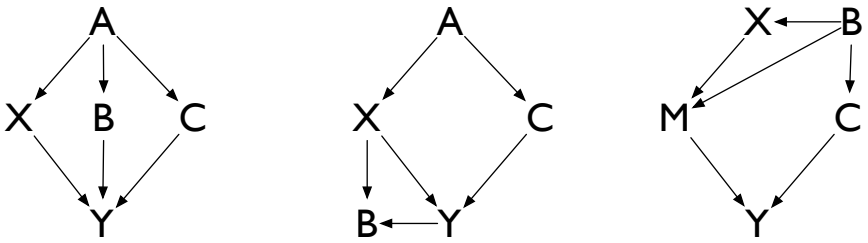
Exercises 13 · Causal inference

To turn in (Due Wednesday, May 6, 2015)

(1) Recovering causal patterns in simulated data

The three panels below depict causal diagrams for three different hypothetical systems.



For each diagram, take the following steps.

(A) Use your own imagination to invent a story that goes with the diagram, where the question of interest in your story is the causal effect of X on Y. For example, if the diagram were $Z \rightarrow X \rightarrow Y$, you might invent the story "Enjoyment of chemistry (Z) leads to greater effort in chemistry class (X) leads to a better chemistry grade (Y)."[1]

[1] My story is lame, so pick better ones!

(B) Now simulate a noisy sample (of size = 100) to go with your story. For the exogenous nodes (with no inward arrows), simulate these directly from some assumed distribution. Use these simulated values, together with a particular "ground truth" model depicting your assumed causal effects, to simulate the childen of the exogenous nodes. Carry on this way until you've simulate values for all nodes in the graph. The `runif` and `rnorm` commands in R will be helpful here.[2] For example, suppose that you decide on a model where

$$Y = 0.5X - 0.25B + 0.4C + \text{noise}\,.$$

[2] Remember that you can always read the help page for a command by typing, e.g., "?runif" at the console prompt.

(Remember to choose the coefficients that make sense in context!) To simulate the Y variable in this model, you would need to have already simulated the X, B, and C variables in your sample. Then you could simulate Y by calling

```
epsY = rnorm(100, mean=0, sd=5)
Y = 0.5*X - 0.25*B + 0.4*C + epsY
```

(C) Now take the simulated data from your causal diagram (each one in turn) and start fitting models. Compare the estimates and error bars for the causal effect of X on Y arising from the following three strategies:

1. The naïve strategy: regress Y on X alone.
2. The "kitchen-sink" strategy: regress Y on everything.
3. The clever strategy: using what you know about front-door and back-door adjustment, run a minimally sufficient analysis that allows you to properly estimate the causal effect of X on Y, without including unnecessary covariates.

Which methods "work" (in the sense of correctly estimating the causal effect of X on Y) from which simulated data sets? Pay close attention to where the kitchen-sink strategy fails, and explain why.

(2) Instrumental variables

Read the following paper: Econometrics in outcomes research: the use of instrumental variables. Newhouse JP, McClellan M. Annu Rev Public Health. 1998;19: 17-34.

To quote directly from the abstract:

> We describe an econometric technique, instrumental variables, that can be useful in estimating the effectiveness of clinical treatments in situations when a controlled trial has not or cannot be done. This technique relies upon the existence of one or more variables that induce substantial variation in the treatment variable but have no direct effect on the outcome variable of interest. We illustrate the use of the technique with an application to aggressive treatment of acute myocardial infarction in the elderly.

Write a short report (aim for 2 pages max, and potentially shorter if you're concise) that addresses the following issues:

1. the fundamental "identifiability" issue that arises in the study of catheterization in treating acute myocardial infarction (AMI);
2. the instrument used by the authors to estimate the causal effect of catheterization on successful treatment of AMI; and
3. the specific mathematical procedure used to fit an instrumental-variable model.

You may write out any equations by hand, if you wish. We will discuss intstrumental varialbles on Monday in class.