Capacity Building workshop on Impact Evaluation of Employment Programs

**Measuring Impact: Part 2**

Maciej Jakubowski, Gdańsk, February 22, 2017
Quasi-experimental methods (require more assumptions)

IE Methods Toolbox

- Randomized Assignment
- Difference-in-Differences
- Regression Discontinuity Design
- Matching
Recap: need the right proxy for the counter-factual

With large sample, random program assignment ensures two groups have very similar characteristics ON AVERAGE

A lottery ensures the two groups are the same
But what if you **cannot** randomize?

TREATMENT GROUP

CONTROL GROUP

not the same
DISCONTINUITY DESIGN
Discontinuity Design

Many programs select participants using an index or score:

- Anti-poverty Programs
  - Targeted to households below a given poverty index/income

- Pensions
  - Targeted to population above a certain age

- Education
  - Scholarships for students with high scores on standardized text
  - Programs for certain age groups (youth, elderly)

- Labor
  - Programs targeted on the duration of unemployment
Discontinuity Design

Compare outcomes $Y$ for units just *above and below* the cut-off point

Units just above the cut-off point are very *similar* to units just below it – *good comparison.*
Regression Discontinuity Design - Baseline

Outcome (e.g. probability of being employed)

Score (e.g. age)

- Not eligible
- Eligible
Regression Discontinuity Design - Post Intervention

Outcome (e.g. probability of being employed)

Score (e.g. age)
For a Discontinuity Design you need...

1) **Continuous eligibility index**
   - e.g. income
   - e.g. age
   - e.g. unemployment spell (months)

2) **Clearly defined cut-off.**

Participants with a score ≤ cutoff are **eligible**
Participants with a score > cutoff are **not-eligible**
Or **vice-versa**
Example with 2 thresholds: Cambodia CCT

- **Eligibility** is based on an index of the probability of dropping out of school.

- 2 **cutoff points** within each school:
  - Applicants with the highest dropout risk offered US $60 per year scholarship
  - Applicants with intermediate dropout risk offered US $45 per year scholarship
  - Applicants with low dropout risk were not offered scholarship by the program
Example with 2 thresholds: Large impact on $45 scholarship

- No scholarship versus $45
- $60 versus $45 scholarship

Advantages of RDD for evaluation

Yields an *unbiased* estimate of treatment effect at the discontinuity

Can take advantage of a known rule for assigning the benefit

- This is common in the design of social interventions
- No need to “exclude” a group of eligible households/individuals from treatment
Warning

• **Need a large enough sample** of people around the cut-off. Because only compare who around the cut-off.

• **Not always generalizable**: it tells us the impact of the program for the people around the cut-off.
  
  – Youth program: would the program have the same impact for very young (16-20 y.o.) and not so young (20-24 y.o.) participants?
  
  – Progresa: would the impact have the same impact for very poor and not so poor households?
Example: Effect of Youth Job Guarantee on employment [Sweden]

**Program**

WHAT? Youth employment program (<25) with in-depth activation once the jobseeker has been registered for more than 90 days at the PES

WHERE? Started in 2007, Sweden

- Open unemployment and registration at the PES (3 months)
- **The Youth Job Guarantee:**
  - In-depth assessment and counseling
  - Job seeking activities with coaching (at least 3 months, normally)
  - Work experience or training
  - Job seeking activities at least 4h/week (12 months, max 15 months in total)

WHY? Increase employment among long-term unemployed youth

**Method/Eligibility**

- Registered jobseekers who are unemployed for more than 90 days, and who are <25 years old are **eligible**
- Registered jobseekers who are unemployed for more than 90 days, but who are >=25 years old are **not eligible**

**Data**

Look at two cohorts: 2008 and 2009

*Combine* data from PES and Health registries
Example: Effect of Youth Job Guarantee on employment [Sweden]

Results

• Participants in the YJG program increases the probability of finding employment during the first 90 days of the unemployment spell by around 2 percentage points.
• Taking into account that about 28 percent of the 25-year-olds find employment within 90 days, this would correspond to an increase of about 7 percent.

IMPACT = 2 percentage points (≈30%-28%)
Poland:
When can you use RDD?

Examples/discussion

• Can you think about programs in Poland that could use a Regression Discontinuity Design (RDD)?

• What would be the eligibility rule (how do you define participants and non-participants)?

