clear

Regional Centers for
Learning on Evaluation and Results

Impact Evaluation: Scope and Limits in the Real World

Claudia Maldonado

17th October 2017

NEC UNDP

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Evaluation

Paraphrasing Wildavsky, evaluation is:

a way of **speaking truth to power**

The use of the tools of social science to answer questions about attributes of public programs/intervention.

Which attributes?

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Learning Objectives

- Understanding of key concepts in impact evaluation literature
- Gain familiarity with experimental and cuasiexperimental approaches
- Critically discuss scope and limits of impact evaluation in the real world

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Warm-up discusion

- What is evidence?
- What is impact?
- What is strong evidence of impact?
- Why is strong evidence expected from evaluation?

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Impact as casual effect

- Causal link = independent effect
- Average treatment effect (ATE)
- Impact evaluation implies the *identification* and *isolation* of the independent causal effect of an intervention on the treated population.

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Impact evaluation

… requires meeting the challenge of making **causal inferences** using the methodological tools of scientific inference.

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

The Clinical Metaphor

- A program is a discrete and exogenous intervention (X_{t1}) on a given problem/situation (Y_{t1}) that seeks to alter the *status quo in some desired direction*. ($Y_{t1} \neq Y_{t2}$)
- If Y is rural poverty, then it is desired that ($Y_{t1} > Y_{t2}$)
- X_{t1}=treatment Y_{t1}= problem

CIDE

Types of evaluation (questions)

- Is Y_{t1} (z_1,z_2,z_3 ,…z_n) a public problem that is susceptible of public intervention? (**Diagnostics**)

- Is X_{t1} and adequate instrument to address Y_{t1}? (**Design**) Xε (*Z)*

- What is the actual change in Y_{t2} directly and exclusively attributable to X_{t1} ?(**Impact**)

Another way of looking at impact

- The empirical verification of a program´s theory of change

- The answer to a counterfactual
- **"What would have happened to the treated population in the absence of the program?**(impact variable, i.e. rural poverty)

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

What type of impact?

- Define impact variable (variable to be observed)
- Define aggregation level (individual, household, community?)
- Define measurement and means of verification (how will you know if and when change occurs?)

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Impact Evaluation: How to identify net impact?

- Δ in impact variable is NOT impact
- Δ=0 does NOT mean there is no impact
- Δ=X does NOT reflect the *magnitude* or *direction* of impact

Why?

CIDE

The problem of attribution

There are factors that affect the problem that coexist with the program,but are not related to the program:

- Time
- External shocks
- Demographic change
- Other programs
- Preferences of the population in the program
- Changes in relative prices
- Characteristics of the treated
- And a long…. very long etcétera.

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Counterfactual

What would have happened if you had chosen not to attend NEC 2017?

Well, we may NEVER know!!!!!

CIDE

Because…in order to answer that question

- We must observe the universe in Yt1 without a program and that same universe in Yt1 with program, at the same time.

- If we observe Yt2 with program and compare it to Yt1 without program (before and after), we are going to observe the effects of the program with the effects of everything else (the world turning around).

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Common problems ….

- **Before and after comparisons**
- **Comparison of non comparable groups/units**
- **Bias** (selection, attrition, measurement error)
- **Observables**
- **Non observables**

CIDE

Exercise 1: the fallacy of attribution

- A community called Penkalingo has suffered from severe unemployment for the last 4 years. Last year, the President of Penkalingo launched a ambitious public works program in order to create 800,000 new temporary jobs, and cut unemployment by half.
- 2012-2016: unemployment 34% average
- 2017: unemployment rate 38% so far.

The problem of attribution

- Since it is impossible to observe two state of things that are mutually exclusive at the same time….
- We need to build a counterfactual ….
- Using the experimental method …

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Experimental Design

- Random assignment of treatment & control group.
- Average impact variable value is compared between treatment and control groups.
- The difference in means is attributable to the program

- This is a randomized controlled trial (RCT)

CIDE

Random selection and assignment

- What is it?
- What does it solve?
- What about internal/external validity?

