

Evaluating Anti-Poverty Programs

Part 1: Concepts and Methods

Martin Ravallion

Development Research Group, World Bank

- 1. *Introduction***
- 2. *The evaluation problem***
- 3. *Generic issues***
- 4. *Single difference: randomization***
- 5. *Single difference: matching***
- 6. *Single difference: exploiting program design***
- 7. *Double difference***
- 8. *Higher-order differencing***
- 9. *Instrumental variables***
- 10. *Learning more from evaluations***

1. Introduction

- Assigned programs
 - some units (individuals, households, villages) get the program;
 - some do not.
- Examples:
 - Social fund selects from applicants
 - Workfare: gains to workers and benefiting communities; others get nothing
 - Cash transfers to eligible households only
- *Ex-post* evaluation

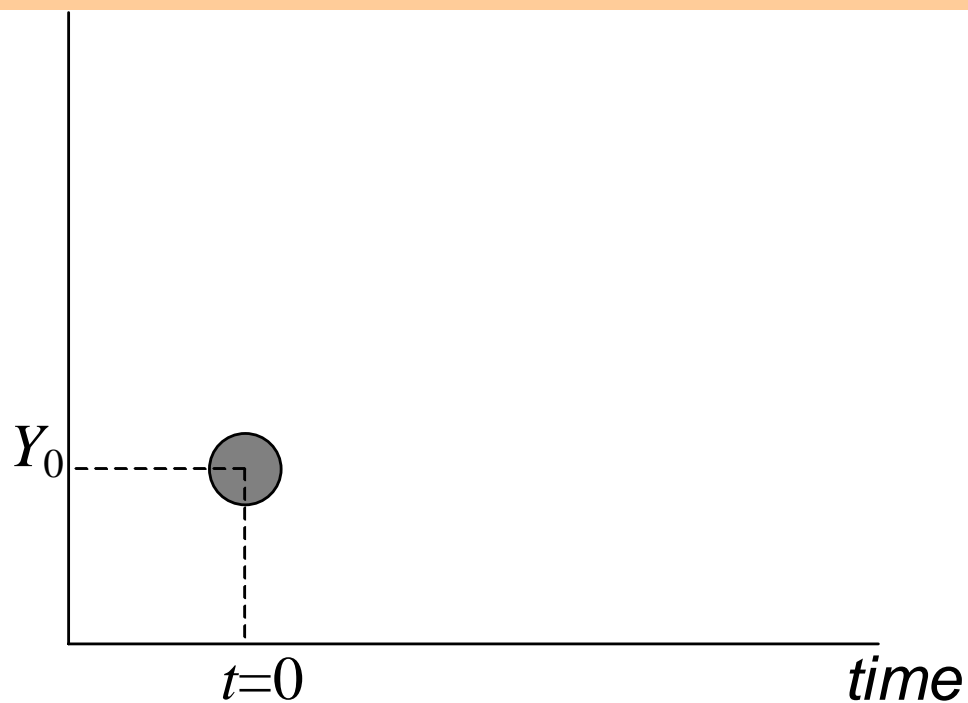
2. The evaluation problem

Impact is the difference between the relevant outcome indicator with the program and that without it.

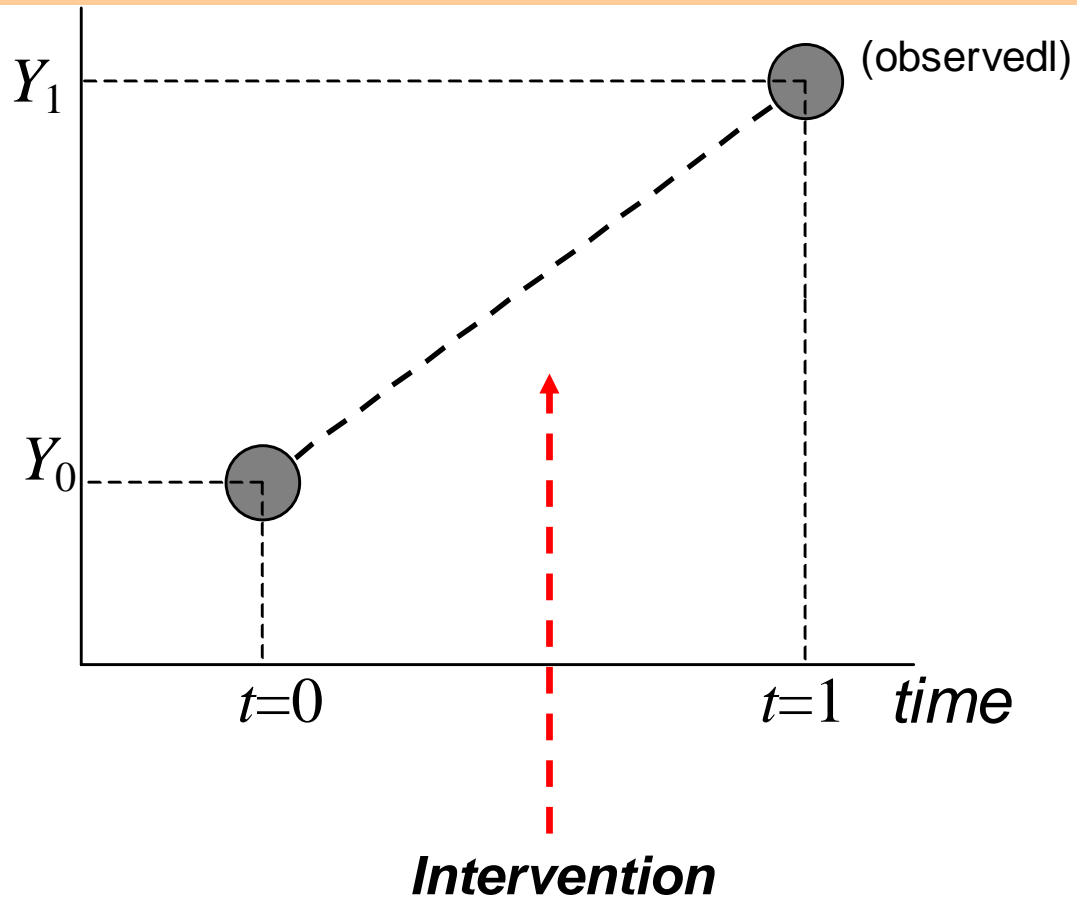
- However, we can never simultaneously observe someone in two different states of nature.
- While a post-intervention indicator is observed, its value in the absence of the program is not, i.e., it is a counterfactual.

So all evaluation is essentially a problem of missing data. Calls for counterfactual analysis.

