Development Learning Lab

What can we learn from Quasi-experimental evaluations?

Jon Einar Flatnes, Chr. Michelsen Institute

June 27, 2023

RCTs are the "first best" option for creating a counterfactual

Without intervention

With intervention

But sometimes you cannot randomize

- Ethical considerations
  - E.g. intentionally withholding potentially beneficial programs that are free or almost free to provide at the margin
- Practical considerations
  - E.g. organization is working with a partner that only operates in certain communities
- Feasibility considerations
  - E.g. intervention has already started

Luckily, there are other “second best” options

→ The objective is still to create a counterfactual, i.e. a measure of what would have happened in the absence of the intervention

→ Other methods can approximate a counterfactual, but are based on several (sometimes strong) assumptions or have other limitations

→ These methods are known as “quasi-experimental” methods

Bad counterfactuals (may lead to incorrect conclusions)

- **Compare before and after (no control):**
  - A change in outcomes over time can be due to many things besides the program (weather, economy, trends, etc.)
- **Non-program recipients as controls (no baseline or other controls)**
  - Program recipients and non-program recipients are often very different, even without the program

Quasi-experimental methods

- Today, we will look at 3 quasi-experimental methods:
  - **Difference-in-differences (DiD)**
  - **Regression discontinuity (RD)**
  - **Propensity score matching (PSM)**

Difference-in-Differences

DiD: Overview

- **What it is**: Compares changes in outcomes between before and after intervention but controlling for the changes in a control group.
  - *Similarly*: Compares difference between treatment and control group, controlling for baseline differences
- **When to use**:
  - When you have baseline & endline data
  - When you have a control group that does not receive the intervention
  - When the control group is similar to the treatment group but may differ by factors that don't change over time

DiD: Example

- Suppose we have data on pre-program/policy incomes (or other variables of interest)

	Control	Treatment
Before	\$1,000	\$1,100
After	\$1,200	\$1,400

- Here, **'Control'** refers to the group of people who did NOT receive/adopt the program/policy, while **'Treatment'** refers to the group who did receive/adopt the program/policy
- $DiD = (y_{treat,after} - y_{contr,after}) - (y_{treat,before} - y_{contr,before})$
- The income increased by \$300 in the group that received the program/policy, but \$200 of those would have happened anyway, as we can see from the control group.

