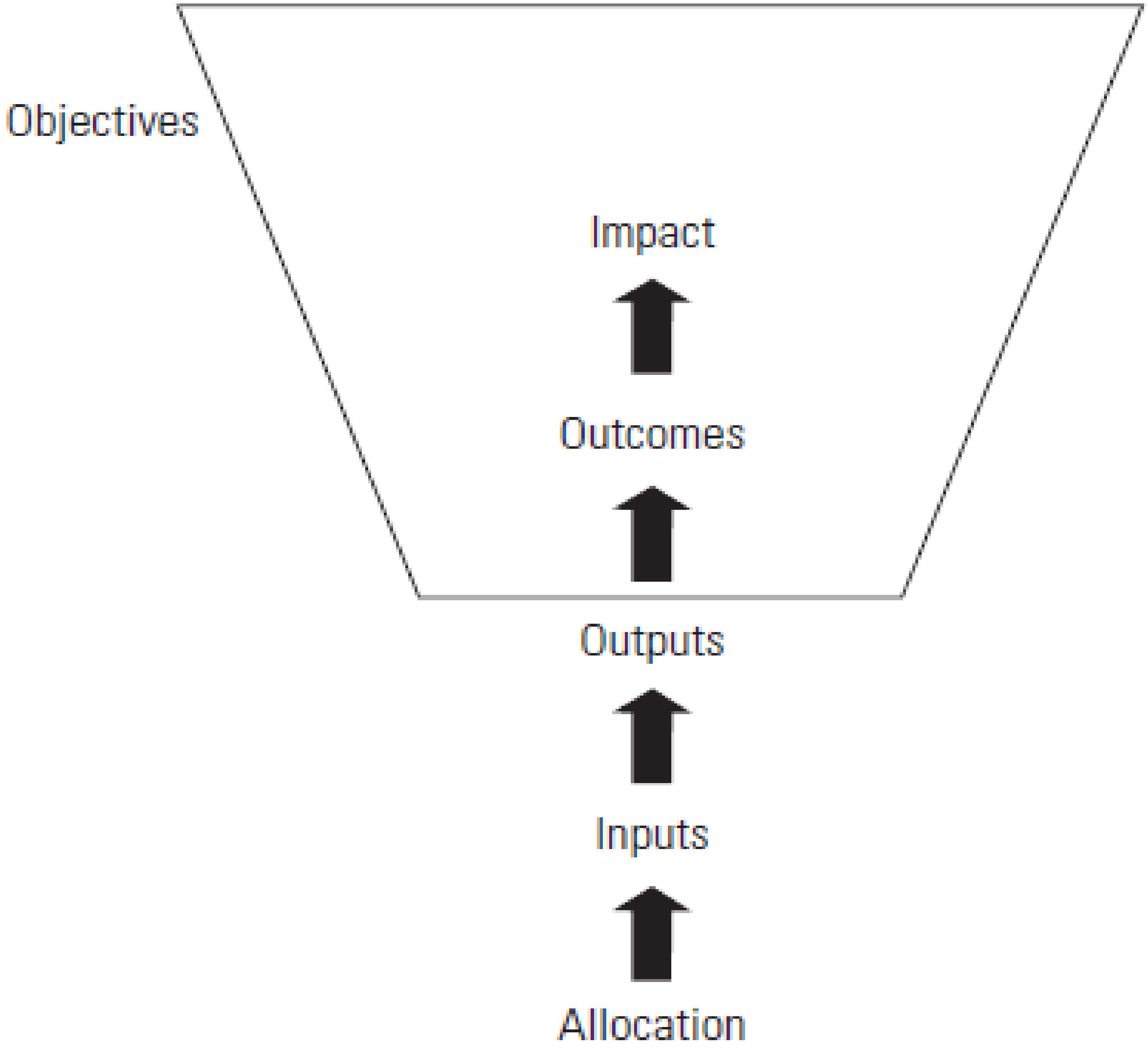
CHAPTER 5: IMPACT EVALUATION: SOME BASIC CONCEPTS

INTRODUCTION

- Identifying the precise effects of a policy is a complex and challenging task.
- Programs might appear potentially promising before implementation yet fail to generate expected impacts or benefits.
- The obvious need for impact evaluation is to help policy makers decide:
  - whether programs are generating intended effects;
  - to promote accountability in the allocation of resources across public programs; and
  - to fill gaps in understanding what works, what does not, and how measured changes in well-being are attributable to a particular project or policy intervention.