• Could there be some issues with the eligibility rule?
Keep in Mind

Discontinuity Design

- Requires **continuous eligibility criteria with clear cut-off**.
- Gives unbiased estimate of the treatment effect: *Observations just across the cut-off are good comparisons*.
- No need to **exclude** a group of eligible households/individuals from treatment.
- Can sometimes use it for programs that already ongoing.
DIFFERENCE-IN-DIFFERENCES
(DIFF-IN-DIFF)
### Difference-in-differences (Diff-in-diff)

\[ \text{Y} = \text{Probability of being employed} \]

\[ \text{P} = \text{Youth training program} \]

<table>
<thead>
<tr>
<th></th>
<th>Enrolled (T)</th>
<th>Not Enrolled (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>After (1)</td>
<td>0.74</td>
<td>0.81</td>
</tr>
<tr>
<td>Before (0)</td>
<td>0.60</td>
<td>0.78</td>
</tr>
<tr>
<td>Difference</td>
<td>+0.14</td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td>+0.03</td>
<td>= 0.11</td>
</tr>
</tbody>
</table>

*Diff-in-Diff: Impact = (Y}_{T1} - Y_{T0} - (Y}_{C1} - Y_{C0})
Impact = (A - B) - (C - D) = (A - C) - (B - D)

Probability of being employed

Enrolled

Not enrolled

Similar trends before the program

B = 0.60

C = 0.81

D = 0.78

A = 0.74

Impact = 0.11
Example: New Deal for Young People [United Kingdom]

**Program**

**WHAT?** Program for 18-24 y.o. who have been claiming Jobseeker’s Allowance (JSA) for 6 months or more. It provides opportunities to work, get new skills and/or get work experience in the voluntary and environmental sectors of the economy.

**WHERE?** Introduced in the UK in 1998.

**WHY?** Help the young unemployed into work and increase their employability.

**Method/Eligibility**

- Individuals who are between 18 and 24 years old, and who have been registered unemployed for more than 6 months are **the treatment group**.
- Individuals who are between 30 and 39 years old, and who have been unemployed for more than 6 months are **the control group**.

**Data**

Joint Unemployment and Vacancies Operating System, (JUVOS), and covers the period up to February 2001, 32 months after the start of the national program.
Problem I: Assume Equal Trends

- **Diff-in-Diff only valid** if both groups had similar trends before the program.
- The change in observed outcomes for those not enrolled would have been the same for those that are enrolled.
- What if attendance for those enrolled would have increased by more than those not enrolled in any case?
Same Trend

Similar trends before the program

Probability of being employed

B = 0.60
C = 0.81
D = 0.78
A = 0.74

T = 0
T = 1
Different Trend

Different trends before the program

Diff-in-Diff cannot measure the impact of the program
What if an event affects only one group?

Case 1: Training program
• Only highly motivated people participated in the program
• Only candidates who are expected to be successful in the training are enrolled by the placement officers

DiD overestimate the effect of the program

Case 2: Grants to set up a business with the UK
• Treatment group = small businesses working with the UK
• Control group = small domestic businesses
• Only the treatment group is affected by the BREXIT

DiD underestimates the effect of the program
To test this, at least 3 observations in time are needed:
- 2 observations before
- 1 observation after.
Problem 2: Changes in group composition over time

- Diff-in-Diff requires that we follow the same types of people over time.

For example, all the skilled people drop out of a training program, because they don’t need the training. So average training outcomes for those in the program is lower at the end of the program.

DiD underestimates the effect of the program.

For example, all the unskilled people drop out of a training program, because they cannot travel to the training location.

DiD overestimates the effect of the program.
Keep in Mind

Difference-in-Differences

Combines *Enrolled & Not Enrolled* with *Before & After*.

**Slope:** Generate counterfactual for change in outcome

**FUNDAMENTAL ASSUMPTION**
Trends –slopes– are the same in treatments and comparisons

To test this, at least 3 observations in time are needed:
- 2 observations *before*
- 1 observation *after*.
Poland:
When can you use Difference-in-Difference?

Examples/discussion

• What are the programs in Poland that may have used Difference-in-difference (DiD)?

• Can you think about programs in Poland that could use Difference-in-difference (DiD)?

• How would you define participants and non-participants?

