PSC 504: Randomized Experiments

Matthew Blackwell

1/23/2013

Randomization and identification

What is identification?

- Indentification tells us what quantities are estimable if we had inifinite data and so we didn't have to worry about random variability. Thus, we are even abstracting away from the idea of uncertainty: could we know this estimand in a situation with basically standard errors of size 0.
- You've probably seen a statistical identification problem before. Let's say you are running a regression on a set of mutually exclusive and exhaustive dummy variables. Say, $M_i = 1$ for male and $F_i = 1$ for not male. You cannot identify the coefficients on both of these variables in the same regression (even if we had infinite sample sizes) because for any value of one coefficient, there is a value of the other coefficient that produces the same conditional mean of the outcome (and thus the same value of the least squares).
- Indentification of statistical models is not inherently about causality because statistical models are not inherently causal. The history of what identification meant in economics is quite interesting and Pearl dedicates part of his book to it. In this class, we'll usually mean "indentification" to mean the statistical identification of causal parameters.

What is the selection problem?

- Let's first look at what we might call the *prima facie* effect, which is just the difference in means between those who take a treatment and those who do not. Let's imagine that this is the average Democratic share of the two-party vote for Democratic Senate candidates that go negative ($A_i = 1$) and those that stay positive ($A_i = 0$).

$$
\begin{aligned}
E[Y_i|A_i = 1] - E[Y_i|A_i = 0] &= E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 0] && (1)\\
&= E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1] && (2)\\
&+ E[Y_i(0)|A_i = 1] - E[Y_i(0)|A_i = 0] && (3)
\end{aligned}
$$

- The second line here is the average treatment effect on the treated and the third line is what we call **selection bias**. It measure how different the treated and control groups are in terms of their potential outcome under control. That is, it measures how different (in terms of potential outcomes) the candidates who went negative are compared to those that remained positive.

- Because of the selection bias, we say that the ATT is **unidentified** because for any value of the ATT there is an amount of selection bias that would create the observed difference in means.
- To see this, imagine we say a negative prima facie effect of negativity. That is, negative Democrats did worse than positive Democrats. This could mean that the ATT is negative and there is a causal effect **OR** it could mean that the ATT is positive and there is an offsetting amount of selection bias.
- Now, you can probably see that if there are bounds on the outcome, the effect isn't completely unidentified because the bounds on Y_i imply that there can only be so much selection bias. This idea is what forms the foundation of nonparametric bounds, which we will talk more about later.

Randomization solves the selection problem

- Randomizing the treatment means that the treated group is a random sample from the population. We know that the mean of a variable in a random sample is an unbiased estimate of that variable in the population. Therefore, the observed mean outcome in a randomly chosen treatment group is the same as the mean outcome in the population.

$$E[Y_i(0)|A_i = 0] = E[Y_i(0)] = E[Y_i(0)|A_i = 1]$$

- Specifically, randomization implies **exchangablility** or **ignorability**: the potential outcomes are independent of the treatments. We write ignorability like this:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i$$

- This is not the same as the treatment being independent of the observed outcomes ($Y_i \perp\!\!\!\perp A_i$). Obviously, if there is a causal effect, then the treatment won't be independent of the outcome in a randomized experiment (even though the randomization guarantees the independence of the potential outcomes and the treatment).
- How does randomization help indentify the causal effect? It ensures that there is no selection bias Note that, because of ingorability:

$$E[Y_i(0)|A_i = 1] - E[Y_i(0)|A_i = 0] = E[Y_i(0)] - E[Y_i(0)] = 0$$

Plugging this in above gives us:

$$\begin{aligned} E[Y_i|A_i = 1] - E[Y_i|A_i = 0] &= E[Y_i(1)|A_i = 1] - E[Y_i(0)|A_i = 1] + 0 \quad (4) \\ &= E[Y_i(1)] - E[Y_i(0)] = \tau \quad (5) \end{aligned}$$

- Randomization in graphs: randomization implies that there is only one arrow into the treatment: that of the randomization.