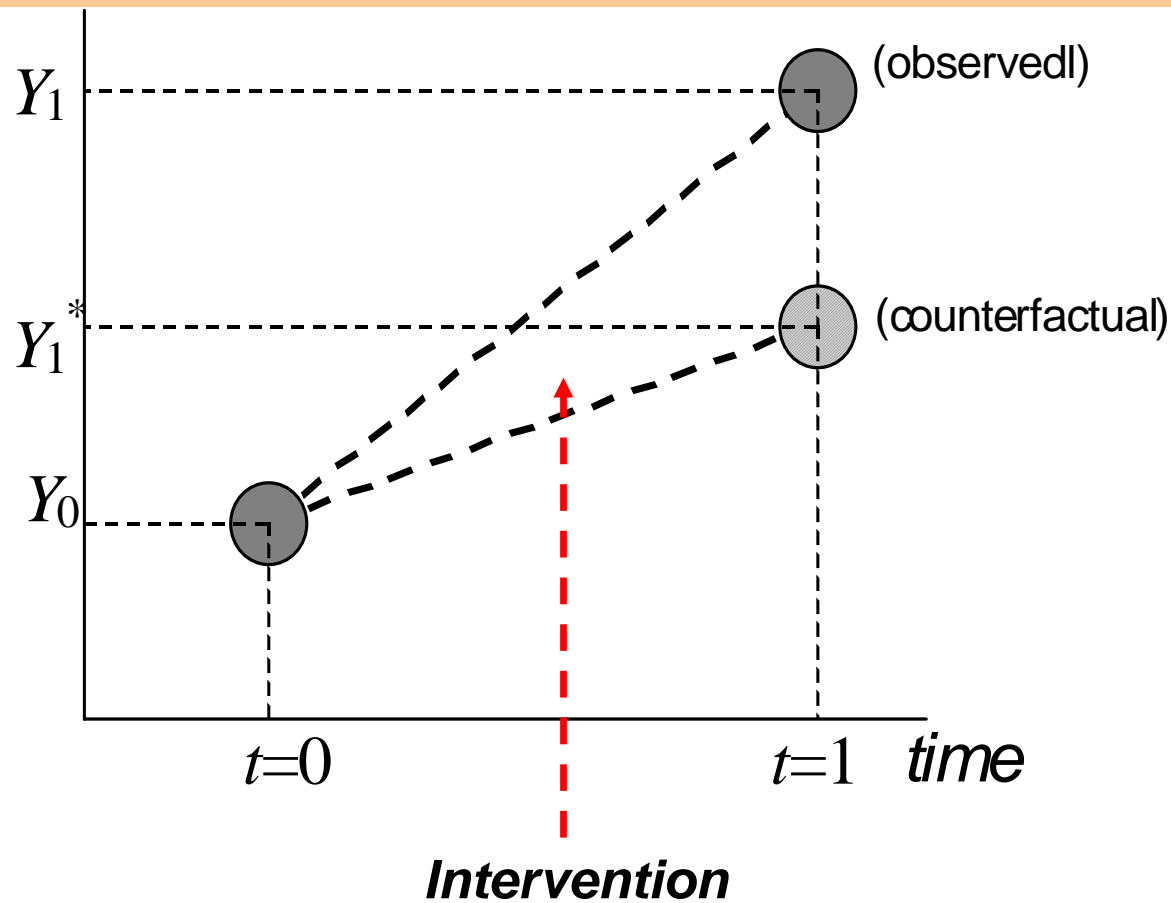
We observe an outcome indicator,



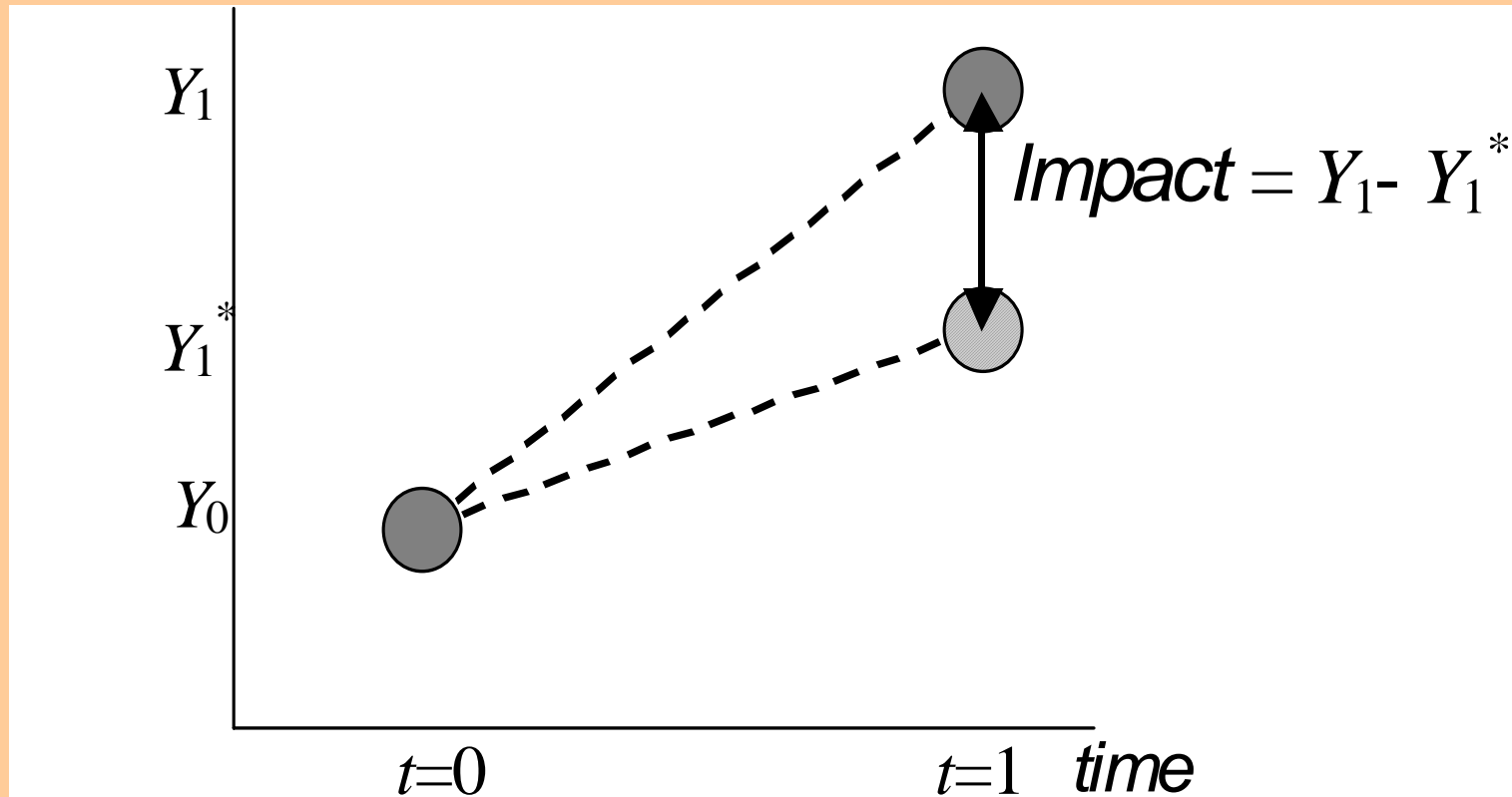
and its value rises after the program:



However, we need to identify the counterfactual...



