

Lecture 4: Social Experimentation and Program Evaluation

Problem: The Missing Counterfactual

We're interested in evaluating the outcome of some treatment (eg employment outcome resulting from a government-funded training program).

Define

y_i = the observed outcome for individual i

y_{1i} = the outcome for individual i if i receives treatment

y_{0i} = the outcome for individual i if i does not receive treatment

$d_i=1$ if i receives treatment

$d_i=0$ if i does not receive treatment

We never simultaneously observe y_{1i} and y_{0i} . The unobserved outcome for a treated (untreated) person without (with) the program is our missing counterfactual.

What we observe is $y_i = y_{0i}(1-d_i) + y_{1i}d_i$. Assuming

$\beta_i = y_{0i}$ and $\alpha_i = y_{1i} - y_{0i}$, we can write

$$y_i = \beta_i + \alpha_i d_i$$

Here α_i is the individual treatment effect. Though we'd like to know the entire distribution of treatment effects α_i , because of the missing counterfactual problem we generally limit our interest to one of the following two measures.

$\bar{\alpha} = E[y_{1i} - y_{0i}] = E[\alpha_i]$, the population mean impact

$\tilde{\alpha} = E[y_{1i} - y_{0i} | d_i = 1] = E[\alpha_i | d_i = 1]$, the mean effect of
treatment on the treated

Note that if the individual treatment effect is constant throughout the population then these are the same.

Our problem is that we only observe y_{1i} for those in the program and y_{0i} for those out of it. We can obtain

$E[y_1 | d=1]$ and $E[y_0 | d=0]$, and thus we could estimate

$$\hat{\alpha} = E[y_1 | d=1] - E[y_0 | d=0].$$

This term is an unbiased measure of $\bar{\alpha}$ only if $E[y_1 | d=1] = E[y_1]$ and $E[y_0 | d=0] = E[y_0]$, which is not generally the case.

Introducing additional explanatory variables by assuming $\beta_i = x_i\beta + \varepsilon_i$, we get the regression form

$$y_i = \alpha d_i + x_i\beta + \varepsilon_i \quad i=1, \dots, n. \quad (1.1)$$

We can now write

$$E[y_i | d_i = 1, x_i] = \alpha + x_i\beta + E[\varepsilon_i | d_i = 1, x_i]$$

$$E[y_i | d_i = 0, x_i] = x_i\beta + E[\varepsilon_i | d_i = 0, x_i]$$

and differencing the two gives the previously suggested estimator of alpha

$$E[y_i | d_i = 1, x_i] - E[y_i | d_i = 0, x_i] = \alpha + (E[\varepsilon_i | d_i = 1, x_i] - E[\varepsilon_i | d_i = 0, x_i])$$

This estimator yields alpha and a bias term. In the rare case in which the bias term in the treatment and control groups is the

same, our simple differencing approach provides an unbiased estimate of alpha.

Randomized Clinical Trial (Experiment)

When participants are randomly assigned to treatment and control groups, this is done to insure that

$$E[y_0 | d = 0] = E[y_0 | d = 1] = E[y_0]$$

The case of random assignment unconditional on x_i implies that

$$E[\varepsilon_i | d_i = 1, x_i] = E[\varepsilon_i | d_i = 0, x_i],$$
 and our bias term in the

difference estimate is zero. The difference estimate yields an unbiased estimate of the population mean treatment effect.

Selection Model

When assignment into the program is nonrandom, dependence between d and ε can lead to selection bias in the estimation of α .

The dependence arises when the treatment status of an individual is related to some characteristic not in x that itself is related to program outcome.

If the characteristic is observable, we call this ‘**selection on observables**’. If it is unobservable, we call this ‘**selection on unobservables**’.

Our selection model is:

$$y_i = \beta_0 + \alpha d_i + \beta_1 x_i + \varepsilon_i \quad (2)$$

$$d_i^* = \delta_0 + \delta_1 x_i + \delta_2 w_i + v_i \quad (3)$$

$$d_i = 1 \text{ if } d_i^* \geq 0 \quad (4)$$

$$0 \text{ if } d_i^* < 0$$

Conventional rule: OLS is biased if ε_i and d_i are not independent.

