Section Chapter 4: Impact Evaluation

1 Impact Evaluation Introduction

The main challenge for impact evaluation is establishing causality between the intervention X and an outcome Y. Let's take as an example an agricultural training program (X) meant to improve consumption (Y).

1.1 Correlation vs. Causation

Correlation of X and Y may mean:

- X causes Y (causality) : Great! Our program works and should be expanded!
- Y causes X (reverse causality): People who are well off are more likely to participate in the training program.
- X causes Y and Y causes X (simultaneity): Program has some effect on income, but also people who are well-off tend to participate more.
- Z causes Y and X (omitted variable bias): People with more education both have higher consumption and are more likely to participate in the program.

1.1.1 Examples of Common Complications:

- *Self-selection:* When participants are free to choose whether or not to join a program, we worry about self-selection as the source of reverse causality or omitted-variable problems.
- *Program Placement:* Even if participants don't have a choice, the government or NGO running the program usually has a choice about who participants, so worry that program placement can be a source of reverse causality or omitted-variable problems.

1.1.2 Relevance to Policy Decisions: In theory, policy changes imply or assume causal inference. For example, if we offer micro-loans to women, we assume women entrepreneurs are constrained by limited access to capital. If this is not the *cause* of low incomes among women - for example, if women entrepreneurs have plenty of capital but few good ways to invest it in productive businesses because the are prevented from participating in certain industries such as transportation or construction - then formulating a policy to offer micro-loans will be useless, i.e. we misunderstood the real cause of low income.

1.2 Counterfactual

To measure the impact of a program, we need to compare what happened with the program to what would have happened without the program (the **counterfactual**). But "what would have happened"

is a hypothetical question. At best, we can choose find a group that we think closely represents the hypothetical counterfactual. Each of the methods of impact evaluation that we will look at are based on different way of representing that counterfactual. **Treatment Effect:** In the most basic terms, the treatment effect is the difference in the outcome of interest between the "treated" group and the "counterfactual" (control) group.

Internal vs. External Validity:

- *Internal:* Is your causal inference valid in the context your are analyzing, i.e. did you screw up the analysis or make bad assumptions?
- *External:* Do your findings/conclusions/lessons learned apply to other cases, other sample, other countries, other groups of people, other time periods or are they only valid and relevant for the specific context you are examining?

Method	*Counterfactual*
Randomized Control Trial	Randomly assigned untreated control group
Differences in Differences (DD)	Projected outcome of treated group using time trend
Propensity Score Matching (PSM)	Group in population that is similar to treated group
Regression Discontinuity (RD)	Untreated group just across a threshold

2 Randomized Control Trial (RCT)

Practical Considerations	***Technical Considerations***
Cost of field experiment	Success of randomization - balanced baseline
Practicality of implementing the control group	Non-random attrition (ex: financial education)
Ethics of a control group	Controlling for spillovers (ex: deworming)
Length of the experiment	External validity (ex: SMS reminders to save)

Examples:

- Progressa: Conditional Cash Transfers in Mexico increased school continuation rate; Randomized at village level during program roll-out.
- Deworming in Kenya: School deworming treatments reduce absenteeism; Randomization by village during rollout; major spillovers discovered.
- Financial education in Peru: Provide business education among microfinance clients in Peru to improve businesses; major attrition due burden of coursework

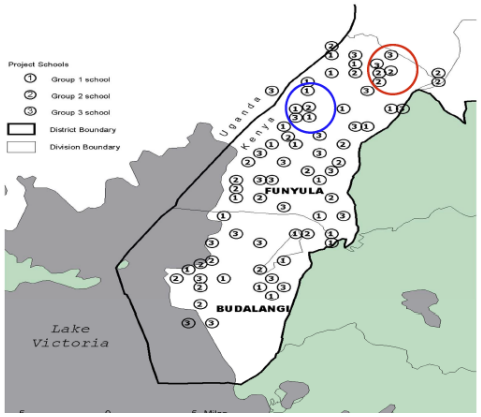


TABLE 6.—RESPONSE RATE BY THE FOLLOW-UP SURVEY BY LOCATION AND RETENTION IN FINCA

	Treatment	Control	Difference	*T*-statistic
Global	75.2	77.9	−2.7	−2.06
By Location				
Lima	77.2	83.5	−6.2	−2.85
Ayacucho	74.5	74.8	−0.3	−0.17
By retention in FINCA				
Clients	83.2	83.9	−0.6	−0.34
Ex-clients	69.9	74.2	−4.3	−2.44

(Sources: Ted Miguel, lecture notes Econ 270b, 2013; Karlan and Valdivia, 2011)

3 (Propensity Score) Matching

Matching is used ex-post when there are many untreated obs. from which we can find good comparison matches for the treatment, and treatment assignment was determined by observable characteristics.