... since only then can we determine the impact of the intervention



However, counterfactual analysis has not been the norm

- 78 “evaluations” by OED of WB projects since 1979 (Kapoor)
- Counterfactual analysis in only 21 cases
- For the rest, there is no way to know if the observed outcomes are in fact attributable to the project
- **We can do better!**

Archetypal formulation

Outcomes with and without treatment:

$$Y_i^T = X_i \beta^T + \mu_i^T \quad (i=1, \dots, n)$$

$$Y_i^C = X_i \beta^C + \mu_i^C \quad (i=1, \dots, n)$$

$$E(\mu_{0i} | X_i) = E(\mu_{1i} | X_i) = 0$$

Gain from the program: $G_i \equiv Y_i^T - Y_i^C$

Archetypal formulation

Outcomes with and without treatment:

$$Y_i^T = X_i \beta^T + \mu_i^T \quad (i=1, \dots, n)$$

$$Y_i^C = X_i \beta^C + \mu_i^C \quad (i=1, \dots, n)$$

$$E(\mu_{0i} | X_i) = E(\mu_{1i} | X_i) = 0$$

Gain from the program: $G_i \equiv Y_i^T - Y_i^C$

ATE: average treatment effect: $E(G_i)$

conditional ATE: $E(G_i | X_i) = X_i(\beta^T - \beta^C)$

ATET: ATE on the treated: $E(G_i | D_i = 1)$

conditional ATET:

$$E(G_i | X_i, D_i = 1) = X_i(\beta^T - \beta^C) + E(\mu_i^T - \mu_i^C | X_i, D_i = 1)$$

The evaluation problem

We cannot observe Y_i^C for $D_i = 1$ or Y_i^T for $D_i = 0$

How then can we estimate the following model?

$$Y_i^T = X_i\beta^T + \mu_i^T \text{ if } D_i = 1$$
$$Y_i^C = X_i\beta^C + \mu_i^C \text{ if } D_i = 0$$

Or the (equivalent) switching regression:

$$Y_i = D_i Y_i^T + (1 - D_i) Y_i^C = X_i \beta^C + X_i (\beta^T - \beta^C) D_i + \varepsilon_i$$

$$\varepsilon_i = D_i (\mu_i^T - \mu_i^C) + \mu_i^C$$

Common effects specification (only intercepts differ):

$$Y_i = (\beta_0^T - \beta_0^C) D_i + X_i \beta^C + \varepsilon_i$$

Alternative solutions

Experimental evaluation (“Social experiment”)

- Program is randomly assigned
- Rare for anti-poverty programs in practice

Non-experimental evaluation (“Quasi-experimental”; “observational studies”)

- Choose between two (non-nested) conditional independence assumptions:
 1. Exogeneous placement conditional on observables
 2. Instrumental variable that is independent of outcomes conditional on program placement and other relevant observables

3. Generic issues

- Selection bias
- Spillover effects
- Data and measurement errors

Selection bias in the outcome difference between participants and non-participants

Observed difference in mean outcomes between participants ($D=1$) and non-participants ($D=0$):

$$E(Y^T | D = 1) - E(Y^C | D = 0) =$$

$$E(Y^T | D = 1) - E(Y^C | D = 1)$$

ATET=average treatment effect on the treated

$$+ E(Y^C | D = 1) - E(Y^C | D = 0)$$

Selection bias=difference in mean outcomes for the comparison group

Sources of selection bias

- Selection on observables
 - Data
 - Linearity in controls?
- Selection on unobservables
 - Participants have latent attributes that yield higher/lower outcomes
- Cannot judge if exogeneity is plausible without knowing whether one has dealt adequately with observable heterogeneity
 - That depends on program, setting and data

Naïve comparisons can be deceptive

- Common practice: compare units (people, households, villages) with and without the anti-poverty program.
- Failure to control for differences in unit characteristics that influence program placement can severely bias such comparisons.

Impacts on poverty?

| | Without (n=56) | With (n=44) | % increase (t-test) |
|---------------|-------------------|----------------|------------------------|
| Case 1 | 43 | 80 | 87% (2.29) |
| | | | |

Percent not poor

Impacts on poverty?

| | Without (n=56) | With (n=44) | % increase (t-test) |
|---------------|-------------------|----------------|------------------------|
| Case 1 | 43 | 80 | 87% (2.29) |
| Case 2 | 43 | 66 | 54% (2.00) |

Percent not poor

| | Without (n=56) | With (n=44) | % increase (t-test) |
|---|-------------------|----------------|------------------------|
| 1: Program yields 20% gain | 43 | 80 | 87% (2.29) |
| 2: Program yields <u>no</u> gain | 43 | 66 | 54% (2.00) |

But even with controls...

OLS only gives consistent estimates under exogenous program placement

- there is no selection bias in placement, conditional on X , i.e.,

$$E(\mu_i^T - \mu_i^C | X_i, D_i = 1) = 0$$

- or (equivalently) that the conditional mean outcomes do not depend on treatment:

$$E[Y_i^C | X_i, D_i = 1] = E[Y_i^C | X_i, D_i = 0]$$

Spillover effects

- Hidden impacts for non-participants?
- Spillover effects can stem from:
 - Markets
 - Non-market behavior of participants/non-participants
 - Behavior of intervening agents (governmental/NGO)
- Example: Employment Guarantee Scheme
 - assigned program, but no valid comparison group.

Measurement and data

Poverty measurement:

- Reinterpret such that $Y=1$ if poor and $Y=0$ if not
- $E(G)$ =impact on headcount index of poverty

Data and measurement errors:

- Discrepancies with NAS
- Under-reporting; noncompliance bias

Under certain conditions unbiased ATE is still possible

- Additive error component common the T and C groups
- This needs to be uncorrelated with X for SD but not DD (later)

4. Randomization

“Randomized out” group reveals counterfactual.

- As long as the assignment is genuinely random, mean impact is revealed: $E(Y^C | D = 1) = E(Y^C | D = 0)$
- ATE is consistently estimated (nonparametrically) by the difference between sample mean outcomes of participants and non-participants.
- Pure randomization is the theoretical ideal for ATE, and the benchmark for non-experimental methods.

Examples for developing countries

- PROGRESA in Mexico
 - Conditional cash transfer scheme
 - 1/3 of the original 500 communities selected were retained as control; public access to data
 - Impacts on health, schooling, consumption
- *Proempleo* in Argentina
 - Wage subsidy + training
 - Wage subsidy: Impacts on employment, but not incomes
 - Training: no impacts though selective compliance

Lessons from practice 1

Ethical objections and political sensitivities

- Deliberately denying a program to those who need it
- And providing the program to some who do not
- Yes, too few resources to go around
- But since when is randomization the fairest solution to limited resources?
- Intention-to-treat helps alleviate these concerns
=> randomize assignment, but free to not participate
- But even then many in the randomized out group may be in great need

=> Constraints on design

- Sub-optimal timing of randomization
- Selective attrition + higher costs

Lessons from practice 2

Internal validity: Selective compliance

- Some of those assigned the program choose not to participate.
- Impacts may only appear if one corrects for selective take-up.
- Randomized assignment as IV for participation
- *Proempleo example*: impacts of training only appear if one corrects for selective take-up

Lessons from practice 3

External validity: inference for scaling up

- Systematic differences between characteristics of people normally attracted to a program and those randomly assigned (“randomization bias”: Heckman-Smith)
- One ends up evaluating a different program to the one actually implemented
- Difficult in extrapolating results from a pilot experiment to the whole population

5. Matching

Matched comparators identify counterfactual

- Match participants to non-participants from a larger survey.
- The matches are chosen on the basis of similarities in observed characteristics.
- This assumes no selection bias based on unobservable heterogeneity.

Propensity-score matching (PSM)

Match on the probability of participation.

- Ideally we would match on the entire vector X of observed characteristics. However, this is practically impossible. X could be huge.
- Rosenbaum and Rubin: match on the basis of the ***propensity score*** =

$$P(X_i) = \Pr(D_i = 1 | X_i)$$

- This assumes that participation is independent of outcomes given X . If no bias give X then no bias given $P(X)$.

