An Overview of Causal Inference

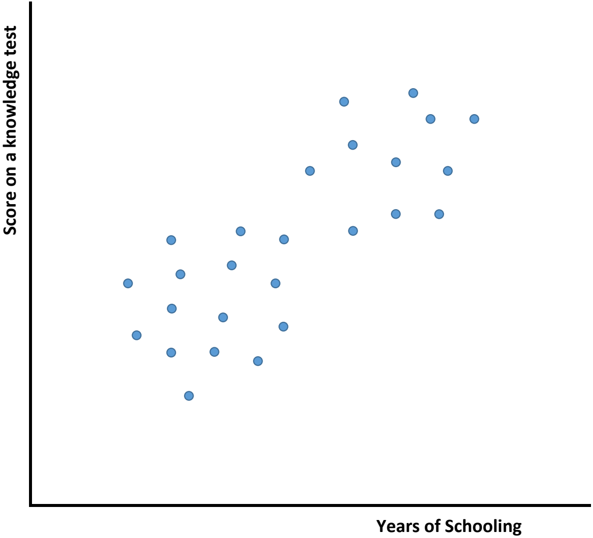
Da Gong

Adopted from Dr.Esterling's slides

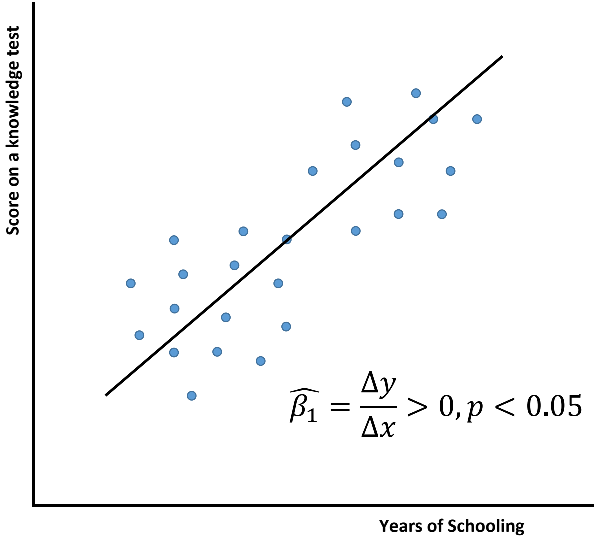
UC Riverside – GradQuant

January 9, 2024

Q: Should we promote college education more broadly?

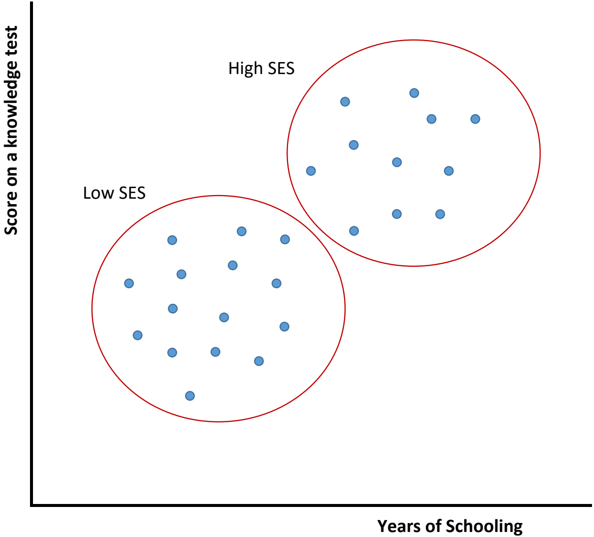


Should we promote college education more broadly?