3.1 Matching

Matching is a generalization of propensity score matching, which - as the name suggests - matches treated observations to control observations according to observable characterisitics that would not have been altered by the treatment (either fixed traits or characteristics before the treatment period). You may match "treated" female farmers to "control" female farmers of the same age within the same village only, i.e. matching on gender, age, and village. Matching requires making some very large and often implausible assumptions that everything other than the variables you are matching on are irrelevant (orthogonal) to the likelihood of the treatment.

3.2 Propensity Score Matching (PSM)

PSM is done in three steps. First, you use observable characteristics X of the treatment and control that would not have been altered by treatment to determine which characteristics predict assignment to treatment. Second, you assign each observation a propensity score - i.e the predicted likelihood of participation based on each observation's characteristics. Third, compare outcomes of each treatment observation to a control observation (or group of them) with the same p-score.

Step 1: Regress treatment status on observables:

$$P(Treat) \quad = \quad \alpha + \beta X_1 + \beta X_2 + \ldots + \beta X_k + \varepsilon$$

Step 2: Estimate P-score using coefficients from regression:

$$P - score \quad = \quad P(\hat{Treat}) = \alpha + \hat{\beta} X_1 + \hat{\beta} X_2 + \ldots + \hat{\beta} X_k$$

Step 3: Calculate treatment effect comparing obs. with similar p-scores

$$Impact \quad = \quad Treat - Control = \frac{1}{N} \sum_{i \in T} \left(Y_i - Y_{m(i)} \right)$$

The method applies best to programs that have only covered part of the potential population of beneficiaries, where selection was based on criteria completely independent of the potential impact of the program, and where personal choice had little to do in participation. Just like the more general method of matching, if treatment status is based on unobservable traits that have led to preferential program placement or self-selection, this method is no longer valid and will give biased estimates.

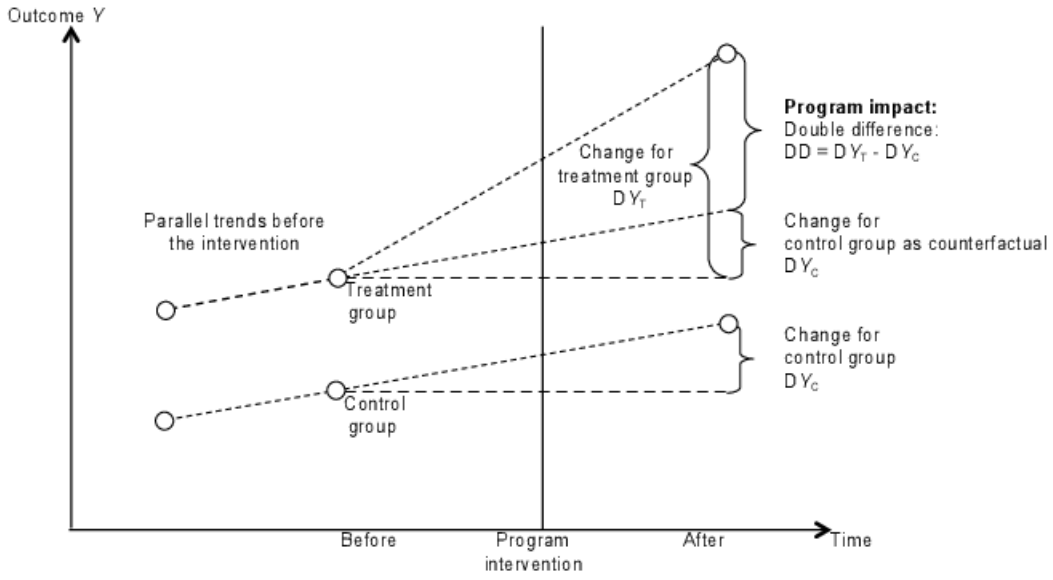
4 Difference in Difference: (dif-in-dif, DD)

