Demystifying Structural Equation Modeling

Jonathan Amburgey and Sean Raleigh, Westminster College (Salt Lake City, UT)

2022-05-18

Contents

Introduction

Welcome to our book on structural equation modeling!

[THIS BOOK IS A WORK IN PROGRESS. FEEL FREE TO PERUSE WHATEVER CONTENT YOU FIND HERE, BUT THE FINAL VERSION WILL NOT BE READY UNTIL SOMETIME IN 2023.]

If you want, you can also download this book as a PDF or EPUB file. Be aware that the print versions are missing some of the richer formatting of the online version.

Some history

In 2016, Jonathan and Sean embarked upon a bold experiment, asking the question, "Is is possible to teach structural equation modeling (SEM) to undergraduates with little statistical background?" To make things even more exciting, we attempted to do so in a special topics course lasting only one month during our May Term at Westminster College (Salt Lake City, UT).

In such an endeavor, we had to temper our expectations, of course. The goal was not to produce competent practitioners who would subsequently go on to do serious research using SEM techniques. We were quite happy that, at the end of May, we had undergraduates who were able to put together a simple final project that required them to find some data, posit a model, fit the model in R, interpret the output, and check a few model fit statistics. Some exposure to the topic and some appreciation of its power were satisfying enough. In fact, we think we got a little more out of it than that: we are reasonably confident that most of our students had developed—at that point right after taking the course—the ability to read a research article with an SEM model and have at least some idea what the article was talking about. We called it a win!

We repeated the experiment with some modifications to our materials and pedagogy in 2018. By that point, it was clear that finding textbooks and articles to assign to students was challenging. There are some great books out there,

but they are mostly aimed at graduate students. Even the ones labeled "introductory" were often far from that for the typical undergraduate with limited statistical training.

We decided that we could write our own textbook that would fill this hole in the literature. The book that follows is the fruit of our efforts.

Sean was granted sabbatical in Spring 2020 and proposed to use that time to start writing the book in preparation for running the May Term course again in May, 2020. And, well, we all know how that went…

Once the pandemic subsided enough for us to offer the course in person again, we attempted it again in May, 2022. [TO BE CONTINUED]

Our philosophy

As we mentioned before, our motivation for writing the book was driven by the difficulty we had finding readings for the students. Perhaps that begs the question, should one even try teaching a topic as "difficult" or "advanced" as structural equation modeling to the audience we had? To be sure, the traditional approaches already on the market seem to assume a lot more background than we had at our disposal. And the books that claim to assume less background…well, sometimes they require more than they let on.

The prerequisite for our class is an intro stats class that covers pretty standard material for such a course: hypothesis testing and confidence intervals for one and two proportions, one and two means (and paired means), ANOVA, chi-squared, and simple linear regression. We also benefit in that our intro course introduces students to R. (For those lacking R background, the first four modules here [FIX THIS LINK ONCE THE DATA 220 BOOK IS ALSO FULLY ONLINE] should suffice as a basic introduction to R and R Markdown sufficient for success in this course.)

To respect our students, we made some very deliberate choices about the way our book would be structured.

- *Make the book free and open source.*

Students have enough trouble in their lives and shouldn't be exposed to the extortionate practices of most textbook publishers. Not only is this book freely available online, it's also published under a permissive open source license (the MIT license) that allows folks to "use, copy, modify, merge, publish, distribute, sublicense, and/or sell" their own versions of the book as desired. Furthermore, any derivative of the book must also abide by the same open standards. So our book is both *libre* and *gratis* (or, in more common parlance, "free as in speech" and "free as in beer").

- *Start from scratch.*

Explain everything from the beginning in terms that are as simple as possible. Some of the first few chapters may look like review for students. Even if it is, of course, that review gives students confidence to tackle upcoming new material. But you might be surprised at some of the novel ways we explain seemingly familiar concepts. All the exposition has an eye toward direct application in later chapters, so what might seem a little idiosyncratic at first is motivated by a desire to smooth the pathways into later concepts.

- *Incorporate active learning into everything.*

The chapters are structured to work as templates for classroom experiences. They intersperse conceptual explanation with activities designed to reinforce those concepts and lead students to important conclusions. These learning activities will appear framed in blue boxes [CHANGE THIS IF WE ESTABLISH A CUSTOM CALLOUT] like this:

Hey, kids! Stop and do this activity here!

- *Do the math and do it well.*

One common thread we see in a lot of SEM books is a tendency to sweep most of the math under the rug. The intention comes from a good place; mathematics can appear intimidating and, therefore, may seem to serve as a deterrent to learning. To be sure, there are some complex mathematical ideas in SEM that are inaccessible to our audience. At the same time—and, in fairness, this may be due to Sean's bias as a mathematician—we truly believe that the mathematics, carefully explained, can illuminate student understanding. The more mathy sections may need additional instructor support for students without a strong math background. But all it takes is some relatively straightforward algebra to nail down some concepts that most books ignore. A good example of this is investing time in the rules for manipulating variances and covariances. This allows students to calculate the "model-implied matrix" that is only cryptically referenced in most textbooks. However, we do skip the math sometimes. For example, a lot of the math behind model fit indices is left unexplained. At the very least, we hope to be transparent about our choices to include or exclude certain mathematical details.

- *Use "nice" data.*

Finding data is hard, so we rely a lot on data sets that other textbooks and R package authors make available (with due attribution, of course). To keep things simple for this course, we work almost exclusively with numerical (quantitative) data. [MODIFY THIS IF WE END UP WORKING WITH BINARY CATEGORICAL EXOGENOUS VARIABLES (CODED 0/1) AT SOME POINT.]

- *Be careful about diagrams.*

Learning about complex models induces a sizable cognitive load. Shortcuts in diagrams tend to confuse students. For example, if error terms are truly latent variables, they should be drawn as circles or ellipses and not hidden, even if an advanced practitioner "knows" they're there. Variances and covariances among exogenous variables should always appear as well. We take the time to build up a consistent pictographic representation of every part of a model. (Each chapter is introduced with an archetypal diagram that illustrates that chapter's content.) Then we stick to that representation throughout the book.

- *Be careful about notation.*

While it may be the industry standard, LISREL notation is needlessly complex for undergraduate students. We take a consistent and simple approach to notation that represents all variables using UPPERCASE names and all path values using lowercase names. Abstract variables tend to be called something like X when exogenous and Y when endogenous. Real-world variables have contextually meaningful names. For those interested in reading the research literature, we have included an appendix describing LISREL notation.

Course structure

We use this book to teach a 2-credit-hour course. (Even though it's a special topics course in our May Term, the number of contact hours for students is equivalent to a semester-long, 2-credit-hour course.)

[ADD INFO HERE AS WE DECIDE HOW MUCH IS REASONABLE TO COVER. IF WE WANT THE BOOK TO BE USABLE IN A 4-CREDIT-HOUR COURSE, WHAT ADDITIONAL MATERIAL SHOULD WE CONSIDER INCLUDING?]

Onward and upward

We hope you enjoy our textbook. Please send us your feedback!

–Jonathan Amburgey (jamburgey@westminstercollege.edu)

–Sean Raleigh (sraleigh@westminstercollege.edu)

Chapter 1

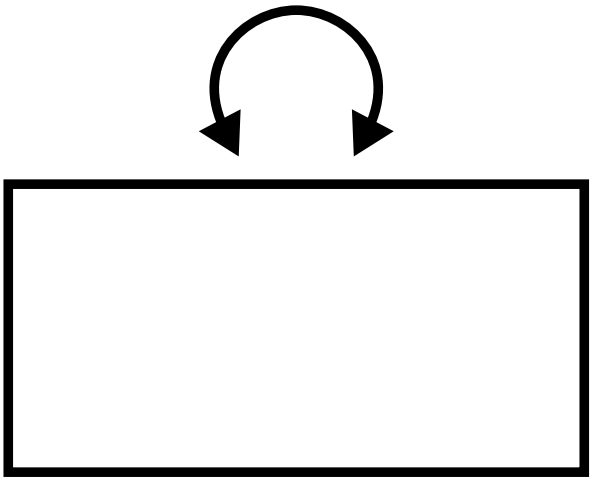
Variables and measurement

1.1 First section

[SOMEWHERE NEED TO MENTION "CONSTANT" VARIABLES, OR VARIABLES THAT TAKE ONLY ONE VALUE.]

Chapter 2

Variance



2.1 A quick refresher on the mean

Most of us were taught how to calculate the mean of a variable way back in elementary school: add up all the numbers and divide by the size of the group of numbers. In a statistics context, we often use a "bar" to indicate the mean of a variable; in other words, if a variable is called X, the mean is denoted $\overline{X}$. Remembering that we always use n to represent the sample size, the formula is

$$\overline{X} = \frac{\sum X}{n}$$

(In case you forgot, the Greek letter Sigma Σ stands for "sum" and means "add up all values of the thing that follows".)

Here is a small data set we'll use throughout this chapter as a simple example we can work "by hand":

3, 4, 5, 6, 6, 7, 8, 9

Calculate the mean of this set of eight numbers.

2.2 Calculating variance

Variance is a quantity meant to capture information about how spread out data is.

Let's build it up step by step.

The first thing to note about spread is that we don't care how large or small the numbers are in any absolute sense. We only care how large or small they are *relative to each other*.

Look at the numbers from the earlier exercise:

3, 4, 5, 6, 6, 7, 8, 9

What if we had the following numbers instead?

1003, 1004, 1005, 1006, 1006, 1007, 1008, 1009

Explain why any reasonable measure of "spread" should be the same for both groups of numbers.

One way to measure how large or small a number is relative to the whole set is to measure the distance of each number to the mean.

Recall that the mean of the following numbers is 6:

3, 4, 5, 6, 6, 7, 8, 9

Create a new list of eight numbers that measures the distance between each of the above numbers and the mean. In other words, subtract 6 from each of the above numbers.

Some of the numbers in your new list should be negative, some should be zero, and some should be positive. Why does that make sense? In other words, what does it mean when a number is negative, zero, or positive?

If the original set of numbers is called X, then what you've just calculated is a new list $\left(X - \overline{X}\right)$. Let's start organizing this into a table:

X	$\left(X - \overline{X}\right)$
3	-3
4	-2
5	-1

X	$(X - \overline{X})$
6	0
6	0
7	1
8	2
9	3

The numbers in the second columns are "deviations" from the mean.

One way you might measure "spread" is to look at the average deviation. After all, if the deviations represent the distances to the mean, a set with large spread will have large deviations and a set with small spread will have small deviations.

Go ahead and take the average (mean) of the numbers in the second column above.

Uh, oh! You should have calculated zero. Explain why you will always get zero, no matter what set of numbers you start with.

The idea of the "average deviation" seems like it should work, but it clearly doesn't. How do we fix the idea?

Hopefully, you identified that having negative deviations was a problem because they canceled out the positive deviations. But if all the deviations were positive, that wouldn't be an issue any more.

There are two ways of making numbers positive:

- Taking absolute values

We could just take the absolute value and make all the values positive. There are some statistical procures that do just that,[1] but we're going to take a slightly different approach…

- Squaring

If we square each value, they all become positive.

Taking the absolute value is conceptually easier, but there are some historical and mathematical reasons why squaring is a little better.[2]

Square each of the numbers from the second column of the table above. This will calculate a new list $\left(X - \overline{X}\right)^2$

[1] This leads to the "mean absolute deviation" or MAD.

[2] If you know calculus, you might think why the square function is much better behaved than the absolute value function.

Putting the new numbers into our previous table:

X	$(X - \overline{X})$	$(X - \overline{X})^2$
3	-3	9
4	-2	4
5	-1	1
6	0	0
6	0	0
7	1	1
8	2	4
9	3	9

Now take the average (mean) of the numbers in the third column above.

The number you got (should be 3.5) is *almost* what we call the variance. There's only one more annoying wrinkle.

When you took the mean of the last column of numbers, you added them all up and divided by 8 since there are 8 numbers in the list. But for some fairly technical mathematical reasons, we actually don't want to divide by 8. Instead, we divide by one less than that number; in other words, we divide by 7.[3]

Re-do the math above, but divide by 7 instead of dividing by 8.

The number you found is the *variance*, written as $Var(X)$. The full formula is

$$Var(X) = \frac{\sum (X - \overline{X})^2}{n - 1}$$

As a one-liner, the formula may look a little intimidating, but when you break it down step by step as we did above, it's not so bad.

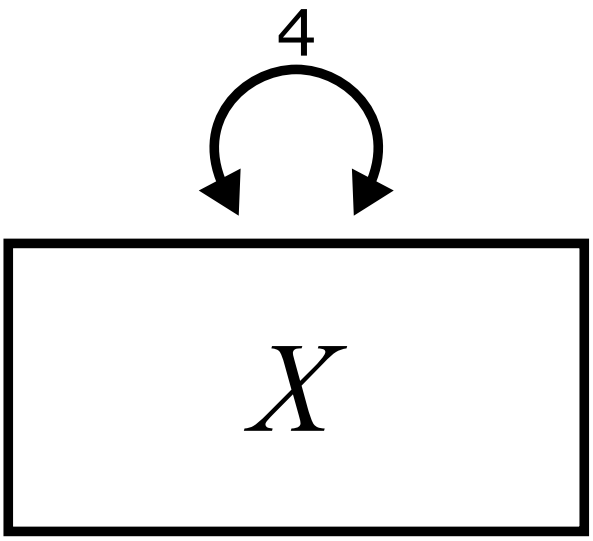
Here is the full calculation in the table:

X	$(X - \overline{X})$	$(X - \overline{X})^2$
3	-3	9
4	-2	4
5	-1	1
6	0	0
6	0	0
7	1	1
8	2	4
9	3	9

[3] For more information on that, search the internet for "sample variance unbiased"

X	$(X - \overline{X})$	$(X - \overline{X})^2$
		Sum: 28
		Variance: $28/7 = \boxed{4}$

In our diagrams, the variance of a variable is indicated by a curved, double-headed arrow, labeled with the value of the variance, like this:



Using the tabular approach, calculate the variance of the following set of numbers:

4, 3, 7, 2, 9, 4, 6

Consider the following two sets of numbers:

A) 1, 2, 5, 8, 9

B) 1, 4, 5, 6, 9

Without doing any calculations, which of the sets has the larger variance?

Once you've decided, then calculate the variance for both sets and check your answer.