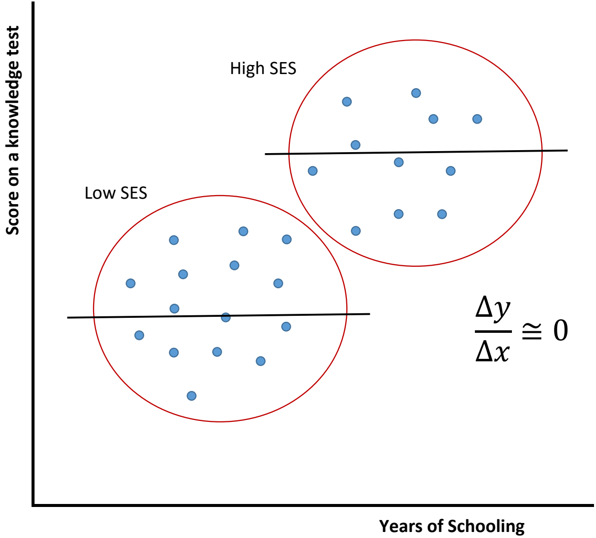


$$\widehat{\beta_1} = \frac{\Delta y}{\Delta x} > 0, p < 0.05$$

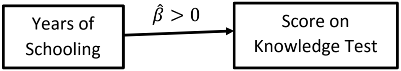
Should we promote college education more broadly?



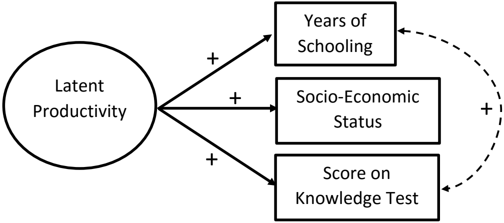
Should we promote college education more broadly?



Should we promote college education more broadly?



Panel A: What you thought you were testing....

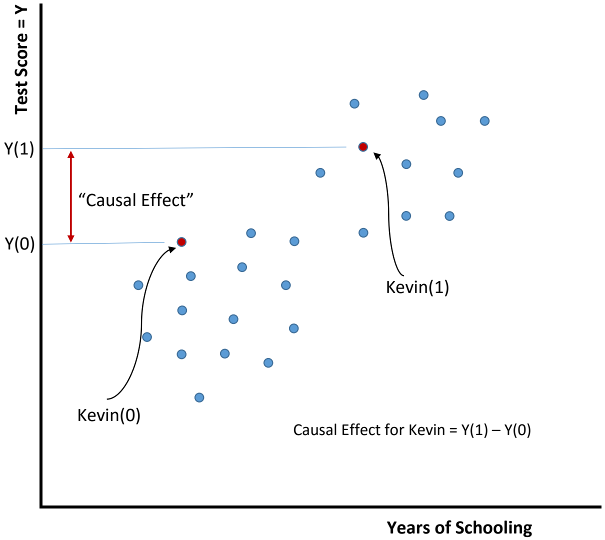


Panel B: ...What was actually true:

People select their own level of schooling.

Causal Analysis

- Question: How can we know if an intervention *caused* an outcome?
- Need to answer two questions:
    - What happened to an individual when exposed to the intervention? and
    - What would have happened had the participant not been exposed to the intervention?



Causal Effect for Kevin = Y(1) – Y(0)

Fundamental problem of causal inference

Causal analysis depends on “counterfactuals”

- Not possible to know both what happened to a participant when exposed to the intervention, and also what would have happened had the participant not been exposed to the intervention.
- The “counterfactual” that you need to see to identify causal effects can never be observed.
- Half of the data you need to do a causal analysis is inherently missing.

Solution to the problem of causal inference

Design-based approach

- Use a research design that creates comparisons among similar people, such as
  - Researcher randomly assigns to “treatment” and “control”
  - NGO uses lottery to allocate spots in program
  - Government agency uses a cutoff rule for eligibility
  - One state exposed to a policy but another nearby state is not
- Use outcomes from a similar group to “fill in” the missing data for the other group
- Requirements:
  - The intervention is manipulated, not fixed
  - People’s exposure to the intervention is not entirely of their own choosing

Defining causality at the individual level

Table: Before the experiment is run

Unit	Z	D	Y(Z=0)	Y(Z=1)
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

$$\text{Causal Effect} = Y_i(Z=1) - Y_i(Z=0)$$

What an omniscient observer would see..

