Identification

Christoph Hanck

Summer 2023

The data generating process

- A general way how scientists think about reality is that there are regular laws that govern the way the universe works.
- These laws are an example of a *data generating process* (DGP). The laws work behind the scenes, doing what they do whether we know about them or not.
- DGPs are unknown but we observe the data that result from them.

We have a two-part concept of DGPs—the characteristics we do know about (by combining everything we already know about a topic and its underlying laws) and those we do not know.

- The parts we do not know are what we are hoping to learn about using research!

DGPs — an example

Example: A DGP for Income

Assume that we have income data that has been generated from the following laws, but we have no idea what the laws are.

- Income is log-normally distributed
- Being brown-haired gives you a 10% income boost
- 20% of people are naturally brown-haired
- Having a college degree gives you a 20% income boost
- 30% of people have college degrees
- 40% of people who don't have brown hair or a college degree will choose to dye their hair brown

DGPs — an example

Example: A DGP for Income

What can we say about the effect of being brown-haired using *conditional distributions*?

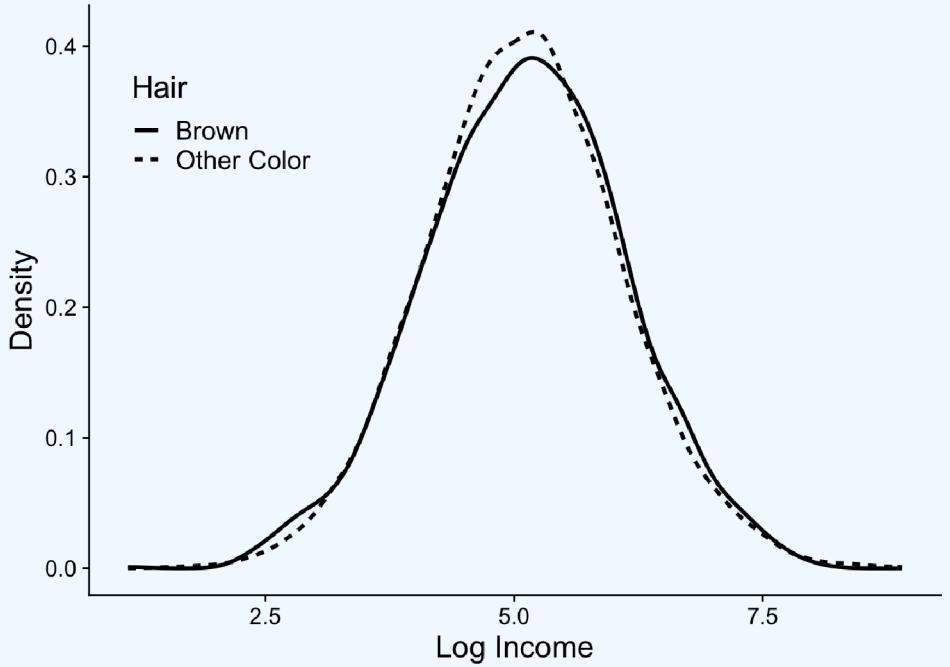

Figure 1: log-income distributions

DGPs — an example

Example: A DGP for Income

What can we say about the effect of being brown-haired using *conditional distributions*?

Hair	**mean log income**
Brown	5.111
Other Color	5.095

Table 1: log-income for all individuals

DGPs — an example

Example: A DGP for Income

- We find that being brown-haired gives you a pay-bump of about 5%:

  We don't see much effect of brown hair because many non-college people have brown hair, but those people don't get the College wage bump, so it looks like brown hair doesn't do much!

- Where can we go from there in order to get the right answer? We have College information in our data, and need an idea of how it fits into the DGP:

  If we know that it's only non-college people who are dying their hair, and that College gives you a bump, we have some alternate explanations for our data.

DGPs — an example

Example: A DGP for Income

Knowing about the DGP also lets us figure out what we need to do to the data to get the right answer. In this DGP, we can notice that among college students, nobody is dying their hair, and so there's no reason we can see why brown hair and income might be related except for brown hair giving you an income boost.

Limiting things just to college students, we see that brown-haired students get a bump of about 10%.