2.3 Calculating variance in R

Once we've done it by hand a few times to make sure we understand how the formula works, from here on out we can let R do the work for us:

```
X1 <- c(3, 4, 5, 6, 6, 7, 8, 9)
var(X1)
```

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

```
X2 <- c(4, 3, 7, 2, 9, 4, 6)
var(X2)
```

```
## [1] 6
```

```
X3 <- c(1, 2, 5, 8, 9)
var(X3)
```

```
## [1] 12.5
```

```
X4 <- c(1, 4, 5, 6, 9)
var(X4)
```

```
## [1] 8.5
```

This is also easier for real-world data that is not highly engineered to produce whole numbers:

```
PlantGrowth$weight
```

```
##  [1] 4.17 5.58 5.18 6.11 4.50 4.61 5.17 4.53 5.33 5.14 4.81 4.17 4.41 3.59 5.87
## [16] 3.83 6.03 4.89 4.32 4.69 6.31 5.12 5.54 5.50 5.37 5.29 4.92 6.15 5.80 5.26
```

```
var(PlantGrowth$weight)
```

```
## [1] 0.49167
```

2.4 Variance rules

In this course, we will need to be able to calculate the variance of various combinations of variables. For example, if X_1 and X_2 are two variables, we can create a new variable $X_1 + X_2$ by adding up the values of the two variables. What is the variance of $X_1 + X_2$?

But before we answer that, let's establish the first rule.

- **Rule 1**

Suppose that C is a "constant" variable, meaning that it always has the same value (rather than being a variable that could contain lots of different numbers). Then,

$$Var\left(C\right) = 0$$

Why is **Rule 1** true? You can either reason through this conceptually, based on how you understand what variance is supposed to measure, or you can do a sample calculation. (Make a table starting with a column that contains many copies of only a single number and work through the calculation.)

Now, back to the example at the beginning of the section of finding the variance of $X_1 + X_2$.

- **Rule 2**

If X_1 and X_2 are independent, then

$$Var\left(X_1 + X_2\right) = Var\left(X_1\right) + Var\left(X_2\right)$$

We're not going to get into a formal definition of *independence* here. For now, it suffices to think of the intuitive definition you may already have in your head of what it means for two things to be independent. The idea is that, to be independent, X_1 and X_2 should have nothing to do with each other. Knowing the values of one should not give you any information about values of the other. In the next chapter [LINK], we'll say more about this rule.

It's important to note that **Rule 2** is an abstract mathematical rule that holds *in theory*. When we have actual data, however, we know that statistics won't always match their theoretical values. For example, even if a true population mean is 42, samples drawn from that population will have sample means that are *close* to 42, but likely not exactly 42.[4]

Let's test this out. Below are two new variables that are defined using random numbers. The first one is normally distributed with mean 1 and standard deviation 2. (If you don't remember standard deviation from intro stats, we talk about it in the next section.) The second one is normally distributed with mean 4 and standard deviation 3. [WHERE DOES SEED INFO GO?] These are independent because the definition of X_5 does not depend on any way on the definition of X_6 and vice versa.

The sample sizes (2000) are large enough that we should get pretty close to the theoretically correct results here.

[4]The exact distribution of sample means around a true population value is something you probably learned about in an intro stats course. Sample means follow a Student t distribution.

```
set.seed(10101)
X5 <- rnorm(2000, mean = 1, sd = 2)
X6 <- rnorm(2000, mean = 4, sd = 3)
```

```
head(X5)
```

```
## [1] -0.7535339 -0.4927789  3.7518296  1.4751639  1.2172549  3.4054426
```

```
head(X6)
```

```
## [1] 2.297279 4.856377 6.661822 1.309892 2.270882 3.827944
```

Use R to calculate the variance of X_5 and X_6 separately. Then use R to add the two numbers you just obtained (the sum of the two variances). Finally, use R to calculate the variance of the sum of the two variables.

Here's an example to help think about this intuitively.

Suppose someone comes along and offers to give you a random amount of money, some number between \$0 and \$100.[5] If the variance is a measure of spread, then it stands to reason that variance reflects something about how uncertain you are about how much money you will have after this transaction. On average, you expect about \$50, but you know that the actual amount of money you will receive can vary greatly.

Okay, now a second person comes along and offers you the same deal, a random dollar gift between \$0 and \$100.[6] At the end of both transactions, how much money will you have? On average, maybe about \$100, but what about your uncertainty? Because the total amount is the result of two random gifts, you are even less sure how close to \$100 you might be. The range of possible values is now \$0 to \$200.[7] Your uncertainty is greater overall.

Of course, all this explains is why the variance of the sum of two variables is larger than the variance of either variable individually. The fact that the variance of the sum of two independent variables is *exactly* the sum of the variances has to be shown mathematically. But hopefully, the intuition is clear.

The next rule is a consequence of the first two rules, so we will not give it a special number

$$Var\left(X + C\right) = Var\left(X\right)$$

[5] To be more concrete, the values are uniformly distributed, meaning that any number between 0 and 100 is equally likely.

[6] Apparently you live in a town with very generous strangers.

[7] To be clear, though, the probabilities are no longer uniform between 0 and 200. To get near 0, you would have to be unlucky twice, and to get near 200 you would have to get lucky twice.

Can you apply **Rule 2** followed by **Rule 1** to see mathematically why $Var(X + C) = Var(X)$?

What is the intuition behind the statement $Var(X + C) = Var(X)$? In other words, can you explain the rule to someone in terms of what it means about shifting the values of a data set up or down by a constant amount?

Rule 3 is similar to **Rule 2**, but it's quite counter-intuitive:

- **Rule 3**

If X_1 and X_2 are independent, then

$$Var(X_1 - X_2) = Var(X_1) + Var(X_2)$$

It is very common for students to think that a minus sign on the left would translate into a minus sign on the right.[8]

What gives?

Let's return to our example of strangers giving you money.[9] The first person still offers you a random amount between \$0 and \$100. But, now, the second person is a robber, and forces you to give them a random dollar value between \$0 and \$100 (of their choosing, of course). How much money do you expect to have after these two events? On average, \$0. (The first person gives you, on average, \$50, and the second person takes away, on average, \$50.) But how certain are you about that amount?

Imagine a world in which the wrong rule prevailed. What if $Var(X_1 - X_2)$ were truly the difference of the two variances. But $Var(X_1)$ and $Var(X_2)$ are the same in this scenario. (Although one person is giving money and one is taking, our uncertainty about the dollar amount is the same in both cases.) And this implies

$$Var(X_1) - Var(X_2) = 0$$

Can this be true? Zero variance means "no spread" which means exact certainty of the value. (Remember **Rule 1**?) Are you 100% confident that you will end both transactions with exactly \$0? No way!

In fact, the amount of money you end up with ranges from -\$100 up to \$100. This is a larger range than in either transaction individually. Our uncertainty has grown because there are two random processes in play, just like in the scenario with two beneficent strangers. In fact, the width of the range of possibilities is the same in both scenarios: \$0 to \$200 and -\$100 to \$100 both span a range of \$200.

[8] This results from many years of developing a Pavlovian response to anything that looks like the distributive law from algebra.

[9] Actually, that sounds a little creepy when put like that.

The next rule, unfortunately, does not have a great intuitive explanation. It will make a little more sense in the next chapter [LINK], and we'll revisit it then.

- **Rule 4**

If a is any number,

$$Var\left(aX\right) = a^2 Var\left(X\right)$$

If you go back to the table, imagine multiplying every number in the first column by a. Every number in the second column will still have a factor of a. But when you square those values, every number in the third column will have a factor of a^2. That's the gist of the rule anyway. But, again, there's not much intuition about why that makes sense.

We can, at least, check empirically that the rule works.

We'll use X_5 as we defined it above, a normally distributed variable with mean 1 and standard deviation 2. The variance of the data is about 4:

```
var(X5)
```

```
## [1] 4.15763
```

Let's use $a = 3$.

In R, calculate $Var\left(3X_5\right)$. (Don't forget that in R, you can't just type `3 X5`. You have to explicitly include the multiplication sign: `3 * X5`.)

Now try calculating $3Var\left(X_5\right)$. You'll see that you don't get the right answer.

But now try $9Var\left(X_5\right)$. That should work.

And that's all the variance rules we'll need!

2.5 Standard deviation

The variance is nice because it obeys all the above rules. The one big downside is that it's not very interpretable.

For example, think of the scenario with people giving/taking money. In that case, the values were measures in units of dollars.

If X is measured in dollars, what are the units of measurement of $\overline{X}$? That seems sensible, right?

What are the units of $\left(X - \overline{X}\right)$? Still sensible, right? (It's not a problem that some of these values will be positive and other negative. Negative dollars still make sense. Just think about your student loans.)

Okay, now here's where things get weird. What are the units of $\left(X - \overline{X}\right)^2$? This no longer makes sense.

Variance is *nearly* the average of a bunch of squared deviations, so for a variable measured in dollars, the units of variance would be "squared dollars", whatever that is.

Variances are not really interpretable directly. How do we make them more interpretable? Well, if variance has "squared" units, we can take the square root to get back to the natural units we started with.

And this is called the standard deviation, $SD(X)$.

$$SD(X) = \sqrt{\frac{\sum \left(X - \overline{X}\right)^2}{n-1}}$$

Or, said more simply,

$$SD(X) = \sqrt{Var(X)}$$

Equivalently,

$$Var(X) = SD(X)^2$$

Due to its interpretability, an intro stats class will focus far more on the standard deviation than on the variance. The downside is that the mathematical rules aren't so nice for standard deviations. For example, what is

$$SD\left(X_1 + X_2\right)?$$

You can work through the definition to see that

$$SD\left(X_1 + X_2\right) = \sqrt{SD\left(X_1\right)^2 + SD\left(X_2\right)^2}$$

But, eww, that's gross.

For SEM, we will focus almost exclusively on variance and switch to standard deviation for only two reasons:

1. We need to communicate something about spread in meaningful units.
2. We need to standardize variables. (See Section 2.7 below.)

Although the standard deviance is just the square root of the variance, it is worth knowing the R command to calculate it. It's just `sd`. For example:

```
sd(PlantGrowth$weight)
```

```
## [1] 0.7011918
```

You can see below that `sd` did the right thing:

```
sqrt(var(PlantGrowth$weight))
```

```
## [1] 0.7011918
```

2.6 Mean centering data

Many of the statistical techniques taught in an intro stats course focus on learning about the means of variables. Structural equation modeling is a little different in that it is more focused on the explaining the variability of data—how changes in one or more variables predict changes in other variables.[10]

A habit we'll start forming now is to mean center all our variables. We do this by subtracting the mean of a variable from all its values.

Let's use X_6 as we defined it before, a normally distributed variable with mean 4 and standard deviation 3. How do we interpret the values of $X_6 - \overline{X_6}$? (Remember, this is just the second column in our variance tables earlier.)

If we shift all the X_6 values to the left by $\overline{X_6}$ units, what is the mean of the new list of numbers?

Let's verify this in R. We'll use the "suffix" `mc` to indicate a mean-centered variable.

```
X6_mc <- X6 - mean(X6)
mean(X6_mc)
```

```
## [1] 2.851573e-16
```

Why does this answer not exactly agree with the "theoretical" answer you came up with in a few lines above? (If you don't already know, the `e-16` in the expression above is scientific notation and means "times 10^{-16}. That's a really small number!)

[10] There are tools in SEM for working with means as well. WILL WE COVER THIS IN A FUTURE CHAPTER?

Take a guess about the variance of `X6_mc`. Verify your guess in R.

So the good news is that **mean centering preserves the variance**. While the mean will be shifted to be 0, the variance does not change, so any statistical model we build that analyzes the variance will not be affecting by mean-centering.

2.7 Standardizing data

After we've mean centered the data, we can go one step further and divide by the standard deviation. This results in something often called a *z-score*. The process of converting variables from their original units to z-scores is called *standardizing* the data.

$$Z = \frac{(X - \overline{X})}{SD(X)}$$

Why is this useful? One reason is that it remove the units of measurement to facilitate comparisons between variables. Suppose X represents height in inches. The numerator $(X - \overline{X})$ has units of inches. The standard deviation $SD(X)$ also has units of inches. So when you divide, the units go away and the z-score is left without units, sometimes called a "dimensionless quantity".

Suppose a female in the United States is 6 feet tall (72 inches). Suppose a female in China is 5'8 tall (68 inches). In absolute terms, the American woman is taller than the Chinese woman. But what if we're interested in knowing which woman is taller *relative* to their respective population?

The mean height for an American woman is 65" with a standard deviation of 3.5" The mean height for a Chinese woman is 62" with a standard deviation of 2.5". (These numbers aren't perfectly correct, but they're probably close-ish.)

Calculate the z-scores for both these women.

Which woman is taller relative to their population?

Although z-scores don't technically have units, we can think of them as measuring how many standard deviations a value lies above or below the mean.

What is the z-score for a value that equals the mean?

What is the meaning of a negative z-score?

The z-score for the American woman was 2. This means that her height measures two standard deviations above the mean.

For real-world data, we will use technology to do this. Here are some temperature measurements from New York in 1974. (These are daily highs across a six-month period.)

```
airquality$Temp
```

```
##   [1] 67 72 74 62 56 66 65 59 61 69 74 69 66 68 58 64 66 57 68 62 59 73 61 61 57
##  [26] 58 57 67 81 79 76 78 74 67 84 85 79 82 87 90 87 93 92 82 80 79 77 72 65 73
##  [51] 76 77 76 76 76 75 78 73 80 77 83 84 85 81 84 83 83 88 92 92 89 82 73 81 91
##  [76] 80 81 82 84 87 85 74 81 82 86 85 82 86 88 86 83 81 81 81 82 86 85 87 89 90
## [101] 90 92 86 86 82 80 79 77 79 76 78 78 77 72 75 79 81 86 88 97 94 96 94 91 92
## [126] 93 93 87 84 80 78 75 73 81 76 77 71 71 78 67 76 68 82 64 71 81 69 63 70 77
## [151] 75 76 68
```

We calculate the mean and standard deviation:

```
mean(airquality$Temp)
```

```
## [1] 77.88235
```

```
sd(airquality$Temp)
```

```
## [1] 9.46527
```

This is an average high of about 78 degrees Fahrenheit with a standard deviation of about 9.5 degrees Fahrenheit.

