6. Experimental and Quasi-Experimental Methods

LPO.7870: Research Design and Data Analysis II

Sean P. Corcoran

Last time

Limited dependent variables: when Y is not a continuous variable, but takes on a limited range or set of values. We looked specifically at approaches for binary outcomes (0-1):

- Linear probability model (LPM) - estimated using OLS and easy to interpret.
- Logit and probit models - estimated using maximum likelihood method. Ensures predicted probabilities remain between 0 and 1, but are a bit harder to interpret.

Tonight

Randomized controlled trials and a preview of quasi-experimental designs.

- A strong design is one that allows us to make **causal inferences** about the effect of a policy, program, practice, or input on an outcome of interest.
- We will look at the advantages and disadvantages of different research designs for causal inference.

Tonight

Theme: randomized controlled trials (RCTs) are the "gold standard" but are often infeasible. So why study them?

- They are, in fact, increasingly used in education and social science. When done well, they are highly influential.
- The ideal RCT provides a conceptual benchmark against which to judge other designs.
- Natural and quasi-experiments provide treatments that are "as if" randomly assigned.

Tonight's sample datasets

We will refer to one dataset tonight (on Github):

1. `star.dta`: select variables from the Tennessee STAR class size experiment.

Causal effects

Causal effects

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 financial aid on college enrollment, 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 Maria's outcomes would be in each of two scenarios.

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

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

Average 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**.

Estimating average treatment effects

What if we had large samples of individuals (like Maria) who attended a charter school and others (like Michael) who attended a traditional public school? We could calculate:

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

Or (equivalently) estimate β_1 from the regression:

$$Y = \beta_0 + \beta_1 charter + u$$

where $charter = 1$ if the student attended a charter and $charter = 0$ otherwise.

Estimating average treatment effects

This would estimate the difference in population means:

$$\beta_1 = E[Y|charter = 1] - E[Y|charter = 0]$$

Does this represent the average treatment effect for Maria-types? For Michael-types? For anyone? Why or why not?

Note: recall $E[\ ]$ refers to a population mean.

Estimating average treatment effects

Problem:

$\bar{Y}_{charter}$ estimates $E[Y(1)]$ for those who chose to go to a charter.

$\bar{Y}_{TPS}$ estimates $E[Y(0)]$ for those who chose to go to a TPS.

Is the latter a good counterfactual for the former? Only if students enrolled in charter and TPS have the same (mean) potential outcomes $Y(1)$, $Y(0)$.

This seems unlikely if families select schools they believe are best for them! It could be that $\tau_{Maria} > \tau_{Michael}$

- Recall interaction effects: an intervention may have larger effects for some than for others.

Estimating average treatment effects

Using terminology from earlier lectures, the regression below likely suffers from **omitted variables bias**:

$$Y = \beta_0 + \beta_1 charter + u$$

charter status is likely correlated with students' potential outcomes. Kids who attend charter schools may have different potential outcomes, on average, than kids who attend TPS. This could be for many reasons: prior achievement, family income, motivation, etc. Effectively, charter and TPS students are different populations, in observable and unobservable ways.

It would help if students attending charter and TPS were random draws from the same population!

Randomized controlled trials

Randomized controlled trials

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

- The population average potential outcome $E[Y(1)]$ should be the same for the treatment and control group.
- The population average potential outcome $E[Y(0)]$ should be the same for the treatment and control group.
- Treatment status is *independent* of potential outcomes.

The difference $E[Y(1)] - E[Y(0)]$ is the ATE for your population! There is no selection / omitted variables bias!

Baseline equivalence

Randomization creates **baseline equivalence**: since the treatment and control group are drawn from the same population, they should have the same mean potential outcomes.

One of the first things researchers do in an RCT is check for baseline equivalence. Does it appear the two groups are on average the same (using what we can observe)?

RCTs: a video introduction

The following is an excellent introduction to RCTs from UNICEF (about 9 minutes long):

`https://www.youtube.com/watch?v=_Mr0s1XpsRY`

Estimating average treatment effects in a RCT

Randomization totally changes the interpretation of the simple regression:

$$Y = \beta_0 + \beta_1 charter + u$$

If assignment to a charter school was randomized, OVB is no longer a concern. $\hat{\beta}_1$ estimates the ATE for this population!

What was once a bad research design is now an excellent one!

Estimating average treatment effects in a RCT

When using regression to estimate the ATE, we can also include additional covariates:

$$Y = \beta_0 + \beta_1 charter + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + u$$

- Including covariates should not have a big effect on your estimate of β_1, as they should not be strongly correlated with *charter* (due to randomization).
- Benefits: reduces standard errors, since the additional X's explain some of the residual variation in Y.