Hair	mean log income
Brown	5.340
Other Color	5.208

Table 2: log-income for college students

DGPs

What did we do?

- **Looking for variation**

  The DGP comprises all mechanisms producing the observed data. We are interested in a part of the observed variation: the variation in income by hair color among college students.

  How can we find the variation we need and focus just on that?

- **Identification**

  How can we use the DGP to be sure that the variation we look at is the right variation? Figuring out what sorts of problems in the data you need to clear away—like how we noticed that the non-college students dying their hair was giving us problems—is the process of identification.

Looking for variation

Example: Avocado sales

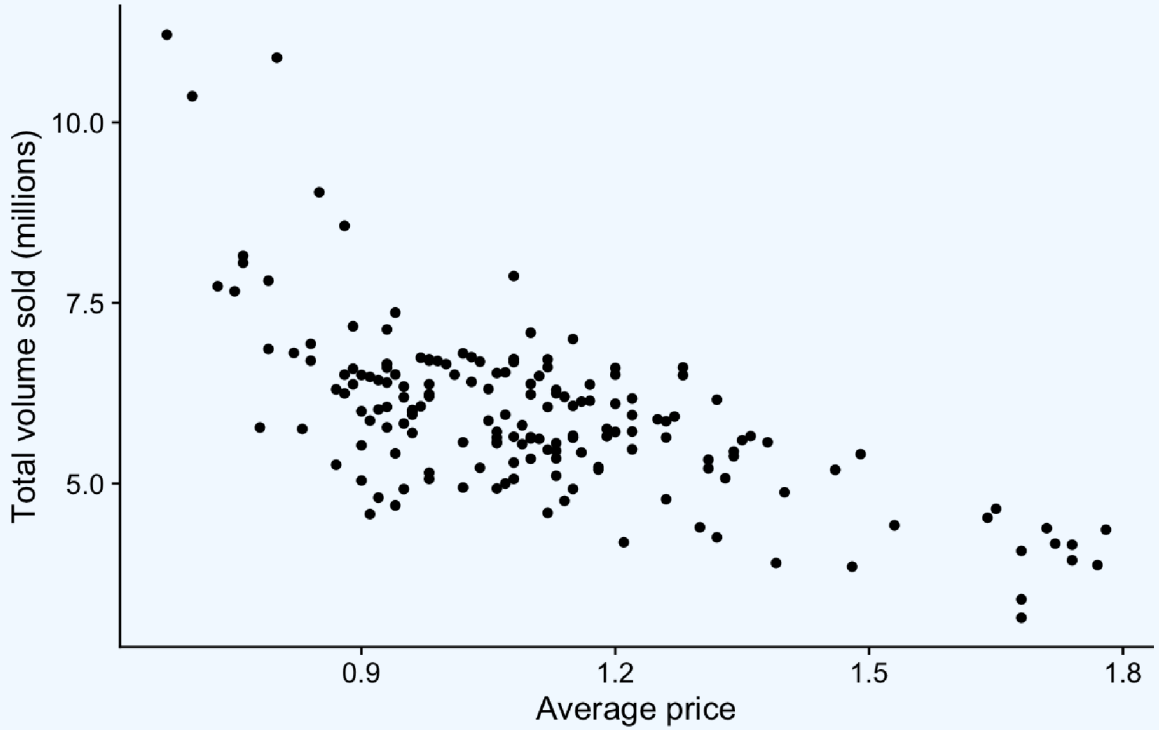

Figure 2: Weekly Sales of Avocados in California, Jan 2015 — March 2018

Looking for variation

Example: Avocado sales

- What conclusion can we draw from Figure 2?

  Avocado sales tend to be lower in weeks where the price of avocados is high and vice versa.
- An interesting research question might be:

  What is the effect of a 10% increase in price on the number of avocados people want to buy?
- Can we answer our research question from using the data shown in Figure 2? Why or why not?

Looking for variation

Example: Avocado sales

- We might be tempted to say something like "an increase in the price of avocados drives down sales." But even that's not actually on the graph!
- All that we *can* see in the graph is the **covariation** or **correlation** between price and quantity of avocados—how they move together or apart. But these variables move around for *all sorts of reasons*!