Causal inference

- If X=1 (treatment), then Yt
- If X=0 (control), then Yt does not ocurr (Yc)
- The difference between Yt and Yc is the average treatment effect (ATE) of the intervention X.

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Summary

- RCTs are an evaluation strategy for impact evaluation that credibly claims to rigorously make causal inferences about what works.
- RCTs answer the counterfactual question.
- RCTs require impact evaluation to be part of the design phase of an intervention, so that the logistical procedures to preserve methodological control are taking place (randomization, internal validity,group integrity, time horizon.
- Despite being considered the gold standard withing the evidence based policy movement, they are less frequent than other evaluation strategies, and tend to answer very specific question about what works.

- What types of questions do impact evaluation with RCT NOT answer?

CIDE

clear

Regional Centers for
Learning on Evaluation and Results

Quasiexperimental Methods

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Criticisms to "RCT"

- Ethical considerations
- Logistical considerations
- External validity problems
- Trade off between internal and external validity

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Plan B?

- The statistical (observational) construction of conditions that are close enough to experimental ones.

- PSM: Propensity score matching

CIDE

PSM step by step …

- Construction of a statistical clone of the participants in the treatment, conditional on their observed characteristics (X)
- In other words, groups that may no be identical on “X”, but are equally likely to have participated in the program, given “X” (equivalently propense).

And how do you know …

- We have a group of participants/treated (T=1) and group of not treated (T=0)

- We gather data on their individual characteristics $I_1(x1,x2,x3,x4)$

- We calculate how each characteristic affects their probability of participating in the program/treatment and we get their overall propensity score, given x.
- We compare the averages of the treated and not treated in the impact variable: the difference is the treatment effect on the treated (TOT).

How is it done?

$P = \alpha + \beta 1x1 + \beta 2x2 + \beta 3x3 + ... \varepsilon$ (e.g. Probit model)

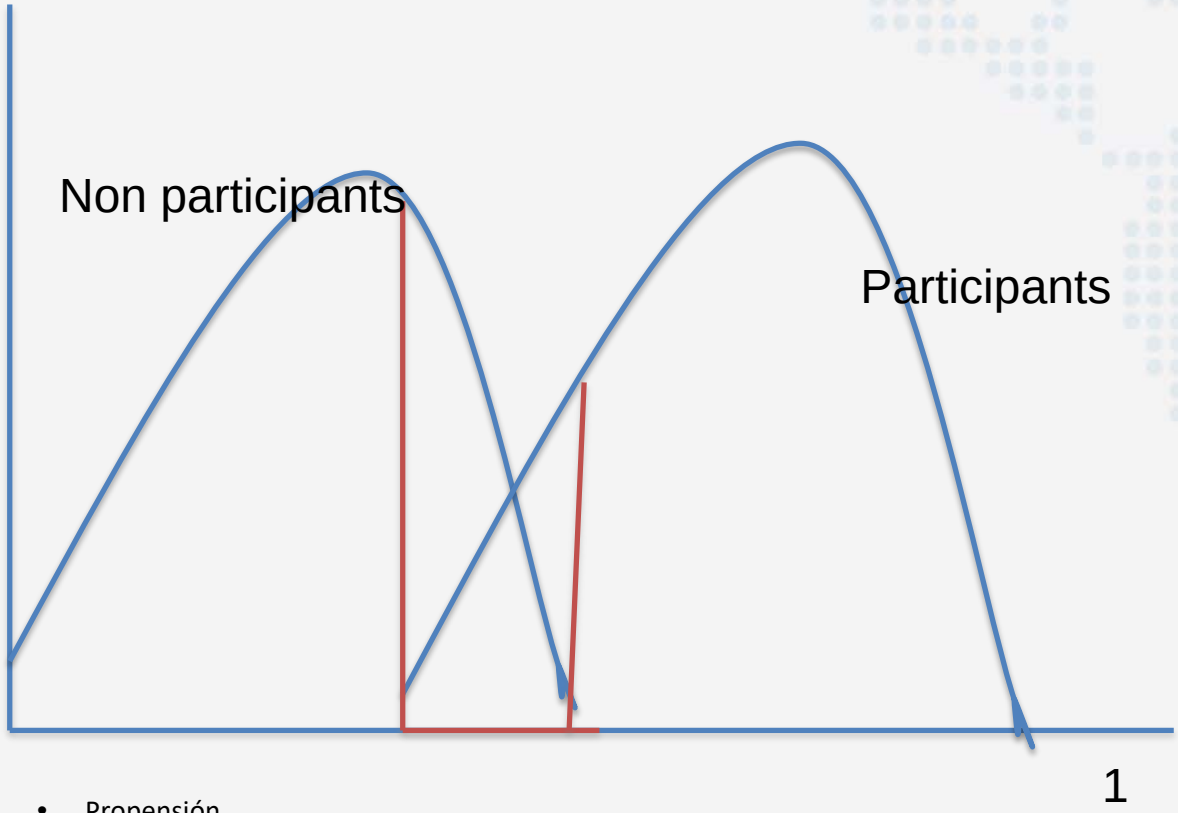
- P = participation (discrete variable)
- β = marginal contribution of a characteristic x on the probability of participating in the program

