Using DAGs to describe & understand causal relations

Jesse Brunner

2022-08-31

A motivating 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 in biomass of **Algae**
- **Density** of tadpoles in the pond earlier in the season

Overall, your hypothesis is that pond size will influence the size of metamorphosing frogs. What should you include in your statistical analysis (e.g., regression) to estimate or test this effect?

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 with a standard deviation of one. I've also simulated the data so everything is a linear regression[1]. This is about as nice and neat as we might hope! This scatter plot-matrix shows how these variables relate to each other. Some are strongly correlated and others much less so.

[1] Note that nothing I present here *requires* relationships to be linear. It is just simpler to work with
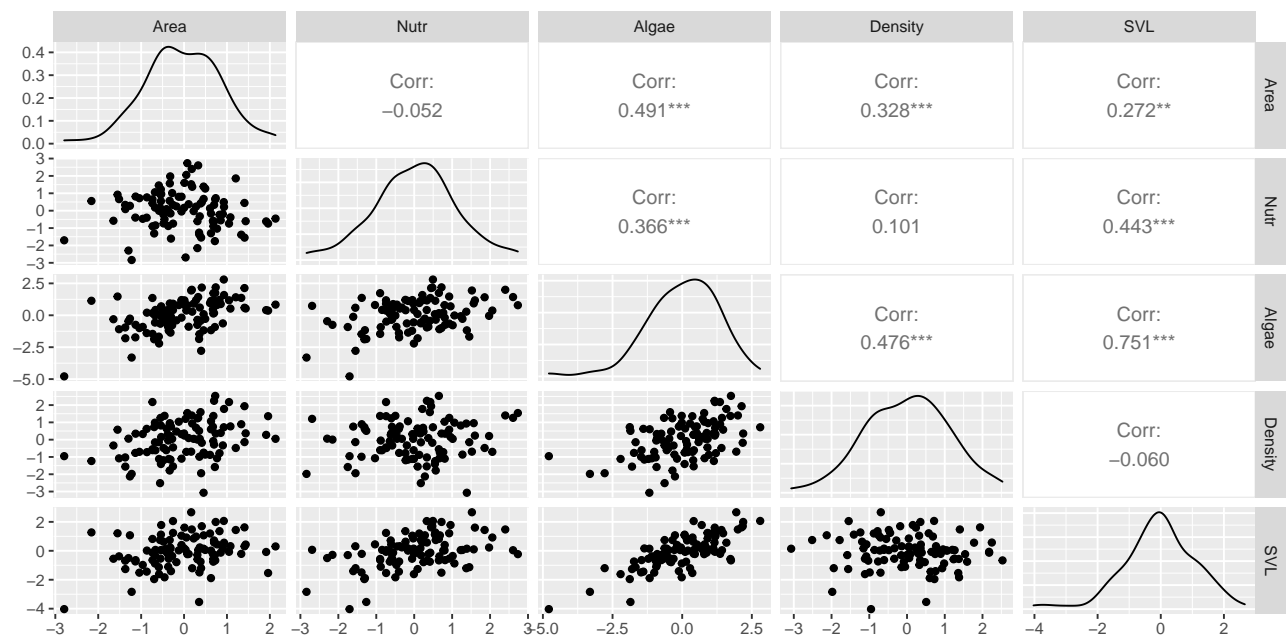
First challenge: *What variable(s) should you include in a regression* to understand the influence of pond size on frogs' size at metamorphosis (SVL)? How would you set up your analyses if this were your question and study?[2]

There are two basic approaches most people would suggest. First, many would suggest using individual predictors in separate regressions. Here are those individual regression result in table form:

Table 1: Parameter estimates from regression with single variables.

	Estimate	Std. Error	t value	Pr(>|t|)
Area	0.3551179	0.1256036	2.8272900	0.0056811
Algae	0.6830361	0.0608276	11.2290508	0.0000000
Nutr	0.4625887	0.0946365	4.8880562	0.0000039
Density	-0.0632511	0.1044847	-0.6053624	0.5463234

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:

Table 2: Parameter estimates from the full regression model with all variables included.

	Estimate	Std. Error	t value	Pr(>|t|)
Area	-0.0168088	0.0706780	-0.2378215	0.8125264
Algae	0.8680179	0.0570414	15.2173261	0.0000000
Nutr	0.1511290	0.0533208	2.8343336	0.0055986
Density	-0.5454214	0.0540551	-10.0900996	0.0000000

A close inspection of the parameter estimates in the single-variable models and the full model illustrates that we are likely to end up with different results depending on how we do our analyses! 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.