Looking for variation

Example: Avocado sales

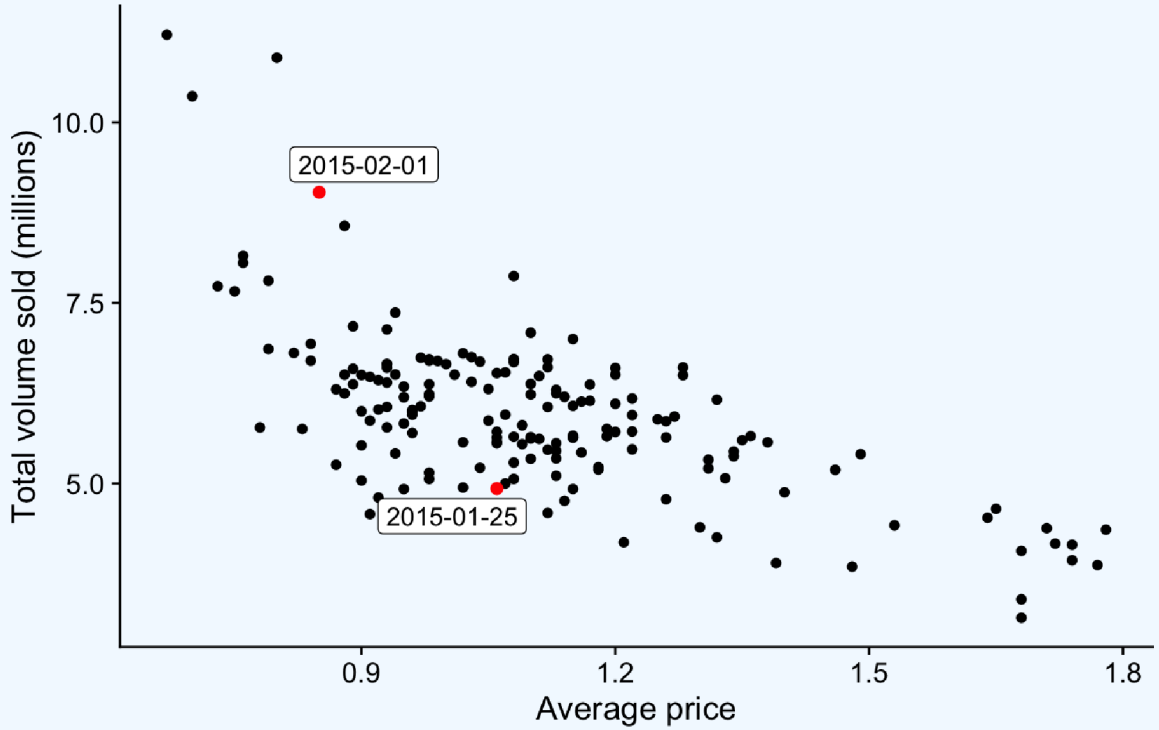

Figure 3: Weekly Sales of Avocados in California, Jan 2015 — March 2018

Looking for variation

Example: Avocado sales

We still see a negative relationship in Figure 3. But why did price drop and quantity rise from January to February that year?

- Is it because a drop in price made people buy more?
- Is it because the market was flooded with avocados so people wouldn't pay as much for them?
- Is it because the high price in January made suppliers bring way more avocados to market in February?

It's probably a little bit of all of them: Variables move around for all sorts of reasons. Those reasons would be reflected in the DGP. But when we have a research question in mind, we are usually only interested in one of those reasons!

Looking for variation

Example: Avocado sales

- How can we find the variation in the data that answers our question?
- We have to ask what is the variation that we want to find? If we want to figure out what the effect of the price is on how many avocados people want to buy, then we want variation in people buying avocados (rather than people selling them) that is driven by changes in the price (rather than, say, avocados becoming less popular).
- We're going to be hopeless at doing this if we don't know anything about the DGP:

  We need to use what we know about the DGP to learn a little more.

Looking for variation

Example: Avocado sales

- Assume that at the beginning of each month, avocado suppliers make a plan for what avocado prices will be each week in that month, and never change their plans until the next month.
- If that's true, then the "suppliers set prices" and "suppliers set quantities" explanations only matter between months. The variation in price and quantity from week to week in the same month will isolate variation in people buying avocados and get rid of variation from people selling avocados.
- Further, because the price is set by the sellers, the variation in quantity we're looking at can only be driven by changes in the price.

