

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

Jesse Brunner

2022-08-31

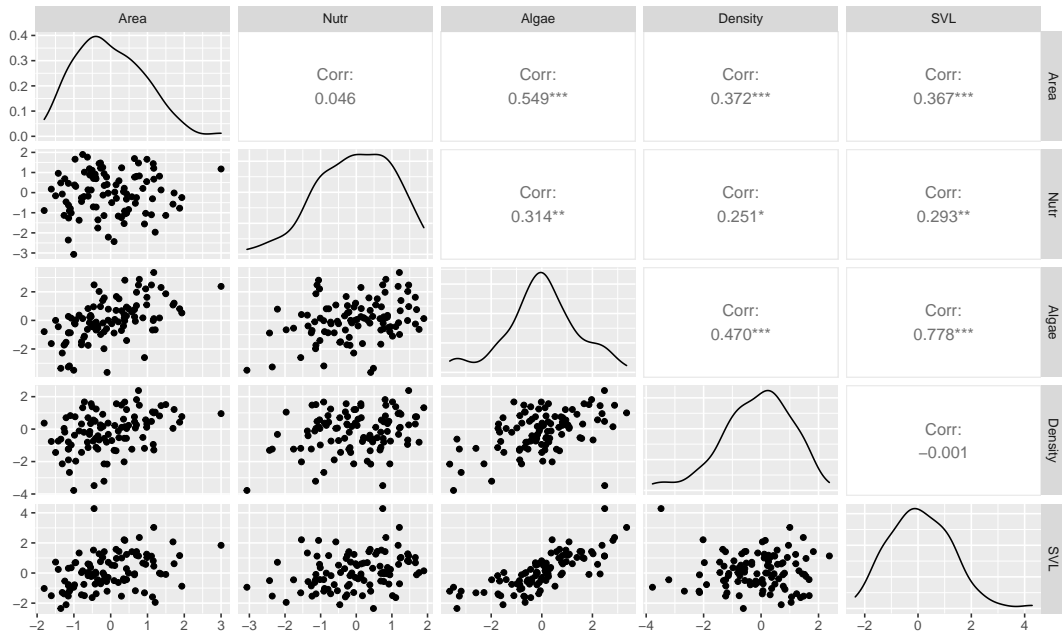
A Froggy Example

You hypothesize size of amphibians at metamorphosis increases with size of vernal ponds.

You have measured:

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

The data, in all its glory



What should you include in your regression?

Estimates effects depend on what is included... why?

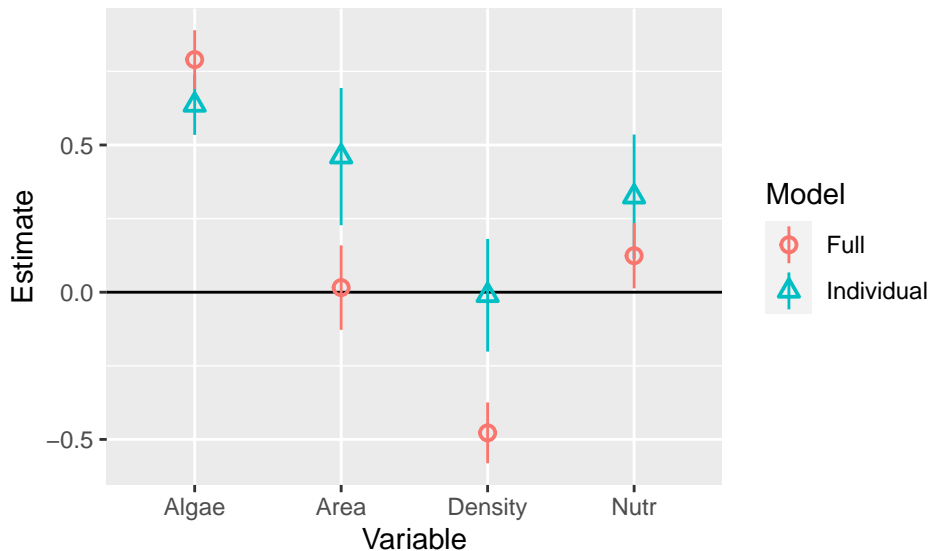


Figure 1: Estimated coefficients when estimated individually or in a full model. Vertical lines are 95 percent CIs.

Statistics are association machines

It is up to us to interpret what they are telling us. We have not (yet) done the hard work of figuring out how our statistics map on to how we think the system works.

Enter the DAG

What is a DAG?

