1. Research design for causal inference

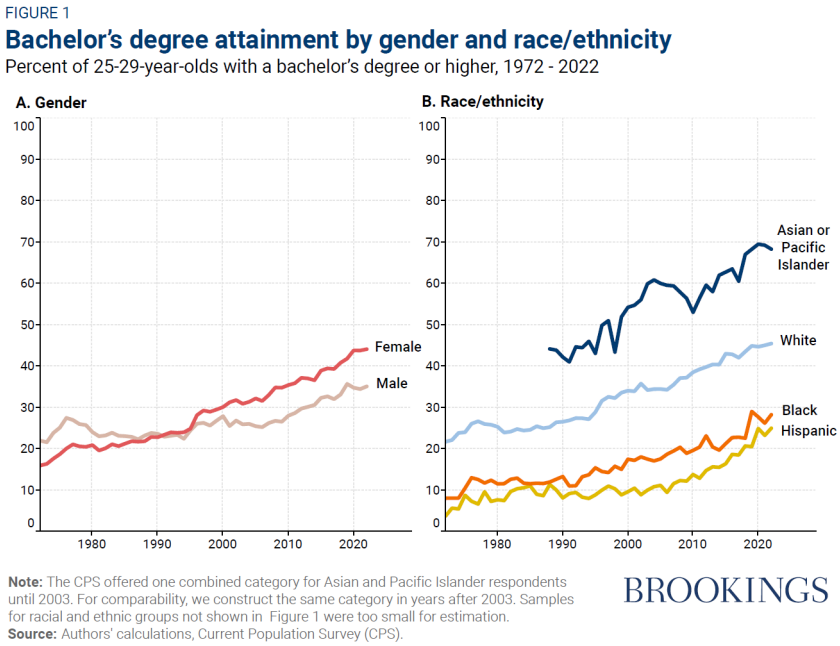
LPO.7870: Research Design and Data Analysis II

Sean P. Corcoran

Introduction to Data II

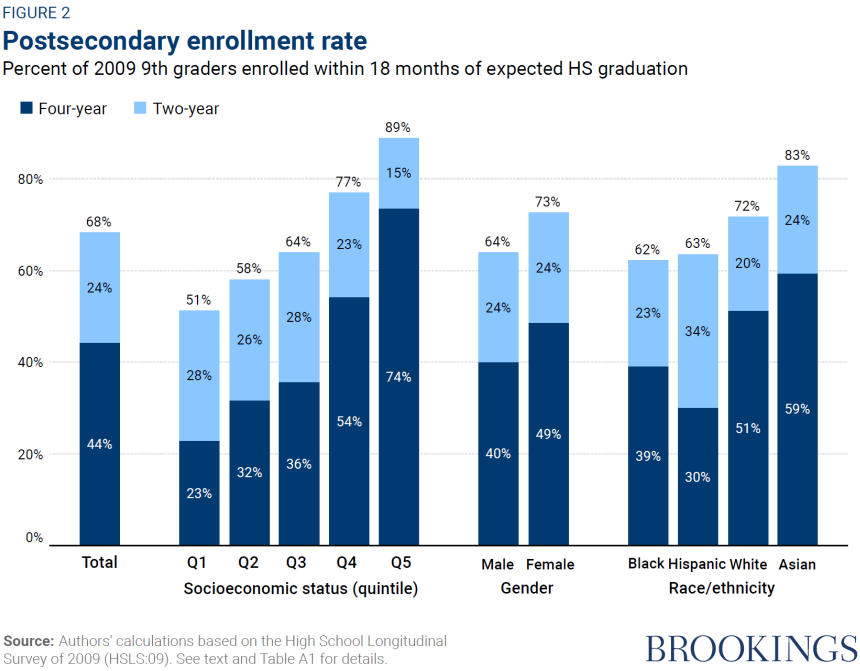
- The primary aim of this class is to further develop your skills as policy and data analysts in the field of education. It builds on what you learned in Data I (LPO 7860).
- A big part of this is understanding what constitutes *good research design*—recognizing it when you see it, knowing how to interpret and critically read published research, and properly applying and implementing strong designs with real data.
- Often, a good research design is one that allows us to make **causal inferences** about the effect of a policy, program, input, or risk factor on an outcome of interest.