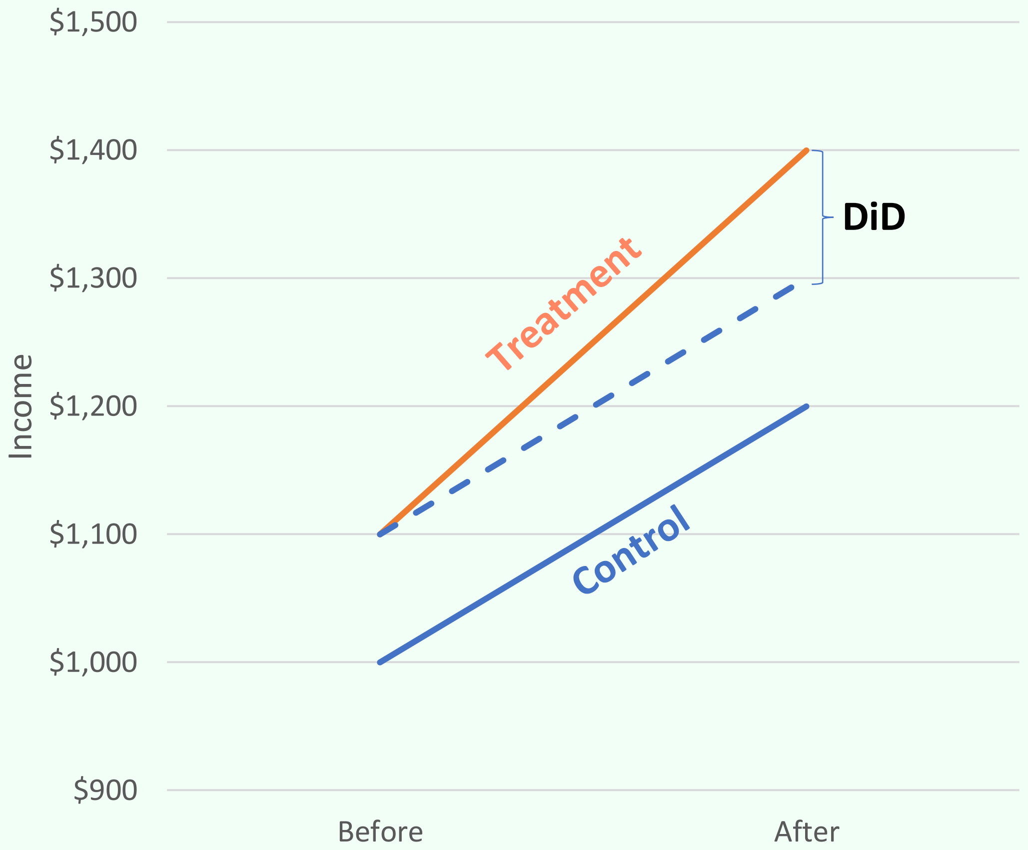
DiD: Example

	Control	Treatment
Before	$1,000	$1,100
After	$1,200	$1,400

→ DiD=(1400-1200)-(1100-1000)=100

OR

→ DiD=(1400-1100)-(1200-1000)=100

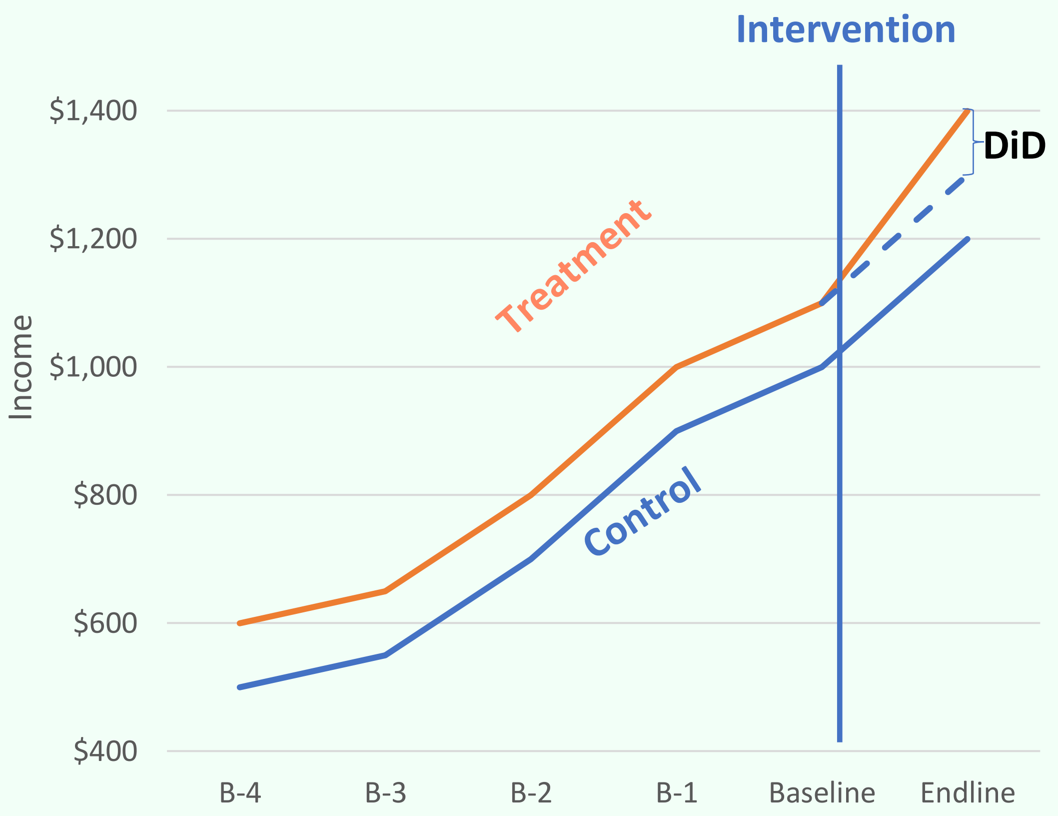


DiD: Assumptions

- Only works if the “**parallel trends assumption**” is satisfied:
  - in the absence of treatment (the program/policy), the two groups (control/treatment) would evolve the same over time
- In other words: the only difference between the treatment and the control groups is the level of the outcome variable(s)

DiD: How to test for parallel trends?

- Need data from before the baseline
- The control group and the treatment group should follow a parallel trend prior to the intervention
- Historic LSMS or government data on village-level may be useful



DiD: Other assumptions & limitations

