EEP/IAS 118 - Introductory Applied Econometrics, Lecture 11

Gregory Lane

July 2017

Difference-in-Difference, Intro

Sometimes an RCT can't be conducted because:

- Impossible (i.e. central bank policy)
- Intervention already happened
- Unethical

Without a formal RCT, we have to assume the treatment was **not** randomly assigned:

- Receiving treatment was dependent on (unobserved) characteristics of the subjects
- If we can't control for these variables we will have **OVB**

What can we try in these cases?

- Use specific characteristics of the policy change to get identification
- E.g. Diff-in-diff, Regression Discontinuity

Difference-in-Difference, Intro

In Indonesia 60,000 schools were constructed in the 1970s. Some areas got many new school (High) and others got fewer schools (Low)

- **Question:** Did this intervention impact education and future labor market outcomes children?
- **Problem:** Construction occurred a long time ago, and building the schools was not random.
- **Idea (Duflo, 2001):** Children who were 12 at the time of the construction (1974) would not be affected. Children who were 6 benefitted fully from the school construction
    - We can use these two groups of children to estimate the effect of the school construction

Difference-in-Difference, Intro

We have data from 1995 on school and labor market outcomes

- The older cohort (12 year olds) are now 33
- The younger cohort (6 year olds) are now 27

We hypothesize that the construction increased education and then labor market outcomes

- Can we estimate this effect just by looking at the differences between these two groups in 1995?
- No, because other factors likely caused education to improve over those 6 years (e.g. income gains) - we can't attribute all gains to the school construction

We need to do something else instead

Difference-in-Difference, Intro

We will leverage the fact that the school construction happened more intensely (Treatment group) in some areas than others (Control group):

- The difference between the age cohorts in the **control** areas can be used to measure the improvement in eduction and labor markets that can be attributed to other factors
- The difference between the age cohorts in the **treatment** areas can be used to measure changes that can be attributed to other factors *and* the school construction combined
- Therefore, if we subtract these two differences from each other, the remainder will be only the effect of the school construction

Diff-in-Diff, Intro

Duflo, 2001 data:

	Years of education		
	Level of program in region of birth		Difference
	Low (C)	High (T)	(T - C)
Before (12-17 in 1974)	9.40	8.02	-1.38
After (2-6 in 1974)	9.76	8.49	-1.27
Before-After changes	0.36	0.47	0.11

Diff-in-Diff, Calculation

Assume we have two groups (T and C) and two time periods (1 and 2). The program was implemented for the treatment group in time period 2

1. Calc difference in the outcome variable Y in the control group across the two time periods:

$$\bar{Y}_{C1} - \bar{Y}_{C0} = \Delta\bar{Y}_C$$

2. Do the same for treatment:

$$\bar{Y}_{T1} - \bar{Y}_{T0} = \Delta\bar{Y}_T$$

3. The impact of the program is measured by the difference in the differences:

$$(\bar{Y}_{T1} - \bar{Y}_{T0}) - (\bar{Y}_{C1} - \bar{Y}_{C0}) = (\Delta\bar{Y}_T - \Delta\bar{Y}_C)$$

DIff-in-Diff, Regression

We can also find the impact of the program through regression:

$$Y_i = \beta_0 + \beta_1 Post_{it} + \beta_2 Treat_{it} + \beta_3 Post_{it} \times Treat_{it} + u_{it}$$

- $Post_{it}$ is a dummy that indicates time period two
- $Treat_{it}$ is a dummy that indicates being in the treatment group

	Pre	Post
Control	Not Treated	Not Treated
Treatment	Not Treated	Treated

This has the benefit of giving us standard errors for the $\hat{\beta}$ so we can run hypothesis tests

