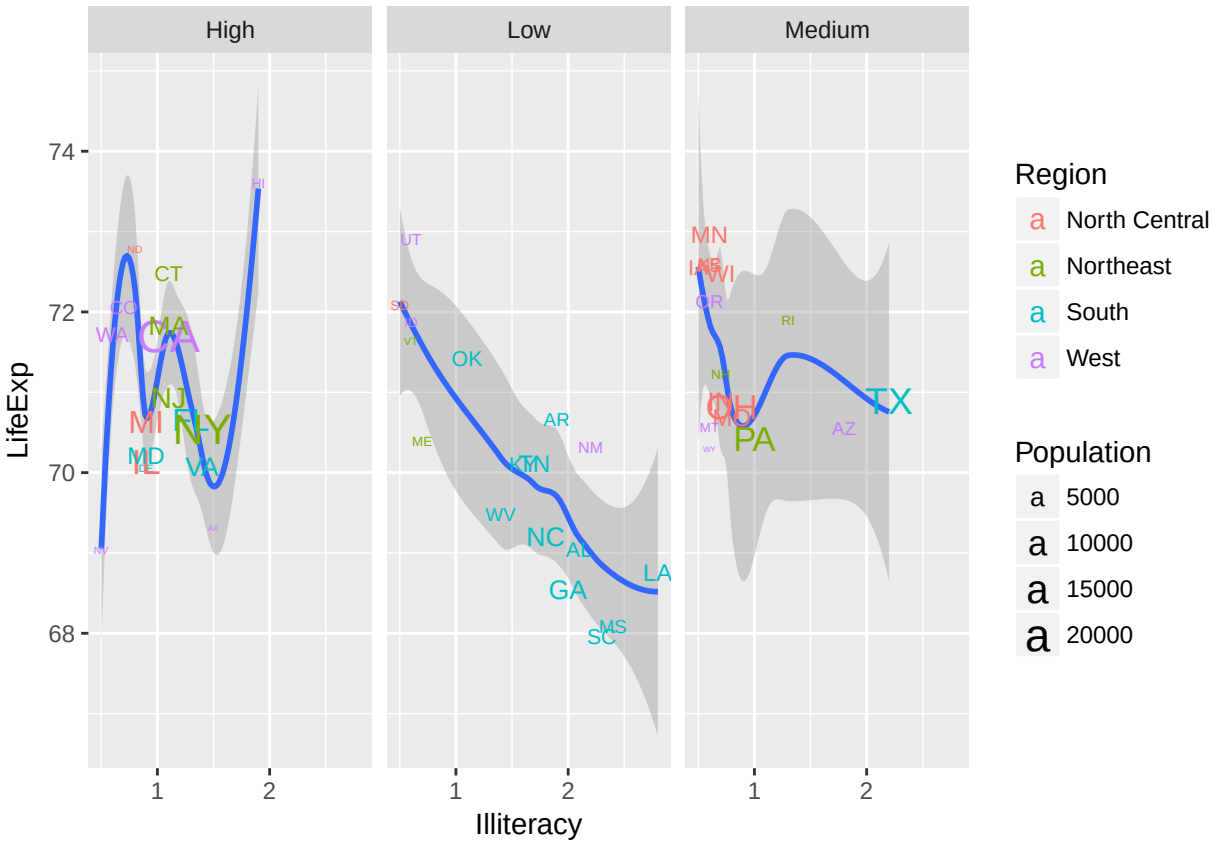
```
ggplot(state_data, aes(x = Population)) +
  geom_density(aes(linetype = IncomeGroup))
```



(I always find these incomprehensible with more than two lines, but maybe that's just me.) You can use multiple aesthetics together, and you can even combine aesthetics with faceting, as in the following example.

```
ggplot(state_data, aes(x = Illiteracy, y = LifeExp)) +
  geom_smooth() +
  geom_text(aes(label = Abbrev, colour = Region, size = Population)) +
  facet_wrap(~ IncomeGroup)
```

```
## `geom_smooth()` using method = 'loess'
```



But the fact that you *can* do something doesn't mean you *should*. That plot is so cluttered that it's hard to extract the relevant information from it. Data visualizations should communicate a clear message to viewers without overwhelming them. To do this well takes practice, patience, and maybe even a bit of taste.

4.5 Appendix: Creating the Example Data

The example data comes from data on U.S. states in 1977 that are included with base R. See `?state`.

```
library("tidyverse")

state_data <- state.x77 %>%
  as_tibble() %>%
  add_column(State = rownames(state.x77),
             Abbrev = state.abb,
             Region = state.region,
             .before = 1) %>%
  rename(LifeExp = `Life Exp`,
         HSGrad = `HS Grad`) %>%
  mutate(IncomeGroup = cut(Income,
```

```
                            breaks = quantile(Income,
                                              probs = seq(0, 1, by = 1/3)),
                            labels = c("Low", "Medium", "High"),
                            include.lowest = TRUE))

write_csv(state_data, path = "state-data.csv")
```

Chapter 5

Bivariate Regression

The goal of empirical social science is usually to learn about the relationships between variables in the social world. Our goals might be descriptive: were college graduates more likely to vote for Clinton in 2016? Or causal: does receiving more education make a person more liberal on average? Or predictive: what kinds of voters should Democrats target in 2020 to have the best chance of victory?

The linear model is one of the simplest ways to model relationships between variables. Ordinary least squares regression is one of the easiest and (often) best ways to estimate the parameters of the linear model. Consequently, a linear model estimated by OLS is the starting point for many analyses. We will start with the simplest case: regression on a single covariate.

5.1 Probability Refresher

Let Y be a random variable that takes values in the finite set $\mathcal{Y}$ according to the probability mass function $f_Y : \mathcal{Y} \to [0, 1]$. The *expected value* (aka *expectation*) of Y is the weighted average of each value in $\mathcal{Y}$, where the weights are the corresponding probabilities:

$$E[Y] = \sum_{y \in \mathcal{Y}} y \, f_Y(y); \tag{5.1}$$

For a continuous random variable Y on $\mathbb{R}$ with probability density function f_Y, the expected value is the analogous integral:

$$E[Y] = \int y \, f_Y(y) \, dy. \tag{5.2}$$

Now suppose (X, Y) is a pair of discrete random variables drawn according to the joint mass

function f_{XY} on $\mathcal{X} \times \mathcal{Y}$, with respective marginal mass functions f_X and f_Y.[1] Recall the formula for conditional probability,

$$\Pr(Y = y \,|\, X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)} = \frac{f_{XY}(x, y)}{f_X(x)}. \tag{5.3}$$

For each $x \in \mathcal{X}$, we have the *conditional mass function*

$$f_{Y|X}(y \,|\, x) = \frac{f_{XY}(x, y)}{f_X(x)} \tag{5.4}$$

and corresponding *conditional expectation*

$$E[Y|X = x] = \sum_{y \in \mathcal{Y}} y \, f_{Y|X}(y \,|\, x). \tag{5.5}$$

For continuous random variables, the conditional expectation is

$$E[Y|X = x] = \int y \, f_{Y|X}(y \,|\, x) \, dy, \tag{5.6}$$

where $f_{Y|X}$ is the conditional density function.

The *variance* of a random variable Y is

$$V[Y] = E[(Y - E[Y])^2]. \tag{5.7}$$

Given a sample $Y_1, \ldots, Y_N$ of observations of Y, we usually estimate $V[Y]$ with the *sample variance*

$$S_Y^2 = \frac{1}{N-1} \sum_n (Y_n - \bar{Y})^2, \tag{5.8}$$

where $\bar{Y}$ is the sample mean and $\sum_n$ denotes summation from $n = 1$ to N.

Similarly (in fact a generalization of the above), the *covariance* between random variables X and Y is

$$\mathrm{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])],$$

which we estimate with the *sample covariance*

$$S_{XY} = \frac{1}{N-1} \sum_n (X_n - \bar{X})(Y_n - \bar{Y}). \tag{5.9}$$

[1] The marginal mass function, if you don't recall, is $f_X(x) = \sum_{y \in \mathcal{Y}} f_{XY}(x, y)$. In the continuous case, the marginal density function is $f_X(x) = \int f_{XY}(x, y) \, dy$.

A fun fact about the sample covariance is that

$$S_{XY} = \frac{1}{N-1}\sum_n (X_n - \bar{X})(Y_n - \bar{Y}) \tag{5.10}$$

$$= \frac{1}{N-1}\left[\sum_n X_n(Y_n - \bar{Y}) + \sum_n \bar{X}(Y_n - \bar{Y})\right] \tag{5.11}$$

$$= \frac{1}{N-1}\left[\sum_n X_n(Y_n - \bar{Y}) + \bar{X}\sum_n (Y_n - \bar{Y})\right] \tag{5.12}$$

$$= \frac{1}{N-1}\sum_n X_n(Y_n - \bar{Y}). \tag{5.13}$$

If we had split up the second term instead of the first, we would see that

$$S_{XY} = \frac{1}{N-1}\sum_n Y_n(X_n - \bar{X})$$

as well.

Since the (sample) variance is a special case of the (sample) covariance, by the same token we have

$$S_Y^2 = \frac{1}{N-1}\sum_n Y_n(Y_n - \bar{Y}). \tag{5.14}$$

5.2 The Linear Model

Suppose we observe a sequence of N draws from f_{XY}, denoted $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$, or $\{(X_n, Y_n)\}_{n=1}^N$ for short. What can we learn about the relationship between X and Y from this sample of data?

If we were really ambitious, we could try to estimate the shape of the full joint distribution, f_{XY}. The joint distribution encodes everything there is to know about the relationship between the two variables, so it would be pretty useful to know. But except in the most trivial cases, it would be infeasible to estimate f_{XY} precisely. If X or Y can take on more than a few values, estimating the joint distribution would require an amount of data that we're unlikely to have.[2]

The first way we simplify our estimation task is to set our sights lower. Let Y be the *response* or the *dependent variable*—i.e., the thing we want to explain. We call X the *covariate* or the *independent variable*. Instead of estimating the full joint distribution, we're just going to try to learn the conditional expectation, $E[Y \mid X]$. In other words, for each potential value

[2]This problem only gets worse as we move from bivariate into multivariate analysis, a phenomenon called the *curse of dimensionality*.

of the covariate, what is the expected value of the response? This will allow us to answer questions like whether greater values of X are associated with greater values of Y.

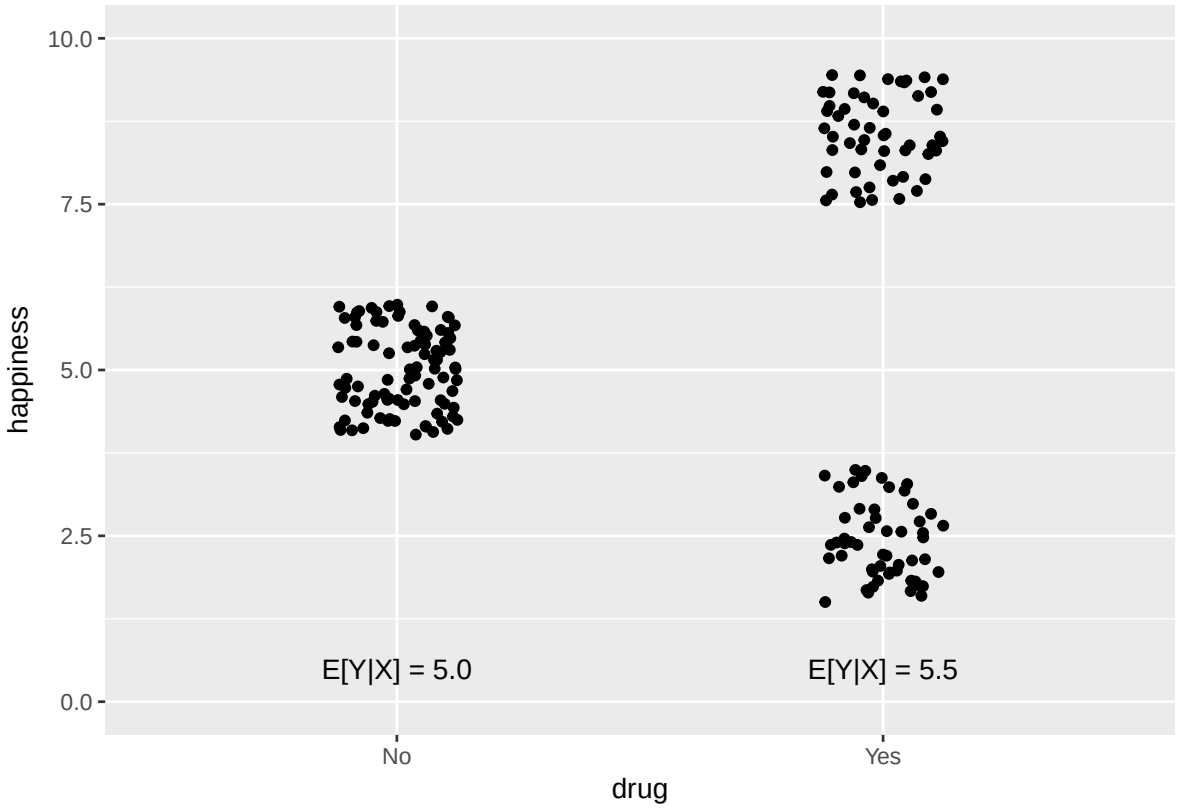
Two important things about the estimation of conditional expectations before we go any further.

1. Statements about conditional expectations are not causal. If Y is rain and X is umbrella sales, we know $E[Y|X]$ increases with X, but that doesn't mean umbrella sales make it rain.

   We will spend some time in the latter part of the course on how to move from conditional expectations to causality. Then, in Stat III, you will learn about causal inference in excruciating detail.

2. The conditional expectation doesn't give you everything you'd want to know about the relationship between variables.

   As a hypothetical example, suppose I told you that taking a particular drug made people happier on average. In other words, $E[\text{Happiness} \mid \text{Drug}] > E[\text{Happiness} \mid \text{No Drug}]$. Sounds great! Then imagine the dose-response graph looked like this:



   The fact that expected happiness rises by half a point doesn't quite tell the whole story.

In spite of these caveats, conditional expectation is a really useful tool for summarizing the relationship between variables.

If X takes on sufficiently few values (and we have enough data), we don't need to model the

conditional expectation function. We can just directly estimate $E[Y|X = x]$ for each $x \in \mathcal{X}$. The graph above, where there are just two values of X, is one example.

But if X is continuous, or even if it is discrete with many values, estimating $E[Y|X]$ for each distinct value is infeasible. In this case, we need to *model* the relationship. The very simplest choice—and thus the default for social scientists—is to model the conditional expectation of Y as a linear function of X:

$$E[Y \mid X] = \alpha + \beta X. \tag{5.15}$$

In this formulation, α and β are the parameters to be estimated from sample data. We call α and β "coefficients," with α the "intercept" and β the "slope." Regardless of how many different values X might take on, we only need to estimate two parameters of the linear model.

Exercise your judgment before using a linear model. Ask yourself, is a linear conditional expectation function at least minimally plausible? Not perfect—just a reasonable approximation. If X is years of education and Y is annual income, the answer is probably yes (depending on the population!). But if X is hour of the day (0–24) and Y is the amount of traffic on I-65, probably not.

To obtain the linear conditional expectation, we usually assume the following model of the response variable:

$$Y_n = \alpha + \beta X_n + \epsilon_n, \tag{5.16}$$

where ϵ_n is "white noise" error with the property

$$E[\epsilon_n \mid X_1, \ldots X_N] = 0. \tag{5.17}$$

You can think of ϵ_n as the summation of everything besides the covariate X_n that affects the response Y_n. The assumption that $E[\epsilon_n \mid X_1, \ldots, X_N] = 0$ implies that these external factors are uncorrelated with the covariate. This is not a trivial technical condition that you can ignore—it is a substantive statement about the variables in your model. It requires justification, and it is difficult to justify.

For now we will proceed assuming that our data satisfy the above conditions. Later in the course, we will talk about how to proceed when $E[\epsilon_n \mid X_1, \ldots, X_N] \neq 0$, and you will learn much more about such strategies in Stat III.

5.3 Least Squares

To estimate the parameters of the linear model, we will rely on a mathematically convenient method called *least squares*. We will see that this method not only is convenient, but also

has nice statistical properties.

Given a parameter estimate $(\hat{\alpha}, \hat{\beta})$, define the *residual* of the n'th observation as the difference between the true and predicted values:

$$e_n(\hat{\alpha}, \hat{\beta}) = Y_n - \hat{\alpha} - \hat{\beta} X_n. \tag{5.18}$$

The residual is directional. The residual is positive when the regression line falls below the observation, and vice versa when it is negative.

We would like the regression line to lie close to the data—i.e., for the residuals to be small in magnitude. "Close" can mean many things, so we need to be a bit more specific to derive an estimator. The usual one, *ordinary least squares*, is chosen to minimize the sum of squared errors,

$$\text{SSE}(\hat{\alpha}, \hat{\beta}) = \sum_n e_n(\hat{\alpha}, \hat{\beta})^2.$$

(Throughout the rest of this chapter, I write $\sum_n$ as shorthand for $\sum_{n=1}^N$.) When we focus on squared error, we penalize a positive residual the same as a negative residual of the same size. Moreover, we penalize one big residual proportionally more than a few small ones.

It is important to keep the linear model and ordinary least squares distinct in your mind. The linear model is a model of the data. Ordinary least squares is one estimator—one among many—of the parameters of the linear model. Assuming a linear model does not commit you to estimate it with OLS if you think another estimator is more appropriate. And using OLS does not necessarily commit you to the linear model, as we will discuss when we get to multiple regression.

To derive the OLS estimator, we will derive the conditions for minimization of the sum of squared errors. The SSE is a quadratic and therefore continuously differentiable function of the estimands, $\hat{\alpha}$ and $\hat{\beta}$. You will remember from calculus that, at any extreme point of a continuous function, all its partial derivatives equal zero. To derive necessary conditions for minimization,[3] we can take the derivatives of the SSE and set them to equal zero.

The derivative with respect to the intercept is

$$\frac{\partial \, \text{SSE}(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} = -2 \sum_n (Y_n - \hat{\alpha} - \hat{\beta} X_n).$$

Setting this to equal zero gives

$$\hat{\alpha} = \frac{1}{N} \sum_n (Y_n - \hat{\beta} X_n) = \bar{Y} - \hat{\beta} \bar{X}.$$

This gives us one important property of OLS: the regression line estimated by OLS always passes through $(\bar{X}, \bar{Y})$.

[3] In fact, since the SSE function is strictly convex, these conditions are sufficient for global minimization.

The derivative with respect to the slope is

$$\frac{\partial \, \text{SSE}(\hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} = -2 \sum_n X_n (Y_n - \hat{\alpha} - \hat{\beta} X_n).$$

Setting this equal to zero and substituting in the expression for $\hat{\alpha}$ we derived above gives

$$\sum_n X_n (Y_n - \bar{Y}) = \hat{\beta} \sum_n X_n (X_n - \bar{X}).$$

As long as the sample variance of X is non-zero (i.e., X is not a constant), we can divide to solve for $\hat{\beta}$:

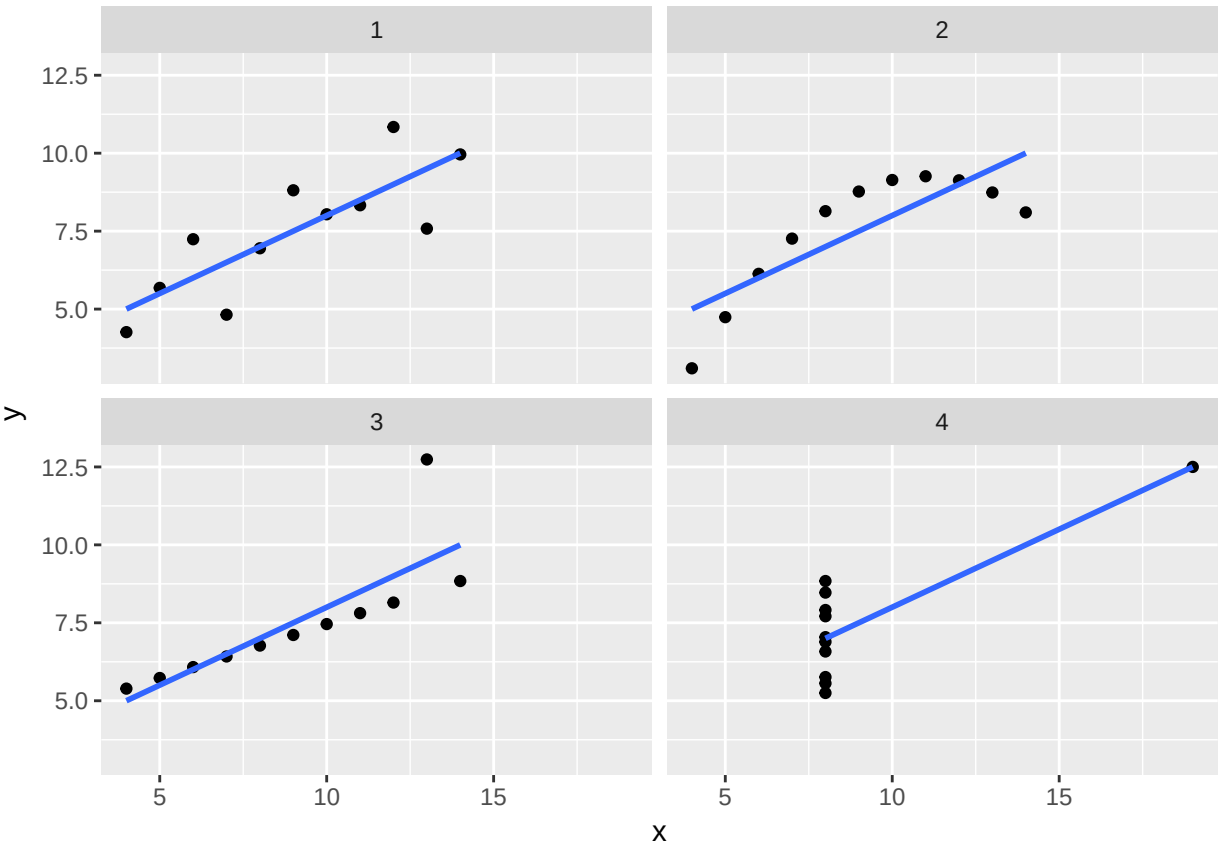
$$\hat{\beta} = \frac{\sum_n X_n (Y_n - \bar{Y})}{\sum_n X_n (X_n - \bar{X})} = \frac{S_{XY}}{S_X^2}.$$

Combining these two results, we have the OLS estimators of the intercept and slope of the bivariate linear model. We write them as functions of $(X_1, \ldots, X_N, Y_1, \ldots, Y_N)$, or (X, Y) for short,[4] to emphasize that an estimator is a statistic, which in turn is a function of sample data. We place the "OLS" subscript on them to emphasize that there are many estimators of these parameters, of which OLS is just one (good!) choice.

$$\hat{\alpha}_{\text{OLS}}(X, Y) = \bar{Y} - \frac{S_{XY}}{S_X^2} \bar{X},$$
$$\hat{\beta}_{\text{OLS}}(X, Y) = \frac{S_{XY}}{S_X^2}.$$

Regression is a convenient way to summarize the relationship between variables, but it is a complement to—not a substitute for—graphical analysis. The statistician Francis Anscombe found that OLS yields nearly identical regression lines for all four of the datasets in the following graph:

[4] This is a bit of an abuse of notation, since previously I used X and Y to refer to the random variables and now I'm using them to refer to vectors of sample data. Sorry.