Monitoring and Evaluation framework



CONT'D

- The question of *causality* makes impact evaluation different from M&E and other evaluation approaches.
- In the absence of data on counterfactual outcomes (i.e., *outcomes for participants had they not been exposed to the program*), impact evaluations can be rigorous in identifying program effects by applying different models to survey data to construct comparison groups for participants.
- The main question of impact evaluation is one of attribution—*isolating the effect of the program from other factors and potential selection bias*.

The Problem of the Counterfactual

- The main challenge of an impact evaluation is to determine what would have happened to the beneficiaries if the program had not existed.
- Example, one has to determine the per capita household income of beneficiaries in the absence of the intervention.
- A beneficiary’s outcome in the absence of the intervention would be its *counterfactual*.
- Example, a program intend to improve income of beneficiaries. Suppose his/here income changes.
- The counterfactual is:
  - Does this change relate directly to the intervention?
  - Has this intervention caused expenditure or employment to grow? Not necessarily.
- In fact, with only a point observation after treatment, it is impossible to reach a conclusion about the impact.
- At best one can say whether the objective of the intervention was met.
- But the result after the intervention cannot be attributed to the program itself.

CONT'D

- The problem of evaluation is that while the program's impact (independent of other factors) can truly be assessed only by comparing actual and counterfactual outcomes, the counterfactual is not observed.
- So the challenge of an impact assessment is to create a convincing and reasonable comparison group for beneficiaries in light of this missing data.
- Ideally, one would like to compare how the same household or individual would have fared with and without an intervention or "treatment."
- But one cannot do so because at a given point in time a household or an individual cannot have two simultaneous existences, i.e., a household or an individual cannot be in the treated and the control groups at the same time.

COUNTERFACTUAL CONT'D

- Finding an appropriate counterfactual constitutes the main challenge of an impact evaluation.
  - How about a comparison between treated and non-treated groups when both are eligible to be treated?
  - How about a comparison of outcomes of treated groups before and after they are treated?
- These potential comparison groups can be "counterfeit" counterfactuals.

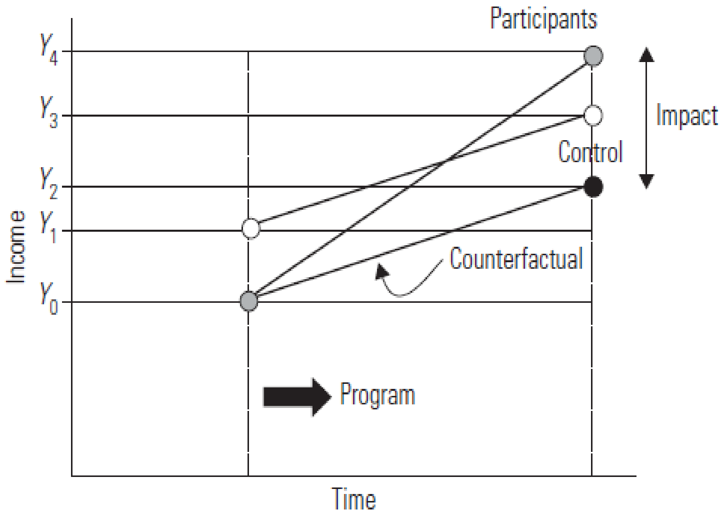
Looking for a Counterfactual: With-and-Without Comparisons

- Consider the case of Grameen Bank's beneficiaries in Bangladesh.
- The Bank offers credit to poor women to improve their food consumption.
- Data, however, show that the per capita consumption among program participants is lower than that of nonparticipants prior to program intervention.
  - Is this a case of failure of Grameen Bank? Not necessarily.
- Grameen Bank targeted poor families because they had lower per capita food consumption to begin with, so judging the program's impact by comparing the food consumption of program participants with that of nonparticipants is incorrect.

CONT'D

- What is needed is to compare what would have happened to the food consumption of the participating women had the program not existed.
- A proper comparison group that is a close counterfactual of program beneficiaries is needed.

