PSC 504 - Observational Studies and Confounding

Matthew Blackwell

2/07/2013

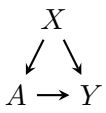
Confounding

Observational studies versus experiments

- What is an observational study? It is a study where the researcher **does not control the treatment assignment.** Because the analyst does not control the assignment he or she cannot guarantee that the treatment and control groups are comparable.
- In the previous weeks, randomization gave us a crucial result: that the treatment and control groups were comparable on any pre-treatment covariate so that any remaining differences were due to causal effects. Once we move to observational studies, this no longer holds by default. We are going to have to work harder to justify our analyses with observational studies.
- Rubin (2008) argues that we should try to "design" our observational studies in the same way we might analyze an experiment where we've lost the randomization procedure. That is, we should ignore the outcomes, try to estimate the randomization procedure (of the ideal experiment we think the data comes from).
- Remember in the DAGs, randomization implies no arrows pointing into the treatment or we know exactly which arrows because we have done a block-randomized experiment.

Backdoor paths and blocking paths

- What is a backdoor path? A **backdoor path** is a non-causal path from A to Y. This is a path that would remain if we were to remove any arrows pointing out of A (these are the potentially causal paths from A, sometimes called **frontdoor paths**). They are "backdoor" paths because they flow backwards out of A: all of these paths point into A.
- Backdoor paths between A and Y generally indicate common causes of A and Y (though not always, see M-bias below). The simplest possible backdoor path is the common confounding situation:



- Here there is a backdoor path $A \leftarrow X \rightarrow Y$, where X is a common cause for the treatment and the outcome. This might represent the relationship between money raised in a campaign by an incumbent (treatment) and the margin of victory for the incumbent (outcomes), where the common cause might be challenger quality.
- When there are unblocked backdoor paths, there are two sources of any association between A and Y: one causal (the effect of A on Y) and one non-causal (from the backdoor path). Thus, with unblocked backdoor paths, it's difficult to know if any association is a result of the causal effect or the backdoor path.
- A path is **blocked** if (a) we control for or stratify a non-collider on that path OR (b) we do not control for a collider. Thus, in the above sample, if we condition on X, then the backdoor path is blocked. Remember that blocked paths have no association following over them. Also remember that for any given path, we only have to have one of these conditions to hold. So, if we see a path with an uncontrolled collider, this path is blocked without conditioning on any other variables.

Backdoor criterion

- How to tell if an effect is identifiable from the graph? From Pearl (2000), we have the **backdoor criterion** which states that an effect of A on Y is identifiable if either:
  1. No backdoor paths from A to Y
  2. Measured covariates are sufficient to block all backdoor paths from A to Y.
- The first situation is only plausible in a randomized experiment, but the second might be plausible in observational studies as well.
- The backdoor criterion is fairly powerful. It can tell us (1) is there confounding given this DAG, (2) if it is possible to removing the confounding, and (3) what variables to condition on to eliminate the confounding.

Ignorability and backdoor paths

- How does the backdoor criterion relate to ignorability? On DAGs we don't have any explicit potential outcomes or counterfactuals. If the graph is causal (in the sense that each of arrows represents a causal effect in the potential outcomes sense), then there is a specific relationship between the backdoor criterion and ignorability.
- Suppose that we use the backdoor criterion and find that a set of variables X blocks all the backdoor paths. This implies the treatment assignment is conditionally ignorable: $Y(a) \perp\!\!\!\perp A|X$.
- Thus, in many cases we refer to the ignorability assumption as "no unmeasured confounders" which is really short hand for no unblockable backdoor paths.

Assumptions to identify effects

- In general, there are two approaches to identifying causal effects in observational studies. In the coming weeks, we will techniques that fall into both camps.

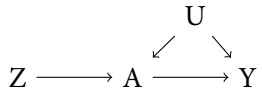
- One thing to note about observational studies: without randomization, it is assumptions that will identify the causal effects. These assumptions will be untestable in general and require subject-matter knowledge to justify. The Acemoglu paper that we read is a good example of picking apart the assumptions that underlie an analysis in terms of subject-matter knowledge. This deep understanding of a place, institution, or set of units allow us to justify and/or criticize causal assumptions.
- Sometimes causal inference is seen as "atheortical" but often theoretical concerns influence what types of assumptions we find plausible and which we do not. For instance, Acemoglu uses theory to argue that agents should have induced preferences over political institutions (since they have preferences over outcomes), which leads him to argue that it will be hard to block all backdoor paths (of course, he doesn't use this language).