Table: Full Compliance, $Z \equiv D$

Unit	Z	D	Y(Z=0)	Y(Z=1)
1	1	1	1	4
2	1	1	2	2
3	1	1	1	1
4	1	1	2	4
5	0	0	3	2
6	0	0	2	2
7	0	0	1	4
8	0	0	1	3
9	0	0	1	2
10	0	0	2	3

Question: What are the unit level treatment effects? Sample treatment effect?

What we actually see..

Table: Full Compliance, $Z \equiv D$

Unit	Z	D	Y(Z=0)	Y(Z=1)
1	1	1	?	4
2	1	1	?	2
3	1	1	?	1
4	1	1	?	4
5	0	0	3	?
6	0	0	2	?
7	0	0	1	?
8	0	0	1	?
9	0	0	1	?
10	0	0	2	?

$\Rightarrow$ "Fundamental problem of causal inference"

Some notation...

For simplicity, I will write $Y_i(1)$ instead of $Y_i(Z = 1)$ for treatment group potential outcomes and $Y_i(0)$ instead of $Y_i(Z = 0)$ for control group potential outcomes. And note, $Y_i(0) \equiv Y_i^0 \equiv Y_{i0}$

Causal Analysis

Design-based approach

- Question: How to fill in missing data for causal inference?
  - We want to find $\tau_i = Y_i(1) - Y_i(0)$
  - BUT: We never can observe both $Y_i(1)$ and $Y_i(0)$
- Answer: Find comparable people, expose only some to the intervention, and compare averages across groups

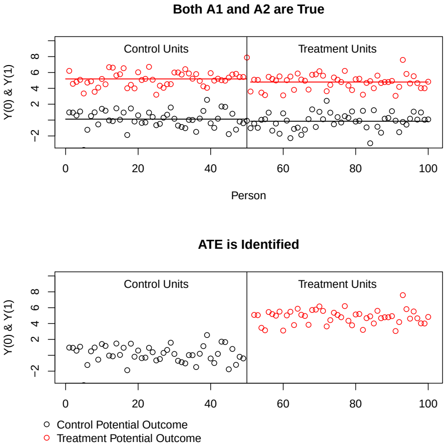


Figure: Both groups can be counterfactuals to each other

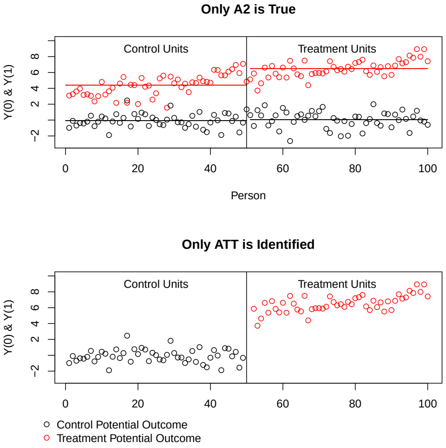


Figure: Controls can be counterfactuals to Treatments

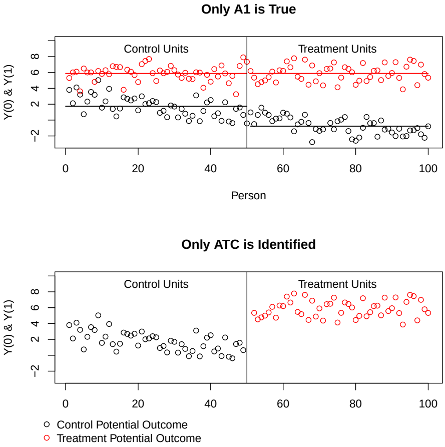


Figure: Treatments can be counterfactuals to Controls

Design #1: Randomized Control Trials

One important strategy to create comparable groups: **randomly assign** participants to experimental conditions.