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?* This question, at least, I can answer.

It is difficult to know what each type regression 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 specify or draw 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.
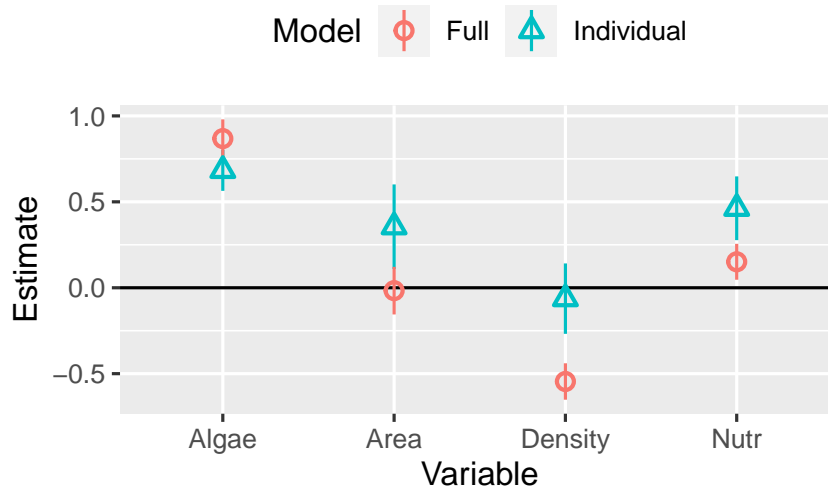
What is a DAG?

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[3]

- Graph: nodes (=variables) connected by arrows (=causal relationships)

[3] We need to return to this issue, later. It is quite important for infectious disease research.

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[4], circled: Ⓤ

- Draw arrows showing (assumed) *causal* relationships connecting variables (e.g., $X \to Y$ means "changes in X *cause* changes in Y")

  - Note that we are not drawing the *order* of things
  - The arrows do not describe the *direction* or *shape*[5] of the relationships, just the influence
  - Arrows do not show interactions, either

[4] This will vary depending on the reference you read, but we will use it here, for clarity.

[5] Again, relationships between variables need not be linear.

- 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.

Lastly, draw alternate versions representing your hypotheses of how the system works. We usually have several ideas of how our system might work, so let's be explicit about these alternate version! Moreover, drawing out DAGs can help refine your uncertainty or help you see the questions you need to ask before you go out an collect data. Plus, contrasting models can be very helpful, as we will see.

Back to our example

Spend 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, force yourself to draw out what seem like reasonable causal relationships. Go for it!

For teaching purposes, let me offer three reasonable versions of DAGs representing three ways this system might work.

I am going to use an R package, `daggity`, to help me work with these DAGs[6]. If you do not use R, you can instead use the online tools at http://dagitty.net/dags.html, a great site that also includes interactive lessons.

```
library(dagitty)

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

```
# to get the variables placed in prettier places,
# we can specify coordinates
coordinates(dag1) <- list(x=c(Area=1, Algae=5, Density=1,
                              Nutr = 1, SVL=10),
                          y=c(Area=1, Algae=5, Density=9,
                              Nutr = 5, SVL=5))
rethinking::drawdag(dag1) # or plot(dag1)
```

This first version implies that pond area ("Area") and the influx of nutrients like nitrogen ("Nutr") both influence algal growth ("Algae"), which in turn influences the SVL of metamorphs. 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.
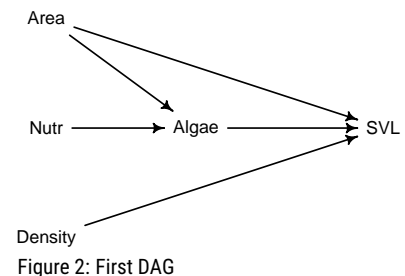

Figure 2: First DAG

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

```
coordinates(dag2) <- list(x=c(Area=1, Algae=5, Density=5,
                              Nutr = 1, SVL=10),
                          y=c(Area=5, Algae=5, Density=9,
```

```
                              Nutr = 1, SVL=5))
rethinking::drawdag(dag2)
```


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. (Recall that Algae refers to algal growth, not quality as a food item.)

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

