Matching Methods & Propensity Scores

Elisabeth Sadoulet
AERC
Mombasa, May 2009

Basic challenge of impact evaluation is to create a counterfactual.

Ideal … randomization → Treatment and Control groups are statistically identical = "No selection into the treatment"

Randomization not always feasible:

- Ex-post evaluation: the program already implemented
- You have no say in implementation

Matching methods to create a comparison group, based on the assumption that selection was based *only on observables*.

Selection on observables (characteristics) **[**misnomer: observed]

Selection on observables:
At given observables, assignation to treatment is "ignorable", i.e., as good as randomization.
i.e., concretely:
We observe all of the variables X that influence both assignation (T or C) and the outcome of interest Y, and assignation is "random" within the sub-populations with same observables.

Selection on non-observables

There exist some non-observables that affect both assignation and outcome of interest.

In a linear econometric model:

$$Y_i = \beta T_i + X_i \alpha + \mu_i$$

outcome treatment observables unobservables

Selection on observables: $E(\mu|T,X) = E(\mu|C,X)$

Control ($T = 0$)

Similar to the concept of exogeneity.
= Condition for getting an unbiased parameter that gives the **causal** effect of T on Y

Recall:
In the randomization case: $Y_i = \beta T_i + \mu_i$ and $E(\mu|T) = E(\mu|C)$

Validity of the assumption

Cannot be verified → we need to assume and argue the case

Selection on observables:

Participation = f(*X*, factors not correlated with determinants of outcome)

Outcome = f(*X*, *T*, factors orthogonal to *T*)

When does it apply?

- A program targeted at people with well-defined characteristics, but that for many reasons (unrelated to the outcome of interest) did not reach all the potential population.
- You have so many variables, including on the behavior of people, that they must capture all of the unobservable effects

- Does not apply when participation is a very clear choice from a subset of the population that were all offered the program.

Examples:

Farmer Field School (FFS) in Peru:
A small pilot extension program promoted by CARE in a few villages in Peru (with intent on expanding it later). Attended by a few potato farmers. But there were thousands of similar farmers who would have qualified and would have chosen to participate if the program had been proposed to them.

Matching Methods & Propensity Scores

Selection on observables (characteristics)
Validity of the assumption
Propensity score matching: Basic idea
Step by step implementation:
Computing Average Treatment on the Treated

Propensity score matching: Basic idea

- Regression framework:

You can estimate the average treatment effect β from the equation:

$$Y_i = \beta T_i + X_i \alpha + \mu_i$$

because, controlling for X, μ is orthogonal to T.
Restrictive because of the imposed linearity.

- Least restrictive would be to only compare observations that all have the exact same values for X, and then define the average treatment effect for this subgroup as:

$$ATE(X) = E(Y|T,X) - E(Y|C,X)$$

But not really practical with many X

- Rosenbaum & Rubin theorem: If assignation is orthogonal to μ, conditional on X, then it is also orthogonal to μ conditional on $p(X) = \Pr(T|X)$, probability of participation given X

Intuition: You do not need to control for each individual X
Nor even for $X\alpha$, which is the influence of X on Y
But only for the correlation between X and T, i.e. $p(X)$

Hence you could get the treatment effect β from the equation:

$$Y_i = \beta T_i + \alpha p(X_i) + \mu_i$$

$p(X) = \Pr(T|X)$ called the *propensity score*

- But then … why keep the linearity restriction?

We can simply compare observations with the same $p(X)$, and define the treatment effect by the difference in their outcomes:

$$ATE(p(X)) = E(Y|T=1, p(X)) - E(Y|T=0, p(X))$$

Step by step implementation:

1. Get representative and comparable data on participants and non-participants

    (ideally using the same survey & a similar time period)

2. Estimate the probability of program participation as a function of observable characteristics

    (using a logit or other discrete choice model)

3. Use predicted values from estimation to generate propensity score $\hat{p}(X_i)$

    for all treatment and comparison group members

4. Match participants: Find a sample of non-participants with similar $\hat{p}(X)$

Restrict samples to ensure *common support*

Determine a *tolerance limit*:

How different can matched control individuals or villages be?

Decide on a *matching technique*

Nearest neighbors, nonlinear matching, multiple matches

5. Once matches are made, we can calculate impact by comparing the means of outcomes across participants and their matches

The difference in outcomes for each participant and its match is the **estimate of the gain** due to the program for that observation.

Calculate the mean of these individual gains to obtain the average overall gain for the participants.