Looking for variation

Example: Avocado sales

- Our ability to find this answer is entirely based on that assumption we made about sellers making their choices between months. The reason I've made this particular assumption is that it helps us isolate (identify) only the variation on the part of consumers, conveniently getting rid of variation on the part of sellers and letting us just look at buyers.
- This assumption, which works so much magic for us, is entirely a fiction. Hopefully you will not find all assumptions made for the purposes of digging out variation to be convenient fictions. Some probably are.

Looking for variation

Example: Avocado sales

- By tossing out any variation between months, we're digging through explanations that rely on that variation and tossing them out. Since sellers only change behavior between months (given our assumption), that explanation gets tossed out when we get rid of between-month variation, leaving us only with buyer behavior.
- If we just look at changes within months, as in Figure 4, we can see that there's still a negative relationship. Note that for each of the months, there's a negative relationship, ignoring any differences between the months. So, given the data and what we know about how sellers operate, an increase in price does reduce how many avocados people want to buy.

Looking for variation

Example: Avocado sales

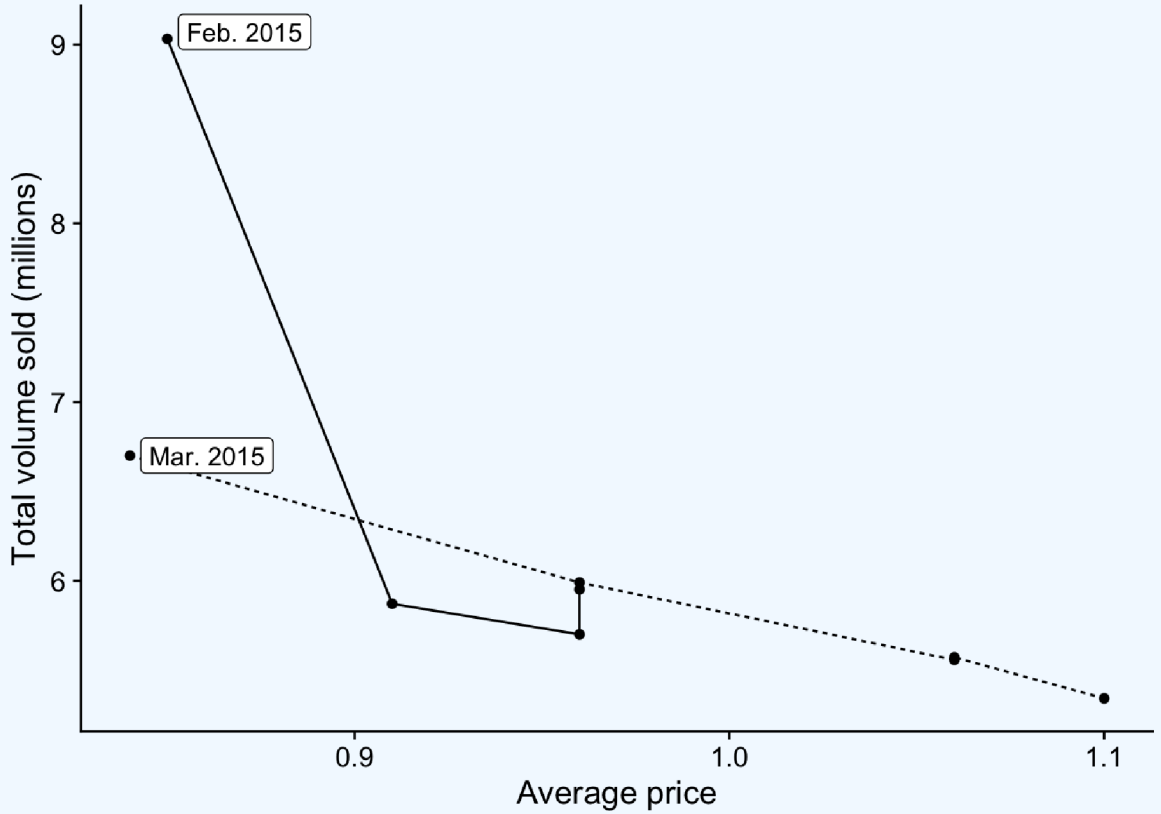

Figure 4: Weekly Sales of Avocados in California, Feb. 2015 — March 2015

Looking for variation

Example: Avocado sales

- The task of figuring out how to answer our research question is really the task of figuring out where your variation is. It's unlikely that the variation in the raw data answers the question you're really interested in. So where is the variation that does answer your question? How can you find it and dig it out? What variation needs to be removed to unearth the good stuff underneath?
- That process - finding where the variation you're interested in is lurking and isolating just that part so you know that you're answering your research question - is called identification.

Identification

- Identification is the process of figuring out what part of the variation in your data answers your research question.
- A **research question takes us from theory to hypothesis**, making sure that the hypothesis we're testing will actually tell us something about the theory.
- **Identification takes us from hypothesis to the data**, making sure that we have a way of testing that hypothesis in the data, and not accidentally testing some other hypothesis instead.

Example: Avocado sales — Identification