Unless your data all lie along a line, the regression line estimated by OLS will not predict the data perfectly. Let the *residual sum of squares* be the squared error left over by OLS,

$$\text{RSS} = \text{SSE}(\hat{\alpha}_{\text{OLS}}, \hat{\beta}_{\text{OLS}}),$$

and let the *total sum of squares* be the squared error that would result from a horizontal regression line through the mean of Y,

$$\text{TSS} = \text{SSE}(\bar{Y}, 0).$$

The R^2 statistic is the proportion of "variance explained" by X, calculated as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

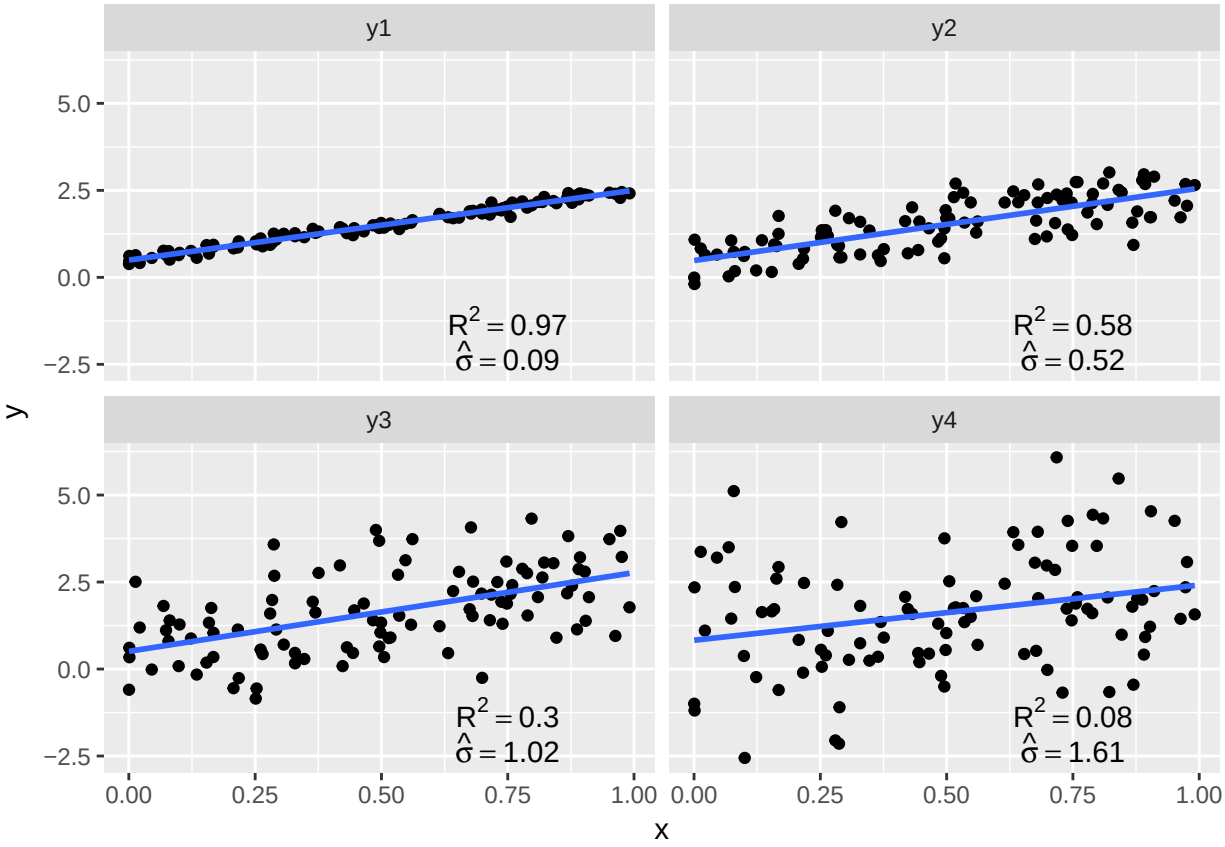
If the regression line is flat, in which case $\hat{\beta}_{\text{OLS}} = 0$ and RSS = TSS, we have $R^2 = 0$. Conversely, if the regression line fits perfectly, in which case RSS = 0, we have $R^2 = 1$.

A statistic that is often more useful than R^2 is the *residual variance*. The residual variance is (almost) the sample variance of the regression residuals, calculated as

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_n e_n(\hat{\alpha}_{\text{OLS}}, \hat{\beta}_{\text{OLS}})^2 = \frac{\text{RSS}}{N-2}$$

Since bivariate regression uses two degrees of freedom (one for the intercept, one for the slope), we divide by $N - 2$ instead of the usual $N - 1$. The most useful quantity is $\hat{\sigma}$,

the square root of the residual variance. $\hat{\sigma}$ is measured in the same units as Y, and it is a measure of the spread of points around the regression line. If the residuals are roughly normally distributed, then we would expect roughly 95% of the data to lie within $\pm 2\hat{\sigma}$ of the regression line.



5.4 Properties

We didn't use any fancy statistical theory to derive the OLS estimator. We just found the intercept and slope that minimize the sum of squared residuals. As it turns out, though, OLS indeed has some very nice statistical properties as an estimator of the linear model.

The first desirable property of OLS is that it is *unbiased*. Recall that an estimator $\hat{\theta}$ of the parameter θ is unbiased if $E[\hat{\theta}] = \theta$. This doesn't mean the estimator always gives us the right answer, just that on average it is not systematically biased upward or downward. In other words, if we could take many many samples and apply the estimator to each of them, the average would equal the true parameter.

We will begin by showing that the OLS estimator of the slope is unbiased; i.e., that $E[\hat{\beta}_{\text{OLS}}(X, Y)] = \beta$. At first, we'll take the conditional expectation of the slope estimator,

treating the covariates $(X_1, \ldots, X_N)$ as fixed.

$$\begin{aligned}
E[\hat{\beta}_{\text{OLS}}(X, Y) \mid X] &= E\left[\frac{S_{XY}}{S_X^2} \,\middle|\, X \right] \\
&= E\left[\frac{\sum_n Y_n (X_n - \bar{X})}{\sum_n X_n (X_n - \bar{X})} \,\middle|\, X \right] \\
&= \frac{\sum_n E[Y_n \mid X](X_n - \bar{X})}{\sum_n X_n (X_n - \bar{X})} \\
&= \frac{\sum_n (\alpha + \beta X_n)(X_n - \bar{X})}{\sum_n X_n (X_n - \bar{X})} \\
&= \frac{\alpha \sum_n (X_n - \bar{X}) + \beta \sum_n X_n (X_n - \bar{X})}{\sum_n X_n (X_n - \bar{X})} \\
&= \frac{\beta \sum_n X_n (X_n - \bar{X})}{\sum_n X_n (X_n - \bar{X})} \\
&= \beta.
\end{aligned}$$

It then follows from the *law of iterated expectation*[5] that

$$E[\hat{\beta}_{\text{OLS}}(X, Y)] = \beta.$$

Then, for the intercept, we have

$$\begin{aligned}
E[\hat{\alpha}_{\text{OLS}}(X, Y) \mid X] &= E[\bar{Y} - \hat{\beta}_{\text{OLS}}(X, Y)\bar{X} \mid X] \\
&= E[\bar{Y} \mid X] - E[\hat{\beta}_{\text{OLS}}(X, Y) \mid X]\bar{X} \\
&= E\left[\frac{1}{N} \sum_n Y_n \,\middle|\, X \right] - \beta \bar{X} \\
&= E\left[\frac{1}{N} \sum_n (\alpha + \beta X_n + \epsilon_n) \,\middle|\, X \right] - \beta \bar{X} \\
&= \frac{1}{N} \sum_n E[\alpha + \beta X_n + \epsilon_n \mid X] - \beta \bar{X} \\
&= \frac{1}{N} \sum_n \alpha + \frac{\beta}{N} \sum_n X_n + \frac{1}{N} \sum_n E[\epsilon_n \mid X] - \beta \bar{X} \\
&= \alpha + \beta \bar{X} - \beta \bar{X} \\
&= \alpha.
\end{aligned}$$

As with the slope, this conditional expectation gives us the unconditional expectation we want:

$$E[\hat{\alpha}_{\text{OLS}}(X, Y)] = \alpha.$$

To sum up: as long as the crucial condition $E[\epsilon_n \mid X_1, \ldots, X_N] = 0$ holds, then OLS is an unbiased estimator of the parameters of the linear model.

[5]For random variables A and B, $E[f(A, B)] = E_A[E_B[f(A, B) \mid A]] = E_B[E_A[f(A, B) \mid B]]$.

Another important property of OLS is that it is *consistent*. Informally, this means that in sufficiently large samples, the OLS estimates $(\hat{\alpha}_{\text{OLS}}, \hat{\beta}_{\text{OLS}})$ are very likely to be close to the true parameter values (α, β). Another way to think of consistency is that, as $N \to \infty$, the bias and variance of the OLS estimator both go to zero.[6]

Of course the bias "goes to" zero, since OLS is unbiased. The real trick to proving consistency is to show that the variance goes to zero. If you wanted to do that for the slope estimate, you'd derive an expression for

$$V[\hat{\beta}_{\text{OLS}}] = E[(\hat{\beta}_{\text{OLS}} - E[\hat{\beta}_{\text{OLS}}])^2] = E[(\hat{\beta}_{\text{OLS}} - \beta)^2]$$

and show that

$$\lim_{N \to \infty} V[\hat{\beta}_{\text{OLS}}] = 0.$$

This takes more algebra than we have time for, so I leave it as an exercise for the reader.

5.5 Appendix: Regression in R

We will be using the **tidyverse** package as always, the **car** package for the `Prestige` data, and the **broom** package for its convenient post-analysis functions.

```
library("tidyverse")
library("car")
library("broom")
```

Let's take a look at `Prestige`, which records basic information (including perceived prestige) for a variety of occupations.

```
head(Prestige)
```

```
##                     education income women prestige census type
## gov.administrators      13.11  12351 11.16     68.8   1113 prof
## general.managers        12.26  25879  4.02     69.1   1130 prof
## accountants             12.77   9271 15.70     63.4   1171 prof
## purchasing.officers     11.42   8865  9.11     56.8   1175 prof
## chemists                14.62   8403 11.68     73.5   2111 prof
## physicists              15.64  11030  5.13     77.6   2113 prof
```

Suppose we want to run a regression of prestige on education. We will use the `lm()` function, which stands for *linear model*. This will employ the "formula" syntax that you previously saw when faceting in ggplot. The basic syntax of a formula is `response ~ covariate`, where `response` and `covariate` are the names of the variables in question. In this case, with `prestige` (note that the variable is lowercase, while the dataset is capitalized) as the response and education as the covariate:

[6] What I am describing here is *mean square consistency*, which is stronger than the broadest definitions of consistency in statistical theory.

```
lm(prestige ~ education, data = Prestige)
```

```
##
## Call:
## lm(formula = prestige ~ education, data = Prestige)
##
## Coefficients:
## (Intercept)    education
##      -10.73         5.36
```

You'll notice that didn't give us very much. If you've previously used statistical programs like Stata, you might expect a ton of output at this point. It's all there in R too, but R has a different philosophy about models. R sees the fitted model as an object in its own right—like a data frame, a function, or anything else you load or create in R. Therefore, to analyze regression results in R, you will typically save the regression results to a variable.

Like any other variable, you'll want to give your regression results meaningful names. I typically call them `fit_` to indicate a fitted model, followed by some memorable description.

```
fit_educ <- lm(prestige ~ education, data = Prestige)
```

When you do this, the output doesn't get printed. To see the default output, just run the variable name, just like you would to see the content of a data frame:

```
fit_educ
```

```
##
## Call:
## lm(formula = prestige ~ education, data = Prestige)
##
## Coefficients:
## (Intercept)    education
##      -10.73         5.36
```

For a more detailed readout, use the `summary()` method:

```
summary(fit_educ)
```

```
##
## Call:
## lm(formula = prestige ~ education, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.040  -6.523   0.661   6.743  18.164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -10.732      3.677   -2.92   0.0043
## education      5.361      0.332   16.15   <2e-16
##
## Residual standard error: 9.1 on 100 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.72
## F-statistic:  261 on 1 and 100 DF,  p-value: <2e-16
```

This prints out a whole boatload of information, including inferential statistics that we're going to wait until later in the course to discuss how to interpret:

- The model you ran
- Basic statistics about the distribution of the residuals
- For each coefficient:
  - Parameter estimate
  - Standard error estimate
  - Test statistic for a hypothesis test of equality with zero
  - p-value associated with the test statistic
- $\hat{\sigma}$ (called the "residual standard error", a term seemingly unique to R)
- R^2 and an "adjusted" variant that accounts for the number of variables in the model
- F statistic, degrees of freedom, and associated p-value for a hypothesis test that every coefficient besides the intercept equals zero

Strangely, `summary()` doesn't give you the sample size. For that you must use `nobs()`:

```
nobs(fit_educ)
```

```
## [1] 102
```

You can use a fitted model object to make predictions for new data. For example, let's make a basic data frame of education levels.

```
my_data <- data_frame(education = 8:16)
my_data
```

```
## # A tibble: 9 × 1
##   education
##       <int>
## 1         8
## 2         9
## 3        10
## 4        11
## 5        12
## 6        13
## 7        14
## 8        15
## 9        16
```

To calculate the predicted level of prestige for each education level, use `predict()`:

```
predict(fit_educ, newdata = my_data)
```

```
##      1      2      3      4      5      6      7      8      9
## 32.155 37.516 42.877 48.238 53.599 58.959 64.320 69.681 75.042
```

When using `predict()`, it is crucial that the `newdata` have the same column names as in the data used to fit the model.

You can also extract a confidence interval for each prediction:

```
predict(fit_educ,
        newdata = my_data,
        interval = "confidence",
        level = 0.95)
```

```
##      fit    lwr    upr
## 1 32.155 29.615 34.695
## 2 37.516 35.393 39.639
## 3 42.877 41.024 44.730
## 4 48.238 46.441 50.034
## 5 53.599 51.627 55.571
## 6 58.959 56.632 61.287
## 7 64.320 61.525 67.116
## 8 69.681 66.353 73.010
## 9 75.042 71.142 78.942
```

One of the problems with `summary()` and `predict()` is that they return inconveniently shaped output. The output of `summary()` is particularly hard to deal with. The **broom** package provides three utilities to help get model output into shape. The first is `tidy()`, which makes a tidy data frame out of the regression coefficients and the associated inferential statistics:

```
tidy(fit_educ)
```

```
##          term estimate std.error statistic    p.value
## 1 (Intercept) -10.7320   3.67709   -2.9186 4.3434e-03
## 2   education   5.3609   0.33199   16.1478 1.2863e-29
```

The second is `glance()`, which provides a one-row data frame containing overall model characteristics (e.g., R^2 and $\hat{\sigma}$):

```
glance(fit_educ)
```

```
##   r.squared adj.r.squared  sigma statistic    p.value df logLik    AIC
## 1    0.7228       0.72003 9.1033    260.75 1.2863e-29  2   -369 744.01
##      BIC deviance df.residual
## 1 751.88     8287         100
```

The third is `augment()`, which "augments" the original data—or new data you supply, as in

`predict()`—with information from the model, such as predicted values.

```
# Lots of output, so only printing first 10 rows
head(augment(fit_educ), 10)
```

```
##              .rownames prestige education .fitted .se.fit  .resid     .hat
## 1   gov.administrators     68.8     13.11  59.549 1.19689  9.2509 0.017287
## 2     general.managers     69.1     12.26  54.992 1.03332 14.1076 0.012885
## 3          accountants     63.4     12.77  57.726 1.12584  5.6736 0.015295
## 4  purchasing.officers     56.8     11.42  50.489 0.92936  6.3108 0.010422
## 5             chemists     73.5     14.62  67.644 1.57269  5.8559 0.029846
## 6           physicists     77.6     15.64  73.112 1.86034  4.4879 0.041763
## 7           biologists     72.6     15.09  70.164 1.70291  2.4363 0.034993
## 8           architects     78.1     15.44  72.040 1.80254  6.0600 0.039208
## 9      civil.engineers     73.1     14.52  67.108 1.54561  5.9920 0.028827
## 10    mining.engineers     68.8     14.64  67.751 1.57814  1.0487 0.030053
##    .sigma    .cooksd .std.resid
## 1  9.1010 0.00924267    1.02511
## 2  9.0372 0.01587881    1.55981
## 3  9.1311 0.00306360    0.62807
## 4  9.1269 0.00255745    0.69688
## 5  9.1296 0.00656112    0.65310
## 6  9.1375 0.00552702    0.50362
## 7  9.1458 0.00134578    0.27244
## 8  9.1280 0.00941104    0.67914
## 9  9.1287 0.00662109    0.66793
## 10 9.1485 0.00021198    0.11697
```

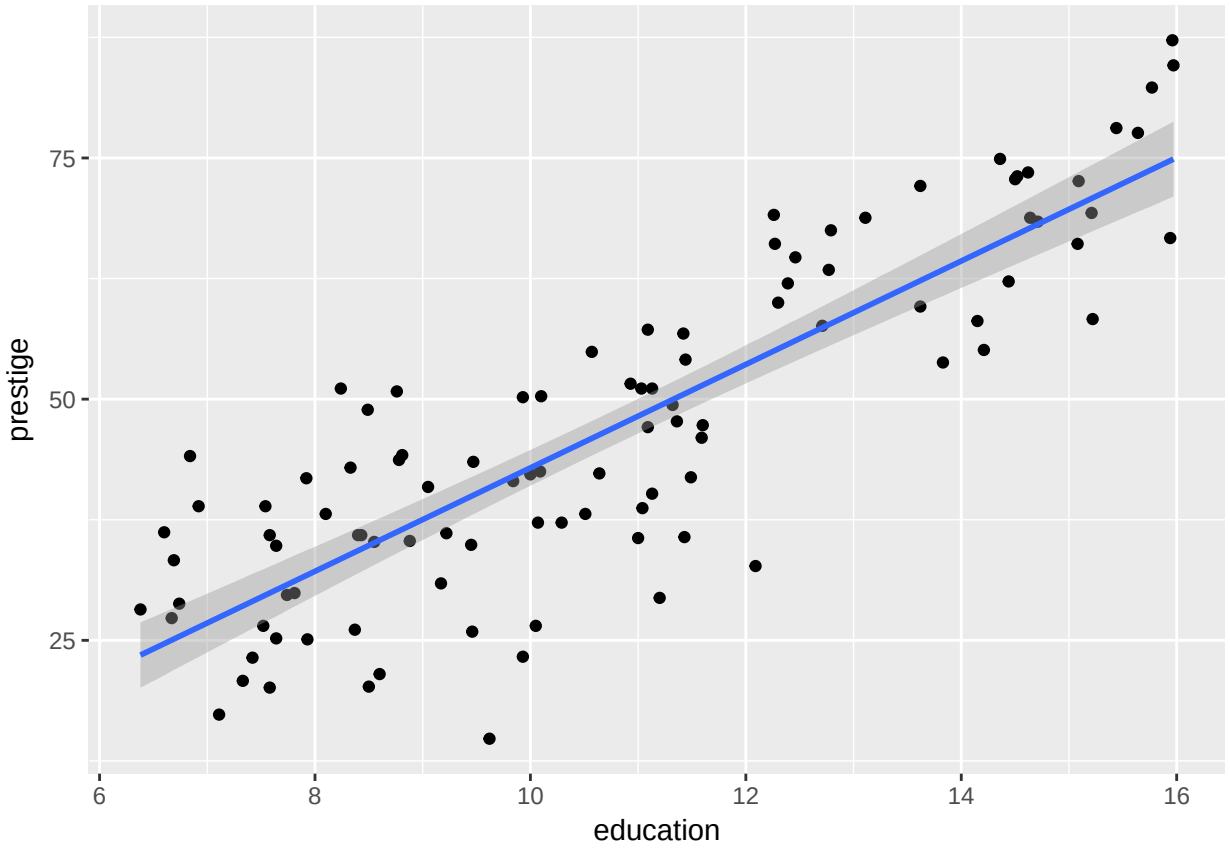
```
augment(fit_educ,
        newdata = my_data)
```

```
##   education .fitted .se.fit
## 1         8  32.155 1.28013
## 2         9  37.516 1.07023
## 3        10  42.877 0.93407
## 4        11  48.238 0.90555
## 5        12  53.599 0.99397
## 6        13  58.959 1.17319
## 7        14  64.320 1.40897
## 8        15  69.681 1.67763
## 9        16  75.042 1.96574
```

Notice that you get back more information for the data used to fit the model than for newly supplied data. The most important is `.fitted`, the predicted value. See `?augment.lm` for what all the various output represents.