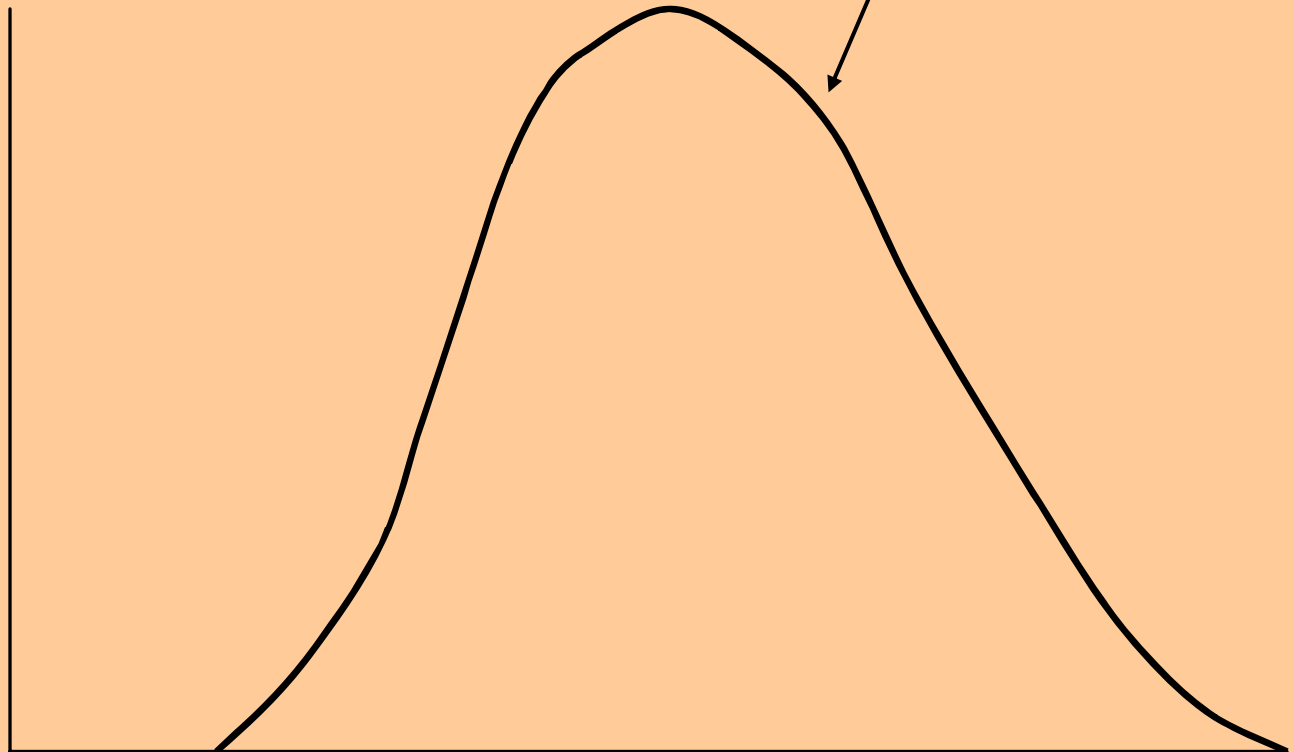
Steps in score matching:

- 1:** Representative, highly comparable, surveys of the non-participants and participants.
- 2:** Pool the two samples and estimate a logit (or probit) model of program participation.
Predicted values are the “propensity scores”.
- 3:** Restrict samples to assure common support

Failure of common support is an important source of bias in observational studies (Heckman et al.)