If we just subtract the mean, we get mean-centered data.

```
airquality$Temp - mean(airquality$Temp)
```

```
##   [1] -10.8823529  -5.8823529  -3.8823529 -15.8823529 -21.8823529 -11.8823529
##   [7] -12.8823529 -18.8823529 -16.8823529  -8.8823529  -3.8823529  -8.8823529
##  [13] -11.8823529  -9.8823529 -19.8823529 -13.8823529 -11.8823529 -20.8823529
##  [19]  -9.8823529 -15.8823529 -18.8823529  -4.8823529 -16.8823529 -16.8823529
##  [25] -20.8823529 -19.8823529 -20.8823529 -10.8823529   3.1176471   1.1176471
##  [31]  -1.8823529   0.1176471  -3.8823529 -10.8823529   6.1176471   7.1176471
##  [37]   1.1176471   4.1176471   9.1176471  12.1176471   9.1176471  15.1176471
##  [43]  14.1176471   4.1176471   2.1176471   1.1176471  -0.8823529  -5.8823529
##  [49] -12.8823529  -4.8823529  -1.8823529  -0.8823529  -1.8823529  -1.8823529
##  [55]  -1.8823529  -2.8823529   0.1176471  -4.8823529   2.1176471  -0.8823529
##  [61]   5.1176471   6.1176471   7.1176471   3.1176471   6.1176471   5.1176471
##  [67]   5.1176471  10.1176471  14.1176471  14.1176471  11.1176471   4.1176471
##  [73]  -4.8823529   3.1176471  13.1176471   2.1176471   3.1176471   4.1176471
##  [79]   6.1176471   9.1176471   7.1176471  -3.8823529   3.1176471   4.1176471
##  [85]   8.1176471   7.1176471   4.1176471   8.1176471  10.1176471   8.1176471
```

```
##  [91]   5.1176471   3.1176471   3.1176471   3.1176471   4.1176471   8.1176471
##  [97]   7.1176471   9.1176471  11.1176471  12.1176471  12.1176471  14.1176471
## [103]   8.1176471   8.1176471   4.1176471   2.1176471   1.1176471  -0.8823529
## [109]   1.1176471  -1.8823529   0.1176471   0.1176471  -0.8823529  -5.8823529
## [115]  -2.8823529   1.1176471   3.1176471   8.1176471  10.1176471  19.1176471
## [121]  16.1176471  18.1176471  16.1176471  13.1176471  14.1176471  15.1176471
## [127]  15.1176471   9.1176471   6.1176471   2.1176471   0.1176471  -2.8823529
## [133]  -4.8823529   3.1176471  -1.8823529  -0.8823529  -6.8823529  -6.8823529
## [139]   0.1176471 -10.8823529  -1.8823529  -9.8823529   4.1176471 -13.8823529
## [145]  -6.8823529   3.1176471  -8.8823529 -14.8823529  -7.8823529  -0.8823529
## [151]  -2.8823529  -1.8823529  -9.8823529
```

But if we also divide by the standard deviation, we get a standardized variable (or a set of z-scores). Note the extra parentheses to make sure we get the order of operations right. We have to subtract first, but then divide that whole mean-centered quantity by the standard deviation.

```
(airquality$Temp - mean(airquality$Temp))/sd(airquality$Temp)
```

```
##   [1] -1.14971398 -0.62146702 -0.41016823 -1.67796094 -2.31185730 -1.25536337
##   [7] -1.36101276 -1.99490912 -1.78361034 -0.93841519 -0.41016823 -0.93841519
##  [13] -1.25536337 -1.04406459 -2.10055851 -1.46666216 -1.25536337 -2.20620791
##  [19] -1.04406459 -1.67796094 -1.99490912 -0.51581762 -1.78361034 -1.78361034
##  [25] -2.20620791 -2.10055851 -2.20620791 -1.14971398  0.32937752  0.11807873
##  [31] -0.19886945  0.01242934 -0.41016823 -1.14971398  0.64632570  0.75197509
##  [37]  0.11807873  0.43502691  0.96327387  1.28022205  0.96327387  1.59717023
##  [43]  1.49152084  0.43502691  0.22372813  0.11807873 -0.09322005 -0.62146702
##  [49] -1.36101276 -0.51581762 -0.19886945 -0.09322005 -0.19886945 -0.19886945
##  [55] -0.19886945 -0.30451884  0.01242934 -0.51581762  0.22372813 -0.09322005
##  [61]  0.54067630  0.64632570  0.75197509  0.32937752  0.64632570  0.54067630
##  [67]  0.54067630  1.06892327  1.49152084  1.49152084  1.17457266  0.43502691
##  [73] -0.51581762  0.32937752  1.38587145  0.22372813  0.32937752  0.43502691
##  [79]  0.64632570  0.96327387  0.75197509 -0.41016823  0.32937752  0.43502691
##  [85]  0.85762448  0.75197509  0.43502691  0.85762448  1.06892327  0.85762448
##  [91]  0.54067630  0.32937752  0.32937752  0.32937752  0.43502691  0.85762448
##  [97]  0.75197509  0.96327387  1.17457266  1.28022205  1.28022205  1.49152084
## [103]  0.85762448  0.85762448  0.43502691  0.22372813  0.11807873 -0.09322005
## [109]  0.11807873 -0.19886945  0.01242934  0.01242934 -0.09322005 -0.62146702
## [115] -0.30451884  0.11807873  0.32937752  0.85762448  1.06892327  2.01976780
## [121]  1.70281962  1.91411841  1.70281962  1.38587145  1.49152084  1.59717023
## [127]  1.59717023  0.96327387  0.64632570  0.22372813  0.01242934 -0.30451884
## [133] -0.51581762  0.32937752 -0.19886945 -0.09322005 -0.72711641 -0.72711641
## [139]  0.01242934 -1.14971398 -0.19886945 -1.04406459  0.43502691 -1.46666216
## [145] -0.72711641  0.32937752 -0.93841519 -1.57231155 -0.83276580 -0.09322005
## [151] -0.30451884 -0.19886945 -1.04406459
```

The easier way to do this in R is to use the `scale` command. (Sorry, the output is a little long. Keep scrolling below.)

```
scale(airquality$Temp)
```

```
##               [,1]
##   [1,] -1.14971398
##   [2,] -0.62146702
##   [3,] -0.41016823
##   [4,] -1.67796094
##   [5,] -2.31185730
##   [6,] -1.25536337
##   [7,] -1.36101276
##   [8,] -1.99490912
##   [9,] -1.78361034
##  [10,] -0.93841519
##  [11,] -0.41016823
##  [12,] -0.93841519
##  [13,] -1.25536337
##  [14,] -1.04406459
##  [15,] -2.10055851
##  [16,] -1.46666216
##  [17,] -1.25536337
##  [18,] -2.20620791
##  [19,] -1.04406459
##  [20,] -1.67796094
##  [21,] -1.99490912
##  [22,] -0.51581762
##  [23,] -1.78361034
##  [24,] -1.78361034
##  [25,] -2.20620791
##  [26,] -2.10055851
##  [27,] -2.20620791
##  [28,] -1.14971398
##  [29,]  0.32937752
##  [30,]  0.11807873
##  [31,] -0.19886945
##  [32,]  0.01242934
##  [33,] -0.41016823
##  [34,] -1.14971398
##  [35,]  0.64632570
##  [36,]  0.75197509
##  [37,]  0.11807873
##  [38,]  0.43502691
##  [39,]  0.96327387
```

```
##  [40,]  1.28022205
##  [41,]  0.96327387
##  [42,]  1.59717023
##  [43,]  1.49152084
##  [44,]  0.43502691
##  [45,]  0.22372813
##  [46,]  0.11807873
##  [47,] -0.09322005
##  [48,] -0.62146702
##  [49,] -1.36101276
##  [50,] -0.51581762
##  [51,] -0.19886945
##  [52,] -0.09322005
##  [53,] -0.19886945
##  [54,] -0.19886945
##  [55,] -0.19886945
##  [56,] -0.30451884
##  [57,]  0.01242934
##  [58,] -0.51581762
##  [59,]  0.22372813
##  [60,] -0.09322005
##  [61,]  0.54067630
##  [62,]  0.64632570
##  [63,]  0.75197509
##  [64,]  0.32937752
##  [65,]  0.64632570
##  [66,]  0.54067630
##  [67,]  0.54067630
##  [68,]  1.06892327
##  [69,]  1.49152084
##  [70,]  1.49152084
##  [71,]  1.17457266
##  [72,]  0.43502691
##  [73,] -0.51581762
##  [74,]  0.32937752
##  [75,]  1.38587145
##  [76,]  0.22372813
##  [77,]  0.32937752
##  [78,]  0.43502691
##  [79,]  0.64632570
##  [80,]  0.96327387
##  [81,]  0.75197509
##  [82,] -0.41016823
##  [83,]  0.32937752
##  [84,]  0.43502691
##  [85,]  0.85762448
```

```
##  [86,]  0.75197509
##  [87,]  0.43502691
##  [88,]  0.85762448
##  [89,]  1.06892327
##  [90,]  0.85762448
##  [91,]  0.54067630
##  [92,]  0.32937752
##  [93,]  0.32937752
##  [94,]  0.32937752
##  [95,]  0.43502691
##  [96,]  0.85762448
##  [97,]  0.75197509
##  [98,]  0.96327387
##  [99,]  1.17457266
## [100,]  1.28022205
## [101,]  1.28022205
## [102,]  1.49152084
## [103,]  0.85762448
## [104,]  0.85762448
## [105,]  0.43502691
## [106,]  0.22372813
## [107,]  0.11807873
## [108,] -0.09322005
## [109,]  0.11807873
## [110,] -0.19886945
## [111,]  0.01242934
## [112,]  0.01242934
## [113,] -0.09322005
## [114,] -0.62146702
## [115,] -0.30451884
## [116,]  0.11807873
## [117,]  0.32937752
## [118,]  0.85762448
## [119,]  1.06892327
## [120,]  2.01976780
## [121,]  1.70281962
## [122,]  1.91411841
## [123,]  1.70281962
## [124,]  1.38587145
## [125,]  1.49152084
## [126,]  1.59717023
## [127,]  1.59717023
## [128,]  0.96327387
## [129,]  0.64632570
## [130,]  0.22372813
## [131,]  0.01242934
```

```
## [132,] -0.30451884
## [133,] -0.51581762
## [134,]  0.32937752
## [135,] -0.19886945
## [136,] -0.09322005
## [137,] -0.72711641
## [138,] -0.72711641
## [139,]  0.01242934
## [140,] -1.14971398
## [141,] -0.19886945
## [142,] -1.04406459
## [143,]  0.43502691
## [144,] -1.46666216
## [145,] -0.72711641
## [146,]  0.32937752
## [147,] -0.93841519
## [148,] -1.57231155
## [149,] -0.83276580
## [150,] -0.09322005
## [151,] -0.30451884
## [152,] -0.19886945
## [153,] -1.04406459
## attr(,"scaled:center")
## [1] 77.88235
## attr(,"scaled:scale")
## [1] 9.46527
```

Although the outputs are formatted a little differently, you can go back and check that these sets of numbers match each other.

What is the mean of a standardized variable? How do you know this?

Let's calculate the variance of a standardized variable. To do so, I'll note that the mean $\overline{X}$ is just a number. Also, the standard deviation $SD(X)$ is just a number. To make the calculation easier to understand, let's just substitute letters that are easier to work with:

$M = \overline{X}$

$S = SD(X)$

Remember, M and S are *constants*.

Now we need to calculate $Var(Z)$. I'll do the first couple of steps. Then you take over and, using the variance rules from earlier in the chapter, simplify the expression until you get to a numerical answer. Be sure to justify each step by citing the rule you invoked to get there.

$$Var(Z) = Var\left(\frac{(X-\overline{X})}{SD(X)}\right) \tag{2.1}$$

$$= Var\left(\frac{(X-M)}{S}\right) \tag{2.2}$$

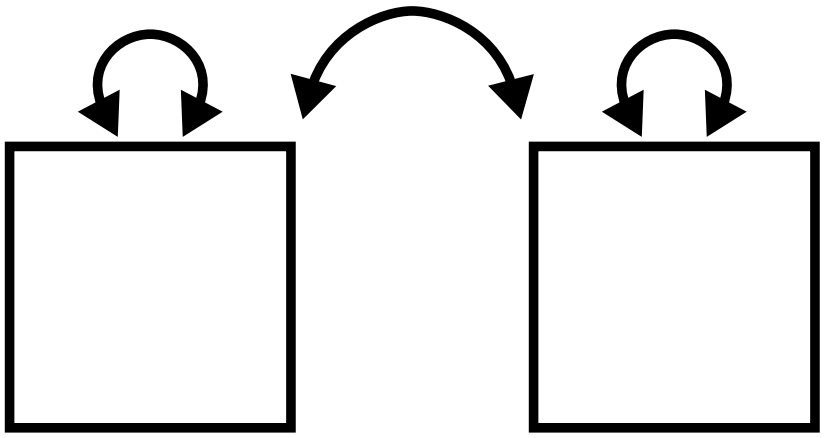
$$= Var\left(\frac{1}{S}(X-M)\right) \tag{2.3}$$

$$= \quad ??? \tag{2.4}$$

You should get the answer 1. A standardized variable always has variance 1. This will be an important fact in future chapters.

Chapter 3

Covariance



3.1 Calculating covariance

The last chapter was about variance, which measures the spread of a single variable. Now we extend this idea to pairs of variables.

We say that two variables "co-vary" when the spread of one variable is related to the spread of another variable. This relationship represents an *association* between the two variables.

We'll call our two variables X_1 and X_2. To keep things simple, let's assume that we have already mean centered our variables.

If X_1 and X_2 are already mean centered, then what are $\overline{X_1}$ and $\overline{X_2}$?

As we did in the last chapter with variance, we'll build up the calculation of

covariance step-by-step using a table to keep track of intermediate quantities we need.

Here are two variables (with $n = 7$) that have been mean centered:

X_1	X_2
-1	-2
-2	2
2	-2
-3	-1
4	2
-1	-2
1	3

Check that the mean of both columns is truly zero.

Something interesting happens when we look at the product X_1X_2.

If X_1 and X_2 both lie above their means, they are both positive numbers. Therefore, their product is positive.

What if both X_1 and X_2 lie below their means? What do we know about their values individually and what do we know about their product?

Here is the chart again, but with the products listed in a new column:

X_1	X_2	X_1X_2
-1	-2	2
-2	2	-4
2	-2	-4
-3	-1	3
4	2	8
-1	-2	2
1	3	3

Now we add up the products across all seven data pairs:

X_1	X_2	X_1X_2
-1	-2	2
-2	2	-4
2	-2	-4
-3	-1	3
4	2	8
-1	-2	2

X_1	X_2	X_1X_2
1	3	3
		Sum: 10

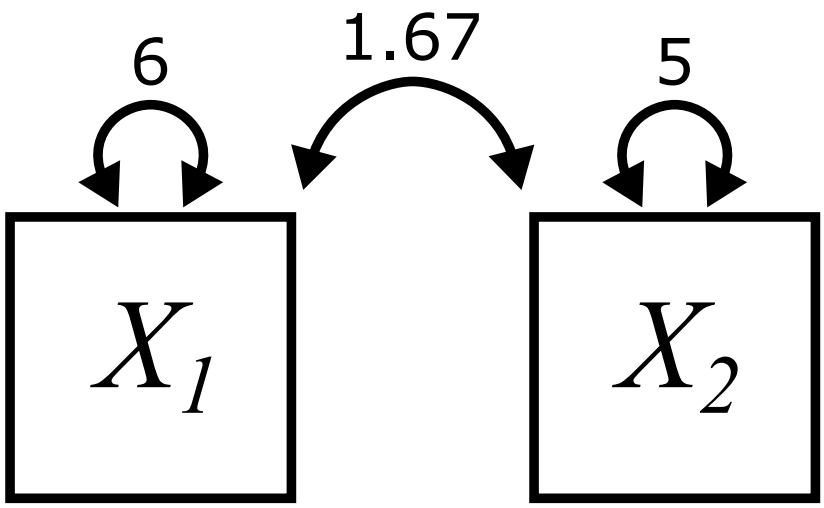
So when X_1 and X_2 tend to have similar values (both positive or both negative), their product is usually positive. It's not true of every pair of values in the table above; some products are negative. But the majority are positive. Therefore, the sum of all such products will be positive.

