7
Drawing Causal Diagrams

Our Idea of the World

We have a basic idea now about how causal diagrams work. But that assumes we have a causal diagram to work with. How can we get one? We'll have to draw one ourselves.[1]

Causal diagrams represent the data-generating process that got us our data. So, drawing our own causal diagram will come down to putting our idea of what the data-generating process is onto paper (or a computer screen).

This can be tricky! It requires that we know as much as possible about that data-generating process before getting started. So the first step in drawing a causal diagram is really to do some research. Lots and lots of research on your topic.

Once you've done that, you can combine what you've learned with your intuition, follow this chapter, and suddenly, hey there it is—a causal diagram. One step closer to answering that research question.

Thinking Through the Data Generating Process

How can we possibly write the world down in a graph? After all, the world is very complex, and a graph is clean and simple. But we'll figure it out.

The key is to focus in as much as possible. Think about our research question and try to live in the world of that research question. We can't possibly put everything in the world on our graph, so we need to work hard to not fall for the trap of trying to do so. What we *can* do is try to put everything relevant to our research question on the graph.

All through the process, we're going to want to keep in mind that we're trying to make a graph that mimics the *data generating process*

[1] This chapter will be focused on how to design and put together a causal diagram. For the actual literal drawing portion of the show, I recommend either doing it by hand or using the website `dagitty.net`.

relevant to our research question. What leads us to observe the data we do? What causes the outcome? What causes the treatment?

It will probably help if we work with an example.

Let's walk through a study that I worked on.[2]. That way I can tell you exactly what sorts of things we were thinking about when considering what we thought the data generating process looked like. In this study, we were interested in the effect of *taking online courses* on *staying in college*, specifically in community college in Washington State.[3,4]

Our first task will be thinking through the list of relevant variables.

First off, just so we're clear, *what is a variable?* We've covered this in previous chapters but it bears repeating since it's easy to forget when applying this stuff to the real world!

A *variable* on a causal diagram is a *measurement* we could take that could result in different values. So, for example, one of the variables relevant to our research question is "online class." Is the class you're taking online?—That's the measurement. It could be "yes," it could be "no"—those are the values. If we like we could call it "type of class" with the values "online class" and "face-to-face class." Notice that we don't have one variable for "online class" and another for "face-to-face class." That's because those aren't two separate variables, they're two separate values that the same variable could take.

So there's one relevant variable—online class. That's our treatment variable. We're interested in the effect of that variable. Another relevant variable will be "dropout"—did the student drop out of college since taking the class? This is our outcome variable. Treatment and outcome are always a good place to start.

Treatment variable. The variable we want to know the effect *of.* How does it affect the outcome?
Outcome variable. The variable we want to know the effect *on.* How does the treatment variable affect it?

Variable list:

- Online class
- Dropout

Alright, now what else?

We want to include all variables relevant to the data-generating process. That means any variable that has something to say about whether we observe online class-taking, or whether we observe dropout, or whether we observe both them together or apart! *Every variable that causes the treatment or outcome, or causes something that causes something that causes the treatment or outcome, or causes something that causes something that causes something...* is a good candidate for inclusion![5]

What are some things that might cause people to take online classes? Well, different students have different preferences for or

[2] Nick Huntington-Klein, James Cowan, and Dan Goldhaber. Selection into Online Community College Courses and their Effects on Persistence. *Research in Higher Education*, 58(3):244–269, 2017

[3] The world is just dripping with context. The data generating process may well be different depending on where, when, or who you're talking about! The diagram in this chapter might look very different if the study were to be done in Oregon. Or Thailand!

[4] Is this a good research question? Sure! We can imagine randomly assigning students into online or face-to-face college classes and recording afterwards whether they stay in college or drop out, so it's a question we can answer with evidence. And learning the answer to this question would certainly help shape a theory about why people stay in school. If online classes lead to dropout, then something about the face-to-face experience might make people want to stick around.

[5] You may want to include some variables that *don't* fit that description too, if they happen to be closely *related* to the treatment or outcome, or to some of the other variables.

against online courses, so Preferences. Those preferences might be driven by background factors like Race, Gender, Age, and SocioeconomicStatus. Those same background factors might influence how much AvailableTime students have—time-pressed students may prefer online courses. And AvailableTime might be influenced by how many WorkHours the student is doing. You also need solid InternetAccess to take online courses.[6]

Now how about things that might cause people to drop out of community college? Some of the same background factors as before might be relevant, like Race, Gender, SES (socioeconomic status), and WorkHours. Your previous performance in school, Academics, is also likely to be a factor.