→ Participants in the control group are not given the intervention, i.e., not moved to the treatment group (and vice versa).

→ If there are unobserved differences between treatment and control at baseline, parallel trends are unlikely to hold, leading to biased (inaccurate) estimates of effects.

→ Any other changes or interventions that affect one group more than the other and occur between baseline and endline can lead to biased (inaccurate) estimates of effects.

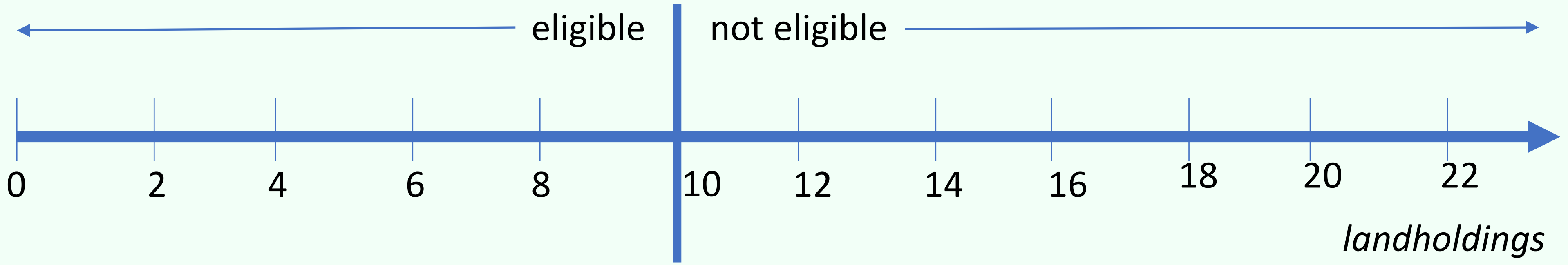
Regression Discontinuity

RD: Overview

- **What it is**: Compares outcomes of people who fall right above and right below some (semi-)continuous eligibility criteria or other program cut-off
- **When to use**:
  - When no baseline data are available (though baseline data helps)
  - When you have one or more clear (semi-)continuous eligibility criteria or cut-offs that are unique to the project and cannot be manipulated