Assumptions

- Conditional independence: no unobservable variables affect P (the probability of participating)
- Common support: some commonalities in propensity score in control and treatments group.
- Problem: only controls for bias due to observables

Common support

- Density



- Propensión

Formally…

$$TOT_{PSM} = {}_{EP(X) \mid T=1}\{E[Y_T \mid T = 1, P(X)] - E[Y_C \mid T = 0, P(X)]\}.$$

Expected value of Y on the treated

Expected value of Y on the not treated (control)

TOT: Treatment effect on the treated

psm: propensity score matching

E =expected vaue

If there is no « invisible hand » or unobserved variable that is affecting the probability of participatin. Voilá!!!! There is our effect!

Las X no son afectadas por el programa (ej. Si es un programa de ingreso/empleo, no puedo usar esa variable X para el paramiento,porque se correlaciona con Y directamente).

Methods for matching

- Closest neighbour:
- Caliper : por radio
- Stratification
- Kernell (estimación no paramétrica de promedios ponderados de no participantes)

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Example: PROGRESA

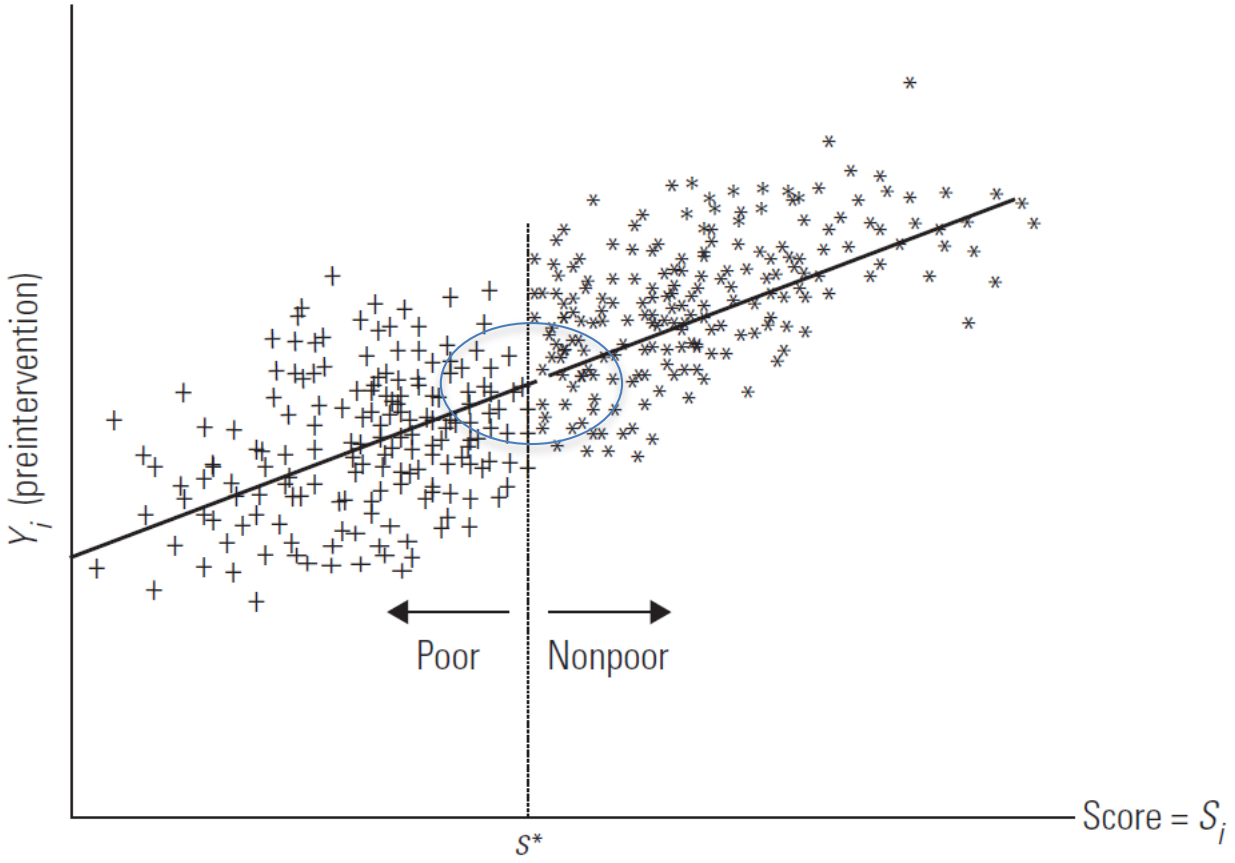
- Díaz y Handa: PSM is reliable when experimental results are compared to quasiexperimental ones. But if data sources (different instruments to measure income, for instance) vary too much, bias is a real problem.

CIDE

Regression discontinuity (RD)

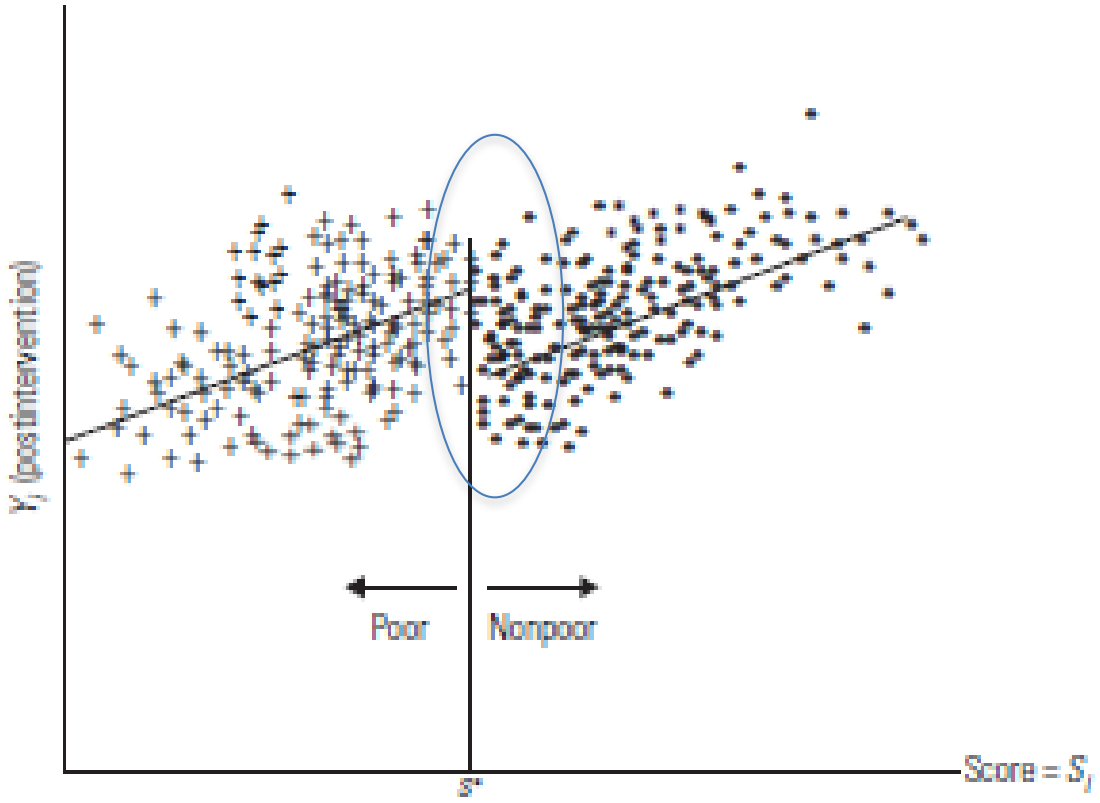
- Non experimental method that uses exogenous rules of eligibility to identify participants and non participants.
- “Controls” for observed and unobserved heterogeneity.
- Compare people/groups before and after the elegibility threshold, asuming that they are fundamentally similar and with the same take up rate.