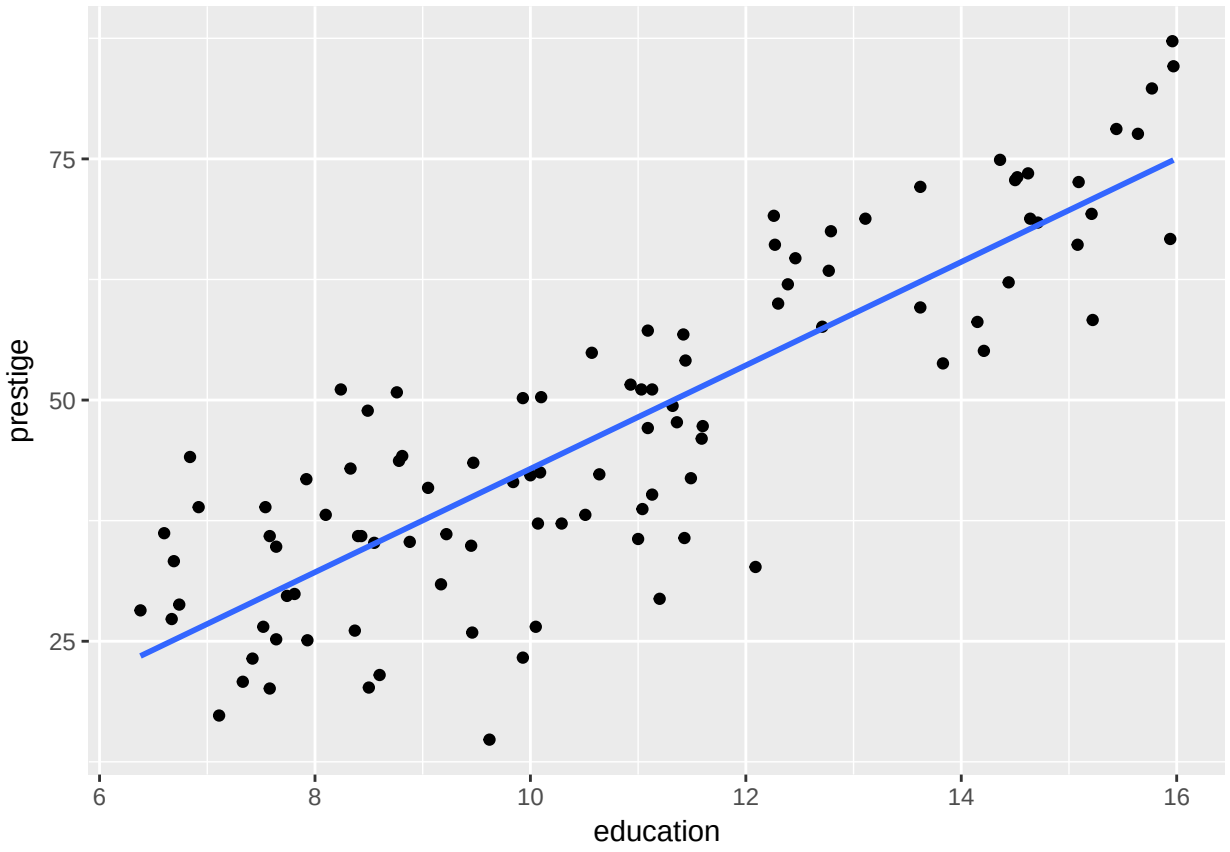
One last note on plotting regression lines with ggplot. Use `geom_smooth(method = "lm")`.

```
ggplot(Prestige, aes(x = education, y = prestige)) +
  geom_point() +
  geom_smooth(method = "lm")
```



To get rid of the confidence interval:

```
ggplot(Prestige, aes(x = education, y = prestige)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```


prestige
75
50
25
6
8
10
12
14
16
education

Chapter 6

Matrix Algebra: A Crash Course

Some material in this chapter is adapted from notes Hye Young You wrote for the math boot camp for the political science PhD program at Vanderbilt.

Matrix algebra is an essential tool for understanding multivariate statistics. You are probably already familiar with matrices, at least informally. The data representations we have worked with so far—each row an observation, each column a variable—are formatted like matrices.

An introductory treatment of matrix algebra is a semester-long college course. We don't have that long, or even half that long. This chapter gives you the *bare minimum* you need to understand to get up and running with the matrix algebra we need for OLS with multiple covariates. If you want to use advanced statistical methods in your research and haven't previously taken a matrix algebra or linear algebra course, I recommend taking some time this summer to catch up. For example, MIT has its undergraduate linear algebra course available online, including video lectures.

6.1 Vector Operations

A *vector* is an ordered array. To denote a vector v of k elements, we write $\mathbf{v} = (v_1, v_2, \ldots, v_k)$, or sometimes

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix}.$$

Notice the convention of using a lowercase bold letter to denote a vector. We will usually be dealing with vectors of real numbers. To denote the fact that $\mathbf{v}$ is a vector of k real numbers, we write $\mathbf{v} \in \mathbb{R}^k$.

A vector can be multiplied by a scalar $c \in \mathbb{R}$, producing what you would expect:

$$c\mathbf{v} = \begin{pmatrix} cv_1 \\ cv_2 \\ \vdots \\ cv_k \end{pmatrix}$$

You can also add and subtract two vectors of the same length.[1]

$$\mathbf{u} + \mathbf{v} = \begin{pmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_k + v_k \end{pmatrix},$$

$$\mathbf{u} - \mathbf{v} = \begin{pmatrix} u_1 - v_1 \\ u_2 - v_2 \\ \vdots \\ u_k - v_k \end{pmatrix}.$$

A special vector is the *zero vector*, which contains—you guessed it—all zeroes. We write $\mathbf{0}_k$ to denote the zero vector of length k. When the length of the zero vector is clear from the context, we may just write $\mathbf{0}$.

The last important vector operation is the *dot product*. The dot product of $\mathbf{u}$ and $\mathbf{v}$, written $\mathbf{u} \cdot \mathbf{v}$, is the sum of the products of the entries:

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_k v_k = \sum_{m=1}^{k} u_m v_m.$$

An important concept for regression analysis is the linear independence of a collection of vectors. Let $\mathbf{v}_1, \ldots, \mathbf{v}_J$ be a collection of J vectors, each of length k. We call $\mathbf{u}$ a *linear combination* of $\mathbf{v}_1, \ldots, \mathbf{v}_J$ if there exist real numbers $c_1, \ldots, c_J$ such that

$$\mathbf{u} = c_1 \mathbf{v}_1 + \cdots + c_J \mathbf{v}_J = \sum_{j=1}^{J} c_j \mathbf{v}_j.$$

A collection of vectors is *linearly independent* if the only solution to

$$c_1 \mathbf{v}_1 + \cdots + c_J \mathbf{v}_J = \mathbf{0}$$

is $c_1 = 0, \ldots, c_J = 0$. Otherwise, we call the vectors *linearly dependent*. Some fun facts about linear independence:

[1] R will let you add and subtract vectors of different lengths, via a technique called "recycling". For example `c(1, 0) + c(1, 2, 3, 4)` will produce `c(2, 2, 4, 4)`. This is kosher in R, but not in mathematical derivations.

- If any vector in $\mathbf{v}_1, \ldots, \mathbf{v}_J$ is a linear combination of the others, then these vectors are linearly dependent.
- A collection of J vectors of length k cannot be linearly independent if $J > k$. In other words, given vectors of length k, the most that can be linearly independent of each other is k.
- If any $\mathbf{v}_j = \mathbf{0}$, then $\mathbf{v}_1, \ldots, \mathbf{v}_J$ are linearly dependent. (Why?)

Examples:

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix};$$

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 14 \\ 12 \\ 0 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix};$$

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 1 \\ 4 \\ 9 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 1 \\ 8 \\ 27 \end{pmatrix}.$$

6.2 Matrix Operations

A matrix is a two-dimensional array of numbers, with entries in rows and columns. We call a matrix with n rows and m columns an $n \times m$ matrix. For example, the following is a 2×3 matrix:

$$\mathbf{A} = \begin{bmatrix} 99 & 73 & 2 \\ 13 & 40 & 41 \end{bmatrix}$$

Notice the convention of using an uppercase bold letter to denote a matrix. Given a matrix $\mathbf{A}$, we usually write a_{ij} to denote the entry in the i'th row and j'th column. In the above example, we have $a_{13} = 2$.

You can think of a vector $\mathbf{v} \in \mathbb{R}^k$ as a $1 \times k$ *row matrix* or as a $k \times 1$ *column matrix*. Throughout this book, I will treat vectors as column matrices unless otherwise noted.

Like vectors, matrices can be multipled by a scalar $c \in \mathbb{R}$.

$$c\mathbf{A} = \begin{bmatrix} ca_{11} & ca_{12} & \cdots & ca_{1m} \\ ca_{21} & ca_{22} & \cdots & ca_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ ca_{n1} & ca_{n2} & \cdots & ca_{nm} \end{bmatrix}$$

Matrices of the same dimension (i.e., both with the same number of rows n and columns m)

can be added …

$$\mathbf{A}+\mathbf{B} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1m}+b_{1m} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2m}+b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}+b_{n1} & a_{n2}+b_{n2} & \cdots & a_{nm}+b_{nm} \end{bmatrix}$$

… and subtracted …

$$\mathbf{A}-\mathbf{B} = \begin{bmatrix} a_{11}-b_{11} & a_{12}-b_{12} & \cdots & a_{1m}-b_{1m} \\ a_{21}-b_{21} & a_{22}-b_{22} & \cdots & a_{2m}-b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}-b_{n1} & a_{n2}-b_{n2} & \cdots & a_{nm}-b_{nm} \end{bmatrix}$$

Sometimes you will want to "rotate" an $n \times m$ matrix into an $m \times n$ one, so that the first row becomes the first column, the second row becomes the second column, and so on. This is called the *transpose*. I write the transpose of $\mathbf{A}$ as $\mathbf{A}^\top$, though you will often also see it written $\mathbf{A}'$. For example:

$$\mathbf{A} = \begin{bmatrix} 99 & 73 & 2 \\ 13 & 40 & 41 \end{bmatrix} \quad \Leftrightarrow \quad \mathbf{A}^\top = \begin{bmatrix} 99 & 13 \\ 73 & 40 \\ 2 & 41 \end{bmatrix}$$

Some of the most commonly invoked properties of the transpose are:

$$\begin{aligned} (\mathbf{A}^\top)^\top &= \mathbf{A}, \\ (c\mathbf{A})^\top &= c\mathbf{A}^\top, \\ (\mathbf{A}+\mathbf{B})^\top &= \mathbf{A}^\top + \mathbf{B}^\top, \\ (\mathbf{A}-\mathbf{B})^\top &= \mathbf{A}^\top - \mathbf{B}^\top. \end{aligned}$$

A matrix is *square* if it has the same number of rows as columns, i.e., it is $n \times n$. Every matrix is special, but some kinds of square matrix are *especially* special.

- A *symmetric* matrix is equal to its transpose: $\mathbf{A} = \mathbf{A}^\top$. Example:

$$\begin{bmatrix} 1 & 10 & 100 \\ 10 & 2 & 0.1 \\ 100 & 0.1 & 3 \end{bmatrix}$$

- A *diagonal* matrix contains zeroes everywhere except along the main diagonal: if $i \neq j$, then $a_{ij} = 0$. A diagonal matrix is symmetric by definition. Example:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

- The $n \times n$ *identity* matrix, written $\mathbf{I}_n$ (or just $\mathbf{I}$ when the size is clear from context), is the $n \times n$ diagonal matrix where each diagonal entry is 1. Example:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

And last we come to matrix multiplication. Whereas matrix addition and subtraction are pretty intuitive, matrix multiplication is not. Let $\mathbf{A}$ be an $n \times m$ matrix and $\mathbf{B}$ be an $m \times p$ matrix. (Notice that the number of columns of $\mathbf{A}$ must match the number of rows of $\mathbf{B}$.) Then $\mathbf{C} = \mathbf{AB}$ is an $n \times p$ matrix whose ij'th element is the dot product of the i'th row of $\mathbf{A}$ and the j'th column of $\mathbf{B}$:

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{im}b_{mj}.$$

Some examples might make this more clear.

$$\mathbf{A} = \begin{bmatrix} 2 & 10 \\ 0 & 1 \\ -1 & 5 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 4 \\ -1 & 10 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} 2 \cdot 1 + 10 \cdot (-1) & 2 \cdot 4 + 10 \cdot 10 \\ 0 \cdot 1 + 1 \cdot (-1) & 0 \cdot 4 + 1 \cdot 10 \\ (-1) \cdot 1 + 5 \cdot (-1) & (-1) \cdot 4 + 5 \cdot 10 \end{bmatrix} = \begin{bmatrix} -8 & 108 \\ -1 & 10 \\ -6 & 46 \end{bmatrix}$$

And here's one that you'll start seeing a lot of soon.

$$\mathbf{A} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ & \vdots & \\ 1 & x_{N1} & x_{N2} \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} \\ \vdots \\ \beta_0 + \beta_1 x_{N1} + \beta_2 x_{N2} \end{bmatrix}$$

Some important properties of matrix multiplication:

- Matrix multiplication is associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.
- Matrix multiplication is distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$.
- For any $n \times m$ matrix $\mathbf{A}$, we have $\mathbf{AI}_m = \mathbf{I}_n\mathbf{A} = \mathbf{A}$. In this way, the identity matrix is kind of like the matrix equivalent of the number one. (More on this when we get to matrix inversion.)
- Matrix multiplication is *not* commutative. In other words, $\mathbf{AB} \neq \mathbf{BA}$ except in very special cases (e.g., one of them is the identity matrix).

This is obvious when we're dealing with non-square matrices. Let $\mathbf{A}$ be $n \times m$ and $\mathbf{B}$ be $m \times p$, so that $\mathbf{AB}$ exists. Then $\mathbf{BA}$ doesn't even exist unless $n = p$. Even then, if $n \neq m$, then $\mathbf{AB}$ is $n \times n$ and $\mathbf{BA}$ is $m \times m$, so they can't possibly be the same.

For an example that $\mathbf{AB} \neq \mathbf{BA}$ even for square matrices:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$
$$\mathbf{AB} = \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}, \mathbf{BA} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

- The transpose of the product is the product of the transposes … but the other way around: $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$.

  This is intuitive, if you think about it. Suppose $\mathbf{A}$ is $n \times m$ and $\mathbf{B}$ is $m \times p$. Then $\mathbf{AB}$ is $n \times p$, so $(\mathbf{AB})^\top$ should be $p \times n$. Therefore, $\mathbf{B}^\top$ must come first.

6.3 Matrix Inversion

We've covered matrix addition, subtraction, and multiplication. What about division?

Let's think about division of real numbers for a second. We know that any division problem can be rewritten as a multiplication problem,

$$\frac{a}{b} = a \times b^{-1},$$

where b^{-1} is the unique real number such that

$$b \times b^{-1} = 1.$$

Similarly, in matrix algebra, we say that the $n \times n$ matrix $\mathbf{C}$ is an *inverse* of the $n \times n$ matrix $\mathbf{A}$ if $\mathbf{AC} = \mathbf{CA} = \mathbf{I}_n$.

Some basic properties of matrix inverses:

- If $\mathbf{C}$ is an inverse of $\mathbf{A}$, then $\mathbf{A}$ is an inverse of $\mathbf{C}$. This is immediate from the definition.
- If $\mathbf{C}$ and $\mathbf{D}$ are both inverses of $\mathbf{A}$, then $\mathbf{C} = \mathbf{D}$. Proof: If $\mathbf{C}$ and $\mathbf{D}$ are inverses of $\mathbf{A}$, then we have

$$\begin{aligned} \mathbf{AC} = \mathbf{I} &\Leftrightarrow \mathbf{D}(\mathbf{AC}) = \mathbf{DI} \\ &\Leftrightarrow (\mathbf{DA})\mathbf{C} = \mathbf{D} \\ &\Leftrightarrow \mathbf{IC} = \mathbf{D} \\ &\Leftrightarrow \mathbf{C} = \mathbf{D}. \end{aligned}$$

  As a consequence of this property, we write the inverse of $\mathbf{A}$, when it exists, as $\mathbf{A}^{-1}$.

- The inverse of the inverse of $\mathbf{A}$ is $\mathbf{A}$: $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
- If the inverse of $\mathbf{A}$ exists, then the inverse of its transpose is the transpose of the inverse: $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$.
- Matrix inversion inverts scalar multiplication: if $c \neq 0$, then $(c\mathbf{A})^{-1} = (1/c)\mathbf{A}^{-1}$.
- The identity matrix is its own inverse: $\mathbf{I}_n^{-1} = \mathbf{I}_n$.

Some matrices are not *invertible*; i.e., their inverse does not exist. As a simple example, think of

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

It's easy to see that, for any 2×2 matrix $\mathbf{B}$, we have

$$\mathbf{AB} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \neq \mathbf{I}_2.$$

Therefore, $\mathbf{A}$ does not have an inverse.

Remember that matrix inversion is kind of like division for scalar numbers. In that light, the previous example is a generalization of the principle that you can't divide by zero. But matrices full of zeroes are not the only ones that aren't invertible. For instance, it may not be obvious at first glance, but the following matrix is not invertible:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

We know that because of the following theorem: *A matrix is invertible if and only if its columns are linearly independent.* In the above example, the second column is 2 times the first column, so the columns are not linearly independent, so the matrix is not invertible.

Consider the general 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

We have a simple criterion for linear independence here. In particular, the columns of $\mathbf{A}$ are linearly independent if and only if $ad \neq bc$, or $ad - bc \neq 0$. We call this the *determinant* of the matrix, since it determines whether the matrix is invertible.[2] Moreover, if $ad - bc \neq 0$ we have

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

I'll leave it to you to convince yourself that's true. For now, let's try a couple of examples.

[2] On the determinants of 3×3 and larger matrices, see your friendly local linear algebra textbook. Calculating the determinant becomes exponentially more complicated with the size of the matrix.

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \mathbf{A}^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix},$$
$$\mathbf{A} = \begin{bmatrix} 4 & 6 \\ 2 & 4 \end{bmatrix}, \mathbf{A}^{-1} = \begin{bmatrix} 1 & -1.5 \\ -0.5 & 1 \end{bmatrix},$$
$$\mathbf{A} = \begin{bmatrix} 10 & 25 \\ 4 & 10 \end{bmatrix}, \mathbf{A}^{-1} \text{ does not exist.}$$

6.4 Solving Linear Systems

You may remember from high school being asked to solve for x_1 and x_2 in systems of equations like the following one:

$$2x_1 + x_2 = 10,$$
$$2x_1 - x_2 = -10.$$

Matrix algebra lets us write this whole system as a single equation, $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 2 & -1 \end{bmatrix},$$
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$
$$\mathbf{b} = \begin{bmatrix} 10 \\ -10 \end{bmatrix}.$$

This suggests a natural way to solve for $\mathbf{x}$:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

In fact, the linear system of equations $\mathbf{Ax} = \mathbf{b}$ has a unique solution if and only if $\mathbf{A}$ is invertible. Otherwise, it has either zero solutions or infinitely many solutions.

Example with zero solutions:

$$x_1 + x_2 = 1,$$
$$2x_1 + 2x_2 = 10.$$

Example with infinitely many solutions:

$$x_1 + x_2 = 1,$$
$$2x_1 + 2x_2 = 2.$$

6.5 Appendix: Matrices in R

We use the `matrix()` command to create matrices.

```
A <- matrix(c(2, 1, 3, 4),
            nrow = 2,
            ncol = 2)
A
```

```
##      [,1] [,2]
## [1,]    2    3
## [2,]    1    4
```

Notice that it fills "down" by column. To fill "across" instead, use the `byrow` argument:

```
B <- matrix(c(2, 1, 3, 4),
            nrow = 2,
            ncol = 2,
            byrow = 2)
B
```

```
##      [,1] [,2]
## [1,]    2    1
## [2,]    3    4
```

There are a few utilities for checking the dimension of a matrix.

```
nrow(A)
```

```
## [1] 2
```

```
ncol(A)
```

```
## [1] 2
```

```
dim(A)
```

```
## [1] 2 2
```

To extract the i'th row, j'th column, or ij'th element, use square brackets.

```
A[1, ]   # 1st row
```

```
## [1] 2 3
```

```
A[, 2]   # 2nd column
```

```
## [1] 3 4
```

```
A[2, 1]  # entry in 2nd row, 1st column
```

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

Notice that when you extract a row or column, R turns it into a vector—the result has only a single dimension. If you dislike this behavior (i.e., you want an extracted column to be a 1-column matrix), use the `drop = FALSE` option in the square brackets.

```
A[, 2, drop = FALSE]
```

```
##      [,1]
## [1,]    3
## [2,]    4
```

Adding and subtracting matrices works as you'd expect.

```
A + B
```

```
##      [,1] [,2]
## [1,]    4    4
## [2,]    4    8
```

```
A - B
```

```
##      [,1] [,2]
## [1,]    0    2
## [2,]   -2    0
```

As does scalar multiplication.

```
5 * A
```

```
##      [,1] [,2]
## [1,]   10   15
## [2,]    5   20
```

```
-1 * B
```

```
##      [,1] [,2]
## [1,]   -2   -1
## [2,]   -3   -4
```

However, the `*` operator performs *element-by-element* multiplication, not matrix multiplication.