Example: FFS in Peru

Steps 1-3: A large household survey that includes all 93 participants and a random samples in the population (~400 potato farmers)

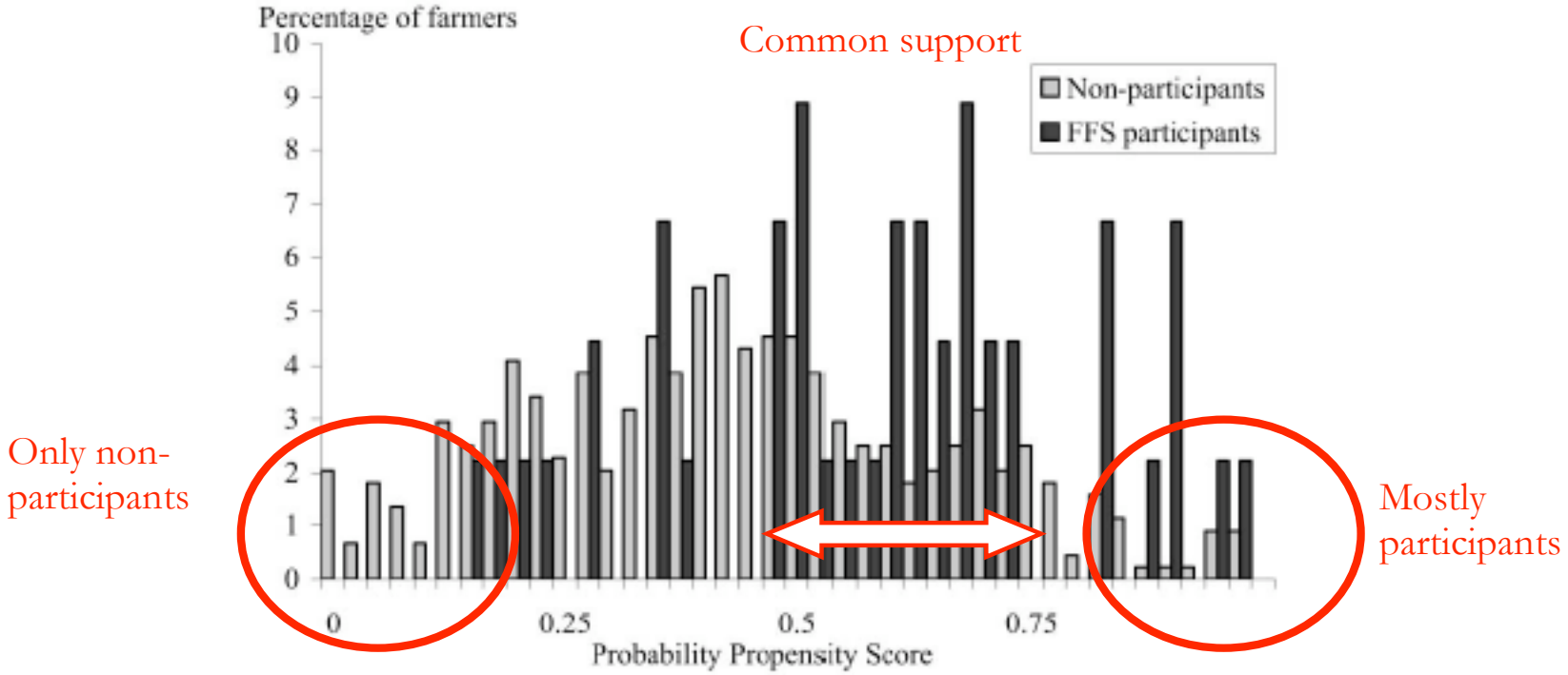
FARMER FIELD SCHOOL PARTICIPATION PROBIT (DEPENDENT VARIABLE: PARTICIPATION [0/1])

	Coefficient	*P*-Value
Education of household head	−.74	.18
Quadratic term for education	.20	.14
Age of household head	−.02	.05
Number of family members	.21	.02
Dependency rate	−.27	.32
Total land ownership (100 ha)	.70	.03
Quadratic term for land ownership	1.05	.41
Value of cattle assets (100 soles)	.01	.76
Number of inherited livestock	.00	1.00
Value of household assets (100 soles)	.22	.00
Quadratic term for household assets	−.01	.00
Value of farm assets (100 soles)	.24	.66
Quadratic term for farm assets	−.23	.21
Fraction of plots lost in El Niño	1.61	.12
Quadratic term for plots lost in El Niño	−2.21	.00
Credit constraint	.16	.69
Constant	.12	.84

+ plot characteristics + community characteristics

Step 4:

Compare the distributions of $\hat{p}(X)$



Eliminate the extremes.