Selection on the observables

- There are many names for this assumption and they vary by discipline. It is "selection on the observables" in economics, "no unmeasured confounders" in epidemiology, "exchangability" or "ingorability" in statistics, and "no omitted variables" in political science.
- Basically, it says that selection into treatment is based only on observable data, X. Or, more specifically, that the treatment assignment, A is independent of the errors in Y, conditional on X. This is a parametric version of the ignorability assumption, $Y(a) \perp\!\!\!\perp A|X$.

Exclusion restrictions

- In many instances, it is difficult to justify ignorability because there is unmeasured confounding between the treatment and the outcome. For instance, with institutions, elites might have preferences for lower levels of redistribution and presidential systems and elites will attempt to achieve both of these and often succeed. Thus, the political institutions are not causing fiscal outcomes, but rather elite preferences and their control over the government is causing both outcomes.
- In these situations, we can still identify causal effects using a different sort of assumption, called an exclusion restriction. These assume that there exists a variable (or set of variables) that affects the treatment and only affect the outcome through their affect on the treatment. Here is a DAG that describes the relationship:



- Here, Z affects A, but has no direct effect or common cause with Y. The latter part of this assumption (no direct effect or common cause) is the exclusion restriction. It's fairly difficult to find valid instruments and some find them more plausible than others in general.
- We call Z an instrument for A and we'll talk more about these instrumental variables approaches later in the term.

Estimating causal effects under no unmeasured confounders

Typical OLS

- Let's say we have the usual regression formula:

$$Y_i = \alpha A_i + X_i'\beta + u_i$$

- Does no unmeasured confounders help us identify the causal parameter α? Let's figure that out. First, note that an equivalent way of running the same regression is to replace each variable with its residual from a regression of itself on X_i:

$$\tilde{Y}_i = \alpha \tilde{A}_i + \tilde{u}_i$$

- Using the usual OLS theory, we can show that the probability limit of the OLS estimator of α is:

$$\begin{aligned} \text{plim}\hat{\alpha}^{\text{OLS}} &= \frac{\text{Cov}(\tilde{A}_i, \tilde{Y}_i)}{\text{Var}(\tilde{A}_i)} && (1) \\ &= \frac{\alpha\text{Cov}(\tilde{A}_i, \tilde{A}_i) + \text{Cov}(\tilde{A}_i, \tilde{u}_i)}{\text{Var}(\tilde{A}_i)} && (2) \\ &= \alpha + \frac{\text{Cov}(\tilde{A}_i, \tilde{u}_i)}{\text{Var}(\tilde{A}_i)} && (3) \end{aligned}$$

- Thus, the key assumption comes from $\text{Cov}(\tilde{A}_i, \tilde{u}_i) = 0$. Note that $\tilde{A}_i$ and $\tilde{u}_i$ are these variables, purged of their relationship with X_i. Thus, under ignorability conditional on X_i, there should be no covariance between these two variables should be o because there are no other common causes after accounting for X_i.

- It is instructive to see what happens when this is violated. For example, if $\tilde{u}_i = \lambda\tilde{L}_i + \omega_i$, with ω_i independent of the treatment but no $\tilde{L}_i$, then the OLS estimator would be inconsistent:

$$\text{plim}\hat{\alpha}^{\text{OLS}} = \alpha + \lambda\frac{\text{Cov}(\tilde{A}_i, \tilde{L}_i)}{\text{Var}(\tilde{A}_i)}$$

- Note that under ignorability, just because we can identify α does not mean that α is, in general, equal to the average treatment effect, τ. In fact, they will be different in most cases. The α here has a causal interpretation, just not one as the average treatment effect. We'll talk more about this in the coming weeks.

Subclassification/stratification

- If we have ignorability conditional on some set of covariates X, then how should we proceed? Remember that conditional ignorability is similar to a block-randomized experiment, where we would estimate effects within the blocks because we had mini-experiments within the blocks. In observational studies, we can do the same exact thing: stratify the data based on X and calculate the condtional average treatment effect $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$. Ignorability ensures that these conditional average treatment effects are identified.