Now what does our list look like?

- OnlineClass
- Dropout
- Preferences
- Race
- Gender
- Age
- SES
- AvailableTime
- WorkHours
- InternetAccess
- Academics

Now is a good time to pause, look at our list, and think hard about whether there's anything important that we've left off. You may be able to think of a thing or two. But for now, let's leave it at that.

How can we tell if something is important enough to be included? I just mentioned we should think about whether there's anything important being left off. But how can we tell if a variable is important or not?

What it really comes down to is *how strong* we think the causal links out of that variable are.[7]

For example, the presence of QuietCafes in someone's area might encourage them to take an online course. A nice quiet place outside

[6] Writing this during the Coronavirus lockdown in 2020 gives me another idea for a *preeeetty* darn important cause of taking online classes. But of course this isn't relevant to the data generating process in the paper since the the data is from long ago—lockdown was not occurring at that time. So we leave it out.

[7] Or how strong the links *into* that variable are if it's a collider anywhere... that's a different chapter.

the house to do schoolwork. That may well be a real thing that encourages a few additional students to take online courses. But it's unlikely to really be a determining factor for too many students.

So yes, it's relevant, but it seems unlikely that it would have anything but a tiny effect, on average, on whether a student takes an online course. So we're probably okay leaving it out.

With our set of variables in hand, we must try to think about which variables *cause* which others.

Conveniently, we've already done most of the work here! When we were thinking about which variables to include, we were asking ourselves what variables might be out there that cause our treatment or outcome variables. So we already have an idea of what might cause those.

What's left is to think about how those variables might cause *each other*, or perhaps be *caused by* the treatment or outcome. We might also want to consider whether any of the variables are related but neither causes the other, in which case they must have some sort of common cause we can include.

We already have some causes of our treatment and outcome, as well as some other causes, from how we described the variables as we introduced them in the previous section:

- OnlineClass: causes dropout
- Dropout
- Preferences: causes OnlineClass
- Race: causes Preferences, Dropout
- Gender: causes Preferences, Dropout
- Age: causes Preferences, Dropout
- SES: causes Preferences, Dropout, InternetAccess
- AvailableTime: causes OnlineClass, Dropout
- WorkHours: causes AvailableTime
- InternetAccess: causes OnlineClass
- Academics: causes Dropout

How about which non-treatment and non-control variables cause each other? Age certainly causes SES, and all of the background variables (Race, Gender, Age, SES) affect AvailableTime and WorkHours. SES probably causes InternetAccess as well.

How about which variables are related to each other without there necessarily being a clear causal arrow in either direction? In this case, we add on common causes we can just call U1, U2, etc., that cause both variables.

Academics and SES are clearly correlated with Race and Gender, so we'll want to have some sort of common cause there. Academics might also be related to the kinds of employment someone has now and so affect WorkHours. InternetAccess is also likely to be caused not just by SES, but also Location, which we've left out up to now. So Location might want to get in that list!

Now what do we have?

- OnlineClass: causes dropout
- Dropout
- Preferences: causes OnlineClass
- Race: causes Preferences, Dropout, AvailableTime, WorkHours, related to Academics, SES
- Gender: causes Preferences, Dropout, AvailableTime, WorkHours, related to Academics, SES
- Age: causes Preferences, Dropout, SES, AvailableTime, WorkHours
- SES: causes Preferences, Dropout, InternetAccess, AvailableTime, WorkHours
- AvailableTime: causes OnlineClass, Dropout
- WorkHours: causes AvailableTime
- InternetAccess: causes OnlineClass
- Academics: causes Dropout, WorkHours, related to Race, Gender
- Location: causes InternetAccess, related to SES

And now that we have our list we can draw a diagram. Although with a list this long and this many causal arrows described, I can warn you it's going to be a little messy. Okay, very messy. That can happen. We'll be coming back to clean it up a bit later.

AND THERE WE HAVE OUR DIAGRAM in Figure 7.1. Now that we do, it's a good time for *revision*. Looking at the diagram, what might be left off? What variables are likely to be relevant? What arrows should probably be there?

There are probably a lot of things we're missing here.[8] One big one

[8] And don't worry... things weren't quite this simple in the actual paper.

