Introduction to Causal Inference

Brenton Kenkel — PSCI 8357
March 24, 2016

So far in this course, I have been careful not to use causal language. Today that ends—not the part about being careful (hopefully) but the self-imposed ban on causal talk. We will:

- Remind ourselves about what it means for random variables to be independent and conditionally independent, which are crucial concepts in causal analysis.
- Get acquainted with the canonical statistical model of causality.
- Learn some basic, if imperfect, methods for estimating causal effects.
- Discuss variable selection for modeling observational data, with frequent reminders that fancy statistics cannot save you from poorly designed research.

Independence and Conditional Independence

Let A, B, and C be random variables. To keep things simple, assume each is discrete. A and B are *independent*, written $A \perp\!\!\!\perp B$, if

$$\Pr(A = a, B = b) = \Pr(A = a)\Pr(B = b)$$

for all a and b. If A and B are independent, then

$$\Pr(A = a \,|\, B = b) = \frac{\Pr(A = a, B = b)}{\Pr(B = b)} = \Pr(A = a)$$

for all a and b, which in turn implies

$$E[A \,|\, B = b] = E[A]$$

for all b.

Conditional independence is a weaker condition than independence. Loosely speaking, if two variables are not independent, but we can account for the source of the dependence between them, they are conditionally independent. Formally, A and B are *conditionally independent* given C, written $A \perp\!\!\!\perp B \mid C$, if

$$\Pr(A = a, B = b \mid C = c) = \Pr(A = a \mid C = c) \Pr(B = b \mid C = c)$$

for all a, b, and c. Conditional independence of A and B given C implies

$$\Pr(A = a \mid B = b, C = c) = \Pr(A = a \mid C = c)$$

for all a, b, and c, and

$$E[A \mid B = b, C = c] = E[A \mid C = c]$$

for all b and c.

Here is a basic example of random variables that are conditionally independent but not independent. Let C be a non-degenerate, real-valued random variable, and let ϵ and η be real-valued random variables that are independent of each other and of C. Define A and B by

$$\begin{aligned} A &= C + \epsilon, \\ B &= C + \eta. \end{aligned}$$

Since A and B both depend on C, they are not independent: the value of B is more likely to be high if that of A is high, and so on. However, they are conditionally independent given C: for any a, b, and c,

$$\begin{aligned} \Pr(A = a, B = b \mid C = c) &= \Pr(\epsilon = a - c, \eta = b - c) \\ &= \Pr(\epsilon = a - c) \Pr(\eta = b - c) \\ &= \Pr(A = a \mid C = c) \Pr(B = b \mid C = c). \end{aligned}$$

This simple example illustrates a more general principle. If C contains all of the factors that are determinants of both A and B, then A and B are conditionally independent given C.

The Potential Outcomes Model

The canonical statistical model of causality is the *potential outcomes model*, which is also sometimes called the Neyman-Rubin model (Holland 1986).[1] The

[1] Yes, it is standard to cite Holland for an attribution to Neyman and Rubin. Life is strange and beautiful.

potential outcomes model assumes that we have a set of units $i = 1, \ldots, N$, each of which receives a treatment T_i and produces a response Y_i. The response is a function of the treatment received, so we write it as $Y_i(T_i)$. Hence the terminology of potential outcomes: each unit has numerous possible responses, one for each treatment it might receive.

Table 1. Potential outcomes.

	$T_i = 0$	$T_i = 1$	$T_i = 2$
Unit 1	$Y_1(0)$	$Y_1(1)$	$Y_1(2)$
Unit 2	$Y_2(0)$	$Y_2(1)$	$Y_2(2)$
$\vdots$			
Unit N	$Y_N(0)$	$Y_N(1)$	$Y_N(2)$

In practice, we cannot observe all of these potential outcomes. Each unit receives a particular treatment and responds accordingly. The other potential outcomes are counterfactual—what would have happened if the treatment had been different.

Table 2. Observed outcomes.

	$T_i = 0$	$T_i = 1$	$T_i = 2$
Unit 1	$Y_1(0)$	$Y_1(1)$	$Y_1(2)$
Unit 2	$Y_2(0)$	$Y_2(1)$	$Y_2(2)$
$\vdots$			
Unit N	$Y_N(0)$	$Y_N(1)$	$Y_N(2)$

From here on, I will assume a binary treatment, valued 0 or 1 for every unit. Everything here carries over to treatments with more values, but the notation gets messier. We may call the set of units with $T_i = 1$ the “treatment group$ and those with $T_i = 0$ the”control group“.