We're almost there. Just like we wanted the average squared deviation to calculate the variance, here we want the average of the products from the third column above. And just like in the case of variance, it's not *quite* the average we calculate. Instead of dividing by n, we divide by $n-1$ for exactly the same esoteric reason. In our example, there are 7 data points (in other words, 7 rows of data), so we divide by 6.

Putting this all together:

X_1	X_2	X_1X_2
-1	-2	2
-2	2	-4
2	-2	-4
-3	-1	3
4	2	8
-1	-2	2
1	3	3
		Sum: 10
		Covariance: 10/6 = $\boxed{1.67}$

In our diagrams, the covariance of two variables is indicated by a curved, double-headed arrow pointing at both boxes and labeled with the value of the covariance, like this:



Note that we still include the variances of each of the individual variables. They are still important to us. We just have one new type of arrow now.

Verify that the variances in the diagram are correct for our example. You can do it by hand if you want, but using R is fine too.

Here is the final formula for covariance, written as $Cov(X_1, X_2)$. This works for all pairs of variables, even if they aren't mean centered. The terms $(X_1 - \overline{X_1})$ and $(X_2 - \overline{X_2})$ do the mean centering:

$$Cov(X_1, X_2) = \frac{\sum (X_1 - \overline{X_1})(X_2 - \overline{X_2})}{n-1}$$

Suppose X_1 tends to be above its mean when X_2 is below its mean and X_1 tends to be below its mean when X_2 is above its mean. What will the product $(X_1 - \overline{X_1})(X_2 - \overline{X_2})$ usually be? Therefore, what will the sum of all such products likely be?

For general variables (not necessarily mean centered), the table will actually look like this:

X_1	X_2	$(X_1 - \overline{X_1})$	$(X_2 - \overline{X_2})$	$(X_1 - \overline{X_1})(X_2 - \overline{X_2})$
17	23	-2	11	-22
25	15	6	3	18
...	...	...	...	...

Calculate the covariance by hand by making a table like the one above. (These variables are *not* mean centered, so you'll have to calculate the mean of each variable in order to fill out the third and fourth columns.)

X_3: 8, 10, 16, 7, 4, 3

X_4: 6, 5, 4, 9, 11, 7

Explain intuitively why the covariance is negative for these two variables.

When calculating variance, the order of the data points does not matter. Why?

When calculating covariance, the order of the data points *does* matter. Why?

What if you keep pairs together, but rearrange the rows of the table. How does that affect the covariance?

3.2 Calculating covariance in R

Once we've done it by hand a few times to make sure we understand how the formula works, from here on out we can let R do the work for us:

```
X1 <- c(-1,-2, 2, -3, 4, -1, 1)
X2 <- c(-2, 2, -2, -1, 2, -2, 3)
cov(X1, X2)
```

```
## [1] 1.666667
```

```
X3 <- c(8, 10, 16, 7, 4, 3)
X4 <- c(6, 5, 4, 9, 11, 7)
cov(X3, X4)
```

```
## [1] -9.2
```

And here's some real world data:

```
cov(airquality$Temp, airquality$Wind)
```

```
## [1] -15.27214
```

3.3 Covariance rules

We'll think of the variance and covariance rules as one big list. We left off on **Rule 4**, so now we'll introduce **Rule 5**.

- **Rule 5**

$$Cov(X, X) = Var(X)$$

In words, **Rule 5** states that the covariance of a variable *with itself* is just the same thing as the variance of that variable. This is quite remarkable! It means that variance is really just a special case of covariance.

Explain why **Rule 5** is true. (Hint: think about how you would calculate $Cov(X, X)$ using either the formula or the table—or both!)

- **Rule 6**

$$Cov\left(X_1, X_2\right) = Cov\left(X_2, X_1\right)$$

In words, we would say that covariance is *symmetric*.

Explain why **Rule 6** is true. (Again, think about the formula or the table—or both!)

The next four rules are analogous to similar rules for variance (**Rule 1**, **Rule 2**, **Rule 3**, and **Rule 4**).

- **Rule 7**

Suppose that C is a "constant" variable, meaning that it always has the same value (rather than being a variable that could contain lots of different numbers). Then,

$$Cov\left(X, C\right) = 0$$

As always, try to explain this rule. Give an intuitive explanation of why this rule "should" be true. Then think about it computationally, thinking of either the formula or the table—or both!

- **Rule 8**

$$Cov\left(X_1 + X_2, X_3\right) = Cov\left(X_1, X_3\right) + Cov\left(X_2, X_3\right)$$

What you should appreciate here is that there is no longer any restriction on the relationships among the variables involved. **Rule 2** only worked when the two variables were independent. On the other hand, **Rule 8** works for any combination of variables, no matter their relation.

Even more satisfying is this next rule:

- **Rule 9**

$$Cov\left(X_1 - X_2, X_3\right) = Cov\left(X_1, X_3\right) - Cov\left(X_2, X_3\right)$$

Yay! The minus sign behaves sensibly now! Of course, since covariances can be positive or negative (unlike variances which are always positive!) we can more safely subtract two of them without worry. And this rule, like **Rule 8**, does not depend on X_1 and X_2 being independent. They can be any two variables.

There are versions of these rules with the addition or subtraction on the other side, but these are just minor variations of **Rule 8** and **Rule 9**, so they're not worth mentioning as a separate rule. Remember that covariance is symmetric, so you can always swap things on the left and right of the comma.

$$Cov\left(X_1, X_2 \pm X_3\right) = Cov\left(X_1, X_2\right) \pm Cov\left(X_1, X_3\right)$$

- **Rule 10**

If a is any number,

$$Cov\left(aX_1, X_2\right) = aCov\left(X_1, X_2\right) = Cov\left(X_1, aX_2\right)$$

This rule is also very sensible. Instead of **Rule 4** that takes a number a and pulls out an a^2, **Rule 10** just pulls out a single factor of a (from either slot).

Just a couple more rules. We were talking about independence in conjunction with **Rule 8** and **Rule 9**. That leads directly to an interesting and super-important rule:

- **Rule 11**

If X_1 and X_2 are independent, then

$$Cov\left(X_1, X_2\right) = 0$$

Why is **Rule 11** true, intuitively?

It's interesting to note that this rule only works one way. In other words, if you know that two variables are independent, then you can conclude their covariance is zero. However, if you know the covariance is zero, that doesn't necessarily mean that the two variables are independent. We'll see an example of this later in the chapter.

Finally, one rule to rule them all:

- **Rule 12**

For *any* two variables X_1 and X_2:

$$Var(aX_1 + bX_2) = a^2Var(X_1) + b^2Var(X_2) + 2abCov(X_1, X_2)$$

This brings practically everything we know together into one rule!

Proving **Rule 12** will give us good practice with the type of manipulation we'll need to do in future chapters. So here goes. For the first few steps, you name what rule we're invoking. Then, you'll pick up the thread and follow it through the last few steps on your own.

$$\begin{aligned} Var(aX_1 + bX_2) &= Cov(aX_1 + bX_2, aX_1 + bX_2) && (3.1)\\ &= Cov(aX_1 + bX_2, aX_1) + Cov(aX_1 + bX_2, bX_2) && (3.2)\\ &= Cov(aX_1, aX_1) + Cov(bX_2, aX_1) + && (3.3)\\ &\quad Cov(aX_1, bX_2) + Cov(bX_2, bX_2) && (3.4)\\ &= \quad ??? && (3.5) \end{aligned}$$

You'll need these rules to do calculations in future chapters. Rather than having to search for them in Chapter 2 and this chapter, we've gathered up all the rules in one convenient place in Appendix A.

3.4 Correlation

The pros and cons for calculating covariance are similar to those for variance. The mathematics is much nicer for covariance, but we lose interpretability.

Let's suppose that X_1 measures salary in dollars and X_2 measures years of education. We would expect there to be some association between these variables, so we calculate the covariance. What is the unit of measurement of the resulting number?

The solution to the problem here is not as simple as it was for variance. Since variance had squared units, all we had to do was take the square root. Covariance has a weird product of units, so we have to do something more clever.

Following up on the activity above, let's suppose we have a covariance with units of "dollar-years". If we divide by a number expressed in dollars, we get rid of those units and we're left with years. But that seems unsatisfying; covariance should express something about both variables that went into it. Likewise, it makes no sense to divide by a number expressed in years as that would leave us just with dollars.

The solution to the dilemma is to accept that we aren't going to be able to keep any units in a meaningful way. Therefore, what we want is something *standardized*, meaning that it has no units.

If X_1 is expressed in dollars, can you think of a statistic that measures spread and is also in units of dollars?

Likewise, if X_2 is measured in years, what statistic that measures spread is also in units of years?

The previous activity gives us an idea. What if we divide the covariance by *both* the standard deviation of X_1 *and* the standard deviation of X_2?

$$\frac{Cov(X_1, X_2)}{SD(X_1)SD(X_2)}$$

Sometimes it's written like this:

$$\frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)}\sqrt{Var(X_2)}}$$

But that's the same thing, right?

This quantity has no units. We call this the *correlation* between X_1 and X_2. We'll either write

$$Corr(X_1, X_2)$$

or, if we need to be more concise,

$$r_{X_1X_2}$$

Yes, this is the same as the correlation coefficient you learned about in your intro stats class, although it wasn't likely presented to you quite this way.[1]

One great thing about correlation is that it has no units, so it serves as a sort of "universal" measure of how two variables co-vary. But the best part is that it has a nice intuitive meaning precisely because it factors out the pieces of the covariance that are only there because of the spread of the two variables individually. In other words, the fact that X_1 and X_2 have their own variability actually *complicates* the notion of covariance. Those individual variances "corrupt" the interpretation of covariance. But after excising them, all that's left in the correlation is the "pure" part of the covariance that expresses the relationship or association between X_1 and X_2.

[1] Karl Pearson is credited with inventing the correlation coefficient. Pearson was a life-long eugenicist and a proponent of using "science" to prove that some races were superior to others. It important to disentangle the truly valuable notion of correlation from the discredited hands that may have first written it down. Therefore, we will not be referring to it in this text as the Pearson correlation coefficient.

3.5 Covariance with standardized data

In the last chapter, you showed that the variance of a standardized variable was 1. What is the covariance between two standardized variables?

Let's standardize both X_1 and X_2. To make the math a little easier, we'll use similar notation to what we used at the end of the last chapter.

$M_1 = \overline{X_1}$

$S_1 = SD(X_1)$

$M_2 = \overline{X_2}$

$S_2 = SD(X_2)$

And we'll write the z-scores in a way that is more amenable to mathematical manipulation (like before):

$$Z_1 = \frac{1}{S_1}(X_1 - M_1)$$

$$Z_2 = \frac{1}{S_2}(X_2 - M_2)$$

This looks a little more intimidating, but if you apply the rules, it works out:

$$Cov(Z_1, Z_2) = Cov\left(\frac{1}{S_1}(X_1 - M_1), \frac{1}{S_2}(X_2 - M_2)\right) \tag{3.6}$$

$$= \quad ??? \tag{3.7}$$

Work this out. Take your time. Apply the rules carefully. So that you know what you're aiming for, you should get

$$Cov(Z_1, Z_2) = \frac{Cov(X_1, X_2)}{S_1 S_2}$$

Okay, now remember that S_1 is just a convenient substitute for $SD(X_1)$ and S_2 is just a substitute for $SD(X_2)$. Wait, does that answer look familiar?

This is cool! Correlation is simply the covariance of two variables after they've been standardized.

This also reinforces the earlier comment about interpreting covariance after removing the extraneous influence of the spread of the individual variables. Standardizing variables makes the spread all all variables 1, so their covariance is now a pure representation of just the association between them.

You probably remember from intro stats that correlation takes on values between -1 and 1. That fact is not obvious from the formula we have. Why should the fraction

$$\frac{Cov(X_1, X_2)}{SD(X_1)SD(X_2)}$$

be bounded by -1 and 1?

Let's go back to standardized variable to keep things simple. The correlation is just the covariance of two standardized variables:

$$Corr(X_1, X_2) = Cov(Z_1, Z_2)$$

Use the rules to calculate this:

$$Var(Z_1 + Z_2)$$

Remember that Z_1 and Z_2 are not necessarily independent. (In fact, we hope they are not. Otherwise, why do we care about their correlation? It would be zero!) So you need **Rule 12**, not **Rule 2**. Keep manipulating until you get

$$2 + 2Corr(X_1, X_2)$$

Since variances are always non-negative, we now know that

$$0 \leq 2 + 2Corr(X_1, X_2)$$

Solve this inequality for $Corr(X_1, X_2)$.

Now follow the exact same steps for

$$Var(Z_1 - Z_2)$$

Very little should change in your answer, but there is one small change. Again, solve the resulting inequality. (Don't forget the key rule when working with inequalities that multiplying or dividing by a negative number changes the direction of the inequality.)

Here is a fact we will state without proof:

Correlation is only interpretable as the strength of **linear** associations.

Why is this? Basically, it boils down to the fact that a "perfect" correlation of 1 or -1 is only achievable when data points lie on a perfectly straight line. Therefore, thinking of correlation as lying between 0 and 1 (or 0 and -1) is only sensible if you are judging how close points are to lying on a straight line. We'll see examples of this in the next section when we plot some data.

To calculation correlation in R, use the `cor` command:

```
cor(airquality$Temp, airquality$Wind)
```

```
## [1] -0.4579879
```

Use R to confirm that the number above is the covariance divided by the product of the standard deviations.

3.6 Visualizing correlation

Covariance is hard to interpret, so when we're visualizing data and we want to understand any association that might exist between two variables, correlation is a much better statistic to calculate. Let's see how correlation relates to the graph of two variables.

Before getting into the graphing, we will need to load some packages. The `tidyverse` is a whole set of commonly used packages that will allow us to work with data frames (or "tibbles" as the cool kids are calling them) and make graphs. Be sure to load the package by typing the following in R before going any further:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

In fact, from here on out, we'll start each chapter by loading any necessary libraries in R that we'll need.