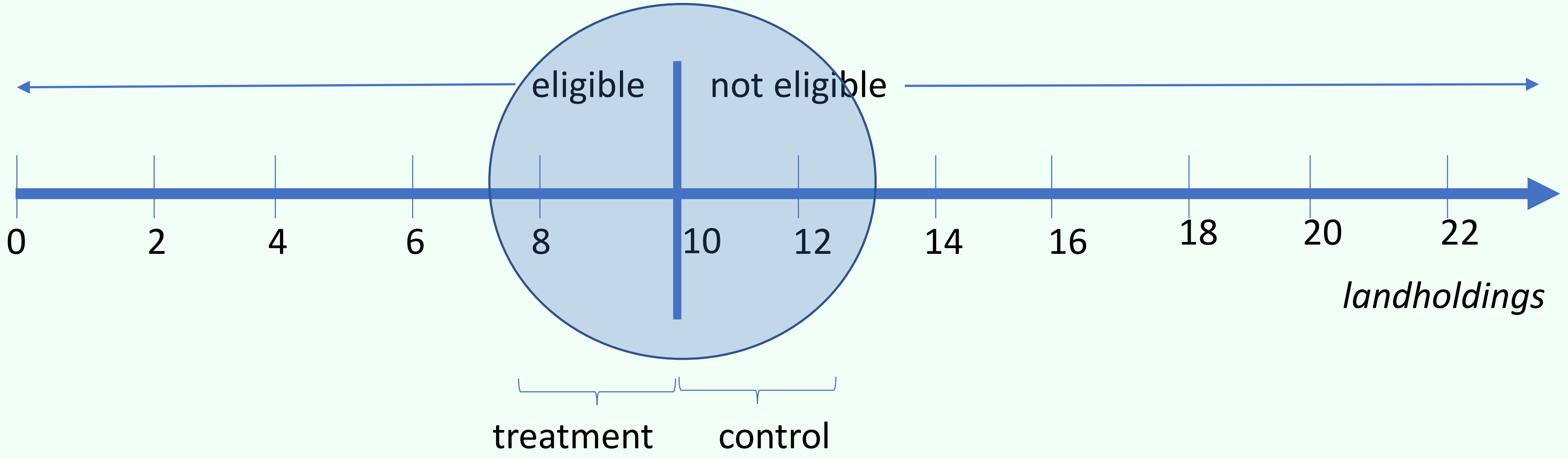
RD: Example

→ Only households with landholdings of 10 acres or less are eligible for an agricultural microfinance loan



