PLSC 502: "Statistical Methods for Political Research"

Multivariate Statistics: A Conceptual Overview
December 6, 2016

Introduction

We'll spend a bit of time discussing relationships among multiple variables today, with a particular focus on the idea of *causality*.

Types of Relationships

Bivariate Recursive and Nonrecursive

The simplest relationship is a *recursive bivariate* relationship:[1]

$$\begin{tikzcd} X \arrow[r] & Y \end{tikzcd} \tag{1}$$

In this model, X influences Y, but not vice-versa. A conceptually (slightly) more complex relationship is a *nonrecursive* one, in which X influences Y and Y also influences X simultaneously:

$$\begin{tikzcd} X \arrow[r, bend left] & Y \arrow[l, bend left] \end{tikzcd} \tag{2}$$

This latter type of relationship occurs regularly lots of places, including (say) economics, where (at least in classical theory) quantities like prices, demand, and supply all react to one another instantaneously.

In the absence of any other information, untangling a causal effect in a relationship like (2) is both logically and practically impossible.

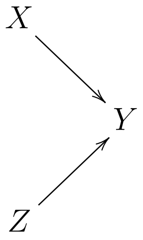
For both types of relationships, we can think of *association* between X and Y as a necessary but not sufficient condition for causality to be present. That is, if either (1) or (2) is the case, then we will observe an association between X and Y, but the observation of such an association could mean either (1), or (2), or neither (see below).

[1] Figures like the one in (1) are often called *path diagrams*, or (more formally) *directed acyclic graphs* ("DAGs"). The latter term is due to UCLA computer scientist Judea Pearl, whose work on causality and its representations is well worth reading; see his 2000 book *Causality* for the definitive treatment of his graphical model of causality. Pearl's work has been very influential in the social sciences, so much so that it has become the standard lingo for discussing causal relationships.

Multivariate, Recursive (“Collider”) and Spurious

Introducing a third variable, Z, complicates things somewhat. Multivariate relationships raise a host of different issues, both statistical and conceptual. We’ll just discuss the former, and only a bit of that.

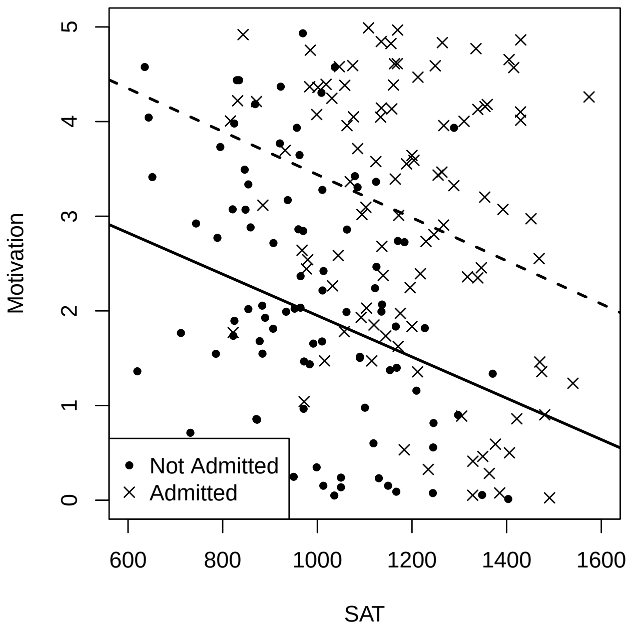
The simplest multivariate relationship is one in which more than one covariate influences the outcome of interest Y, but does so independently of all the others. Figure (3) illustrates this.



(3)

Pearl refers to Y in this context as a “collider” variable, because it has the effect of “blocking” any induced relationship between X and Z. Importantly, because the effect of X on Y is independent of Z’s effect on Y, **one need not condition on Z to accurately estimate the relationship between X and Y.**

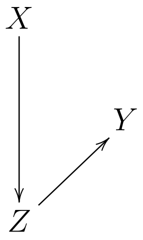
Figure 1: Conditional Dependence with a Collider Variable



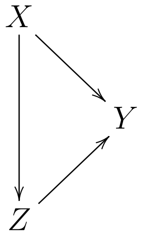
Here, think of Y as admittance to college (where a "×" indicated the individual got in, and a "•" indicates they did not), X as SAT scores, and Z as "motivation."[2] X and Z are unconditionally unrelated; however, conditioning on Y induces a (negative) relationship between X and Z within subcategories of Y (the dotted lines).

