Using Directed Acyclic Graphs (DAGs) to describe and understand causal relations

Jesse Brunner

2021-11-02

An Example

Imagine you are interested in explaining the size of amphibians at metamorphosis in a series of vernal ponds. You have measured:

- The snout-vent-length, or **SVL**, of metamorphosing frogs
- **Area** of the ponds
- **Nutr**ient concentrations entering the ponds (say, all sources of nitrogen, for simplicity)
- The growth of algal biomass as **Algae**
- **Density** of tadpoles in the pond

Overall, your hypothesis is that pond size will influence the size of metamorphosing frogs. What should you include in your regression?

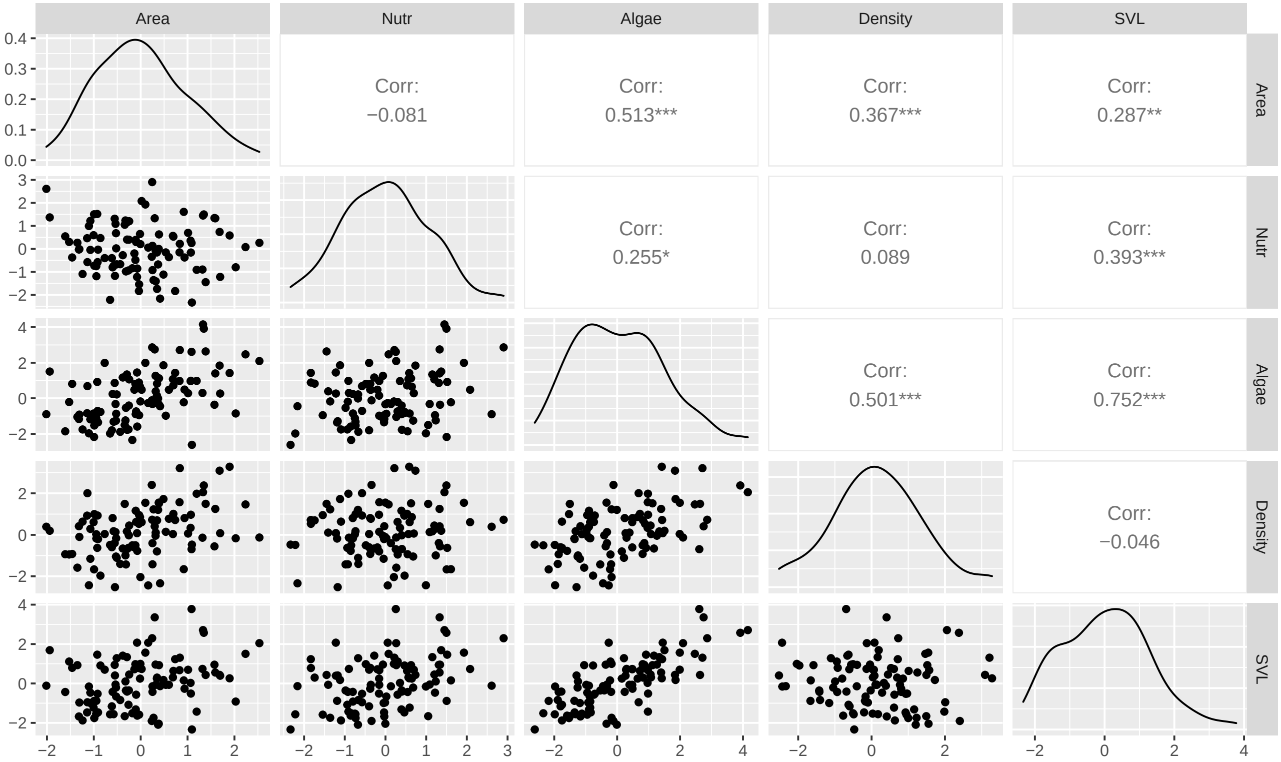
I have made up data so our example can be concrete. Note that every variable is normally distributed and standardized so that it is centered on zero and one represents one standard deviation from the mean. I've also simulated the data so everything is a linear regression. This is about as nice and neat as we might hope!