```
coordinates(dag3) <- list(x=c(Area=1, Algae=5, Density=5,
                              Nutr = 5, Q = 7.5, SVL=10),
                          y=c(Area=5, Algae=5, Density=9,
                              Nutr = 1, Q = 3, SVL=5))
rethinking::drawdag(dag3)
```


Figure 4: Third DAG

In the third version we've included this effect of nutrients ("Nutr") on SVL due to its influence on food quality ("Q"), which is unobserved.

Implied conditional independencies

What have we gained by drawing out these DAGs? There are a few benefits, but let me focus on one: the testable implications of the DAGs. That is, what associations are implied? What variables should be independent of each other?

If you look back at the first DAG you can see that the Density and Nutrients are independent of each other; there is no way that Density affects Area, or vice versa. But just importantly, from a statistical point of view, it suggests that learning the values of Density in a pond tells us nothing about the Area of that pond; they should be statistically independent. The same is true of Density and Area, Area and Nutrients, and Density and Algae. That is, if you were to look
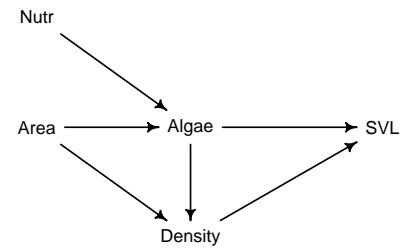
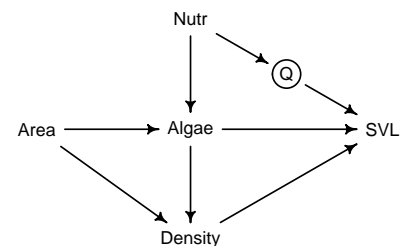for some statistical association, this DAG suggests you should not find any between any of these pairs of variables[7].

There is one more implied conditional independence: Nutrients should be independent of SVL *if* we condition on (or include in our regression model) 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 or transmitted 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, and thus metamorphs with larger SVLs. According to this model, that could only happen if the pond area were large, and so even if we condition on (know) the value of Algae, Nutrients still tell us something about (are correlated with) SVL simply because they imply something about pond area. Again, SVLs can only be large in nutrient-poor ponds *if* the Area is large, so knowing what the nutrient conditions are like, even if we already know its influence on Algae growth, tells us a bit about the Area of the pond and thus SVL. It's a bit headache-inducing, I know, 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 mechanical and so computers can do it just fine. In the R package `dagitty`, there is a function with the catchy name, `impliedConditionalIndependencies`. It can tell you those implied conditional independencies.

```
impliedConditionalIndependencies(dag1)
```

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

Two comments 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 similarly find the implied conditional independencies for the other two models.

```
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 entirely the same between the three DAGs. 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. Such independence would be consistence with the first DAG, but not the second or third, so determining whether Area ⊥⊥ Density would be quite interesting!

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
## -3.2809 -0.7451  0.0489  0.7853  2.4529
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02498    0.10494   0.238 0.812312
## Area         0.40930    0.11913   3.436 0.000868 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.049 on 98 degrees of freedom
## Multiple R-squared:  0.1075, Adjusted R-squared:  0.09839
## F-statistic:  11.8 on 1 and 98 DF,  p-value: 0.0008684
```

In this case, it looks like Density does increase discernibly with Area, which is evidence that the version of the system represented by the first DAG is probably not correct. You, the researcher, have to decide whether this is sufficient evidence to reject this first DAG. But is this finding not useful?

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[8]. 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 relieve us scientists of the duty to sort out causal relationships[9].

In a very real sense, our understanding of causation happens in our thinking, our conversations with colleagues, and 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[10], but they sure do help!

The four elemental relationships

We can gain further insights into and from 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.

$$X \rightarrow Z \rightarrow Y$$

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
```

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

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

[10] 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[11].

[11] Does your brain hurt yet?

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, however, only true if we do *not* condition on Z. If we do know (or condition on) Z, then information flows between X and Y.

This takes a bit of thought, or perhaps an example[12]. 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 working electric battery attached. If all you know is that the light switch is on ($X = 1$), you know nothing about whether there is a working battery ($Y =?$). If, however, you also knew that the light bulb was shining ($Z = 1$), then you could easily infer that there must be a working battery connected to the circuit ($Y = 1$) and if it were not, you would know the battery was not working ($Y = 0$). Knowing the value of (or conditioning on) the collider, Z, lets information flow between X and Y[13].