Example: gaps in postsecondary enrollment



Source: Reber & Smith (2023)
`https://www.brookings.edu/articles/college-enrollment-disparities/`

Example: gaps in postsecondary enrollment



Source: Reber & Smith (2023)
`https://www.brookings.edu/articles/college-enrollment-disparities/`

Example: gaps in postsecondary enrollment

These figures raise a lot of questions:

- What explains gaps in postsecondary enrollment by income, gender, and race?
- What explains observed *changes* in postsecondary enrollment gaps over time?
- What programs and/or policies would be effective in *reducing* gaps in postsecondary enrollment? Which would be most cost effective?

These are inherently causal questions, and call for more than just correlational (descriptive) evidence.

Examples of other questions in education policy

- Do smaller class sizes lead to better student outcomes?
- How important is teacher experience to student learning?
- Do charter schools produce better educational outcomes than traditional public schools?
- Will "letter grade" accountability improve schools?
- How will a publicly-funded voucher for private school enrollment affect student outcomes and traditional public schools?
- Do conditional cash transfer programs improve childrens' schooling and health outcomes in low-income countries?

Features of high-quality research

Whether a causal design or not, high-quality empirical research in the social sciences and education shares some common features:

- Clear statement of the research question
- An underlying theoretical framework
- Clear statement of the population of interest
- Representative data
- Appropriate measures of:
  - The outcome of interest (sometimes called "dependent variable")
  - The key explanatory or predictor variable (sometimes called "independent variable")
  - Covariates or control variables

The importance of theory

- This course will focus on the empirical tools used to provide answers to these questions. These tools cannot be used in isolation, however.
- *Theory* is needed as a framework for understanding causal questions.
- Theory guides what questions to ask, key constructs to measure, and hypothesized relationships. It is also often informative about potential *mechanisms*.
- Theory may be a formal, well-developed model, or a more informally reasoned *theory of change* or *logic model*.

Application to gaps in postsecondary enrollment

What kind of theory might we use to think about gaps in postsecondary enrollment by income, race, and gender?

Economics has developed a useful model for thinking about investments in education: the **human capital model**. In this model, the decision to obtain additional education depends on its *benefits* and *costs*. How might that be applied here?

Let's take one specific hypothesis generated by this framework. What specific relationship would we want to examine empirically? What data and measures would we need?

Causal inference in education research

Causal inference about the effect of a policy, program, input, or risk factor is difficult!

- A **causal effect** is a change in some outcome (Y) that is the result of a change in some other (manipulable) factor (X).
- For simplicity, assume for now the factor X is a binary "treatment." Example: suppose we are interested in the effect of aspirin on headache pain, getting a vaccine on contracting COVID-19, or attending a charter school on academic achievement.
- Causal effects involve a **counterfactual** comparison between two different states of the world: e.g., Y whenever $X = 1$ versus Y whenever $X = 0$, assuming all else is held constant.

Potential outcomes and treatment effects

Example: take one individual (Maria)

$Y_{Maria}(1)$ = Maria's achievement if she attends a charter school.

$Y_{Maria}(0)$ = Maria's achievement if she attends a traditional school.

These two values are called **potential outcomes** since they represent what will happen to Maria in each of two scenarios.

$\tau_{Maria} = Y_{Maria}(1) - Y_{Maria}(0)$ is the **treatment effect** of attending a charter school for Maria.

Potential outcomes and treatment effects

If we knew τ for every individual in our population of interest, we could average them to get an **average treatment effect** (ATE) of attending a charter school:

$$ATE = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$$

Problem: we can't observe individuals like Maria in both conditions at the same point in time! If she attended a charter school, we don't know what her achievement would have been had she attended a traditional school and vice versa. This the **fundamental problem of causal inference**.