I've plotted a scatter plot-matrix and you can see some variables are strongly correlated and others much less so.

So my question for you is, *what variable(s) should you include in a regression* to understand the influence of pond size on frogs' size at metamorphosis (SVL)?

I would suggest that there are two basic approaches many people would suggest. First, many would suggest using individual predictors in separate regressions. Here are those (in R, but hopefully you can find familiar terms in the output) individual regression result:

```
##
## Call:
## lm(formula = SVL ~ Area - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7480 -0.6344 -0.0766  0.8086  3.3670
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
```



```
## Area   0.3777     0.1266   2.984  0.00358 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.214 on 99 degrees of freedom
## Multiple R-squared:  0.08252,    Adjusted R-squared:  0.07325
## F-statistic: 8.904 on 1 and 99 DF,  p-value: 0.003584
```

```
##
## Call:
## lm(formula = SVL ~ Algae - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08543 -0.58781  0.02514  0.45395  2.24655
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## Algae  0.66700    0.05861   11.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.834 on 99 degrees of freedom
## Multiple R-squared:  0.5668, Adjusted R-squared:  0.5624
## F-statistic: 129.5 on 1 and 99 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = SVL ~ Nutr - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1521 -0.7468 -0.0532  0.8904  3.6577
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## Nutr   0.4690     0.1103    4.25 4.84e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.165 on 99 degrees of freedom
## Multiple R-squared:  0.1543, Adjusted R-squared:  0.1457
## F-statistic: 18.06 on 1 and 99 DF,  p-value: 4.845e-05
```

```
##
## Call:
## lm(formula = SVL ~ Density - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3551 -0.9326  0.0041  0.8380  3.7477
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## Density -0.04203    0.10307  -0.408    0.684
##
## Residual standard error: 1.266 on 99 degrees of freedom
## Multiple R-squared:  0.001677,   Adjusted R-squared:  -0.008407
## F-statistic: 0.1663 on 1 and 99 DF,  p-value: 0.6843
```

The other approach is to throw every measured variable into the regression and let the statistics sort it out. Here is the output from this *full* model:

```
##
## Call:
## lm(formula = SVL ~ Area + Nutr + Algae + Density - 1)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -1.24652 -0.28528  0.06842  0.38049  1.13694
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## Area       0.00354    0.06578   0.054    0.957
## Nutr       0.22973    0.05255   4.372 3.12e-05 ***
## Algae      0.86894    0.04887  17.782  < 2e-16 ***
## Density -0.56727    0.04942 -11.478  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5189 on 96 degrees of freedom
## Multiple R-squared:  0.8373, Adjusted R-squared:  0.8306
## F-statistic: 123.6 on 4 and 96 DF,  p-value: < 2.2e-16
```

A close inspection suggestions that we might end up with different results depending on how we do it. Let me make it more obvious by plotting the parameter estimates from the regression from the individual models and the corresponding parameter from the full model.

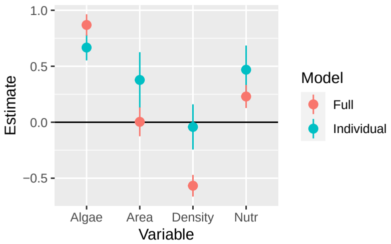


Figure 1: Estimated coefficients when estimated individually or in a full model. Vertical lines are 95 percent CIs.

So what is the right answer to the question of how pond area affects size at metamorphosis? (Or similarly, if we were interested in any of the other variables, which model should you listen to?)

Maybe a better question is, why is this so hard? That one, at least, I can answer now. It is difficult to know what each type of model is telling us because we have not specified how we think things work in this system. Statistical models, including linear regressions, are simply association machines. No matter what you have been told, regressions cannot tell us what caused what, at least not by themselves. We need to graph out these relationships ourselves, outside of the statistics. They can then help us understand what the regressions are telling us (contingent on our graphs or models being right!). We will call these DAGs.

