

Structural Equation Modeling

Ross Larsen

Table of Contents

Table of Contents	1
Foreword	3
Chapter 1. Introduction to Structural Equation Modeling	5
Chapter 2. Center and Spread	7
Chapter 3. Type of Data, Distributions, Graphs	13
Chapter 4. Covariance and Correlation	19
Chapter 5. Directionality and Causality	25
Chapter 6. Standard Errors and p-values	31
Chapter 7. Linear Regression	37
Appendix A. Introduction to SPSS	47
Appendix B. Introduction to AMOS	53
Appendix C. Common Formulas	57



Ross Larsen

Brigham Young University

Dr. Ross Larsen is an Associate Professor of Instructional Psychology & Technology at Brigham Young University.



This content is provided to you freely by EdTech Books.

Access it online or download it at <https://edtechbooks.org/sem>.

Table of Contents

Structural [a] Equation Modeling: An Introduction (stats is a language).....	2 [b] [c] [d]
Chapter 2: Center and Spread.....	10
Univariate Data Characteristics: Mean, Median, Mode, Standard [e] [f] deviation, Variance, Range, Min, Max .	15
Measures of association: Covariance, Correlation.....	20
Structural Equation Modeling Diagrams.....	25
Causality.....	30
Standard errors, and p-values	35
Linear Regression	40
More Practice	50
Sample projects to practice your skills	60
Common Pitfalls.....	65 [g]
Just a taste of more advanced statistical models that you can learn in [h] future courses.....	70 [i] [j]

[\[a\]](#) I think the table of contents doesn't have all the contents listed? Also, if this is an introductory course for people with no background in statistics, the chapter titles should be something relatable and accessible to that audience. These are suggestions, and I barely know what the current chapter titles are talking about, so these might be dumb examples. Just take in the principle, please.

Chapter 1: Statistics is a language that describes relationships.

Chapter 2: Statistics gathers data about a group, but not the whole group, because that would be too much data.

Chapter 3: Statistics represents that data in graphs and charts.

Chapter 4: Statistics compares sets of data to form conclusions.

Chapter 5: Statistics can show us if one thing causes another thing.

[\[b\]](#) I really want a clear vision of what context you want this book to be used in. I know that you want it to be a free source textbook and that is super useful. Do they do this by themselves? With a teacher? Is it for a full year course? A half year course? This will help you know how slowly or how quickly to take the different topics. It will give you an idea of how long the book should be also. You can add more examples or less examples, more practice or less. You can design a dbi things for it too that will really help scaffold the information. It will help a lot when you are writing the chapters to have all these thoughts in mind.

[\[c\]](#) As a rough draft how about we assume a regular one semester University statistics class. The students will read one chapter and do the exercises for that chapter before each class? Then you will know roughly how much material to put

into each chapter. Because you don't want it to be too much work or too little. Also, you will say there are usually this many classes in the semester so this is about how many chapters you want to have.

You will want the chapters to explain the topic well enough that students can work the exercises on their own after reading and they can go over sticking points or difficulties with their teacher in class. How does that feel?

That's one way to organize the book but certainly not the only way, and maybe not even the best way. What are your thoughts?

[d]@tatilarsen@gmail.com I love your comments and insights on this document. You ask fantastic questions about the learners and learning environment. Have you considered doing (or have you done) a master's in IP&T? You talk about DBL and scaffolding like a professional, so maybe you have.

[e]I think you should make this into at least 3 chapters. You can still call the chapters univariate data characteristics: mean mode and median. This would help the students understand the organization. It would even be cool to make as dbl style decision tree that shows all of statistics and you could zoom in on a certain part of the tree for each chapter in a graphic. This could show them how SEM really explains all of statistics together in one.

[f]Another great comment!

[g]Appendix: Introduction to SPSS

[h]Another thought for when the book is done, it wouldn't hurt to do a little bit of advertising for it. Introduce it to all your stats professors and collaborators, send it around to the UVU and BYU stats departments. BYU pathways might be interested. You could also send it to different high school teachers. A little work here might benefit a lot of people.

[i]I want this chapter to get them excited about all the cool things they can learn to do to solve different problems with their data. This would be best if they are problems that have come up in previous chapters and you said you can't do that or you weren't ready to address it. This could be a cool little appendix or you could do a whole chapter with examples and exercises depending on how many of these trouble shooting things there are.

[j]This is a good idea and I don't know how to wrap my head around it. What can get people excited about all the cool things they can do with statistics when those people are absolute novices in statistics, and possibly even a little afraid of it?



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/table_of_contents.

Foreword

To provide the reader with a unifying structure to understand statistics, data analysis, and inference that spans the enormity of the science in a way that is accessible to the layman.

Intended audience: graduate students who have not been exposed to statistics

Learning Objectives

- How to describe one variable (Univariate Descriptives)
 - Describe data distributions via graphical (e.g., histograms) and summary statistics (e.g., mean, standard deviation) techniques that evaluate center and spread.
- How to describe how two variables relate (Bivariate associations)
 - Describe bivariate relationships via graphical (e.g., scatterplots) and summary statistics (e.g., covariance, correlation) techniques that evaluate the strength of linear associations.
- How to construct graphics to show relationships between variables (Structural Equation Modeling Diagrams)
 - Show hypothesized causal and associative relationships between variables through accepted SEM diagramming techniques.
- Sampling Distributions
 - Describe the process of the creation of sampling distributions and describe their distributional properties (e.g., spread or standard error).
- Statistical language fluency
 - Demonstrate mastery of statistical terminology (e.g., p-value, standard errors) through the application and the implication on the stories we can tell from data.
- Regression Betas
 - Recite from memory the definition of an unstandardized and standardized beta in a linear regression context and apply its definition in understanding its implication in data analysis.
- Relate to traditional statistics
 - Demonstrate mastery of how SEM techniques are equivalent to traditional statistical methodology (e.g., ANOVA, t-test, etc).





This content is provided to you freely by EdTech Books.

Access it online or download it at <https://edtechbooks.org/sem/purpose>.

Chapter 1

Introduction to Structural Equation Modeling

Modeling

Structural Equation Modeling

Learning Outcomes

By the end of this chapter, students will be able to:

- Identify the definitions of the chapter vocabulary.
- Summarize an overview of structural equation modeling.
- Draw, label, and explain a two variable path diagram based on a story problem.

We are writing this book to graduate students everywhere who have no (or little) experience with statistics. We will address the reader as "you." We are writing to you.

Our purpose in this book is to change the way you think about statistics. If you are like many new graduate students, you have been dreading this class. And, given the way that most statistics classes teach the subject, we don't blame you. Somehow those classes make statistics seem impossible. It is not impossible. It is not voodoo or abstract nonsense.

Statistics describe relationships. It tries to predict how one thing affects another thing.

Structural equation modeling (SEM) is a branch of statistics that uses diagrams and numbers to describe how one thing predicts something else. Users of SEM search for truth in a way that is organized and **quantifiable**. (That means that it can be measured with numbers.) Throughout this book, you will learn how to gather, organize, and analyze **data** that will inform and transform the world. You will come to understand and enjoy statistics and the stories it tells through numbers.

Let's say that you want to know the *heart rate* of people from Mars. (*Heart rate* is an example of a **variable**. We will italicize the variables in this chapter to help you identify them.) Because you are looking for an answer to a question, you are a researcher. The answer that you are looking for is a number.

You ask a nurse to measure the *heart rate* of people from Mars. (All the people from Mars is called the **population**.) But you can't test everyone in the population, because that is thousands of people. That would take a lot of time. Instead,

you can test a group of people from Mars. (This is called a **sample**.) If you choose a good sample group, their information represents all of the people from Mars.

So you get a group of people from Mars and the nurse measures the *heart rate* of each person in the sample group. The *heart rate* for an individual is a number. You now have a set of numbers that tells you the *heart rate* of the individuals in this group. (This is called a **data set**.)

That is nice information. But now you want to know more. You want to know how *gender* predicts *heart rate*. Then you want to know how *gender* and *age* predict *heart rate*. And then you want to know how *gender* and *age* and *ethnicity* predict *heart rate*.

All of these predictive relationships between variables can be described using structural equation modeling (SEM). You take your data, organize it, and then analyze it to figure out the story the numbers tell. When done correctly, the data shows what is really happening. You will know how *gender*, *age*, and *ethnicity* predict the *heart rate* of a person from Mars.

This module explains and shows examples.

MODULE

Let's review and practice what you have learned so far.

Vocabulary Check

Practice Set 1

Practice Set 2

Learning Outcomes Check

I can identify the definitions of the chapter vocabulary.

I can summarize an overview of structural equation modeling.

I can draw, label, and explain a two variable path diagram based on a story problem.



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/chapter_1_introducti.

Chapter 2

Center and Spread

In this chapter you will learn how to:

- input and read data into SPSS,
- have SPSS produce measures of the center of a dataset (e.g., mean, median),
- and have SPSS produce a variety of spread estimates (e.g., range, standard deviation, variance, max, min).

[\[a\]](#)

Inputting the data into SPSS

In a hypothetical classroom of 3 students you decide to study their test scores. One way is to create a table in Excel or even on paper as shown below in figure 2.1.

Figure 2.1

Hypothetical test scores from a classroom of 3 students.

Name	Scores
Mike	76
Amy	85
Jane	98

Start entering this data into SPSS by opening the program. Start a new project by clicking File→New→Data. [\[b\]](#)

Figure 2.2

Graphic User interface (GUI) for starting a new project in SPSS.

Then select “variable view” from the tabs at the bottom of the screen.

Figure 2.3

Graphic User Interface (GUI) for the variable view in SPSS.

The first column is called “name”, enter the names of the two columns seen in Figure 2.1 in the first two rows. These columns are called “variables” in statistics. Click in the 2nd column which is called “type” in the first row that has the word “numeric” in it.

Figure 2.4

Graphic User Interface (GUI) for variable type in SPSS.

A box with 3 dots will appear. Click on it and select the 'string' option from the menu.

Figure 2.5

Graphic User Interface (GUI) for variable type in SPSS.

This allows you to put the names in that column that use letters. Now the program knows that these are not numbers. Leave the 2nd row as "numeric" in this field.

Figure 2.6

Screenshot of SPSS screen after the creation of two variables.

Now click on "data view" from the tabs at the bottom of the screen and in rows 1-3 recreate the names and scores in Figure 2.1.

Figure 2.7

Screenshot of SPSS after input of data from Figure 2.1.

Calculating the Center

One of the primary ways to describe some data is to calculate where the center of the data is. In statistics the two most popular measures of center are the mean and the median. The mean is the arithmetic average, while the median is the number in the center. Meaning half of the other data points lie below the median and half lie above the median [\[6\]](#). In SPSS this can be calculated easily. In SPSS click on the 'analyze' menu on the top of the screen. Click on the 'descriptive statistics' option and then click on 'frequencies'.

Figure 2.8

Graphic User Interface (GUI) for frequencies in SPSS part one.

Now move 'Scores' from the list of the variables in the left box to the 'Variable(s)' box.

Figure 2.9

Graphic User Interface (GUI) for frequencies in SPSS part two.

Now click on the 'statistics' box on the right.

Figure 2.10

Graphic User Interface (GUI) for frequencies in SPSS part three.

On the right-hand side there is submenu called 'Central Tendency'. Check the 'Mean' and 'Median' boxes. Click the 'Continue' button at the bottom of the GUI.

Figure 2.11

Graphic User Interface (GUI) for frequencies in SPSS part four.

Now click 'Ok'. SPSS will bring up a new 'Output' window and you should be see what is shown in figure 2.12.

Figure 2.12

Mean and Median from classroom of 3 students.