Mediated

A *mediated* relationship is one in which a third variable Z is caused/influenced by X and, in turn, influences Y. We can think of two types of mediated relationships:



(4)



(5)

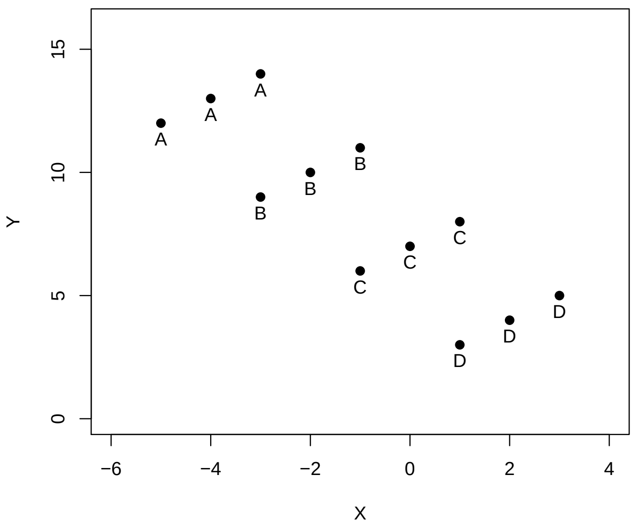
Figure (4) is a special case of figure (5), where the (conditional on Z) effect of X on Y is zero. The former is known as a *completely mediated* relationship, the latter is *partially mediated.*

Mediating variables serve to complicate estimation of the main relationship of interest (between X and Y). For example, a commonly-referenced phenomenon is *suppression*, where controlling for the presence of a mediating variable Z actually serves to *increase* the strength of the relationship between X and Y. In extreme cases (known colloquially as "Simpson's paradoxes"), the relationship between X and Y can actually change signs once conditioning on Z occurs. This is illustrated in Figure 2; note that the overall relationship between X and Y is negative, but that within each group (A vs. B vs. C, etc.) – that is, conditioning on a variable Z defined as group identity – the relationship is positive.[3]

[2] The idea for this figure was borrowed from Morgan and Winship (2007, Fig. 3.4).

[3] Pearl shows convincingly that the "paradox" isn't really a paradox at all, but a failure of causal logic.

Figure 2: Simpson's Paradox



Confounded

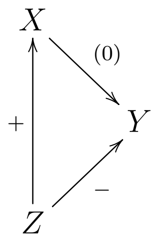
If the relationship between X and Y is *confounded*, that means that there exists a third variable Z that is correlated with both X and Y. In its simplest form, confounding looks like this:

$$\begin{tikzcd}
X \arrow[dr] & \\
& Y \\
Z \arrow[uu] \arrow[ur] &
\end{tikzcd} \tag{6}$$

Confounding has the effect of altering the perceived (bivariate) relationship between X and Y. For example, if the correlation between Z and X, and between Z and Y, is positive, then the relationship between X and Y may appear stronger than it "really is":

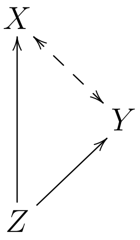
$$\begin{tikzcd}
X \arrow[dr, Rightarrow, "(+)"] & \\
& Y \\
Z \arrow[uu, "+"] \arrow[ur, "+"'] &
\end{tikzcd}$$

Conversely, if Z correlates (say) positively with X, but negatively with Y, it may have the effect of "masking" an actual (positive) relationship between X and Y:



Spurious

A "spurious" relationship is a particular form of confounding, in which the absence of a relationship between X and Y is "masked" by the appearance of one induced by their joint dependence on Z:

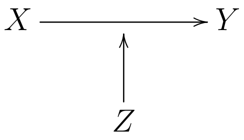


(7)

(7) illustrates the case where the association between X and Y is entirely due to their dependence on Z. In such a model, conditioning on Z is critical to uncovering the true (null) relationship between X and Y; failure to control for a confounder of this nature is usually presented as the #1 cause of Type I errors.