Having an idea of the DGP (the different ways that prices and quantities might be related), we know what work we need to do for identification:

We closed off undesirable explanations by getting rid of between-month variation driven by sellers. The only way for price to affect quantity at that point is through the consumers.

Identification

Identification requires statistical procedures in order to isolate the variation we are interested in. Just as important, **identification critically relies on the theory and the assumptions about DGP**.

Example: Avocado sales — Identification

Consider again the avocado example:

- We used our knowledge about markets to realize that sellers might be setting price in response to the quantity—an alternate explanation we need to deal with in isolating the variation of interest.
- We then used an assumption about how sellers set prices to figure out how we can block out this alternate explanation by looking within-month.

Identification

Summary:

If we want to identify the part of our data that gives the answer to our research question, we must:

1. Using theory, paint the most accurate picture possible of what the DGP looks like
2. Use that DGP to figure out the reasons our data might look the way it does that *do not* answer our research question
3. Find ways to block out those alternate reasons and to isolate the variation we need

This process is a lot more difficult than just “look at the data and see what it says.” But if we don’t go the extra mile of following these steps, we can end up with confusing, inconsistent, or just plain wrong results.

Let us see what may happen when we do not take identification seriously enough...

Study: Alcohol and mortality

A. M. Wood et al. (2018)[1] investigate the relationship between drinking and outcomes like mortality and cardiovascular disease among nearly 600,000 people.

- no benefit of small amounts of drinking.
- Amount of alcohol it took to start noticing increased risk for serious outcomes was at about 100 grams of alcohol per week (which is about a drink per day, and below current guidelines in some countries)

Figure 5 shows the relationship they found between weekly alcohol consumption and the chances of mortality. Mortality starts to rise at around 100 grams of alcohol per week, and goes up sharply from there.

[1] Wood, Angela M, Stephen Kaptoge, Adam S Butterworth, and 239 more. 2018. “Risk Thresholds for Alcohol Consumption: Combined Analysis of Individual-Participant Data for 599,912 Current Drinkers in 83 Prospective Studies.” The Lancet 391 (10129): 1513–23. https://doi.org/10.1016/S0140-6736(18)30134-X.