Density

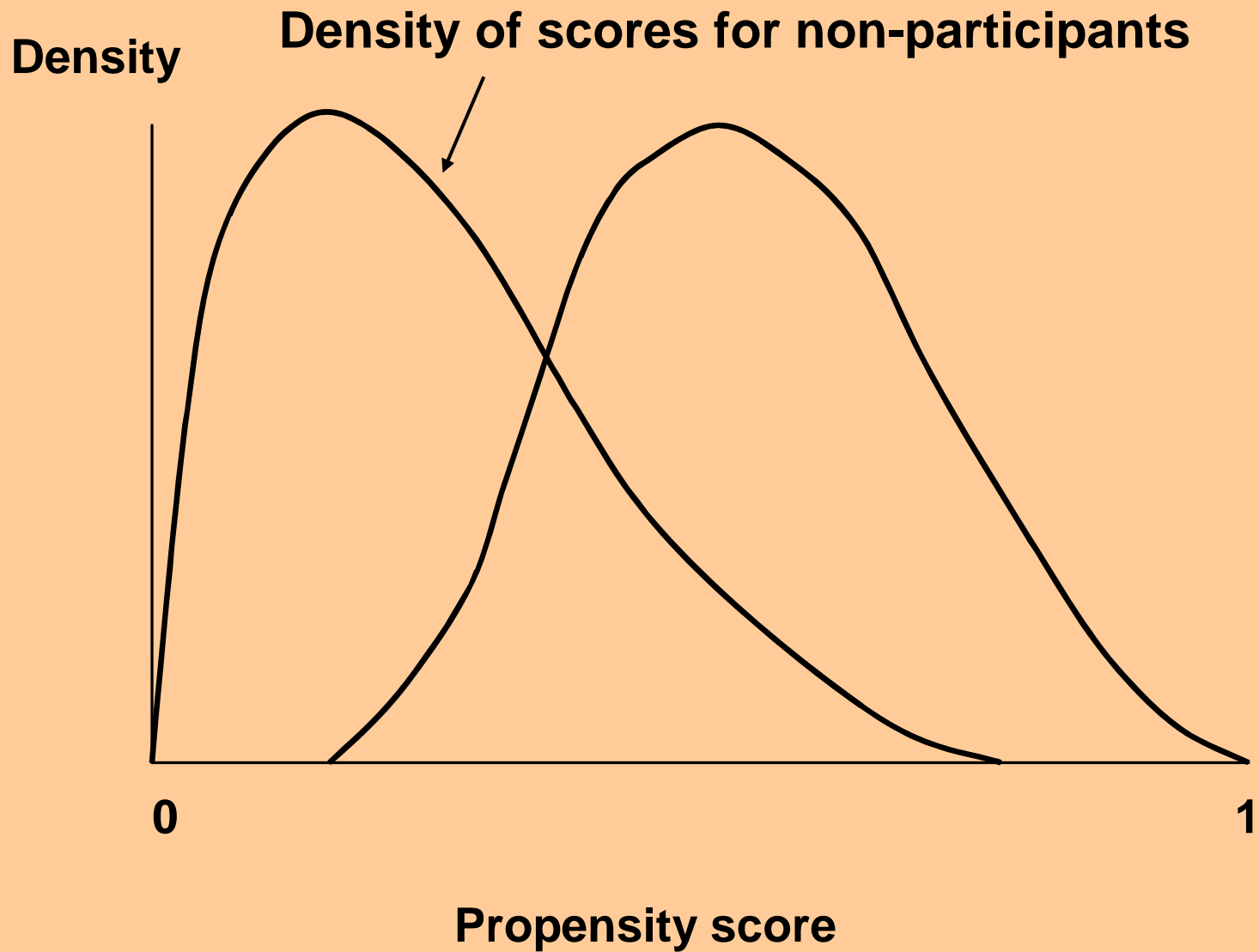
Density of scores for participants

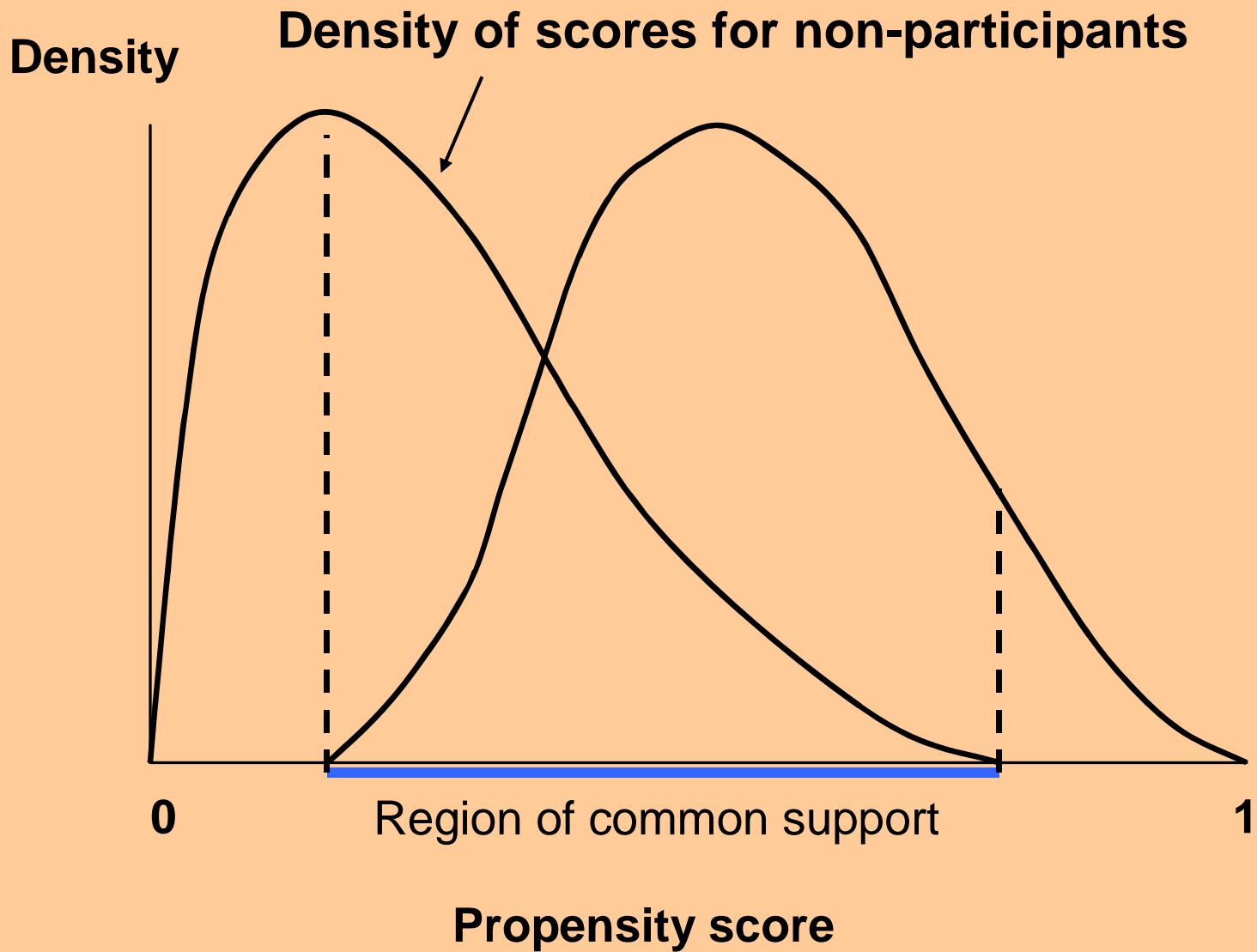


0

1

Propensity score






- 5:** For each participant find a sample of non-participants that have similar propensity scores.
- 6:** Compare the outcome indicators. The difference is the estimate of the gain due to the program for that observation.
- 7:** Calculate the mean of these individual gains to obtain the average overall gain. Various weighting schemes.

The mean impact estimator

$$\bar{G} = \sum_{j=1}^P (Y_{j1} - \sum_{i=1}^{NP} W_{ij} Y_{ij0}) / P$$

Various weighting schemes:

 Nearest k neighbors

 Kernel-weights (Heckman et al.):

$$K_{ij} = K[P(X_i) - P(X_j)]$$

$$W_{ij} = K_{ij} / \sum_{j=1}^P K_{ij}$$

How does PSM compare to an experiment?

- PSM is the observational analogue of an experiment in which placement is independent of outcomes
- The difference is that a pure experiment does not require the untestable assumption of independence conditional on observables.
- Thus PSM requires good data.
- Example of Argentina's *Trabajar* program
 - Plausible estimates using SD matching on good data
 - Implausible estimates using weaker data

How does PSM perform relative to other methods?

- In comparisons with results of a randomized experiment on a US training program, PSM gave a good approximation (Heckman et al.; Dehejia and Wahba)
- Better than the non-experimental regression-based methods studied by Lalonde for the same program.
- However, robustness has been questioned (Smith and Todd)

Lessons on matching methods

- When neither randomization nor a baseline survey are feasible, careful matching is crucial to control for observable heterogeneity.
- Validity of matching methods depends heavily on data quality. Highly comparable surveys; similar economic environment
- Common support can be a problem (esp., if treatment units are lost).
- Look for heterogeneity in impact; average impact may hide important differences in the characteristics of those who gain or lose from the intervention.

6. Exploiting program design 1

Discontinuity designs

- Participate if score $M < m$
- Impact=

$$E(Y_i^T | M_i = m - \varepsilon) - E(Y_i^C | M_i = m + \varepsilon)$$

- Key identifying assumption: no discontinuity in counterfactual outcomes at m

Exploiting program design 2

Pipeline comparisons

- Applicants who have not yet received program form the comparison group
- Assumes exogeneous assignment amongst applicants
- Reflects latent selection into the program

Lessons from practice

- Know your program well: Program design features can be very useful for identifying impact.
- But what if you end up changing the program to identify impact? You have evaluated something else!

7. Difference-in-difference

Observed changes over time for non-participants provide the counterfactual for participants.

Steps:

1. Collect baseline data on non-participants and (probable) participants before the program.
2. Compare with data after the program.
3. Subtract the two differences, or use a regression with a dummy variable for participant.

This allows for selection bias but it must be time-invariant and additive.