Types of randomizations/experiments

- Bernoulli trials: flip coins for each person in the experiment. Problematic because there could be very large or very small treated groups.
- Completely randomized experiments: choose a number of treated units N_t and randomly choose N_t units from the N units in the population. Fixes the number of treated units, but all units have the same marginal probability of being treated. Problem: if there are covariates available, then you might get very unbalanced randomizations.
- Stratified randomized experiment: form J blocks, b_j, $j = 1, \ldots, J$, based on the covariates and then use completely randomized assignment in each block. This eliminates the possibility of "bad randomizations" since the treatment is by design balanced within blocks. This type of experiment leads to conditional ignorability: $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i | B_i$, where B_i is the blocking variable.
- Pair randomized experiments: a stratified randomized experiments where each block has 2 units, one of which receives the treatment. An extreme version of the stratified/blocked randomized experiment.
- What type of experiment was the Gerber, Green, and Larimer paper?
- Natural experiment: experiment where treatment is randomized in some fashion, but that randomization was not under the control of the researcher.
- Natural experiments obviously have lots of pitfalls, because we didn't perform the randomization, it's more difficult to justify. How does the Hyde paper do at justifying this? She claims that the election observers mimicked random assignment. Does that seem right?

Effect modification

- We might think that the effect of negativity might vary by whether or not the candidate is an incumbent. That is, for a given covariate, X_i and two different levels of that covariate, x and x^*, we have

$$\tau(x) \equiv E[Y_i(1) - Y_i(0) | X_i = x] \neq E[Y_i(1) - Y_i(0) | X_i = x^*] \equiv \tau(x^*).$$

- The difference between $\tau(x)$ and $\tau(x^*)$ might be causal, in which case call this a causal effect modifier and it might be non-causal in which case we call it a surrogate effect modifier. These surrogate modifiers are indicators of some other variable which is truly the causal modifier.
- In the Wantchekon paper, he looks at effect modification by gender, but this is surely a surrogate effect modifier.
- Now, can we identify these effects, sometimes called **conditional average treatment effects (CATE)**. In a radominzed experiement, yes we can. Note that in a completely randomized experiment we have ignorability: $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i$. Also, ignorability implies conditional ignorability: $(Y_i(1), Y_i(0)) \perp\!\!\!\perp A_i | X_i$.
- By the same logic as before and using the fact that ignorability implies conditional ignorability, we know that

$$E[Y_i | A_i = 1, X_i = x] - E[Y_i | A_i = 0, X_i = x] = \tau(x)$$

- In stratified randomized experiments, we can estimate effect modification by the blocks (and using any variation within blocks since there is just a mini randomized experiment).
- Why do we care about effect modification? Because it's important for external validity. They can also give us some purchase on how the effect works.

Estimation and Inference

Samples versus Populations

Large sample/population/super-population parameters

- We are often interested in making inferences about a large population of which we have a random sample. Let's call the population V.
- We will define the **population average treatment effect (PATE)** as the population average of the individual treatment effects:

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

- We can define the population average treatment effect on the treated (PATT, or τ_{att}) similarly, in addition to the conditional average treatment effect $\tau(x)$.

Finite sample results

- Sometimes instead of making inference about a population, we would rather make inference about the sample that we actually observed. This might make more sense in a lot of political science, where we don't have a larger super population in mind.
- Suppose that we have a sample, S, of units, $i = 1, \ldots, N$ where N_t of the units are treated.
- For this, we can define the **sample average treatment effect (SATE)** as the in-sample average of the potential outcomes:

$$SATE = \tau_S = \frac{1}{N} \sum_{i \in S} Y_i(1) - Y_i(0)$$

- The SATE is the in-sample version of the PATE and for any given sample, won't equal the PATE. In fact, the SATE varies over samples from the population. We're going to ignore this variation when conducting in-sample causal inference and just focus on estimate the SATE for our sample.
- Once we assign some groups to treatment and some to control we do not actually observe $Y_i(1)$ and $Y_i(0)$ and so we cannot actually observe SATE. We can, however, estimate it:

$$\hat{\tau}_S = \frac{1}{N_t} \sum_{i:A_i=1} Y_i - \frac{1}{N_c} \sum_{i:A_i=0} Y_i$$

- Note that, conditional on the sample, the only variation in $\hat{\tau}_S$ is from the treatment assignment. Unconditionally, there are two sources of variation: the treatment assignment and the sampling procedure.

- We can show that, with a completely randomized experiment assignment, $\hat{\tau}_S$ is unbiased for τ_S and, in fact, τ:

$$E[\hat{\tau}_S|S] = \frac{1}{N_t}\sum_{i:A_i=1} E[Y_i|A_i=1,S] - \frac{1}{N_c}\sum_{i:A_i=0} E[Y_i|A_i=0,S] \tag{6}$$

$$= \frac{1}{N_t}\sum_{i:A_i=1} E[Y_i(1)|S] - \frac{1}{N_c}\sum_{i:A_i=0} E[Y_i(0)|S] \tag{7}$$

$$= \frac{1}{N_t}N_t E[Y_i(1)|S] - \frac{1}{N_c}N_c E[Y_i(0)|S] \tag{8}$$

$$= E[Y_i(1) - Y_i(0)|S] = \frac{1}{N}\sum_{i\in S} Y_i(1) - Y_i(0) = \tau_S \tag{9}$$

- By the law of iterated expectations, we also know that $E[E[\hat{\tau}_S|S]] = E[\tau_S] = \tau$. Thus, the difference in means is also unbiased for the PATE.