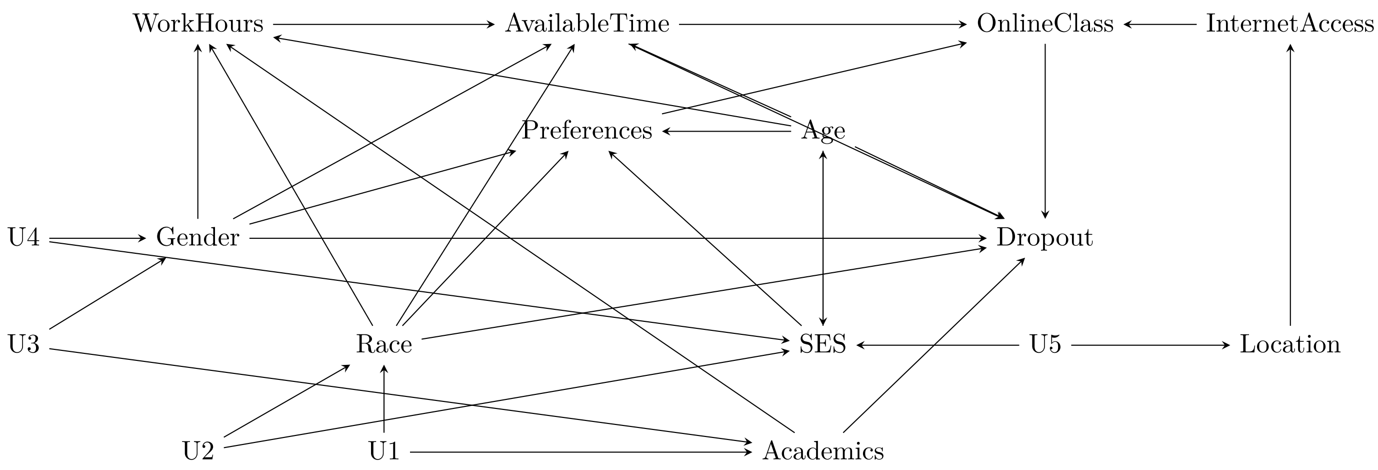

Figure 7.1: A Messy Diagram of the Effect of Online Classes on Dropout

is the specific community college being attended—some offer lots of online courses and some don't! So that's likely to be a big cause of OnlineClass, and caused by plenty of other things on the list, especially Location.

But that's not all. It's a good exercise to look at a causal diagram and think carefully about what's both important and missing.

Simplify

The real world is complex. And so the true data-generating process is too.

But of course that leads us to a problem. The whole point of having a model like a causal diagram is to help us make sense of the data-generating process and, eventually, figure out how we can use it to identify the answer to our research question.

But if the diagram we end up with looks like what we have in the previous section, we're going to be very hard-pressed to make any sort of sense of it. It would be handy if we could simplify it in some way.

Why is it important to simplify? Ultimately, the more complex a causal diagram is, the less helpful it is likely to be. Imagine you were asking someone directions to the next gas station and instead of saying "it's two exits North on the freeway, then next to the Wendy's" they handed you a giant atlas where each page is so intricately detailed that it only covers a single square mile. Sometimes less information is more information.

The trick, of course, will be to simplify where we can without getting *so* simple that our diagram no longer represents the true data generating process. It's a bit too far in the other direction to take the atlas away and just say the nearest gas station is "on Earth somewhere."

HOW CAN WE HIT THAT GOLDEN MEAN OF SIMPLE BUT NOT TOO SIMPLE? We can apply a few simple tests to see if there's any needless complexity in our diagram.

Unimportance We've already discussed this one—If the arrows coming in and out of a variable are likely to be tiny and unimportant effects, we can probably remove the variable.

Redundancy If there are any variables on the diagram that occupy the *same space*—that is, they have the arrows coming in and going out of them from/to the same variables—we can probably combine them and describe them together (this works even if there are arrows between some of the variables being grouped together).

Mediators If one variable is *only* on the graph as a way for one variable to affect another (i.e. B in A → B → C where nothing else connects to B), then we can probably remove it and just have A → C directly.

Irrelevance This one will take some additional knowledge about causal paths from Chapter 8. Some variables are an important part of the data generating process but irrelevant to the research question at hand. If a variable isn't on any path between the treatment and outcome variables, we can probably remove the variable.[9]

Can we apply these steps to our diagram above? We've already done Unimportance and left a few variables off for that reason. Let's leave Irrelevance for now since we haven't gotten to Chapter 8 yet. Can we do Redundancy or Mediators?