Case 1: Evaluation Using a With-and-Without Comparison

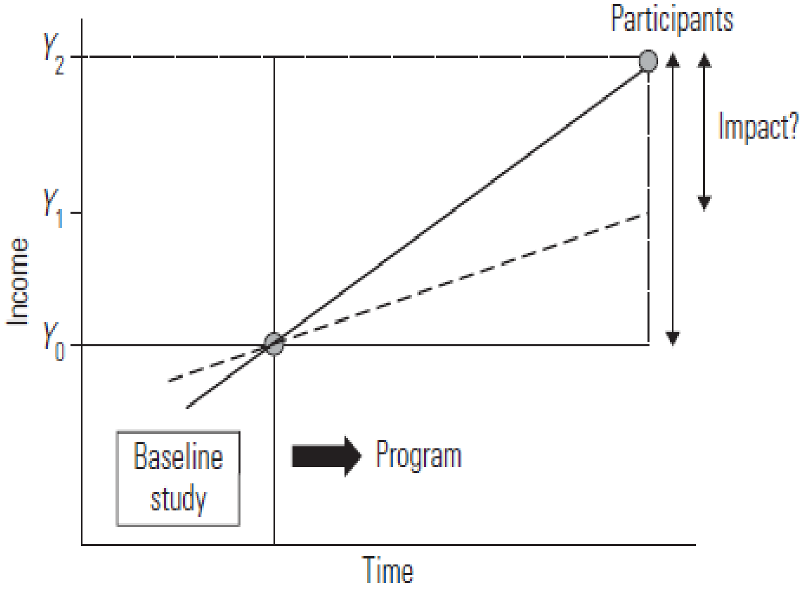


- Suppose:
  - Income of:
  - Participants is Y_4
  - Non-participants is Y_3
  - Program effects: Y_4-Y_3
- Is this measure a right estimate of program effect? Selection to the program may matter.

If we know the counterfactual (Y_0, Y_1): the true impact of the program is: $Y_4 - Y_2$, not Y_4- Y_3

CASE 2: LOOKING FOR A COUNTERFACTUAL: BEFORE-AND-AFTER COMPARISONS

- Another counterfeit counterfactual could be a comparison between the pre- and post program outcomes of participants.
- Compare data before and after or use retrospective data.



- Pre-intervention income is Y_0 and post intervention income is $Y_{2.}$
- The program's effect might be estimated as $(Y_2 - Y_0)$.
- *This is called reflexive method* of impact.
- But this may not be the true impact of the project.
- It will be $Y_2 - Y_1$.

5.1. Impact assessment basics

- Several approaches can be used to evaluate programs.
  - *Monitoring* tracks key indicators of progress over the course of a program as a basis on which to evaluate outcomes of the intervention.
  - *Operational evaluation* examines how effectively programs were implemented and whether there are gaps between planned and realized outcomes.
  - *Impact evaluation* studies whether the changes in well-being are indeed due to the program intervention and not to other factors.

Cont'd

- Impact evaluation could involve:
  - Qualitative and quantitative methods,
  - ex ante and ex post methods.
- **Quantitative results:** involves survey data collection or simulations before or after a program is introduced.
  - It can be generalizable, the qualitative results may not be.
- **Qualitative analysis**: seeks to gauge potential impacts that the program may generate, the mechanisms of such impacts, and the extent of benefits to recipients from in-depth and group-based interviews.
  - qualitative methods generate information that may be critical for understanding the mechanisms through which the program helps beneficiaries.

Quantitative Impact Assessment

- **Quantitative impact assessment could be conducted:**
  - Ex Post or
  - Ex Ante Impact Evaluations
- **An ex ante impact evaluation** attempts to measure the intended impacts of future programs and policies, given a potentially targeted area's current situation, and may involve simulations based on assumptions about how the economy works.
  - Many times, ex ante evaluations are based on structural models of the economic environment facing potential participants.
  - It predicts program impacts using data before the program intervention.
- **Ex post evaluations**: measure actual impacts accrued by the beneficiaries that are attributable to program intervention.
  - It examines outcomes after programs have been implemented.
  - One form of this type of evaluation is the treatment effects model.
- The main challenge across different types of impact evaluation is to find a good counterfactual, i.e., the situation a participating subject would have experienced had he or she not been exposed to the program.

IMPACT EVALUATION VS M& E

- Impact evaluation can or should not necessarily be conducted independently of M&E.
- M&E assesses how an intervention evolves over time, evaluating data available from the project management office in terms of initial goals, indicators, and outcomes associated with the program.
- Although M&E does not spell out whether the impact indicators are a *result* of program intervention, impact evaluations often depend on knowing how the program is designed, how it is intended to help the target audience, and how it is being implemented.
- Such information is often available only through operational evaluation as part of M&E.
- M&E is necessary to understand the goals of a project, the ways an intervention can take place, and the potential metrics to measure effects on the target beneficiaries.
- Impact evaluation provides a framework sufficient to understand whether the beneficiaries are truly benefiting from the program—and not from other factors.