Estimating average treatment effects in a RCT

You can include an interaction term if you think the treatment effect may be different for a certain subgroup:

$$Y = \beta_0 + \beta_1 charter + \beta_2 X_2 * charter + u$$

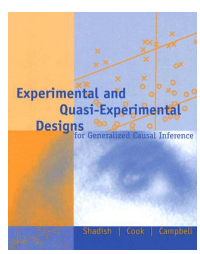
When randomization depends on covariates (e.g., a blocked or cluster randomized trial), need to include those covariates in the regression.

Internal and external validity

Evaluating RCTs

We often evaluate RCTs—and other research designs—through the lens of threats to validity.

- **Internal validity** means the research design adequately addresses selection bias and can be interpreted causally for the population being studied.
- **External validity** means the research design's findings can be generalized to other populations or settings.



Shadish, Cook, & Campbell (2002) is a classic reference on threats to validity.

Threats to internal validity (RCTs)

1. **Failed randomization**: randomization did not result in baseline equivalence.
2. **Partial compliance**: treatment and control subjects did not fully comply with their treatment assignment.
3. **Attrition**: subjects dropped out of the study after randomization.
4. **Experimental effects**: the conduct of the experiment itself had an effect on outcomes.
5. **Small sample sizes**: does not result in bias, but estimates are imprecise and large-sample assumptions inappropriate.

(1) Failed randomization

If assignment to treatment was not fully random but related to potential outcomes.

- Could happen by chance, or due to a poor randomization procedure (e.g., based on last name or time of day).
- If treatment assignment (e.g., *charter*) is related to potential outcomes, we are back to a situation of OVB.

Researchers will usually conduct **balance tests**. For example, regress X (treatment) on baseline characteristics and then do an F-test for joint significance of these characteristics.

(2) Partial compliance

When study subjects do not fully comply with their treatment assignment (treatment or control).

- "No shows" = assigned to treatment but did not receive it.
  "Crossovers" = assigned to control but received treatment anyway.
- When compliance is related to potential outcomes (non-random), we are back to a situation of OVB.
- Can regress Y on treatment *assignment* rather than treatment *receipt*. The result is interpreted as an **intent-to-treat** (ITT) effect—the effect of *offering* the treatment.

(2) Partial compliance

Researchers should keep good records of compliance with treatment assignment, whenever possible.

In some circumstances it is possible to estimate the effect of **treatment-on-the-treated** (TOT) when there is partial compliance.

(3) Attrition

When study subjects drop out of the study at some point after randomization, but before the outcome of interest is measured.

- When attrition is related to potential outcomes (non-random), we are back to a situation of OVB. (Example: job training program)
- Can't estimate "intent-to-treat" here, since outcome data is missing for those who dropped out.
- Even random attrition is bad if it reduces sample sizes by a lot.

Researchers should track attrition and conduct tests for selective attrition (e.g., LPM for attrition using baseline characteristics as predictors).

(4) Experimental effects

Sometimes the conduct of an experiment itself has effects on outcomes.

- **Hawthorne effect**: when subjects behave differently because they know they are in an experiment.
- **Disappointment effects**: when subjects not selected for treatment are "disappointed" and have worse outcomes than otherwise.
- **Compensatory effects**: for example, when subjects not selected for treatment are given other assistance in compensation.
- **Double blind** studies can help with this, but this is hard when the RCT involves an intervention.

Threats to external validity (RCTs)

1. **Nonrepresentative sample**: the population studied differs systematically from the population of interest.
2. **Nonrepresentative program or policy**: the treatment studied is dissimilar from the treatment of interest (e.g., small scale versus scaled-up study).
3. **General equilibrium effects**: when scaling up a program or policy changes the underlying operating environment (e.g., job training or class size reduction).

RCT examples

Tennessee STAR

- Well-known large scale RCT in the late 1980s (4 years, $12m)
- Students randomized to regular class, small class, or regular class with teacher's aide.
- Randomization was within-school, at kindergarten. Students were expected to remain in their assigned treatment through 3rd grade.

Glazerman et al., 2006

Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes

Steven Glazerman
Daniel Mayer
Paul Decker

Abstract

This paper reports on a randomized experiment to study the impact of an alternative teacher preparation program, Teach for America (TFA), on student achievement and other outcomes. We found that TFA teachers had a positive impact on math achievement and no impact on reading achievement. The size of the impact on math scores was about 15 percent of a standard deviation, equivalent to about one month of instruction. The general conclusions did not differ substantially for subgroups of teachers, including novice teachers, or for subgroups of students. We found no impacts on other student outcomes such as attendance, promotion, or disciplinary incidents, but TFA teachers were more likely to report problems with student behavior than were their peers. The findings contradict claims that such programs allowing teachers to bypass the traditional route to the classroom harm students. © *2006 by the Association for Public Policy Analysis and Management*

Glazerman et al., 2006

- Impact of TFA teachers on achievement and other outcomes.
- Block random assignment: randomization of students to teachers *within school and grade* (grades 1-5) in 6 regions.
- Estimated main effects but also compared TFA and subsets of non-TFA teachers (e.g., novice and veteran).
- Careful attention to "crossovers" using roster checks.
- Careful discussion of attrition (student and teacher).
- Careful discussion of what the "treatment" is (recruitment + training effects) and the counterfactual condition.