The standard graph of two numerical variables is a scatterplot. Let's start with a straight line relationship. First, we define two variables. We'll use some shortcuts here to make our lives a little easier. The `seq` command just generates a sequence of numbers.

```
X5 <- seq(1, 9)
X5
```

```
## [1] 1 2 3 4 5 6 7 8 9
```

Then we can establish a linear relationship just by declaring one in a formula:

```
X6 <- 3 + 0.5 * X5
X6
```

```
## [1] 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5
```

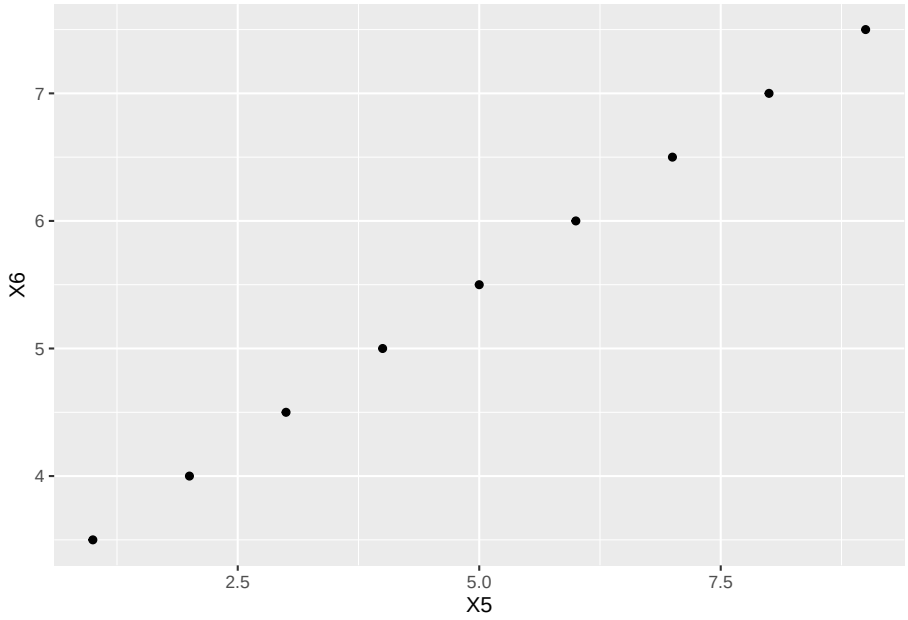
To put both variables into the same graph, it helps to make them both columns in a single tibble.

```
linear_data <- tibble(X5, X6)
linear_data
```

```
## # A tibble: 9 x 2
##      X5    X6
##   <int> <dbl>
## 1     1   3.5
## 2     2   4
## 3     3   4.5
## 4     4   5
## 5     5   5.5
## 6     6   6
## 7     7   6.5
## 8     8   7
## 9     9   7.5
```

And here is the graph:

```
ggplot(linear_data, aes(y = X6, x = X5)) +
    geom_point()
```



Now the correlation:

```
cor(X5, X6)
```

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

It is 1, as expected.

What about a perfectly straight line with a negative slope?

```
X7 <- 5 - 0.2 * X5
X7
```

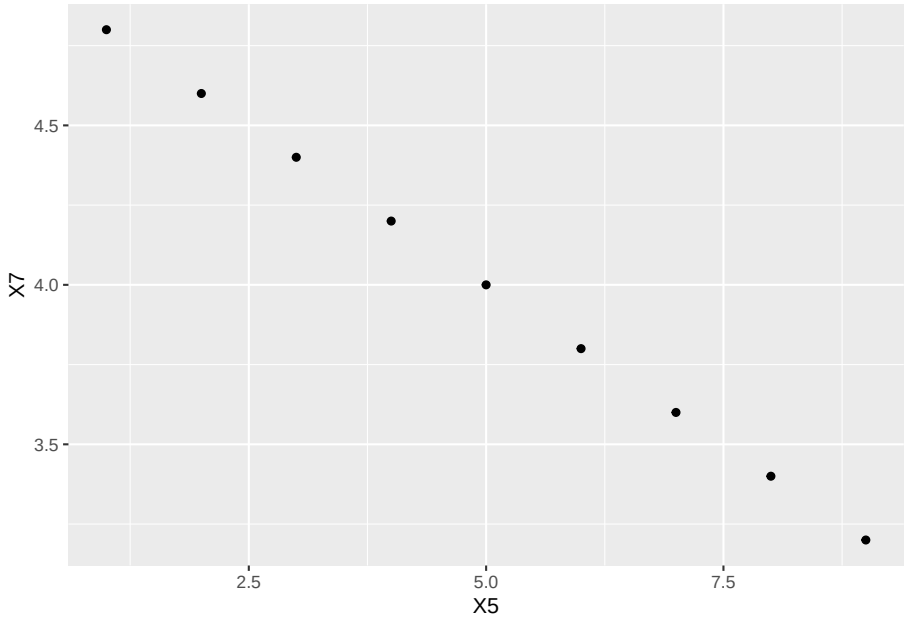
```
## [1] 4.8 4.6 4.4 4.2 4.0 3.8 3.6 3.4 3.2
```

Throw this new variable into the tibble we already have (for convenience). To explain the syntax below, the `%>%` symbol is called a "pipe" and it tells R to pass the `linear_data` tibble on to the next row to process it. And the processing itself is dictated by the `bind_cols` command which tells R to "bind a new column" to the tibble. The part that says `X7 = X7` may be a little confusing. It says to add the new column `X7`, but also still call it `X7`.

```
linear_data <- linear_data %>%
    bind_cols(X7 = X7)
linear_data
```

```
## # A tibble: 9 x 3
##      X5    X6    X7
##   <int> <dbl> <dbl>
## 1     1   3.5   4.8
## 2     2   4     4.6
## 3     3   4.5   4.4
## 4     4   5     4.2
## 5     5   5.5   4
## 6     6   6     3.8
## 7     7   6.5   3.6
## 8     8   7     3.4
## 9     9   7.5   3.2
```

```
ggplot(linear_data, aes(y = X7, x = X5)) +
    geom_point()
```




```
cor(X5, X7)
```

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

Again, that is what we expected.

What happens if we plot random data? The `runif` command just chooses random numbers uniformly between 0 and 1.[2]

```
set.seed(1234)
X8 <- runif(20)
X9 <- runif(20)
```

```
X8
```

```
##  [1] 0.113703411 0.622299405 0.609274733 0.623379442 0.860915384 0.640310605
##  [7] 0.009495756 0.232550506 0.666083758 0.514251141 0.693591292 0.544974836
## [13] 0.282733584 0.923433484 0.292315840 0.837295628 0.286223285 0.266820780
## [19] 0.186722790 0.232225911
```

```
X9
```

```
##  [1] 0.31661245 0.30269337 0.15904600 0.03999592 0.21879954 0.81059855
##  [7] 0.52569755 0.91465817 0.83134505 0.04577026 0.45609148 0.26518667
## [13] 0.30467220 0.50730687 0.18109621 0.75967064 0.20124804 0.25880982
## [19] 0.99215042 0.80735234
```

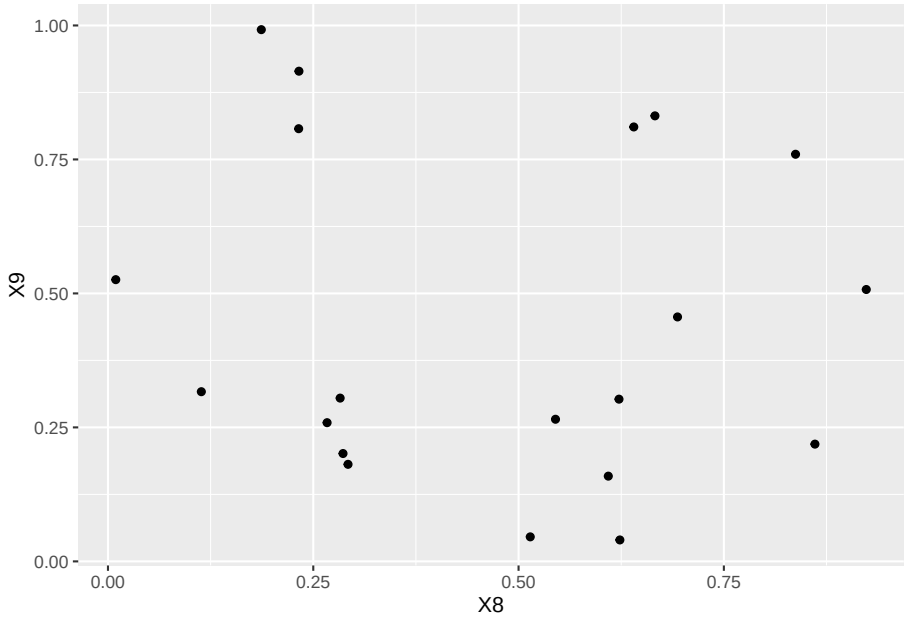
```
random_data <- tibble(X8, X9)
random_data
```

```
## # A tibble: 20 x 2
##         X8     X9
##      <dbl>  <dbl>
##  1 0.114   0.317
##  2 0.622   0.303
##  3 0.609   0.159
##  4 0.623   0.0400
##  5 0.861   0.219
##  6 0.640   0.811
##  7 0.00950 0.526
##  8 0.233   0.915
##  9 0.666   0.831
## 10 0.514   0.0458
## 11 0.694   0.456
## 12 0.545   0.265
## 13 0.283   0.305
```

[2] Sean's brain always want to parse this command as "run if". Run if what? No, no, it's "r unif".

```
## 14 0.923   0.507
## 15 0.292   0.181
## 16 0.837   0.760
## 17 0.286   0.201
## 18 0.267   0.259
## 19 0.187   0.992
## 20 0.232   0.807
```

```
ggplot(random_data, aes(y = X9, x = X8)) +
    geom_point()
```



What do you guess is the correlation between X_8 and X_9?

Now calculate it using R? Did you get the *exact* answer you guessed? If not, why not?

What about data that follows a perfect mathematical relationship that is not a straight line? For example, here is a part of a parabola.

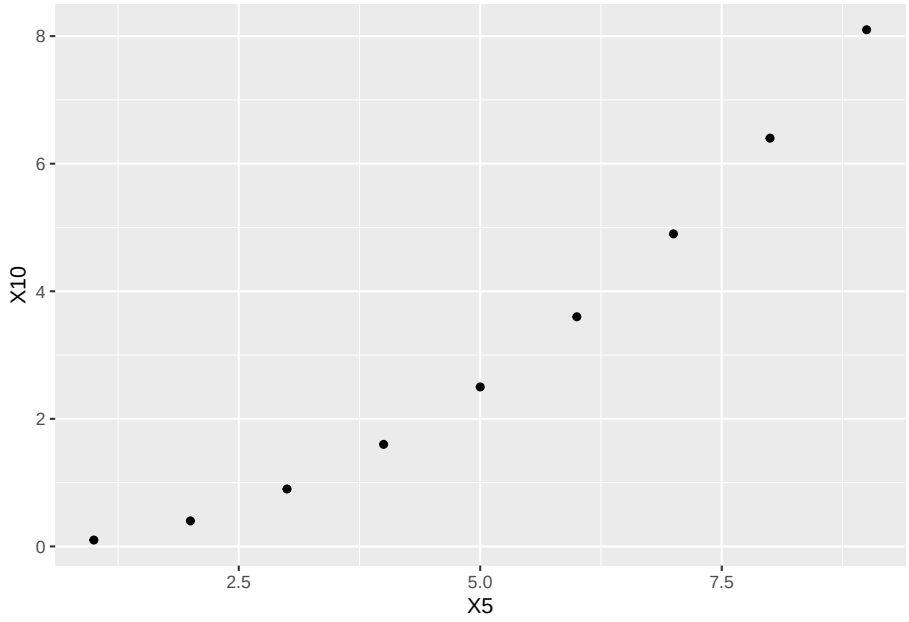
```
X10 <- 0.1 * X5^2
X10
```

```
## [1] 0.1 0.4 0.9 1.6 2.5 3.6 4.9 6.4 8.1
```

```
nonlinear_data <- tibble(X5, X10)
nonlinear_data
```

```
## # A tibble: 9 x 2
##      X5   X10
##   <int> <dbl>
## 1     1   0.1
## 2     2   0.4
## 3     3   0.9
## 4     4   1.6
## 5     5   2.5
## 6     6   3.6
## 7     7   4.9
## 8     8   6.4
## 9     9   8.1
```

```
ggplot(nonlinear_data, aes(y = X10, x = X5)) +
    geom_point()
```



Now for the correlation:

```
cor(X5, X10)
```

```
## [1] 0.975281
```

This is a large correlation, but it is not exactly 1, even though the points follow a precise mathematical relationship. That relationship is not linear.

Here's a fascinating example. For this, we'll want a parabola that goes down and then up.

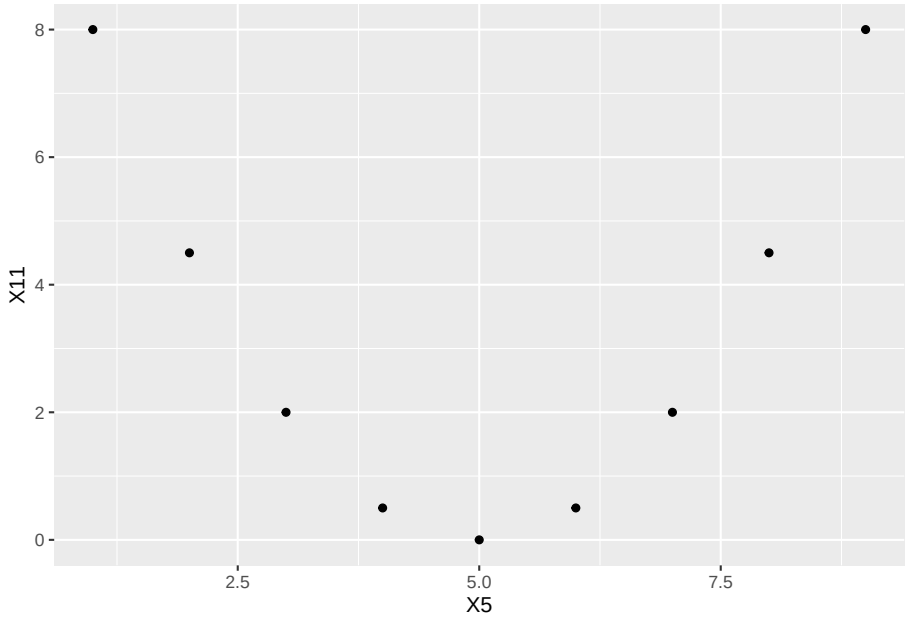
```
X11 <- 0.5 * (X5 - 5)^2
X11
```

```
## [1] 8.0 4.5 2.0 0.5 0.0 0.5 2.0 4.5 8.0
```

```
nonlinear_data <- nonlinear_data %>%
    bind_cols(X11 = X11)
nonlinear_data
```

```
## # A tibble: 9 x 3
##      X5   X10   X11
##   <int> <dbl> <dbl>
## 1     1   0.1   8
## 2     2   0.4   4.5
## 3     3   0.9   2
## 4     4   1.6   0.5
## 5     5   2.5   0
## 6     6   3.6   0.5
## 7     7   4.9   2
## 8     8   6.4   4.5
## 9     9   8.1   8
```

```
ggplot(nonlinear_data, aes(y = X11, x = X5)) +
    geom_point()
```



Before looking at the answer, what is your guess for the correlation between X_5 and X_{11}?