- **Question:** Interpret what each of the β above

Diff-in-Diff, Assumption

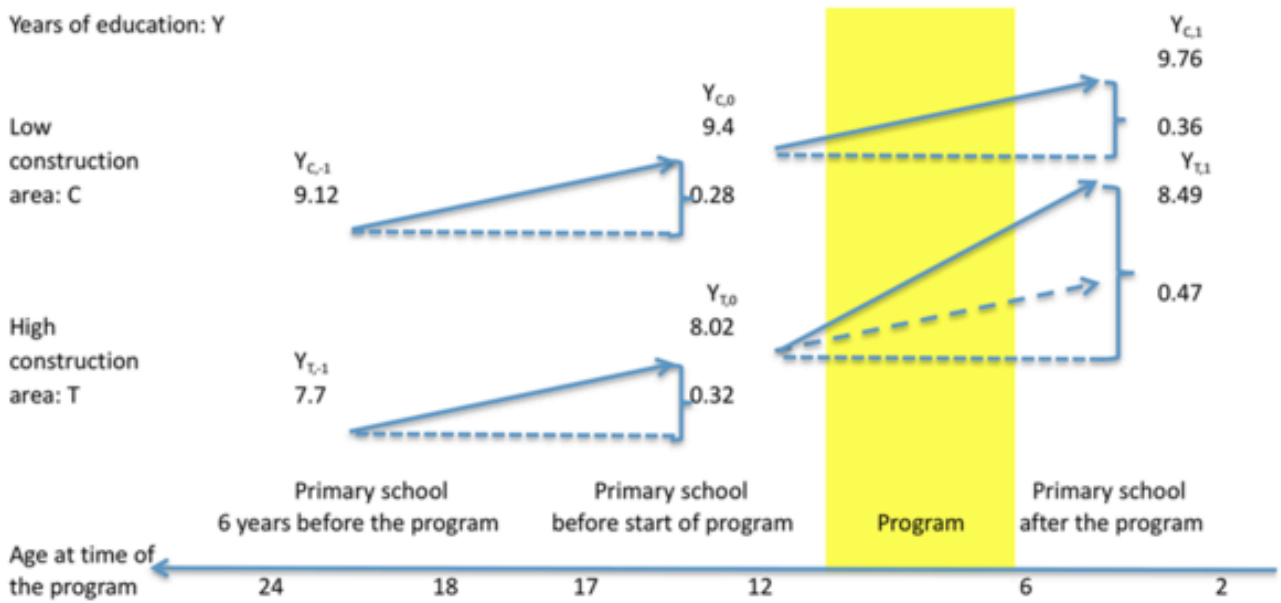
Key assumption for this to work:

- *The difference between before and after in the control group is a good counterfactual for the treatment group*

In other words, the change in outcomes for the control group is what we would have observed in the treatment group absent the policy

Diff-in-Diff, Testing Assumption

- Assumption is fundamentally untestable
- Best we can do is analyze pre-trends
- In order for the diff-in-diff to be valid, we want to see **parallel trends**:



Diff-in-Diff, Testing Assumption

To test for parallel trends, we need at least one more year of data before the intervention (*pre-pre-period*)

Then in a regression framework we can run the following estimation using data only from **BEFORE** treatment:

$$Y_i = \beta_0 + \beta_1 Pre_i + \beta_2 Treat_i + \beta_3 Pre_i \times Treat_i + u_i$$

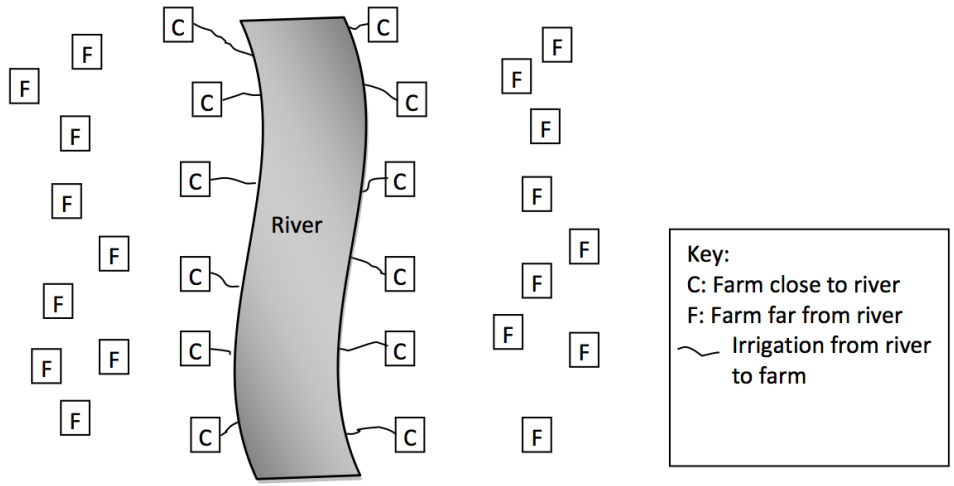
- Pre_i is a dummy that indicates being in the pre-period (as opposed to the pre-pre-period)
- $Treat_i$ is a dummy that indicates being in the treatment group