We do have some variables that occupy the same space on the diagram and so might be redundant. In particular, Gender and Race have the exact same set of arrows coming in and going out.[10] So we can combine those into one, which we can call Demographics.[11]

How about Mediators? We have a few here, the most prominent of which is Preferences. Instead of having Gender, Race, SES, and Age affect Preferences and then have Preferences affect OnlineClass, we can just have those four variables affect OnlineClass directly. Another one here is Location → InternetAccess → OnlineClass. We can chuck InternetAccess right out and lose nothing!

The last one is a bit less certain. WorkHours affects OnlineClass through AvailableTime. WorkHours and AvailableTime don't quite fall under Redundancy, since Academics affects WorkHours but not

[9] This also requires that we aren't planning to use the variable as an instrument (Chapter 20), and that the variable isn't a collider (Chapter **??**) that we've controlled for.

[10] Other than the U1, U2, etc. But for this purpose we an ignore those. The new U1, U2, etc. that we include in the diagram will just technically be a mix of the old ones.

[11] Combined variables like this can make diagrams nicer but make things slightly trickier when we get to controlling for things in later chapters. We have to remember that Demographics was made up of Gender and Race, and so if we want to control for Demographics we need to control for *both* Gender and Race.

AvailableTime! And they don't quite fall under Mediators, because other variables besides WorkHours affect AvailableTime.

However, the other variables besides WorkHours that cause AvailableTime *also* cause WorkHours. So if we got rid of AvailableTime and just had WorkHours affect OnlineClass directly (Mediators), we'd still have all those same AvailableTime causes affecting WorkHours and wouldn't lose anything (Redundancy). The only sticky thing is that Academics doesn't cause AvailableTime. But that's fine in this case, because Academics *does* cause AvailableTime... through WorkHours! We do lose a little bit of information with this simplification because of the Academics variable, so we'd have to think carefully about whether we're okay with that.

Now we have the much-better-looking, although still slightly messy, causal diagram below:

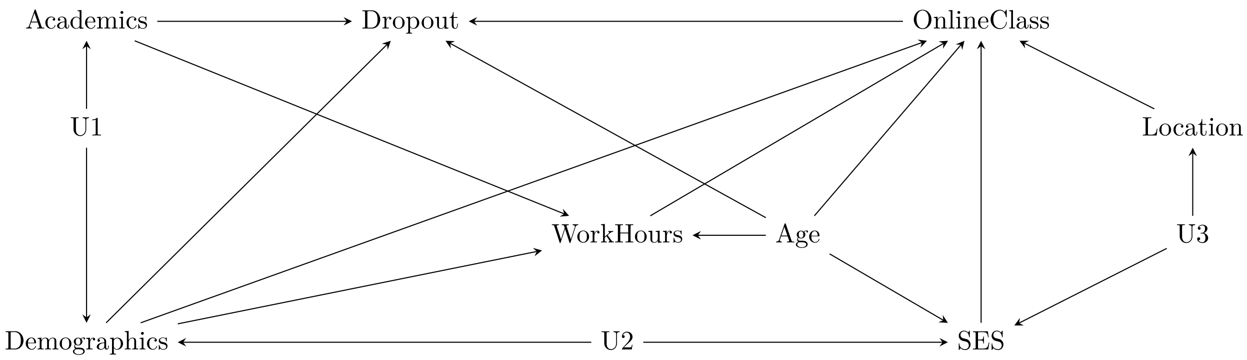

Figure 7.2: A Cleaner Diagram of the Effect of Online Classes on Dropout

While these steps can come in very handy, pay close attention to the use of "probably" in each of them. We can't just apply these blindly. Even if a variable is subject to one of these steps, we don't want to remove it if it's key for our research design or crucial for communicating what's going on.

For example, let's say we're interested in the effect of exercise on your lifespan. One reason exercise might lengthen your lifespan is because it raises your heart rate, and another reason is that it develops muscle. Heart rate and muscle development might be subject to the Mediator step—if the only arrows pointing to them are from exercise, and the only ones out are to lifespan, then we could eliminate both and just have exercise point to lifespan. But if we're interested in *why* exercise works (is it heart rate or is it muscle?) we'd have no hope of answering that question if heart rate and muscle development aren't actually on the diagram. So that would be a simplification too far.

To give another example, just a few paragraphs ago we eliminated InternetAccess because it was a mediator. But in the original study we're talking about, the fact that InternetAccess caused OnlineClass was crucial because it acted as an "instrument" (Chapter 20). If we did that simplification in our study there would be no study!

Avoiding Cycles

There is one thing a causal diagram cannot abide. And that is *cycles*.[12] That is, you shouldn't be able to start at one variable, follow down the path of the arrows, and end up back where you started.