```
A * B
```

```
##      [,1] [,2]
## [1,]    4    3
## [2,]    3   16
```

To perform matrix multiplication, use the `%*%` operator.

```
A %*% B
```

```
##      [,1] [,2]
## [1,]   13   14
## [2,]   14   17
```

To invert a matrix or solve a linear system, use the `solve()` function.

```
# Invert A
solve(A)
```

```
##      [,1] [,2]
## [1,]  0.8 -0.6
## [2,] -0.2  0.4
```

```
# Solve for x in Ax = (3, 2)
solve(A, c(3, 2))
```

```
## [1] 1.2 0.2
```

Here is a not-so-fun fact about matrix inversion in R: it's not entirely exact. To see this, let's invert a matrix with some decimal elements.

```
X <- matrix(c(1.123, 2.345, 3.456, 4.567), 2, 2)
Y <- solve(X)
Y
```

```
##          [,1]     [,2]
## [1,] -1.53483  1.16145
## [2,]  0.78808 -0.37741
```

Now let's see what we get when we multiply `X` and `Y`.

```
X %*% Y
```

```
##          [,1]       [,2]
## [1,] 1.00e+00 1.4798e-16
## [2,] 8.21e-17 1.0000e+00
```

That's not an identity matrix! The issue here is *floating point error*, the fact that decimal numbers are not stored exactly on computers. Notice that the off-diagonal elements here, which are supposed to be exactly zero, are instead very very tiny numbers, on the order of 10^{-16}, or 0.0000000000000001.

Let's check that our result is *numerically* equal to what we expected. By numerically equal, I mean, loosely speaking, that any differences are less than the amount of error you would expect due to floating point error. First we'll use `diag()` to generate a 2×2 identity matrix, then we'll compare numerical equality using `all.equal()`.

```
I <- diag(2)
all.equal(X %*% Y, I)
```

```
## [1] TRUE
```

Whereas the traditional `==` operator is stricter, checking for exact equality.

```
X %*% Y == I
```

```
##       [,1]  [,2]
```

```
## [1,]  TRUE FALSE
## [2,] FALSE FALSE
```

Moral of the story: when comparing decimal numbers, use `all.equal()` rather than `==`. When `all.equal()` is not `TRUE`, it returns a message indicating how far apart the numbers are. This is annoying if you want to use `all.equal()` in, say, an `if/else` statement. To get around that, we have the `isTRUE()` function.

```
all.equal(1.0, 1.5)
```

```
## [1] "Mean relative difference: 0.5"
```

```
isTRUE(all.equal(1.0, 1.5))
```

```
## [1] FALSE
```

One last thing. If `solve()` throws an error that says "reciprocal condition number…" or "system is exactly singular", that means you tried to invert a matrix that is not invertible.

