Demystifying Structural Equation Modeling

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

2022-05-09

Contents

Introduction

Welcome to our book on structural equation modeling!

If you want, you can also download this book as a PDF or EPUB file.

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 TEST

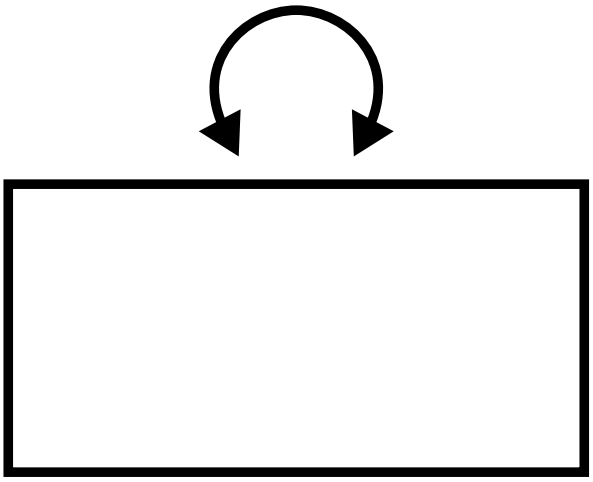
Note here

```
2 + 2
```

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

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 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 five 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 will 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	$(X - \overline{X})$	$(X - \overline{X})^2$
		$\overline{28}$

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

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

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

Once you've done it by hand a few times to make sure you 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
```

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
```

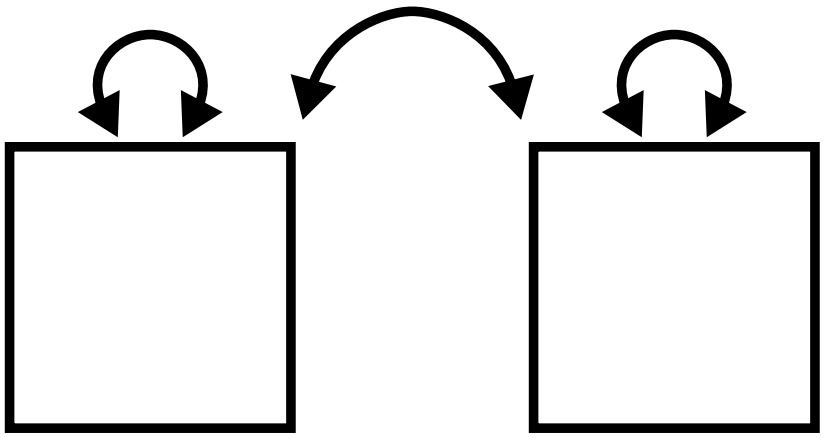
2.3 Mean centering data

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.[4]

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

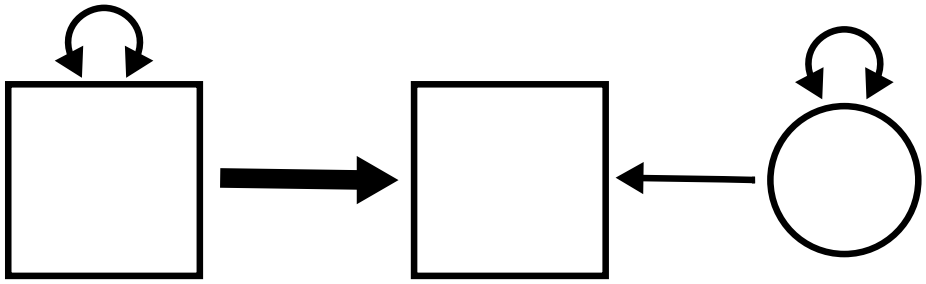
Chapter 3

Covariance



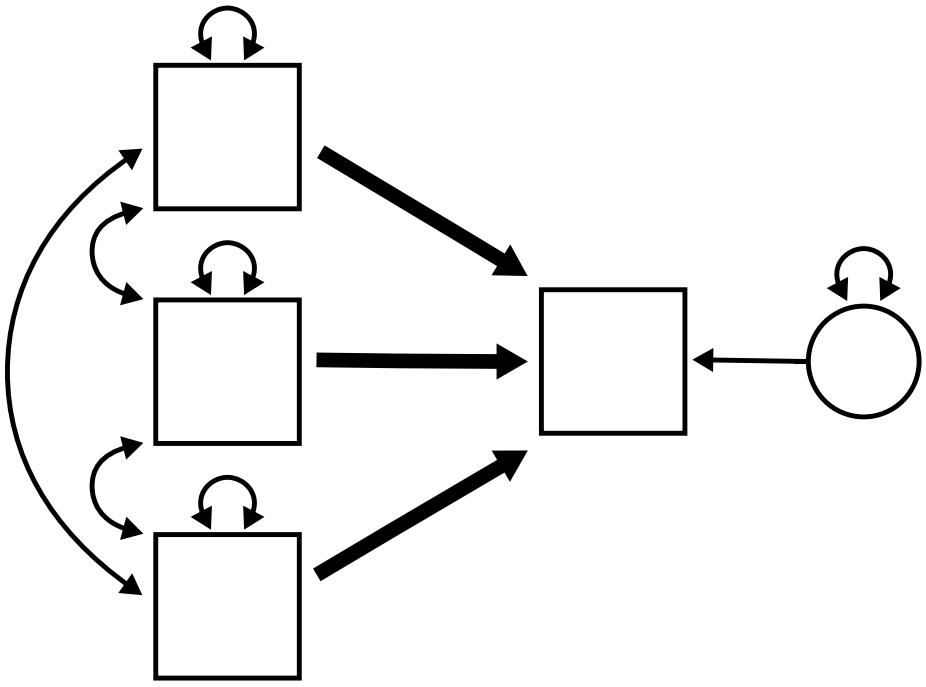
Chapter 4

Simple regression



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

LISREL notation