What is a DAG?

A “DAG” is a **d**irected, **a**cyclic **g**raph.

- directed: we are using arrows to describe causal influence
- acyclic: no cycles or loops, where $A \to B \to C \to A$
  - positive or negative feedbacks means what you expect to see depends on *when* in the process you are looking
- graph: nodes (=variables) connected by arrows (=causal relationships)

Drawing causal relationships

The basics of drawing a causal graph or diagram are simple:

- Write out the variables that are important in your little piece of the system
  - include both "predictors" and "responses" (Remember, our statistics do not "know" which is which!)
  - By convention, things you have measured are unadorned: e.g., X, Y, Z
  - Things you have not measured (or are unobserved) are, for the purposes of this handout, circled: Ⓤ
- Draw arrows showing (assumed) *causal* relationships connecting variables (e.g., $X \rightarrow Y$ means "changes in X causes changes in Y")
  - Note that we are not drawing the *order* of things
  - The arrows do not describe the *direction* or *shape* of the relationships, just the influence
  - Arrows do not show interactions, either
- Keep it simple. While you can, of course, draw whatever web of causal relationships you like, just as with any other model, the more complicated it is, the more difficult it is to understand and work with.

Also, try drawing alternate versions representing your hypotheses of how the system works.

Back to our example

It would be worth spending a moment thinking about how *you* would draw a DAG for our size-at-metamorphosis example. Even if you are uncertain about how the system might work, trying to draw a DAG can help refine your uncertainty or help you see the questions you need to ask. But that said, let me offer a couple reasonable versions.

```r
library(dagitty)

dag1 <- dagitty("dag{
Area -> Algae -> SVL
Area -> SVL
Nutr -> Algae
Density -> SVL
Area [exposure]
SVL [outcome]
}")
```

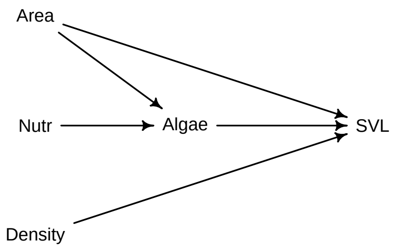


Figure 2: First DAG

This first version suggestions that pond area ("Area"), perhaps simply because there is more sunlight available, and the influx of nutrients like nitrogen

("Nutr") influence algal growth ("Algae"), which in turn influences SVL. Pond area also has a direct effect on SVL, as does the density of tadpoles ("Density"). Does that sound reasonable? Me, I might wonder how pond area directly affects the size of metamorphosing tadpoles.

```
dag2 <- dagitty("dag{
Area -> Density -> SVL
Area -> Algae -> Density
Nutr -> Algae -> SVL
Area [exposure]
SVL [outcome]
}")
```

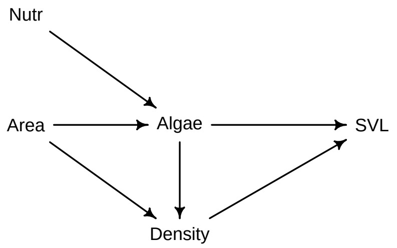


Figure 3: Second DAG

The second version suggests that the amount of algal growth is determined by the area of the pond and nutrients flowing into the pond. Both the area of the pond and algal growth affect the density of tadpoles; perhaps larger ponds attract more breeding females in the spring and more food keeps more tadpoles alive. I would guess that greater algal growth increases the size at metamorphosis (SVL) and that higher densities decrease it. That seems a bit more reasonable. But perhaps the nitrogen influx into a pond has a direct effect on size at metamorphosis because the algae are of higher quality.

```
dag3 <- dagitty("dag{
Area -> Density -> SVL
Area -> Algae -> Density
Nutr -> Algae -> SVL
Nutr -> Q -> SVL
Area [exposure]
SVL [outcome]
Q [unobserved]
}")
```