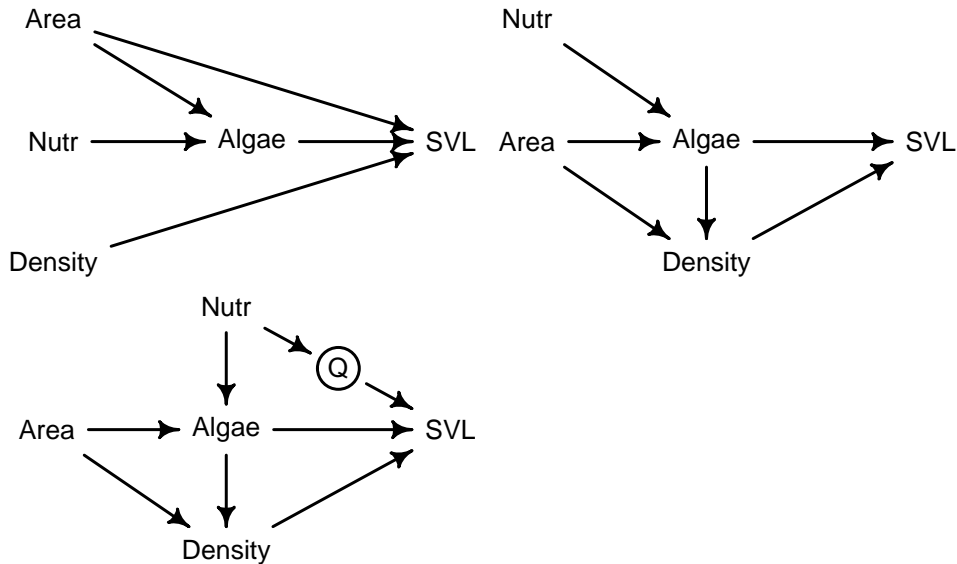
A “DAG” is a **d**irected, **a**cyclic **g**raph.

- ▶ directed: arrows describe causal influence
- ▶ acyclic: no cycles or loops, no positive or negative feedbacks
- ▶ graph: nodes (=variables) connected by arrows (=causal relationships)

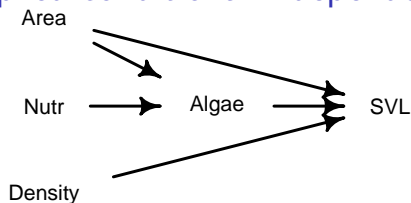
Drawing a DAG

- ▶ Write out the important variables (both “predictors” and “responses”)
 - ▶ measured variables are unadorned: e.g., X, Y, Z
 - ▶ unmeasured (or are unobserved) variables are circled: \bigcirc
- ▶ Draw arrows defining (assumed) *causal* relationships connecting variables (e.g., $X \rightarrow Y$ means “changes in X causes changes in Y ”)
 - ▶ We are not drawing the *order* of things
 - ▶ We are not describing the *direction* or *shape* of relationships
 - ▶ Arrows do not show interactions, either
- ▶ Keep it simple.
- ▶ Can draw different versions representing different hypotheses

Three possible DAGs for our frog example



Implied conditional independencies



Use `library(dagitty)` in R or <http://dagitty.net/dags.html>

```
impliedConditionalIndependencies(dagitty("dag{  
  Algae <- Area -> SVL  
  Nutr -> Algae -> SVL <- Density  
}"))
```

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

What do we mean by independent?

Quick and dirty definition: parameter estimate is essentially zero

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

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-0.1299641	0.1125465	-1.154759	0.2509972305
## Area	0.4827452	0.1217846	3.963927	0.0001401506

Since Density ∥ Area, DAG1 seems wrong...

Remember...

DAGs just tell us the (implied) consequences of the causal model we *assume*.

We, as scientists, have to sort out what are reasonable models, interpret model outputs, etc.

The four elemental relationships

1. Pipe: $X \rightarrow Z \rightarrow Y$

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

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

2. Confound: $X \leftarrow Z \rightarrow Y$

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

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


Notice they have the same conditional independencies! *Causation* flows one way, *Information* flows both ways.

The four elemental relationships

3. **Collider:** $X \rightarrow Z \leftarrow Y$ (Opposite of confound.)

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

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

4. **Descendant:**  $Z \rightarrow D$

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

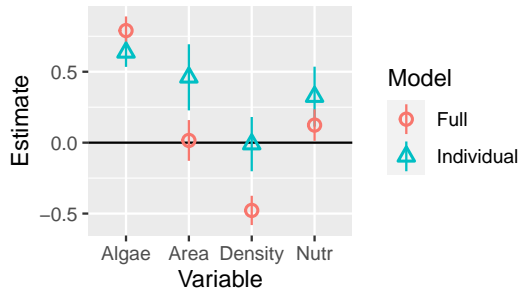
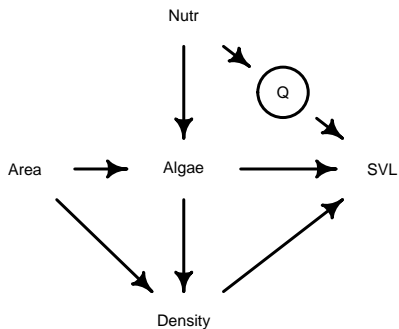
```
## D _||_ X | Z
```

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

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

Back to our example: What happened?