If parallel trends holds, we expect the coefficient on the interaction term to be statistically insignificant

Example: Irrigation

Want to evaluate the effect of irrigation on farm yields.

- Naive estimation would just compare yields for farms close enough to the river to get irrigation to those who were not:

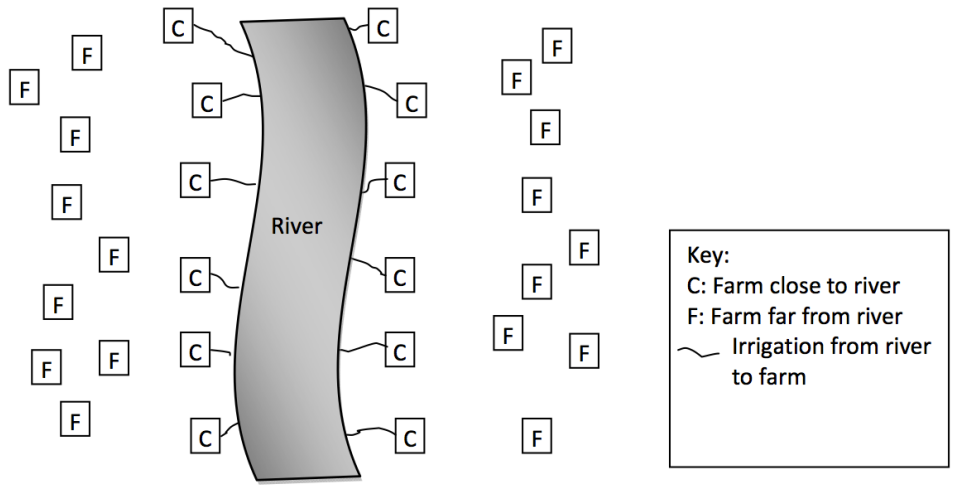


$$yield_i = \beta_0 + \beta_1 irrigation_i + u_i$$

- **What's the problem with this approach?**

Example: Irrigation

- Naive estimation would just compare yields for farms close enough to the river to get irrigation to those who were not:

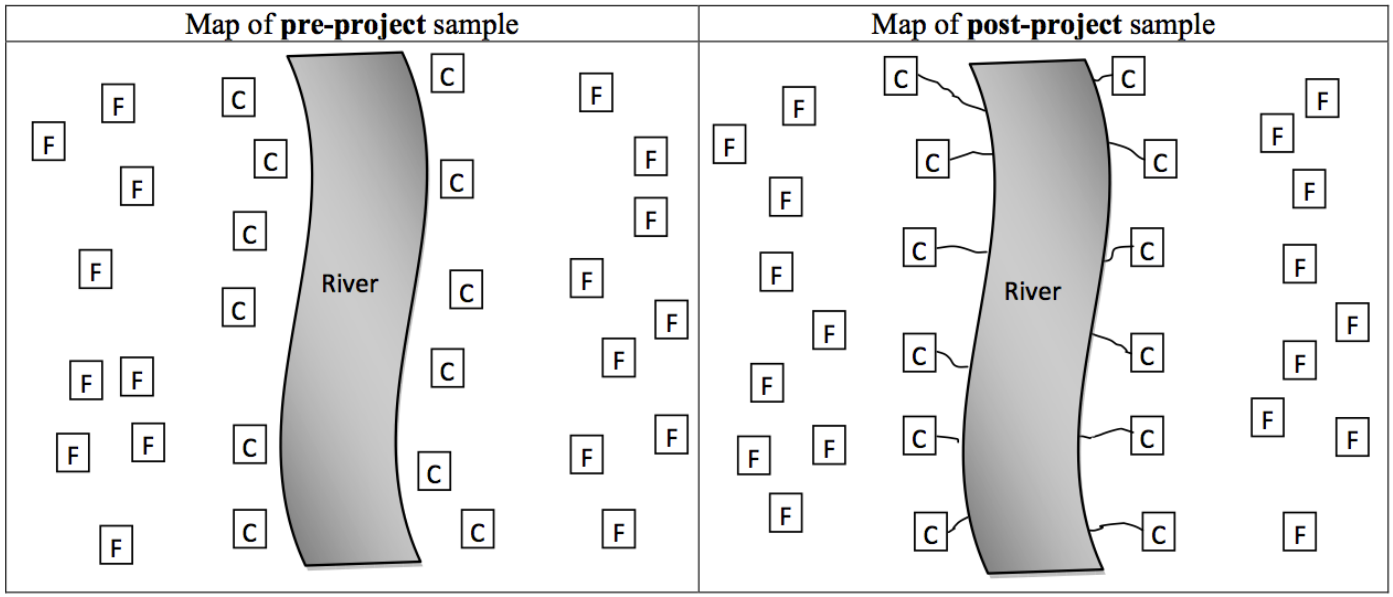


$$yield_i = \beta_0 + \beta_1 irrigation_i + u_i$$

- **Problem:** The farms that got irrigation are the farms that are close to the river! There are probably a lot of things that vary between C and F farms besides irrigation

Example: Irrigation

A better design would use two waves of data, one before the project was started and one after it was completed:



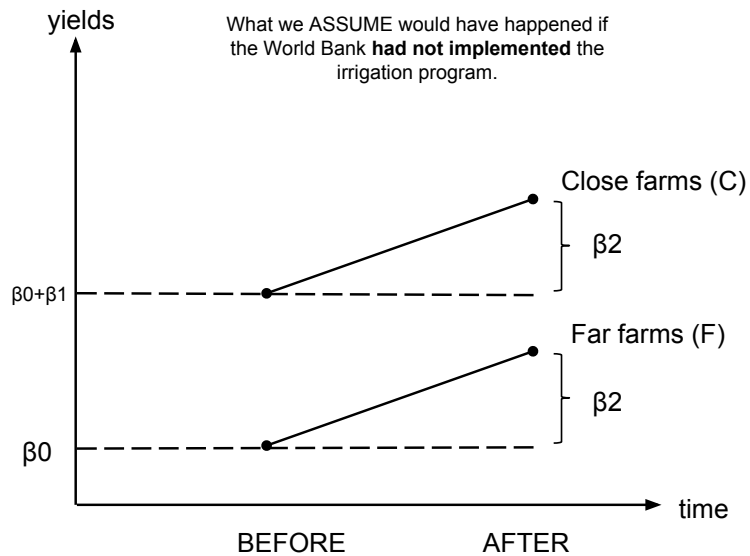
$$yield_i = \beta_0 + \beta_1 irrigation_i + \beta_2 post_i + \beta_3 (irrigation * post)_i + u_i$$

- This accounts for differences (some of which we can't observe) between the C and F farms, getting around the fact that the irrigation was not randomly assigned across farms.

Example: Irrigation

What is the identifying assumption for this estimation strategy?