(i) Selection on Observables

ε_i and ν_i independent, but w_i not independent of ε_i .

Solution: We assume a form of the conditional expectation (a 'control function') $h(w_i) = E[\varepsilon | w_i]$ and we include this in (2). For example, we assume $\varepsilon_i = \gamma_0 + \gamma_1 w_i + \nu_i$ with $E[d_i \nu_i | x_i, w_i] = 0$, and we estimate

$$y_i = \bar{\beta}_0 + \alpha d_i + \bar{\beta}_1 x_i + \bar{\beta}_2 w_i + \bar{\varepsilon}_i$$

using OLS. The addition of w_i to the expression proxies for the role of the selection equation and absorbs the bias.

(ii) Selection on Unobservables

ε_i and w_i independent, but ν_i not independent of ε_i . w_i is an instrument (exclusion restriction).

Instrumental variables estimation (Heckman & Robb (1985), Angrist (1991)):

For identification of treatment effects without parametric assumptions on the joint error distribution, we require at least one regressor in the selection equation (3 here) that isn't in the outcome equation (2). w_i is our exclusion restriction.

Suppose that w_i represents inclusion in a particular policy, so that $w_i = 1$ if offered a policy, $w_i = 0$ if not. (Consider a schooling voucher offered to students in one district but not those in the

adjacent one.) Let \bar{d} be the proportion of those offered the policy who take it up. Then the IV estimate of the average effect of the treatment on the treated is

$$\alpha_{IV} = \alpha_{2SLS} = \frac{\overline{y_1} - \overline{y_0}}{\bar{d}},$$

where

$\overline{y_1}$ = mean of y for those with $w_i = 1$

$\overline{y_0}$ = mean of y for those with $w_i = 0$

The exogenous variation in eligibility for the program is used as an instrument, allowing us to control for selection into the program through the difference in outcomes between eligibles and ineligibles, along with the participation rate among eligibles.

Note that our estimator throws out all information on the within- w_i group correlation of y and d , and only uses between- w_i group variation.

This approach can also be used for randomized clinical trials in which nonparticipation is allowed. Here the experimental group becomes those who are *offered* the program, whether or not they take it up. The control is those who are not offered the program.

There exist several approaches to correction for selection on unobservables in the absence of fully exogenous variation in program eligibility. I mention some of them here, with related references and limited discussion.

(a) Nonparametric bounds on the treatment effect (Manski)

Manski uses a series of assumptions on the bounds on possible Y_0 and Y_1 , which vary from the most general to more restrictive assumptions, to find tighter and tighter bounds on the possible ranges of $E[Y_0 | X, D=1]$ and $E[Y_1 | X, D=0]$. From these he derives narrowing bounds on the possible sizes of the treatment effect.

Examples of assumptions used to create first weak and then stronger bounds are:

- (i) If we're willing to assume the X 's and $Y_0 | X$ are bounded, what bounds can we set on $E[Y_0 | X, D=1]$ and $E[Y_1 | X, D=0]$?
- (ii) If we're willing to assume that agents select the treatment with the better outcome, what bounds can we set on $E[Y_0 | X, D=1]$ and $E[Y_1 | X, D=0]$?

Obviously one major interest of this approach is what assumptions are required in order to set bounds that exclude a 0 value of the treatment effect.

(b) SATE-LATE (Angrist and Imbens, EMA 1995)

Angrist and Imbens take as their objective the estimation of the 'selected average treatment effect' (SATE), or the mean treatment effect on the treated

$$E(Y_1 - Y_0 | d=1).$$

They demonstrate that in the presence of selection on unobservables, if there exists an instrument Z such that the outcome absent treatment is independent of Z and there is some positive probability that a given Z value is such that the individual is excluded from the treatment group, then using Z one can consistently estimate the SATE.

(c) Propensity score matching (eg Heckman, Ichimura, & Todd REStud 1998)

The propensity score matching method uses a treatment propensity score $P(D=1/X) = P(X)$ to compare the outcomes of treated and untreated individuals whose observable characteristics are similar in the sense that they imply similar propensities for being observed in the treated group.