Outcome indicator: $Y_{it}^T = Y_{it}^* + G_{it}$

where

G_{it} = *impact* (“gain”);

Y_{it}^* = *counterfactual*;

Y_{it}^C = *comparison group*

Diff-in-diff: $E[\Delta(Y_{it}^T - Y_{it}^C)] = G_{it}$

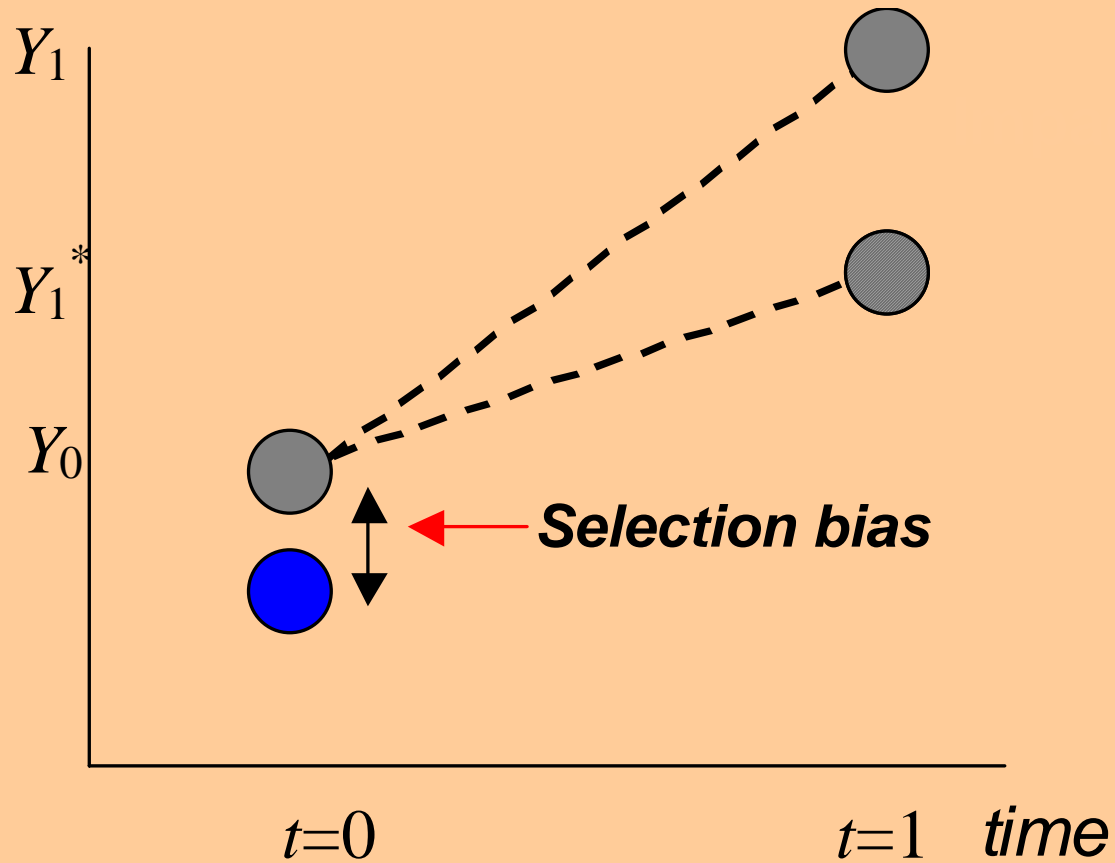
if (i) change over time for comparison group reveals counterfactual

$$E\Delta Y_{it}^C = E\Delta Y_{it}^*$$

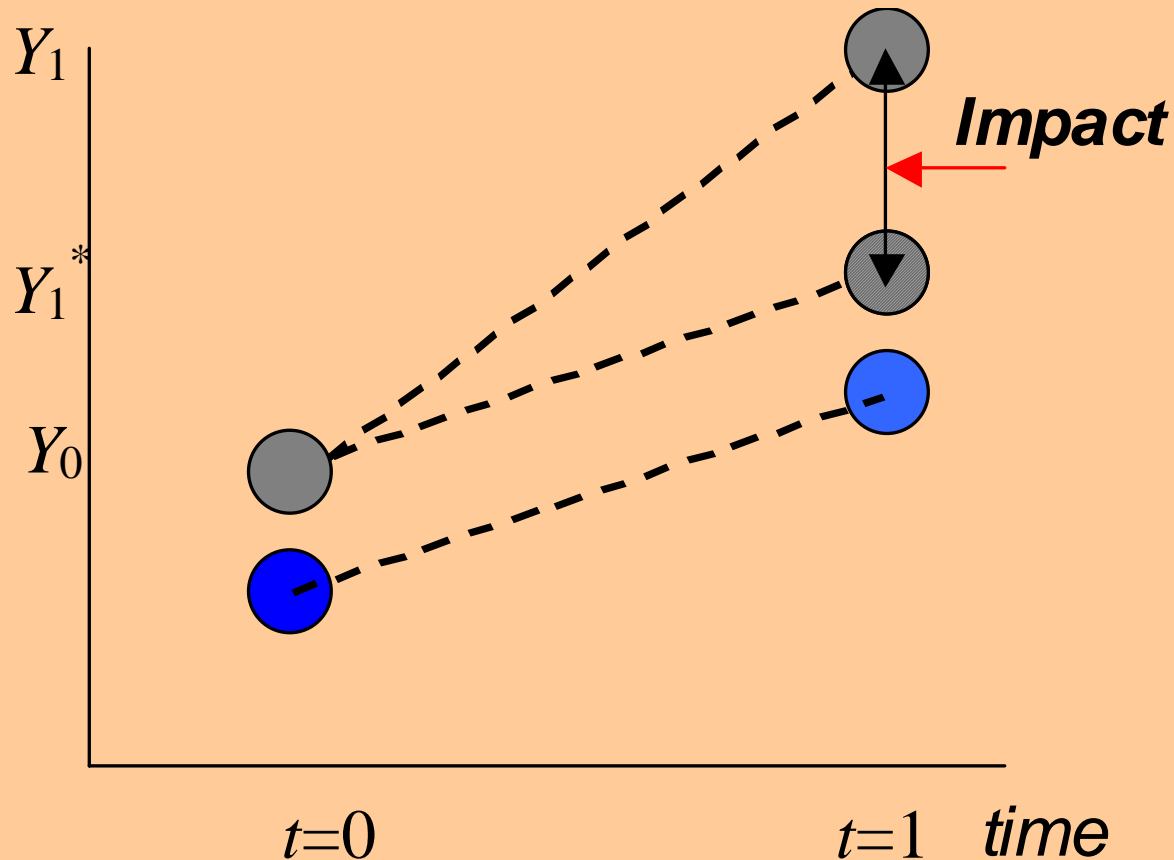
and (ii) *baseline is uncontaminated by the program,*