- Under randomization, all individuals have the same probability of being assigned to treatment or control.
- Selection into the treatment is unrelated to all observed and unobserved characteristics of individuals.
- With some mild assumptions, participants' outcomes in each treatment arm can be counterfactuals to the other treatment arm

RCM requires 3 core assumptions to identify causal effects

Assumptions

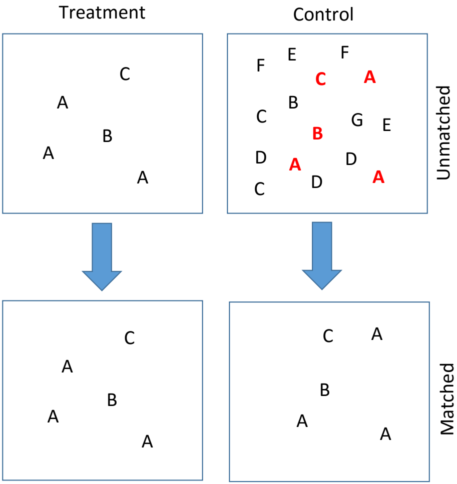
- Randomization
- Stable Unit Treatment Value (SUTVA)
- Exclusion Restriction

Designs for Observational Data

Sometimes we do not have strong design elements that we can use to justify assumptions to identify causal effects, for example if the participants select themselves into treatment and control. In this situation, we can possibly use:

- Matching/Stratification
- Difference-in-Difference
- Regression Discontinuity Design

Design #2: Matching



Matching creates balance between treatment and control groups

Matching as a research design, cont.

- If you match control units to the treatment units, then you identify the ATT
- You also can match treatment units to the control units to identify the ATC
- And you can take a weighted average of the ATT and ATC to estimate the ATE, where the weight is the proportion p of units in the treatment group

Matching requires the assumption of strong ignorability

If we are willing to assume *selection on observables*, we can make valid comparisons between treatment and control groups within levels of the stratifying variables.

How to do matching in practice?

Question: What do you do if, as is the general case, you have many stratifying variables, each with many values?

- If you have four stratifying variables, each with five values, then you have $5^4 = 625$ strata. Even if you have a large sample, most of the strata will have treatment observations but no control observations, and vice versa.
- In this situation, you can't make direct comparisons between treatment and control observations within strata.
- What do you do?

Propensity score

You can construct a *propensity score* for each person in your dataset, which is the probability that each person would select into the treatment, conditional on their covariate values.

- The propensity score for person i is $p(D = 1|S_1, S_2, \ldots S_k)$
- You can estimate the propensity score using a procedure called "logit."
- Assuming selection on observables, the propensity score summarizes all of the information contained in the stratifying variables that is relevent for selection in a single number (!).

Matching mechanics

- In general, no two units will have exactly equal propensity scores
- How do you know that two units are close enough for a match?
    - Small difference in propensity scores
    - Ratio of propensity scores that is near 1
    - Mahalanobis distance is small
- There are many ways to choose matching units
    - Exact matching
    - Nearest neighbor matching
    - Kernel matching
    - Optimal matching (e.g., GenMatch)

Design #3: Difference-in-Difference

Often we want to evaluate the effect of a policy that is implemented within a specific jurisdiction

- What is the effect of a seat belt law in Connecticut on highway fatalities?
- How can we know?