• Could there be some issues with Difference-in-difference (DiD)?
PROPENSITY SCORE MATCHING
(PSM)
Matching

• The group that enrolled is, on average, different from the group that did not enroll.
• However, some individuals are similar.
• So, can match similar individuals with each other.
Group Exercise

Can everyone stand up?
Try to compare outcomes for similar people

<table>
<thead>
<tr>
<th>NOT ENROLLED</th>
<th>ENROLLED</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO EDUCATION</td>
<td></td>
</tr>
<tr>
<td>PRIMARY EDUCATION</td>
<td></td>
</tr>
<tr>
<td>SECONDARY EDUCATION</td>
<td></td>
</tr>
<tr>
<td>TERTIARY EDUCATION</td>
<td></td>
</tr>
</tbody>
</table>
More Complicated in Practice

• Match on all observable characteristics (e.g. income, gender, education...)

• Comparison group: non-participants with similar characteristics

• Create one aggregate Propensity Score to match:
  – Compute everyone’s probability of participating, based on their observable characteristics
  – Choose matches that have the same probability of participation as the treatments
Try to compare outcomes for similar people

<table>
<thead>
<tr>
<th>NOT ENROLLED</th>
<th>ENROLLED</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO EDUCATION</td>
<td>ENROLLED</td>
</tr>
<tr>
<td>PRIMARY EDUCATION</td>
<td>ENROLLED</td>
</tr>
<tr>
<td>SECONDARY EDUCATION</td>
<td>ENROLLED</td>
</tr>
<tr>
<td>TERTIARY EDUCATION</td>
<td>ENROLLED</td>
</tr>
</tbody>
</table>
Problem Two: Can Only Match on Observables

MATCHING DOES NOT OVERCOME SELECTION PROBLEM!

What if we can’t collect data on people characteristics that are relevant for program participation and outcomes?
Keep in Mind

Matching

Requires large samples and good quality data.

Matching at baseline can be very useful:
- Know the assignment rule and match based on it
- combine with other techniques (i.e. diff-in-diff)

Ex-post matching is risky:
- If there is no baseline, be careful!
- matching on endogenous ex-post variables gives bad results.
Example: Active Labor Market Programs [Poland]

Programs

WHAT? Three ALMP programs to combat unemployment in Poland:
(i) training: for workers with redundant or no skills → skills in high demand
(ii) “intervention works”: wages subsidies in the amount of the unemployment benefit
(iii) public works: jobs created by local government/municipalities


Method/Eligibility

- Individuals who declare that they participated in one of the three programs are the treatment group
- Individuals who did not participate, but have similar characteristics before the program took place are the control group

Data

Special supplement of the Polish Labor force Survey (LFS) of 1996 which includes retrospective data

Results

(i) Training = performs well, (ii) “intervention works” = none for women, negative for men, (iii) public works = negative for men
Poland:
When can you use Matching?

Examples/discussion

• What are the programs in Poland that may have used some sort of matching techniques?

• Can you think about programs in Poland that could use Propensity Score Matching (PSM)?

• How would you define participants and non-participants?

• Could there be some issues with Propensity Score Matching (PSM)?
1. Clear understanding of the intervention
   Research question: “what are we evaluating?”