5.2 Methodologies in impact evaluation

- The different approaches to impact evaluation are:
  - Randomized evaluations,
  - Propensity score matching,
  - Double-difference methods,
  - Instrumental variables, and
  - Regression discontinuity and pipeline approaches.
- Each of these methods involves a different set of assumptions in accounting for potential selection bias in participation that might affect construction of program treatment effects.

1. Randomized evaluation

- Allocating a program or intervention randomly across a sample of observations is one solution to avoiding selection bias, provided that program impacts are examined at the level of randomization.
- Careful selection of control areas (or the counterfactual) is also important in ensuring comparability with participant areas and ultimately calculating the treatment effect (or difference in outcomes) between the two groups.
- The treatment effect can be distinguished as the *average treatment effect* (ATE) between participants and control units, or the *treatment effect on the treated* (TOT), a narrower measure that compares participant and control units, conditional on participants being in a treated area.

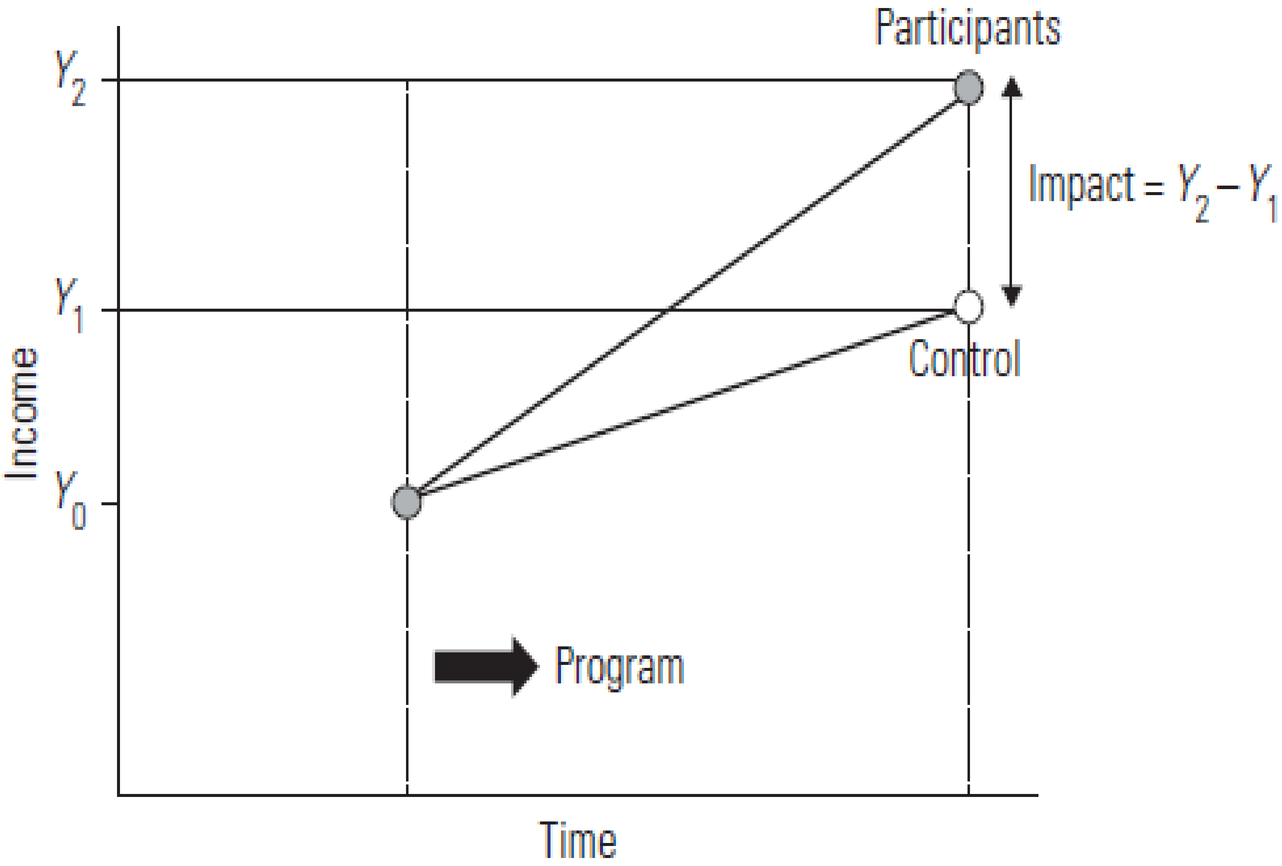
Randomization cont'd

- Randomization could be conducted
  - Purely randomly: where treated and control units have the same expected outcome in absence of the program
    - This method requires ensuring external and internal validity of the targeting design.
  - Partial randomization settings: where treatment and control samples are chosen randomly, conditional on some observable characteristics,
    - for example, landholding or income.
- If these programs are exogenously placed, conditional on these observed characteristics, an unbiased program estimate can be made.

Statistical Design of Randomization

- In finding the counterfactual, it can be very difficult to ensure that a control group is very similar to project areas, that the treatment effects observed in *the sample are generalizable*, and that *the effects themselves are a function of only the program itself*.
- Statisticians have proposed a two-stage randomization approach outlining these priorities.
- In the first stage, a sample of potential participants is selected randomly from the relevant population.
  - This sample should be representative of the population, within a certain sampling error.
  - This stage ensures *external validity* of the experiment