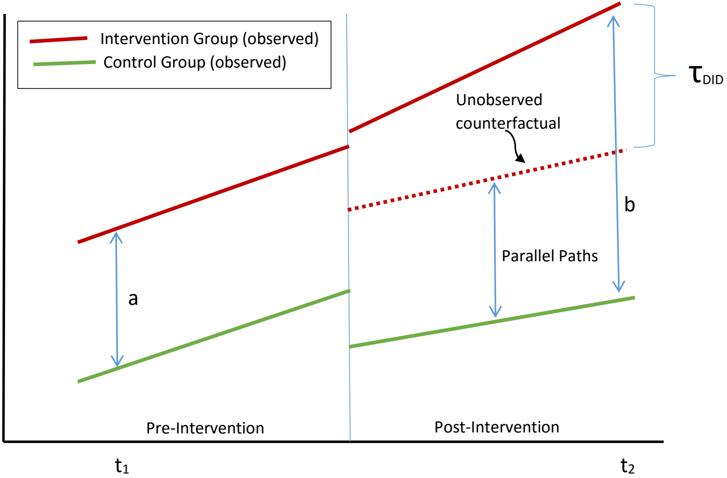
Motivation for Difference-in-Difference

Two naïve approaches to evaluate the policy

- “Within-state” comparison
  - Idea: compare Connecticut highway fatalities before and after the policy was adopted
  - Problem: Many things can change at the time of the intervention
- “Between-state” comparison
  - Idea: compare Connecticut fatalities after the law passed to Rhode Island fatalities at the same time
  - Problem: Many things are different between Connecticut and Rhode Island

The problem is that states are not randomly assigned to policies; they self-select policies

Difference-in-Difference Design



$$\tau_{DID} = b - a$$

Identification Assumption

Parallel Path Assumption

The potential outcomes for the treated and control units would have followed parallel paths in the absence of the intervention. That is, traffic fatalities in Connecticut would have followed a similar trend in Connecticut as they did in Rhode Island, if only Connecticut had not enacted the seat belt law.

Why the assumption works

The assumption allows you to net out the effect of the state's selection into the policy. E.g., Connecticut may have passed the law because it has more highway fatalities. And it nets out changes over time.

Design #4: Regression Discontinuity Design

Sometimes an agency will establish eligibility based on a cut off

- A politician wins an election with 50% + 1 votes
- Can't buy liquor before 21st birthday
- College loans available before a date but not after

RDD compares those just below and just above the cut off to test the effect of the program.

Two versions of RDD

There are two types of regression discontinuity designs:

- **Sharp RD** is when the rule fully determines assignment
- **Fuzzy RD** is when the rule only partly determines assignment, for example when the agency uses discretion in when to apply the rule, or when recipients are able to find loopholes to evade the rule

First we will describe the sharp design, and then we will discuss how to implement the fuzzy design

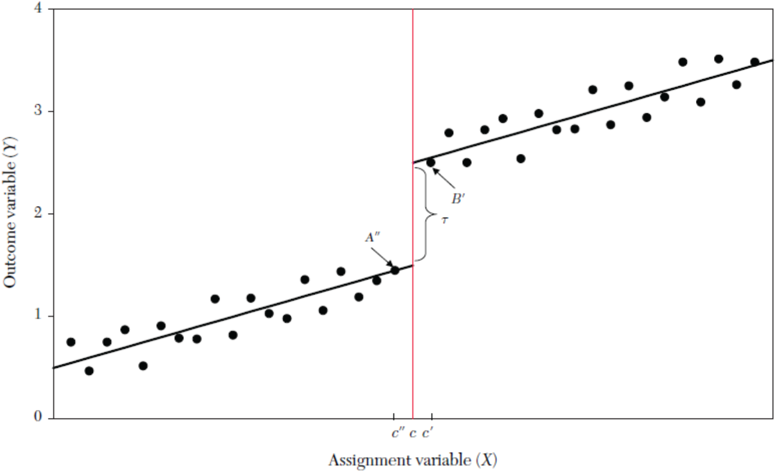
Defining the Sharp RDD

In an RDD, exposure to the program (D) is determined by a variable that determines eligibility (X).