There are two examples of graphs with cycles below in Figure 7.3. In the first one, you can go A → B → C → A. In the second, the variables cause each other, and so you can just go A → B → A.[13]

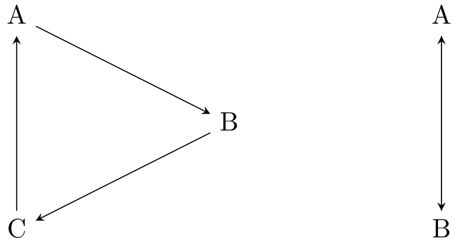

Figure 7.3: Two Causal Diagrams with Cycles

Why can't we have this? Because if we do, then a variable can *cause itself*, and suddenly we've lost all hope of ever isolating the cause of anything, since we can't separate the effect of B on A from the effect of A on B on A from the effect of B on A on B on A... and so on.

But hold on a minute. Surely there are plenty of real-world data generating processes with feedback loops like that. The rich get richer, objects have momentum, and if I punch you that makes you punch me, which makes me punch you.

Surely we're not just going to have to give up any time this happens?

Well, no. But that's because in the true data generating process there *can't* really be any cycles, if you think about it right. That's because of *time*.

Let's think about that punching feedback loop. It certainly seems like the diagram should look like Figure 7.4.

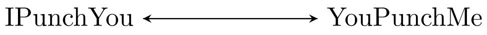

Figure 7.4: A Diagram About Punching

But that's not quite right. After all, if I punch you, and you punch

[12] The formal name for a causal diagram is *directed acyclic graph*. It's even right there in the name! Acyclic!

[13] Some people, when drawing causal diagrams, will use a double-headed arrow like this to indicate that these two variables share a common cause, i.e. A ← U1 → B. But we don't do that in this book.

me back, that doesn't cause me to send the *first* punch—it can't, I already did it. It might, however, cause me to send a *later* punch.

Let's pay attention to *when* these punches are thrown. As is common in statistical applications where time is a factor, let's refer to these time periods as t, $t+1$, $t+2$, and so on, where t is "some particular time," $t+1$ is "the time right after that," and so on. Now the diagram looks like Figure 7.5 and the cycle is gone. The diagram as is only has t and $t+1$ but we could keep going out to the right with $t+2$, $t+3$, and so on, if we wanted!

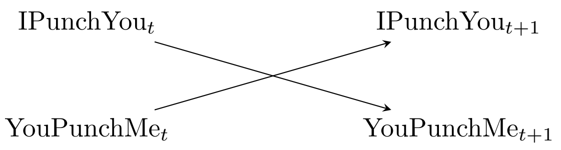

Figure 7.5: Cycles are Even Worse than Punching. But also don't punch people.

Whenever we have a cycle in our diagram, we can get out of it by thinking about adding a time dimension. And it has to work—the cycles pop up because the arrows loop back on themselves. But time's arrow only moves in one direction![14]

Getting Comfortable With Assumptions

In this chapter, I've emphasized that your causal diagram should be based as much as possible about real-world knowledge and prior research. But since we can't possibly know everything about every part of the data generating process,[15] it also contains a lot of assumptions.

That's both necessary and scary! Writing down a diagram like this means sticking your neck out. *This* causes *that*, you have to say. *That* doesn't cause *this*, or at least not enough to draw an arrow. *This other thing* isn't even worth including on the diagram. You think surely this will draw the pitchforks outside your door, or at least a slight disapproving glance from a professor.[16]

But in order to progress, the assumptions do have to be made. And the quality of your research will hinge on how accurate those assumptions are.[17] So how can we get comfortable with the idea that we have to make assumptions, and how can we make those assumptions as accurate as possible?

The convenient thing about empirical work is that assumptions are rarely right or wrong. They're more on a scale of probably-false to probably-true.

After all, if we have to make an assumption it's usually because there's a gap in our knowledge. There's no way to know for sure.

[14] Time travel nonwithstanding.

[15] If we did we wouldn't need to bother doing research, now would we?

[16] And which, you wonder, is worse?

[17] And even great researchers make bad assumptions. Plenty of research goes back to look at old work and finds flaws in its assumptions, replacing them with better ones to progress the field. Heck, the economist Gary Becker was a genius, founded like four major subfields of economics, and probably would have won a second econ Nobel if you were allowed to do that. But pretty much everywhere he made a mark, people have spent decades showing how wrong his assumptions were.