- In the second stage, individuals in this sample are randomly assigned to treatment and comparison groups, ensuring *internal validity* in that subsequent changes in the outcomes measured are due to the program instead of other factors.

The Ideal Experiment with an Equivalent Control Group



Calculating Treatment Effects

- Randomization can correct for the selection bias by randomly assigning individuals or groups to treatment and control groups.
- Consider the classic problem of measuring treatment effects.
  - Let the treatment, T_i , be equal to 1 if subject i is treated and 0 if not.
  - Let $Y_i(1)$ be the outcome under treatment and $Y_i(0)$ if there is no treatment.
- Observe Y_i and T_i, where $\text{Y}_\text{i} = [T_\text{i}Y_\text{i}(1) + (1 - T_\text{i})Y_\text{i}(0)]$.
- Strictly speaking, the treatment effect for unit i is $Y_i(1) - Y_i(0)$, and the ATE is:

$$\text{ATE} = E[Y_i(1) - Y_i(0)],$$

- This formulation assumes that everyone in the population has an equally likely chance of being targeted.

CONT'D

- But in practice we only observe:
- The average outcomes of the treated, conditional on being in a treated area:

$$E[Y_i(1) \mid T_i = 1]$$

- The average outcomes of the untreated, conditional on not being in a treated area,:

$$E[Y_i(0) \mid T_i = 0]$$

- With nonrandom targeting and observations on only a subsample of the population:
  - $E[Y_i(1)]$ is not necessarily equal to $E[Y_i(1) \mid Ti = 1]$, and $E[Y_i(0)]$ is not necessarily equal to $E[Y_i(0) \mid T_i = 0]$.
- Alternate treatment effects are observed TOT:

$$TOT = E[Y_i(1) - Y_i(0) \mid T_i = 1]$$

- The difference in outcomes from receiving the program as compared with being in a control area for a person or subject *i* randomly drawn from the treated sample.

CONT'D

- The TOT reflects the average gains for participants, conditional on these participants receiving the program.
- Suppose the area of interest is the TOT:

$$\mathrm{TOT} = \mathrm{E}[\mathrm{Y_i}(1) - \mathrm{Y_i}(0) \mid \mathrm{Ti} = 1]$$

- If T_i is nonrandom, a simple difference between treated and control areas will not be equal to the TOT.

$$\mathrm{D} = \mathrm{E}[\mathrm{Y_i}(1) \mid \mathrm{T_i} = 1] - \mathrm{E}[\mathrm{Y_i}(0) \mid \mathrm{T_i} = 0]$$

- The discrepancy between the TOT and this D will be $E[Y_{\mathrm{i}}(0) \mid T_i = 1] - E[Y_i(0) \mid T_i = 0]$, which is equal to the bias B in estimating the treatment effect.