- X is known as the *assignment variable* and is also sometimes called the *forcing variable*. These two expressions mean the same thing.
- Let's say that eligibility requires $X_i > x_0$. Exposure is a function of the assignment variable:

$$D_i = \begin{cases} 0, & \text{if} X_i \leq x_0 \\ 1, & \text{if} X_i > x_0 \end{cases} \tag{1}$$

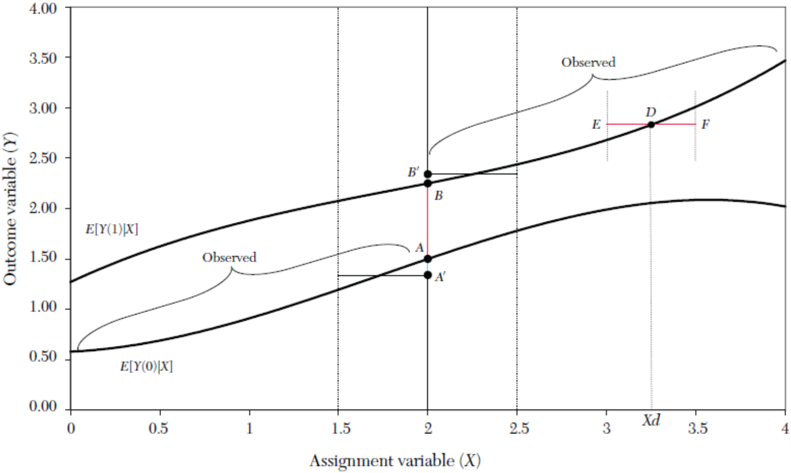
Regression Discontinuity Design



RDD Identification Assumption

- We can write $\tau_{RDD} = Y_i(D = 1|X = x_0) - Y_i(D = 0|X = x_0)$
- The treatment effect estimand compares $Y_i = Y_i(1)$ values for those just to the right of the cutoff and $Y_i = Y_i(0)$ for those just to the left
- For this comparison to be valid, the functions $E[Y_i(1)|X_i]$ and $E[Y_i(0)|X_i]$ must be *continuous* in X at x_0
- This works, for example, if each person had a small "random component" to their X value that placed them just above or just below the cutoff, or if people are not aware of the value of the cutoff

Regression Discontinuity Design



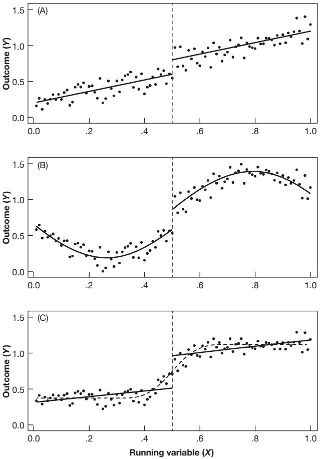
Sharp RDD Estimation Issues

Some things to note:

- τ_{RDD} identifies the local average treatment effect (LATE) only at x_0
- Rarely do we have enough observations right at x_0 to have sufficient power. Instead, we specify a *bandwidth* around x_0 and include the observations within that interval
  - The wider the bandwidth, the more power but the less internal validity
  - We usually specify a *kernel function* that gives more weight to observations closer to x_0
- Be careful not to confuse a nonlinear change at x_0 with a shift at x_0

RDD Issues

FIGURE 4.3
RD in action, three ways



Notes: Panel A shows RD with a linear model for $E[Y_i|X_i]$; panel B adds some curvature. Panel C shows nonlinearity mistaken for a discontinuity. The vertical dashed line indicates a hypothetical RD cutoff.

Fuzzy RDD

- In some instances, the agency uses discretion in applying the cutoff rule, or participants have some ability to find loopholes in gaining eligibility
- In this case, often the cutoff changes the probability of treatment status and we can use the discontinuity as an instrument for treatment status
- This identifies the LATE of the effect of the program on compliers, that is, on those whose eligibility status actually changes at the cutoff

Summary: Potential Outcomes Framework

- Framework to identify causal effects from data by stating requirements for how to fill in the missing data
  - States assumptions that are *necessary for identification* for causal effects ("Rubin Causal Model")
  - "Identification" implies your research analysis has enough constraints for a unique solution
  - Identification is a purely conceptual problem; estimation is a separate question entirely ("estimand" versus "estimator")
- Benefits of working within the POF
  - Directs attention to research design
  - Simplifies analysis and improves credibility
  - Assumptions clarify how *hard* it is to make causal statements; why social science is so hard to do