In Figure 2.12 you see SPSS has produced the arithmetic average called the 'mean' (86.3333) and the center data point called the median (85.0000). In this case the mean and median are almost the same. Let's take another hypothetical classroom. We will replace the student 'Mike' with the student 'George'. Go ahead and input this new dataset into SPSS and recalculate the mean and median [\[d\]](#)

Figure 2.13

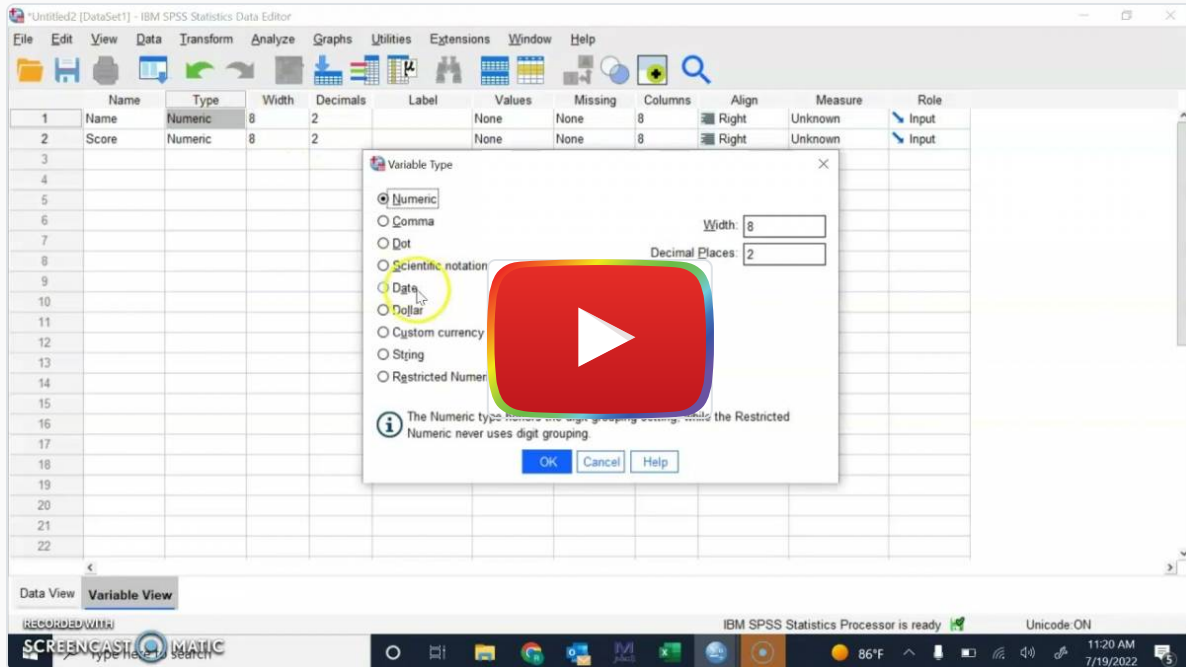
Hypothetical test scores from a classroom of 3 students with Mike replace by George

Name	Scores
George	10
Amy	85
Jane	98

Figure 2.14

Mean and Median from the classroom of 3 students where Mike is replaced by George.

Video: Getting the Mean and Median in SPSS



[Watch on YouTube](#)

Notice how the mean and median is affected by replacing Mike with George. The mean is now 64.3333 while the Median stays the same value of 85.00. Now the two measures of center disagree. The arithmetic average is affected by the fact that George's score is so much lower than Amy's or Jane's. George's score is what is called an 'outlier'^[f]. Notice that median is not affected by the fact that George's score is an outlier. The median is called 'robust' to outliers, in other words it is not affected by outliers. In this case the median would be a truer measure of center than the mean as most of the students did score about an 85 with just one outlier who scored much lower. ^[g]If the mean and median are about the same, as they are in the class with Mike instead of George, the mean is used because of better mathematical properties of the mean as compared to the median.

Measures of Spread

Let's go back to our classroom with Mike as seen in Figure 2.1. Go through the same process you did in calculating the mean and the median but in addition to clicking the 'mean' and 'median' boxes, in 'Dispersion' click the 'Std. deviation', 'Minimum', 'Maximum', and 'Range' buttons. Click 'Continue' and 'Ok' like normal.

Figure 2.15

Graphic User Interface (GUI) for the statistics submenu in SPSS.

Figure 2.16

Mean, Median, Standard Deviation, Range, Minimum, and Maximum, from the classroom of 3 students with Mike

The minimum and maximum are self-explanatory, while the Range is the Max-Min. Standard deviation of 11.06044 is the average distance from the mean. Please memorize this. The standard deviation is the average distance from the mean. [b]SPSS chooses its own decimal places sometimes, if your output looks different than mine in the decimal places don't worry. It is calculated thusly:

$$\text{Standard Deviation} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Let's break down the numerator of this equation: $\sum (x_i - \bar{x})^2$, x_i refers to each data point in the dataset, \bar{x} , refers to the mean, and Σ is the capital greek letter sigma which means to sum the results. Thus, in our dataset we already have the mean calculated (86.3333) from figure 2.3. Let's do the operation for each of our data points:

$$(76 - 86.3333)^2 = (-10.3333)^2 = 106.7777$$

$$(85 - 86.3333)^2 = (-1.3333)^2 = 1.7778$$

$$(98 - 86.3333)^2 = (11.6667)^2 = 136.1112$$

Now we do the Σ operation and add these three numbers together to get our numerator (106.7777 + 1.7778 + 136.1112)=244.6667.

The denominator has n which refers to the number of data points we have, in this case 3. Subtract 1 from 3 to get 2 and divide the numerator (244.6667) by the denominator (2) to get (244.6667/2)=122.3333. Referring back to the equation we have to take the square root of this number $\sqrt{122.3333}$ which results in 11.06044. This is just what SPSS tells us the standard deviation is. If you look at the dataset this makes sense as Mike's score of 76 and Jane's score of 98 are about 11 points away from the mean of 86.3333.

Let's go back to the classroom shown in figure 2.14 with George and recalculate our measures of spread.

Figure 2.17

Mean, Median, Standard Deviation, Range, Minimum, and Maximum, from the classroom of 3 students with George

Notice in this dataset the standard deviation is 47.50088. Just like the mean is affected by George, the outlier, the standard deviation is also not robust to outliers.

Instructions on how to make a youtube video [i]

<https://edtechbooks.org/-sgPP>

Exercises

1. Without notes write down the standard deviation formula. [Answer](#)
2. By hand, calculate the standard deviation of this dataset (1,2,3). Show your work. [Answer](#)
3. In your own words, describe when you would use the median rather than the mean to describe the center of the data. [Answer](#)

[a]Needs a motivating example.

[b] I think this can be embedded in the Ed Tech book? Or linked? Something easy.

[c] This is an important definition and is buried in this paragraph. Maybe definitions could be in callout boxes? Have big, red arrows pointing to them? Something that indicates this is an important definition--remember it.

[d] This part seems sort of video-y, like it would be good as a video.

[e] More videos later on.

[f] I just realized that we didn't define outlier very explicitly.

[g] This is another important point, and it is buried in the paragraph. Important points need to be visually emphasized.

[h] Good definitions and important information. This should be more important visually so the reader knows they are important.

[i] Introduce AMOS



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/center_and_spread.

Chapter 3

Type of Data, Distributions, Graphs

Up to this point we have been working with continuous data. Continuous data can be any value within a range. Good examples of continuous data are blood pressure, height, and weight. Categorical data, on the other hand, can only take specific values that define certain categories. For example, fruit. In SPSS it is possible to have a variable called fruit with 1=apple, 2=banana, 3=orange, etc. In this context a 1.5 would make no sense. It makes more sense to think of these numbers separately as different categories that a fruit (or something else) can fall into. Contrast this with height. A person who is 65 inches tall is just a bit shorter than a 65.5 inch tall person, who is in turn also just a bit shorter than a 66 inch tall person. This type of data gradually builds without abrupt breaks, so it is called continuous data. The ways we handle data depend heavily on what type of data it is, continuous or categorical. This next section we will learn how to show continuous data graphically.

Histograms

One of the most commonly used ways to graphically show continuous data is through histograms.

Figure 3.1 shows the histogram of how positive the climate was in a kindergarten classroom, as recorded by an outside observer. It is produced by SPSS.

[\[a\]](#)

Figure 3.1

Positive Climate histogram

The histogram shows the distribution of scores. A distribution is defined as all the possible values a variable can take on and how often those values occur. This histogram has, on the x-axis the possible values (and more) that positive climate can take on and the bars show how often those values occur on the y-axis. The bars only appear for values for 3 through 7 with most of the values appearing in the '6' bar. That means that no outside observer recorded the Kindergarten classroom had a climate of 2 or 8. These bars, also called bins, are ranges of values that SPSS combines for easier presentation. Thus the 6 bar or bin represents all the values from 6.0 to 6.9999. SPSS also provides helpful values in the top right hand of the histogram graphic. Some of these values are the mean, the standard deviation, and n which refers to the number of data points in your study. You can see these numbers in the top right hand of the graphic.

Different shapes of a Distribution

The histogram in 3.1 is just one example of a shape a distribution can take on. Another example is found in figure 3.2

Figure 3.2

Example of a left skewed distribution

A left skewed distribution like seen in Figure 3.2 means most of the data is clustered towards high values of the variable with a 'tail' to the left. The downward slope of the data on the left hand side sort of resembles a tail and is often referred to this way.^{[b][g]} Remember, left skewed means the tail is on the left. Notice how the Mean is drawn towards the tail more than the Median. This is similar to the behavior of the Mean and Median in the presence of outliers^[d] like we discussed in the previous chapter. Similar to that case, the Median is considered a more robust measure of center in this case than the Mean. The Mode is the value that is repeated the most in the dataset and is generally not used as a measure of center.

Figure 3.3

Example of right skewed distribution

Figure 3.4

Example of symmetric distribution with no skew

Figures 3.3 and 3.4 show other examples of shapes of a distribution, where figure 3.3 shows a right skewed distribution with its tail on the right.^[e] Figure 3.4 shows a symmetric distribution with no skew where the mean, median and mode are all equal. Figure 3.4 shows a special distribution called a 'normal' or 'gaussian' distribution it is a symmetrical and bell-shaped. This type of distribution occurs frequently.

Your turn

Click on this link to download this .sav (which is an SPSS datafile) file.

[Example dataset 3.1](#)

Once you have downloaded this dataset open it in SPSS, go to the variable view and you should see the following as shown in figure 3.5

Figure 3.5

Screenshot of example dataset 3.1

Now go to Analyze→Descriptive Statistics→Frequencies and move the third variable 'wk continuous' into the variables box, click on charts and select histograms and click continue.

Figure 3.6

Graphic User Interface (GUI) for frequencies in SPSS.

Figure 3.7

Graphic User Interface (GUI) for the charts submenu in SPSS.

Now uncheck the 'Display frequency tables' box on the lower left side and click 'Ok.' This will produce the output seen in Figure 3.8.

Figure 3.8

Histogram from example dataset.

This is a measure of the childrens 'SES' measure or socioeconomic status which is a measure of wealth. How would you describe the shape of this distribution? Would you use the mean or median to describe [\[f\]\[g\]](#) its center?

Bar Charts

Now you will produce a 'bar chart', which is a way to describe categorical data [\[h\]](#). Just like we saw in Figure 3.1, a bar chart is a representation of how often different values occur in your data. This is a great tool to be able to tell at a glance the general story of the data. Go back to example dataset 3.1 and repeat the process of Analyze→Descriptive Statistics→Frequencies but this time pick the second variable instead of the third that starts with 'Child Composite Race' [\[i\]](#). Depending on the width, you may only see the words "Child Composite." Go to Charts, but instead of selecting histogram select bar chart. You will produce the output seen in Figure 3.9. [\[j\]](#)

[\[k\]](#)

Figure 3.9

Selected output of Bar Chart produced from Example data 3.1 of student race.