RD: Example

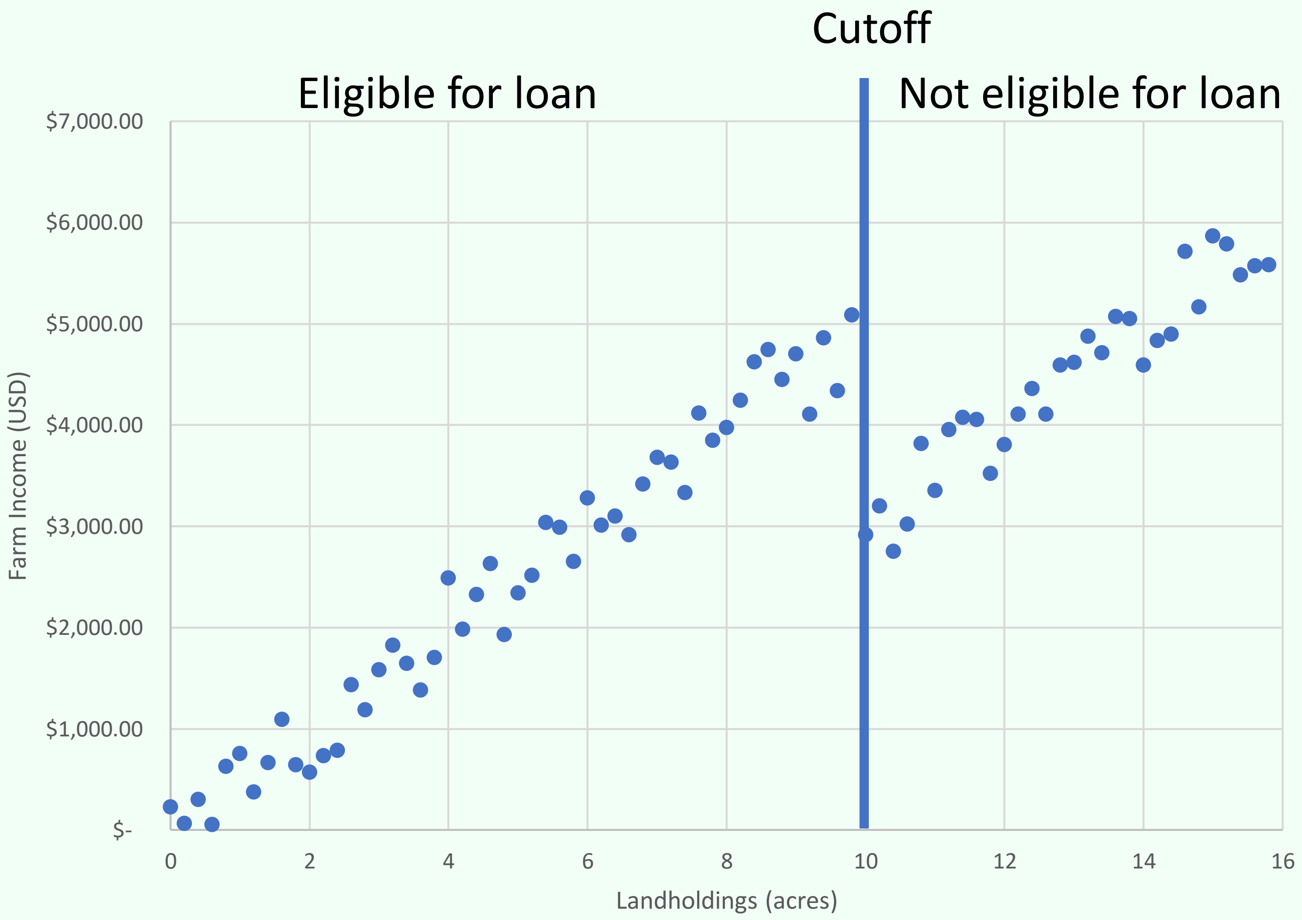
- Only households with landholdings of 10 acres or less are eligible for an agricultural microfinance loan