Now calculate the correlation in R.

Again, there's a perfect mathematical relationship between these two variables. They are most definitely associated. So why is the correlation 0?

Recall the earlier promise to discuss **Rule 11**. If two variables are independent, then their covariance is zero, and, therefore, their correlation is also zero. However, this rule doesn't work the other way around. The claim is that knowing the covariance/correlation is zero does not imply (necessarily) that the two variables are independent. Here is the promised example of that phenomenon. X_5 and X_{11} have zero correlation. And yet, X_5 and X_{11} are definitely *not* independent.

This is important enough for a fancy box:

When you see that the correlation between two variables is zero or near zero, be careful not to conclude that the variables are independent.

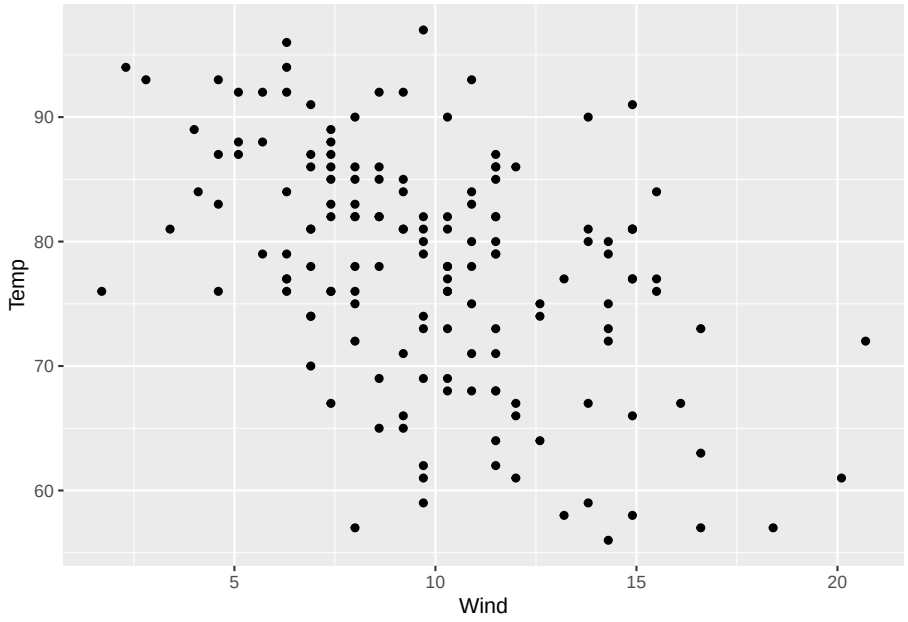
A zero or near-zero correlation indicates only the lack of a *linear* association between two variables. There may be nonlinear associations. That's why it's always a good idea to graph your data.

Real data is, of course, much messier and it's just not possible to have perfect correlations between two variables measured out there in the real world. (If you do find a perfect correlation between two columns of your data, chances are

that you either recorded the same column twice, or the second column is some simple transformation of first column, like multiplying every value by the same number or something like that.)

Here is a plot of the temperature (degrees Fahrenheit) and wind speed (mph) from the New York air quality data set.

```
ggplot(airquality, aes(y = Temp, x = Wind)) +
    geom_point()
```



Just looking at the scatterplot (without calculating anything), is the correlation between these two variables positive or negative? Try guessing the exact value of the correlation.

Now calculate the exact value of the correlation to see how close you were.

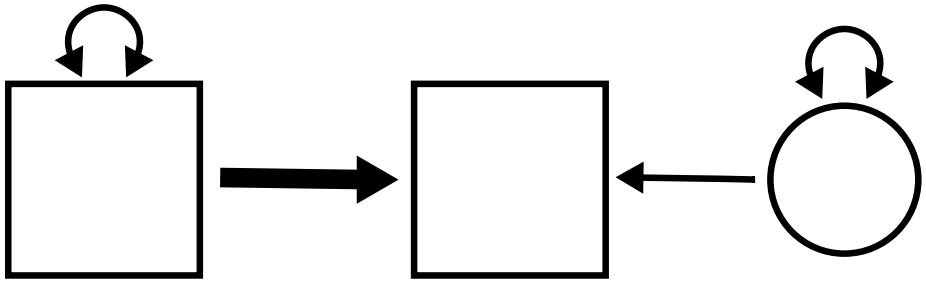
If you want some practice with looking at scatterplots and guessing the correlation, try this online game:

Guess the Correlation

Turn up the sound! If the whole class plays at the same time, your classroom will sound like an arcade. Compete with your classmates to see who can get the high score.

Chapter 4

Simple regression



Preliminaries

We need to load the packages we will need for this chapter. The `tidyverse` package has all sorts of utilities for working with tibbles (data frames). This will also be our first introduction to the `lavaan` package that will be used throughout the rest of the book.

```
library(tidyverse)
library(lavaan)
```

```
## This is lavaan 0.6-11
## lavaan is FREE software! Please report any bugs.
```

4.1 Prediction

One of the most important tasks in statistics is *prediction*. Given some data, can we predict the value of something important about a population of interest?

Suppose you have gathered some data on anxiety among Utah high school students. There are various instruments available for measuring anxiety, so say you have administered the Beck Anxiety Inventory. This instrument assigns a score from 0 to 63, with lower numbers indicating less anxiety and higher numbers indicating more.

You take care to make sure your sample is as close to a simple random sample as possible so that it's representative of the population (all high school students in the state of Utah). From your sample data, you can calculate summary statistics. For example, you might find that the mean anxiety score for Utah high school students is 7.1 with a standard deviation of 3.9.

A random Utah high school student walks through the door. You don't know anything about them. Can you say anything about their anxiety? What is your best guess as to what their score might be on the Beck Anxiety Inventory?

We can do a lot better if we have another variable we can measure. For example, let's suppose our data records not only anxiety, but also the minutes of smart phone usage per day.

In theory, why would having information about smart phone usage potentially help us make better predictions of anxiety?

Do you suspect that the association between anxiety and smart phone usage is positive or negative? (You can Google this question to check if there is any empirical evidence out there for your guess.)

Now imagine that another random Utah high school student walks through the door. This time, I tell you that their smart phone usage is average (sitting at the mean). What is your best prediction for their anxiety score? (Give an exact value.)

What if I told you that the student who walked through the door had *higher* than average smart phone usage? What would be your prediction of their anxiety score? (You can't give an exact value here, but give a qualitatively sensible answer.)

What if I told you that the student who walked through the door had *lower* than average smart phone usage? What would be your prediction of their anxiety score? (Again, just give a qualitatively sensible answer.)

4.2 Regression terminology

When we have one variable we suspect may help us predict another variable, one way to study it is using a simple regression model.

This is related to, but somewhat different from, covariance. Covariance is symmetric, so it expresses the idea that two variables are mutually related. But there is no "directionality" to that relationship. By way of contrast, a simple regression model asserts that one of the variables is "explanatory" and the other is "response". In other words, we start with the values or properties of the explanatory variable and try to deduce what we can about the values or properties of the response variable.

Keep in mind that "directionality" is not the same as "causality". While it's possible that one variable causes another, there needs to be a data collection process (often a carefully controlled experiment) and a clear scientific rationale that justifies a causal relationship between variables before we can start thinking about inferring causality. For purposes of much of this book, directionality just means that we wish to establish a predictive relationship wherein we start with the properties of one variable and try to predict the properties of another variable. There is often a "sensible" order in which to do this based on the research questions asked or the hypotheses posed.

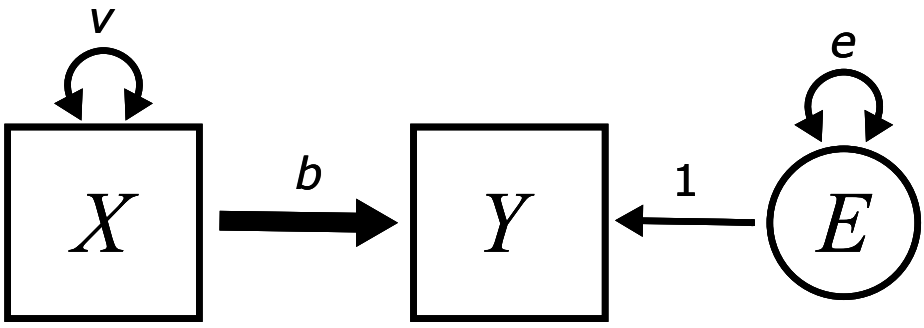
There are many different terminological conventions in statistics, so be aware that "explanatory" variables are also called—often depending on the discipline and the context—predictors, features, covariates, controls, regressors, inputs, or independent variables. In fact, in the context of structural equation modeling, we will use the term "exogenous" to refer to variables that play this role. (That term has a much more precise definition that we'll discuss in future chapters.) And "response" variables might be called outcomes, outputs, targets, explained variables, or dependent variables, among others. In this book, we will often use the term "endogenous" (again, in a very specific way yet to be explained). If there is a data collection process and a clear scientific rationale that justifies a causal relationship between variables, then we might be able to refer to variables as either "cause" or "effect".

Keep in mind that it's the scientific question we want to ask that determines the explanatory/response relationship. A different researcher with a different hypothesis might use the same two variables but with the roles reversed.

In the anxiety/smart-phone example above, which variable is explanatory and which is response, at least according to the way we stated the scenario?

4.3 The simple regression model

Here is the figure from the top of the chapter, only now we have decorated it with some letters (and a number):



The goal of this section is to explain all these.

The variable names are X and Y. X is the exogenous variable and Y is the endogenous variable. For example, X might be smart phone usage and Y might be the anxiety score from the example above. In this section, we're going to do some concrete calculations using the example from the last chapter about wind speed and temperature from the `airquality` data set. In the last chapter, we simply calculated the (symmetric) correlation between wind speed and temperature. Here, we will consider wind speed as exogenous and temperature as endogenous. In other words, our goal is to use the wind speed as a predictor of temperature.

The letter v requires no further explanation. This is the variance of the variable X, so we already know about it.

The parameter b is supposed to measure something about the predictive relationship between X and Y. It is attached to an arrow that is drawn a little thicker than the other arrows in the diagram. More to say about that in a moment.

The really weird, new part is the circle on the right. This will be the "error" term.

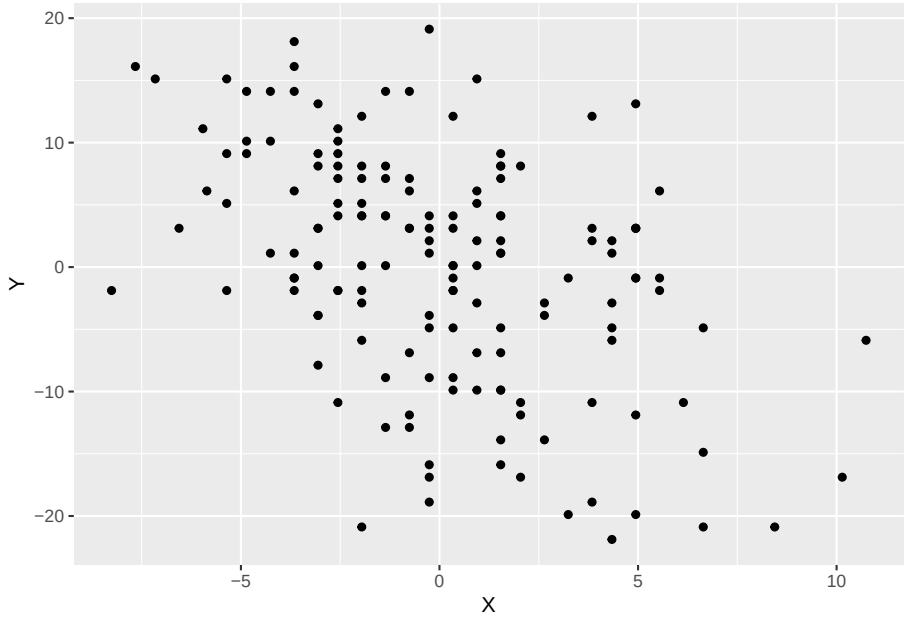
What is "error" and why is it here? To illustrate, let's plot wind speed against temperature. Before plotting and analyzing these variables, we are going to mean-center them and put them in a tibble.

```
X <- airquality$Wind - mean(airquality$Wind)
Y <- airquality$Temp - mean(airquality$Temp)
airquality_mc <- tibble(X, Y)
airquality_mc
```

```
## # A tibble: 153 x 2
##        X      Y
##    <dbl>  <dbl>
##  1 -2.56 -10.9
```

```
##  2 -1.96  -5.88
##  3  2.64  -3.88
##  4  1.54 -15.9
##  5  4.34 -21.9
##  6  4.94 -11.9
##  7 -1.36 -12.9
##  8  3.84 -18.9
##  9 10.1  -16.9
## 10 -1.36  -8.88
## # ... with 143 more rows
```

```
ggplot(airquality_mc, aes(y = Y, x = X)) +
    geom_point()
```

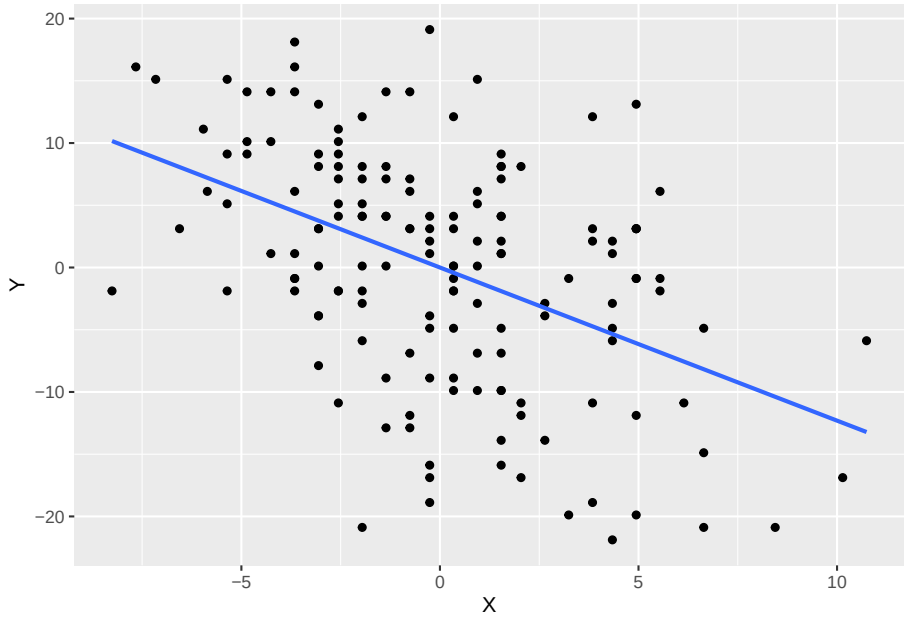


Note that the exogenous variable (wind speed) is on the x-axis and the endogenous variable (temperature) is on the y-axis.