```
Z <- matrix(c(1, 1, 2, 2), 2, 2)
solve(Z)
```

```
## Error in solve.default(Z): Lapack routine dgesv: system is exactly singular: U[2,2] = 0
```

Sad!

Chapter 7

Reintroduction to the Linear Model

Having learned some matrix algebra, let us now return to the world of statistics. We are going to take what we learned about regression and ordinary least squares in the bivariate case, then generalize it to a setting with potentially many variables. To make that task feasible, we will rely on the tools of matrix algebra that we learned last week.

7.1 The Linear Model in Matrix Form

We have a sequence of observations indexed by $n \in \{1, \ldots, N\}$. Each observation consists of a response, Y_n, a real number; and a vector of K covariates,

$$\mathbf{x}_n = \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nK} \end{pmatrix}.$$

Just like in bivariate regression, our goal is to estimate the conditional expectation of the response given the covariates, $E[Y_n \,|\, \mathbf{x}_n]$. To make that task feasible, we will assume the relationship is linear,

$$E[Y_n \,|\, \mathbf{x}_n] = \beta \cdot \mathbf{x}_n,$$

where β is the $K \times 1$ vector of coefficients,

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}.$$

Our data model is

$$Y_n = \beta \cdot \mathbf{x}_n + \epsilon_n,$$

where ϵ_n is "white noise" error that is uncorrelated with the covariates. (More on this in a second.)

This data model looks a little bit different than our bivariate linear model, which you'll recall was

$$Y_n = \alpha + \beta x_n + \epsilon_n.$$

What happened to α, the intercept? When working with the multivariate linear model, it will make our lives easiest to treat the intercept like any other coefficient. Specifically, we will assume $x_{n1} = 1$ for all n, and we will treat β_1 as the intercept. With $K = 2$, our multivariate model becomes

$$\begin{aligned} Y_n &= \beta \cdot \mathbf{x}_n + \epsilon_n \\ &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ x_{n2} \end{pmatrix} + \epsilon_n \\ &= \beta_1 + \beta_2 x_{n2} + \epsilon_n, \end{aligned}$$

which is the same as our bivariate regression model, replacing the intercept α with β_1, the slope β with β_2, and the covariate x_n with x_{n2}.

If we were to stack up all of our data, we would have N equations,

$$\begin{aligned} Y_1 &= \beta \cdot \mathbf{x}_1 + \epsilon_1, \\ Y_2 &= \beta \cdot \mathbf{x}_2 + \epsilon_2, \\ &\vdots \\ Y_N &= \beta \cdot \mathbf{x}_N + \epsilon_N. \end{aligned}$$

Like any system of linear equations, we can write this one more easily in matrix form. Let $\mathbf{Y}$ be the $N \times 1$ vector that collects the response,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}.$$

Let $\mathbf{X}$ be the $N \times K$ matrix that collects the covariates,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1K} \\ 1 & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N2} & \cdots & x_{NK} \end{bmatrix}.$$

The n'th row of $\mathbf{X}$, which we will write $\mathbf{x}_n$ (lowercase), contains the covariates for the n'th observation. The k'th column of $\mathbf{X}$, which we will write $\mathbf{X}_k$ (uppercase), contains the value of the k'th covariate for every observation. Finally, we will collect the error terms in an $N \times 1$ vector,

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}.$$

We can now write a model of the full data,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

It is worth pausing to clarify what is known and unknown here.

- The covariate matrix $\mathbf{X}$ and the response vector $\mathbf{Y}$ are known. They are our data.
- The regression parameters β are unknown. They are what we are trying to learn from the data.
- The error term ϵ is also unknown. We can think of each observation of Y_n as being a combination of "signal", $\mathbf{x}_n \cdot \beta$, and "noise", ϵ_n. The fundamental problem is that we don't know exactly what the signal is and what the noise is.

7.2 The OLS Estimator

Consider the linear model with three covariates,

$$Y_n = \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \epsilon_n.$$

Let's do like we did with bivariate regression, and imagine estimating the parameters of the model by least squares. Let (b_1, b_2, b_3) denote an estimate of the parameters.[1] We will set up the sum of squared errors as a function of the parameters,

$$\text{SSE}(b_1, b_2, b_3) = \sum_n (Y_n - b_1 x_{n1} - b_2 x_{n2} - b_3 x_{n3})^2.$$

Just as we did to derive the bivariate OLS estimator, let's begin by taking the partial derivative of the SSE with respect to the first regression coefficient, then equalizing it to zero.

$$\frac{\partial \,\text{SSE}}{\partial b_1} = -2 \sum_n x_{n1}(Y_n - b_1 x_{n1} - b_2 x_{n2} - b_3 x_{n3}) = 0.$$

Dividing each side by -2 and rearranging terms gives us

$$\sum_n x_{n1}(b_1 x_{n1} + b_2 x_{n2} + b_3 x_{n3}) = \sum_n x_{n1} Y_n.$$

If we break up the left-hand sum into three individual sums, we get

$$\left(\sum_n x_{n1}^2\right) b_1 + \left(\sum_n x_{n1} x_{n2}\right) b_2 + \left(\sum_n x_{n1} x_{n3}\right) b_3 = \sum_n x_{n1} Y_n,$$

[1] I'm using b_k instead of $\hat{\beta}_k$ simply because it's exhausting to type all those `\hat{}`s.

which is a linear condition on (b_1, b_2, b_3). If we go through the same steps with $\partial \,\mathrm{SSE}\,/\partial b_2$ and $\partial \,\mathrm{SSE}\,/\partial b_3$, we obtain the linear system

$$\begin{aligned}
\left(\sum_n x_{n1}^2\right) b_1 + \left(\sum_n x_{n1}x_{n2}\right) b_2 + \left(\sum_n x_{n1}x_{n3}\right) b_3 &= \sum_n x_{n1}Y_n, \\
\left(\sum_n x_{n2}x_{n1}\right) b_1 + \left(\sum_n x_{n2}^2\right) b_2 + \left(\sum_n x_{n2}x_{n3}\right) b_3 &= \sum_n x_{n2}Y_n, \\
\left(\sum_n x_{n3}x_{n1}\right) b_1 + \left(\sum_n x_{n3}x_{n2}\right) b_2 + \left(\sum_n x_{n3}^2\right) b_3 &= \sum_n x_{n3}Y_n.
\end{aligned}$$

This is a linear system of three equations in three unknowns, namely (b_1, b_2, b_3). We can write it as $\mathbf{Ab} = \mathbf{c}$, where $\mathbf{b}$ is the 3×1 column vector we are trying to solve for. You'll remember from last week that we use matrix algebra to solve linear systems like this one.

Let's take a closer look at the coefficient matrix we have here,

$$\mathbf{A} = \begin{bmatrix} \sum_n x_{n1}^2 & \sum_n x_{n1}x_{n2} & \sum_n x_{n1}x_{n3} \\ \sum_n x_{n2}x_{n1} & \sum_n x_{n2}^2 & \sum_n x_{n2}x_{n3} \\ \sum_n x_{n3}x_{n1} & \sum_n x_{n3}x_{n2} & \sum_n x_{n3}^2 \end{bmatrix}$$

Notice that each ij'th element is

$$a_{ij} = \sum_n x_{ni}x_{nj} = \mathbf{X}_i \cdot \mathbf{X}_j,$$

the dot product of the i'th and j'th columns of our $\mathbf{X}$ matrix. Of course, the i'th column of $\mathbf{X}$ is the i'th row of $\mathbf{X}^\top$. If the ij'th entry of $\mathbf{A}$ is the dot product of the i'th row of $\mathbf{X}^\top$ and the j'th column of $\mathbf{X}$, that means

$$\mathbf{A} = \mathbf{X}^\top\mathbf{X}.$$

Similarly, let's take a look at our right-hand side,

$$\mathbf{c} = \begin{bmatrix} \sum_n x_{n1}Y_n \\ \sum_n x_{n2}Y_n \\ \sum_n x_{n3}Y_n \end{bmatrix}.$$

Each i'th entry of $\mathbf{c}$ is

$$c_i = \sum_n x_{ni}Y_n = \mathbf{X}_i \cdot \mathbf{Y}.$$

the dot product of the i'th column of $\mathbf{X}$ (i.e., the i'th column of $\mathbf{X}^\top$) and the vector $\mathbf{Y}$. Therefore, we have

$$\mathbf{c} = \mathbf{X}^\top\mathbf{Y}.$$

Our linear system of equations, $\mathbf{Ab} = \mathbf{c}$, is equivalent to

$$(\mathbf{X}^\top\mathbf{X})\mathbf{b} = \mathbf{X}^\top\mathbf{Y}.$$

Consequently, the solution to the system is

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Although we got here via the 3×3 case, this formula works for any number of covariates. The *OLS estimator* of the linear model coefficients from covariate matrix $\mathbf{X}$ and response vector $\mathbf{Y}$ is

$$\hat{\beta}_{\text{OLS}}(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

When you see this formula, your hackles should be raised. *Wait a minute*, you ought to be saying. *How do we know the inverse of* $\mathbf{X}^\top \mathbf{X}$ *exists?* That's an excellent question! Luckily, there's a simple condition: $\mathbf{X}^\top \mathbf{X}$ is invertible if and only if the columns of $\mathbf{X}$ are linearly independent.

The linear independence condition isn't just a technical thing that we need to satisfy. It goes to the heart of what we're doing in linear regression. If the columns of $\mathbf{X}$ aren't linearly independent, then the question you're asking of OLS—to learn something about the coefficients from the data—is ill-defined.

Imagine you have a linear dependency between two variables, so one is just a scalar multiple of the other. For example, a regression of a person's weight on their height in inches and height in centimeters. Or a regression of whether it rains on temperature Fahrenheit and temperature Celsius. It is absurd to think that the relationship between temperature and rain might be different depending on how you measure it. But that's exactly what you're asking for when you run this regression—separate estimates for the effect of degrees Fahrenheit and the effect of degrees Celsius.

7.3 Vector-Valued Random Variables

Before we can talk about the properties of OLS in the multivariate case, we need to refresh ourselves on how basic statistical operations (expected value and variance) translate when we're dealing with vectors of random variables.

Let A and B be random variables with means $\mu_A = E[A]$ and $\mu_B = E[B]$ respectively. Let C be the column vector whose first value is A and whose second value is B:

$$C = \begin{pmatrix} A \\ B \end{pmatrix}.$$

As a function of random variables, C is itself a random variable. Unlike those we've encountered before, though, it is a *vector-valued* random variable.

Assume A and B take values in the finite sets $\mathcal{A}$ and $\mathcal{B}$ respectively. The expected value of

C is

$$\begin{aligned} E[C] &= \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \begin{pmatrix} a \\ b \end{pmatrix} \Pr(A = a, B = b) \\ &= \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}. \end{aligned}$$

I encourage you to prove this on your own—the proof just relies on simple facts about vector addition and joint probability that we've already covered in this class. It is easiest to prove in the finite case, but it remains true that $E[C] = (\mu_A, \mu_B)$ in the more general case.

You might expect the variance of C to be a vector too. You would be wrong—it's a 2×2 matrix.

$$\begin{aligned} V[C] &= E\left[(C - E[C])(C - E[C])^\top\right] \\ &= E\left[\begin{pmatrix} A - \mu_A \\ B - \mu_B \end{pmatrix} \begin{pmatrix} A - \mu_A & B - \mu_B \end{pmatrix}\right] \\ &= E\left[\begin{bmatrix} (A - \mu_A)^2 & (A - \mu_A)(B - \mu_B) \\ (A - \mu_A)(B - \mu_B) & (B - \mu_B)^2 \end{bmatrix}\right] \\ &= \begin{bmatrix} E[(A - \mu_A)^2] & E[(A - \mu_A)(B - \mu_B)] \\ E[(A - \mu_A)(B - \mu_B)] & E[(B - \mu_B)^2] \end{bmatrix} \\ &= \begin{bmatrix} V[A] & \text{Cov}[A, B] \\ \text{Cov}[A, B] & V[B] \end{bmatrix}. \end{aligned}$$

This is what we call the *variance* matrix, or *variance-covariance matrix*, of a vector-valued random variable. The i'th element along the main diagonal gives us the variance of the i'th element of the vector. The ij'th off-diagonal element gives us the covariance of the i and j'th elements. Consequently, since $\text{Cov}[A, B] = \text{Cov}[B, A]$, the variance matrix is always symmetric.

7.4 Properties of OLS

Just like in the bivariate case, the "good" properties of OLS depend on whether the process that generated our data satisfies particular assumptions. The key assumption, which we call *strict exogeneity*, is

$$E[\epsilon \mid \mathbf{X}] = \mathbf{0}.$$

In other words, the error term must be uncorrelated with the covariates. Remember that the error for the n'th observation, ϵ_n, collects everything that affects Y_n but is not included in $\mathbf{x}_n$. So what we're saying when we impose this condition is either that there's nothing else out there that affects $\mathbf{Y}$ besides $\mathbf{X}$ (unlikely!), or that anything else that affects $\mathbf{Y}$ is uncorrelated with $\mathbf{X}$ (also unlikely, but slightly less so!).

In the '90s and '00s, as more data became available and computing power increased, political scientists labored under the delusion that the way to make strict exogeneity hold was to throw every covariate you could imagine into each regression. This approach was statistically

illiterate (Clarke, 2005) and scholars have since begun to favor *design-based* approaches. The basic idea is to collect data with relatively little unobservable heterogeneity, whether through experiments or through careful observational work, rather than to try to eliminate it through endless controls. We'll talk more about design when we get to causal inference, and it will be a major source of discussion in Stat III.

For now, let us proceed imagining that strict exogeneity holds. Then, just as in the bivariate case, OLS is unbiased. In fact, it's even easier to prove now. First, notice that under strict exogeneity, we have

$$\begin{aligned} E[\mathbf{Y} \mid \mathbf{X}] &= E[\mathbf{X}\beta + \epsilon \mid \mathbf{X}] \\ &= \mathbf{X}\beta + E[\epsilon \mid \mathbf{X}] \\ &= \mathbf{X}\beta. \end{aligned}$$

It follows that

$$\begin{aligned} E[\hat{\beta}_{\text{OLS}}(\mathbf{X}, \mathbf{Y}) \mid \mathbf{X}] &= E[(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y} \mid \mathbf{X}] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top E[\mathbf{Y} \mid \mathbf{X}] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top (\mathbf{X}\beta) \\ &= (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{X})\beta \\ &= \beta, \end{aligned}$$

which is the definition of unbiasedness.

Unbiasedness is a small-sample property. No matter the sample size, if strict exogeneity holds, the OLS estimator is unbiased. OLS also has some asymptotic (or large-sample) properties under strict exogeneity that we won't prove, but are worth mentioning:

- OLS is *consistent*. Informally, what this means is that as N grows larger, the distribution of the OLS estimator becomes tighter around the population parameter β. In other words, with a sufficiently large sample, it becomes highly unlikely that you will draw a sample $(\mathbf{X}, \mathbf{Y})$ such that $\hat{\beta}_{\text{OLS}}(\mathbf{X}, \mathbf{Y})$ is far from the true value.

  Of course, you can't know that the OLS estimate from any particular sample is close to the truth. But you're much more likely to get an estimate close to the truth if $N = 100{,}000$ than if $N = 10$.

- OLS is *asymptotically normal*. Informally, what this is means is that if N is large enough, the sampling distribution of $\hat{\beta}_{\text{OLS}}$ (i.e., its distribution across different possible samples) is roughly normal. This makes the computation of inferential statistics fairly simple in large samples. More on this in two weeks.

Unbiasedness and consistency are nice, but frankly they're kind of dime-a-dozen. Lots of estimators are unbiased and consistent. Why is OLS so ubiquitous? The reason is that it is *efficient*, at least under a particular condition on the error term. Unlike unbiasedness and consistency, efficiency is defined with reference to other estimators. Given some class or collection of estimators, one is efficient if it has the lowest standard errors—i.e., it is the least sensitive to sampling variation, and thereby the most likely to come close to the true parameter value.

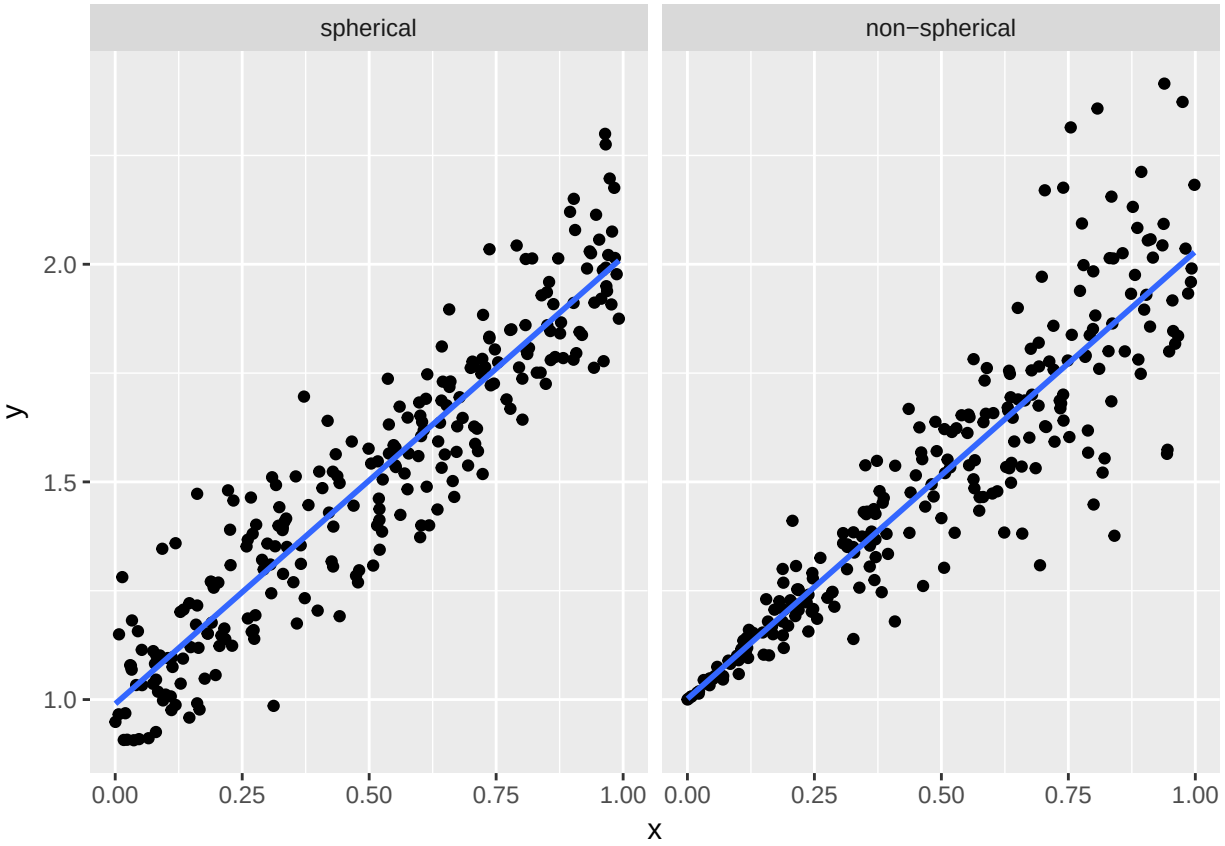
The condition we need to hold is that we have *spherical errors*:

$$V[\epsilon \,|\, \mathbf{X}] = \sigma^2 \mathbf{I}_N = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}.$$

Spherical errors summarizes two important conditions:

- The variance of each ϵ_n—i.e., the expected "spread" of the points around the regression line—is the same for every observation. This is also known as *homoskedasticity*.
- For $n \neq m$, there is no correlation between ϵ_n and ϵ_m. In other words, the fact that Y_n lies above the regression line doesn't tell us anything about whether Y_m lies above or below the regression line. This is also known as *no autocorrelation*.

Spherical errors holds if each ϵ_n is independent and identically distributed, though it is possible for non-i.i.d. errors to satisfy the condition. The illustration below compares spherical and non-spherical errors.



Notice that in the right-hand graph, the distribution of errors around the regression line is uneven—the spread is much greater at greater values of the covariate.

According to the Gauss-Markov theorem, if the errors are spherical, then OLS is the *best linear unbiased estimator (BLUE)* of the linear model parameters β. By "best," we mean

that it is efficient—any other linear unbiased estimator has larger standard errors. In other words, under the spherical error condition, any estimator $\hat{\beta}$ with a smaller standard errors than OLS must either be:

- Biased: $E[\hat{\beta}] \neq \beta$.
- Nonlinear: $\hat{\beta}$ cannot be written as a linear function of Y.

Much later in the course, we will encounter ridge regression, a linear estimator that has lower standard errors than OLS. The Gauss-Markov theorem tells us that we're making a tradeoff when we use ridge regression—that we're taking on some bias in exchange for the reduction in variance.

Chapter 8

Specification Issues

I lied to you about the linear model last week. Like the grade-school teachers who told you everyone thought the world was flat before Columbus proved them wrong, I had good intentions—but it was a lie nonetheless.

I claimed that the linear model assumed that the conditional expectation of the response was a linear function of the covariates. That is false. A data model is a linear model, can be estimated consistently and without bias by OLS, and all that good stuff, as long as it is linear in the *parameters*.

For example, the following is a linear model.

$$Y_n = \beta_1 + \beta_2 x_n + \beta_3 x_n^2 + \beta_4 x_n^7 + \epsilon_n.$$

The conditional expectation of Y_n is a nonlinear function of x_n (holding β fixed) but a linear function of β (holding x_n fixed). Therefore, assuming strict exogeneity holds, OLS is an unbiased, consistent, asymptotically normal estimator of β.

The following is not a linear model.

$$Y_n = 2^{\beta_1} + 2^{\beta_2} x_n + \epsilon_n.$$

Holding β fixed, this is a linear function of the covariate x_n. But, holding x_n fixed, this is not a linear function of β. OLS is not an appropriate estimator for the parameters of this model.

This week, we will talk about linear models with non-standard covariate specifications—those that aren't just a linear function of continuous variables.

8.1 Categorical Variables

Using the linear model, we write the conditional expectation for the n'th response as

$$E[Y_n \,|\, \mathbf{x}_n] = \mathbf{x}_n \cdot \beta + \epsilon_n,$$

where $\mathbf{x}_n$ is the vector of K covariates (including the intercept) and β is the vector of K coefficients we wish to estimate.

This makes sense with numerical variables, but not so much with categorical variables. For example, think of the relationship between party identification and one's vote in the 2016 presidential election. Suppose our response variable is one's vote (1 for Trump, 0 for non-Trump), and our party ID variable records whether the respondent is a Republican, Democrat, or independent. The resulting linear model equation,

$$\text{Trump}_n = \beta_1 + \beta_2 \text{Party ID}_n + \epsilon_n,$$

doesn't really make sense, because party ID isn't a number.[1]

To incorporate a categorical variable into the linear model, we break each category into its own binary variable. For example, with our party ID variable, we go from

$$\text{Party ID} = \begin{pmatrix} \text{R} \\ \text{R} \\ \text{I} \\ \text{I} \\ \text{D} \\ \text{D} \end{pmatrix}$$

to

$$\text{Republican} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \text{Independent} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \text{Democratic} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}.$$

These are called *dummy variables* or, preferably, *indicator variables*.

Having turned our categorical variable into a set of indicators, you may be tempted to rewrite the model as

$$\text{Trump}_n = \beta_1 + \beta_2 \text{Republican}_n + \beta_3 \text{Independent}_n + \beta_4 \text{Democratic}_n + \epsilon_n.$$

But take a look at the matrix of covariates, or *design matrix*, that would result if we set up the model this way:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

[1] Oldish-school political scientists, by which I mean just about anyone who got their PhD between 1990 and 2010, would now be clutching their pearls at the mere thought of using a linear model for vote choice. But so-called "binary dependent variable" models like logistic regression are at best overrated, as you read in *Mostly Harmless Econometrics*.

The columns of the design matrix are linearly dependent: the constant term is equal to the sum of three party ID indicators. (A useful exercise is to calculate $\mathbf{X}^\top\mathbf{X}$ and confirm that its columns are linearly dependent too.) This means we can't include all three when estimating β via OLS—we have to drop one category.

In one sense, which category we drop is immaterial—our regression will make the same predictions either way. However, in order to interpret the results of a regression on categorical variables, it is important that we know what the categories are, and which one has been dropped.

For example, imagine we drop the Republican category, so we have the following linear model:

$$\text{Trump}_n = \beta_1 + \beta_2\text{Independent}_n + \beta_3\text{Democratic}_n + \epsilon_n.$$

For a Republican voter, the Independent and Democratic variables will both equal zero, so we will have

$$E[\text{Trump}_n \,|\, \text{Party ID}_n = \text{R}] = \beta_1.$$

In other words, the intercept will be the predicted probability that a Republican votes for Trump. For an Independent voter, we will have

$$E[\text{Trump}_n \,|\, \text{Party ID}_n = \text{I}] = \beta_1 + \beta_2.$$

So the coefficient on the Independent indicator is not the predicted probability that an Independent votes for Trump. Instead, it is the *difference* in probability of a Trump vote between Independents and the baseline category (in this case, Republicans).

- If Independents are less likely than Republicans to vote for Trump, the coefficient on Independent will be negative.
- If Independents are more likely than Republicans to vote for Trump, the coefficient on Independent will be positive.
- If Independents are equally likely as Republicans to vote for Trump, the coefficient on Independent will be zero.

Similarly, for a Democratic voter, we have

$$E[\text{Trump}_n \,|\, \text{Party ID}_n = \text{D}] = \beta_1 + \beta_3.$$

The interpretation of the coefficient on Democratic is the same as for the coefficient on Independent.

Take a look at the following results of a hypothetical regression.

Coefficient	Estimate
(Intercept)	0.9
Independent	-0.4
Democratic	-0.75

- Republicans have a 90% chance of voting for Trump. We see that by looking at the intercept, since Republicans are the omitted category.
- We see from the coefficient on Independent that an Independent's chance of voting for Trump is 40% lower than a Republican's. This means that an Independent has a 50% chance of voting for Trump.
- Similarly, we see from the coefficient on Democratic that a Democrat's chance of voting for Trump is 75% lower than a Republican's, for a 15% chance overall.

Had we instead omitted Independent, we would get different coefficients, but the same predictions.

Coefficient	Estimate
(Intercept)	0.5
Republican	0.4
Democratic	-0.35

Same story, different numbers, had we omitted Democratic.

Coefficient	Estimate
(Intercept)	0.15
Republican	0.75
Independent	0.35

Given that the results are substantively the same no matter what, does it matter which category we choose to drop? Yes, for the purpose of communicating your results. The omitted category should serve as a meaningful baseline. For this example, all of our three categories are substantively meaningful, so any choice will do. But imagine replacing our part ID variable with a race variable that has the following categories:

- White
- Black
- Hispanic
- Asian
- Other

You may be tempted to make "Other" the excluded category, so that you obtain a coefficient for each of the specific racial groups. But that's actually the worst choice possible. The coefficient on the White variable would then represent the difference in probability of voting for Trump between a white voter and a voter in the "other" category—which is hard to interpret. Whereas if we instead omitted the Black category, the coefficient on the White variable would represent the difference between white and black voters.

When in doubt, I recommend omitting whichever category is largest in the data.

Now let's introduce covariates into the mix. Consider a regression of Trump vote on party ID (Republican as omitted category) and age, producing the following results.

Coefficient	Estimate
(Intercept)	0.8
Independent	-0.4
Democratic	-0.75
Age	0.002

Remember what the coefficient of 0.002 on Age means: if we compared one voter to another who was otherwise identical (in this case, same Party ID) except five years older, we would expect the latter voter to have a 1% greater chance of voting for Trump.

More specifically, we have three different regression lines—one for each group:

$$\begin{aligned} E[\text{Trump}_n \,|\, \text{Party ID}_n = \text{R}, \text{Age}_n] &= 0.8 + 0.002\text{Age}_n, \\ E[\text{Trump}_n \,|\, \text{Party ID}_n = \text{I}, \text{Age}_n] &= 0.4 + 0.002\text{Age}_n, \\ E[\text{Trump}_n \,|\, \text{Party ID}_n = \text{D}, \text{Age}_n] &= 0.05 + 0.002\text{Age}_n. \end{aligned}$$

Notice that the slope is the same in each regression line. Only the intercept varies across groups. When we include a categorical variable in a regression model, it's like allowing the intercept to differ across categories.

8.2 Interaction Terms

When political scientists or economists describe their regression results, they will often talk about the marginal effects of different variables. Formally, the *marginal effect* of the k'th covariate, x_{nk}, is

$$\frac{\partial E[Y_n \,|\, \mathbf{x}_n]}{\partial x_{nk}},$$

the partial derivative of the conditional expectation with respect to the k'th covariate.

The marginal effect answers the following question: Suppose we have two observations that differ in the k'th covariate by one unit, but are otherwise identical. How much greater, or less, would we expect the response to be for the observation with the one-unit-greater value of x_{nk}?

If we were sure the relationship we were modeling were causal, we could phrase the above question more succinctly. We could ask: Given a one-unit change in the k'th covariate, holding all else fixed, what change in the response should we expect? But we haven't yet gotten to the point where we can make our claims causal. Hence I will often refer to *so-called marginal effects*, since I don't want the "effect" terminology to deceive us into thinking we're drawing causal inferences. So-called marginal effects are just a nice way to summarize the relationship between individual covariates and the conditional expectation.

The bare-bones linear model has the (sometimes appealing, sometimes not) feature that it assumes constant marginal effects. For each covariate x_{nk}, we have

$$\frac{\partial E[Y_n \,|\, \mathbf{x}_n]}{\partial x_{nk}} = \beta_k,$$

the coefficient on that covariate. This encodes two critical assumptions:

1. The marginal effect of the k'th covariate does not depend on the value of any other covariates.
2. The marginal effect of the k'th covariate does not depend on its own value.

It is easy to think of scenarios where each of these might be questionable.

1. Imagine a study of individual voters' choices in U.S. House races, where we model voting for the incumbent as a function of how often the voter goes to church. The marginal effect of religiosity is probably different if the incumbent is a Republican than if the incumbent is a Democrat.
2. Imagine a study of individual voters' turnout decisions, where we model turnout as a function of the voter's ideology. Suppose ideology is measured on a 7-point scale, where 1 is most liberal and 7 is most conservative. We know the most ideologically extreme voters are the most likely to turn out. So, all else equal, we'd expect moving from 1 to 2 (very liberal to pretty liberal) to decrease one's probability of voting, but we'd expect moving from 6 to 7 (pretty conservative to very conservative) to increase one's probability of voting.

Let's start with the first case, where the (so-called) marginal effect of one variable depends on the value of another variable. To allow for this in our models, we include the product of the two covariates in our model.

For example, suppose we are interested in whether the relationship between education and voting for Trump is different between whites and non-whites. We would include three terms in the model (plus an intercept): education, an indicator for white, and their product.

$$\text{Trump}_n = \beta_1 + \beta_2 \text{Education}_n + \beta_3 \text{White}_n + \beta_4 (\text{Education}_n \times \text{White}_n) + \epsilon_n.$$

The so-called marginal effect of education is now

$$\frac{\partial E[\text{Trump}_n \,|\, \mathbf{x}_n]}{\partial \text{Education}_n} = \beta_2 + \beta_4 \text{White}_n.$$

This equation tells us three things.

- β_2 is the marginal effect of education for non-white voters.
- β_4 is the difference between the marginal effect of education for white voters and the effect for non-white voters.
- $\beta_2 + \beta_4$ is the marginal effect of education for white voters.

Another way to think of it is that we have two regression lines:

$$E[\text{Trump}_n \,|\, \text{White}_n = 0, \text{Education}_n] = \beta_1 + \beta_2 \text{Education}_n,$$
$$E[\text{Trump}_n \,|\, \text{White}_n = 1, \text{Education}_n] = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)\text{Education}_n.$$

We saw before that including a categorical variable is like allowing a different intercept for each category. Including an interaction with a categorical variable is like allowing a different slope for each category.

At this point, you might ask, why not just run two separate regressions? I can think of at least two reasons not to.

- You might want to include other covariates whose effects you don't think are dependent on race (e.g., age). If you ran separate regressions, you would estimate race-dependent effects for every covariate, at a potential loss of efficiency.
- You might want to formally test the hypothesis that the effect of education is equal for whites and non-whites. This is easiest to do if you have a single model. Next week we will talk about the tools you would need to undertake this sort of test.

One frequent source of confusion with interaction terms is whether you need to include lower-order terms in the model. For example, if we are only interested in how the effect of education differs with race, why can't we just include education and its product with race in the specification? The equations above give you the answer. Leaving the white indicator out of the model is like fixing $\beta_3 = 0$. This means you're forcing the regression lines for whites and non-whites to have the same intercept, which there's no good reason to do.

If you're not yet persuaded on the necessity of including constitutive terms of interactions in your regressions, see Braumoeller (2004).

For an example of interaction terms, imagine the following example. Suppose education is measured in years of schooling.

Coefficient	Estimate
(Intercept)	0.3
Education	0.01
White	0.4
Education * White	-0.03

We would interpret these in the following way.

- A hypothetical non-white voter with zero years of education has a 30% chance of voting for Trump. For each additional year of education, the probability of voting for Trump goes up by 1%.
- A hypothetical white voter with zero years of education has a 70% (0.3 + 0.4) chance of voting for Trump. For each additional year of education, the probability of voting for Trump goes down by 2% (0.01 - 0.03).

What about an interaction between two continuous variables? For example, imagine an interaction between age and education in our model of voting for Trump.

Coefficient	Estimate
(Intercept)	0.4
Education	-0.02
Age	0.002
Education * Age	0.0002

One simple way to interpret the effect of each variable is to hold the other one fixed at various values. For example, for a 20-year-old, we have

$$E[\text{Trump}_n \mid \text{Age}_n = 20, \text{Education}_n] = 0.44 - 0.16\text{Education}_n,$$

whereas for an 80-year-old, we have

$$E[\text{Trump}_n \mid \text{Age}_n = 80, \text{Education}_n] = 0.56 - 0.04\text{Education}_n.$$

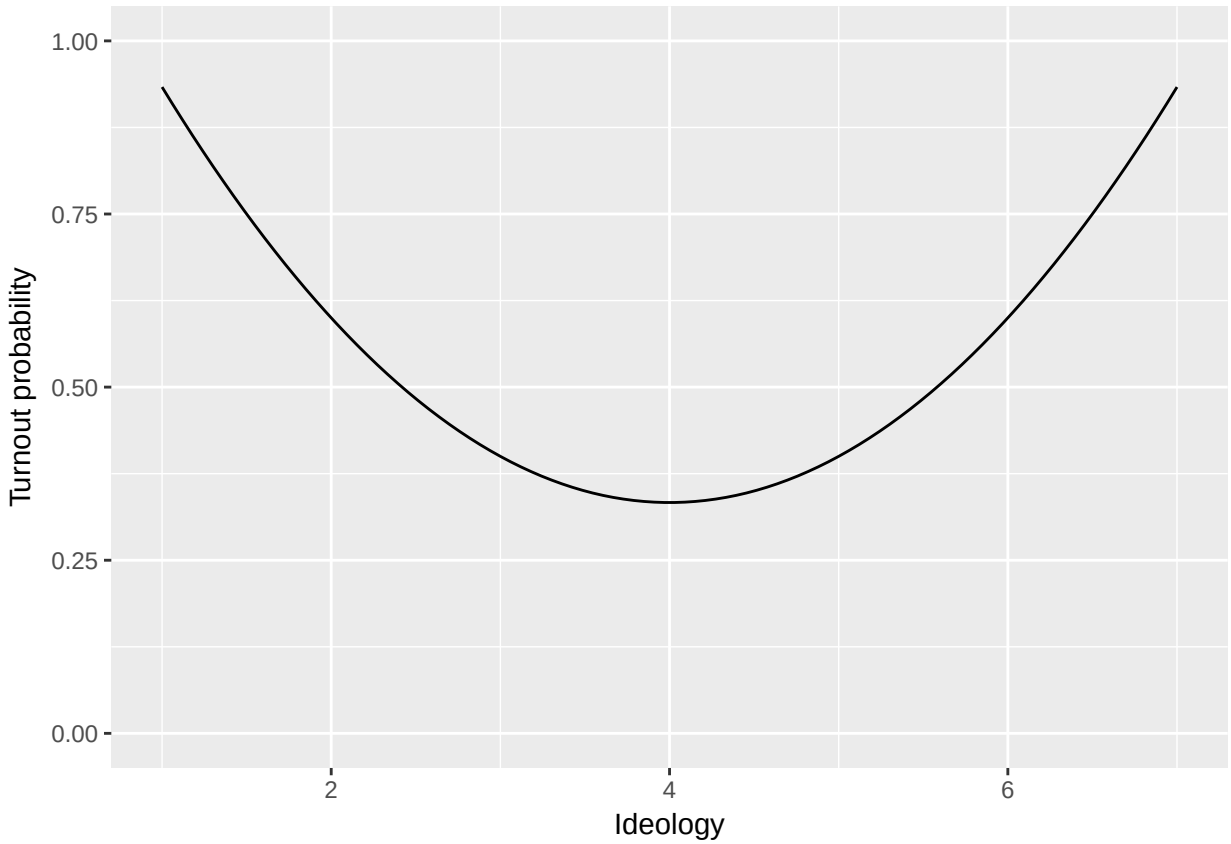
These results would seem to imply that (1) older people have a higher baseline probability of voting for Trump, and (2) the magnitude of the negative relationship between education and voting for Trump is weaker for older voters.

Always remember: when in doubt, take the partial derivative of Y_n (or, more precisely, its conditional expectation) with respect to the variable you're interested in. For example, here we have

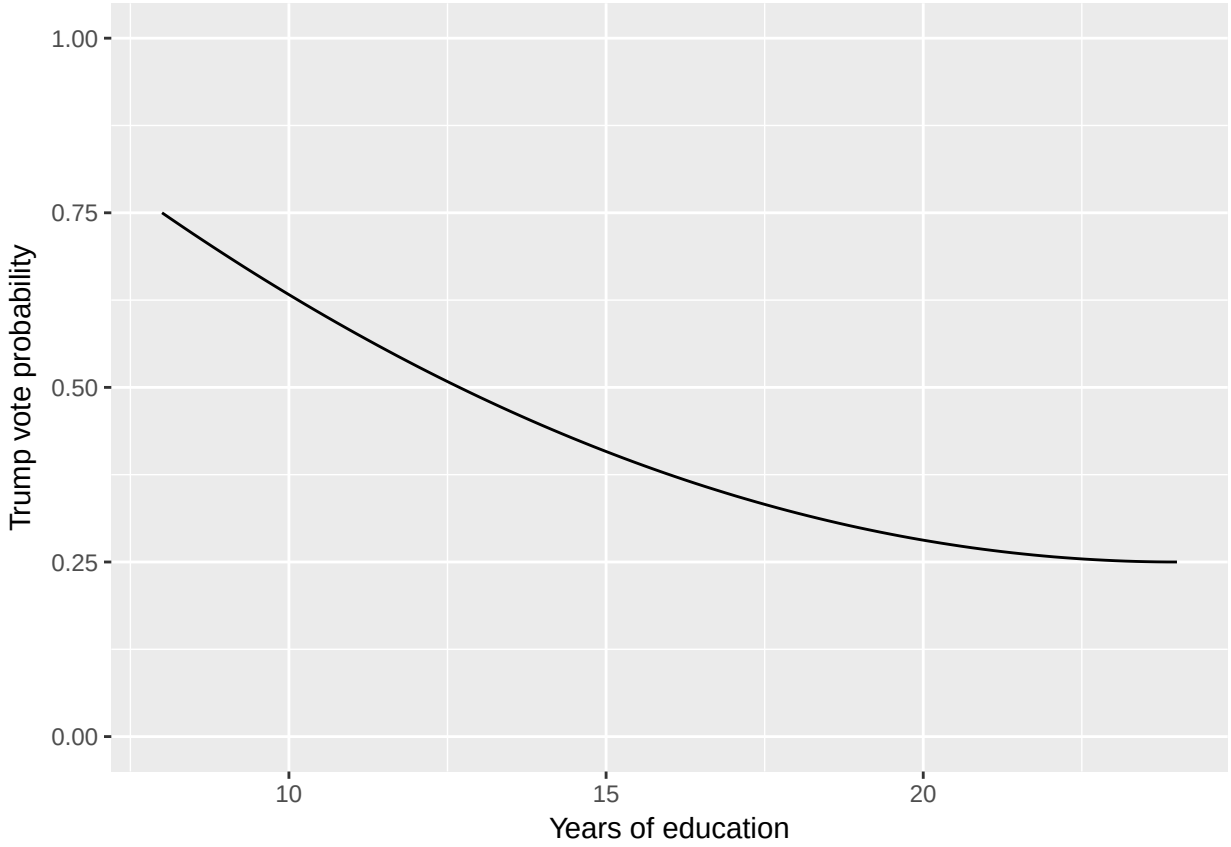
$$\frac{\partial E[\text{Trump}_n \mid \mathbf{x}_n]}{\partial \text{Education}_n} = -0.02 + 0.0002\text{Age}_n.$$

8.3 Quadratic and Logarithmic Terms

We use interaction terms when the marginal effect of one variable depends on the value of another variable. But sometimes the marginal effect of a variable depends on its *own* value. The most stark example is a "U-shaped" relationship, such as we expect between ideology and voter turnout.



We call this kind of relationship *non-monotonic*, since it is neither increasing everywhere nor decreasing everywhere. However, even with a monotonic relationship, the marginal effect might depend on the value of the variable. (In other words, while every linear function is monotonic, not every monotonic function is linear.) For example, think of a hockey-stick shaped relationship.



If we model the relationship between years of education and voting for Trump as linear, then we impose the assumption that the difference between voters with 16 years of education and 12 years of education (college versus high-school graduates) is the same as between those with 24 and 20 years of education (got the PhD slowly versus got the PhD quickly). Depending on our sample and the goal of our study, that may not be a reasonable assumption (or approximation).

We typically use quadratic models for non-monotonic relationships. In the example above, this would entail a regression model like

$$\text{Turnout}_n = \beta_1 + \beta_2 \text{Ideology}_n + \beta_3 \text{Ideology}_n^2 + \epsilon_n.$$

Under this model, the marginal effect of Ideology is

$$\frac{\partial E[\text{Turnout}_n \mid \text{Ideology}_n]}{\partial \text{Ideology}_n} = \beta_2 + 2\beta_3 \text{Ideology}_n.$$

- If β_3 is positive, that means the effect of Ideology increases with the value of Ideology, representing a U-shaped relationship.
- If β_3 is negative, that means the effect of Ideology decreases with the value of Ideology, representing an inverse-U-shaped relationship.
- If $beta_3$ is zero, that means the effect of Ideology is constant.

The other main way to model a nonlinear relationship in a single variable is with a logarithmic model. Remember that, in the standard bivariate linear model,

$$Y_n = \beta_1 + \beta_2 x_n + \epsilon_n,$$

we say that a 1-unit difference in x_n is associated with a β_2-unit difference in Y_n. If we were to instead model the natural logarithm of the response,

$$\log Y_n = \beta_1 + \beta_2 x_n + \epsilon_n,$$

then we would say that a 1-unit difference in x_n is associated with a β_2-*percent* difference in Y_n. So, for example, if x_n and Y_n are exactly proportional to each other (tripling x_n leads to a tripling of Y_n, for example), we would have $\beta_2 = 1$. Conversely, if we were to model the response as a function of the natural logarithm of the covariate,

$$Y_n = \beta_1 + \beta_2 \log x_n + \epsilon_n,$$

then we would say a 1-percent difference in x_n is associated with a β_2-unit difference in Y_n. Finally, in a full log-log model,

$$\log Y_n = \beta_1 + \beta_2 \log x_n + \epsilon_n,$$

we would say that a 1-percent difference in x_n is associated with a β_2-percent difference in Y_n.

How do you decide which logarithmic model, if any, to use?

- You may let theory be your guide—develop expectations, on the basis of your substantive knowledge, about whether the relevant changes in conditional expectation will be in terms of levels or proportions.
- Or you may be inductive—make scatterplots of all four possibilities, and choose the specification under which the relationship is closest to linear.

A final note on logarithmic models. The logarithm of a number $c \leq 0$ does not exist. Therefore, a logarithmic model is only appropriate for a strictly positive response/covariate. For non-negative variables that include zeroes, some people try to "fix" this by doing $\log(Y_n + 1)$, but in this situation it is generally better to follow the procedure advocated by Burbidge et al. (1988).

8.4 Appendix: Nonstandard Specifications in R

We will use the following packages:

```
library("tidyverse")
library("broom")
library("forcats")
library("interplot")
```

forcats contains convenience functions for *factors*, which are R's way of representing categorical variables. **interplot** is for plotting marginal effects from interactive models.

Once again, we will work with the occupational prestige data from the **car** package. We will also convert it to a tibble, so that it will more closely resemble the kind of data frame we would get had we read it in with `read_csv()`.

```
library("car")
data(Prestige)
Prestige <- as_tibble(Prestige)

Prestige
```

```
## # A tibble: 102 × 6
##    education income women prestige census   type
## *      <dbl>  <int> <dbl>    <dbl>  <int> <fctr>
## 1      13.11  12351 11.16     68.8   1113   prof
## 2      12.26  25879  4.02     69.1   1130   prof
## 3      12.77   9271 15.70     63.4   1171   prof
## 4      11.42   8865  9.11     56.8   1175   prof
## 5      14.62   8403 11.68     73.5   2111   prof
## # ... with 97 more rows
```

8.4.1 Categorical Variables

You will notice that the `type` column of the prestige data is listed as a `<fctr>`, which stands for factor. A factor is R's representation of a categorical variable. Let's take a closer look.

```
Prestige$type
```

```
##   [1] prof prof prof prof prof prof prof prof prof prof prof prof prof prof prof
##  [15] prof prof prof prof prof prof prof prof prof prof prof prof prof prof bc
##  [29] prof prof wc   prof wc   <NA> wc   wc   wc   wc   wc   wc   wc   wc
##  [43] wc   wc   wc   wc   wc   wc   wc   wc   wc   wc   <NA> bc   wc   wc
##  [57] wc   bc   bc   bc   bc   bc   <NA> bc   bc   bc   <NA> bc   bc   bc
##  [71] bc   bc   bc   bc   bc   bc   bc   bc   bc   bc   bc   bc   bc   bc
##  [85] bc   bc   bc   bc   bc   bc   bc   bc   bc   bc   bc   prof bc   bc
##  [99] bc   bc   bc   bc
## Levels: bc prof wc
```

This looks kind of like—but isn't quite—a character variable. The most noticeable difference is that R tells us its *levels*: the set of categories available. We can extract these directly with the `levels()` function.

```
levels(Prestige$type)
```

```
## [1] "bc"   "prof" "wc"
```

This brings us to an important difference between `read.csv()` (with a dot, the built-in R function) and `read_csv()` (with an underscore, the tidyverse version). The old way, `read.csv()`, by default treats any column of character strings as a factor. The tidyverse way, `read_csv()`, never creates factors by default—you must make them explicitly.

To create a factor variable, you can use the `factor()` function:

```
example_vector <- c(4, 1, 3, 3, 1)

factor(example_vector,  # vector to convert to factor
       levels = 1:4,    # possible values of the vector
       labels = c("ONE",
                  "two",
                  "Three",
                  "FOUR!"))  # label corresponding to each value
```

```
## [1] FOUR! ONE   Three Three ONE
## Levels: ONE two Three FOUR!
```

Returning to the prestige data, let's run a regression of occupational prestige on occupational category.

```
fit_type <- lm(prestige ~ type, data = Prestige)
fit_type
```

```
##
## Call:
## lm(formula = prestige ~ type, data = Prestige)
##
## Coefficients:
## (Intercept)     typeprof       typewc
##       35.53        32.32         6.72
```

`lm()` automatically converts the factor into a set of indicators, and automatically omits one category from the design matrix. In particular, it omits whichever level is listed first (which may not be the first level to appear in the data!). If you want to have a different category omitted, you need to reorder the levels, placing the category you want to omit first. You can do that with `fct_relevel()` from the **forcats** package. Let's make white-collar (`wc`) the omitted category.

```
Prestige$type <- fct_relevel(Prestige$type, "wc")
levels(Prestige$type)
```

```
## [1] "wc"   "bc"   "prof"
```

```
fit_type_relevel <- lm(prestige ~ type, data = Prestige)
```

We can confirm by checking out the model fit statistics that which category we omit makes no difference to the overall fit of the model, or the predicted values.

```
glance(fit_type)
```

```
##   r.squared adj.r.squared  sigma statistic    p.value df  logLik    AIC
## 1   0.69763       0.69126 9.4986    109.59 2.1168e-25  3 -358.15 724.29
##      BIC deviance df.residual
## 1 734.63   8571.3          95
```

```
glance(fit_type_relevel)  # Should be the same
```

```
##   r.squared adj.r.squared  sigma statistic    p.value df  logLik    AIC
## 1   0.69763       0.69126 9.4986    109.59 2.1168e-25  3 -358.15 724.29
##      BIC deviance df.residual
## 1 734.63   8571.3          95
```

```
# fitted() extracts the fitted value for each observation
all.equal(fitted(fit_type), fitted(fit_type_relevel))
```

```
## [1] TRUE
```

One more thing—it's rare you should need to do this, but you can use `model.matrix()` to extract the design matrix for a fitted regression model.

```
X <- model.matrix(fit_type)
dim(X)
```

```
## [1] 98  3
```

```
head(X)
```

```
##                     (Intercept) typeprof typewc
## gov.administrators            1        1      0
## general.managers              1        1      0
## accountants                   1        1      0
## purchasing.officers           1        1      0
## chemists                      1        1      0
## physicists                    1        1      0
```

8.4.2 Interaction Terms

The syntax for an interactive model is pretty intuitive. Let's look at the joint effect of education and income on occupational prestige.

```
fit_interactive <- lm(prestige ~ education * income, data = Prestige)
fit_interactive
```

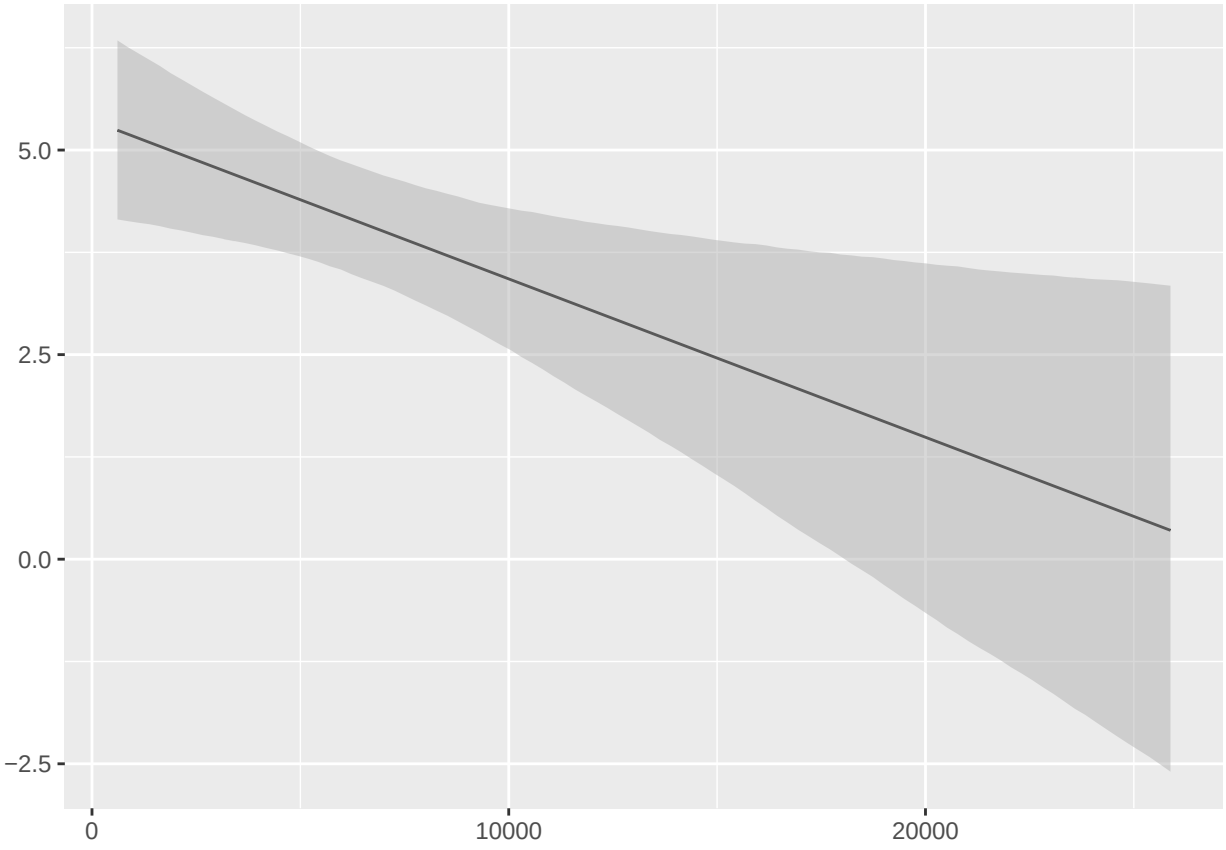
```
##
## Call:
## lm(formula = prestige ~ education * income, data = Prestige)
```

```
##
## Coefficients:
##      (Intercept)         education            income  education:income
##        -2.21e+01          5.37e+00          3.94e-03         -1.96e-04
```

Notice that `lm()` automatically included the lower-order terms for us, so we didn't have to remember which terms to put in. This is particularly useful when you are interacting with a categorical variable that has many categories, or when you are including higher-order interactions.

If you want to plot the (so-called) marginal effect of education as a function of income, you can use the handy function from the **interplot** package.

```
interplot(fit_interactive, var1 = "education", var2 = "income")
```



We see that the marginal effect of education on prestige is high for low-income occupations, but almost nil for an occupation that earns around $25,000/year. (The bars represent a confidence interval—of course, we haven't done any inferential statistics yet.)

8.4.3 Quadratic and Logarithmic Models

The syntax for a quadratic model is a bit weird. You would think you could use a formula like `y ~ x + x^2`, but that won't work. Instead, you have to write `y ~ x + I(x^2)`, as in

the following example.

```
fit_quad <- lm(prestige ~ education + I(education^2) + income, data = Prestige)
fit_quad
```

```
##
## Call:
## lm(formula = prestige ~ education + I(education^2) + income,
##     data = Prestige)
##
## Coefficients:
##    (Intercept)       education  I(education^2)          income
##       10.97951         0.77477         0.15373         0.00128
```

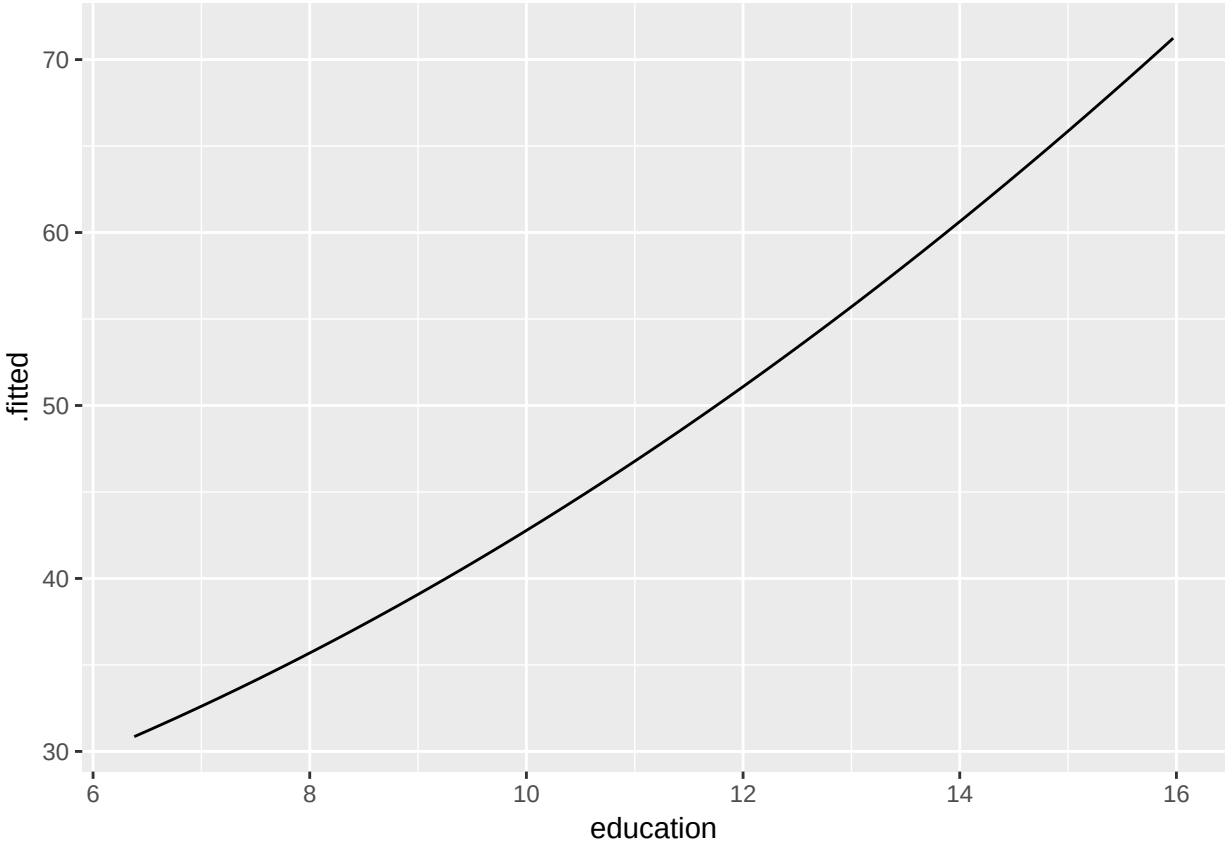
The easiest way to visualize the results of a quadratic model is to create a synthetic dataset, where you vary the relevant variable across its range while holding all the other variables fixed at the same value. Then plug the synthetic dataset into the model to get predicted values.

```
synthetic_data <- data_frame(
    education = seq(min(Prestige$education),
                    max(Prestige$education),
                    length.out = 100),
    income = mean(Prestige$income)
)

synthetic_data <- augment(fit_quad, newdata = synthetic_data)
head(synthetic_data)
```

```
##   education income .fitted .se.fit
## 1    6.3800 6797.9  30.855  2.2842
## 2    6.4769 6797.9  31.122  2.1959
## 3    6.5737 6797.9  31.391  2.1104
## 4    6.6706 6797.9  31.663  2.0280
## 5    6.7675 6797.9  31.939  1.9486
## 6    6.8643 6797.9  32.217  1.8723
```

```
ggplot(synthetic_data, aes(x = education, y = .fitted)) +
    geom_line()
```



In this case, within the range spanned by the sample data, the estimated quadratic relationship is only barely nonlinear.

Finally, to run a logarithmic model, just put `log()` around the variables you want to log.