- It turns out that the sampling variance of the difference in means estimator is:

$$V(\hat{\tau}_S|S) = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{\tau_i}^2}{N},$$

where S_c^2 and S_t^2 are the in-sample variances of $Y_i(0)$ and $Y_i(1)$, respectively. We can use sample variances within levels of A_i to estimat these. The last term, $S_{\tau_i}^2$ is the in-sample variance of the individual treatmente effects. Obviously, we don't observe any individual treatment effects, so we can't estimate a sample variance of this quantity. If the treatment effect is constant, then this term equals zero.

- It turns out that the overall variance of the estimator is simply:

$$V(\hat{\tau}_S) = \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t},$$

which can be estimated with this simple variance estimator:

$$\hat{V} = \frac{\hat{s}_c^2}{N_c} + \frac{\hat{s}_t^2}{N_t}$$

- This estimator is unbiased for the variance of the difference in means in the population OR a conservative estimate of the variance of the difference in means in the sample.

Can we use regression with experiments?

- We can just run a regression of the outcome on a binary treatment indicator. Note that this works even if the outcome is binary because this is just a difference in means test.

- First, let's remember how we relate the potential outcomes to the observed outcome:

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0) \tag{10}$$
$$= A_i Y_i(1) + (1 - A_i) Y_i(0) + E[Y_i(0)] - E[Y_i(0)] + A_i E[Y_i(1) - Y_i(0)] - A_i E[Y_i(1) - Y_i(0)] \tag{11}$$
$$= E[Y_i(0)] + A_i E[Y_i(1) - Y_i(0)] + (Y_i(0) - E[Y_i(0)]) + A_i \cdot ((Y_i(1) - Y_i(0)) - E[Y_i(1) - Y_i(0)]) \tag{12}$$
$$= \alpha + A_i \tau + \epsilon_i \tag{13}$$

- See that $\alpha = E[Y_i(0)]$ and remember that $\tau = E[Y_i(1) - Y_i(0)]$. And also the residual here is the deviation for the control group plus the treatment effect hetergeneity.
- Let's check to see if the errors here are independent of the treatment, which would imply that a regression estimator $\hat{\tau}_{\text{ols}}$ would be unbiased for τ:

$$E[\epsilon_i | A_i = 0] = E[Y_i(0) - E[Y_i(0)] | A_i = 0] = E[Y_i(0) | A_i = 0] - E[Y_i(0)] = 0$$

and

$$E[\epsilon_i | A_i = 1] = E[Y_i(1) - E[Y_i(0)] + E[Y_i(1) - Y_i(0)] | A_i = 1] = E[Y_i(1) | A_i = 1] - E[Y_i(1)] = 0$$

- Thus, just using the randomization assumption, we have justified the use of regression.
- Randomization implies that we don't have to adjust for any covariates when estimating causal effects.
- But we know that, on average, the treatment will be uncorrelated with any covariates, so adding them to a regression won't change the consistency of the estimator, even if the regression is misspecified, for instance, because the true population regression function is nonlinear. It does have two other effects, though. One is that it can make the treatment effect estimates more precise. The other is that it can add finite sample bias.
- The finite sample bias is due to the fact that even if randomization implies the population correlation between the treatment and covariates will be zero, it won't be zero in finite samples, which will bias our estimates.
- Should we use a logit or a probit for a binary outcome? With an experiment that focuses on estimating the treatment effect, probably not.

Details about Experiments

Diagnostics

- There are we can check to see that the randomization worked by comparing the treatment groups on a host of background covariates. The treatment status should be unrelated to these background covariates. Now this doesn't mean that the treatment is unrelated to the potential outcomes, but it does give us confidence that the randomization worked at it was supposed to on the observables. Also note that some of the covariates could be related to the treatment by random chance. With 20 variables, one of them should have a p-value below 0.05 by random chance.

Blocking

- Sometimes with a completely randomized experiment, you end up with bad randomizations.
- SATE for blocked estimators is just the weighted average of the within-block ATE estimate. The weights are the size of the blocks relative to the size of the sample.