In causal analysis, we are interested in the effect of receiving the treatment versus receiving the control. In other words, what is the difference between a unit’s potential response if it receives the treatment and its potential response

if it receives the control? We define this as the causal effect for unit i, or

$$\tau_i = Y_i(1) - Y_i(0).$$

What Holland (1986) calls the *fundamental problem of causal inference* is that these unit-specific causal effects are unobservable. We only observe one outcome per unit and thus cannot directly measure the differences between potential outcomes.

Instead of unit-specific causal effects, what about the average causal effect across units? Define the *average treatment effect* as

$$\tau = E[\tau_i] = E[Y_i(1) - Y_i(0)],$$

where the expectation is taken with respect to the population distribution of units. Our goal will be to estimate τ.

If each unit is randomly assigned to the treatment or control group, then it is easy to estimate τ. What makes it easy is that, under random assignment, treatment status T_i is independent from the potential outcomes $Y_i(0)$ and $Y_i(1)$. Everyone is just as likely to receive the treatment, so we have:

$$\Pr(T_i = 1 \mid Y_i(0), Y_i(1)) = \Pr(T_i = 1).$$

So we can estimate τ without bias just by taking the difference of the average response within the treatment and control groups. Let N_1 and N_0 denote the number of treated and control units respectively. The expected value of the naïve difference of means estimator is

$$\begin{aligned} E\left[\frac{1}{N_1}\sum_{i:T_i=1} Y_i - \frac{1}{N_0}\sum_{i:T_i=0} Y_i\right] &= E[Y_i(1) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0] \\ &= E[Y_i(1)] - E[Y_i(0)] \\ &= E[Y_i(1) - Y_i(0)] \\ &= \tau. \end{aligned}$$

The same logic does not hold in situations without random assignment, such as when treatment is self-selected. Imagine using observational data to estimate the effect of an expensive GRE preparation program on GRE scores. Only high-income or highly motivated students will choose to shell out for the program,

but we would also expect income and motivation themselves to lead to higher test scores regardless of the effect of the program. So in this case, T_i is not independent of $Y_i(0)$ and $Y_i(1)$, and the naïve difference of means estimator is most likely biased. The bias comes from the *confounding variables*, income and motivation, which affect both treatment assignment and the potential outcomes.

This does not mean all is lost, though. Let X_i denote the full set of confounding variables. Anything that affects both treatment assignment and either or both of the potential outcomes goes into X_i. By accounting for all these factors, we remove the sources of dependence between treatment assignment and the potential outcomes. In other words, we have the conditional independence condition

$$T_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \,|\, X_i.$$

If this conditional independence condition holds, along with the overlap condition

$$0 < \Pr(T_i = 1 \,|\, X_i) < 1 \quad \text{for all feasible } X_i,$$

then we say that treatment assignment is *strongly ignorable* given X_i (Rosenbaum and Rubin 1983). Under strong ignorability, we can estimate treatment effects without bias by properly adjusting for the confounding variables X_i. But what do we mean by "properly adjusting"? And exactly which variables do we need to adjust for? These are the topics of the next two sections.

Estimating Treatment Effects

With nonrandom assignment, the problem with the unadjusted difference of means estimator is that we're comparing dissimilar units. The control group and the treatment group differ in terms of their background characteristics, and the difference of means picks up those characteristics' effects in addition to those of the treatment.

We want to estimate the average treatment effect while holding these background characteristics fixed. Our estimator should only compare units that are similar in terms of confounding factors, so as to isolate the effect of the treatment. There are a mind-boggling number of estimators that claim to do so. We will consider two of the simplest. The first is subclassification, which always compares apples to apples but is only applicable in limited circumstances. The

second is our longtime friend regression, which is broadly applicable but may sometimes compare apples to oranges.

Subclassification

We can take the notion of only comparing similar observations to an extreme—only compare those with identical covariate values. *Subclassification* involves calculating the difference of means within each group defined by a unique combination of covariates, then averaging over those groups.

The subclassification estimator is motivated by the observation that, under strong ignorability,

$$\tau = \sum_x \Pr(X_i = x)\left(E[Y_i(1) \mid T_i = 1, X_i = x] - E[Y_i(0) \mid T_i = 0, X_i = x]\right).$$

In words, the average treatment effect is the weighted average of the differences of means within each covariate grouping, where the weights are the population proportions of each grouping.