$$G_{i0} = 0$$

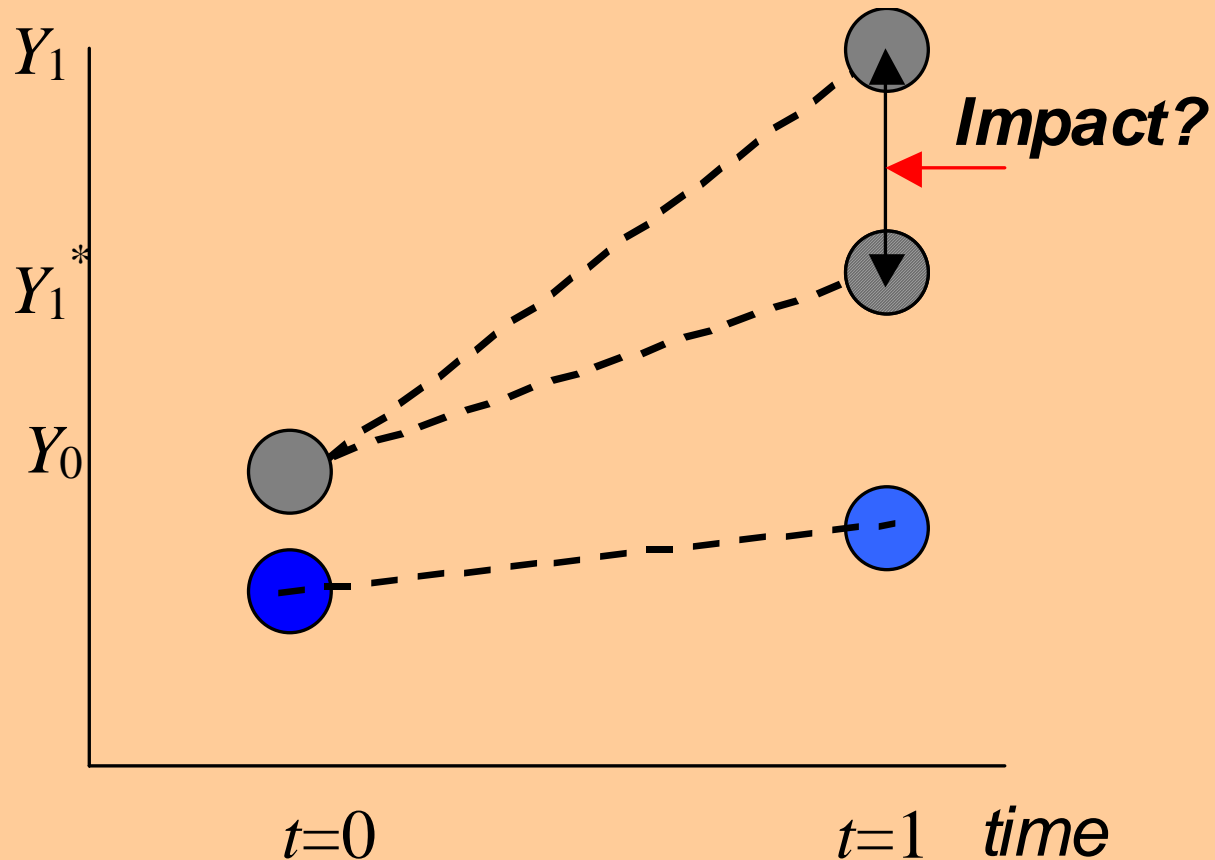
Selection bias



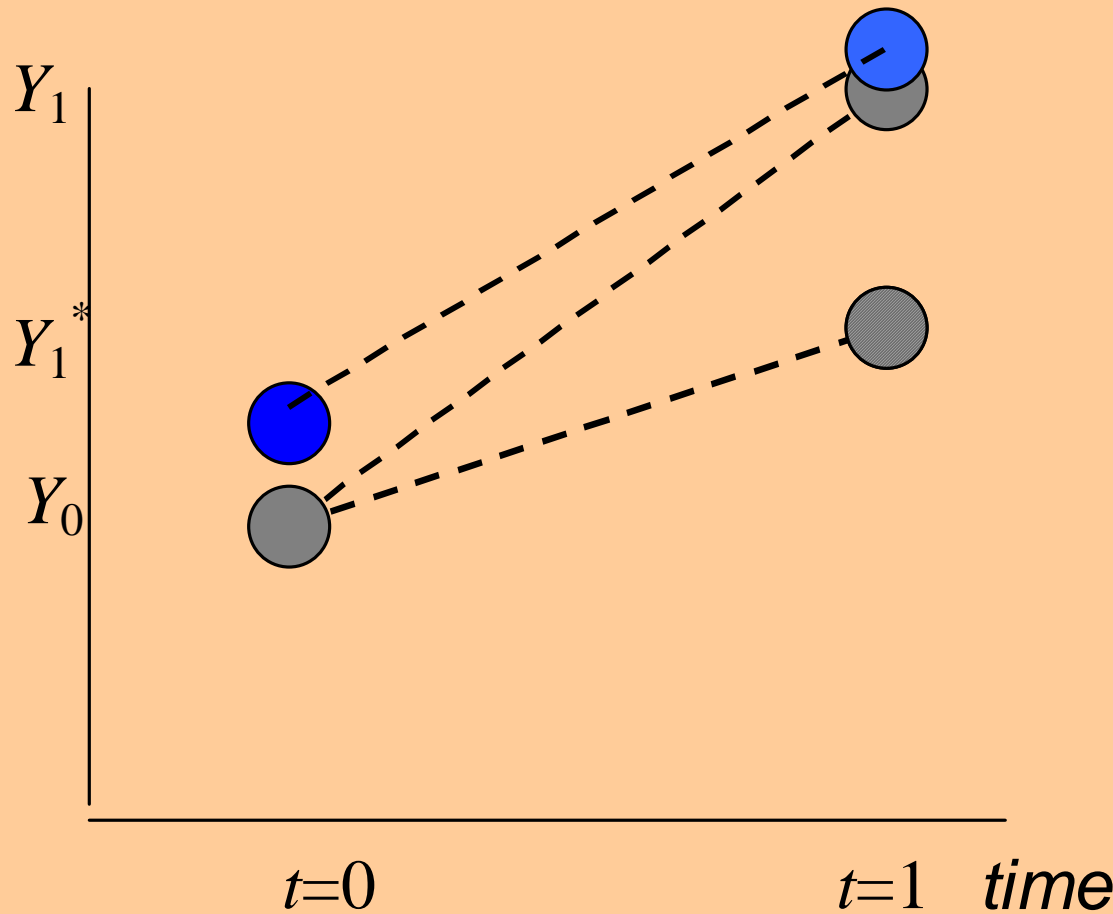
Diff-in-diff requires that the bias is additive and time-invariant



The method fails if the comparison group is on a different trajectory

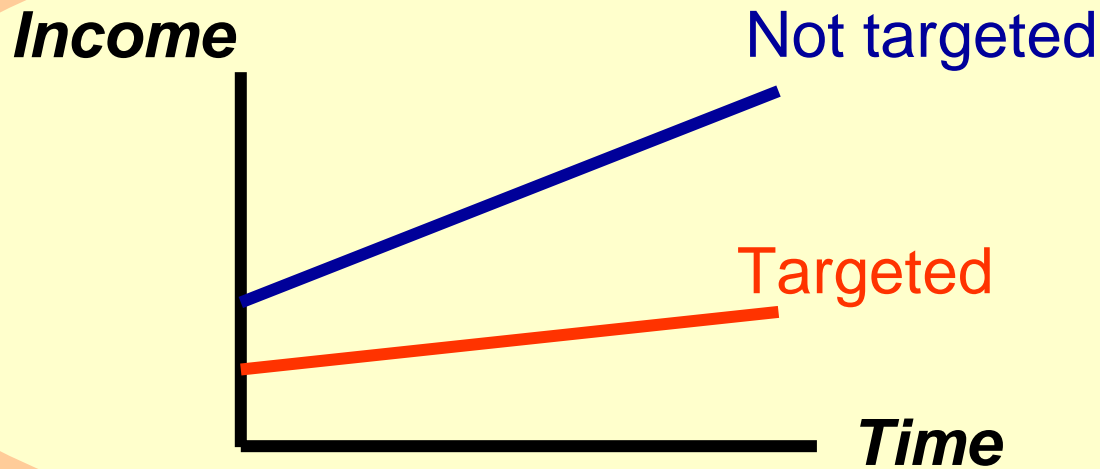


Or...