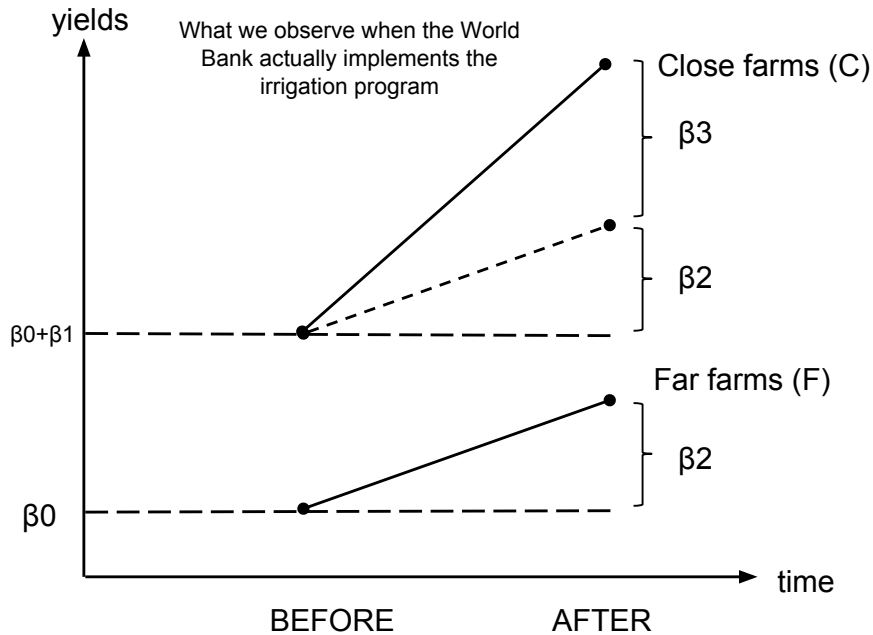
- The difference between before and after in the comparison group is a good counterfactual for the treatment group
- We can also draw a picture to understand the diff in diff assumptions and strategy:



$$yield_i = \beta_0 + \beta_1 irrigation_i + \beta_2 post_i + \beta_3 (irrigation * post)_i + u_i$$

Example: Irrigation

$$yield_i = \beta_0 + \beta_1 irrigation_i + \beta_2 post_i + \beta_3 (irrigation * post)_i + u_i$$



- The diff-in-diff strategy assumes that the entire difference in the slope of these two lines is due to the treatment (because we are assuming that the slopes *would* have been the same without the program).

Example: Irrigation

Exercise: Using this four data points for average yield, calculate what values for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, and $\hat{\beta}_3$ we would obtain from a diff-in-diff regression:

Yields in Kg/acre

	Pre-Period	Post-Period
Far (Not Irrigated)	40	60
Close (Irrigated)	70	80

$$yield_i = \beta_0 + \beta_1 irrigation_i + \beta_2 post_i + \beta_3 (irrigation * post)_i + u_i$$

Example: Irrigation

Yields in Kg/acre

	Pre-Period	Post-Period
Far (Not Irrigated)	40	60
Close (Irrigated)	70	80

$$yield_i = \beta_0 + \beta_1 irrigation_i + \beta_2 post_i + \beta_3 (irrigation * post)_i + u_i$$

- $\hat{\beta}_0 = 40$
- $\hat{\beta}_1 = 30$
- $\hat{\beta}_2 = 20$
- $\hat{\beta}_3 = -10$

Example: Irrigation

How do we test the validity of the diff-in-diff assumption?

- We want to show parallel trends hold
- Need more years of pre-period data
- With this data, we could see whether the slope, or trends, in yields were the same for both groups leading up to the introduction of the irrigation.
- If the slopes are similar in the pre-period, then is more reasonable to assume they would have *continued* to have similar slopes

Regression Discontinuity: Intro

In an RD design, we take advantage of policy quirks where treatment was assigned based on some threshold value of a "running variable". Examples include:

- Age
- Test scores
- Poverty line

Basic idea of an RD is to compare the outcome variable for observations just below and just about the threshold.