Pre-intervention scenario



Escenario post-intervención

Figure 7.2 Outcomes after Program Intervention



Source: Authors' representation.

Regression Discontinuity

- PRE
- POST

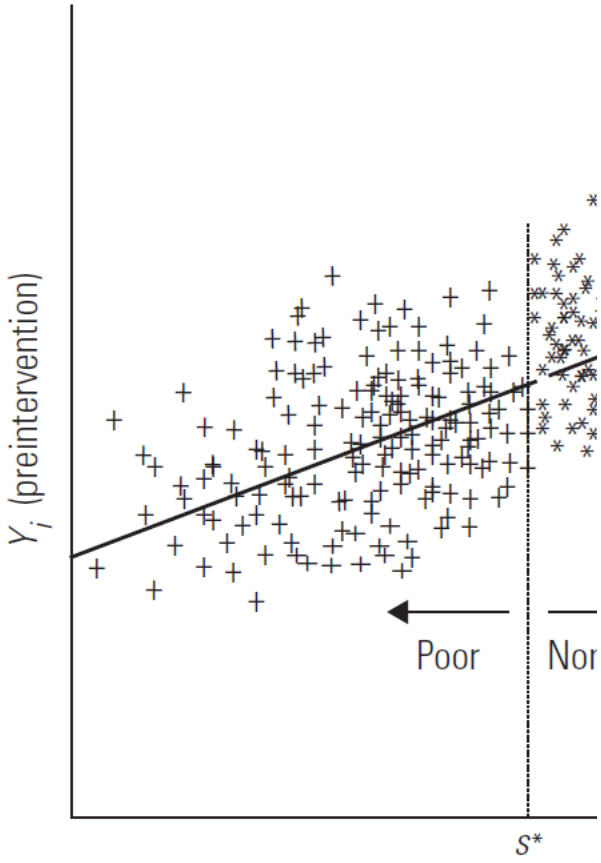
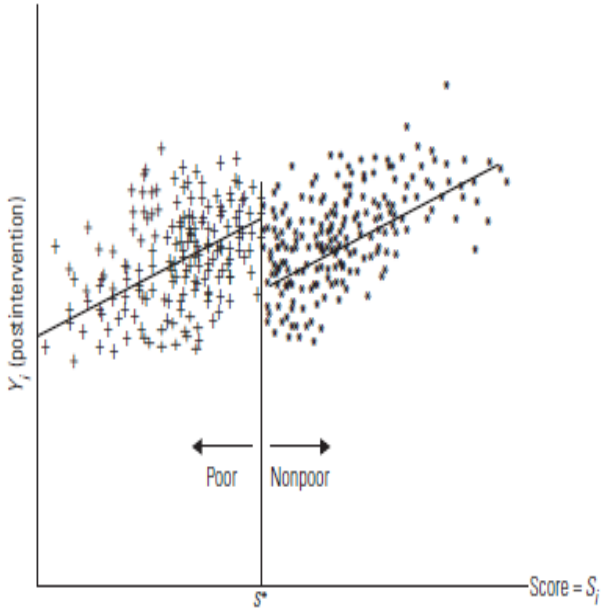