We can see a negative and reasonably linear association between these variables, so let's add a line of best fit to the data.

```
ggplot(airquality_mc, aes(y = Y, x = X)) +
    geom_point() +
    geom_smooth(method = lm, se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The line passes right through the origin $(0, 0)$. Why?

The slope of this line is -1.23. We'll see how to calculate this slope in a bit. But what does this slope mean?

Look at the help file for the `airquality` data set. (Either use the Help tab in RStudio or type `?airquality` at the Console.)

What are the units of measurement of X? What are the units of measurement of Y? Since slope is "rise over run", what are the units of the slope?

So, the idea is that for every additional mile per hour of wind speed, we predict that the temperature goes down by 1.23 degrees Fahrenheit.

Why is the following sentence incorrect?

> For every additional mile per hour of wind speed, the temperature goes down by 1.23 degrees Fahrenheit.

The point is that the line is a *model* that makes predictions. As long as X and Y are both mean-centered, the equation of this line is

$$\hat{Y} = bX$$

There is a new piece of notation here: $\hat{Y}$. This symbol represents the *predicted* value of Y according to the model. We will not get the *actual* value of Y from this piece of the model because the actual values of Y differ from the model

because real-world data doesn't lie on a perfect straight line. More on that in a moment.

According to the information above, we can estimate that the value of b is -1.23:

$$\hat{Y} = -1.23X$$

This is a multiplicative effect. Again, as long as X and Y are both mean-centered, knowing the value of X allows us to predict the value of Y by multiplying by b.

But those predictions will almost always be wrong. On any given day, given an increase of 1 mile per hour wind in wind speed, it will very rarely happen that the temperature will drop by *exactly* 1.23 degrees. That's just a sort of "average" over time. On average, there's a slight temperature change associated with 1 mph change in wind speed, and the number -1.23 is the best estimate of that average change across our whole data set. We need to be *especially* clear that an increase in wind speed does not necessary *cause* a drop in temperature. I mean, that might be partially true, but we can't prove it from our observational data. There are all sorts of other reasons to explain both an increase in wind speed and a drop in temperature (like a cold front moving in).

Since our predictions are average effects and not specific guarantees, every prediction we make will be wrong by some amount. (We could get extremely lucky, but even then, it's difficult to imagine a situation in which our prediction is precisely correct to, say, 10 decimal places or something like that.) Therefore, there is error in our prediction. The new equation—accounting for error—is

$$Y = bX + E$$

or

$$Y = -1.23X + E$$

Now we use Y instead of $\hat{Y}$. Once we include the error, we can recover the exact value of Y, so this is no longer just a prediction from the straight-line model. Remember this:

If you write down a regression equation for an endogenous variable that includes all incoming arrows, *including the residual error term*, use Y.

If you write down a regression equation for an endogenous variable that includes all incoming arrows, *excluding the residual error term*, use $\hat{Y}$.

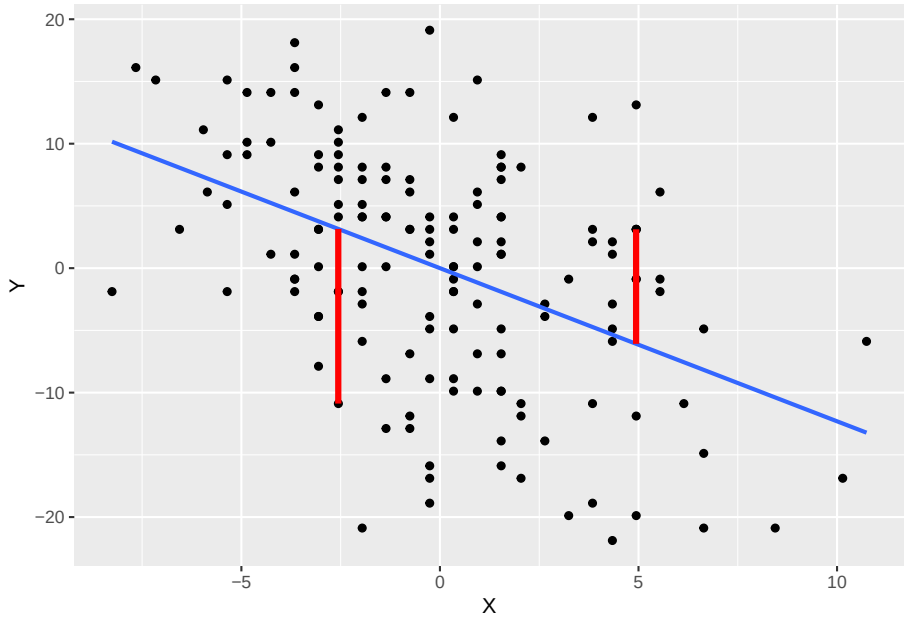
Error is a funny word because it has a negative connotation. It sounds like we made a mistake. Well, the model does make mistakes. Every model prediction is technically wrong. But this is not the kind of mistake that results from doing our arithmetic wrong or anything like that. It's simply the "natural" error that results from the messiness of the real world and the impossibility of predicting

anything with certainty. For this reason, we will often prefer the term "residual". It's what is "left over" after we have made a prediction. It's the extra change in temperature, for example, that is not accounted for by the model with wind speed alone.

The residuals are evident in the plot above. If there were no residuals, every data point would lie on a perfect straight line. But the data points are all either a little above or below the line. Those vertical distances between the data and the line are the residuals or errors. Here is an example of two residuals plotted below in red.

```
ggplot(airquality_mc, aes(y = Y, x = X)) +
    geom_point() +
    geom_smooth(method = lm, se = FALSE) +
    annotate("segment",
             x = -2.56, y = 3.15,
             xend = -2.56, yend = 3.15 - 14.03,
             color = "red", size = 1.5) +
    annotate("segment",
             x = 4.94, y = -6.08,
             xend = 4.94, yend = -6.08 + 9.20,
             color = "red", size = 1.5)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Points below the line have negative residuals and points above the line have positive residuals.

The residuals do not appear as observed or measured variables in our data. They are a consequence of a variety of *unmeasured* factors that determine temperature aside from wind speed. An unmeasured variable that appears in a model is called a *latent variable*. We will discuss latent variables in far greater detail in Chapter 8. For now, just know that latent variables are indicated by circles in the diagram. That's why there is a circle with the letter E inside.

The equation

$$Y = bX + E$$

can also be written as

$$Y = bX + 1 \cdot E$$

How is that "1" represented in the diagram?

The only thing in the diagram that hasn't been explained yet is e.

Where does e appear in the diagram? Given where it appears, what does it represent mathematically?

We know that curved arrows represent variance. But what does it mean to measure the variance of a variable we can't observe?

What would the scatterplot look like if the error variance were very small. What about if the error variance were large?

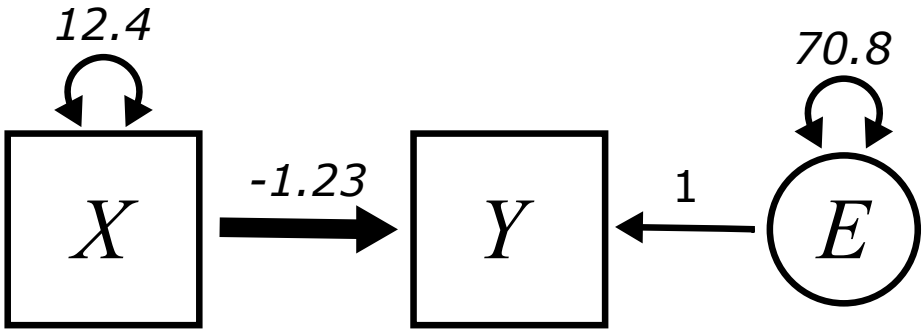
There is variability in the size of the residuals. Some are small (points that are close to the line) and some are large (points that are far from the line). This spread of residuals can be estimated from data, just like any other variance calculation.[1] It turns out to be about 70.8. We'll see how to calculate that below.

Let's put everything together into a diagram.

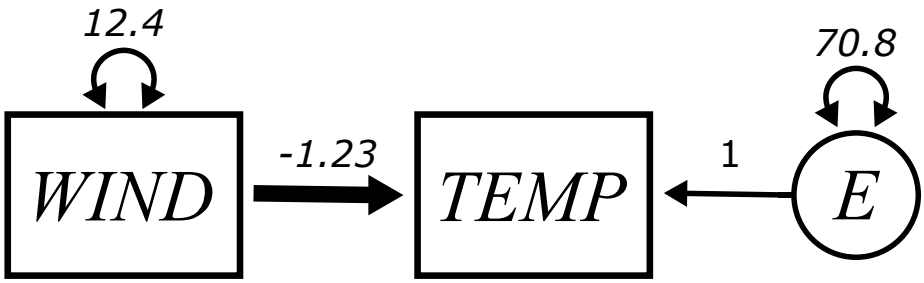
First, we need the variance of X (wind speed). Calculate it in R. You should get 12.4.

Does it matter if you calculate the variance of the variable `Wind` from the original `airquality` data, or the variance of the X variable from `airqualilty_mc`? Why or why not?

[1] We know that variance is close to the average squared deviation, except we divide by $n-1$ instead of n. Well, residuals are a little more weird still. To get an unbiased estimate, you have divide by $n-2$. However, the calculation that appears next is the one that uses $n-1$. This is for reasons that, regrettably, we'll have to sweep under the rug here.



While it's helpful to see X and Y as generic prototypes for any simple regression model, in most applied problems from now on we'll refer to variables using contextually meaningful names. The final diagram looks like this:



Is the error variable E exogenous or endogenous?

One final note about this diagram: you may have noticed that the Y variable (or *TEMP*) does *not* have a variance term attached. There is no double-headed arrow on that box. Why not? The point here is that we are trying to explain the reasons that Y varies using other elements of the model. In other words, Y has a variance, but that variance is partially explained by X. And the rest of the variance not explained by X is explained by the error term E. So all the variance is Y is accounted for through the contribution of X and E combined.

4.4 Simple regression assumptions

All the calculations we need to do, and our ability to interpret the results, depend on certain assumptions being met.

If you look up regression assumptions, you might find a huge list of requirements. Some of these requirements relate to calculating statistics like P-values

for regression parameters. For now, we are content simply to know that it makes sense to interpret the parameters in the model above.

For that, we really only need four assumptions:

1. The data should come from a "good" sample.
2. The relationship between X and Y should be approximately linear.
3. The residuals should be independent of the X values.
4. There should be no influential outliers.

Let's address these one at a time:

1. What do we mean by a "good" sample? While a simple random sample is the gold standard, it's usually not possible to obtain one in the real world. So we make our sampling process as random as possible and ensure that the resulting sample is as representative of the population we're trying to study as possible.
2. You can check linearity with a scatterplot. Just make sure the pattern of dots doesn't have strong curvature to it.
3. There should be no patterns in the residuals at all. They should be randomly scattered around the best-fit line and the average size of the residuals should not change radically from one side of the graph to the other.[2]
4. Check the scatterplot for outliers. If there are serious ones, assess them to make sure they are not data entry mistakes. If they correspond to valid data, you cannot just throw them away.[3] Often, the solution is to run the analysis both including and (temporarily) excluding the outliers to make sure their presence doesn't radically alter the parameter estimates.

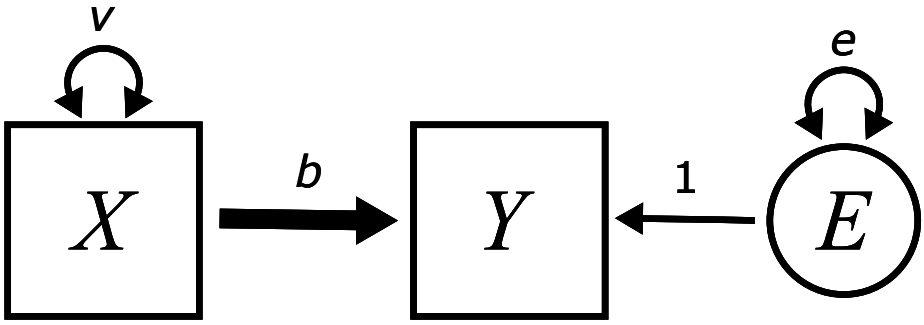
4.5 Calculating regression parameters

Now we'll show one way to calculate the regression parameters (the numbers) in the above diagram. This isn't the only way to do it. In fact, this is not the approach that is used in most intro stats classes. But this approach will be helpful to illustrate the way we will do these calculations in future chapters.

Let's go back to the diagram without the numbers:

[2]This property of similarly-sized residuals is called *homoskesdasticity*. The violation of that condition is called *heteroskedasticity* which is one of Sean's favorite words ever!

[3]In other words, don't follow the all-too-common "rule of thumb" for outliers, which is just covering them up with your thumb and pretending they don't exist.



The letter v is easy because it's just the variance of X:

$$Var(X) = v$$

We can estimate it directly from the data. (You already did it in R above for the temperature and wind speed data.) Through completing the activities below, we will also calculate b and e.

To get the other parameters, we have to set up a few equations. The first observation we need to make is that, as important as the arrows are in a diagram, it's just as important where the arrows are *not*.

Are there any arrows directly connecting X and E? What does that imply about the relationship between X and E? Which regression assumption is related to this question?

Given all that, what is $Cov(X, E)$?

Next, because Y is the combination of X and E, we've already seen that we can write

$$Y = bX + E$$

Therefore, we can calculate $Var(Y)$ according to this formula using the established rules. (A convenient list of all of them in one place is located in Appendix A.)

Keep simplifying the following as much as possible:

$$\begin{aligned} Var(Y) &= Var(bX + E) && (4.1) \\ &= \quad ??? && (4.2) \end{aligned}$$

You should end up with

$$b^2v + e$$

Don't forget that $Var(X) = v$ and $Var(E) = e$ in the diagram!

We also need to use information about the covariance between X and Y. Keep simplifying the calculation below:

$$\begin{aligned} Cov(X, Y) &= Cov(X, bX + E) && (4.3)\\ &= \quad ??? && (4.4) \end{aligned}$$

You should end up with

$$bv$$

Use R to calculate $Var(X)$, $Var(Y)$ and $Cov(X, Y)$.