An important distinction between a bar chart and a histogram can be seen in the x-axis. Instead of numbers put into bins you have categories. The order does not matter in a bar chart while it does matter in a histogram. In this case we can tell at a glance that the vast majority of the students in this dataset are 'White, non-hispanic' followed by a distant second of 'Black or African American, Non-hispanic'. The smallest category is 'Native Hawaiian, other Pacific Islander'.

Boxplots

Now we will combine both categorical and continuous data to produce Boxplots. Boxplots are a birds eye view of a histogram where you can compare different categories side-by-side. For example, refer to figure 3.9 which is the SES data shown in figure 3.6 but where the viewer of the data is above it.

[\[l\]](#)

Figure 3.10

SES data as seen in a boxplot rather than a histogram (figure 3.6).

The bottom line of the data is at -1.00. This horizontal line is called a 'whisker'. The line that connects the whisker to the blue box represents the lower 25% of the data. The blue box shows where the majority of the data or where the middle 50% of the data lie. The line in the middle of the blue box is the median. The last line up to the last whisker is the upper 25% of the data. Some of the data in the tail is determined to be outliers according to an arbitrary standard that SPSS uses. These are represented by dots. The numbers associated with those dots are the row numbers of those datapoints in the SPSS spreadsheet. This boxplot shows you that SES is right skewed.

Go to our example dataset 3.1 and go to Graphs→Legacy Dialogs→Box plots. You will see the following as shown in Figure 3.11

Figure 3.11

Graphic User interface (GUI) for boxplots in SPSS part one.

Don't change the defaults and click on 'Define'. In the next GUI put the 2nd variable "Child Composite Race" in 'Category axis' and the third variable "WK Continuous SES" in 'Variable' as seen in figure 3.9.

Figure 3.12

Graphic User interface (GUI) for boxplots in SPSS part two.

You will produce the output as shown in figure 3.13.

Figure 3.13

Boxplot of SES vs. Race from example dataset 3.1.

As mentioned previously, the boxplot is a birdseye view of a histogram set side-by-side according to categories, in this case SES (wealth) vs. Race. From Figure 3.13 we can tell the 'Asian' group has the highest median wealth followed by 'White, non-hispanic' and all the races SES are right skewed.

[a] Put in SES.

[b] I felt the reference to tail wasn't clear so I added this sentence. Not certain if this helped.

[c] This might be better done with an arrow labeling the tail in the graphic than with my not-overly clear clarification sentence.

[d] I just realized we didn't define outlier explicitly in chapter 2. We have a good example, but with such a small data set, I'm not super confident that everyone will catch what it means.

[e] To me it feels counter intuitive that the tail on the right means right skewed when the preponderance of data is on the left, that's why I'm pointing this out in both of the skewed graphs. Don't include this if you don't like it.

[f] right skewed, median

[g] Can give them the answer if you want.

[h] Might be helpful to discuss why you would want to do, like what it's useful for, or what it means, before jumping in to doing it.

[i] you wanted race not gender right?

[j] Tell them to click frequency tables on if you want the exact chart you have to come up.

[k] I got the same blue bars, but not the chart of numbers above it.

[l] Show how to get this in SPSS



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/data_distributions_graphs.

Chapter 4

Covariance and Correlation

Chapter 4: Covariance and Correlation

A great way to understand how two continuous variables relate is through a scatterplot. A scatterplot shows one of the variables on the y-axis and one on the x-axis. Lets take for example, the continuous variables height and weight. Height is on the x-axis on weight is on the y-axis. [\[a\]](#)

Figure 4.1. Dataset of height vs. weight. n=3.

To create this graphically, points are created where the values are for each individual, so in this case there would be three points at (60,150), (62,175), and (63,170).

Figure 4.2. Scatterplot of height vs. weight. n=3.

Notice the x and y-axes are clearly labeled to show which variable they represent. The user can tell at a glance, even with a dataset this small, that the relationship between height and weight is positive, meaning as height increases, so does weight, in general. Create the dataset in SPSS and create the scatterplot via Graphs → Legacy Dialogues → Scatter/Dot → Simple Scatter.

Now click define. Next put "Weight" in the 'Y Axis' and "Height" in the 'X Axis' and press OK. Your scatterplot should look like Figure 4.2.

Covariance

In the scatterplot above we could visually see that as height increased in our data set, weight also increased. This made our points into a shape approximating a line. We call this type of association between our x and y variables a linear association. Because datasets are more complicated than the one above, it is useful to have a number to summarize the strength of the linear association between x and y. A good way to visually determine this linear association is to draw an oval around the points on your scatter plot. The longer and skinnier the oval is, the stronger the linear association is. One [\[b\]](#) such measure is the covariance, which is defined in the following formula

Where x_i and y_i are each individual data point for x and y, \bar{x} and \bar{y} are the means of x and y, and n is the sample size. For example x could be height and y could be weight. Therefore \bar{x} would be the mean of height and \bar{y} would be the mean of weight for the data set.

$$-\infty < \text{Covariance}(x,y) < \infty$$

Covariance can go as low as negative infinity and as high as positive infinity, with a 0 value signifying no linear association between x and y. In the case of our data of height and weight found in figure 4.1 the covariance is 17.5, which shows a positive relationship but does not tell us much about the strength of the relationship between height and weight. In order to tell more about the strength we use the Pearson correlation coefficient.

Pearson Correlation Coefficient

The Pearson correlation coefficient is the covariance of a pair of variables but it is standardized. Instead of going from $-\infty$ to ∞ like covariance, Pearson correlation goes just from -1 to 1.

$$-1 < r_{xy} < 1$$

Here is what it looks like in equation form. Pearson correlation between x and y is generally expressed as r_{xy} .

$$r_{xy} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}$$

Where σ_x and σ_y are the standard deviations of x and y. Now the bounds of the Pearson correlation coefficient are -1 and +1.

Where a Pearson correlation coefficient of -1 is a perfect negative linear association. This relationship can be seen in Figure 4.3. In this type of correlation as your variable y gets smaller, your x gets larger. A Pearson correlation coefficient of 0 means there is no linear association between x and y as can be seen in Figure 4.4. A Pearson correlation coefficient of 1 means there is a perfect positive linear association as can be seen in Figure 4.5. This means that as y increases, x also increases.

Figure 4.3. Scatterplot of a perfect negative Pearson correlation coefficient ($r_{xy}=-1$).

Figure 4.4. Scatterplot of a relationship where there is no linear association between x and y ($r_{xy}=0$).

Figure 4.5. Scatterplot of a perfect positive linear association ($r_{xy}=1$).

Using the Pearson correlation coefficient will allow us to tell the magnitude of the strength of the association between x and y.^[4] While magnitude cutoffs are arbitrary generally it is regarded that an $r_{xy} > |.8|$ is considered a strong correlation. An r_{xy} of $|.5|$ is considered a moderately strong correlation, and an $r_{xy} < |.2|$ is considered a weak correlation. These cutoffs will vary by discipline, with the harder sciences such as chemistry or physics a relationship can be considered strong if it is $> .9$ while in education a correlation of $.5$ would be considered very strong. A good way to visually determine the correlation is to draw an oval around the points on your scatter plot. The longer and skinnier the oval is, the stronger the correlation is. Figure 4.6 has a weak correlation relationship between x and y, while Figure 4.7 has a strong correlation relationship. For the height and weight dataset the correlation is $.866$ signifying a strong relationship between height and weight. Calculate the Pearson correlation coefficient yourself via Analyze→Correlate→Bivariate Correlation. Put both height and weight into the "Variables" box on the left and click OK.

Your turn.^[d]

Click on this link to download this .sav (which is an SPSS datafile) file.

[Example dataset 4.1.sav^{\[e\]}](#)

Once you have downloaded this dataset, open it in SPSS, go to the variable view, and you should see the following as shown in figure 4.6.

Figure 4.6. Screenshot of example dataset 4.1

Now go to Graphs → Legacy Dialogues → Scatter/Dot → Simple Scatter and click 'Define'. Move the first variable "Age in years" to the 'X Axis' box and the second variable "Number of books..." to the 'Y Axis' box, as seen figure 4.7. Now click 'OK'.

Figure 4.7. Graphic User interface (GUI) for simple scatterplot in SPSS.

You will produce the output shown in figure 4.8.

Figure 4.8. Scatterplot of age vs. number of books read in past year. n=20

Now, draw a circle around the points in your scatterplot. How would you describe the correlation between these variables?

Repeat the steps you just used with the variable "Hours spent on social..." on the 'X Axis' and variable "Depression score..." on the 'Y Axis'. Your scatterplot will look like Figure 4.9.

Figure 4.9. Scatterplot of hours spent on social media per week vs. depression score on the DASS-21. n=20

How is this scatterplot different than the scatterplot in Figure 4.8? What does the shape of the oval say about the correlation?

Now, repeat the steps using variable "Depression score..." on the 'X Axis' and "Anxiety score..." on the 'Y Axis'. Your scatterplot will look like Figure 4.10.

Figure 4.10. Scatterplot of depression score on the DASS-21 vs. Anxiety score on the DASS-21. n=20

How does this scatterplot compare to the previous two? What does the shape of the oval say about the correlation?

Now go to Analyze → Correlate → Bivariate, as seen in figure 4.11.

Figure 4.11. Graphic User Interface (GUI) for bivariate correlation in SPSS.

Move variables "Age in years" and "Number of books..." to the 'Variables' box and select 'Show only the lower triangle', as shown in Figure 4.12. Now click 'OK'.

Figure 4.12. Graphic User Interface (GUI) for bivariate correlation in SPSS.

Your correlation table will look like Figure 4.13. What does this correlation say about the relationship between age and the number of books read in the past year?

Figure 4.13. Correlation table for age in years vs. number of books read in past year.

Now, create a correlation table for the variables "Hours spent on social..." and "Depression score...". Your correlation table will look like Figure 4.14.

Figure 4.14. Correlation table for hours spent on social media per week and depression score on the DASS-21.

What kind of correlation is present between these two variables? Why might that be? How does it differ from the previous example? Finally, create a correlation table for the variables "Depression score..." and "Anxiety score...". Your correlation table will look like Figure 4.15.

Figure 4.15. Correlation table for depression score on the DASS-21 and anxiety score on the DASS-21.

What is the strength of this correlation? Does it make sense for that kind of relationship to exist between the variables?

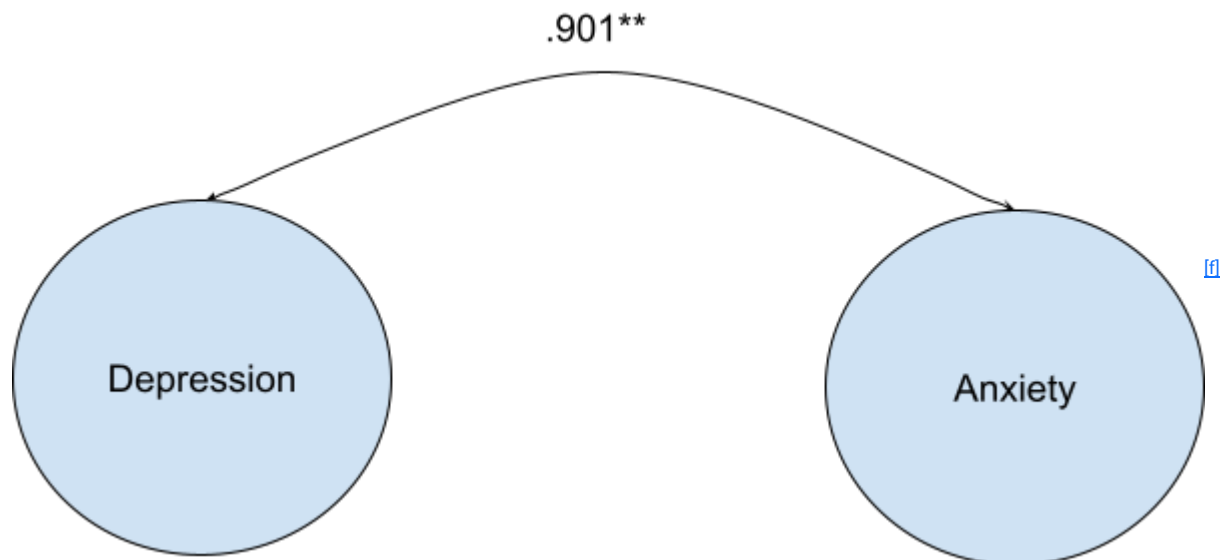


Figure 4.16. Structural equation modeling (SEM) diagram of the correlation between anxiety and depression on the DASS-21.