For each participant, select the non-participant with the closest $\hat{p}(X)$

Match only observations that have difference in $\hat{p}(X) < 0.001$

Should also compare the obtained samples on their characteristics, which is called *balancing tests*:

Step 5:
Compare the average outcome (knowledge) in the participant and matched non-participant groups

% that knows	*T*	*C*	Difference
Knowledge on late blight	35.1	25.2	9.9
Knowledge on Andean potato weevil	25.3	8.5	16.8
Knowledge on potato tuber moth	14.9	4.1	10.9
Pesticide knowledge	29.1	20.8	8.3
Knowledge on resistant varieties	49.4	16.0	33.5
Total test score	34.0	18.7	15.3

Computing Average Treatment Effect on the Treated (ATT)

This method gives an average treatment effect on *the population that is used to compute the difference in means*.

What population is this? The participants or a representative sample of the participants. Hence you obtain an average treatment on the treated.

Matching Methods & Propensity Scores

Selection on observables (characteristics)
Validity of the assumption
Propensity score matching: Basic idea
Step by step implementation
Computing Average Treatment on the Treated
Other method based on propensity score
Common use of propensity score methods
Conclusion

Other method: propensity score weighting to obtain ATE.

- Use all the observations (from T and C)
- Compute a weighted average of their outcome, with the weight proportional to $\frac{1}{\hat{p}(X)}$ for the treated and $\frac{1}{1-\hat{p}(X)}$ for the comparison observations
- Obtain the average treatment effect in the population (if the original sample was representative)

$$ATE = \sum_{i \in T} \omega_i Y_i - \sum_{j \in C} \omega_j Y_j$$

(based on the same idea as the weighting in a stratified sample)

Common uses of Propensity Score methods

1. Ex-post matching for estimating the impact of a program with no baseline - Be cautious
   - Can match participants with non-participants using time-invariant characteristics.
   - Can't use variables that change due to program participation (i.e., endogenous variables)
   - Could use pre-determined variables. Usually not available at the individual levels.
   - Can use many village- and neighborhood-level variables.

Example: FFS

2. As a method to select a counterfactual, in conjunction with double difference (ex-post, with panel data)

Similar to what was suggested with an "imperfect" randomization. Using panel data, match observations in the baseline, and then do double difference.

Example:

Evaluation of Fadama II, a CDD in Nigeria, with components of rural infrastructure and advisory services (IFPRI + Nigerian researchers)

Program implementation in 2005, Survey in 2006, with *recall* data for 2004 and 2005 --> Info for one year before the program and one year after the program

3750 households, ~1200 matched

Find that it has an important effect on the beneficiaries, increasing access to services, assets, and incomes.

Need to be cautious with recalls (assets OK, income ?)

3. As a method to select samples (ex-ante)

Select non-randomized (but matched) evaluation samples ex-ante, when randomization is not acceptable/feasible.

Example:

Yemen - Evaluation of Rainfed Agriculture and Livestock Program.

First phase income generation projects in 200 villages.

Village selection done, but project not started and baseline not done yet.

Regions selected by (?), districts and villages from agroecological information and census data on structure of production, by the Social Funds technicians themselves. Hence use the same source of information and find matches to create the comparison sample. Then do baseline survey and will do follow-up.

Summary

Identification Assumption

Selection on Observables: After controlling for observables, treated and control groups are not systematically different

Data Requirements

Rich data on as many observable characteristics as possible.
Large sample size (so that it is possible to find appropriate matches)

Advantages

- Might be possible to do with existing survey data. Doesn't require randomization/experiment

- Might be possible even completely ex-post, with no baseline. However, much better if you can combine with a double difference that will control for the time invariant unobservables

- Allows estimation of heterogeneous treatment effects because we have individual counterfactuals, instead of just having group averages.

- Doesn't require assumption of linearity

Disadvantages

- Strong identifying assumption: That there are no unobserved differences.

    But if individuals are otherwise identical, then why did some participate and others not?

- Requires good quality data. Need to match on as many characteristics as possible

- Requires sufficiently large sample size. Need a match for each participant in the treatment group

Matching is a useful way to control for OBSERVABLE heterogeneity. However, it requires relatively strong assumptions

TABLE 7

FFS: BALANCING TEST RESULTS FOR PPS METHODS

	Definition of Control Group		
	Method 1		Method 2
	Stratum 1	Stratum 2	
Education of household head	.42	.77	.46
Age of household head	.55	.15	.10
Number of family members	.58	.47	.23
Dependency rate	.27	.19	.20
Total land ownership	.31	.52	.41
Value of cattle assets	.82	.85	.38
Number of inherited livestock	.85	.63	.99
Value of household assets	.39	.74	.52
Value of farm assets	.30	.48	.31
Fraction of plots lost in El Niño	.76	.73	.31
Credit constrained	.45	.36	.88
Number of observations	22	22	45

Note. In method 1, the control for each participant is the average of the five nonparticipants with closest PPS (within .01 PPS) under common support. In method 2, the control is the kernel-weighted average of all nonparticipant farmers under common support.