```
fit_log <- lm(log(prestige) ~ log(income) + education, data = Prestige)
fit_log
```

```
##
## Call:
## lm(formula = log(prestige) ~ log(income) + education, data = Prestige)
##
## Coefficients:
## (Intercept)  log(income)    education
##      0.3373       0.2991       0.0789
```

To visualize the resulting relationship, you can use the same technique as for quadratic models.

Chapter 9

Drawing Inferences

You can think of regression as a descriptive statistic or data reduction method—a simple way to summarize trends and relationships in multivariate data. But for better or worse, most social scientists view regression as a tool for hypothesis testing. This week, we will learn what it is that we're going when we gaze at the "stars" that accompany our regression output.

9.1 The Basics of Hypothesis Testing

Remember the general procedure for testing against a null hypothesis.

1. Choose a test statistic and significance level.
2. Derive the sampling distribution of the test statistic under the null hypothesis.
3. Calculate the value of the test statistic for our sample.
4. Compare the sample test statistic to the sampling distribution under the null hypothesis. If the probability of obtaining a result as least as extreme as ours is at or below the significance level, reject the null hypothesis.

Imagine the null hypothesis is true. Given that, imagine 100 labs run independent tests of the null hypothesis, each using a significance level of 0.05. If they follow the procedure above, on average 5 of the labs will reject the null hypothesis, and 95 will fail to reject the null hypothesis.

What if the null hypothesis is false? What percentage of the labs will falsely reject it? That's the *power* of the test, and it depends on a number of factors: how far off the null hypothesis is, what size sample each lab is drawing, and the significance level.

Before we get into hypothesis tests for regression, let's refresh ourselves on how we draw inferences from a random sample about the population mean.

Suppose we have a sequence of N i.i.d. draws of the random variable X, which we will denote $X_1, \ldots, X_N$, and we are interested in testing the null hypothesis

$$H_0 : E[X] = \mu_0.$$

Let $\bar{X}$ denote the sample mean and S_X denote the sample standard deviation. Define the t statistic as

$$t = \frac{\bar{X} - \mu_0}{S_X/\sqrt{N}}.$$

The denominator of the t statistic is the *standard error*—our estimate of the standard deviation of the sampling distribution under the null hypothesis. The greater the standard error, the more the statistic varies across samples, and thus the less reliable it is in any given sample. Naturally enough, our standard errors decrease with our sample size; the more data we have, the more reliably we can draw inferences.

If X is known to be normally distributed—an assumption that, in the realms political scientists deal with, is usually implausible—then the sampling distribution of t under the null hypothesis is t_{N-1}, the Student's t distribution with $N-1$ degrees of freedom.

If X is not known to be normally distributed, but our sample size is "large" (in practice, $N \geq 30$), we can rely on the Central Limit Theorem. As $N \to \infty$, the distribution of t under the null hypothesis is approximately $N(0,1)$, the normal distribution with mean zero and variance one.

9.2 Variance of OLS

Now let us return to the world of the linear model,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

and the OLS estimator of β,

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

In order to draw inferences on OLS results, the first thing we need to know is the variance of the OLS estimator. You will remember from Stat 1 that the variance of the sample mean, $\bar{X}$, is

$$V[\bar{X}] = \frac{V[X]}{N},$$

whose square root ends up in the denominator of the t-test statistic. We will do something similar for OLS.

Throughout this week, we will maintain the following two assumptions on the error term:

- Strict exogeneity: $E[\epsilon \,|\, \mathbf{X}] = \mathbf{0}$.
- Spherical errors: $V[\epsilon \,|\, \mathbf{X}] = \sigma^2 \mathbf{I}_N$, where $\sigma^2 > 0$.

Without the first assumption, OLS is hopeless to begin with. Without the second assumption, OLS is unbiased and consistent, but not efficient. In a couple of weeks, we will discuss how to draw inferences in the presence of non-spherical errors.

The OLS estimator is a $K \times 1$ vector, so its variance won't be a single number—it will be a $K \times K$ matrix,

$$V[\hat{\beta}] = \begin{bmatrix} V[\hat{\beta}_1] & \text{Cov}[\hat{\beta}_1, \hat{\beta}_2] & \cdots & \text{Cov}[\hat{\beta}_1, \hat{\beta}_K] \\ \text{Cov}[\hat{\beta}_2, \hat{\beta}_1] & V[\hat{\beta}_2] & \cdots & \text{Cov}[\hat{\beta}_2, \hat{\beta}_K] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[\hat{\beta}_K, \hat{\beta}_1] & \text{Cov}[\hat{\beta}_K, \hat{\beta}_2] & \cdots & V[\hat{\beta}_K] \end{bmatrix}.$$

Specifically, the variance of the OLS estimator (treating the covariates as fixed) is

$$V[\hat{\beta} \,|\, \mathbf{X}] = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

See the appendix to this chapter—or any graduate-level econometrics textbook—for how we derive this result.

If we knew σ^2, the variance of the error term, then we could just use the above formula to draw inferences. Realistically, though, we will need to estimate σ^2. We will do so using the residual variance,

$$\hat{\sigma}^2 = \frac{\sum_n (Y_n - \mathbf{x}_n \cdot \hat{\beta})^2}{N - K} = \frac{\text{SSE}}{N - K}.$$

Why do we divide by $N - K$? Remember that when we estimate the variance of a random variable, we divide the squared deviations from the mean by $N-1$ to correct for the degree of freedom we used to estimate the sample mean. The resulting estimator is unbiased. Similarly, when estimating the residual variance of a linear regression model, we need to correct for the K degrees of freedom we used to estimate the model coefficients. Hence we must divide by $N - K$ in order for $\hat{\sigma}^2$ to be unbiased.

Under the spherical error assumption, our estimate of the variance of OLS will therefore be the $K \times K$ matrix

$$\hat{\Sigma} = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

A very important note. If the errors are not spherical, this formula produces a *biased* and *inconsistent* estimate of the sampling variability of OLS. This is true even though the OLS estimate itself is unbiased and consistent. In other words, if we use OLS in the presence of non-spherical errors:

- Our estimates will not be systematically biased away from the population parameter, and the probability of our estimate being any meaningful distance away from the population parameter goes to zero as our sample size increases.
- Our hypothesis tests will **not** perform as advertised—typically, they will lead us to reject the null hypothesis more often than we should—and this problem does **not** go away as our sample size increases.

For the remainder of this week, we will proceed under the assumption of spherical errors. In a couple of weeks, we will discuss how to draw inferences appropriately when this assumption fails to hold.

9.3 Single Variable Hypotheses

Consider a null hypothesis about the population value of a single coefficient, of the form

$$H_0 : \beta_k = b,$$

where b is a fixed constant. Usually, though not always, political scientists concern themselves with null hypotheses of the form $\beta_k = 0$; i.e., the k'th variable has zero (so-called) marginal effect on the response.

We will test this hypothesis using the familiar t-statistic. The *estimated standard error* of $\hat{\beta}_k$ is

$$\text{SE}(\hat{\beta}_k) = \sqrt{\hat{\Sigma}_{kk}},$$

where $\hat{\Sigma}_{kk}$ denotes the k'th element of the diagonal of $\hat{\Sigma} = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$. The "standard errors" that appear alongside your regression output are calculating by taking the square root of the diagonal of this matrix.

The t statistic for the test of the null hypothesis H_0 is in the familiar "estimate divided by standard error" form,

$$t = \frac{\hat{\beta}_k - b}{\text{SE}(\hat{\beta}_k)} = \frac{\hat{\beta}_k - b}{\sqrt{\hat{\Sigma}_{kk}}}.$$

Our mode of inference from there depends on the sample size and on whether we are willing to make a normality assumption.

- If ϵ is normally distributed, then the sampling distribution of our test statistic is t_{N-K}, the t distribution with $N - K$ degrees of freedom.
- As our sample size grows large, the sampling distribution of our test statistic is approximately $N(0, 1)$, the normal distribution with mean zero and variance one. This follows from the Central Limit Theorem, and it holds even if ϵ is *not* normally distributed.

So if you have a small sample, the validity of the standard hypothesis test depends on a normality assumption that may or may not be palatable, depending on the circumstances. Other techniques exist for this situation, but they are beyond the scope of this course. Of course, if your sample is so small that you need to resort to non-standard hypothesis testing techniques in order to draw appropriate inferences, you should probably go back to the drawing board on your study design.

Regardless of your sample size, regression software will compare your test statistics to t_{N-K} to calculate p-values and the results of hypothesis tests. This is innocuous even if you don't

assume normality, since if N is large the t_{N-K} distribution is approximately the same as the $N(0,1)$ distribution (with infinitesimally fatter tails).

Our method of constructing confidence intervals for a single parameter is also analogous to what we do with the sample mean. Let z_α be the critical value of the sampling distribution of our test statistic for our chosen significance level α. For example, the critical value of $N(0,1)$ for significance $\alpha = 0.05$ is $z_\alpha = 1.96$. Then the $(1-\alpha)$-confidence interval around $\hat{\beta}_k$ is

$$\text{CI}_{1-\alpha}(\hat{\beta}_k) = [\hat{\beta}_k - z_\alpha \,\text{SE}(\hat{\beta}_k), \hat{\beta}_k + z_\alpha \,\text{SE}(\hat{\beta}_k)].$$

9.4 Multiple Variable Hypotheses

It is common, especially (though not exclusively) when working with categorical variables or higher-order terms, to have hypotheses involving multiple variables. For example, think of our model from last week,

$$\text{Trump}_n = \beta_1 + \beta_2 \text{Independent}_n + \beta_3 \text{Democratic}_n + \beta_4 \text{Age}_n + \epsilon_n.$$

Remember that β_2 denotes the expected difference between Independents and Republicans (the omitted category) of the same age in their propensity to vote for Trump, and β_3 denotes the expected difference between Democrats and Republicans of the same age.

If our null hypothesis were that Independents and Republicans of the same age had the same chance of voting for Trump, we would state that as

$$H_0 : \beta_2 = 0.$$

But what about the null hypothesis were that Independents and Democrats had the same chance of voting for Trump? We would have to phrase that in terms of multiple coefficients,

$$H_0 : \beta_2 = \beta_3,$$

or equivalently,

$$H_0 : \beta_2 - \beta_3 = 0.$$

Or what if our null hypothesis were that party identification made no difference at all? That would mean Independents and Democrats are both no different than Republicans on average, or in our model notation,

$$H_0 : \begin{cases} \beta_2 = 0, \\ \beta_3 = 0. \end{cases}$$

Each of these, including the simple single-variable hypothesis, is a linear system in β. In other words, we can write each of these hypotheses in the form

$$H_0 : \mathbf{R}\beta - \mathbf{c} = \mathbf{0},$$

where $\mathbf{R}$ is a fixed $r \times K$ matrix (where r is the number of restrictions we intend to test) and $\mathbf{c}$ is a fixed $r \times 1$ vector.

Perhaps the easiest way to test a hypothesis of this form is the *Wald test*. We form the test statistic

$$W = (\mathbf{R}\beta - \mathbf{c})^\top (\mathbf{R}\hat{\Sigma}\mathbf{R}^\top)^{-1}(\mathbf{R}\beta - \mathbf{c}),$$

which, despite all the matrices involved, works out to be a scalar. Under H_0, the asymptotic sampling distribution of W is χ^2_r, the chi-squared distribution with r degrees of freedom.[1] Regression software like R will usually report an F statistic, since the exact (not asymptotic) distribution of W follows an F distribution in the special case of normal residual error.

The Wald test is not just an aggregation of the individual t tests of the coefficients. Two coefficients might each individually be statistically insignificant, yet the Wald test may lead us to reject the null hypothesis that both are zero. Conversely, one of a group of coefficients might be statistically significant, and yet the Wald test may not have us reject the null hypothesis that all are zero.

We already saw a couple of examples of how to use the Wald test with a model with a categorical variable. Let's also quickly consider its use in some other models with less-common specifications.

Imagine an interactive model,

$$Y_n = \beta_1 + \beta_2 X_n + \beta_3 Z_n + \beta_4 (X_n \times Z_n) + \epsilon_n.$$

The interaction term captures how the (so-called) marginal effect of X_n depends on the value of Z_n, and vice versa. If your null hypothesis is that the marginal effect of X_n does not depend on Z_n, you could use perform a t test of

$$H_0 : \beta_4 = 0.$$

But what if your null hypothesis is that the marginal effect of X_n is *always* zero, regardless of the value of Z_n? Some people make the unfortunate mistake of testing this via the null hypothesis

$$H_0 : \beta_2 = 0,$$

but that only means the marginal effect of X_n is zero *when* $Z_n = 0$. What you want is the composite null

$$H_0 : \begin{cases} \beta_2 = 0, \\ \beta_4 = 0, \end{cases}$$

or, in matrix form,

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

[1] You will notice that, for a single-variable hypothesis $H_0 : \beta_k = b$, the Wald statistic reduces to the square of the t statistic. Since the asymptotic distribution of the t statistic is standard normal under the null hypothesis, it follows that the asymptotic distribution of its square is χ^2_1 under the null hypothesis.

Similarly, think of the quadratic model,

$$Y_n = \beta_1 + \beta_2 X_n + \beta_3 X_n^2 + \epsilon_n.$$

The null hypothesis that the (so-called) marginal effect of X_n is constant is equivalent to

$$H_0 : \beta_3 = 0.$$

But if we wanted to test against the null hypothesis that the marginal effect of X_n is *always* zero, we would have to use a composite null,

$$H_0 : \begin{cases} \beta_2 = 0, \\ \beta_3 = 0. \end{cases}$$

9.5 Appendix: Full Derivation of OLS Variance

We will assume strict exogeneity,

$$E[\epsilon \,|\, \mathbf{X}] = \mathbf{0},$$

and spherical errors,

$$V[\epsilon \,|\, \mathbf{X}] = E[\epsilon\epsilon^\top \,|\, \mathbf{X}] = \sigma^2 \mathbf{I}.$$

A useful thing to know is that since $\mathbf{X}^\top\mathbf{X}$ is symmetric, so is its inverse:

$$[(\mathbf{X}^\top\mathbf{X})^{-1}]^\top = (\mathbf{X}^\top\mathbf{X})^{-1}.$$

Let's start by deriving the variance from our formula for a vector-valued random variable,

$$V[C] = E\left[(C - E[C])(C - E[C])^\top\right].$$

For the OLS estimator $\hat{\beta}$, we have

$$\begin{aligned}
V[\hat{\beta} \,|\, \mathbf{X}] &= E\left[\left(\hat{\beta} - E[\hat{\beta}]\right)\left(\hat{\beta} - E[\hat{\beta}]\right)^\top \,|\, \mathbf{X}\right] \\
&= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^\top \,|\, \mathbf{X}\right] \\
&= E\left[\hat{\beta}\hat{\beta}^\top - 2\beta\hat{\beta}^\top + \beta\beta^\top \,|\, \mathbf{X}\right] \\
&= E\left[\hat{\beta}\hat{\beta}^\top \,|\, \mathbf{X}\right] - 2\beta E\left[\hat{\beta}^\top \,|\, \mathbf{X}\right] + \beta\beta^\top \\
&= E\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\mathbf{Y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \,|\, \mathbf{X}\right] - \beta\beta^\top \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E\left[\mathbf{Y}\mathbf{Y}^\top \,|\, \mathbf{X}\right]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} - \beta\beta^\top.
\end{aligned}$$

Since $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, we have

$$\begin{aligned}
E\left[\mathbf{Y}\mathbf{Y}^\top \,|\, \mathbf{X}\right] &= E\left[(\mathbf{X}\beta + \epsilon)(\mathbf{X}\beta + \epsilon)^\top \,|\, \mathbf{X}\right] \\
&= E\left[\mathbf{X}\beta\beta^\top\mathbf{X}^\top + 2\epsilon\beta^\top\mathbf{X}^\top + \epsilon\epsilon^\top \,|\, \mathbf{X}\right] \\
&= \mathbf{X}\beta\beta^\top\mathbf{X}^\top + 2\underbrace{E[\epsilon \,|\, \mathbf{X}]}_{=\mathbf{0}}\beta^\top\mathbf{X}^\top + E[\epsilon\epsilon^\top \,|\, \mathbf{X}] \\
&= \mathbf{X}\beta\beta^\top\mathbf{X}^\top + \sigma^2\mathbf{I}.
\end{aligned}$$

Continuing from above, we have

$$\begin{aligned} V[\hat{\beta} \,|\, \mathbf{X}] &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E\left[\mathbf{Y}\mathbf{Y}^\top \,|\, \mathbf{X}\right] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} - \beta\beta^\top \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta\beta^\top \mathbf{X}^\top + \sigma^2 \mathbf{I}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} - \beta\beta^\top \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X}) \beta\beta^\top (\mathbf{X}^\top \mathbf{X}) (\mathbf{X}^\top \mathbf{X})^{-1} \\ &\quad + \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} - \beta\beta^\top \\ &= \beta\beta^\top + \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} - \beta\beta^\top \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

A slightly easier way to get there would be to apply two of the helpful properties of variance. You'll remember from the study of scalar-valued random variables that, for a random variable A and scalars c and d,

$$\begin{aligned} V[cA] &= c^2 V[A], \\ V[A + d] &= V[A]. \end{aligned}$$

Similarly, for an $m \times 1$ vector random variable $\mathbf{A}$, a fixed $n \times m$ matrix $\mathbf{C}$, and a fixed $m \times 1$ vector $\mathbf{d}$, we have

$$\begin{aligned} V[\mathbf{C}\mathbf{A}] &= \mathbf{C} V[\mathbf{A}] \mathbf{C}^\top, \\ V[\mathbf{A} + \mathbf{d}] &= V[\mathbf{A}]. \end{aligned}$$

Consequently,

$$\begin{aligned} V\left[\hat{\beta} \,|\, \mathbf{X}\right] &= V[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \,|\, \mathbf{X}] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top V[\mathbf{Y} \,|\, \mathbf{X}] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top V[\mathbf{X}\beta + \epsilon \,|\, \mathbf{X}] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top V[\epsilon \,|\, \mathbf{X}] \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\sigma^2 \mathbf{I}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

9.6 Appendix: Regression Inference in R

As usual, we will rely on the **tidyverse** and **broom** packages. We will also be using the **car** package, not only for data but also for its hypothesis testing functions.

```
library("tidyverse")
library("broom")
library("car")
```

Using the `Prestige` data, let us regress occupational prestige on the type of occupation (blue collar, white collar, or professional) and its average income and education.

```
data("Prestige", package = "car")
fit <- lm(prestige ~ type + education + income, data = Prestige)
```

`summary()` prints the "regression table" containing the following information:

- Estimate of each coefficient.
- Estimated standard error of each coefficient estimate.
- t statistic for each coefficient estimate, for the test against the null hypothesis that the population value of the corresponding coefficient is zero ($H_0 : \beta_k = 0$).
- p value (two-tailed)[2] for the aforementioned hypothesis test.

```
summary(fit)
```

```
##
## Call:
## lm(formula = prestige ~ type + education + income, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.953  -4.449   0.168   5.057  18.632
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.622929   5.227525   -0.12     0.91
## typeprof     6.038971   3.866855    1.56     0.12
## typewc      -2.737231   2.513932   -1.09     0.28
## education    3.673166   0.640502    5.73  1.2e-07
## income       0.001013   0.000221    4.59  1.4e-05
##
## Residual standard error: 7.09 on 93 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.828
## F-statistic:  118 on 4 and 93 DF,  p-value: <2e-16
```

This is useful for a quick check on the output, but not so much if you want to use the standard errors for further calculations (e.g., making a plot). To extract the standard errors, use the `tidy()` function from **broom**. This returns a data frame whose columns correspond to the `summary()` output.

```
fit_results <- tidy(fit)
fit_results
```

```
##          term   estimate  std.error statistic    p.value
## 1 (Intercept) -0.6229292 5.22752549  -0.11916 9.0540e-01
## 2    typeprof  6.0389707 3.86685510   1.56173 1.2175e-01
## 3      typewc -2.7372307 2.51393240  -1.08882 2.7904e-01
## 4   education  3.6731661 0.64050162   5.73483 1.2052e-07
## 5      income  0.0010132 0.00022092   4.58628 1.4049e-05
```

[2]One-tailed tests, while unproblematic in theory, in practice are usually a signal that the two-tailed test was insignificant and the author is fudging it. Don't send a bad signal; always use two-tailed tests.

```
fit_results$std.error
```

```
## [1] 5.22752549 3.86685510 2.51393240 0.64050162 0.00022092
```

To extract the full (estimated) variance matrix, use the `vcov()` function.

```
vcov(fit)
```

```
##             (Intercept)    typeprof      typewc   education      income
## (Intercept)  2.7327e+01  1.6637e+01  7.39848733 -3.1965e+00  9.9963e-05
## typeprof     1.6637e+01  1.4953e+01  6.79707308 -2.1239e+00 -4.9843e-06
## typewc       7.3985e+00  6.7971e+00  6.31985613 -1.1062e+00  1.3108e-04
## education   -3.1965e+00 -2.1239e+00 -1.10618421  4.1024e-01 -4.3335e-05
## income       9.9963e-05 -4.9843e-06  0.00013108 -4.3335e-05  4.8805e-08
```

Luckily, you shouldn't often need to extract the individual standard errors or the variance matrix. R has convenience functions to perform most of the calculations you would care about.

To obtain confidence intervals for the regression coefficients, use the `confint()` function.

```
confint(fit)
```

```
##                    2.5 %     97.5 %
## (Intercept) -1.1004e+01  9.7579004
## typeprof    -1.6398e+00 13.7177785
## typewc      -7.7294e+00  2.2549408
## education    2.4013e+00  4.9450753
## income       5.7449e-04  0.0014519
```

```
confint(fit, level = 0.99)  # Changing the confidence level
```

```
##                    0.5 %     99.5 %
## (Intercept) -1.4370e+01 13.1240626
## typeprof    -4.1298e+00 16.2077638
## typewc      -9.3482e+00  3.8737381
## education    1.9888e+00  5.3575138
## income       4.3224e-04  0.0015941
```

```
confint(fit, "education")   # Only for one coefficient
```

```
##             2.5 % 97.5 %
## education 2.4013 4.9451
```

You may also be interested in testing against null hypotheses other than each individual coefficient being zero. This is where the `linearHypothesis()` function from the car package comes in. For example, suppose we wanted to test against the null hypothesis that the population coefficient on education is 3.

```
linearHypothesis(fit, "education = 3")
```

```
## Linear hypothesis test
##
## Hypothesis:
## education = 3
##
## Model 1: restricted model
## Model 2: prestige ~ type + education + income
##
##   Res.Df  RSS Df Sum of Sq   F Pr(>F)
## 1     94 4737
## 2     93 4681  1      55.6 1.1    0.3
```

We care about the last two columns of the bottom row. `F` gives us the test statistic,[3] and `Pr(>F)` gives us the associated p-value. Here our p-value is 0.3, so we wouldn't reject the null hypothesis that the population coefficient on education is 3.

Two quick notes on the `linearHypothesis()` function:

- Make sure to place your hypothesis in quotes. (Or if you have multiple hypotheses, that you use a vector of quoted strings.) The function will not work if you run something like `linearHypothesis(fit, education = 3)`.
- Make sure the name of the coefficient(s) you're testing are exactly the same as in the regression output. This requires particular care when you're dealing with factor variables, interactions, or quadratic terms.

Of course, for a univariate hypothesis test like this one, we could have just used the confidence interval to figure out the answer. The real value of `linearHypothesis()` comes in simultaneously testing hypotheses about multiple coefficients—i.e., the Wald test.

For example, let's test the null hypothesis that the population coefficients on white-collar and professional are the same.

```
linearHypothesis(fit, "typewc = typeprof")
```

```
## Linear hypothesis test
##
## Hypothesis:
## - typeprof  + typewc = 0
##
## Model 1: restricted model
## Model 2: prestige ~ type + education + income
##
```

[3] Why an F statistic instead of a t statistic? If the random variable Z has a t distribution with n degrees of freedom, then Z^2 has an F distribution with $1, n$ degrees of freedom. The F statistic generalizes better to the case of multiple hypotheses.

```
##   Res.Df  RSS Df Sum of Sq  F Pr(>F)
## 1     94 5186
## 2     93 4681  1       505 10 0.0021
```

We would reject this null hypothesis except under particularly stringent significance levels (less than 0.002).

What about the composite hypothesis that the population coefficients on white-collar and professional both equal zero? To test this, we pass a *vector* of hypotheses.

```
linearHypothesis(fit, c("typewc = 0", "typeprof = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## typewc = 0
## typeprof = 0
##
## Model 1: restricted model
## Model 2: prestige ~ type + education + income
##
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     95 5272
## 2     93 4681  2       591 5.87  0.004
```

The p-value corresponding to this hypothesis test is 0.004. This illustrates a key feature of composite hypothesis tests. You'll remember from the original regression output that neither of the occupational indicators were significant on their own.