- The expectation is that people just below and above the threshold are identical in all observable and non-observable characteristics, except for program participation

Regression Discontinuity: Intro

If people just above and below the threshold are "as good as randomly assigned", then we can do the following:

- Use regression to estimate the relationship between the running variable and the outcome we care about
- We can do this both above and below the threshold
- Any "jump" in this relationship at the threshold we can attribute to the program

Regression Discontinuity: Estimation

We operationalize this idea by estimating this model:

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 (RunningVar_i - threshold)$$

$$+ \beta_3 T_i \times (RunningVar_i - threshold) + u_i$$

- $RunningVar_i$ is the running variable
- $threshold$ is the threshold value for being treated or not treated
- T_i is a dummy variable if the the observation has a value of the running variable that indicates it received treatment

Question: What coefficient tells us the effect of the treatment?

Regression Discontinuity: Estimation

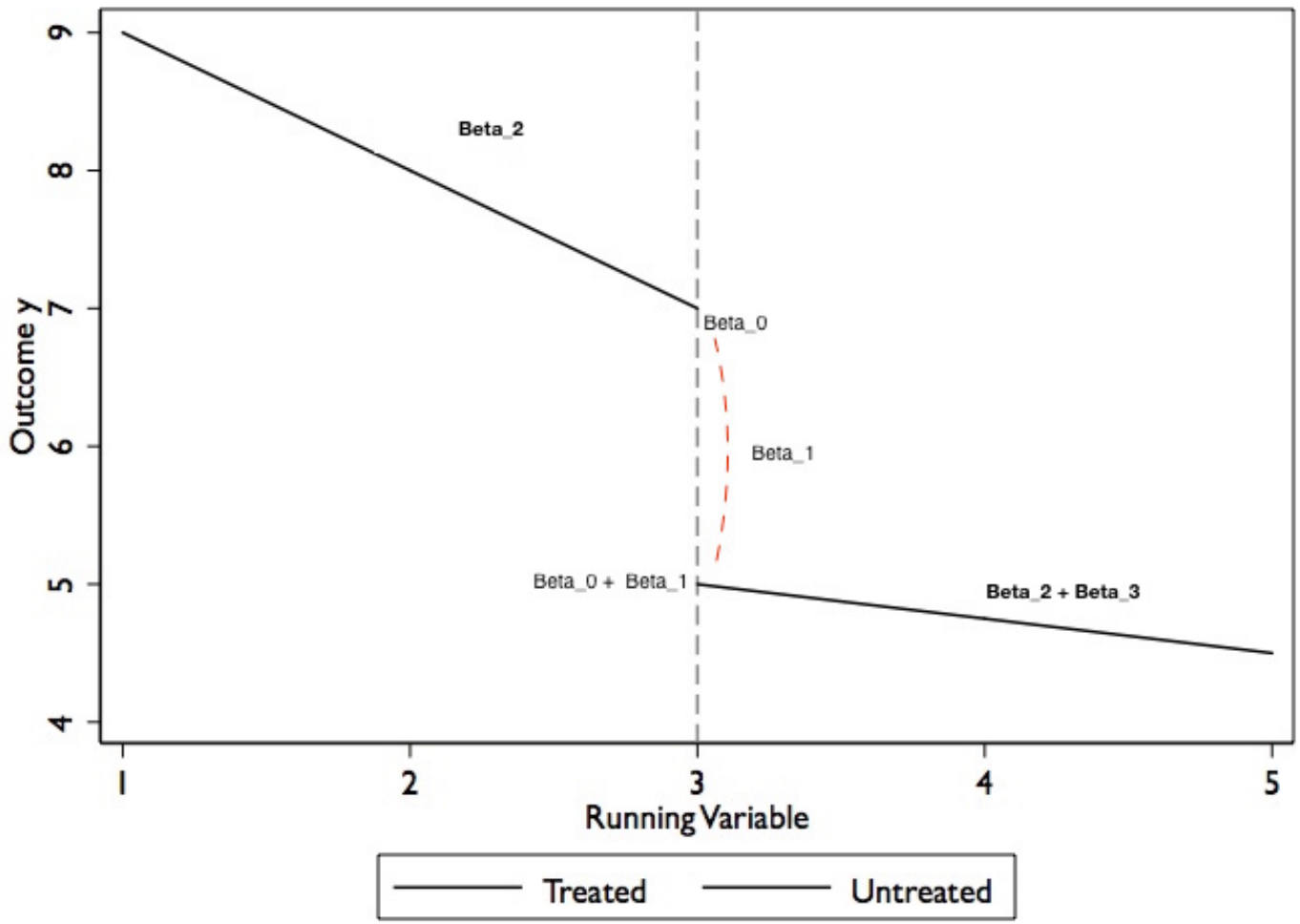
$$y_i = \beta_0 + \beta_1 T_i + \beta_2 (RunningVar_i - threshold)$$
$$+ \beta_3 T_i \times (RunningVar_i - threshold) + u_i$$