- As Rubin (2008) points out, with ignorability, we know that within levels of X, the treatment and control groups should be **balanced** with respect to measured and unmeasured confounders. Again, this depends crucially on the ignorability assumption. What does balance mean? It means that the distribution of a variable (or set of variables) is the same in the treated and control groups:

$$f(Y_i(1)|A_i = 1, X_i = x) = f(Y_i(0)|A_i = 0, X_i = x)$$

- The classic example here is the effect of cigarette smokers versus cigar/pipe smokers. In the raw data, death rates are higher for cigar/pipe smokers compared to cigarette smokers. Of course, there is one very important confounder: age. Pipe/cigar smokers are likely to be much older than cigarette smokers. And when Cochran stratified the data into age-based strata and compared smokers of similar ages, he found that cigarette smokers had higher death rates.

- In this example, Cochran divided age into k different strata, $S_i \in s_1, s_2, \ldots, s_k$, where s_1 might be 18-25, s_2 might be 26-35, and so on. The key assumption here is that there is balance on X_i within these strata. That is, the distribution of X_i is the same across levels of the treatment within the strata:

$$f(X_i|A_i = 1, S_i = s) = f(X_i|A_i = 0, S_i = s)$$

- When this holds along with ignorability, we know that ignorability holds on just S_i: $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i|S_i$. This is useful because it means that we don't have to worry about the continuous nature of age in this case.

- What about when X has has many dimensions? Even if we stratify as above, there will be very few, if any, units in a given stratum of X_i. So, how do we calculate effects? One approach involves stratification on what we call the **propensity score**, which is the unit's individual probability of receiving treatment, condition on the covariates:

$$e_i = \Pr[A_i = 1|X_i]$$

- Rosenbaum and Rubin (1983) showed that if we correctly estimate the e_i, balancing the treatment and control groups on this estimated propensity score is the same as balancing with respect to the the entire set of covariates. That is, if we create some strata based on e_i and we have balance within strata: $f(e_i|A_i = 1, S_i = s) = f(e_i|A_i = 0, S_i = s)$, then this guarantees that X_i is balanced as well.

- Of course, in observational studies, we don't know the propensity score. Thus, we can use a logistic regression to estimate the propensity score, then group the units in strata based on the estimated propensity score and then estimate the average effects within these strata.
- We would run a parametric model with parameters γ to estimate the propensity scores. First, we estimate $\hat{\gamma}$, then use those estimates to get the predicted probabilities, which are simply the propensity scores:

$$\hat{e}_i = \Pr[A_i = 1 | X_i; \hat{\gamma}]$$

- For instance, in R, we could easily calculate the propensity scores using the `glm` function:

```
pscores <- glm(treat ~ var1 + var2 + var3, data = mydata, family = binomial())$fitted.values
```

- What variables do we include in the propensity score model? Any set of variables that blocks all the backdoor paths from A_i to Y_i. Why? Because conditioning or balancing on these variables ensures balance on the potential outcomes.
- One common diagnostic for this subclassification approach is to check the balance (usually the standardized difference in means) of each of the covariates within the strata defined by the propensity score.

Standardization/direct adjustment

- Above we calculated the CATE, $\tau(x)$, but what if we want the average treatment effect, τ? Let $\mathcal{X}$ be the support of X_i. That is, $\mathcal{X} = \{x : f(x) > 0\}$, where f is the probability density/mass function for x. It turns out that we can estimate this by simply taking the average of the CATEs weighted by the distribution of X_i:

$$\tau = \sum_{x \in \mathcal{X}} E[Y_i(1) - Y_i(0) | X_i = x] \Pr[X_i = x]$$

- If X_i is continuous with c.d.f. $F(x)$, then we have the integral:

$$\tau = \int_{x \in \mathcal{X}} E[Y_i(1) - Y_i(0) | X_i = x] dF(x)$$

- When X_i is low dimensional and discrete, we can easily calculate $\Pr[X_i = x]$ with its empirical distribution: $\frac{1}{N} \sum_i^N \mathbb{I}(X_i = x)$.
- For subclassification on the propensity score, you simply weight by the size of each stratum.