Glazerman et al., 2006

Table 1. Study sample.

Region	Number of Schools	Number of Comparison Blocks	Number of Classes Taught by: TFA Teacher	Novice Control Teacher	Veteran Control Teacher	Number of Students Taught by:[a] TFA Teacher	Novice Control Teacher	Veteran Control Teacher
Baltimore	3	6	7	1	8	137	18	147
Chicago	3	7	7	2	5	139	42	105
Houston	3	7	7	3	7	126	56	114
Compton	2	6	6	6	4	97	111	72
Mississippi Delta	3	6	12	2	10	201	31	146
New Orleans	3	5	5	1	7	85	21	117
Total	17	37	44	15	41	785	279	701

Source: TFA Evaluation Project tracking system.
[a] Includes students in the research sample who completed the spring achievement test.

Table 2. Mobility rates of control and TFA students (percentages).

Mobility Type	Control Students	TFA Students	Difference[a]	Total
Stayer	87.8	86.0	–1.8	87.3
Crossover[b]	3.7	4.3	0.7	4.0
Mover within district	5.2	5.6	0.4	5.4
Mover out of district	2.3	2.9	0.6	2.5
Mover other/unknown	1.3	1.3	0.0	1.3
Sample size	1,094	875		1,969

Source: Student tracking system.
[a] Chi-squared test fails to reject the null hypothesis of equal distributions ($p = 0.898$); that is, the differences between

Glazerman et al., 2006

Table 4. Impacts on test scores, teacher subgroups (NCEs).

	Mathematics		Reading		Sample Size		
Subgroup Comparison	Impact Estimate	Standard Error	Impact Estimate	Standard Error	Blocks	Classes	Students
Full sample	2.43***	(0.73)	0.56	(0.62)	37	100	1,715
Experience							
Novice TFAs versus novice controls	4.13***	(1.24)	1.06	(1.19)	11	25	432
All TFAs versus veteran controls	2.71***	(0.97)	0.45	(0.70)	31	79	1,370
First-year TFAs versus all controls	1.81	(1.70)	-0.90	(0.99)	12	32	526
Second-year and veteran TFAs versus all controls[b]	2.55***	(0.74)	1.09	(0.71)	29	77	1,320
All TFAs versus recently hired controls	2.46***	(0.79)	1.12	(0.75)	24	57	994
Certification							
All TFAs versus certified controls	1.92*	(0.94)	0.01	(0.75)	27	70	1,216
All TFAs versus uncertified controls	3.12**	(1.11)	1.01	(0.95)	14	36	620
Uncertified TFAs versus all controls	3.21***	(0.98)	-0.34	(0.92)	19	58	973

Source: Scores from the Iowa Test of Basic Skills, administered by Mathematica Policy Research, Inc.
Note: All test scores are expressed in NCEs, whose average score nationally is 50 and standard deviation is 21.06.
[a] Control group means and impacts are regression-adjusted. The regression model controls for base-

Downsides to RCTs

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.
- Recruiting participation may be difficult.
- They are time consuming; it may be a long time before results are known.
- Many potential threats to validity (e.g., crossovers, attrition, experimental effects)

Quasi-experimental designs

Quasi-experiments

In a **quasi-experiment**, units are assigned to treatment and control conditions not by the researcher but by variation in individual conditions. For example:

- Policy changes or program implementation
- Differences in legal institutions
- Location, boundaries
- Birth dates
- Weather, natural disasters

These are often called **natural experiments**.

Quasi-experiments

Types of quasi-experimental designs that exploit "natural" variation:

- Regression discontinuity
- Difference-in-differences
- Within-panel designs
- Instrumental variables

Matching designs that use a "constructed" control group are sometimes called quasi-experiments, although I would not put them in this category.