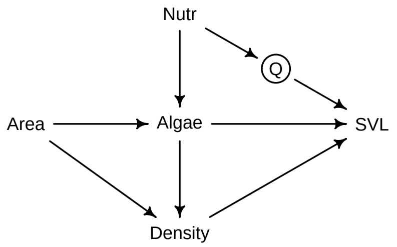


Figure 4: Third DAG

In the third version we've included this relationship from nutrients ("Nutr") to the unobserved variable "Q", for food quality, to SVL.

Implied conditional independencies

What, you are likely asking, have we gained by drawing out these DAGs? There are a few gains, but let me first focus on the testable implications of the DAGs. If you look back at the first DAG, you can see that the two variables, Density and Nutrients are independent of each other, as are Density and Area, Area and Nutrients, and Density and Algae. That is, if you were to look for some statistical association, this DAG suggests you should not find any. Or put another way, knowing the value of one of the variables in each pair tells you nothing about the value of the other.

(Importantly, all of these “conditional independencies” are, well, conditional on, in this case, not knowing SVL. If we included SVL in our statistical model [e.g,. regression] then Area and Density would no longer be independent of one another. If, say, we know the pond produces large metamomorphs and we also know the pond is small, we could guess with some confidence that the density in that pond must be pretty low, too. We’ll come back to this soon.)

There is one more implied conditional independence: Nutrients will be independent of SVL *if* we condition on or include in our regression Algae and Area. Let’s think about why. Because Nutrients act through their influence on Algae, our DAG says that if we already know what Algae is, then Nutrients do not add any more information. This is called *conditioning on a mediator* because the effects of Nutrients are mediated by algal growth.

Why, you might ask, do we also need to condition on Area to make Nutrients and SVL independent? This is because Area also has an influence on Algae. Imagine a nutrient poor environment that still had moderately high algal growth. According to this model, that could only happen if the pond area were large, and so knowing if we do not also know pond Area, Nutrients still tell us something about SVL even if we know Algae. I know, it’s a bit headache-inducing, but it will get easier with practice.

While it is good to try to puzzle out these independencies yourself, it turns out the logic of them is pretty rote and so computers can do it just fine. In the R package `dagitty`, which I’ve been using to plot the DAGs[1] there is a function with the catchy name, `impliedConditionalIndependencies`. It can tell you those implied conditional independencies. (If you do not use R, you can draw and analyze your DAGs at http://dagitty.net/dags.htm. There are also interactive lessons at http://dagitty.net/learn/index.html.)

```
impliedConditionalIndependencies(dag1)
```

```
## Alga _||_ Dnst
## Area _||_ Dnst
## Area _||_ Nutr
## Dnst _||_ Nutr
## Nutr _||_ SVL | Alga, Area
```

Two notes on the notation The “$X \perp\!\!\!\perp Y$” (`X _||_ Y`) notation means that X is independent of Y. ($X \not\!\perp\!\!\!\perp Y$ would mean that X is *not* independent of Y.) Second, the $|$ (`|`) symbol means “given” or “conditioned on” the stuff to the right. So `Nutr _||_ SVL | Alga, Area` means Nutrients are independent of SVL conditioned on (or given knowledge of) Algae and Area.

We can find the implied conditional independencies for the other two models.

[1] Full disclosure, I’m actually using a plotting function in the package `rethinking`, because it circles unobserved variables.

```
impliedConditionalIndependencies(dag2)
```

```
## Area _||_ Nutr
## Area _||_ SVL | Alga, Dnst
## Dnst _||_ Nutr | Alga, Area
## Nutr _||_ SVL | Alga, Dnst
## Nutr _||_ SVL | Alga, Area
```

```
impliedConditionalIndependencies(dag3)
```

```
## Area _||_ Nutr
## Area _||_ SVL | Alga, Dnst, Nutr
## Dnst _||_ Nutr | Alga, Area
```

Notice that the implied conditional independencies are not the same between the three versions of the DAG. This can give us a way to test and contrast our various DAGs. For instance, we could test if pond area is independent of density.