- We can compare the people who fell right below the eligibility criteria to the those who fell right above

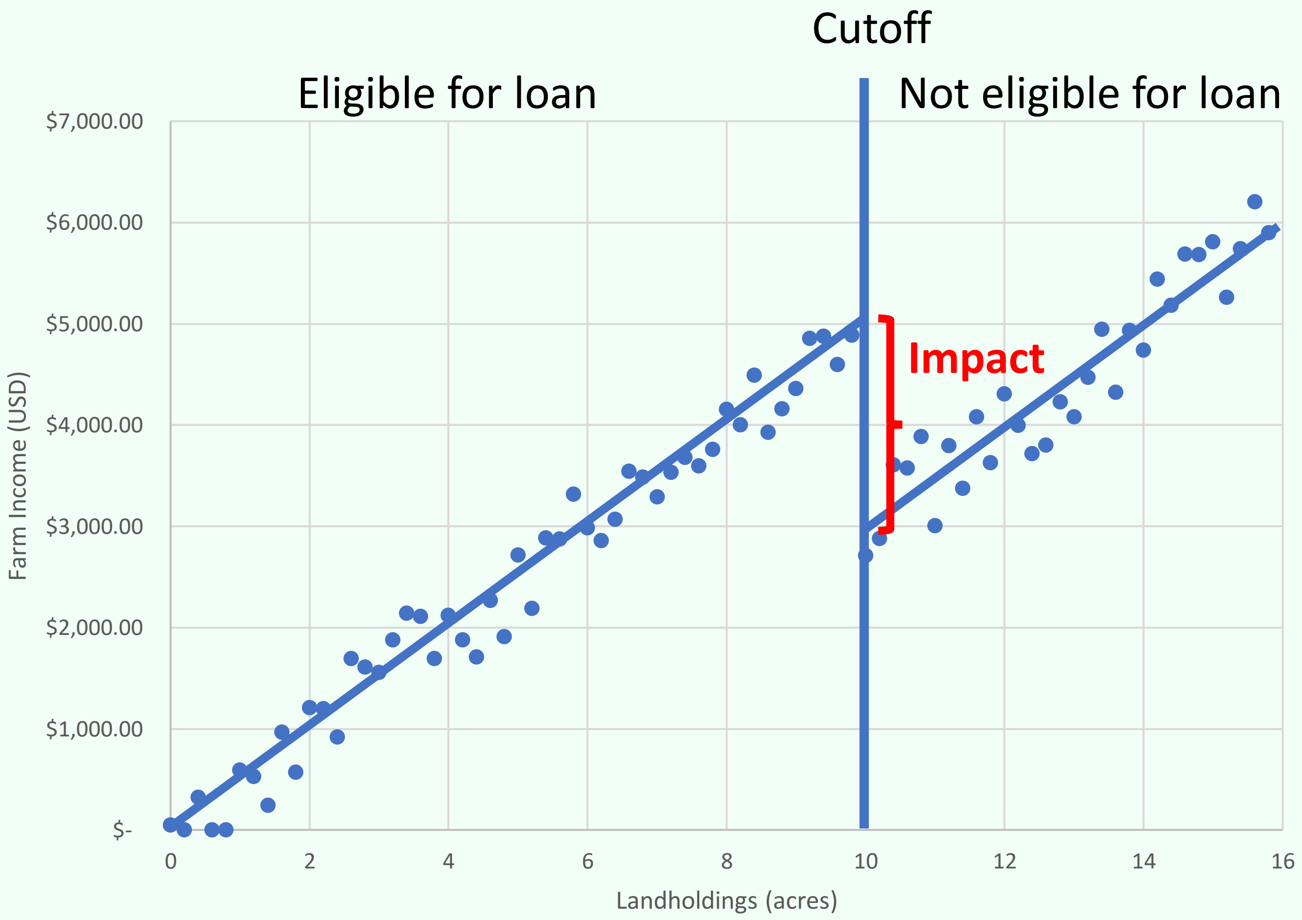
RD: Example

→ Suppose we plot farm income against landholdings using the collected data



RD: Example

- We notice a clear break in the trend at the cutoff
- The break represents the impact of the program
- We can estimate the magnitude of this break using statistics



RD: Assumptions

- People right above and right below the cutoff are similar
  - Baseline data would be useful to test this
- The cutoff should be unique to the project, i.e. there should be no other projects, apart from the project to be evaluated, that uses the same cutoff score
- The eligibility rule and cutoff should be strictly enforced and not be able to be manipulated

RD: Limitations

- → RD does not measure impact of the project for participants that are farther away from the cutoff.
  - → Results may not be generalizable to the entire population.
- → It needs a large sample size to have enough statistical power.
- → Eligibility criteria involving non-numerical categories (e.g. sex) or a limited number of numerical categories (e.g. # of ag. plots) cannot be used

RD: Examples of usable eligibility criteria

- → **Income** (e.g. only those with incomes below $1,000)
- → **Age** (e.g. only people below age 40)
- → **Test scores** (e.g. only students who scored above 70%)
- → **Landholdings** (e.g. only households with less than 10 acres)
- → **Geography** (e.g. only households within a specified polygon on the map; should be careful with political boundaries, such as counties or districts, as there are often other systematic differences between such units)

Propensity score matching

PSM: Overview

- **What it is**: A method of creating a control group by matching each observation in the treatment group with one or several observations from the sample who did not receive the treatment, based on observable characteristics.
- **When to use**:
  - When you have a large, high-quality dataset of many observable characteristics
  - When unobservable characteristics between treatment and control groups have no impact on project allocation
  - When baseline data don't exist (but works much better if baseline data are available)

PSM: Example

Received intervention							Did not receive intervention					
HHID	Age	Sex	Income ($)	Land (ac.)	etc.		HHID	Age	Sex	Income ($)	Land (ac.)	etc.
1	40	M	5,387	4	###		101	45	M	8,567	2	###
2	25	M	2,908	2	###		102	23	F	3,452	5	###
3	75	F	10,608	14	###		103	57	F	2,765	2	###
4	56	M	3,005	4	###		104	75	F	9,868	15	###
5	73	F	1,154	10	###		105	34	M	1,345	12	###