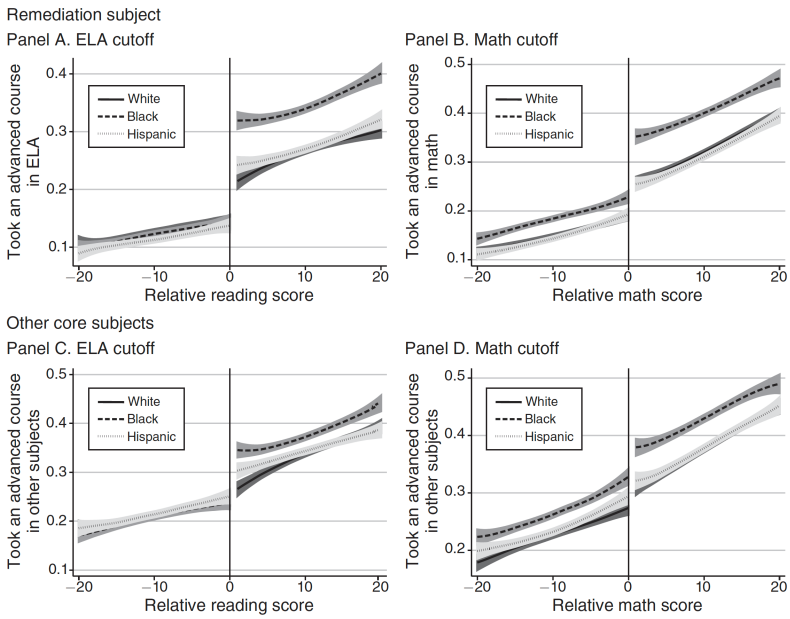
Regression discontinuity

RD can be used when a **precise** rule based on a **continuous** characteristic determines treatment assignment. Examples:

- **Test scores**: can determine school admission, financial aid, summer school, remediation, graduation.
- **Income or poverty score**: eligibility for income assistance or benefits, community eligibility for a means-tested anti-poverty program.
- **Date**: age cutoff for retirement benefits, health insurance, school enrollment (PK or KG).
- **Elections**: fraction that voted for a particular candidate or ballot measure (e.g., school bond)

Regression discontinuity

Figlio & Ozek (2024) looked at a policy in Florida that placed middle school students into remedial classes based on their test score.



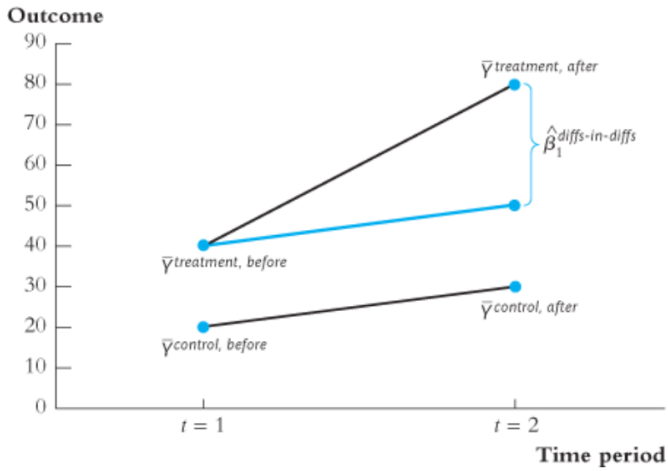
Difference-in-differences

Difference-in-differences can be used when you have multiple observations for two (or more) groups over time. Some of these groups are "treated" at one point while others are not.

Under certain assumptions, the change over time for the untreated group can serve as a counterfactual for the change over time for the treated group.

Difference-in-differences

Figure 13.1 The Differences-in-Differences Estimator



The posttreatment difference between the treatment and control groups is $80 - 30 = 50$, but this overstates the treatment effect because before the treatment $\overline{Y}$ was higher for the treatment group than the control group by $40 - 20 = 20$. The differences-in-differences estimator is the difference between the final and initial gaps, so $\hat{\beta}_1^{diffs-in-diffs} = (80 - 30) - (40 - 20) = 50 - 20 = 30.$. Equivalently, the differences-in-differences estimator is the average change for the treatment group minus the average change for the control group; that is, $\hat{\beta}_1^{diffs-in-diffs} = \Delta\overline{Y}^{treatment} - \Delta\overline{Y}^{control} = (80 - 40) - (30 - 20) = 30.$

Panel methods

When panel data are available, can sometimes use changes to identify causal effects. In this case, treated units are counterfactuals for themselves.

Example: school "switchers"

Instrumental variables

When an external ("exogenous") force assigns treatments to units, can use that variable as an "instrument" for treatment. In this case, you are using variation that is only due to the exogenous force.

Examples: rainfall, temperature, proximity to a hospital or college

Quasi-experiments: internal and external validity

Each of these designs should be assessed for their internal and external validity.

- All of the above have strong claims to internal validity, under certain assumptions (more on this later).
- They vary in their claims to external validity. On the one hand, the study sample may be representative of a large population of interest. On the other hand, the effects estimated may be localized to a specific subset of that population (e.g., RD).

Next time

Regression discontinuity designs

- See *Mastering Metrics* chapter 4 and *Mixtape* chapter 5. Also Bloom (2012).
- Sample studies, including Dee and Wyckoff (2015).