2. Well defined outcomes (impacts)
   What is/are the goal(s) of our policy?

3. Credible identification strategy/ies (definition of the counterfactuals)
   What methodology should we choose?

4. Reliable data
<table>
<thead>
<tr>
<th>METHODOLOGY</th>
<th>DESCRIPTION</th>
<th>WHO IS IN THE COMPARISON GROUP?</th>
<th>THE METHODOLOGY IS ONLY VALID IF...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before-and-after</td>
<td>Measure how program participants improved (or changed) over time.</td>
<td>Program participants themselves—before participating in the program.</td>
<td>The program was the only factor influencing changes in the outcome over time. If the program did not exist, outcomes would be the same before and after the study period.</td>
</tr>
<tr>
<td>Enrolled vs not-enrolled</td>
<td>Measure the difference between program participants and non-participants after the program is completed.</td>
<td>Individuals who did not participate in the program (for any reason), but for whom data were collected after the program.</td>
<td>Non-participants and participants were equally likely to enter the program before it started. Non-participants are identical to participants, except they did not participate in the program.</td>
</tr>
<tr>
<td>Randomized Control Trial (RCT)</td>
<td>Random assignment (e.g. a coin toss or random number generator) determines who may participate in the program so that those assigned to participate in the program are, on average, the same as those who are not, in both observable and unobservable ways. Since the participants and nonparticipants are comparable, except that one group received the program, any differences in outcomes result from the causal effect of the program.</td>
<td>Participants who are randomly assigned to not participate in the program. This is often called the “control” group.</td>
<td>Randomization “worked” and the two groups are statistically identical (on observed and unobserved factors). The effects of the treatment do not spill over to the control group. Any behavioral changes are driven by the program—not by the evaluation itself, or by the fact that the participants or non-participants are being studied. If outcome data are missing, data for the same types of individuals are missing from both the control and treatment groups.</td>
</tr>
<tr>
<td>METHODOLOGY</td>
<td>DESCRIPTION</td>
<td>WHO IS IN THE COMPARISON GROUP?</td>
<td>THE METHODOLOGY IS ONLY VALID IF...</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Regression Discontinuity Design (RDD)</td>
<td>Individuals are ranked or assigned a score based on specific, measurable criteria. A cutoff determines whether an individual is eligible to participate in the program. Participants who are just above the cutoff are compared to non-participants who are just below the cutoff.</td>
<td>Individuals who are close to the cutoff, but fall on the “wrong” side of that cutoff, and therefore do not get the program.</td>
<td>After adjusting for the eligibility criteria (and other observed characteristics), the individuals directly below and directly above the cut-off score are statistically identical. The cutoff criteria must have been strictly adhered to. The cutoff must not have been manipulated to ensure that certain individuals qualify for the program.</td>
</tr>
<tr>
<td>Difference-in-differences (DiD)</td>
<td>Measure the before-and-after change in outcomes for the program participants, then subtract the before-and-after change in outcomes of the non-participants to find the relative change in outcomes for program participants.</td>
<td>Individuals who did not participate in the program (for any reason), but for whom data were collected both before and after the program.</td>
<td>If the program had not existed, the participants and non-participants would have experienced identical trajectories during the study period. Any differences in characteristics between the treatment and control group do not have more or less of an effect over time on outcomes.</td>
</tr>
<tr>
<td>Propensity Score Matching (PSM)</td>
<td>Individuals who received a program are compared to similar individuals who did not receive it.</td>
<td>Propensity score matching: For each participant, a non-participant with the same likelihood of participating, as predicted by known characteristics such as age, gender, and occupation.</td>
<td>The characteristics that were not included (because they are unobservable or have not been measured) either do not affect outcomes or do not differ between participants and non-participants.</td>
</tr>
</tbody>
</table>
Choosing your IE method(s)

<table>
<thead>
<tr>
<th>Money</th>
<th>Excess demand</th>
<th>No Excess demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Targeted</td>
<td>Universal</td>
<td>Targeted</td>
</tr>
<tr>
<td>Timing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phased Roll-out</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ Randomized assignment + RDD</td>
<td>+ Randomized assignment + Randomized promotion + DD with matching</td>
<td>+ Randomized assignment + RDD</td>
</tr>
<tr>
<td>Immediate Roll-out</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ Randomized assignment + RDD</td>
<td>+ Randomized assignment + Randomized Promotion + DD with matching</td>
<td>+ RDD</td>
</tr>
</tbody>
</table>

If less than full take-up:
+ Randomized Promotion + DD with matching
Test
Q1: What is the short-coming(s) of difference-difference?

A. Those enrolled in the program might have a different trend over time as those not enrolled.

B. It does not have a counter-factual.

C. Sample size might be too small.

D. People who are different to comparison group might drop out of the program.

E. Both A and C

F. Both A and D.
Q2: You are evaluating a school management reform program that targets poor school. You decide to perform a diff-diff, comparing target schools with schools that did not receive the program.

Over the same period government deployed more teachers to poor areas. **Would this over-estimate or under-estimate the program?**

A. Over-estimate  
B. Under-estimate  
C. Neither
Q3: What is the biggest short-coming of propensity match scoring?

A. Cannot match on observables characteristics
B. Cannot match on unobservables characteristics
C. Different trends between treatment and comparison groups.
When is it possible to do regression discontinuity design?

A. When there is a continuous eligibility criteria with a clear cut-off.

B. When there is a comparison group of people who do not receive the program.

C. When government randomly assigns some to receive the program and some not.
Thank you!