To define the subclassification estimator, suppose there are finitely many combinations of covariates, $X_i = x_1, \ldots, x_K$. Let N_{k0} and N_{k1} denote the number of control and treated observations with $X_i = x_k$, and assume these are all nonzero. Then the subclassification estimator is

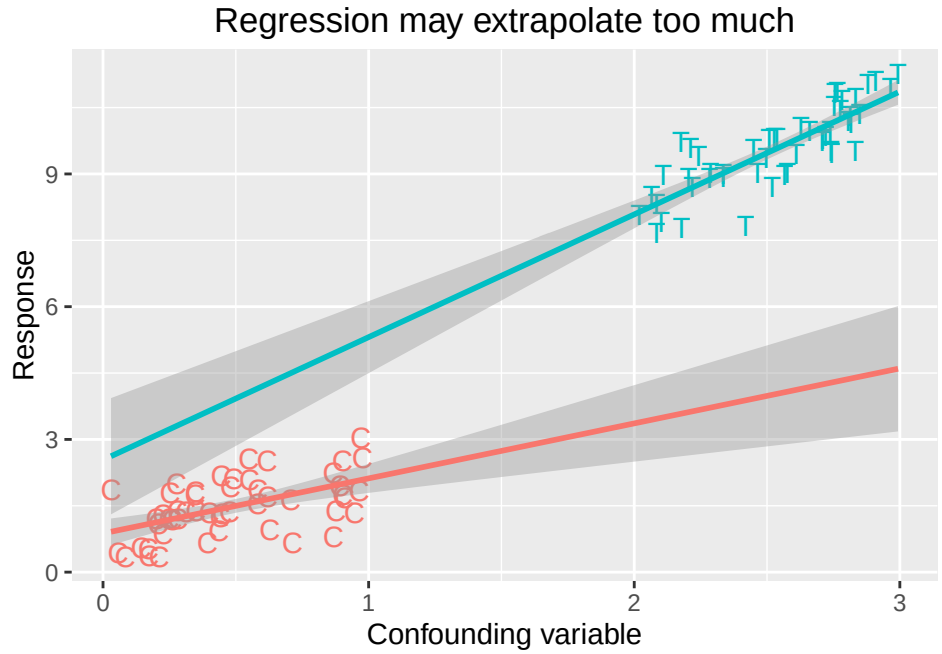
$$\hat{\tau} = \sum_{k=1}^{K} \frac{1}{N_{k0} + N_{k1}} \left(\frac{\sum_{i:X_i = x_k, T_i = 1} Y_i}{N_{k1}} - \frac{\sum_{i:X_i = x_k, T_i = 0} Y_i}{N_{k0}} \right).$$

The problem is, the N_k's might not all be zero. If you have p binary covariates, then there are 2^{p+1} unique combinations of covariates and treatment. For example, with $p = 9$, there are 1,024 groupings. In any reasonably sized sample, at least one of these is liable to be empty. It gets even worse with continuous covariates. These you must discretize just to make subclassification feasible. The discretization must be coarse enough not to leave empty cells, but coarser schemes correspond to more biased estimators (Cochran 1968).

Regression

With many covariates or continuous covariates, a more practical solution is regression: estimate τ as the coefficient on T_i in a regression of Y_i on (T_i, X_i).

Unlike subclassification, regression does not guarantee that each unit is compared only to like units. A particularly extreme contrived example is illustrated below.



If we tried this with subclassification, we would quickly see that there is no overlap in the confounding variable between the treatment and control groups, meaning drawing any inference we draw about treatment effects will be based on extrapolation. But if we used OLS thoughtlessly, we would not be alerted to any problem. The lesson here is not "Don't use OLS", but "Don't use OLS thoughtlessly". Look at your data, pick a specification carefully, check the quality of the fit, and so on.

On that note, another problem with OLS is that the relationship between X_i and the expected response might not be linear. Subclassification bypasses this, since it doesn't model the relationship between X_i and the expected response. Obviously, just as in non-causal regression modeling, you should be cautious about nonlinearities and use higher-order terms judiciously when using OLS

to estimate treatment effects.

Selecting Covariates

In general, the quality of your data matters more than your choice of estimator. If you have a large sample and strong ignorability is satisfied, then most decent estimators will yield similar results. But we are rarely so lucky. It is hard to think of observational data in political science where we have successfully identified and measured all possible confounding variables.

If we cannot eliminate bias due to confounding, how can we best deal with it? The best way is to design our studies well from the outset—to collect data on samples of mostly comparable units, well tailored to our hypotheses, such that potential confounders are minimal, identifiable, and measurable (Freedman 1991). But sometimes we cannot find a natural experiment and we must resign ourselves to the second-best, dealing with messy, heterogeneous data. In this unfortunate circumstance, how should we choose which covariates to adjust for? It is tempting to say "all of them", so as to maximize our chance of achieving strong ignorability. On the contrary, as long as some confounders remain unobserved, adding more variables to the model does not necessarily decrease bias (Clarke 2006). Not to mention the problems for efficiency and interpretability.