Exercises

You can use the following datasets to practice creating correlations in SPSS. The variable names you should use are in *italics*. The independent variable (x) is listed first, followed by the dependent variable (y). For each item:

1. Create a scatterplot of the two variables
2. Find the Pearson correlation
3. Characterize the correlation (strong/weak, positive/negative)

Practice problems ([Answer key](#))

1. Use this [dataset](#) and find the Pearson correlation between the diamonds' weight (Carat_Weight) and price (Price).
2. Use this [dataset](#) to find the Pearson correlation between the cars' curb weight (curbweight) and miles per gallon in a city (citympg).
3. Use this [dataset](#) to find the Pearson correlation between the cars' height in inches (height) and horsepower (horsepower).
4. Use this [dataset](#) to find the Pearson correlation between the cars' engines' peak revolutions per minute (peakrpm) and price in dollars (price).
5. Use this [dataset](#) to find the Pearson correlation between the irises' sepal length (sepal_length) and petal length (petal_length).
6. Use this [dataset](#) to find the Pearson correlation between the air temperature in Celsius (temp) and relative humidity in percent (RH).
7. Use this [dataset](#) to find the Pearson correlation between the wind speed in km/hour (wind) and rain in millimeters (rain).

Next, you will find these Pearson correlations in AMOS. These [instructions](#) will walk you through how to do this using practice problem #1. ([AMOS output for practice problems 2 to 7](#)).

[a] Abrupt beginning

[b] Parabola, $r=0$

[c] This should be made consistent. Either capital or lowercase

[d] My assignment: Simulate 3 datasets, create the scatterplots, draw the oval, calculate the correlation.

[e] When I first clicked on this link I got an error and had to push a bunch of buttons. Then it didn't work. So I closed down all my spss windows and tried again after which the link downloaded and opened perfectly. Not sure what happened. Just thought I'd mention it.

[f] Introduce first AMOS example. We actually have the data, repeat in AMOS.



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/covariance_correlation.

Chapter 5

Directionality and Causality

Chapter 5: Directionality and Causality

Last chapter we introduced structural equation modeling diagrams (SEM) where we showed you how to draw a picture showing the correlation between depression and anxiety (see figure 4.16). In order to draw more complicated diagrams we need to discuss directionality and causality. Notice in Figure 4.16 that the curved line has an arrowhead pointing to both depression and anxiety.^{[a][b]} This type of model that involves just a correlation makes no causal or directional claims between the two variables. It could be that depression causes anxiety, or anxiety causes depression, both of those would be causal or directional claims. Causal and directional claims in this context are synonyms. It could also be that depression and anxiety both do not cause the other and there is a third variable that is not included in the model that causes both. In other words, the model represented in Figure 4.16 makes no causal or directional claims about any of the variables involved. In order to make causal claims we introduce a new diagramming technique and discuss the theoretical justification that needs to be included to justify causal or directional claims..

Drawing Causality or Directionality

Figure 5.1 shows a causal relationship between x and y.

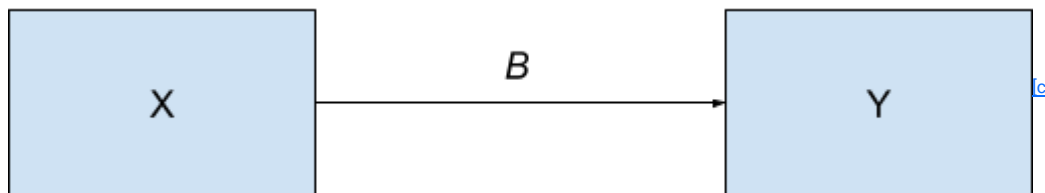


Figure 5.1. An SEM diagram showing a causal or directional relationship between x and y. The strength of the relationship is shown by β .

This model, unlike the model shown in Figure 4.16, makes a strong causal or directional claim from x to y. If you want to change y you can directly influence it by changing x. Note, the only important aspect of this diagram is which way the arrow points. The position of the box that contains x being on the left and the box that contains y being on the right is arbitrary and could be switched with no change of meaning.^[d]

A real world example would be hours of reading a day (x) affects a child's reading fluency (y) as measured by a standardized test as shown in Figure 5.2.

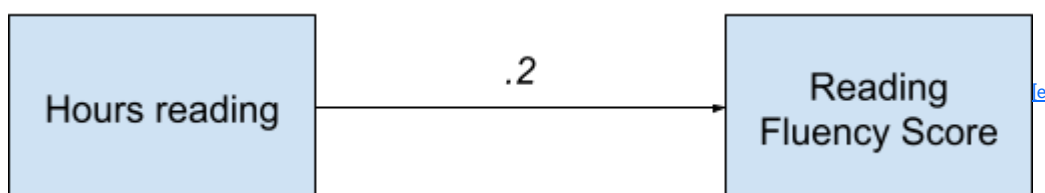


Figure 5.2. An SEM diagram showing a causal relationship between hours reading and reading fluency score.

In models like Figures 5.1 and 5.2, hours reading (x) will always increase by 1 unit in its interpretation while the $.2(\beta)$ will be estimated (or calculated from the data, see Chap X). Thus, Figure 5.2 is interpreted as: for every 1 hour increase in hours reading there is a .2 predicted increase in reading fluency scores. The predicted increase in reading fluency score is caused (a causal or directional relationship) by the increase in hours reading. Note that in this model the line connecting hours reading and reading fluency score is straight and the arrowhead only points towards reading fluency score. The straight line and the arrowhead, show that a causal relationship between hours reading and reading fluency score is being assumed in the model. This type of model can be very useful to the applied researcher as it shows where to intervene in order to achieve an outcome, in this case encourage more hours reading for the child in order to increase reading fluency score.

Justifying Causality or Directionality

In order to justify the validity of a model like the one shown in Figure 5.2, it needs to be theoretical backing. The study of causality or directionality has occupied both philosophers and data analysts for centuries, and there is broad literature covering the topic. Nevertheless, the main ideas of causality or directionality we need consider can be summarized into these three conditions:

1. A correlation must exist between x and y .
2. There must be temporal precedence of the change in x before the change in y .
3. The relationship is non-spurious (meaning it cannot be explained any other way).

Condition 1: A correlation must exist between x and y .

In order for a researcher to argue that x causes a change in y there must be some kind of association between the two that can be measured. In chapter 4, we introduced the idea of the Pearson correlation coefficient as a good measure of linear association between two variables. So, assuming the relationship is linear and has no curves, it is easy to check condition 1 (A correlation must exist between x and y). You do this by running the bivariate correlation procedure in SPSS as you did in chapter 4. If the Pearson correlation coefficient is important (meaning both statistically significant and big enough to be meaningful) then condition 1 is satisfied. In Figure 4.14 and Figure 4.15 we do just that, we calculate the correlation between depression and anxiety. That correlation is statistically significant ($p < .001$), and is quite large ($r = .901$) and thus can be considered important.

Condition 2: There must be temporal precedence of the change in x before the change in y .

In order for directionality to be justified the change in x must come before the change in y . That is temporal precedence, when one event happens before another. In order for the model shown in Figure 5.2 (increased hours reading results in increased reading fluency score) to be valid the researcher cannot have measured hours reading after they measure reading fluency score. You cannot say that a change in hours reading in the future has affected reading fluency score in the past. This can pose a challenge in real world research as many times data is collected on all variables simultaneously. For example, the data collected to create Figure 4.16 is probably taken from a survey where the researcher asked questions about both the subjects' depression and anxiety. Thus, there is no way from the data to determine temporal precedence, instead the researcher must do this with logic and theoretical considerations.

Sometimes the relationships are bidirectional, that is, they both go from x to y and from y to x simultaneously. For example, for the model shown in Figure 5.2 (hours reading causes change in reading fluency score) it could be argued that additionally to the argument that x changes y , that y also changes x . Increasing reading fluency score could also result in a change in hours reading because a child with a higher mastery of reading as measured by their reading fluency score may read more hours because they enjoy it more. Longitudinal studies where x is gathered before y , especially across several time points, can help isolate temporal precedence. This can create an argument that is not only theoretical justified but also justified by empirical evidence.

Condition 3: The relationship is non-spurious.

In order for the argument to be made that x causes y , there cannot be any missing variables or alternative reasons why such a relationship exists. Let's take for example the model shown in Figure 5.3.

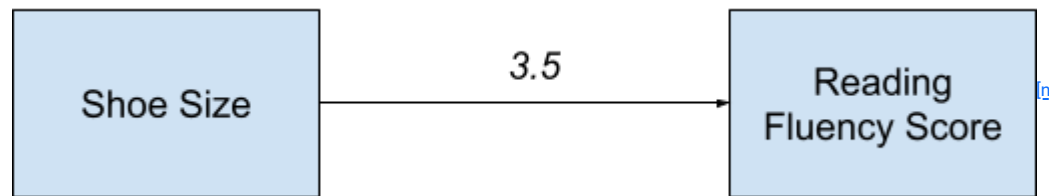


Figure 5.3. An SEM diagram showing a causal relationship between shoe size of a child and reading fluency score.

Figure 5.3 can be interpreted as for every 1 unit increase in shoe size, a predicted child's reading fluency score increases by 3.5 points. This relationship, on its face, seems ridiculous or, in more technical terms, lacks face validity. A researcher could argue that this relationship satisfies condition 1: A correlation must exist between x and y ; and also condition 2: There must be temporal precedence of the change in x before the change in y . They are correct. If you run a bivariate correlation between shoe size and reading fluency score you will indeed get a statistically significant correlation. It will also be a large correlation and therefore it can be judged as a correlation that is important. The researcher advocating for the causality of the model can also argue that shoe size was measured before reading fluency score therefore satisfying condition 2. They are again correct. Nevertheless, critics of the model can rightly point out that there is a missing variable in the model that affects both shoe size and reading fluency score, namely age, as can be seen in Figure 5.4.

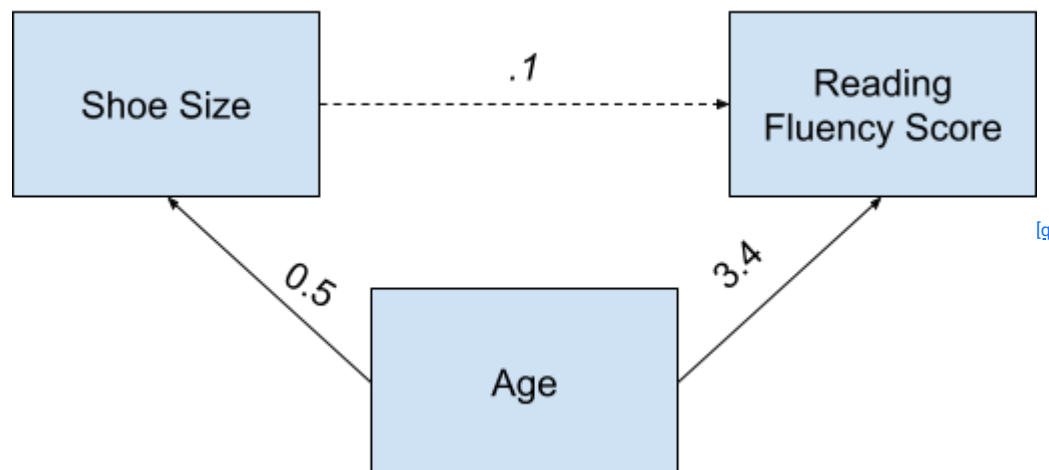


Figure 5.4. SEM diagram showing two causal relationships from age of child to both shoe size and reading fluency score.