The propensity score matching literature comes from applied statistics, and is preceded by a straight matching approach. In it treated and untreated individuals are paired based on their X vector “cells” and their outcomes are compared, generating an estimate of the treatment effect.

A remaining issue in this literature is that it is unclear whether the efficient approach to estimation, in terms of minimizing the variance of the matching estimator, is to match on X or on $P(X)$.

Matching studies abound. The debate over the viability of and choice among non-experimental estimators of program effects includes several matching and other non-experimental estimates of the effect of the National Job Training Partnership Act (JTPA) Study, including work by LaLonde, Heckman (both discussed below), Ichimura, Smith, Todd, Dehejia & Wahba.

(d) Regression Discontinuity (Hahn, Todd and Van der Klaauw, 2000)

The regression discontinuity approach is a quasi-experimental design in which the probability that treatment is offered to an individual is a discontinuous function of one or more underlying variables. Some examples of regression discontinuity studies include

Van der Klaauw (1997), in which discontinuities in the function of (for eg) SAT scores that a college uses to allocate financial aid are used to estimate the effect of aid on a student's decision to attend a particular college

and Black (1999), in which the RD approach is used to estimate parents' willingness to pay for school quality by comparing housing prices near school district borders.

(e) and many, many others. See your syllabus for a few useful references.

The Debate Over Social Experimentation

reference: G. Burtless, "The Case for Randomized Field Trials in Economic and Policy Research," JEP 1995, J. Heckman, "Assessing the Case for Social Experiments," JEP 1995.

Social experiment definition "In the context of social science, a randomized field trial (or social experiment) is simply an experiment that takes place outside a laboratory setting, in the usual environment where social and economic interactions occur. In the simplest kind of experiment, a single treatment is assigned to a randomly selected subsample (the treatment group) and withheld from the remainder (the control or null-treatment group)."

(Burtless) Note that in an experiment observation units are randomly assigned to treatment and control. A demonstration is similar to an experiment, except that it includes no control.

2 types of experiments:

Structural experiment Example: the Seattle NIT experiment tested a variety of income guarantees and tax rates in order to estimate the labor supply functions of the low-income population. The advantage of the structural experiment is that estimation of

behavioral parameters allows extrapolation from the actual experiment performed to other related policy steps.

Black box experiment The intervention tested represents a unique policy intervention. Even in the case of multiple interventions, the steps cannot be parameterized as points along a policy continuum. Therefore the results of the experiments demonstrate the likely results of the tested reform, but do not yield estimated behavioral parameters. The results cannot be used to extrapolate from the outcome of the actual experiment the likely outcome of related but nonidentical policy reforms. Most recent experiments have been of this type.

Points of agreement

Burtless and Heckman generally agree that well-designed social experiments in some cases can be useful in that they get around the selection problems discussed in this lecture.

Burtless's discussion of arguments in favor of social experimentation

- (1) The assignment procedure assures us of the direction of causality between treatment and outcome: differences in average outcome among treatment groups are caused by differences in treatment, and not v/v .
- (2) Experiments permit analysts to measure the effects of economic stimuli that have not been observed previously. For example, many politicians believe that employers would hire many more of the disabled if offered a large subsidy for employing these workers. Variation in subsidies for employing the disabled to date is unlikely to provide enough information about responses to large

expansions in subsidies, and so an experiment could be informative in addressing this question.

- (3) The simplicity of experiments (which involve evident control groups and few identifying assumptions to be explained) offers notable advantages in making results convincing to other social scientists and understandable to policymakers.

This last point has been particularly important in recent years, and experimentation has become increasingly influential in the construction of US policy.

Finally, Burtless cites studies by LaLonde (1986) and Fraker and Maynard (1987) in which the researchers used data from a true randomized trial, the National Supported Work Demonstration, and built both experimental and nonexperimental estimates of the treatment effect. They find the nonexperimental estimators unreliable in predicting the treatment effect found in the randomized trials. Based on this evidence, Burtless questions the ability of econometrics to provide a fix for the selection problem.