[12] I'm stealing this and much else from Richard McElreath's excellent book, Statistical Rethinking.

[13] 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.

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.

$$Y \searrow \atop X \nearrow \quad Z \to D$$

Again, X and Y are independent of each other unless you 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 no longer independent.

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

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

Closing the right doors[14]

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? That is, we are not interested in the separate influence of Area acting through Algae or other paths, but instead on the effect of Area on SVL through *all* paths. In that case we want to look at the relationship between Area and SVL, *without* conditioning on anything else. How do I know this?

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`.

[14] 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(dag1)
```

```
##  {}
```

```
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! (Go back to Figure 1 to see the "correct" estimate of this effect.)

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

```
# Changing the exposure from Area to Algae
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. And it is certainly better than throwing all of the variables you have measured into a statistical model and letting the statistics "think" for you!

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 recreate this cautionary tale[15].

Here's the DAG and a plot parameter estimates of the effect of X on Y assuming we condition on different variables. Which covariate(s) would you include? Which version is "right" if you simply wanted to know that bottom path?
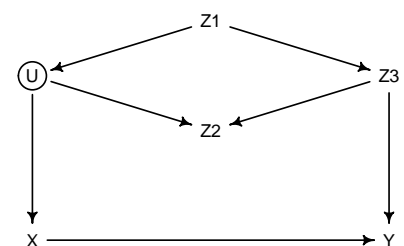

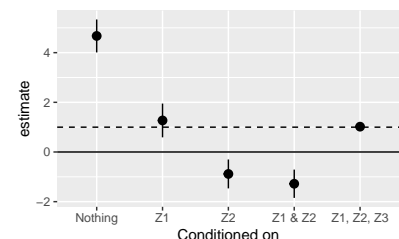Figure 5: The DAG assumed in one version of Simpson's paradox

Figure 6: Estimated effect of X on Y when including various parameters in the regression model. Vertical lines represent the 95 CI.

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

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[16]!

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 be the ones that give you 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[17]. 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 may simply be unable to sort 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 determine causal relationships. I hope that giving you a brief introduction to DAGs might help in this important goal.

[16] Known in this case, because we simulated it.

[17] Is that enough _real_s?

Another example: Pond conditions and die-offs

Imagine we conducted a study across a collection of ponds in an a region and wanted to see if certain characteristics predicted the probability of a die-off. Here are the variables, and some explanation of why each might be important.

- **Nutrients**: Higher nutrient concentrations (e.g., nitrogen) lead to higher food quality which means tadpoles are in better condition and thus less susceptible infections.
- **Salinity**: Salinity is an important cause of physiological stress in freshwater environments, so elevated salinity may reduce condition and susceptibility.
- Distance to nearest **Road**: Road salt can leach into ponds. Ponds near roads are also more likely to see animals introduced into them (or infected researchers![18])
- Distance to nearest **Pond** with a prior die-off: The pathogen comes from somewhere. Perhaps it is moved by animals moving between ponds.
- **Die-offs**: this is your response variable of interest. Think about it as a binary outcome (implying a logistic regression approach), or as the frequency of die-offs over some period of time.

What would you DAG or DAGs look like? Could you discern the effect of salinity on the probability of a die-off?

A challenge: Instead of using Die-offs as the response variable, what if we measured prevalence of infection at some point in the pond? How would you DAG change? Would you still be able to use these methods?[19]

[18] Clean your gear, please!

[19] Let me suggest you return to the definition of the DAG and think about whether it applies if we use prevalence.

Another example: Virulence of ranavirus isolates in cell culture

My graduate student is interested in the cause(s) of virulence in ranavirus infections. For simplicity, she is starting with cell culture experiments and so **Virulence** is defined as the amount of damage done to host cells by some time point. One model is that virulence is an incidental byproduct of virus **replication** rate; presumably virus isolates that replicate more quickly use host cells (or the resources inside) more quickly and so cause more damage (i.e., virulence). In this case, host immune responses such as **interferon** responses slow viral replication and thus reduce virulence. Another model is that the interferon response, itself, causes damage itself; after all, it leads to host cells undergoing preemptive apoptosis. She would like to know the relative importance of both of these pathways. What should her DAG(s) look like?