→ If we only had a few variables, we could possibly find non-treated households who matched each treated household exactly (or almost exactly)

PSM: Overview

- But with many variables, it is impossible to find exact, or almost exact, matches
- Instead, we can calculate a propensity score, which is an estimated probability that a given household/person received the intervention. This propensity score is calculated using statistics
    - E.g. if the intervention was targeted to low-income farmers (but not perfectly), the propensity score will be higher for these individuals and lower for e.g. high-income business owners

PSM: Example

Received intervention								Did not receive intervention						
HHID	Age	Sex	Income ($)	Land (ac.)	etc.	**Prop. Score**		HHID	Age	Sex	Income ($)	Land (ac.)	etc.	**Prop. Score**
1	40	M	5,387	4	###	**0.87**		101	45	M	8,567	2	###	**0.05**
2	25	M	2,908	2	###	**0.08**		102	23	F	3,452	5	###	**0.45**
3	75	F	10,608	14	###	**0.64**		103	57	F	2,765	2	###	**0.9**
4	56	M	3,005	4	###	**0.97**		104	75	F	9,868	15	###	**0.76**
5	73	F	1,154	10	###	**0.71**		105	34	M	1,345	12	###	**0.24**

→ For each observation in the dataset, we estimate a propensity score (i.e. how likely is the person to have received the intervention, assuming we didn't know treatment status)