Unlike a math problem it's not up to us to prove we're right, but rather to get a critical reader to think "okay, that sounds plausible. I buy it."

So it's not our job to *prove* we're right, at least not in the mathematical sense of prove, but it *is* our job to get that critical reader to buy it.

That narrows down our work for us. For a given assumption, ask yourself: "Is this probably true? What evidence can I provide to push this away from *possible* and towards *probable?*"

Put yourself in the head of that critical reader. Why might they not believe that assumption? What evidence could they be shown to convince them? Then, produce whatever evidence you can!

For example, let's say you're drawing a diagram of whether door-knocking for a candidate actually increases votes for that candidate. And on your diagram there's no arrow between "how much money the candidate has" and "having a door-knocking campaign."

A reader might think "hold on... surely candidates with more money can more easily afford a door-knocking campaign, right? There should be an arrow there." And to that you would look at whatever evidence you had—prior studies about the effects of candidate campaign coffers, looking in your data, if you have the data, to see if there's a correlation between money and door-knocking. Neither of these things would truly prove that the arrow shouldn't be there, but they might help tip that reader's scales from skeptical to buying-it (and gives you an opportunity to find out that your assumption was in fact wrong so you can fix it).

THAT'S A LOT OF WHAT IT COMES DOWN TO. Think about whether our assumptions are reasonable, try to base them as much on well-established knowledge and prior research as possible, and if we think there's reason to be skeptical of them, ask what evidence would support the assumption and try to provide that evidence.

There are a few other approaches we can take that can help.

The first is just to get another set of eyes on it. It can be hard to be skeptical of your own assumptions—you made them, after all, so you probably think they're pretty reasonable. But maybe there are reasons to be skeptical that you didn't think of! Show your model to another person, especially a person who knows something about the setting or topic you're trying to make a causal diagram for. Or just describe some of the assumptions you made and see what they think. You might be surprised with what they are and are not okay with!

There are also some more formal tests you can do. One nice thing about causal diagrams is that they produce *testable implications* for us. That is, once we have the diagram written down, it will tell us some relationships that *should be* zero. And we can check those relationships in our actual data using basic correlations.[18] If they're not zero, something about our diagram is wrong! We'll talk more about these formal tests in Chapter 8.

Chapter Problems

1. What is a variable in a causal diagram? What types of variables are relevant to the data generating process (i.e., what variables would you want to include in a causal diagram)?

2. Consider this research question: Does defunding public schools cause drop in student achievement for students in the US?

   (a) What is the treatment and what is the outcome of interest?

   (b) Write down a list of relevant variables. Which variables cause/influence the treatment, and which cause/influence the outcome?

   (c) Are the treatment and outcome variables related for reasons other than the causal effect of the treatment on the outcome? Specify.

3. Describe the situations in which each of the following could be used to simplify a causal diagram, or the kind of variable(s) they apply to.

   (a) Unimportance

   (b) Redundancy

   (c) Mediators

   (d) Irrelevance

4. For the list of variables generated in Question 2b, draw a causal diagram.

   (a) Write down if any of the relationships depicted in the diagram fall under the following steps: Unimportance, Redundancy, Mediators, Irrelevance

   (b) Try to simplify based on your answer to Question 4b. Does the simplified diagram still represent the true data generating process? If not, why?

5. How can a causal diagram be modified so as to avoid cyclic relationships?

6. Think of a research question in your field of interest.

   (a) What is the cause variable and what is the outcome variable?

[18] Formally, we should check more kinds of statistics than just correlations, like ones that allow for non-linear relationships. See Chapter 4.

  (b) Write down a list of between 5 and 10 relevant variables in the data generating process. Write down the relationships between all of the variables.

  (c) Draw a causal diagram.

7. Consider the diagram below. It depicts a cyclical relationship between student achievement and motivation. If students achieve more (i.e., score well on exams), then their motivation goes up, and if their motivation goes up, they achieve more. Change the diagram so that the relationship is not cyclic anymore.

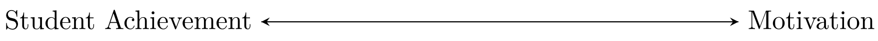

Figure 7.6: Cyclical Causal Diagram for Question 7

8. Consider the research question: Do positive online reviews increase sales of an item?

  (a) Draw a causal diagram with 5-10 relevant variables. Try to simplify the diagram as appropriate.

  (b) Write down a list of questionable assumptions you made in the diagram. Can you support these assumptions with evidence? How might you test them?