If Maria attended a charter school, $Y_{Maria}(0)$ is the **counterfactual**.

Potential outcomes and treatment effects

Suppose Maria attended a charter school. Michael attended a traditional public school. What if we estimated the effect of attending a charter school (for Maria) as:

$$Y_{Maria} - Y_{Michael}$$

What would be wrong with this?

Potential outcomes and treatment effects

The above can be written:

$$Y_{Maria} - Y_{Michael} = Y_{Maria}(1) - Y_{Michael}(0)$$

This is the same as $Y_{Maria}(1) - Y_{Maria}(0)$ only if $Y_{Michael}(0)$ is the same as $Y_{Maria}(0)$. There are lots of reasons to believe it wouldn't be. Such as?

In the real world, people generally **self-select** into "treatments"—they choose a path that makes the most sense for them.

Note: selection need not be *purposeful*. It may just be incidental (i.e., a result of other factors). We sometimes say treatment is **endogenous**.

Potential outcomes and treatment effects

Let's re-write the above:

$$Y_{Maria}(1) - Y_{Michael}(0) = Y_{Maria}(1) - Y_{Michael}(0) + \underbrace{Y_{Maria}(0) - Y_{Maria}(0)}$$

$$= \underbrace{Y_{Maria}(1) - Y_{Maria}(0)}_{\tau_{Maria}} + \underbrace{Y_{Maria}(0) - Y_{Michael}(0)}_{\text{selection bias}}$$

The difference between Maria and Michael's outcome is the treatment effect for Maria plus **selection bias**. Selection bias is how much Maria and Michael's outcomes differ in the "untreated" (traditional school) state.

Potential outcomes and treatment effects

If $Y_{Maria}(0) > Y_{Michael}(0)$, this means Maria would have had better outcomes than Michael even if she had not attended a charter school. We could call this "positive selection."

If $Y_{Maria}(0) < Y_{Michael}(0)$, this means Maria would have had *worse* outcomes than Michael in the traditional school setting. We could call this "negative selection."

The main challenge to causal research is dealing with selection bias. What are some other examples of selection bias you can think of?

Potential outcomes and treatment effects

The above example focused only on two individuals, Maria and Michael. Would it help to have large random samples of individuals who attended charter schools (like Maria) and of individuals who attended traditional public schools (like Michael)? E.g., could we instead calculate:

$$\bar{Y}_{charter} - \bar{Y}_{TPS}$$

Will this solve the problem? Why or why not?

Potential outcomes and treatment effects

The same issue arises when comparing group averages. Let $T = 1$ for treated cases (e.g., charter school) and $T = 0$ for untreated cases (traditional public school).

Let **E[ ]** be the **expectation** operator—i.e., the population mean.

$$E[Y|T = 1] - E[Y|T = 0]$$

is the difference in means for those with $T = 1$ and those with $T = 0$

Potential outcomes and treatment effects

We can write $E[Y|T=1] - E[Y|T=0]$ as:

$$
\begin{aligned}
&= E[Y(1)|T=1] - E[Y(0)|T=0] \\
&= E[Y(1)|T=1] - E[Y(0)|T=0] + E[Y(0)|T=1] - E[Y(0)|T=1] \\
&= \underbrace{E[Y(1)|T=1] - E[Y(0)|T=1]}_{ATT} + \underbrace{E[Y(0)|T=1] - E[Y(0)|T=0]}_{\text{selection bias}}
\end{aligned}
$$

ATT is the **average treatment effect for the treated**. Selection bias is systematic differences between the $T=1$ and $T=0$ groups.

Potential outcomes and treatment effects

Note that these are *population* means. Having large random samples of the $T=1$ and $T=0$ groups will not help! Even a comparison of population means would be fraught.

Research design for causal inference

Research design for causal inference is all about eliminating selection bias. This involves a search for "good counterfactuals" or techniques for "controlling for" factors that make the treated and untreated cases different.

These other factors—sometimes called **confounders**—are "alternative explanations" for an observed association between a "treatment" and an outcome. They are the reason we say correlation $\neq$ causation!