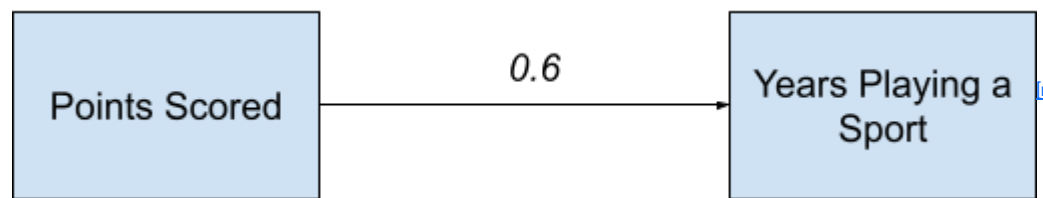
Figure 5.4 is interpreted as: (1) for every one year increase in age a predicted child's shoe size increases by 0.5 units, (2) for every one year increase in age controlling for shoe size the predicted reading fluency score for a child increases by 3.4 units, and (3) for every one unit increase in shoe size controlling for age the predicted reading fluency score increases by .1 units. Note that the directional arrow from shoe size to reading fluency score is dashed rather than being solid. This is frequently done in SEM diagrams to show that the relationship between x and y is statistically nonsignificant or could have easily have happened by chance alone. Also note, that the interpretation of the diagram changes if one of the squares has more than one arrow pointing to it. In this case reading fluency score has arrows from both shoe size and age pointing to it. Thus, the model takes into account that there are two possible causes of change in reading fluency scores. When we are taking into account two possible causes for the change, that is called controlling for.

Now, with this revised model shown in figure 5.4, critics can argue that the relationship between shoe size and reading fluency score is indeed not causal as there is no statistically significant association between shoe size and reading fluency score once the age of the child is accounted for. The previous model shown in Figure 5.3 would therefore be considered spurious. That model is actually very wrong, or in technical terms the model is misspecified. Age must be included in order for the model to be defensible. The real reason there seems to be a relationship between shoe size and reading fluency score is because students who have larger shoe sizes are older, and older students have higher reading fluency score. Age, in this context, would be called a confounding variable. A confounding variable affects both x and y and, in its presence, the relationship between x and y changes.

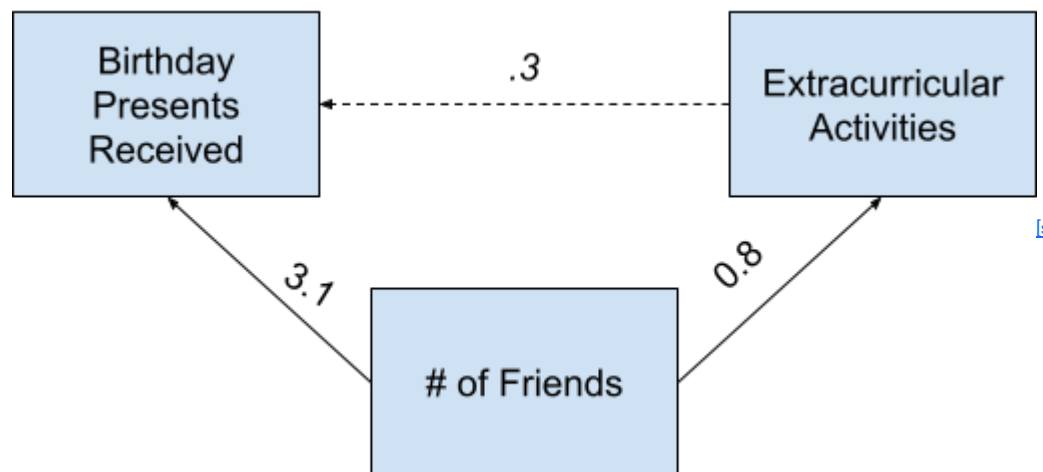
Exercises

Include several SEM diagrams of models and ask the students to evaluate the causal claims. Also create an answer key.

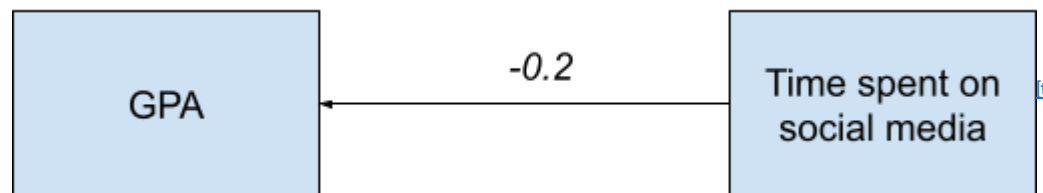
1. Describe the causal relationship between points scored and years playing a sport.



2. Describe the causal relationships from number of friends to both extracurricular activities and birthday presents received.



3. Describe the causal relationship between time spent on social media and GPA.



Answer Key

1. For every one unit increase in points scored, the predicted amount of years playing a sport increases by 0.6 years.
2. a) For every one unit increase in number of friends the predicted number of extracurricular activities increases by 0.8 units.

b) For every one unit increase in number of friends controlling for extracurricular activities the predicted number of birthday presents increases by 3.1 units.

c) For every one unit increase in extracurricular activities controlling for number of friends the predicted number of birthday presents received increases by 0.3 units.

3. For every one unit increase in time spent on social media, predicted GPA decreases by 0.2 points.

[a] Right before this sentence, I would start a new paragraph and define directionality and causality, then go into explaining Figure 4.16

[b] Or we could do the yellow box thing to the side with the definition.

[c] Fix the italicized B to the greek letter Beta in the final product. And get rid of the yellow background.

[d] Insert the backward image as an example.

[e] Fix the italicized B to the greek letter Beta in the final product. And get rid of the yellow background.

[f] Include a sidebox where a intuitive explanation of what a beta is (Beta is a slope, or a rate fo change, for every 1 unit change in Y.

[g] Abby, please make sure that all mentions of variables x and y are lower case and italicized.

[h] Done

[i] Abby, In chapter 4 can you fix all references to correlation so that they say, Pearson correlation coefficient.

[j] introduce this concept possibly in its own chapter. Chronologically before this chapter.

[k] I think this definition is helpful what do you think?

[l] These variables need to be italicized whenever they are mentioned.

[m] Put in a side box.

[n] Fix the italicized B to the greek letter Beta in the final product. And get rid of the yellow background.

[o] I think this should be shoe size, not shoes size. Minor thing...

[p] Do we want to say Pearson correlation?

[q] Fix the italicized B to the greek letter Beta in the final product. And get rid of the yellow background.

[r] Fix the italicized B to the greek letter Beta in the final product. And get rid of the yellow background.

[s] Fix the italicized B to the greek letter Beta in the final product. And get rid of the yellow background.

[t] Fix the italicized B to the greek letter Beta in the final product. And get rid of the yellow background.





This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/directionality_causality.

Chapter 6

Standard Errors and p-values

Chapter 6: Standard Errors and p-values

In order to make a case for causality it is necessary to make a case for correlation. The correlation must not be spurious, meaning it could not have occurred by chance alone. Statistics has a formal framework called Null Hypothesis Significance Testing (NHST) that helps determine whether or not a correlation is spurious. In order to use this framework researchers assume the null hypothesis. This assumption generally means that there is no correlation between two variables ($r = 0$ meaning correlation=0) [\[a\]](#). For this text, remember that the null hypothesis means that there is no correlation between two variables.

In some cases it is quite obvious whether the null hypothesis is true or not. Here is an example: a researcher wants to know something very narrow, say, what is the correlation between height and weight in a particular high school class. To answer the question a researcher would record every student's height and weight for that classroom in SPSS. Then they would run the bivariate correlation procedure as you have learned to do in chapter 4 and examine it. In our example, SPSS calculated the bivariate correlation to be quite small, say $r = .1$. [\[b\]](#) Because $.1$ does not equal 0, the null hypothesis is false, meaning there is a clear though small correlation between height and weight in this classroom. The vast majority of research questions are not so narrow. Usually research tries to take a sample of subjects and make inference [\[c\]](#) to a larger population. Take our above classroom study, to infer the correlation between height and weight in the general population, researchers would use the same process above. Assuming the sample classroom of students is representative of the larger high school population, researchers would calculate $r = .1$ and claim that the null hypothesis is false because $.1$ does not equal 0.

The example researcher would have a problem though, because r is small, they could be open to criticism by those who could say that my classroom correlation may just be a fluke. The correlation between height and weight in the general population may indeed be $r = 0$, and by chance alone this particular classroom happens to have a correlation of $r = .1$. To contradict this criticism, it would be great to know how often this sort of fluke could happen. How often, if the correlation really is 0, would a random sample like the one in our example produce $r = .1$ or larger? That question is exactly what a p-value answers. Please memorize this, a p-value is the probability, by chance alone, of getting results as extreme or more extreme than your results, assuming the null hypothesis.

Let's substitute the example's facts into this definition. This example's p-value is: the probability of getting a correlation of height and weight of $r = .1$ or larger, given that the true population correlation is $r = 0$.

Here is another example of a p-value. Figures 4.13 and Figure 5.1 both show the sample correlation of Age in years and Number of Books read in the Same Year. [\[d\]\[e\]](#)


Figure 6.1. Correlation table for age in years vs. number of books read in the past year.

The Pearson Correlation is $.198$, which is small, but could this small value have happened by chance alone given that the null hypothesis $r = 0$ is true? Notice the row below the $.198$ estimate. The label of this row is "Sig (2-tailed)", which is

short for statistical significance (2-tailed). The value in this row is .403. This value is the p-value for this hypothesis question. It is interpreted as: The probability of getting a correlation of .198 or larger (more extreme) if the null hypothesis of the population correlation of age and books is 0 is 40.3%. Generally, all p-values that are greater than 5% are insufficient evidence to reject the null hypothesis. Critics of this correlation can rightly say that with a p-value of over 5% this study is insufficient evidence to say that the true population correlation is anything but 0. You would fail to reject the null hypothesis. It is common to find that correlations could be the result of chance alone. [\[f\]\[g\]](#)


Standard error[\[h\]](#)

Standard error is a measure of the preciseness of your estimate and is inextricably linked to the p-value. In fact the first step in calculating a p-value involves the standard error. Standard errors are a function of the standard deviation of the distribution of your variable. A standard deviation unit, as mentioned in chapter 2, is a measure of the average distance from the mean in a distribution. The general form of the standard error is

Standard error = 

Where n is the sample size of your data.

Step one to get a p-value is to calculate what is called a test statistic. A test statistic is a transformation of the raw statistic (in the example above the raw statistic would be $r = .198$) into standard deviation units. Here is the general form

test statistic [\[i\]](#) = 

Once this transformation is finished, the computer translates the standard deviation unit into the probability of the tail of the null hypothesis distribution. In other words to show how likely your statistic would have occurred given the null hypothesis is true.