Figure 7.2 Outcomes after Program Intervention



Source: Authors' representation.

Examples?

- We need a clearcut threshold that participants cannot manipulate as elegibility rule.

- If that is the case how different can people be around the cut point?

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Examples

- 65 and older
- Piso firme
- GRE scores for admissions
- Others?

CIDE

Advantages

- Identifies unbiased estimate around the cut point
- Uses program rules in its favour
- Does not imply rationing for control group
- BUT:
- The rule MUST be binding and exogenous.

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Disadvantages

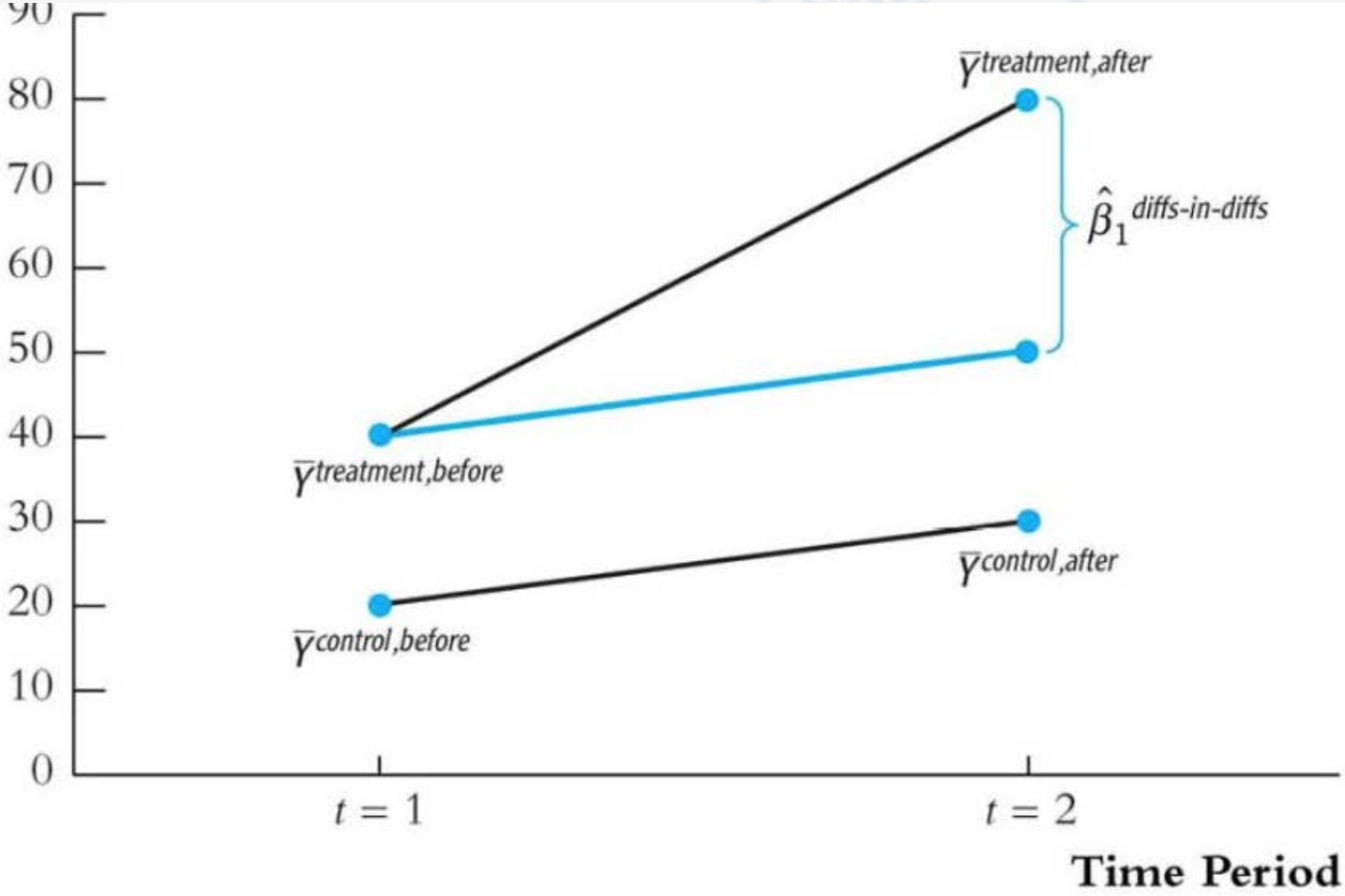
- Identifies average “local” impact, not always generalizable.
- Fewer observations around the cut point. (vs. experimental)
- Results can be less robust (i.e. functional form or interactions).