You should get 12.4, 89.6, and -15.3, respectively.

Now we can set up all the equations we need to solve for the various letters we want. Here are the three equations we have established:

$$\begin{aligned} 12.4 &= v && (4.5)\\ 89.6 &= b^2v + e && (4.6)\\ -15.3 &= bv && (4.7) \end{aligned}$$

Time to do a little algebra. You know v. Using that value, solve for b first (using the last equation). Then, using both values of b and v, solve for e in the second equation.

Check that the values you got are the same as the ones from the earlier diagram (with the possibility of a little rounding error).

Now let's go through that again, but this time, in full generality:

$$\begin{aligned} Var(X) &= v && (4.8)\\ Var(Y) &= b^2v + e && (4.9)\\ Cov(X, Y) &= bv && (4.10) \end{aligned}$$

Therefore,

$$v = Var(X)$$

$$b = \frac{Cov(X,Y)}{Var(X)}$$

$$e = Var(Y) - \left(\frac{Cov(X,Y)}{Var(X)}\right)^2 Var(X)$$

4.6 The model-implied matrix

It will be convenient in future chapters to collect up all these numbers we need in an array of terms called a *sample covariance matrix*. (Sometimes this is called a variance-covariance matrix.) The idea is to take the covariance of all possible pairs of observed variables and arrange them as follows:

$$\begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix}$$

There are some immediate simplifications to make.

1. Since $Cov(Y,X) = Cov(X,Y)$, there is no point in writing it twice. We will just use a dot ($\bullet$) to replace $Cov(Y,X)$.
2. We can replace the upper-left and lower-right entries (the entries on the so-called "diagonal" of the matrix) with variances.

This is our final sample covariance matrix:

$$\begin{bmatrix} Var(X) & Cov(X,Y) \\ \bullet & Var(Y) \end{bmatrix}$$

With the data we have, we can calculate numbers for all these quantities.

There is another important matrix called the *model-implied matrix*. Given the model, what does the covariance matrix look like? From the calculations above, we know that the model-implied matrix is

$$\begin{bmatrix} v & bv \\ \bullet & b^2v + e \end{bmatrix}$$

The letters b, v, and e are unknowns. Okay, v is not *very* unknown. It's an unknown in the sense of being a parameter in the model, but you don't have to work very hard to find it. The point is that model parameters are estimated by equating the covariance matrix (calculated from the data) with model-implied matrix and trying to solve for all the unknown parameters.

There is no new math to do in this section. The matrices are just convenient ways to organize the work we've already done. All parameter estimation in structural equation modeling is essentially setting these two matrices (the sample covariance matrix and the model-implied matrix) equal to each other and attempting to solve for the unknown parameters.

$$\begin{bmatrix} Var(X) & Cov(X,Y) \\ \bullet & Var(Y) \end{bmatrix} = \begin{bmatrix} v & bv \\ \bullet & b^2v + e \end{bmatrix}$$

Don't forget that the matrix on the left—the sample covariance matrix—consists of numbers that we calculate from data. The matrix on the right—the model-implied matrix—contains letters, which are the unknown parameters we're trying to find.

4.7 Error variance in terms of correlation

The formulas we derived are fine as far as they go. They allow you to take quantities calculated from data (variances and covariances of observed variables) and translate that into estimates of model parameters.

The formula for the error variance e is a little gross, though. With a little trickery, we can simplify that formula quite a bit.

Here is the starting point:

$$e = Var(Y) - \left(\frac{Cov(X,Y)}{Var(X)}\right)^2 Var(X)$$

Explain why the right-hand side can be rewritten as

$$Var(Y) - \frac{Cov(X,Y)^2}{Var(X)}$$

We're going to multiply the top and bottom of the fraction on the right by $Var(Y)$:

$$Var(Y) - \frac{Cov(X,Y)^2 Var(Y)}{Var(X)Var(Y)}$$

Then we can factor out a common term of $Var(Y)$ from both pieces:

$$Var(Y)\left(1 - \frac{Cov(X,Y)^2}{Var(X)Var(Y)}\right)$$

Why would we do such a thing? In other words, does the new fraction on the right look familiar in any way?

We hope you recognize that the fraction on the right is just the correlation coefficient squared. The whole equation can now be written as

$$e = Var(Y)\,(1 - r_{XY}^2)$$

There is a nice consequence of this last equation. The term in parentheses $(1 - r_{XY}^2)$ is a number between 0 and 1, right? Since we are multiplying this by the variance of Y, we can think of the term in parentheses as a *proportion*. All the variance of Y is explained in our model in one of two ways. The thick arrow coming in from the left uses X to predict some of the variance of Y. All the rest of the variance of Y is left over in the error term e. Therefore, $(1 - r_{XY}^2)$ is the proportion of the variance of Y left over as error.

And if that is true, it must also be the case that r_{XY}^2 is the proportion of the variance of Y explained by X. Calculating one minus a proportion gives the complementary proportion. For example, if $(1 - r_{XY}^2) = 0.3$, then 30% of the variance of Y is left over as error. But that implies that 70% of the variance of Y is explained by X. $1 - 0.3 = 0.7$.

Most authors will write R^2 instead of r^2 for some reason.

4.8 Regression with standardized variables

Recall that if we convert our variables to z-scores, variances are all 1 and covariances become correlation coefficients. In other words, the covariance matrix becomes a *correlation matrix* and looks like this:

$$\begin{bmatrix} 1 & r_{XY} \\ \bullet & 1 \end{bmatrix}$$

The model-implied matrix does not change. Solving for the parameters as before is the same, then, except we can now replace $Var(X)$ and $Var(Y)$ with 1, and $Cov(X, Y)$ with r_{XY}.

$$\begin{bmatrix} 1 & r_{XY} \\ \bullet & 1 \end{bmatrix} = \begin{bmatrix} v & bv \\ \bullet & b^2v + e \end{bmatrix}$$

$$1 = v \tag{4.11}$$

$$r_{XY} = bv \tag{4.12}$$

$$1 = b^2v + e \tag{4.13}$$

Therefore,

$$v = 1$$

$$b = r_{XY}$$

$$e = 1 - r_{XY}^2$$

We'll use the `scale` command to create standardized variables for temperature and wind speed and put them in a new tibble.

```
X_std <- scale(airquality$Wind)
Y_std <- scale(airquality$Temp)
airquality_std <- tibble(X_std, Y_std)
airquality_std
```

```
## # A tibble: 153 x 2
##    X_std[,1] Y_std[,1]
##        <dbl>     <dbl>
##  1    -0.726    -1.15
##  2    -0.556    -0.621
##  3     0.750    -0.410
##  4     0.438    -1.68
##  5     1.23     -2.31
##  6     1.40     -1.26
##  7    -0.385    -1.36
##  8     1.09     -1.99
##  9     2.88     -1.78
## 10    -0.385    -0.938
## # ... with 143 more rows
```

Modify the `ggplot` code from earlier in the chapter to create a scatterplot of the new standardized variables along with a best-fit line. What is the slope of this line? (Hint: calculate the correlation coefficient between the two standardized variables.)

4.9 Simple regression in R

The straightforward way to run regression in R is to use the `lm` command. This stands for "linear model". It uses a special symbol, the tilde ~, to express the relationship between the endogenous variable and the exogenous variable. The

endogenous (response) variable always goes on the left, before the tilde. The exogenous (predictor) variable goes on the right, after the tilde. Finally, there is a `data` argument to tell `lm` where to find the variables to model.

```
lm(Y ~ X, data = airquality_mc)
```

```
##
## Call:
## lm(formula = Y ~ X, data = airquality_mc)
##
## Coefficients:
## (Intercept)            X
##   1.117e-14   -1.230e+00
```

Which of the two numbers above is the slope b?

We haven't talked about the intercept yet, but according to this output, what is it? (Hint: it's not literally 1.117×10^{-14}. What does that number really mean?)

Run the `lm` command, but this time using the standardized variables from the `airquality_std` tibble. The value of the slope should not surprise you. Explain why it is what it is.

We will also introduce you briefly to the `lavaan` package. While it's totally overkill for simple regression, getting used to the syntax now will make it easier to continue to build up your confidence in using it when it will be the only tool we use.

A `lavaan` model is built in a similar way to `lm` using the tilde ~ notation. One big difference is that the model needs to be specified inside quotation marks first and assigned to a name like this:

```
TEMP_WIND_model <- "Y ~ X"
```

Then we pass that model text to the `sem` function from `lavaan`:

```
TEMP_WIND_fit <- sem(TEMP_WIND_model, data = airquality_mc)
```

The model is now stored as `TEMP_WIND_fit`. One way to learn about the model is to use the `parameterEstimates` function.

```
parameterEstimates(TEMP_WIND_fit)
```

```
##   lhs op rhs    est    se      z pvalue ci.lower ci.upper
## 1   Y  ~   X -1.230 0.193 -6.373      0   -1.609   -0.852
## 2   Y ~~   Y 70.337 8.042  8.746      0   54.575   86.098
## 3   X ~~   X 12.330 0.000     NA     NA   12.330   12.330
```

There is a lot of output here, and we're not going to talk about most of it now. Focus on the "est" column.

You should recognize these three estimates. Explain what these numbers represent.

In particular, pay close attention to the second line. If you are hasty, you may think this is the variance of Y, but that is not correct.

We can also produce the standardized estimates.

```
standardizedSolution(TEMP_WIND_fit)
```

```
##   lhs op rhs est.std    se      z pvalue ci.lower ci.upper
## 1   Y  ~   X  -0.458 0.060 -7.577      0   -0.576   -0.340
## 2   Y ~~   Y   0.790 0.055 14.273      0    0.682    0.899
## 3   X ~~   X   1.000 0.000     NA     NA    1.000    1.000
```

Again, explain these three numbers. (They are now listed in a column called `est.std` for "standardized estimates".)

Verify that the second line is actually the error variance. (Hint: remember $1 - r^2_{XY}$.) How do we know it's not the standardized variance of Y? (In other words, what do you actually know to be the standardized variance of Y?)

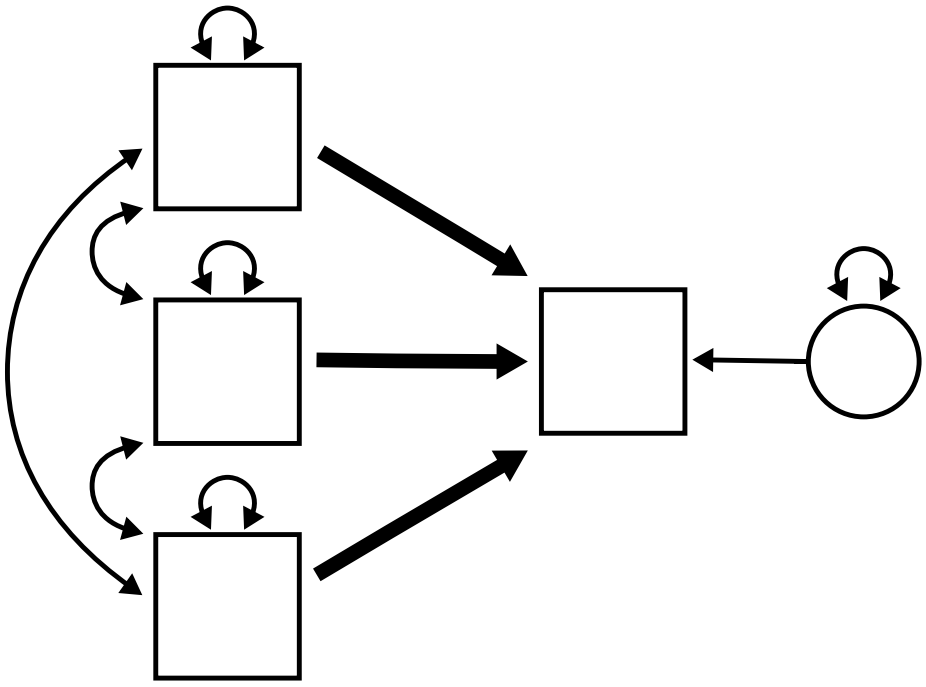
4.10 What about intercepts?

If you are familiar with regression from another course, you may be wondering where the intercepts went. Because we mean-centered and/or standardized all the data, there were no intercepts. The regression line always passes through $(0, 0)$ for mean-centered or standardized data.

[PUT A REFERENCE HERE IF WE DECIDE TO COVER MEAN STRUCTURE IN A FUTURE CHAPTER.]

Chapter 5

Multiple regression



Chapter 6

Mediation

Chapter 7

Path analysis

Chapter 8

Latent variables

Chapter 9

Confirmatory factor analysis

Chapter 10

Structural equation models

Chapter 11

Structural causal models

Appendix A

Variance/covariance rules

- **Rule 1**

If C is constant, then

$$Var\left(C\right) = 0$$

- **Rule 2**

If X_1 and X_2 are independent, then

$$Var\left(X_1 + X_2\right) = Var\left(X_1\right) + Var\left(X_2\right)$$

Consequence of **Rule 1** and **Rule 2**:

$$Var\left(X + C\right) = Var\left(X\right)$$

- **Rule 3**

If X_1 and X_2 are independent, then

$$Var\left(X_1 - X_2\right) = Var\left(X_1\right) + Var\left(X_2\right)$$

- **Rule 4**

If a is any number,

$$Var\left(aX\right) = a^2 Var\left(X\right)$$

- **Rule 5**

$$Cov(X, X) = Var(X)$$

- **Rule 6**

$$Cov\left(X_1, X_2\right) = Cov\left(X_2, X_1\right)$$

- **Rule 7**

If C is constant, then

$$Cov\left(X, C\right) = 0$$

- **Rule 8**

$$Cov\left(X_1 + X_2, X_3\right) = Cov\left(X_1, X_3\right) + Cov\left(X_2, X_3\right)$$

- **Rule 9**

$$Cov\left(X_1 - X_2, X_3\right) = Cov\left(X_1, X_3\right) - Cov\left(X_2, X_3\right)$$

Consequence of **Rule 6**, **Rule 8**, and **Rule 9**:

$$Cov\left(X_1, X_2 \pm X_3\right) = Cov\left(X_1, X_2\right) \pm Cov\left(X_1, X_3\right)$$

- **Rule 10**

If a is any number,

$$Cov\left(aX_1, X_2\right) = aCov\left(X_1, X_2\right) = Cov\left(X_1, aX_2\right)$$

- **Rule 11**

If X_1 and X_2 are independent, then

$$Cov\left(X_1, X_2\right) = 0$$

- **Rule 12**

For *any* two variables X_1 and X_2:

$$Var(aX_1 + bX_2) = a^2Var(X_1) + b^2Var(X_2) + 2abCov(X_1, X_2)$$

Appendix B

LISREL notation