China: targeted poor areas have intrinsically lower growth rates (Jalan and Ravallion)

Poor area programs: areas not targeted yield a biased counter-factual



- The growth process in non-treatment areas is not indicative of what would have happened in the targeted areas without the program
- Example from China (Jalan and Ravallion)

Matched double difference

Matching helps control for time-varying selection bias

- Score match participants and non-participants based on observed characteristics in baseline
- Then do a double difference
- This deals with observable heterogeneity in initial conditions that can influence subsequent changes over time

Lessons from practice

- Single-difference matching can be severely contaminated by selection bias
 - Latent heterogeneity in factors relevant to participation
- Tracking individuals over time allows a double difference
 - This eliminates all time-invariant additive selection bias
- Combining double difference with matching:
 - This allows us to eliminate observable heterogeneity in factors relevant to subsequent changes over time

8. Higher-order differencing

Pre-intervention baseline data unavailable

e.g., safety net intervention in response to a crisis

Can impact be inferred by observing participants outcomes in the absence of the program after the program?

New issues

- Selection bias from two sources:
 1. decision to join the program
 2. decision to stay or drop out
- There are observed and unobserved characteristics that affect both participation and income in the absence of the program
- *Past* participation can bring *current* gains for those who leave the program

Double-Matched Triple Difference

- ☰ Match participants with a comparison group of non-participants
- ☰ Match leavers and stayers
- ☰ Compare gains to continuing participants with those who drop out
- ☰ Ravallion et al.

$$\textit{Triple Difference (DDD) = DD for stayers - DD for leavers}$$

Outcomes for participants: $Y_{it}^T = Y_{it}^* + G_{it}$

Single difference: $E[Y_{it}^T - Y_{it}^C]$

Double difference: $E[\Delta(Y_{it}^T - Y_{it}^C)] = \Delta G_{it}$

Triple difference:

$$E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 1] - E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 0]$$

“stayers”
in period 2

“leavers”
in period 2

$$E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 1] - E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 0] =$$

$$[E(G_{i2} | D_{i2} = 1) - E(G_{i2} | D_{i2} = 0)] \quad \text{net gain from participation}$$

$$- [E(G_{i1} | D_{i2} = 1) - E(G_{i1} | D_{i2} = 0)] \quad \text{selection bias}$$

$$E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 1] - E[\Delta(Y_{i2}^T - Y_{i2}^C) | D_{i2} = 0] =$$

$$[E(G_{i2} | D_{i2} = 1) - E(G_{i2} | D_{i2} = 0)] \quad \text{net gain from participation}$$

$$- [E(G_{i1} | D_{i2} = 1) - E(G_{i1} | D_{i2} = 0)] \quad \text{selection bias}$$

Joint conditions for DDD to estimate impact:

- no current gain to ex-participants; $E(G_{i2} | D_{i2} = 0) = 0$
- no selection bias in who leaves the program; $E(G_{i1} | D_{i2} = 1) = E(G_{i1} | D_{i2} = 0)$

Sign of the selection bias? If leavers have lower gains then DDD underestimates impact

Test for whether DDD identifies gain to current participants

Third round of data allows a test: **mean gains in round 2 should be the same whether or not one drops out in round 3**

$$DDD = E(G_{i2} | D_{i2} = 1, D_{i3} = 1) = E(G_{i2} | D_{i2} = 1, D_{i3} = 0)$$

Gain in round 2 for
stayers in round 3

Gain in round 2 for
leavers in round 3

Lessons from practice

1. Tracking individuals over time:
 - addresses some of the limitations of single-difference on weak data
 - allows us to study the dynamics of recovery
2. “Baseline” can be after the program, but must address the extra sources of selection bias
3. Single difference for leavers vs. stayers can if exogeneous program contraction

9. Instrumental variables

Identifying exogenous variation using a 3rd variable

Outcome regression: $Y_i = \beta D_i + \varepsilon_i$

$D = 0,1$ is our program – not random

- “Instrument” (Z) influences participation, but does not affect outcomes given participation (the “exclusion restriction”).
- This identifies the exogenous variation in outcomes due to the program.

Treatment regression: $D_i = \gamma Z_i + u_i$

Reduced-form outcome regression:

$$Y_i = \beta(\gamma Z_i + u_i) + \varepsilon_i = \pi Z_i + v_i$$

where $\pi = \beta\gamma$ and $v_i = \beta u_i + \varepsilon_i$

Instrumental variables (two-stage least squares) estimator of impact:

$$\hat{\beta}_{IVE} = \hat{\pi}_{OLS} / \hat{\gamma}_{OLS}$$

IVE is only a 'local' effect

- IVE identifies the effect for those induced to switch by the instrument (“local average effect”)
- Suppose Z takes 2 values. Then the effect of the program is:

$$\beta_{IVE} = \frac{E(Y | Z = 1) - E(Y | Z = 0)}{E(D | Z = 1) - E(D | Z = 0)}$$

- Care in extrapolating to the whole population

Valid instruments can be difficult to find; exclusion restrictions are often questionable.

Sources of instrumental variables

- Partially randomized designs as a source of IVs
- Non-experimental sources of IVs
 - Geography of program placement (Attanasio and Vera-Hernandez)
 - Political characteristics (Besley and Case; Paxson and Schady)
 - Discontinuities in survey design

Endogenous compliance: Instrumental variables estimator

$D = 1$ if treated, 0 if control

$Z = 1$ if assigned to treatment, 0 if not.

$$D_i = Z_i\pi_1 + \eta_{1i} \quad \textbf{Compliance regression}$$

$$Y_i = Z_i\pi_2 + \eta_{2i} \quad \textbf{Outcome regression}$$

(“intention to treat effect”)

$$\frac{\hat{\pi}_2}{\hat{\pi}_1}$$

2SLS estimator (=ITT deflated by compliance rate)

Lessons from practice

Partially randomized designs offer great source of IVs

The bar has risen in standards for non-experimental IVE

- Past exclusion restrictions often questionable in developing country settings
- However, defensible options remain in practice, often motivated by theory and/or other data sources

10. Learning from evaluations

Can the lessons be scaled up?

What determines impact?

Is the evaluation answering the relevant policy questions?

Scaling up?

- Contextual factors
 - Example of Bangladesh's Food-for-Education program
 - Same program works well in one village, but fails hopelessly nearby
- Institutional context => impact; *"in certain settings anything works, in others everything fails"*
- Partial equilibrium assumptions are fine for a pilot but not when scaled up
 - PE greatly overestimates impact of tuition subsidy once relative wages adjust (Heckman)

What determines impact?

- Replication across differing contexts
 - Example of Bangladesh's FFE: inequality etc within village => outcomes of program
- Intermediate indicators
 - Example of China's SWPRP
 - Small impact on consumption poverty
 - But large share of gains were saved
- Qualitative research/mixed methods
 - Test the assumptions ("theory-based evaluation")
 - But poor substitute for assessing impacts on final outcome

Policy-relevant questions?

- Choice of counterfactual
- Policy-relevant parameters?
 - Mean vs. poverty (marginal distribution)
 - Average vs marginal impact
 - Joint distribution of Y^T and Y^C (Heckman et al.), esp., if some participants may be worse off: ATE only gives net gain for participants
- “Black box” vs. Structural parameters
 - Simulate changes in program design
 - Example of *PROGRESA* (Attanasio et al.)
 - Modeling schooling choices using randomized assignment for identification
 - Budget-neutral switch from primary to secondary subsidy would increase impact