Why was $\text{Area} \perp\!\!\!\perp \text{SVL} \mid \text{Algae}, \text{Density}, \text{Nutr}$?



- Had conditioned on intermediaries in pipes!

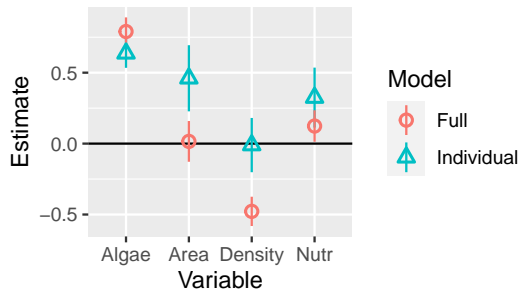
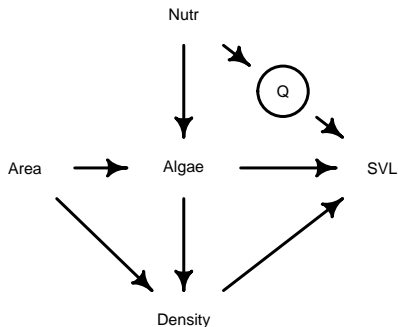
Back to our example: What do we want?

We were interested in effect of Area on SVL

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

```
## {}
```

- All we need to do was regress SVL on Area and nothing else!

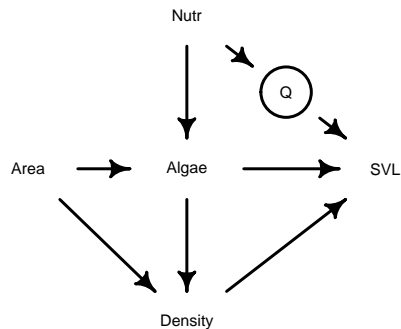


Back to our example: What do we want?

If instead we were interested in influence of Algae on SVL (in 3rd DAG)

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

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



Simpson's paradox

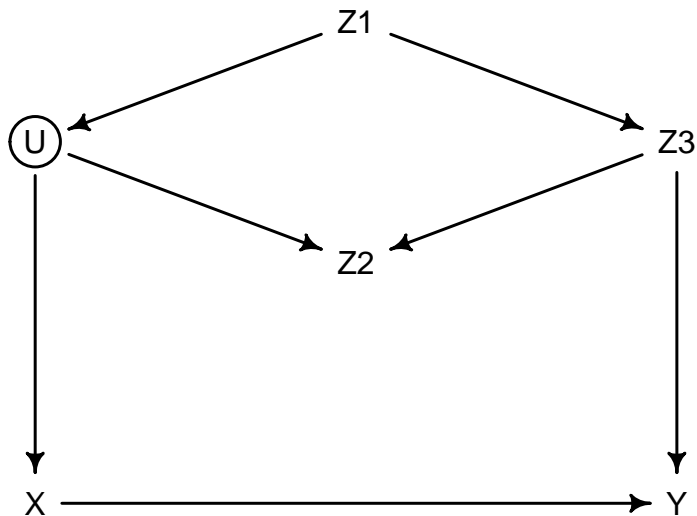
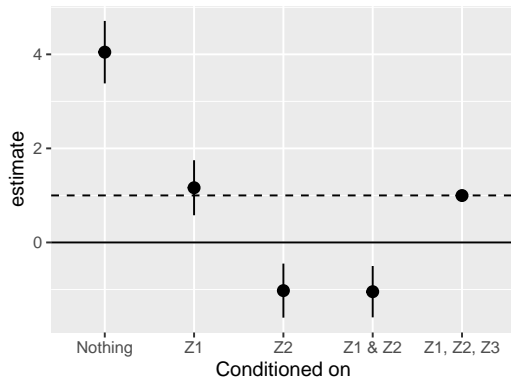
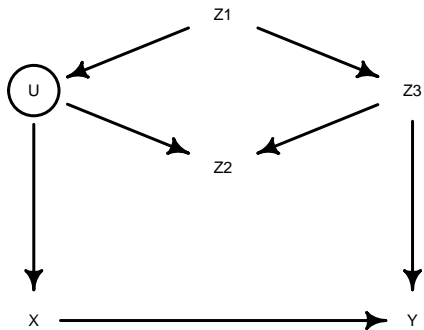


Figure 2: A DAG in one version of Simpson's paradox

Simpson's paradox



Magnitude and *sign* of estimated effect of X on Y depends on what else is in the model!

► Throw in variables at your peril!

Some final thoughts

- ▶ DAGs can help make sense of statistical associations between variables
 - ▶ help you focus on what is reasonable and what you *actually* want to know
 - ▶ *Sometimes* can help you test causal models (implied conditional independencies)
 - ▶ *Usually* can help you find the meaning of parameter estimates (assuming model is right)
- ▶ DAGs are useful in planning studies
 - ▶ determine what variables you need
 - ▶ useful for simulating data (and then analyzing)
- ▶ **But** DAGs are always assumed; you must decide what is reasonable