$$\mathrm{TOT} = E[Y_i(1) - Y_i(0) \mid T_i = 1]$$

$$= E[Yi(1) \mid Ti = 1] - E[Yi(0) \mid Ti = 1]$$

$$= D = E[Yi(1) \mid Ti = 1] - E[Yi(0) \mid Ti = 0] \text{ if } E[Yi(0) \mid Ti = 0] = E[Yi(0) \mid Ti = 1]$$

$$\Rightarrow \mathrm{TOT} = D \ \text{ if } B = 0.$$

2. Propensity Score Matching

- When a treatment cannot be randomized, the next best thing to do is to try to mimic randomization, i.e., try to have an observational analogue of a randomized experiment.
- With matching methods, one tries to develop a counterfactual or control group that is as similar to the treatment group as possible in terms of *observed* characteristics.
- Find from a large group of nonparticipants, individuals who are *observationally similar* to participants in terms of characteristics not affected by the program.
- Each participant is matched with an observationally similar nonparticipant, and then the average difference in outcomes across the two groups is compared to get the program treatment effect.
- If one assumes that differences in participation are based solely on differences in observed characteristics, and if enough nonparticipants are available to match with participants, the corresponding treatment effect can be measured even if treatment is not random.

PSM CONT'D

- The problem is to credibly identify groups that look alike.
- Identification is a problem because even if households are matched along a vector, *X*, of different characteristics, one would rarely find two households that are exactly similar to each other in terms of many characteristics.
- Because many possible characteristics exist, a common way of matching households is **propensity score matching.**
- In PSM, each participant is matched to a nonparticipant on the basis of a single propensity score, reflecting the probability of participating conditional on their different observed characteristics *X*.
- PSM avoids the "curse of dimensionality" associated with trying to match participants and nonparticipants on every possible characteristic when *X* is very large.

PSM Method in Theory

- The PSM approach tries to capture the effects of different observed covariates *X* on participation in a single propensity score or index.
- Then, outcomes of participating and nonparticipating households with similar propensity scores are compared to obtain the program effect.
- Households for which no match is found are dropped because no basis exists for comparison.
- PSM constructs a statistical comparison group that is based on a model of the probability of participating in the treatment *T* conditional on observed characteristics *X*, or the propensity score:

$$\mathrm{P}(X) = \Pr(T = 1 \,|\, X).$$

  - Rosenbaum and Rubin (1983) show that, under certain assumptions, matching on *P*(*X*) is as good as matching on *X*.

PSM Cont'd

- The necessary assumptions for identification of the program effect are:
  - Conditional independence, and
  - Presence of a common support.
- **Assumption of Conditional Independence**
  - *Conditional independence* states that given a set of observable covariates *X* that are not affected by treatment, potential outcomes *Y* are independent of treatment assignment T.
  - If Y^{T}_{i} represent outcomes for participants and Y^{C}_{i} outcomes for nonparticipants, conditional independence implies: $(Y_i^T, Y_i^C) \perp T_i \mid X_i.$
  - To estimate TOT as compared to ATE a weaker assumption is required: $Y_i^C \perp T_i \mid X_i$

Assumption of Common Support

- A second assumption is the *common support* or *overlap condition*:

  $$0 < P(T_i = 1 \mid X_i) < 1$$

- This condition ensures that treatment observations have comparison observations "nearby" in the propensity score distribution.

- Specifically, the effectiveness of PSM also depends on having a large and roughly equal number of participant and nonparticipant observations so that a substantial region of common support can be found.

- For estimating the TOT, this assumption can be relaxed to:

  $$P\ (T_i = 1 \mid X_i) < 1$$

The TOT Using PSM

- If conditional independence holds, and if there is a sizable overlap in $P(X)$ across participants and nonparticipants, the PSM estimator for the TOT can be specified as the mean difference in Y over the common support, weighting the comparison units by the propensity score distribution of participants.
- A typical cross-section estimator can be specified as follows:

$$\text{TOT}_{\text{PSM}} = E_{P(X)\,|\,T=1}\{E[Y^T\,|\,T=1, P(X)] - E[Y^C\,|\,T=0, P(X)]\}$$

- More explicitly, with cross-section data and within the common support, the treatment effect can be written as follows:

$$\Rightarrow \text{TOT}_{\text{PSM}} = \frac{1}{N_T}\left[\sum_{i\in T} Y_i^T - \sum_{j\in C} \omega(i,j)Y_j^C\right]$$

- where N_T is the number of participants i and $\omega(i, j)$ is the weight used to aggregate outcomes for the matched nonparticipants j.

Application of the PSM Method

- Step 1: Estimating a Model of Program Participation.
- Step 2: Defining the Region of Common Support and Balancing Tests
- Step 3: Matching Participants to Nonparticipants
  - *Nearest-neighbor matching*
  - *Caliper or radius matching*
  - *Stratification or interval matching*
  - *Kernel and local linear matching*
  - *Difference-in-difference matching*