What do I mean by independent of? A quick and dirty definition would be that the parameter estimate for, say, a regression of Density on Area is essentially indistinguishable from zero.

```
summary(lm(Density ~ Area))
```

```
##
## Call:
## lm(formula = Density ~ Area)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6743 -0.7181  0.1220  0.8862  2.6795
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1436     0.1146   1.253 0.213024
## Area          0.4669     0.1195   3.908 0.000172 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.146 on 98 degrees of freedom
## Multiple R-squared:  0.1348, Adjusted R-squared:  0.126
## F-statistic: 15.27 on 1 and 98 DF,  p-value: 0.0001716
```

In this case, it looks like Density does increase discernibly with Area, which would suggests that the version of the system represented by the first DAG is probably not correct.

It is important to note that some DAGs will not have testable implications. Also, sometimes different DAGs will have essentially identical implied conditional independencies meaning one cannot differentiate the DAGs based only on these associations. DAGs are useful tools, not magic.

A note on causation and statistics

You have probably noticed that our DAGs have not, so far, told us about causation[2]. That is not an accident; our understanding of causation does not come from a DAG or any other model. Rather, DAGs just tell us the (implied) consequences of the causal model we assume. That is super useful, but it does not get us, as scientists, off the hook for sorting out causal relationships[3].

In a very real sense, our understanding of causation happens in our thinking, our conversations with colleagues, the interplay of different studies. We come to understand causal relationships by *consensus*, not by *statistics*.

There might be one sort of exception to this: experiments. Of course our statistics do not know whether we did an experiment or just an observational study and we have no way of telling them. But experiments are a bit magical because they break the associations between variables. Rather than conditioning on, say, algal growth, you can see what happens when you add or remove algae *while keeping everything else the same*. That let's you discern the effect of algae by itself, free of all the correlated changes. That's powerful! We can and often do sort out causal relationship in the absence of experiments[4], but they sure do help!

The four elemental relationships

We can gain further insights into our DAGs by thinking about how information flows between variables. This is easier if we identify the basic ways that variables can be related. It turns out that there are only four ways that three variables can be related, which makes it easy. (As I describe them, look back at the previous DAGs and see if you can identify each of them. Note: they might not all be present in all of the DAGs.)

1. **Pipe**: Here the causal influence of X on Y is through the intermediate variable Z.

$$X \to Z \to Y$$

   This means that if we were to condition on the intermediate, Z, X and Y should be independent of each other.

```
impliedConditionalIndependencies(dagitty("dag{X -> Z -> Y}"))
```

```
## X _||_ Y | Z
```

[2] Well, we can get a sense of which DAGs might be *wrong*!

[3] There might be multiple DAGs and thus multiple causal models consistent with our data, for instance. And all of them are simplifications of reality.

[4] I am a big fan of the late Sir Austin Bradford Hill "criteria" for thinking about evidence of causal relationships when experiments are not possible. His original paper on this topic is very readable.

2. **Confound**: In this case the variable Z affects both X and Y.

$$X \leftarrow Z \rightarrow Y$$

You might be surprised to see that the implied conditional independence is the same as for the pipe. (I was when I first learned these!)

```
impliedConditionalIndependencies(dagitty("dag{X <- Z -> Y}"))
```

```
## X _||_ Y | Z
```

Why are X and Y *not* independent unless we condition on Z? After all, X does not *cause* anything to do with Y, and vice versa. The reason is that while causation might flow in one direction (or in this case, in two directions away from Z), information flows both ways. Think of it this way: Imagine X and Y both increase with Z. Thus, if we know that X is small, that implies that Y must also be small. This also works if X is positively related to Z and Y is negatively related, or vice versa; knowing the value of one gives us information about the other. This flow of information is only interrupted it we know (condition on) Z. In that case, knowing X does not give us any extra information about Y that is not already given to us by knowing Z[5].