Figure 6.2. Illustration of a p-value in a distribution. [\[k\]](#)

That is just one use of the standard errors. P-values have come under attack as being too simplistic to summarize a statistic. The p-value is simplistic in a way. It just answers one question about the statistic: could this value have happened by chance alone if the null hypothesis is true? In contrast, the standard error can be used to create a plausible range of values called [\[l\]](#) a confidence interval for the parameter from the test statistic. Here is the general form of the confidence interval

confidence interval = statistic \pm (level of confidence [\[m\]](#))*standard error

In the case of age vs. books read the statistic [\[n\]](#) is .198. The level of confidence is a value chosen by the researcher indicating how confident they want to be in the interval. Typically a “95% confidence” level is chosen which translates into plugging in the number 2 for level of confidence. If the standard error for books vs. [\[o\]](#) statistic is .4, then [\[p\]](#) the confidence interval is as follows $.198 + 2*.4$, which results in .998. Then $.198 - 2*.4$ and get approximately -.6. Thus, the confidence interval would be reported as (-.6, .998). That means, researchers are 95% confident that the true population correlation of age vs. books read lies between -.6 and .998. In other words, there is no clear picture of what the true population correlation is as it almost spans the entire correlation range. This shows (a) that our estimate is very imprecise and (b) As 0 is a plausible value for our parameter (0 is in the confidence interval) the null hypothesis can't be rejected ($p > .05$) [\[q\]](#). In an alternative universe where the standard error for age vs. books read is much smaller researchers could end up with a confidence interval of say (.09, .11). This would be interpreted the same way: we are 95% confident that the true population parameter lies between .09, and .11. In this universe, we would be quite sure of our answer, and we could easily reject the null hypothesis. That is because, 0 is not in the interval ($p < .05$). A critic could attack this alternative universe result by saying that even though the correlation is not 0, and researchers are very sure of what the true value is, it is still too small to be relevant. The critic would have a strong case that would need to be discussed subjectively. For example, the researcher could show correlations between income vs. [\[r\]](#) books read per year

or level of education vs. books read per year. If those correlations of what are considered in the literature to be important variables in the life of a student are even smaller than the small correlation of (.09, .11) between age vs. books read is still worth looking at. And the debate would continue. Nevertheless, p-values and standard errors are useful tools in making an empirical case that a statistic is "important".

Exercises

Put examples of p-values, ask the student to interpret. Answer key

P-values

1. How would you interpret the p-value [f\[s\]](#) or the correlation between depression score on the DASS-21 and anxiety score on the DASS-21?
2. How would you interpret the p-value for the correlation between hours spent on social media per week and depression score on the DASS-21?
3. How would you interpret the p-value for the correlation between weight and hours spent on social media per week?

Confidence intervals

Put examples of confidence intervals, ask student to interpret. Answer key

1. Calculate the 95% confidence interval for the correlation between depression score on the DASS-21 and anxiety score on the DASS-21. Interpret your results. The standard error is 0.48. [\[t\]\[u\]](#)
2. Calculate the 95% confidence interval for the correlation between hours spent on social media per week and depression score on the DASS-21 [\[v\]](#). Interpret your results. The standard error is 0.29. [\[w\]](#)
3. Calculate the 95% confidence interval for the correlation between weight and hours spent on social media per week. Interpret your results. The standard error is 0.69 [\[x\]](#).

Answer Key

P-values

1. The probability of getting a correlation of .901 or larger (more extreme) if the null hypothesis of the population correlation of depression score and anxiety score is 0 is 0.1%. That's smaller than .5 so it means it's significant. [\[y\]](#)
2. The probability of getting a correlation of .551 or larger (more extreme) if the null hypothesis of the population correlation of hours spent on social media per week and depression score is 0 is 1.2%.
3. The probability of getting a correlation of -0.294 or larger (more extreme) if the null hypothesis of the population correlation of weight and hours spent on social media per week is 0 is 20.9%.

Confidence intervals

1. Lower band: $(0.901 - 2.00 * 0.48) = -0.059$

Upper band: $(0.901 + 2.00 * 0.48) = 1.861$

$(-0.059, 1.861)$

We are 95% confident that the true population correlation of depression score vs. anxiety score lies between -0.059 and 1.861. [\[z\]\[aa\]](#)

2. Lower band: $(0.551 - 2.00 \cdot 0.29) = -0.029$

Upper band: $(0.551 + 2.00 \cdot 0.29) = 1.131$

$(-0.029, 1.131)$

We are 95% confident that the true population correlation of hours spent on social media per week vs. depression score lies between -0.029 and 1.131.

3. Lower band: $(-0.294 - 2.00 \cdot 0.69) = -1.674$

Upper band: $(-0.294 + 2.00 \cdot 0.69) = 1.086$

$(-1.674, 1.086)$

We are 95% confident that the true population correlation of weight vs. hours spent on social media per week lies between -1.674 and 1.086.

[a]do we know that *r* (italicized) means correlation?

[b]this example would be nice with pictures.

[c]define

[d]I added a transition. What do you think?

[e]Should we do hyperlinks to skip back to these tables? They are a few chapters away. Or maybe just put them in here again so we can look at them.

[f]I don't know if you need this sentence, but you did say this happens often and I didn't like the phrase right there so I redid the paragraph.

[g]I would take it out, unless its important to have them keep in mind their research will often find nothing with a good p-value.

[h]Pu n standard error link

[i]Insert section of sampling distributions.

[j]When you dive into these formulas I get very nervous and and struggling not to gloss over them. I don't really understand.

[k]please label the p-value and put and arrow if necessary. I see not p-value in the above histogram

[l]values of what?

[m]Is this correct? I added the confidence interval to the equation.

[n]raw statistic? test statistic? correlation?

[o]are we missing a word here? can you have a standard error on the values you entered for books read?

[p]Where does this number come from? is it the 40% p-value? A picture would be helpful here to show where the numbers come from.

[q]what does this mean?

[r]is this supposed to have a period? If so this should be consistent throughout.

[s]Am I supposed to create a confidence interval? Am I supposed to say this is $p < .05$ so we reject the null-hypothesis and the correlation is unlikely to be caused by chance alone?

[t]confidence interval = $.901 + 2 \cdot .48, .901 - 2 \cdot .48$

(0.059, 1.861)

0 is not part of the series so the null hypothesis is rejected. It may be that the correlation is small however. We need to address this by looking at subjective literature of correlations that are well accepted.

[u]Say which table this refers to.

[v]again say the table name, or show the table again might be even better.

[w].551+2*.29

.551-2*.29

.58

(-.029, 1.13)

The null hypothesis can't be rejected, because 0 is part of this series. That means it's impossible to tell from our study if there is a non-spurious correlation between social media use and depression.

[x]-.294+2*.69

-.294-2*.69

(-1.674, 1.806)

That's a really big range so we say, we are 95% confident that the correlation between weight and social media use is some number. Also, we can't rule out the null-hypothesis so that nebulous correlation might be caused by chance alone.

[y]Do you want them to remember the relation between this and the $p < .5$ thing?

[z]Should we mention the null hypothesis?

[aa]Also do you want them to comment on how useful this information is likely to be? Like I did in my comments, which I see were the wrong answers... haha



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/se_pvalues.

Chapter 7

Linear Regression

Chapter 7: Linear Regression

Linear regression is the mathematical model behind the path diagrams introduced in chapter 1. Here is a path diagram. [\[a\]](#)

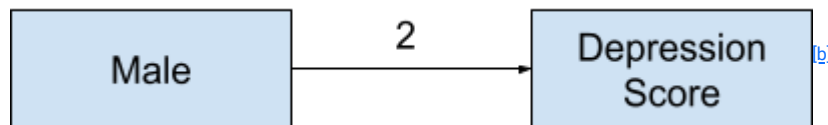


Figure 7.1. A basic Path Diagram showing the relationship between being male and an individual's Depression Score.

Linear regression has two purposes:

1. Prediction: Given the various linear regression parameters a researcher can give a prediction of the next person's depression score based on their gender.
2. Explanation: Looking at Figure 7.1 a researcher can tell males are more depressed than females by 2 points. Thus they may claim that gender causes an increase in depression score. This explains some of the variation in depression scores. [\[c\]](#)

Much of the scientific process is an effort to explain why certain phenomena happen. Linear regression and SEM are powerful tools in achieving that.

The Equation

As mentioned in chapter 1, path diagrams like Figure 7.1 can be more intuitive than the following mathematical equation. Many people would have seen the following equation before

$$y_i = mx_i + b \quad \text{[d][e][f]}$$

where y_i is the predicted outcome for subject i , x_i is the independent variable for subject i , m is the slope and b is the intercept, the value of y_i when $x_i = 0$.

Figure 7.2 refers to the quantity m , where $m=2$. The quantity b is not shown.

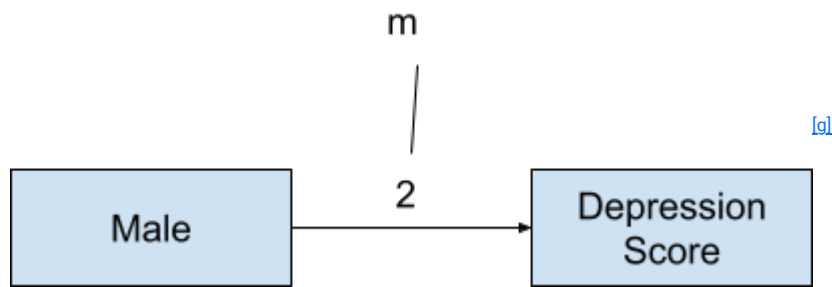


Figure 7.2 A basic Path Diagram showing the relationship between being male and an individual's Depression Score with m labeled and b not shown.

Statistics changes the formula slightly, please follow along in Table 7.1:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Because this equation is not particularly intuitive, the following figures will help clarify the equation.

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Subject ID i	Depression Score y	Male X	Predicted value (y-hat)	error (observed - predicted)
1	24.55605025	0	23	1.556050251
2	21.59820667	0	23	-1.401793331
3	21.51240205	0	23	-1.487597952
4	27.29930048	0	23	4.299300482
5	27.6748042	0	23	4.674804196
6	25.2803203	1	25	0.2803203041
7	23.07122925	1	25	-1.928770747
8	27.9169782	1	25	2.916978201
9	20.87062081	1	25	-4.129379189
10	28.17649223	1	25	3.17649223

estimated from data

Subject i in this case refers to the study ID for a specific individual. Notice there is a y , x , and ϵ_i that each have that little i next to them. Because that is a subscript, that is pronounced sub i , as in y sub i .

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

estimated from data

Subject ID i	Depression Score y	Male X	Predicted value (y-hat)	error (observed - predicted)
1	24.55605025	0	23	1.556050251
2	21.59820667	0	23	-1.401793331
3	21.51240205	0	23	-1.487597952
4	27.29930048	0	23	4.299300482
5	27.6748042	0	23	4.674804196
6	25.2803203	1	25	0.2803203041
7	23.07122925	1	25	-1.928770747
8	27.9169782	1	25	2.916978201
9	20.87062081	1	25	-4.129379189
10	28.17649223	1	25	3.17649223

In this equation, y_i is the observed outcome score, or the actual value for the variable depression score for subject i .