The key to dealing with confounding is to condition on any and all confounders that influence *both* X and Y.

Interactive



(8)

This is the case where the presence, magnitude, and potentially even the direction of the effect of X on Y varies according to the value of Z. Take, for example, a simple unidimensional spatial model, where the probability of a voter V voting for an alternative A over the status quo (SQ, where the location of SQ is normalized to zero) is a function of quadratic utility:

$$\Pr(\text{Yea}) = f[-(V - A)^2]. \tag{9}$$

One might want to know: What is the effect on Pr(Yea) of making the alternative more conservative (that is, pushing it farther to the "right" – making A larger)? Intuitively, of course, we know that the answer depends on "where the voter is" relative to the status quo and the alternative. We can express this by rewriting equation (9) as:

$$\Pr(\text{Yea}) = f[-(V^2 + A^2 - 2AV)]. \tag{10}$$

The effect of A on Pr(Yea) can then be thought of as something like:

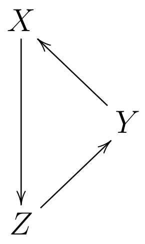
$$\frac{\partial A}{\partial \Pr(\text{Yea})} = f'[-(2A - 2V)].$$

In other words, the effect of the location of A depends both on A itself and on the "location" of the voter. Moreover, the sign of the effect is as we'd expect it to be: As V gets larger (more positive), the effect of moving A in a positive direction also gets more positive. The opposite is true, however, as V gets smaller (that is, more negative).

We will discuss interactive relationships at great length next term, in PLSC 503...

Cyclic

Finally, *cyclic* relationships are multivariate forms of a nonrecursive relationship:

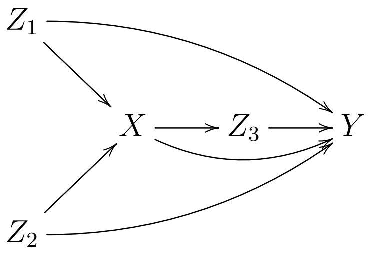


(11)

As with a bivariate nonrecursive relationship, in the absence of additional information, we can't untangle any causal effect for a cyclical relationship.

Combinations

These simple trivariate examples are basic exemplars of the sorts of more complex relationships that one finds in actual social scientific work. Common cases include multiple confounding and intervening variables, e.g.:



(12)

multiple interactions, and so forth.

A Few Words About Causality

Agresti and Finlay offer three helpful criteria in assessing causality:

1. **Association**. If X causes Y, then (once confounders have been conditioned on), there should be an association between X and Y.

2. **Temporal Order**. Causes generally occur prior to effects; therefore, assessing the temporal ordering of phenomena can be useful in establishing causality. Note a few points about this:

    - Some things (race, say, or colonial history) are pretty-well known to be fixed and antecedent to Y; those are always good places to start looking for causal effects.
    - Experiments and quasi-experiments also let us assess causality by controlling time order of measurement...

   - Note, however, that if actors are forward-looking / "sophisticated," temporal ordering can be reversed (e.g., it may appear that high-quality challengers negatively impact members' of Congress reelection chances, but if such challengers "cherry pick" races where incumbents are weak, the effect may well run the other way...).

3. **Elimination of Alternative Explanations**.
   - Controlling for / conditioning on as many confounds as we can think of.
   - This is what we do when we can't really do anything about #2...
   - It's the weakest strategy, in the sense that one can never *completely* rule out *all* competing explanations for most phenomena.

While we may often want to do #2 as well, we often end up doing only #1 and #3...

Some Tools of Analysis

X, Z, And Y Observed

- *Multivariate Regression* analysis...
- *Instrumental Variables* approaches...
- Models for *Causal Inference* (matching, differences-in-differences, etc.).

X and Y Observed, Z Unobserved

- Models for *Unobserved Heterogeneity*
  - Errors-In-Variables Models
  - Fixed/Random Effects ("Frailty") Models

X and Z Observed, Y Unobserved

- Factor Analysis / Principal Components
- Item Response Theory / Measurement Models
- Latent Class Analysis