```
summary(fit)
```

```
##
## Call:
## lm(formula = prestige ~ type + education + income, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.953  -4.449   0.168   5.057  18.632
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.622929   5.227525   -0.12     0.91
## typeprof     6.038971   3.866855    1.56     0.12
## typewc      -2.737231   2.513932   -1.09     0.28
## education    3.673166   0.640502    5.73  1.2e-07
## income       0.001013   0.000221    4.59  1.4e-05
##
## Residual standard error: 7.09 on 93 degrees of freedom
```

```
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.835,  Adjusted R-squared:  0.828
## F-statistic:  118 on 4 and 93 DF,  p-value: <2e-16
```

So, operating at conventional significant levels, we would:

- Fail to reject the null that the coefficient on professional is zero
- Fail to reject the null that the coefficient on white-collar is zero
- Reject the null that the coefficients on both are zero

This seems contradictory, but it just illustrates the limits of drawing inferences from a finite sample. What the results of these hypothesis tests are telling us is that we have enough information to conclude that there are differences in prestige across occupational categories, holding average education and income fixed. However, we do not have enough information to identify exactly where those differences are coming—i.e., exactly which of the two categories it is that differs from the omitted baseline.

Chapter 10

The Statistical Crisis in Science

As a general rule, I do not believe the statistical results reported in political science publications. More to the point, absent compelling evidence to the contrary, I assume that:

- Reported effect sizes are biased upward in magnitude. The marginal effect of the key variable(s) in the population is likely less dramatic than what is reported.
- Reported p-values are biased downward. The probability of making a Type I error, in case the null hypothesis were false and one followed the *actual* procedure that led to the reported results, is greater than what is reported.

My position is not one of blind skepticism. It follows, perhaps ironically, from empirical research showing that far more than 5 percent of studies do not replicate upon re-testing (Open Science Collaboration, 2015). Ioannidis (2005) puts it bluntly: "Most published research findings are false." Today we'll discuss why.

I'll follow Young et al. (2008)'s economic analogy. Consider scientific publication as an economic activity, where researchers "sell" findings to journals in exchange for prestige.[1] The *demand-side problem* is that journals will only "buy" statistically significant findings. Even absent any effects on author behavior, this practice makes published findings a biased sample of actual findings. But of course there are effects on author behavior. The *supply-side problem* is that authors try to produce "statistically significant" findings instead of scientifically sound findings.

There is far more out there on the replication crisis in science than we can cover in a week. Sanjay Srivastava's faux syllabus, "Everything is fucked", provides a more comprehensive treatment.

[1] Have you ever wondered why academic journal subscriptions cost unholy amounts of money, yet the authors, reviewers, and (usually) editors are unpaid?

10.1 Publication Bias

If you open up an issue of any empirically oriented political science journal, you will not read many abstracts that conclude "We were unable to reject the null hypothesis of no effect." You probably won't see any. The prevailing attitude of reviewers and editors is that only significant results are interesting and only interesting results are worth publishing—so only significant results get published.

Consequently, published empirical findings are not a representative sample of all empirical findings. Andrew Gelman calls this the *statistical significance filter*: the publication process only reveals the findings of some studies, namely those that achieve statistical significance. If you draw your beliefs from scientific journals (particularly prestigious ones, as Ioannidis (2008) notes), you will end up with some false ideas about how the world works.

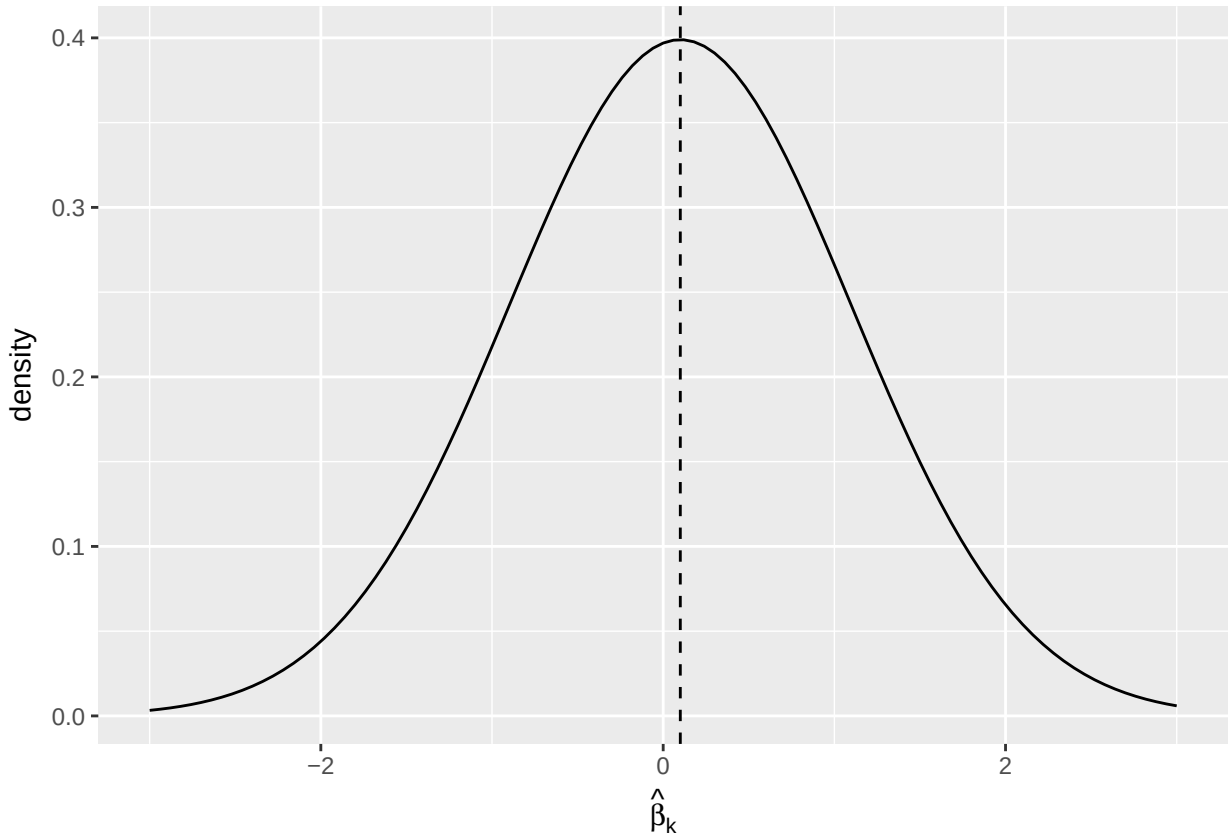
Some of these beliefs will be Type I errors: you will reject null hypotheses that are true. Suppose there is a treatment T that has no effect on an outcome Y, and 100 labs run separate experiments of the effect of T on Y. We would expect about 95 of these experiments to (correctly) fail to reject the null hypotheses, and about 5 to (incorrectly) reject it. But if some of the significant findings get published and none of the insignificant ones do, you will end up incorrectly believing the treatment affects the outcome.

But the statistical significance filter has another, less obvious—and thus more pernicious—effect on our inferences. Assume that the null hypothesis is indeed false: that the treatment T has an effect on the outcome Y. Suppose once again that 100 labs run separate experiments of the effect of T on Y. Depending on the power of the experiments, some proportion of them will (incorrectly) fail to reject the null hypothesis, and the remainder will (correctly) reject it. Because of the statistical significance filter, only the ones that reject the null hypothesis will get published.

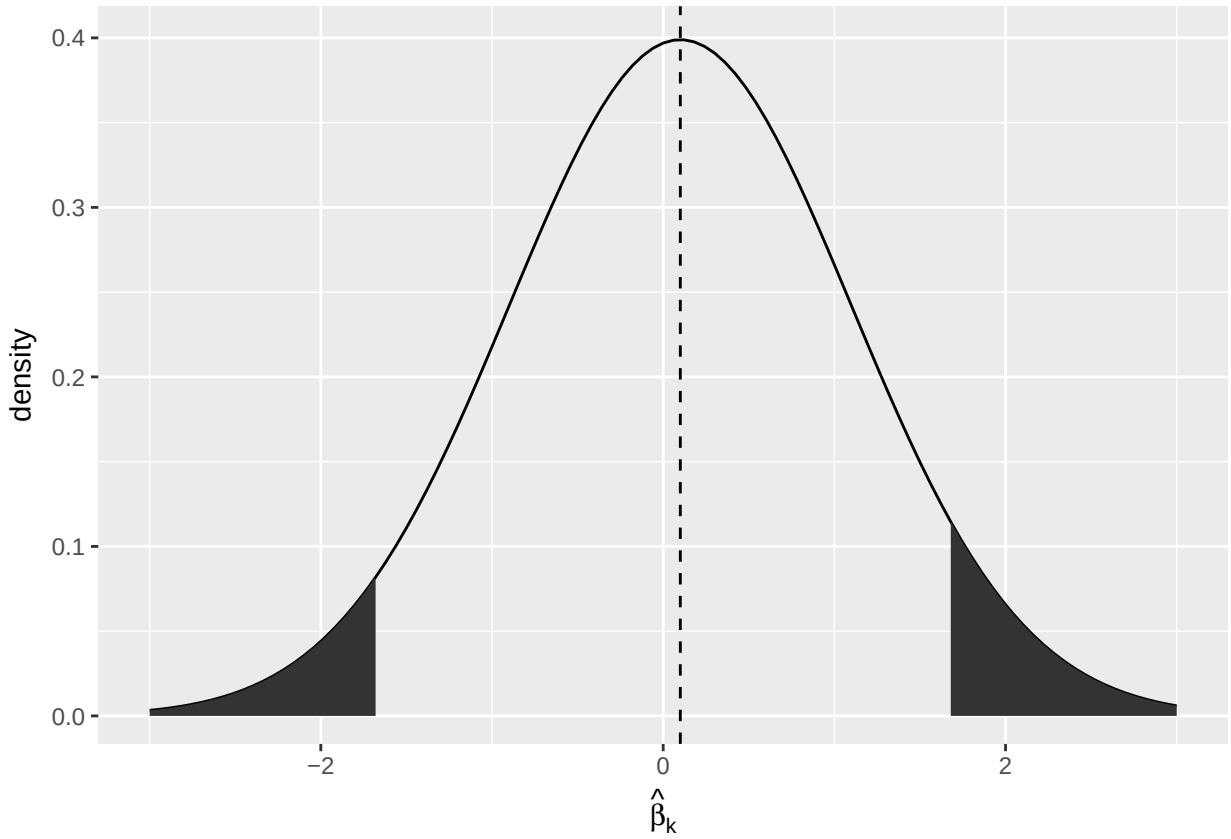
That's not so bad, right? Only the studies that reject the null hypothesis get published, but the null hypothesis is wrong! The problem comes in when we want to evaluate the size of the effect—what political scientists like to call "substantive significance."[2] On average, the statistically significant studies will tend to overestimate the magnitude of the effect. Viewing studies through the statistical significance filter, we will correctly infer that there is an effect, but we will systematically overestimate how strong it is.

Why does the statistical significance filter result in an overestimate of effect magnitudes? Imagine that the population regression coefficient is real but small, say $\beta_k = 0.1$. Then the sampling distribution of $\hat{\beta}_k$ will look something like the following graph.

[2]Mini rant: In my admittedly short career in political science, I have seen zero talks or papers claim to have found a statistically significant but substantively insignificant result. I have, however, seen talks that claimed a 0.001% increase constituted a substantively significant finding. Without a threshold for substantive significance that is decided on *before* the results are obtained, any claim about substantive significance is incredible.



Since the population parameter is close to zero, we are rather likely to yield a sample estimate close to zero. With a small sample size, sample estimates close to zero are likely to be statistically insignificant. Under the statistical significance filter, only the results in the "tails" of the distribution will end up being published.



The first time I read about this result, on Andrew Gelman's blog, I didn't believe it. (I *should* have believed it, because he's a professional statistician and I'm not.) So I fired up R and ran a simulation to answer: if we only report our estimate of β_k when it's statistically significant, will we overestimate its magnitude on average? In your R session this week, you will run a version of this same simulation.

10.2 *p*-Hacking

The statistical significance filter is a demand-side problem. The demand (by journals) for "insignificant" findings is too low. This in turn creates supply-side problems. Scientists' careers depend on their ability to publish their findings. Since there is no demand for insignificant findings, scientists do what they can to conjure up significant results. In the best case scenario, this means devoting effort to projects with a high prior probability of turning up significant, rather than riskier endeavors. In the worst case, it means engaging in vaguely-to-definitely unethical statistical practices in a desperate search for significance.

One way to *p*-hack is to just define the significance level *post hoc*.

Luckily, this is pretty transparent. The convention, for better or worse, is a significance level of 0.05, and it's easy to notice deviations from the convention. Look for the "daggers" in people's regression tables, or language like "comes close to statistical significance". Matthew Hankins' blog post "Still Not Significant" is a comprehensive compendium of the weasel

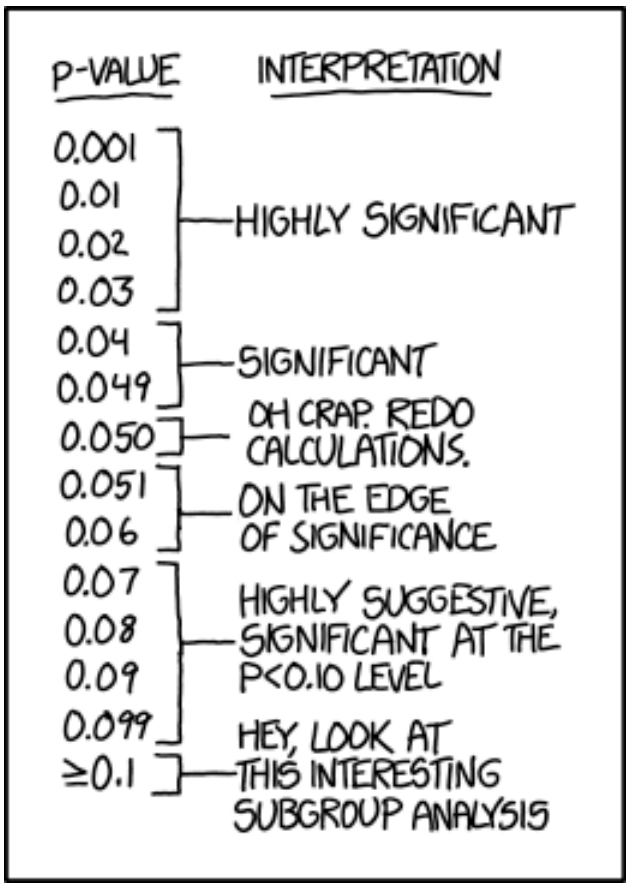


Figure 10.1: XKCD #1478.

language people use to try to dress up their insignificant findings. See also Pritschet et al. (2016).

The more pernicious form of p-hacking is going fishing for something significant after one's original hypothesis test "fails." Let us once again imagine a lab performing an experiment. They are interested in the effect of a treatment T on an outcome Y. To make it concrete, suppose the treatment is reading a particular editorial, and the outcome is where the respondent places himself or herself on a left-right ideological scale ranging between 0 and 1. The lab spends a lot of time and money recruiting subjects, running the experiment, and tabulating the data. They get their spreadsheet together, load their data into R, test for a treatment effect … and fail to reject the null hypothesis.

Damn. All that effort wasted, for a result that can't be published. But wait! The op-ed was written by a man, and his picture appeared next to it. It seems plausible that it might only have an effect on men, or only one on women. So just to see, the lab re-runs the test once just for men and once just for women. They get a p-value just below 0.05 for the male subsample! Hooray! This is at least potentially a publishable finding!

What's wrong with this picture? Let's go back to the formal definition of the significance level.

> The significance level of a hypothesis test is the probability of rejecting the null hypothesis when the null hypothesis is true.

If the null hypothesis is true, and 100 labs run the same experiment on it, we should expect about 5 of them to end up incorrectly rejecting the null hypothesis. Similarly, go back to the formal definition of a p-value.

> The p-value of a test statistic is the probability of yielding a test statistic at least as extreme when the null hypothesis is true.

If the null hypothesis is true, we should expect only about 10 out of 100 labs to end up with $p \leq 0.10$, 5 out of 100 to have $p \leq 0.05$, and so on.

The problem with this hypothetical procedure—testing *post hoc* for effects within subgroups after the main test comes back insignificant—is that the stated significance level is not the real significance level. If you run three different tests and reject the null hypothesis if *any* of them comes back with $p \leq 0.05$, you will reject the null hypothesis more often than 5% of the time. In our running hypothetical example, the lab's reported p-value of 0.05 is a lie.

There are many ways to p-hack:

- Splitting up data by subgroups *post hoc*
- Changing the set of variables you control for
- Changing the operationalization of the covariate of interest or the response variable
- Changing the time period of the analysis
- Stopping data collection as soon as $p \leq 0.05$

What all these have in common is that the final test you report depends on the result of some earlier test you ran. All standard hypothesis tests assume that you didn't do anything

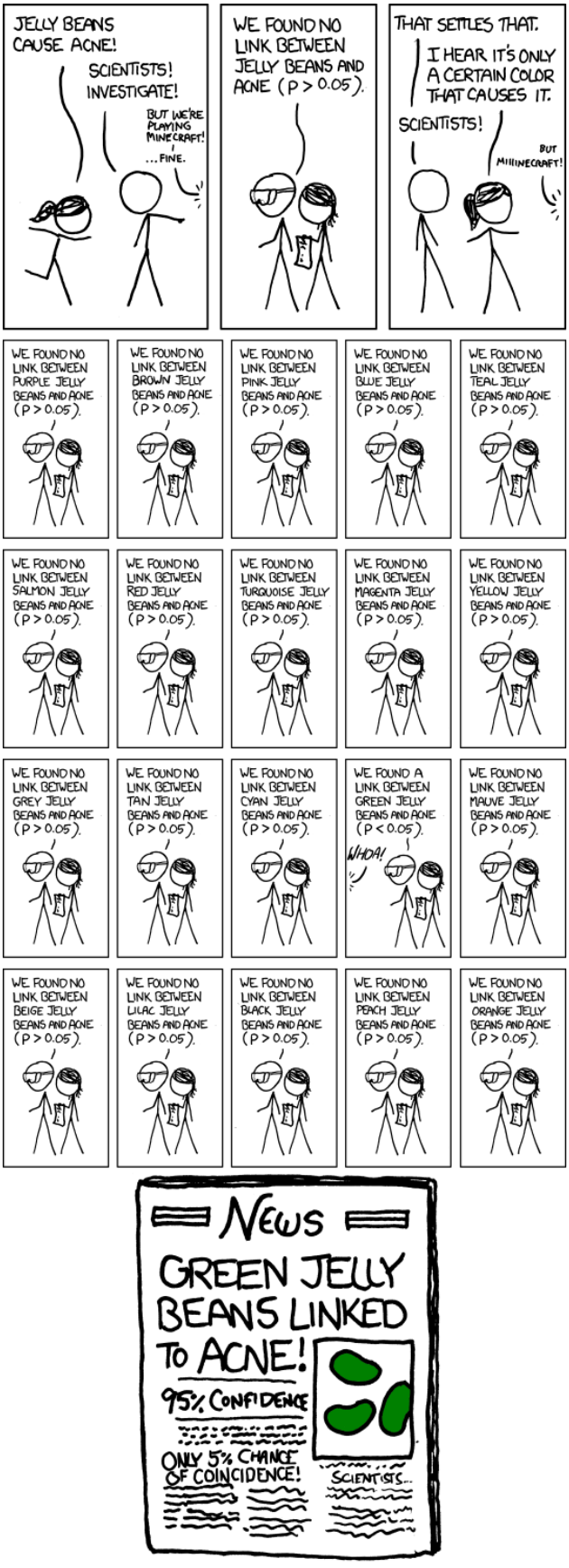


Figure 10.2: XKCD #882.

like this—that this was the only test you ran, that your initial results didn't influence your choice of further tests. It is unethical to report the nominal p-value (i.e., the value your computer spits out) from a p-hacked test, because the true probability of getting a result at least as extreme is greater than the nominal value.

10.3 What to Do

At a macro level, we should probably:

- Assume the magnitudes of published results are exaggerated, and adjust our own beliefs accordingly.
- Collect new data to replicate published findings, and adjust our beliefs in the direction of the replication results.
- When reviewing others' papers, don't judge on the basis of significance. Try to be "results-blind." Assess whether the research design is well suited to address the question at hand, not whether it turned up the results the author wanted, or the results you want, or interesting or surprising or counterintuitive results, etc.
- When writing your own papers, focus on research designs that are clever and novel. Write papers that will be interesting to the political science community regardless of whether the results are statistically significant.

And at a more micro level, to avoid p-hacking in your own research:

- Decide exactly which hypothesis you want to test and which test to run before you collect your data, or at least before running any analysis on it.
- Report every test you perform on the data, and only highlight results that are robust across tests.
- Randomly split your sample before performing any tests. Go wild with the first half of the sample looking for an interesting hypothesis. Then test that hypothesis on the other half of the sample (and report the results whether they come out in your favor or not). Equivalently, hack your pilot data and then go out and collect new data to try to replicate your hacked initial hypothesis.
- Apply a correction for multiple testing problems, or use computational methods (e.g., bootstrap) to calculate the distribution of a data-conditional test statistic under the null hypothesis.

Bibliography

Bowers, J. (2011). Six Steps to a Better Relationship with Your Future Self. *The Political Methodologist*, 18(2):2–8.

Braumoeller, B. F. (2004). Hypothesis testing and multiplicative interaction terms. *International organization*, 58(04):807–820.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83(401):123–127.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, 22(4):341–352.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, 2(8):e124.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648.

Leek, J. (2015). *The Elements of Data Analytic Style*. Leanpub.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Pritschet, L., Powell, D., and Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological science*, 27(7):1036–1042.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10).

Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., and Wilson, P. (2014). Best Practices for Scientific Computing. *PLOS Biology*, 12(1):e1001745.

Young, N. S., Ioannidis, J. P., and Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Med*, 5(10):e201.