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

estimated from data

Subject ID i	Depression Score y	Male X	Predicted value (y-hat)	error (observed - predicted)
1	24.55605025	0	23	1.556050251
2	21.59820667	0	23	-1.401793331
3	21.51240205	0	23	-1.487597952
4	27.29930048	0	23	4.299300482
5	27.6748042	0	23	4.674804196
6	25.2803203	1	25	0.2803203041
7	23.07122925	1	25	-1.928770747
8	27.9169782	1	25	2.916978201
9	20.87062081	1	25	-4.129379189
10	28.17649223	1	25	3.17649223

The variable x_i in this case is the individual's Male score (0=female, 1=Male).

w

Column 4 is the Predicted value or \hat{y}_i . It is not part of the equation, however, it is used to calculate the error or ε_i . It is the value the model predicts for subject i given their male value x_i . Notice the predicted value has a "hat" over the y to show it is the predicted value. Look at the subject who has the ID=1 and is found in the second row, their

Male variable is 0, meaning they are female. Their predicted depression score (\hat{y}_i) is 23 which is found in column 4. This is true for all females in the dataset (rows 1-5). Now, look at the subject who has ID=6 whose Male variable is 1, meaning they are male. Their predicted depression score (\hat{y}_i) is 25 which is found in column 4. This is true for all males in the dataset. The equation that is used to calculate this predicted value will be discussed below. The predicted value is wrong for every subject in the dataset as can be seen by comparing their observed depression score (y_i) in column 2 to their predicted score (\hat{y}_i) in column 4. The difference between what is observed (y_i) and what is predicted (\hat{y}_i) by the model is called the error term (ϵ_i).

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Subject ID i	Depression Score y	Male X	Predicted value (\hat{y} -hat)	error (observed - predicted)
1	24.55605025	0	23	1.556050251
2	21.59820667	0	23	-1.401793331
3	21.51240205	0	23	-1.487597952
4	27.29930048	0	23	4.299300482
5	27.6748042	0	23	4.674804196
6	25.2803203	1	25	0.2803203041
7	23.07122925	1	25	-1.928770747
8	27.9169782	1	25	2.916978201
9	20.87062081	1	25	-4.129379189
10	28.17649223	1	25	3.17649223

The parameter ϵ_i , pronounced epsilon sub i , is the error term. Linear regression predicts a depression score for subject i based on their gender. The error term is the difference between that predicted depression score and the subject's actual observed depression score.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Subject ID i	Depression Score y	Male X	Predicted value (\hat{y} -hat)	error (observed - predicted)
1	24.55605025	0	23	1.556050251
2	21.59820667	0	23	-1.401793331
3	21.51240205	0	23	-1.487597952
4	27.29930048	0	23	4.299300482
5	27.6748042	0	23	4.674804196
6	25.2803203	1	25	0.2803203041
7	23.07122925	1	25	-1.928770747
8	27.9169782	1	25	2.916978201
9	20.87062081	1	25	-4.129379189
10	28.17649223	1	25	3.17649223

The parameter β_0 is the predicted value of the outcome when the predictor, x_i is 0. In this case, for the variable male, someone who has a score $x_i = 0$ is female. Thus β_0 is equivalent to b in $y = mx + b$. β_1 means that for every one unit increase in x_i the predicted score will change by β_1 . In our case, x_i can only take on two values, 0 meaning female and 1 meaning male. Therefore as the variable changes from female to male, the predicted depression score increases by 2 points. β_1 is equivalent to m in $y = mx + b$.

Taking the information from Figure 7.1 and plugging it into the general formula results in:

Table 7.1

Subset of data used in calculating the model where male predicts depression score

$$(\text{depression score})_i = \beta_0 + 2 * (\text{male})_i + \epsilon_i$$

Memorize the following:

- y_i is the observed outcome score for subject i .
- β_0 is the predicted value of the outcome when the predictor, x_i is 0.
- For every one unit increase in x_i the predicted score will change by β_1 .

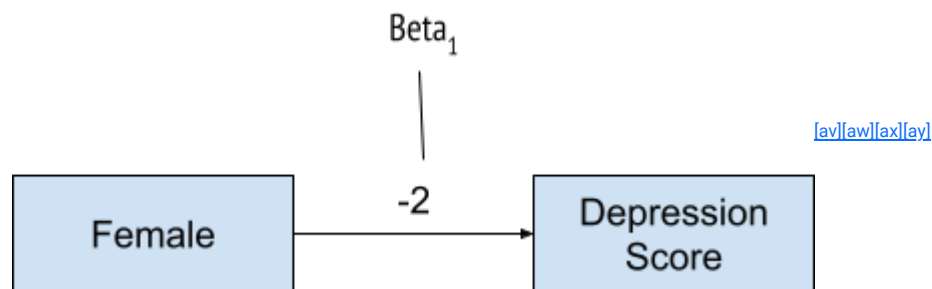


Figure 7.3 A basic path diagram showing the relationship between being male and an individual's depression Score with β_1 labeled and β_0 not shown.

y_i is the depression score for person i . β_0 is the predicted depression score for person i if they were female $x=0$. β_1 is the effect of gender on depression score and is estimated to be 2. This is interpreted exactly as above: Being male will result in a 2 point increase in predicted depression score. Note that graphic shown in Figure 7.1 gives no prediction for β_0 , the predicted depression score for a female. This is because β_0 is generally not a value that researchers are focused on. This study is focused on predicting the effect of gender on depression score. It is not focused on predicting the actual depression score. If the predicted depression score is of interest then the figure could be modified as seen in Figure 7.2.

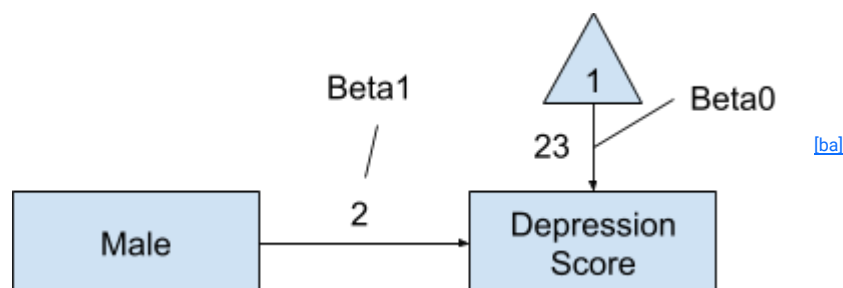


Figure 7.2. A basic path diagram with the estimated β_0 associated with the triangle shape

From Figure 7.2 the triangle with a 1 in the middle has been added with an arrow pointing to the outcome (depression score). The 1 in a triangle symbolizes that this value is a constant for all subjects. The estimated value 23 shows the value for this constant. Thus, we would interpret the 23 as the predicted score for person i who has value "0" for all

predictors. More complicated models might have more than one predictor. In this case there is only 1 predictor (male) with the values (1:Male, 0:Female). With that extra information we can further contextualize the number 23 and say it is the predicted depression score for woman i . If a person is a man then you have to take the 23 and add the effect of being male to it (in this case 2). Thus, the predicted depression score for a man is 25. ^[bb] This is how column 5 in Table 7.1 is calculated.

The linear regression equation in this case would be:

$$(\text{depression score})_i = 23 + 2 * (\text{male})_i + \epsilon_i$$

Note, that in Figures 7.1 and 7.2 the error term is not represented. Generally the error term is considered a nuisance variable and thus can be safely ignored. In this case, every shape that receives an arrow is assumed to have an error term. If, for whatever reason the error term is desired, the figure can be represented as in Figure 7.3.



Figure 7.3. A basic path diagram with the error component shown explicitly

The circle with the ϵ_i in it represents the error component of the prediction. It is a circle because it is not directly observed like the variables male or depression score are, but must be calculated from the model.

Unstandardized and Standardized Coefficients

In Figure 7.3 and in the corresponding equation the calculated or realized value of β_1 ^[bd] in this case “2” is called the unstandardized beta or the unstandardized coefficient. An unstandardized coefficient is in the natural metric of both x_i and y_i . Thus, for every one unit increase in Male, or as a new subject is male instead of female, the predicted depression score increases by 2 depression score units. If y_i was height in inches instead of depression score then a subject who is considered Male (their value was 1) would be predicted to be 2 inches taller than those who were not Male (their value was 0). An unstandardized beta or unstandardized coefficient is useful if the consumer of the information is familiar with the metrics involved. Thus, an expert in depression score would know whether the value of “2” is considered large or important. For those who are not experts in the metric of depression score another way to judge whether a beta or coefficient is large or important is helpful. Thus, we come to standardized coefficients. Standardized coefficients are created by changing the natural metric of y_i and, when appropriate, x_i to standard deviations. In the case of depression scores and Male, it makes sense to rescale depression scale to standard deviations instead. On the other hand, for the variable Male, standard deviations don’t make sense as Male is dichotomous and can take on only two values (0 or 1). After transforming just y_i into standard deviation units the results are found in Figure 7.4.

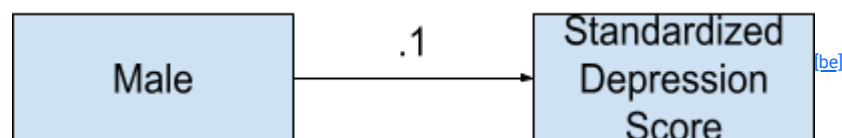


Figure 7.1. A basic Path Diagram showing the relationship between being male and an individual’s ^[bf] depression score when depression score is standardized.

Now we would interpret the relationship as: Males score .1 standard deviations higher than females on Depression score, a small value. This is much more informative to those consumers than the natural metric. Thankfully, SPSS automatically calculates this value for us as seen subsequently. ^[bg]

Your turn. [\[bh\]](#)[\[bi\]](#)

In this example, we'll look at how to predict a person's anxiety score based on their gender. The equation for this example is still:

$$y_i = \text{img}0 + \text{img}1x_i + \text{img}i$$

And the model is:



Click on this link to download this .sav (which is an SPSS datafile) file.

[Insert file]

Once you have downloaded this dataset, open it in SPSS. Now go to Analyze → Regression → Linear.

Figure 7. [\[bi\]](#) # Screenshot of example dataset

In the GUI for linear regression, move the variable “Anxiety score on the DASS-21” to the ‘Dependent’ box, and move the variable “Gender” to the ‘Independent’ box, as seen in Figure #. Then click ‘OK’ to run your regression.

Figure 7. [\[bk\]](#) #. Graphic User interface (GUI) for linear regression in SPSS.

SPSS produces a series of tables when you run your regression. Look at the ‘Coefficients’ table to see the predicted effect of gender on anxiety score.

Figure 7.#. Coefficients table for the linear regression of anxiety score on the DASS-21 on gender

The output of the ‘Coefficients’ table tells us the following:

- (the predicted value of the outcome when the predictor x_i is 0) is 18.400.
- For males you would predict a 5.70 point decrease in anxiety score on the DASS-21. [\[bl\]](#)[\[bm\]](#)
- The probability of obtaining an unstandardized beta of -5.70 or more extreme when the null hypothesis is true is 15.9%. [\[bn\]](#)

[\[a\]](#) AMOS examples.

[\[b\]](#) bring the figure down, have a label for 2 connecting it with Beta1, Male=X, Depression Score=Y

[\[c\]](#) Make this example real.

[\[d\]](#) Create a dual equation figure where we have arrows from m to B1 etc.

[\[e\]](#) Ignore if not familiar

[\[f\]](#) Think about Keith Verbosity, appendices???

[\[g\]](#) bring the figure down, have a label for 2 connecting it with Beta1, Male=X, Depression Score=Y

[h] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[i] Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[j] Introduce p-value, refer to chapter 6 in parenthesis

[k] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[l] Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[m] Introduce p-value, refer to chapter 6 in parenthesis

[n] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[o] Define the subject.

[p] Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[q] Introduce p-value, refer to chapter 6 in parenthesis

[r] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[s] Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[t] Introduce p-value, refer to chapter 6 in parenthesis

[u] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[v] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[w] Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[x] Introduce p-value, refer to chapter 6 in parenthesis

[y] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[z] Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[aa] Introduce p-value, refer to chapter 6 in parenthesis

[ab] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[ac] Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[ad] Introduce p-value, refer to chapter 6 in parenthesis

[ae] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[af]Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[ag]make these all say male, not gender

[ah]Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[ai]Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[aj]Introduce p-value, refer to chapter 6 in parenthesis

[ak]I do like that B(1) is the first thing in this sentence. I don't like that it is repeated twice in the sentence. I do think it helps keep track of where we are in the confusing equation, but I also don't think it's terribly grammatical. Any suggestions?