This method is used when treatment and control groups are not perfectly comparable (imperfect randomization or imperfect matching) but we can observe changes over time (before and after) for both groups. A key assumption is that the difference in Y between "before" and "after" in comparison group is a good counterfactual for that same difference in the treatment group. The method is applied as follows:

- First compute difference in average outcome Y after (Y_1) and before (Y_0) for the control group (C) and get: $\bar{Y}_{C1} - \bar{Y}_{C0} = D\bar{Y}_C$
- Then, we do the same for the treated and get: $\bar{Y}_{T1} - \bar{Y}_{T0} = D\bar{Y}_T$
- Then the impact of the program will be the difference in differences:

$$Impact = \left(\bar{Y}_{T1} - \bar{Y}_{T0}\right) - \left(\bar{Y}_{C1} - \bar{Y}_{C0}\right) = D\bar{Y}_T - D\bar{Y}_C = DD$$

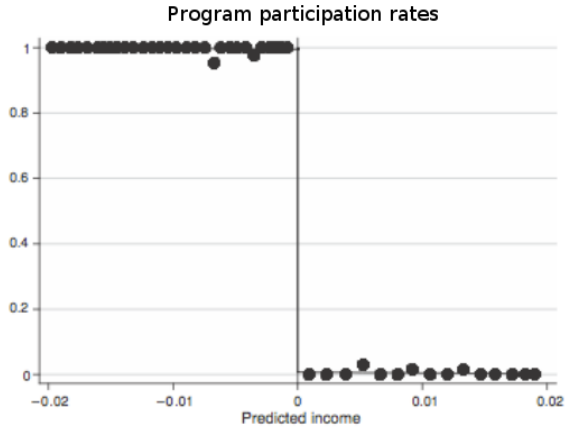
An advantage with this method is that the treatment and comparison groups do not have to start at the same *level*, since we are comparing the *differences* over time. A challenge, though, is that we need to provide evidence of *parallel trends* - that the treatment and comparison groups would likely have changed by the same amount over time if there had been no treatment. We generally do this by showing that changes in treatment and control were the same in the pre-period leading up to the baseline, meaning that we need data for at least two points in time before the treatment and at least one after. In Duflo's Indonesian school construction paper referenced in the chapter, there is only one cross-section of data, but the two "before" values for T (high-intensity construction) and C (low-intensity construction) come from education and wage data for the age cohort just barely too old to benefit from newly constructed schools and the age cohort just before that.

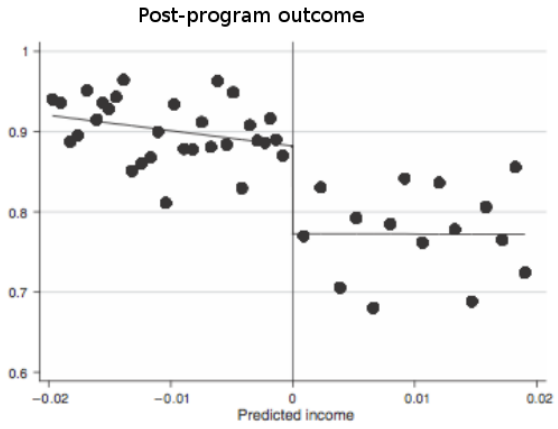


5 Discontinuity Design (regression discontinuity):

This method is used when treatment eligibility is based on a well-defined, continuous characteristic (for example, asset score, age in days, or distance from a geographic point), but the cutoff threshold is arbitrary, such that individuals just inside and just outside of the cutoff are statistically similar in all other observable and unobservable characteristics in the absence of treatment. Observations just excluded serve as comparison for those just included in the treatment. The key assumption for the validity of the method is that the outcome would be a continuous function of the indicator used for eligibility around the threshold, if it were not for the program.The treatment effect can be estimated in a number of ways, but the important thing to remember is that we are estimating the *local average treatment effect* (LATE) - the impact for the types of individuals around the threshold. While this may be useful for making decisions about program expansion (since those at the margin would be the next set group admitted), it doesn't let us estimate the average impact for all participants. Example:

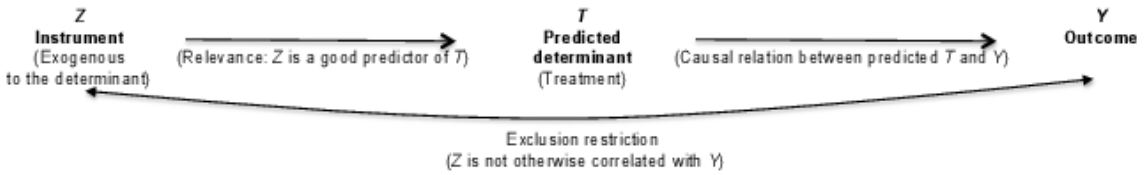
Do transfers to the poor buy political support in Uruguay (Manacorda et al., 2011)?





6 Instrumental Variables (IV):

The IV approach is particularly useful when you want to control for unobservables without a comparison group. Intuitively, the IV only takes advantage of the fraction of the sample of observations that are assigned to the "treatment" in a quasi-random way from the perspective of the outcome of interest and only takes advantage of the effect of the treatment that is due to this quasi-random assignment (again from the perspective of the outcome of interest). You can think of this as throwing out some of the potentially interesting variation between treatment and control in order to guarantee that the variation you do use to estimate any differences between treatment and control is valid and not the result of omitted variables, reverse causality, etc.



Three Essential Assumptions:

1. Relevance: Z is a sufficiently strong predictor of X
2. Exogenity: Z is exogenous to X
3. Exclusion Restriction: Z correlated with Y, but only through its influence on X, not independently of X

Examples:

- Weather effect income, which effects conflict using subtle random variations in weather
- Mortality rate of European settlers effects likelihood of robust European settlement and institution building, which effects future economic growth, but historic mortality rates do not affect future economic growth through any other mechanism

Review: Impact Evaluation Methods

Randomized Control Trial

Counterfactual: Randomly assigned untreated control group
Advantages: Gold standard for determining causality
Disadvantages: Ethical/logistical concerns; Treatment must be designed with IE in mind;
Key assumptions: Successful randomization, *no spillovers, no selective attrition*
Test for internal validity: Verify T and C not different in pre-treatment characteristics

Differences in Differences (DD)

Counterfactual: Projected outcome of treated group had it not been treated, based on an untreated group and assumption of constant differences across time or between eligibles and ineligibles
Advantages: Can be used with existing data *(ex-post facto)*
Disadvantages: Needs survey at both baseline and follow-up for differences in differences across time, plus additional pre-baseline data for testing trends
Key assumptions: Needs constant differences across time for treatment and control
Test for internal validity: Test for *parallel trends*

Propensity Score Matching (PSM)

Counterfactual: Untreated group in general population similar to treated group in all observable characteristics
Advantages: Can be used with existing data *(ex-post facto)*
Disadvantages: Differences in unobservable characteristics could still create bias
Key assumptions: Selection solely on observable characteristics, no self-selection
Test for internal validity: Treatment based on observed characteristics; arbitrary order of rollout to regions; Test balance in observed characteristics between T and C in various p-score ranges.

Regression Discontinuity (RD)

Counterfactual: Untreated group just across a threshold ("just barely ineligible")
Advantages: If exact placement of threshold is arbitrary, observations just above and just below should be identical in both observable and unobservable characteristics (i.e. as good as randomly assigned participation); Can be used with existing data *(ex-post facto)*
Disadvantages: Impact is only measured around the threshold (local average treatment effect); Only possible to use if participation in intervention is based on a threshold
Key assumptions: Threshold for treatment is arbitrary, not based on a natural discontinuity in characteristics

Test for internal validity: Generally show "smooth" distribution of eligibility characteristics right around the threshold value

Instrumental Variable (IV)

Counterfactual: Those predicted not to be treated by the instrument
Advantages: Controls for unobservabels without a true comparison group
Disadvantages: Impact is only measured for the fraction of the treatement predicted by the instrument (weighted local average treatment effect)
Key assumptions: Exclusion restriction, i.e. Z only affects Y through X
Test for internal validity: There is no good test. You often have to tell a convincing story about the exclusion restriction.