3. **Collider**: This is the opposite of the confound, where Z is influenced by both X and Y.

$$X \rightarrow Z \leftarrow Y$$

```
impliedConditionalIndependencies(dagitty("dag{X -> Z <- Y}"))
```

```
## X _||_ Y
```

In this case there is no information flow from X to Y (or vice versa); knowing X tells us nothing about Y. That is, *unless* we condition on Z. If we know Z, then information flows between X and Y.

This takes a bit of thought, or perhaps an example[6]. Imagine Z is a light bulb, either on or off, and X is a light switch (again, on or off) and Y indicates whether there is a source of electricity working. If all you know is that the light switch is on ($X = 1$), you know nothing about whether there is an electric source ($Y = ?$). If, however, you also knew that the light bulb was shining ($Z = 1$), then you could easily infer that there must be electricity available ($Y = 1$). Knowing the value of (or conditioning on) the collider, Z, lets information flow between X and Y[7].

4. **Descendant**: This is like the collider, but now instead of focusing on (or conditioning on) Z we have a variable than comes from Z. It is sort of a half-way collider.

$$\begin{tikzcd}
Y \arrow[dr] & & \\
 & Z \arrow[r] & D \\
X \arrow[ur] & &
\end{tikzcd}$$

[5] Does your brain hurt yet?

[6] I'm stealing this from Richard McElreath's excellent book, Statistical Rethinking.

[7] Try thinking through more scientifically interesting examples, like $G \rightarrow H \leftarrow E$, where G is genetics, E is the environment, and H is height.

Again, X and Y are independent of each other unless one were to condition on D (or Z). If, however, you were to include or condition on Z then D would be independent of X and Y, but of course then you'd be ensuring X and Y were not independent.

```
impliedConditionalIndependencies(dagitty("dag{X -> Z <- Y; Z -> D}"))
```

```
## D _||_ X | Z
## D _||_ Y | Z
## X _||_ Y
```

Closing the right doors[8]

Given these four elemental relationships you have a bit better sense of how information flows between variables. This is important because it allows us to a) better understand what a parameter in a regression is telling us and, if we're lucky, b) what to condition on to ensure the parameter estimate means what we want it to mean.

For instance, in the full regression model, in which we conditioned on *everything*, we saw that the regression coefficient for Area was essentially zero. We might now recognize that in the second and third DAGs this would be expected, because by conditioning on Density and Algae (DAG 2) or Density, Algae, and Nutr (DAG 3) we have made Area independent of SVL. We have, if we believe these DAGs, demonstrated that Area does not have a direct influence on SVL, it only acts through its influences on Algae and Density. (In DAG 1 we would still expect to see a direct influence of Area on SVL, so if this DAG were "correct," it would suggest a very small direct effect.)

Now what if our questions is simply what is the total influence of Area on SVL? In that case we want to look at the relationship between Area and SVL, *without* conditioning on anything else.

We can again use software to help us identify the covariate(s) we need to condition on to obtain an unbiased estimate of the causal effect of one variable on another, *assuming the DAG is correct*. Notice that when I defined the DAGs above I wrote `exposure = Area` and `outcome = SVL`. This was how we tell the software what is the "exposure" or putative cause and what is the response or "outcome." We can then use the function `adjustmentSets` in `dagitty`.

```
adjustmentSets(dag1)
```

```
##  {}
```

[8] Closing a path through which information can flow is called "closing a door." Then there's the "backdoor rule," where information flows through a non-causal path and the "single door rule" and so on. I'll let you look those up.

```
adjustmentSets(dag2)
```

```
##  {}
```

```
adjustmentSets(dag3)
```

```
##  {}
```

In each case we get the empty set. That means that in these DAGs we do not want to condition on anything to understand the effect of Area on SVL. The simple model would do it!