Study: Alcohol and mortality

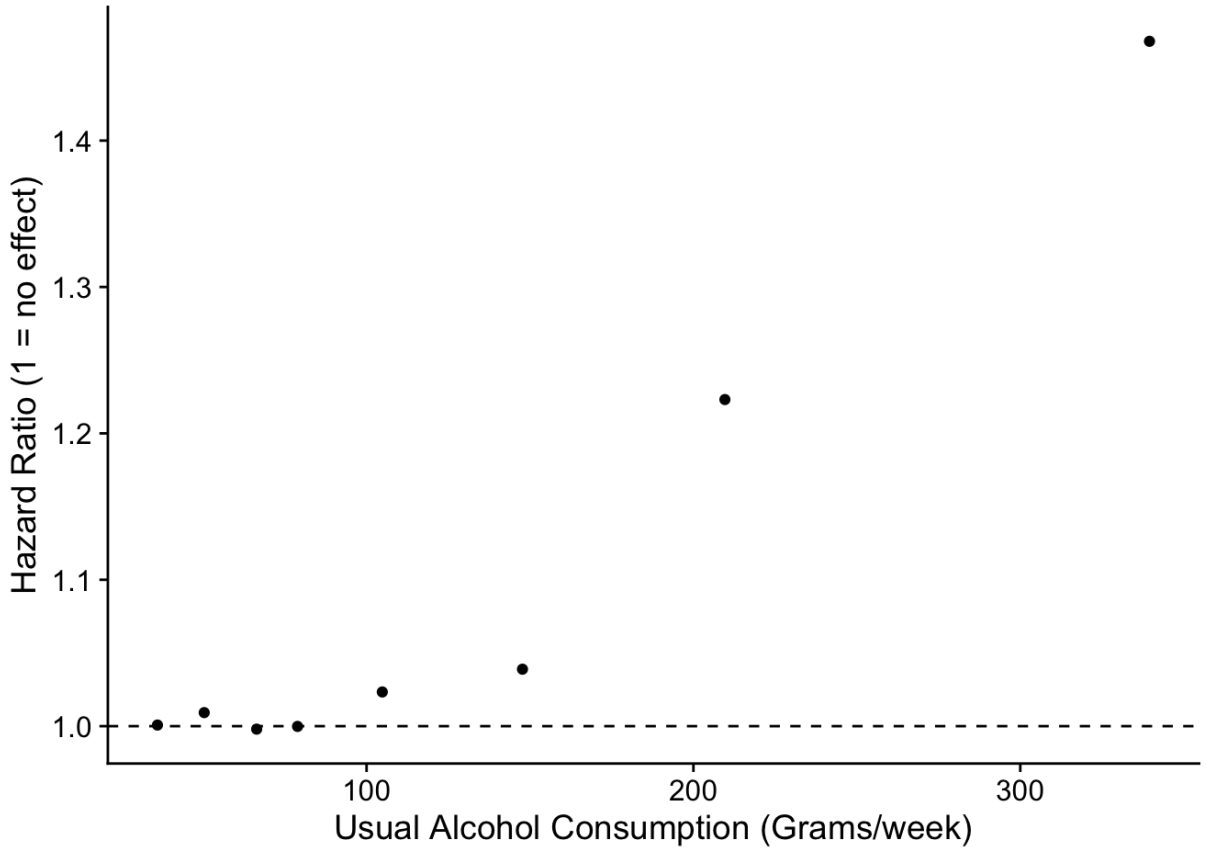

Figure 5: Alcohol consumption and mortality from Wood et al. (2018). Minor changes from original.

Study: Alcohol and mortality

Think about the DGP:

- What leads us to observe people drinking?
- What leads us to observe them dying?
- What reasons might there be for us to see an association between alcohol and mortality?

  Think about determinants of drinking and mortality:

  - Tendency to take more risks may results in more drinking and other unhealthy activities, e.g., smoking
  - Increased mortality risk may result from many bad-health indicators (including drinking and smoking)

Study: Alcohol and mortality

Think of alternate explanations

Smoking is likely related to alcohol consumption and high mortality and thus may be an alternate explanation for the results.

- Smokers could be more likely to drink:

  The relationship between drinking and mortality might just be that smokers tend to drink, and also die earlier because of their smoking.

- Anything else that ends up on both lists is going to give us an alternate explanation.

Study: Alcohol and mortality

Would non-drinkers would have very low mortality rates?

Maybe, but people

- ... may choose not to drink at all because their health is too poor to handle it.
- ... will not drink if they are recovering alcoholics. In these cases, we may actually see worse mortality for non-drinkers, but that relationship is almost certainly *not* because not-drinking is bad for them.

Study: Alcohol and mortality

Did Woods et al. (2018) manage to deal with some of these alternate explanations?

Yes:

- Note that Figure 5 does not contain non-drinkers. They have been left out of the study to block out the too-sick-to-drink and ex-alcoholic alternate explanations.

  This is one reason why the study does not find a positive effect of a little alcohol while other studies do—those that leave in the non-drinkers!
- Woods et al. (2018) also use statistical adjustment to account for further alternate explanations like smoking, age, gender, and a few health variables like BMI and diabetes indicators

Study: Alcohol and mortality

Did Woods et al. (2018) find the "true" effect of drinking on mortality?

