Lecture 4 Notes: Experiments

Econ 4523: Economics of Education
Prof. Tyler Ransom
University of Oklahoma

Fall 2019

Review from last class

- Correlation formula
- Interpretation of positive correlation: "units with large values of X have *on average* large values of Y"
- Correlation $\not\Longrightarrow$ Causation
- Causation $\Longrightarrow$ Correlation
- Confounders and the selection problem
- Treatment effect $= Y_T - Y_C$
  - But we only observe Y_T or Y_C (the other one is the counterfactual)

Today's Big Questions

- What is a controlled experiment?
- How is it different from a randomized experiment?
- How does an experiment help you learn about causality?

1 Controlled experiments

Controlled experiments are fundamental to understanding causality. How do they work? Here's an example from the food industry.

- Take two units, say cartons of yogurt

- These two units are exactly the same in every way (they came from the same batch in the factory, etc.)
- Put one carton in a refrigerated chamber with the tempreature set to 40°F
- Put the other carton in a (slightly) heated chamber with the tempreature set to 75°F
- After awhile, measure the level of microbes in each carton
- Any difference in the level of microbes will be solely due to differences in temperature.
- Treatment effect = microbes in yogurt B − microbes in yogurt A

∗ Food companies actually use this method to help them figure out what the expiration date on their food should be. Pretty neat!

Q: Are controlled experiments possible in social science? Why or why not?

1.1 Controlled experiments in social sciences

There are a couple of problems with performing controlled experiments in the social sciences:

1. It's very difficult to find a reliable measurement for variables that matter in social science contexts (e.g. competitiveness, effort, perseverance). If you can't measure these, you can't be sure that the two units you're comparing are the same.
2. There are a massive number of variables that matter in social science contexts. So it's impossible to find two units (e.g people) who have the exact same value of every single variable, but one is treated and the other is not.
    - In the yogurt example above, there's really only a few variables that matter for the production of microbes: temperature, humidity, and the contents of the food (water content, pH, etc.). So it's much easier to isolate these few variables.

Q: What's the solution to this controlled experiment conundrum in social sciences?

2 Randomized experiments

Here's the basic intuition for a randomized experiment:

- Take a large sample of units (1,000+)
- Split them into two groups: treatment and control
- Units are assigned to the groups at random (e.g. flip a coin)

- Everyone in the treatment group gets treated, everyone in the control group does not
- If the split was random and the sample is sufficiently large, then the two groups should look the same *on average*.
- In this sense, the control group (as a whole) serves as a counterfactual for the treatment group (as a whole)
- We say that randomization creates **balance** across the two groups, meaning that there are no systematic differences in any of the pre-treatment (a.k.a. control) variables across the two groups.
- Thus, any systematic difference in outcomes is solely due to the treatment.
- The **average treatment effect** (ATE) is then the average difference in the outcome between the treatment and control groups.
- Remember: you can't compare individual units across treatment and control, you can only compare the two *groups*

In math terms, the ATE is:

$$\text{Average Treatment Effect} = \text{average outcome in treatment group} - \text{average outcome in control group}$$

2.1 Example: Project STAR

Project STAR (Stduent/Teacher Achievement Ratio) was an experiment to better understand how class size affects student learning. It was implemented in Tennessee from 1985-1989 for students in grades K-3.

- Treatment group: Students placed into smaller classes
- Control group: Students remain in previous-sized classes (i.e. no change)
- Outcome: Achievement test scores at the end of each grade

From Schanzenbach (2006), Table 2:

Characteristic	Treatment	Control	P value
Free lunch	0.47	0.48	0.46
White or Asian	0.68	0.67	0.66
Age in 1985	5.44	5.43	0.38
Female	0.49	0.49	0.87
Attrition rate	0.49	0.52	0.01
Days absent	10.00	10.50	0.01
Class size in kindergarten	15.10	22.40	0.00
Standardized test score in kindergarten	0.17	0.00	0.00

So the ATE of class size on test scores is $0.17 - 0.00 = 0.17$ where here the units of test score are standard deviations. So the interpretation is that being in a smaller-sized class in kindergarten raises students' test scores **on average** by 17% of a standard deviation. For a test like the SAT, the standard deviation is about 100 points. So this would translate to a 17-point increase in the SAT score.

2.2 Validity of experiments

There are two types of validity that we speak of when discussing the results of experiments:

1. Internal validity: is the ATE unbiased?
2. External validity: is the ATE generalizable to other contexts?

Factors that affect internal validity:

- Differential non-response or attrition rates between the two groups
- Sampling error (problematic for experiments with small sample sizes)
- Hawthorne effect (i.e. I act differently if I know I am in the treatment group)
- Noncompliance with treatment assignment (i.e. I don't want to be in the treatment group, so I purposely don't take the medicine)

Most experimental results have nothing to say about external validity. That said, a broad enough experiment could be externally valid. The factors here are:

- Sample size (unlikely that sample is representative of population)
- Sample specificity (results for college graduates not representative of non-college graduates; results for Oklahoma not representative of New York)

3 Why don't we always use experiments?

There are a number of reasons why experiments won't always work:

1. Impossible to implement (e.g. too expensive or unethical to implement)
    - In fact, before conducting an experiment, researchers need approval from an ethical review board (Institutional Review Board)
    - Ethics often affect the scope of the research
2. Research limitations: It may be that the hypothesis you'd like to test just isn't possible to quantify using an experiment
    - This is particularly pertinent to settings where treatment at the moment affects outcomes in the future. Economists call these *dynamic settings*.

    - In dynamic settings, treatment in the past can be confounded by variables in the future. This greatly complicates the analysis

3. We want to know about policy mechanisms that we can't randomize on
    - Going back to the Project STAR example, suppose we want to know *why* kindergarteners' test scores increased when they were assigned to smaller classes?
    - Perhaps it was because teachers in small classes are able to use different teaching methods than teachers in larger classes
    - Or perhaps being in a smaller class allows students to concentrate more on the teacher because there are fewer distrcations
    - We can't use an experiment to distinguish between these two explanations

3.1 How can we learn about causality when we can't use an experiment?

When we can't use an experiment, we must use observational data. There are a variety of non-experimental and so-called quasi-experimental methods that can be used to recover causal effects in these settings. We'll cover these in the next two lectures.

Video links

- https://www.youtube.com/watch?v=S5TVIPknDI4
- https://www.youtube.com/watch?v=crpuBZv6XtA
- https://www.youtube.com/watch?v=SY6m8ghxz7A
- https://www.youtube.com/watch?v=JnD3h82IxrA
- https://www.youtube.com/watch?v=XiVk8uwptCw
- https://www.youtube.com/watch?v=GGYLNvaHjOY

Other links

- https://www.princeton.edu/news/2017/08/22/orange-new-green-how-orange-peels-revived
- https://www.washingtonpost.com/news/speaking-of-science/wp/2016/11/04/can-the-fear-
utm_term=.5d0e4a4c42da

References

Schanzenbach, Diane Whitmore. 2006. "What Have Researchers Learned from Project STAR?" *Brookings Papers on Education Policy* (9):205–228.