If, however, we were interested in the influence of Algae on SVL in the third DAG we could use this code:

```
adjustmentSets(dag3, exposure = "Algae", outcome = "SVL")
```

```
## { Area, Nutr }
```

This means we would want to condition on both Area and Nutrients, but *not* Density. Useful, no?

Again, you may not always have a simple solution. Perhaps the causal structure is just tangled or you didn't or couldn't measure some important variable. In certain cases you may not be able to obtain unbiased estimates of the effects you want. But I think it is better to know this than to proceed in the dark.

Simpson's paradox

Proceeding by intuition or worse, just throwing variables into a regression and hoping for the best, can lead to problems, big ones. A classic example is called "Simpson's paradox." In it, the model structure is such that the estimated effect of X on Y not only changes, but reverses *sign*, depending on which other variables are included in the regression (i.e., conditioned on). It's meant to serve as a warning, so let's see and then head this cautionary tale[9].

Here's the DAG.

Remember that we are interested in understanding the effect of X on Y. If it were you doing this, which covariate(s) would you include?

As you can see, if you just regress Y on X you see a very strong, positive effect. However, if you condition on (include in the model) $Z2$ or both $Z1$ and $Z2$ you get a *negative* effect of X on Y! Only when you condition on either $Z1$ or $Z1$, $Z2$, and $Z3$ do you get the right sign and magnitude of the effect[10]!

There are a few points to make. First, in this case we got the right answer when we threw all of the measured variables into the model, but this is not always the case. Sometimes those extra variables will the ones that give you

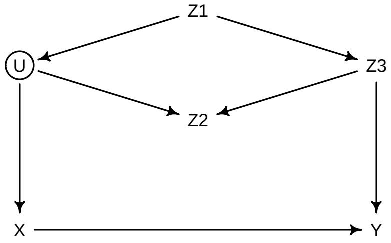


Figure 5: The DAG assumed in one version of Simpson's paradox

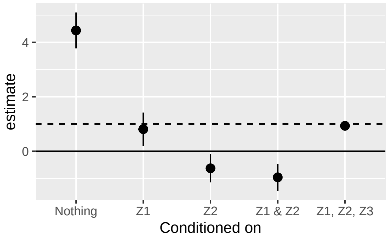


Figure 6: Estimated effect of X on Y when including various parameters in the regression

[9] This example and code is coming straight from http://dagitty.net/learn/simpson/index.html.

the wrong magnitude or sign of the effect. Second, this does happen in real-world situations. It is not simply an edge-case meant to scare you, but a real effect that can really happen[11]. Third, we knew the right answer because we simulated the data, but if you were working with real-world data would you know the right answer? Probably not! All we will know are the data we collected and the DAG(s) we are willing to assume.

Some final notes

Using DAGs can help you make sense of the many statistical associations between variables. They can help you focus on what you think is reasonable and what you actually want to know. Sometimes they can help you toss out or provisionally accept as consistent with the data certain causal models. Other times they can help you see why you are hosed in sorting out the independent effects you seek.

DAGs can also be useful in planning studies, sorting out what data you will need to make the inference you desire. For instance, see what happens if you treat a variable as observed vs. unobserved. DAGs become even more useful if you use them to simulate data. That gives you a chance to see if your planned analyses can distinguish between alternative causal models or provide unbiased estimates of the causal influence of key variables. If you can recover the True estimates from simulated data, this should give you some confidence that you might be similarly successful with real data. If you can't, perhaps you need to redesign your study.

Finally, it is worth beating into our collective psyche that models cannot, by themselves, tell us anything about causation. They can simply quantify associations, in the case of statistical models, or show us the consequences of the assumptions we are making, as with DAGs or other scientific models. It takes us—hard working, harder thinking scientists—to decide on causal relationships. I hope that giving you a brief introduction to DAGs might help in this important goal.

[11] Is that enough reals for you?