- $RunningVar_i$ is the running variable
- $threshold$ is the threshold value for being treated or not treated
- T_i is a dummy variable if the the observation has a value of the running variable that indicates it received treatment

$\hat{\beta}_1$ **captures the effect of of the treatment**

Regression Discontinuity: Estimation

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 (RunningVar_i - threshold)$$
$$+\beta_3 T_i \times (RunningVar_i - threshold) + u_i$$



Regression Discontinuity: Assumptions

Key Assumption:

- Relationship between outcome and running variable would be continuous around the threshold if it were not for the treatment

This assumption might be violated if:

- Participants in the program can manipulate the value of their running variable (e.g. mis-report income to receive subsidy)

Regression Discontinuity: Test Assumption

Test Assumption:

- Check to make sure that the running variable distribution is "smooth" across the threshold. Concerned about manipulation
- Test there are no discontinuities around the running variable threshold for relevant variables **other** than the treatment and the outcome variables
- Look at the averages of observable characteristics of household just above and below the threshold and make sure they're similar (kind of like in RCT)

$$x_i = \beta_0 + \beta_1 T_i + \beta_2 (RunningVar_i - threshold)$$
$$+ \beta_3 T_i \times (RunningVar_i - threshold) + u_i$$

Here we want to find a coefficient of zero for our estimated $\hat{\beta}_1$

Regression Discontinuity: LATE

A key part of the RD identification strategy is that we are only comparing people just above and below the threshold. This has important implications about how we interpret the result:

- Treatment effect from RD analysis is only applicable to individuals that are around the threshold
- We can this a "Local Average Treatment Effect" or **LATE**
- If there is **no** treatment effect heterogeneity along the running variable, then LATE will equal the overall Average Treatment Effect (ATE)
- Treatment effect found from an RD might not be applicable to a similar program that uses a different threshold for eligibility

RD Example

Manacorda & Miguel (2011) studied whether government transfers affects support for a political party. They used an RD design:

- A proxy means test in Uruguay was used to target an anti-poverty cash-transfer program
- Proxy means: use observed characteristics about a household to predict their income (via regression).
- If a household's predicted income is below the threshold, they receive the cash transfers

RD Example

- The running variable is the predicted income level
- Outcome is political support for the ruling party
- Run the following regression

$$y = \beta_0 + \beta_1 T + \beta_2 (Income - thresh) + \beta_3 (Income - thresh) T + u$$

They find that beneficiary households are 11 - 13 percentage points more likely to support the current government

RD Example II

Suppose there was a California Program that gave free college tuition in UC schools to high-schoolers who scored above a 750 on their SAT. We wonder what the effect of this scholarship is on future salaries.

- Write the RD equation that you would use to test this causal relationship
- Describe a test you might run to test the validity of the RD assumption (i.e. the data you want and the process you would use)

RD Example II

- Write the RD equation that you would use to test this causal relationship:

$$income_i = \beta_0 + \beta_1 scholarship_i + \beta_2(SAT_i - 750) +$$

$$\beta_3 scholarship_i \times (SAT_i - 750) + u_i$$

- Describe a test you might run to test the validity of the RD assumption (i.e. the data you want and the process you would use)
  - We would need other observables about the students from when they were in high school. We would then check for smoothness across the threshold
  - E.g. looks at parents' income:

$$parents\ income_i = \beta_0 + \beta_1 scholarship_i + \beta_2(SAT_i - 750) +$$

$$\beta_3 scholarship_i \times (SAT_i - 750) + u_i$$