Not necessarily:

- It is impossible to account for all conceivable alternate explanations related to risk-taking
- Omission of all non-drinkers disregards non-drinkers who are neither sick nor ex-alcoholics.

  What if some very sick people just choose to drink less rather than not at all?

Thus it might be a little premature to take these results, despite the hundreds of thousands of people they examined, and use them to conclude that we have now *identified* the effects of alcohol on mortality.

If it feels like they did their part in addressing some of the alternate explanations and what's left over feels trivial, keep in mind that these alternate explanations can lead us to very strange conclusions...

Study: Alcohol and mortality

If the methods can give us Figure 6 by Auld (2018)[2], then even if there is really an effect of alcohol on mortality, how close do we think Figure 6 is to identifying that effect?

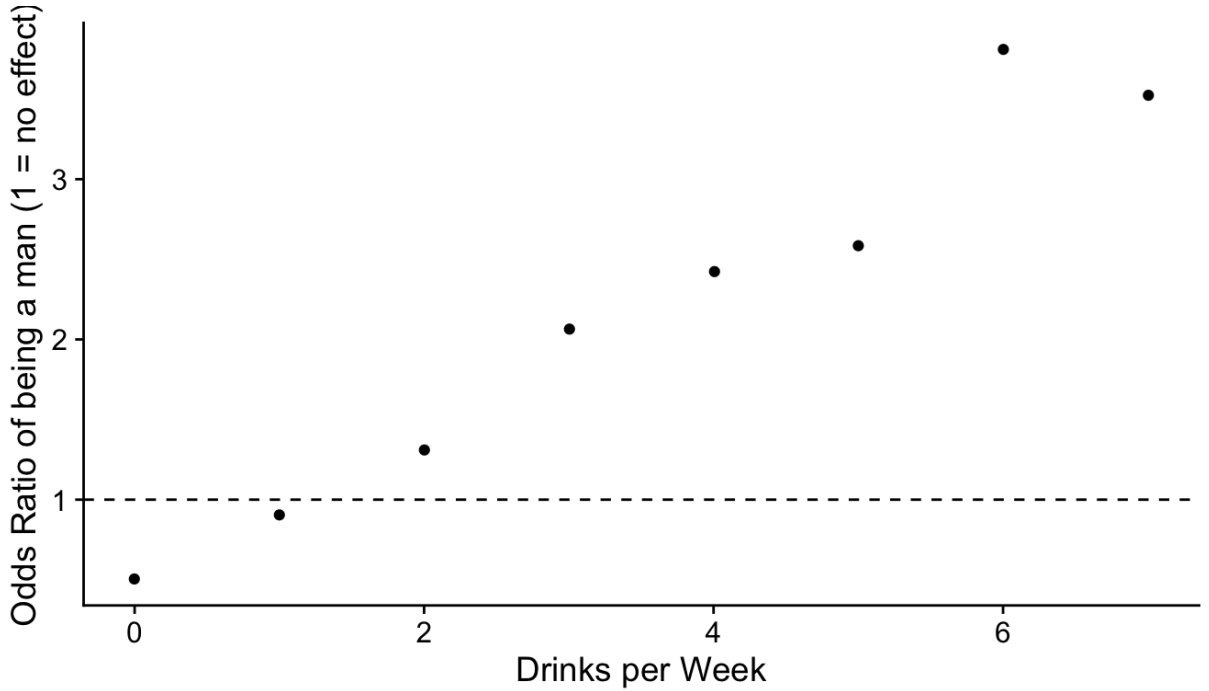

Figure 6: Alcohol Consumption and Being a Man

[2] Auld, Chris. 2018. “Breaking News!” https://twitter.com/Chris_Auld/status/1035230771957485568.

Do we need to be *omniscient*?

No research project is perfect.

We should try to

- ... learn what we can about the context so that we do not miss any hugely important part of the DGP.
- ... be careful to acknowledge what assumptions we are making, and think about how they might be wrong.
- ... spot gaps in our knowledge about the DGP, and make some realistic guesses about what might be in that gap.
- ... aim for getting as close as we can, instead of aiming for perfection.