Randomized controlled trials

In a **randomized controlled trial (RCT)**, assignment to the treatment group is randomized by the researcher. What does this accomplish?

$$\underbrace{E[Y(1)|T=1] - E[Y(0)|T=1]}_{ATT} + \underbrace{E[Y(0)|T=1] - E[Y(0)|T=0]}_{\text{selection bias}}$$

Selection bias should be zero! There is no reason to expect $Y(0)$ to differ on average between the $T=1$ and $T=0$ groups. This is the case for both observable *and unobservable* characteristics.

This is why the RCT is often considered the "gold standard" of social science (and education) research.

Randomized controlled trials

Table 3.2 Examples of Random Assignment Experiments in Education

Experiment	Basic Strategy
Project Star	Random assignment of students in Tennessee to small and large classes in grades K–3
Perry preschool	Random assignment of poor children to an intensive preschool program
Computer instruction: Fast ForWord	Evaluation of program designed to improve language and reading skills
H&R Block financial aid information	Random assignment of low-income students and parents to receive information about financial aid and assistance with financial aid forms
Private school tuition vouchers	Experiments in Milwaukee and New York City that provide vouchers to low-income youth to attend private schools
Expanding College Opportunities Project	Provided information about college application strategies, college quality, and application fee vouchers to a randomly selected set of low-income, high-achieving high school students

Source: Lovenheim & Turner, *Economics of Education*

Randomized controlled trials

RCTs are increasingly common, but they are not always feasible (or ethical). They also have their own challenges and limitations.

- They are expensive and may be complicated to carry out.
- They are time consuming; it may be a long time before results are known.
- Subjects do not always comply with random assignment; there may be "crossovers."
- Subjects may behave differently in an experiment (e.g., "Hawthorne effects").

Randomized controlled trials

RCTs have high internal validity but often have low external validity.

- **Internal validity** means the research design adequately addresses selection bias and can be interpreted causally.
- **External validity** means the research design's findings can be readily generalized to other settings.

We must often work with **non-experimental** or **observational** data. However–even with non-experimental settings—it is often useful to think about what the "ideal" experiment would be!

Quasi-experiments

Quasi-experiments are research designs in which assignment to treatment is not done by the researcher, but by some outside factor that (plausibly) mimics a RCT. For example:

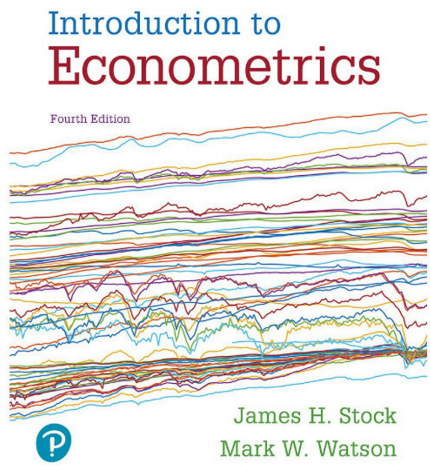
- An unanticipated "shock" assigns some units to treatment but not others (e.g., a policy change, natural disaster, or recession)
- Assignment to treatment is based on idiosyncratic rules (e.g., birthdate, geographic boundary, test score or income threshold)

These are often called **natural experiments** and are a

Regression

Multiple regression is a powerful tool used with non-experimental data to estimate the effect of some variable X on some outcome Y, *holding constant* other factors.

While regression alone often does not solve the problem of selection bias, it is the basic building block of other designs that do.



Regression

This course will develop your skills in this basic building block and introduce you to some of the most common designs for causal inference.

Fundamentals

- Lecture 2: Review of descriptive and inferential statistics
- Lecture 3: Multiple regression fundamentals
- Lecture 4: Statistical power
- Lecture 5: Nonlinear models and limited dependent variables

Applications

- Lecture 6: Experimental and quasi-experimental methods
- Lecture 7: Regression discontinuity designs
- Lecture 8: Time series and interrupted time series designs
- Lecture 9: Panel data methods