Rubin's Potential Outcome Framework

Paul L. Hebert, PhD

Motivation

- Rubin's potential outcomes model has been the theoretical framework that statisticians/econometricians use to demonstrate under what conditions/assumptions their analytic models generate causal interpretations—i.e., that treatment A is not just correlated with better outcomes but causes better outcomes
- The math can get dense, but the core concepts are fairly simple.
- An introduction to them is useful for
  - Framing your own work
  - Understanding concepts like heterogeneous treatment effects, the Average Treatment Effect (ATE), and the effect of the Treatment on the Treated (TT)

Example: Crossfit versus boot camp

Treatment (D): Crossfit vs. Boot camp

D=0 boot camp

D=1 crossfit

My potential outcomes from crossfit or bootcamp

My potential outcome after crossfit

Mean result of crossfit for people with characteristics X (i.e., people like me).

My idiosyncratic result from crossfit

$$Y_1=\mu_1(X) + U_1$$

$$Y_0=\mu_0(X) + U_0$$

My potential outcome after boot camp

Mean result of boot camp for people with characteristics X (i.e., people like me).

My idiosyncratic result from boot camp

Counterfactual treatments and counterfactual outcomes

$$Y_1=\mu_1(X) + U_1$$

$$Y_0=\mu_0(X) + U_0$$

- NOTE: You cannot observe both Y_1 and Y_0; only Y_1 OR Y_0
- If I take crossfit for a year, the outcome is Y_1; the counterfactual treatment is boot camp, and the counterfactual outcome is Y_0.
  - I can't go back in time and take boot camp for that year
- Since we can't observe the counterfactual outcome, we have to come up with a way of estimating it.
  - Good practice to ask yourself what the counterfactual is before you start an analysis. Sometimes the answer is more complex than you think.

My Potential Outcomes

The effect of crossfit compared to boot camp for me

My idiosyncratic benefit from crossfit over boot camp. This is treatment heterogeneity

$$\Delta=(Y_1-Y_0)=\underbrace{\{\mu_1(X)-\mu_0(X)\}}+(U_1-U_0)$$

- Average difference in outcomes for crossfit over boot camp for someone like me.
- This is the Average Treatment Effect.
- This is what you'd get from a Randomized Controlled Trial.

Source: Basu A, Heckman J, et al (2007)

The treatment effect on the treated (TT)

- Although we mostly concentrate on the average treatment effect (ATE), the TT is also interesting. For people with observed covariates x, TT(X=x) is defined as:

$$TT(x) = E(\Delta | X=x, D=1) + \underbrace{E(U_1-U_0 | X=x, D=1)}_{\text{This is the idiosyncratic gain from crossfit among those who chose crossfit over boot camp.}}$$

Source: Basu A, Heckman J, et al (2007)

From potential outcomes to observed outcomes and something you can estimate

If D=1 (i.e., I go to crossfit) the Y you see is Y_1; if D=0 (i.e., I went to boot camp) Y= Y_0

$$Y=D^*Y_1+(1-D)^*Y_0$$

Now average over everyone, so we're not just talking about me anymore but the population

$$Y=\mu_0(X)+E(\Delta|X)^*D + \underbrace{\{D^*(U_1-U_0) + U_0\}}_{\text{Whether the coefficient on D gives the ATE depends on these last terms}}$$

This is something you can estimate because you can observe Y and X and D. E(Δ|X) would be the coefficient on treatment (i.e., D)

Source: Basu A, Heckman J, et al (2007)

Randomized Controlled Trials

Now take mean difference in Y between participants in a randomized trial of crossfit vs, boot camp: $E(Y|D=1)-E(Y|D=0)=$
$=[\mu_1(X)+E(U_1|D=1)]-[\mu_0(X)+E(U_0|D=0)]-$
$=E(\Delta|X)+E(U_1-U_0|D)$

- The coefficient on D is the ATE if either U_1-U_0=0 (i.e., there is no unobserved treatment heterogeneity—the treatment effects everyone exactly the same way), or D is independent of (U_1-U_0) then.
- In an RCT you expect both.
  - Randomization gives some reason to believe D is uncorrelated with U_1-U_0
  - Only people who join an RCT are those who think U_1-U_0=0 for them.

OLS regression

$$E(Y|X=x,D=1)-E(Y|X=x,D=0)=$$
$$=E(\Delta|X) +[E(U_1-U_0|X=x,D=1)+ \{E(U_0|X=x,D=1)-E(U_0|X=x,D=0)\}]$$

- Whether an OLS regression gives you the ATE depends on the error term in the square bracket
  - Blue part reflects whether people who would gain more from crossfit tend to choose crossfit over boot camp
  - Red part difference in unobserved potential outcome after boot camp among people who chose crossfit versus boot camp. (E.g, even if they had taken boot camp, people who took crossfit would be able to do more push-ups than boot camp people, maybe).
- Is it reasonable to assume the term in braces is zero?

Source: Basu A, Heckman J, et al (2007)

Propensity Matching

- Propensity matching says after matching on the propensity score, which is a function of the X's, $E(U_1-U_0|P(X),D)=0$
- That is, there is someone in boot camp that is just like me and has the same idiosyncratic gain to crossfit over boot camp.
- Reasonable assumption?

Instrumental variables

- Assumes that you can find some variable (an instrumental variable) that divides your sample into groups that differ only by the type of treatment they got.
- Provides an estimate of the ATE if the instrument is independent of selection on potential gains.
- Whether or not this is reasonable depends entirely on the instrument
  - More on IVs on Wed and Thursday

Summary

- Potential outcomes framework is the basis for most of the theoretical work on causal modeling
- Important concepts include:
  - Counterfactual outcomes
  - Treatment heterogeneity
  - Average treatment effects (ATE)
  - Treatment effects on the treated (TT)

References & Resources

1. Basu, A., Heckman, J. J., Navarro-Lozano, S., & Urzua, S. (2007). Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Econ, 16(11), 1133-1157. doi: 10.1002/hec.1291
2. Slide 3 contains images from:
   1. lululemon athletica's: http://www.flickr.com/photos/lululemonathletica/4883555927/ (Creative Commons 2.0)
   2. Math_dum's: http://www.flickr.com/photos/crossfit_capop/5490895210/ (Creative Commons 2.0)
   3. CrossFit H2O: http://www.flickr.com/photos/crossfit_h2o/6110119419/ (Creative Commons 2.0)