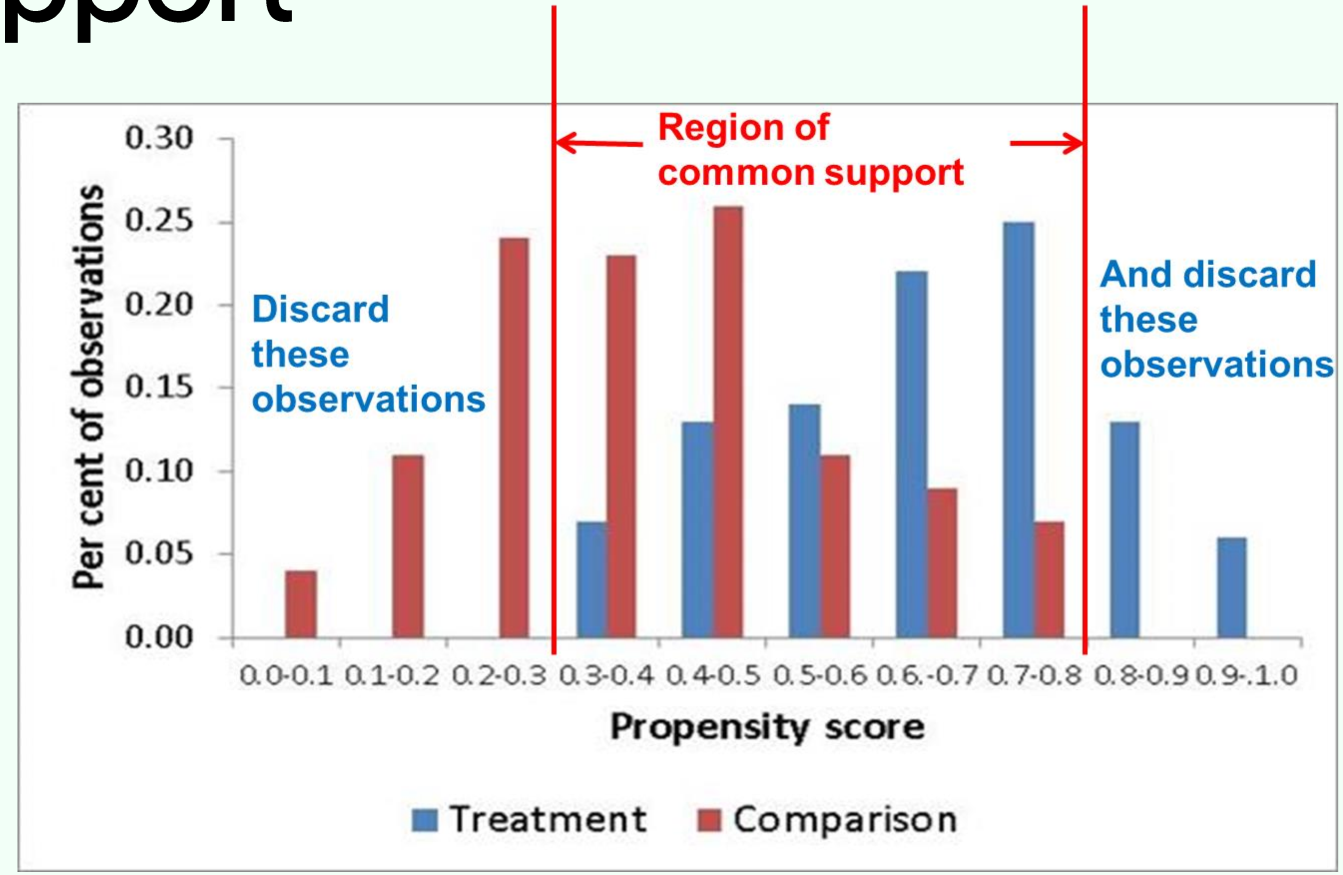
PSM: Example

Received intervention								Did not receive intervention						
HHID	Age	Sex	Income ($)	Land (ac.)	etc.	**Prop. Score**		HHID	Age	Sex	Income ($)	Land (ac.)	etc.	**Prop. Score**
1	40	M	5,387	4	###	**0.87**		101	45	M	8,567	2	###	**0.05**
2	25	M	2,908	2	###	**0.46**		102	23	F	3,452	5	###	**0.45**
3	75	F	10,608	14	###	**0.64**		103	57	F	2,765	2	###	**0.9**
4	56	M	3,005	4	###	**0.97**		104	75	F	9,868	15	###	**0.76**
5	73	F	1,154	10	###	**0.71**		105	34	M	1,345	12	###	**0.24**

→ We can then match observations from the treatment dataset with observations from the control dataset with similar propensity scores (several methods exists for how to match)

PSM: Common support

- Need to have sufficient observations in the control group with similar propensity scores to those in the treatment group
- I.e., for each person who received the intervention, there should exist a person who did not receive the intervention but would have been equally likely to have received it



PSM: Assumptions

→ Assumes that you have a very high-quality, large dataset

→ Assumes that there are no systematic differences in unobservable characteristics between treatment and control groups

→ Assumes that there are enough treatment and control participants with the same propensity score match (common support)

PSM: Limitations

- → Can lead to biased estimates if unobservable characteristics determine program participation
- → If there are many participants with no propensity score match, we may not have enough statistical power
- → Results are conditional on structure of control group and may not be generalizable to the whole population.
- → Potential implication that data are collected but not used

Summary

Method	Need baseline data	Need control (non-program recipients)	Need strict eligibility criteria/cut-off	Observables must explain any difference b/t treatment and control	Main limitation(s)
Randomized Controlled Trial	No, but helps	Yes	No	No	Not always ethical, practical, or feasible
Difference-in-difference	Yes	Yes	No	No, but helps	The control group and the treatment group should follow a parallel trend prior to the intervention
Regression Discontinuity	No, but helps	Yes	Yes	No	->Need a large sample size, especially around the cutoff ->Results may not be generalizable
Propensity Score Matching	No, but helps	Yes	No	Yes	->Can lead to biased estimates if unobservable characteristics determine program participation ->Need a large sample size and common support

Development
Learning
Lab

CMI CHR. MICHELSEN INSTITUTE

UNIVERSITAS BERGENSIS
UNIVERSITY OF BERGEN

NHH

SNF

Group work

- If randomization is not feasible, which methods would you use to establish a valid comparison group?
  - Any quasi-experimental methods that can be used?