Heckman's Response to Arguments for Social Experimentation

(First note that Heckman observes that recent trends in social experimentation favor the black box study over the structural experiment, so that results do not provide means of predicting the outcomes of similar but nonidentical policy steps.)

The selection problem is universal, and nonexperimental methods cannot solve it

In this discussion Heckman responds to LaLonde (1986).

As noted in the discussion of Burtless's points in support of social experimentation, LaLonde uses an experimental evaluation of the National Supported Work Demonstration (NSW) as a benchmark, against which he compares nonexperimental estimates based on comparison of NSW treatment groups to CPS and PSID nonexperimental comparison groups.

Heckman's argument is that flaws in LaLonde's study limit the generalizability of his results.

Heckman's first argument is that selection bias comes from missing data on the common factors that affect both participation and outcomes, and that 'the most convincing way to solve the problem is to collect better data'. Recent studies by Heckman, Ichimura, Smith and Todd (1995) and by Heckman and Roselius (1994) show that sufficient data on individuals eligible for the JTPA and located in the same labor market can be used to create a nonexperimental comparison group that's virtually identical to the control group in the JTPA.

Heckman notes data limitations in LaLonde's nonexperimental comparison group including small sample size, lack of information on eligibility criteria and lack of geographical information, concluding 'the selection bias problem documented by LaLonde (1986) arises at least in part from the crudity of his data'.

Additionally, Heckman points out that the LaLonde study does not include model-selection strategies standard specification tests. Heckman and Hotz (1989) revisit the LaLonde data and add to the study a series of model specification tests. All nonexperimental models excepting those that reproduce the inference obtained by the experimental approach are rejected by the specification tests.

Finally, Heckman emphasizes that substantial progress has been made in nonexperimental approaches since 1986, particularly in the realm of semi- and nonparametric estimation.

Experiments are based on more plausible assumptions

Heckman compares the sets of assumptions required by the experimental and nonexperimental approaches to treatment evaluation:

Experimental

(1) Randomization must not alter the process of selection into the program; those who participate during an experiment must not differ from those who would have participated absent an experiment.

(2) Members of the experimental and control groups cannot obtain close substitutes for the treatment elsewhere. ie, there is no 'substitution bias'. In the presence of substitution bias, the control group no longer corresponds to the set of persons who wanted but did not receive treatment.

Nonexperimental

(1) A model of the outcome process can be determined, along with
(2) the relationship between the outcome process and the process of selection into the program. eg for a training program, economic theory must supply a model of the earnings behavior of the population served by the program, and the effects of earnings on selection into the program must be determined.

[Heckman notes here that nonexperimental evaluations 'can build on cumulative knowledge about earnings and selection processes

from prior studies' and information about selection and outcome process from the current data.]

Heckman argues that both sets of assumptions are easy to understand.

Experimental results are easier to explain to policymakers

Heckman contends that in the presence of randomization or substitution bias, explaining the results of an experiment to a congressional committee is likely to be just as difficult as explaining the results of a nonexperimental study.

Heckman's comments: 'The hard fact is that some evaluation problems have intrinsic levels of difficulty that render them incapable of expression in sound bites. Delegated expertise must therefore play a role in the formation of public policy in these areas, just as it already does in many other fields. It would be foolish to argue for readily understood but incompetent studies, whether they are experimental or not.'

Additionally, Heckman argues that the ability of nonexperimental studies to derive behavioral parameters used to predict the effects of similar but nonidentical policy interventions on outcomes can be of more use to politicians in shaping policy.

Experiments produce a consensus

Experiments tend to produce 'one number', rather than an array of confusing nonexperimental estimates. Heckman argues that this is consensus produced by monopoly over the data, as opposed to consensus produced by scholarship. We can expect to see only one estimate of the program effect if the organization conducting the experiment does not share the resulting data.

Next time:

--Regression discontinuity in general, from Hahn, Todd, & van der Klaauw (EMA 2000),

--& an RD application to higher education financial aid: van der Klaauw (1997).