[al]This would still have B1 twice in one sentence, but it could say something like:

B1 can be interpreted as: For every one unit increase in x_i , the predicted score will change by B1.

[am]Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[an]Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[ao]Both figure and table captions are above the thing now.

[ap]define the subject.

[aq]Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[ar]Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[as]Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[at]Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[au]Show some examples using the spreadsheet of what $x(i)$ and all the others are.

[av]Suggestion: just reverse the gender.

[aw]bring the figure down, have a label for 2 connecting it with Beta1, Male=X, Depression Score=Y

[ax]blue eyes vs. other? instead of male.

[ay]also, gender is sooo common in the field, probably should keep it.

[az]Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[ba]Include beta0 arrow/label

[bb]make bullet pointy

[bc] Have the student do this in SPSS long way, and then AMOS. Add more variables. Split chapters into Simple Linear, and Multiple Linear Regression chapter. Just tease the multiple linear regression, tell them to take stats 2.

[bd] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[be] bring the figure down, have a label for 2 connecting it with Beta1, Male=X, Depression Score=Y

[bf] Change formatting, from paragraph format to more of a single sentence, then the figure, repeat. Have a final paragraph synthesis.

[bg] Include an example where x is continuous

[bh] Somewhere in this chapter, we probably need to include something about standardized vs. unstandardized output in SPSS and when to use each one.

Also, how to interpret p-values in the context of regression. Should we do it in this example or before the example?

[bi] We probably also need to talk about centering continuous IVs

[bj] #?

[bk] #?

[bl] Insert a sidebar that says that for here the unit is gender but if the x variable is inches, the unit would be inches.

[bm] Theoryish chapter/applied chapter, AMOS.

[bn] Include another example.



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/linear_regression.

Appendix A

Introduction to SPSS

Appendix A: Introduction to SPSS

The main software you will be using for this course is SPSS Statistics ("SPSS").

SPSS is available to BYU students using a cloud-based workspace called Cloud Apps.[\[a\]](#)

Opening the Software

To open SPSS, visit cloudapps.byu.edu. Then click 'Log Into Cloud Apps.'

On the next page, enter your BYU email address (netid@byu.edu) and password into the login box. This should be the same information you use to log into Learning Suite and other BYU programs. Once you have entered your information, click 'Log On.'

Once you log in, you'll see a list of recently used apps. To see the full list of apps, select Apps → All Apps.

This page shows all of the apps you can access through Cloud Apps. Scroll down to find the SPSS logo.

Click the icon to launch the software. It should open in a new window in your browser. It may take several minutes to load.

Getting Familiar with SPSS

This section is designed to familiarize you with the SPSS graphical user interface (GUI). This welcome dialog box is the first screen you see when you open SPSS. It includes the option to:

- Open a new dataset
- Restore a recent file
- View sample files
- Or view SPSS tutorials, among other things

To get to the main screen, close this window. Then, you will see the following screen. This is a blank data file.

Data View and Variable View

There are two ways of viewing data files in SPSS: Data View and Variable View.

Data View: This is where your spreadsheet's data is kept. It is laid out in a traditional column and row format, much like Microsoft Excel.

Variable View: This houses information about the variables in your dataset, including their name, description (AKA label), missing values, and more.

So, how do you know which view you are in? At the bottom of the data file, you'll see the following.

Whichever tab is underlined in blue represents the view you are currently in. Select the other tab to change the view.

Creating New Files

Data

To create a data file in SPSS, go to File → New → Data. This will create a new, blank data file where you can enter data.

Syntax

As you can see, SPSS also gives the option to create a new syntax file. What is syntax?

SPSS Syntax: Code that tells SPSS what to do and how to do it. Much like in other statistical software, there is code behind each action you perform in SPSS.

You do not have to know how to code to analyze data in SPSS. However, on most actions in SPSS, you will see the option to 'Paste.' By clicking paste, SPSS will add the syntax for the action you want to perform to a syntax file where you have a record of your changes. It is best practice to 'Paste' your syntax as you go. Save your syntax file in case you need to reference it later or re-run something.

When should you create a new syntax file?

When you want to designate a location for the syntax you 'Paste.' This is helpful if you are working on a project and want to track all of the changes in one place.

Output

When you run analyses or create tables/graphs/charts in SPSS, what you create will be added to an output file. If you would like to designate a location for your output, you can create a new file specifically for your project.

Opening Pre-existing Files

To open a pre-existing data file in SPSS, go to File → Open → Data and locate the file on your computer or in Box.

When you launch SPSS, you will be prompted to authenticate Box using the same login you use to get into Cloud Apps. After authenticating Box, you will have access to files saved in Box and won't need to re-authenticate on subsequent logins.

Importing Data

Creating Dummy Variables

Many data sets include nominal variables: variables with multiple categories, but no intrinsic order to those categories (e.g., gender, race, favorite fruit).

To use a nominal variable in analysis (such as linear regression), you may need to dummy code.

Dummy Coding: Turning the response options from a nominal variable into dichotomous variables (which have only two response options).

Example: Let's say our data set includes the variable 'FavoriteFruit,' which has four response options:

1 = Apples

2 = Bananas

3 = Oranges

4 = Other

By dummy coding 'FavoriteFruit,' we will create four new variables:

FavoriteFruit_Apples

FavoriteFruit_Bananas

FavoriteFruit_Oranges

FavoriteFruit_Other

Each new variable will have two response options – represented by 1s and 0s. For the new variable 'FavoriteFruit_Apples,' a value of 1 will represent people who said their favorite fruit is apples. A value of 0 will represent those who listed a fruit other than apples as their favorite. The value coded as 1 becomes the reference category in a regression equation.

How to Dummy Code

To dummy code a variable, go to Transform → Create Dummy Variables.

Then, select the variable you wish to dummy code and click the arrow to move it to the 'Create Dummy Variables for:' box.

Next, add a name in the 'Root Names' box. SPSS will use this as the first part of the name in the dummy coded variables. Under 'Measurement Level Usage,' select 'Create dummies for all variables.'

After we click 'OK' or 'Paste' and run our syntax, SPSS should create four new variables.

By double clicking on the variable name in the 'Name' column, we can rename the variables so it's easier to identify which is which.

You can double check that the variables have been dummy coded by looking at the Data View tab. The new variables should only include 1s and 0s.

Recoding ID Variables

Many data sets include unique identifiers for each participant. These are often a combination of letters, numbers, and characters, also known as a string variable.

Why would I need to recode a string variable? If you'll only be analyzing your data in SPSS, you probably do not need to work about recoding your string variables. However, if you plan on using other statistical software (e.g., Mplus) to analyze your data, you must recode. Mplus will not run files with string variables.

By recoding your string variables, you can preserve a unique identifier without limiting yourself to one software.

How to Recode String Variables

Go to Transform → Automatic Recode

In the pop-up box, move your ID variable to the 'Variable->New Name' box. Then, type a name for your new variable into the 'New Name' box. Press 'Add New Name' to add it.

Click 'OK' or 'Paste' and run your syntax. Your new variable should only include numbers. You may wish to save a copy of your file that includes the old and new ID variables so you know which is which. Delete your old ID string variable from whatever file you plan to export for use in software like Mplus.

Descriptive Statistics

It is often helpful to have details about the variables in your dataset. This may include the minimum, maximum, standard deviation, mean, median, mode, and more. To generate these descriptive statistics, go to Analyze → Descriptive Statistics → Frequencies.

Then, select the variables you wish to see the descriptive statistics for and click the arrow to move them to the 'Variable(s)' box. Select 'Statistics...' to choose which descriptive statistics you wish to generate.

Here we have selected mean, median, standard deviation, minimum, and maximum, but you can select whatever you wish. Click 'Continue.'

You may wish to deselect 'Display frequency tables.' Click 'OK' or 'Paste' and run your syntax. SPSS will generate tables with the data you requested.

Recode into same

Scatterplots

How to Generate a Scatterplot

Histograms

Histograms are often used to visualize continuous data. Histograms display a distribution of scores, with each bar representing a different value.

How to Generate a Histogram

Missing Data

Data sets usually include missing data. For example, when a survey participant skips several questions on a survey, the responses to those skipped questions are 'missing.' You may notice missing data in your dataset because the cells will have a period in them.

Why do I need to tell SPSS I have missing data? Although you may know what data is missing, SPSS does not know this intuitively. It only knows that some cells are filled with periods (AKA system missing). As a result, you need to tell SPSS what cells the software should read as missing. It is helpful to fill missing cells with a value like -999 that is easily identifiable as missing.

What are the consequences of not addressing missing data? When handled incorrectly, missing data can skew analyses and produce confusing or incorrect results.

There are several ways to tell SPSS that you have missing data.

How to Specify Missing Data

There are two steps to identifying missing data in SPSS. The first involves filling your blank and period-filled boxes with a value that

Go to Transform

How Does SPSS Handle Missing Data?

Some statistical software (including SPSS) uses listwise deletion, which excludes a participant from the analysis if they are missing data on one or more of the variables you are trying to analyze. Listwise deletion is problematic because it reduces the sample size and statistical power of your survey. In some analyses in SPSS, you can choose to

How can I learn more about how to handle missing data? You may want to take a class specifically about how to handle missing data.

Missing data - go take a class, labeling

Exporting Data

[\[a\]](#) Should we include this portion about Cloud Apps?



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/intro_to_spss.

Appendix B

Introduction to AMOS

Appendix B: Introduction to AMOS

AMOS is also available through BYU Cloud Apps. Once you log in and select All Apps, scroll down until you see the AMOS logo.

AMOS:

Click the icon to launch the software. It should open in a new window in your browser. It may take several minutes to load.

AMOS has menus you can use to access its features and an array of buttons on the left side. The buttons are explained below:

Draw an observed variable

Draw an unobserved (latent) variable

Draw a latent variable or add an indicator to a latent variable

Draw paths with a single-headed arrow

Draw covariances with a double-headed arrow

Add a unique variable to an existing variable

Figure captions

List variables in model

List variables in data set

Select one object at a time

Select all objects
Deselect all objects
Duplicate objects
Move objects
Erase objects
Change the shape of objects
Rotate the indicators of a latent variable
Reflect the indicators of a latent variable
Move parameter values
Scroll through path diagram
Touch up a variable (makes your path diagram prettier)
Select data file(s)
Analysis properties
Calculate estimates
Copy the path diagram to the clipboard
View text output
Save the current path diagram
Object properties
Drag properties from object to object
Preserve symmetries
Zoom in on an area you select

Zoom in

Zoom out

Show the entire image on screen

Resize the diagram to fit on a page

Zoom in closely on the diagram

Bayesian

Multiple group analysis

Print the selected path diagram(s)

Undo

Redo

Specification search



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/intro_to_amos.

Common Formulas

R-Squared

$$R^2 = \frac{N \sum xy - \sum x \sum y}{\sqrt{[N \sum x^2 - (\sum x)^2] [N \sum y^2 - (\sum y)^2]}}$$

F Test

$$F = \frac{\text{Variance of set 1}}{\text{Variance of set 2}} = \frac{\sigma_1^2}{\sigma_2^2}$$

See [Variance](#).

Chi-Square

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Population Mean

$$\mu = \frac{\sum X_i}{N}$$

Mean

$$\bar{x} = \frac{\sum x}{n}$$

Variance

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$$

Standard Deviation

$$S = \sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

Linear Regression

$$y = a + bx$$

Where a (or the intercept) is:

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{(\sum x^2) - (\sum x)^2}$$

And b (or the slope) is:

$$b = \frac{n \sum xy - (\sum x) (\sum y)}{n \sum x^2 - (\sum x)^2}$$



This content is provided to you freely by EdTech Books.

Access it online or download it at https://edtechbooks.org/sem/common_formulas.