CIDE

Using Diff in Diff as contrafactual: DD

- Comparison of groups (treatment vs control) according to their trayectories before and after the intervention.
- Baseline of both groups.
- Before and after analysis.
- The difference in the differences IS THE PROGRAM´S EFFECT.

EXAMPLE


90
80
70
60
50
40
30
20
10
0
$\bar{Y}$treatment,after
$\hat{\beta}_1$ diffs-in-diffs
$\bar{Y}$treatment,before
$\bar{Y}$control,after
$\bar{Y}$control,before
$t = 1$
$t = 2$
Time Period

DD

$$DD = E\,(Y_1t - Y_0t \mid T_1 = 1) - E(Y_1c - Y_0c \mid T_1 = 0)$$

Assumptions

- Combined with PSM: time invariant unobservables
- *Solves bias for unobserved heterogeneity iff:*
- *Biases are time invariant*
- *Parallel trayectories*

Ashenfelter´s dip

- Efect of capacity building for labour inclusion
- Participants just suffered a temporal negative shock.
- Treatment groups will anyway show an increasing income over time…

How does this affect the estimate of impact?

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Beyond the black box

- Correlation is NOT causation!!!!!

- In development is it as important to know **what works** as to know **how it works** or doesn´t

- We need causal mechanisms and causal narratives …

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Some examples

- Condoms and pregnancy prevention
- Retroviral and adherence rate
- Preventing infant mortality in indigenous areas
- Influencers and healthy habits

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Criticisms to EBPP

- *Boundary conditions and replicability.*
- Empirical record (CCTs)
- Pilots or prototypes?
- What does impact really mean?

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Qualitative response

- Old fashioned concept of causality from the C. XIX! (positivismo *naive*)
- Contingency and randomness are not the same
- Sampling (representative vs. intentional)
- *Process-tracing* and nested inference
- Interventions are open ended processes in themselves, not treatment like pills.

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Examples

- CCTs in indigenous areas
- No health impacts of proven nutritional supplement
- Others?

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Use of experimental impact evaluation

- Credibility and rigour widely recognized
- Legitimacy
- Evidence for scaling up (replicability)
- Knowledge accumulation on what works
- Global learning and difussion of good practice
- New trend: delivery sciences

CIDE

clear
Regional Centers for Learning on Evaluation and Results
www.clear-la.cide.edu

Final Remarks?

- Gold standard of what?
- What about communication and decision making value?
- Sociology of knowledge matters …
- There is no single right question to ask in an evaluation, there are reasonable and credible ways to answer a specific question, but what really matters is what you want the answer for…

CIDE