Affects Treatment and Response: Yes

A confounding variable, by definition, is one that directly affects both treatment assignment and the potential outcomes. These are the biggest threats to causal inference, and these are what we should take the most care to adjust for in our empirical analyses. Again, our work will be most convincing if we deal with confounding through the research design rather than through post hoc statistical adjustments. If that is not possible, though, it is important to identify and control for the variables with the strongest confounding effects.[2]

[2]Under certain conditions, controlling for weak confounders can be dicey if there are stronger unobserved confounders. My working paper with Kevin Clarke and Miguel Rueda, "Misspecification and the Propensity Score: The Possibility of Overadjustment" (title to be changed *very* soon), gives the details.

Only Affects Response: No

There is no need to control for a variable that affects the response but does not at all affect treatment assignment. Such variables are not a threat to strong ignorability and thus do not induce bias in our estimates. We can sometimes estimate causal effects more efficiently by including them, but the consequences for bias are nil. Plus, in most observational settings, any variable that affects the outcome but cannot possibly affect treatment assignment is probably post-treatment and therefore should be left out of the model in any case (see below).

My friends who actually do empirical work tell me that reviewers love suggesting that you include variables like these in your model. If there is no way the variable in question could affect the assignment of the relevant treatment, just note in a footnote (or your response to the reviewers) that the variable is not a confounder and thus need not be controlled for. Cite Rosenbaum and Rubin (1983) and the formal definition of strong ignorability if you have to.

Only Affects Treatment: No (But Hold on to It...)

There is no need to control for a variable that affects treatment assignment but does not at all directly affect the potential outcomes. We call these variables *instruments*. As we will see next week, instruments are very useful for estimating treatment effects when strong ignorability is violated, but we should not "control for" them the same way we would an ordinary confounder.

Affected by Treatment: NO!

Not only is there no need to control for variables that come after the treatment—you should *not* do so. Rosenbaum (1984) gives a formal exposition on why not to control for post-treatment variables. I will instead pursue a proof by example.

Imagine that we want to estimate the effect of smoking on lung cancer. Furthermore, imagine that smoking causes lung cancer through precisely one channel: buildup of tar in the lungs. You smoke, it fills your lungs with tar, you get cancer. According to our counterfactual model of causality, the average treatment effect of smoking on cancer is positive. For each smoker, had they not smoked,

they would have less tar in their lungs and a lower chance of contracting cancer (and vice versa for the non-smokers).

If we were to estimate the regression equation

$$\text{Lung Cancer}_i = \beta_0 + \beta_1 \text{Smoking}_i + X_i'\beta + \epsilon_i,$$

where X_i is the full set of confounding variables, we would yield a positive coefficient on smoking. We would correctly conclude that smoking causes lung cancer. But now imagine we included the post-treatment variable tar buildup in our regression equation, estimating

$$\text{Lung Cancer}_i = \beta_0 + \beta_1 \text{Smoking}_i + \beta_2 \text{Tar}_i + X_i'\beta + \epsilon_i.$$

Remember we assumed that the only path between smoking and cancer was through tar buildup. So in our model where we control for the post-treatment factor, we would yield a coefficient close to 0 on smoking and conclude, absurdly, that smoking does not cause cancer. Ergo, via proof by extreme example, we should not control for post-treatment variables.

Reviewers will tell you to control for post-treatment variables. Don't do it (and cite Rosenbaum as your reason not to). You will read papers that control for post-treatment variables. Don't believe their results. Their authors will say, yeah, but post-treatment bias shrinks the estimated effect and just makes it harder to pass the significance test; I passed the significance test anyway so my hypothesis must be right! Tell them, no, post-treatment bias is not necessarily conservative (http://cyrussamii.com/?p=730) and then ask them why they're more interested in getting stars than accurately estimating the causal effect of their treatment.

Summary

We've covered a lot of ground today.

- We began to speak in the counterfactual language of causality.
- We discussed a couple of simple estimators for causal effects when strong ignorability holds.
- We worked through some basic heuristics for variable selection with observational data.

- Do control for pre-treatment confounders.
- Safe to leave out other pre-treatment variables.
- Definitely leave out post-treatment variables.

Next time, we'll work on using instrumental variables to obtain valid estimates even when strong ignorability is violated.

References

Clarke, Kevin A. 2006. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22 (4): 341–52.

Cochran, W G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24 (2): 295.

Freedman, David A. 1991. "Statistical Models and Shoe Leather." *Sociological Methodology* 21 (January): 291–313.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Rosenbaum, Paul R. 1984. "The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society. Series A (General)* 147 (5): 656–66.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70 (1): 41–55.