Project Monitoring and Evaluation

MAPMS-711

Admas University

School of Postgraduate Programs

Misgina Asmelash (Ph.D.)

Asst. Prof. of Agricultural Economics and Management at AIGaD, ECSU

misgie2008@yahoo.com

Sep 2019

Methods of Impact Evaluation: building the CF

1. Field Experiments/Randomized evaluations
2. Propensity score matching (PSM)
3. Double-difference (DID) methods
4. Instrumental variable (IV) methods
5. Regression discontinuity (RD) design

Randomized Experiment

- **A theory of observational studies** must have a clear view of **the role of randomization**, so it can have an equally clear view of **the consequences of its absence** (Rosenbaum, 2002).
- Fisher, **The *Design of Experiments*** (1935/1971) book, introduced the principles of **randomization**, demonstrating them with the example of testing a British lady's tea tasting ability.

Key Assumption: randomized assignment

Consider an experimental program design with random selection:

- One group participates in the program (the ***treatment*** group)
- While the other group does not (the ***control*** group)

Key assumption of randomization is ***mean independence***:

$$Y_i^1, Y_i^0 \perp T_i$$

Because of randomization, the average outcomes are independent of treatment assignment

$$E\left(Y_i^1 | T_i = 1\right) = E\left(Y_i^1 | T_i = 0\right) = E\left(Y_i^1\right)$$
$$E\left(Y_i^0 | T_i = 1\right) = E\left(Y_i^0 | T_i = 0\right) = E\left(Y_i^0\right)$$

Key Assumption: randomized assignment

- Hence, the selection effect is removed

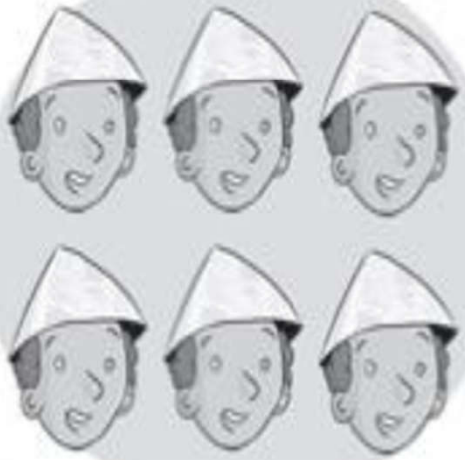
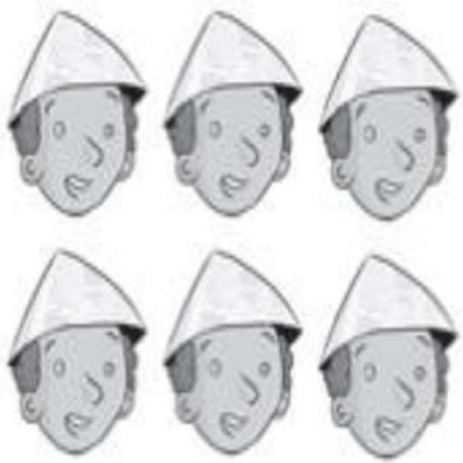
$$E\left(Y_i^0 | T_i = 1\right) - E\left(Y_i^0 | T_i = 0\right) = 0$$

- The observed outcomes yield the treatment effects

$$\begin{aligned}\hat{\beta}_{ra} &= E\left(Y_i^1 | T_i = 1\right) - E\left(Y_i^0 | T_i = 0\right) \\ &= E\left(Y_i^1 | T_i = 1\right) - E\left(Y_i^0 | T_i = 1\right) \\ &= E\left(Y_i^1 - Y_i^0 | T_i = 1\right) \\ &= E\left(Y_i^1 - Y_i^0\right)\end{aligned}$$

- Thus, a randomized experiment identifies $ATE = ATET$
- In practice, we can simply apply OLS

Estimating Impact under Randomized Assignment

	Treatment	Comparison	Impact
	Average (Y) for the treatment group = 100	Average (Y) for the comparison group = 80	Impact = ΔY = 20
Enroll if, and only if, assigned to the treatment group			

Randomized Experiment

- **Individuals/communities/firms are randomly assigned into participation**
- **Advantages:**
  - Often addressed to as the "gold standard": by design: **selection bias is zero on average and mean impact is revealed**
  - Perceived as a fair process of allocation with limited resources

Randomization

❑ Disadvantages:

- ➢ Ethical issues, political constraints, feasibility constraints
- ➢ Internal validity (exogeneity): the validity of inferences about whether the relationship between two variables is causal
- ➢ External validity (generalizability): usually run controlled experiment on a **pilot, small scale**. Difficult to extrapolate the results to a larger population.
- ➢ Does not always solve problem of **spillovers & contamination**

Example: impact of credit on incomes (STATA)

- Treatment with **pure randomization**
- Let Y_i *represent the income per capita for household i.*
- *For* participants, $T_i = 1$, *and the value of* Y_i *under treatment is represented as* $Y_i(1)$.
- *For* non-participants, $T_i = 0$, *and* Y_i *can be represented as* $Y_i(0)$.
- *If* $Y_i(0)$ *is used across nonparticipating* households as a comparison outcome for participant outcomes $Y_i(1)$, *the* average effect of the program can be represented as follows:

$$Y_i = \beta_0 + \boldsymbol{\beta} T_i + \varepsilon_i ,$$

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

- If treatment is random (then $\boldsymbol{T}$ and $\boldsymbol{\varepsilon}$ are independent) → OLS

Propensity Score Matching

Overview

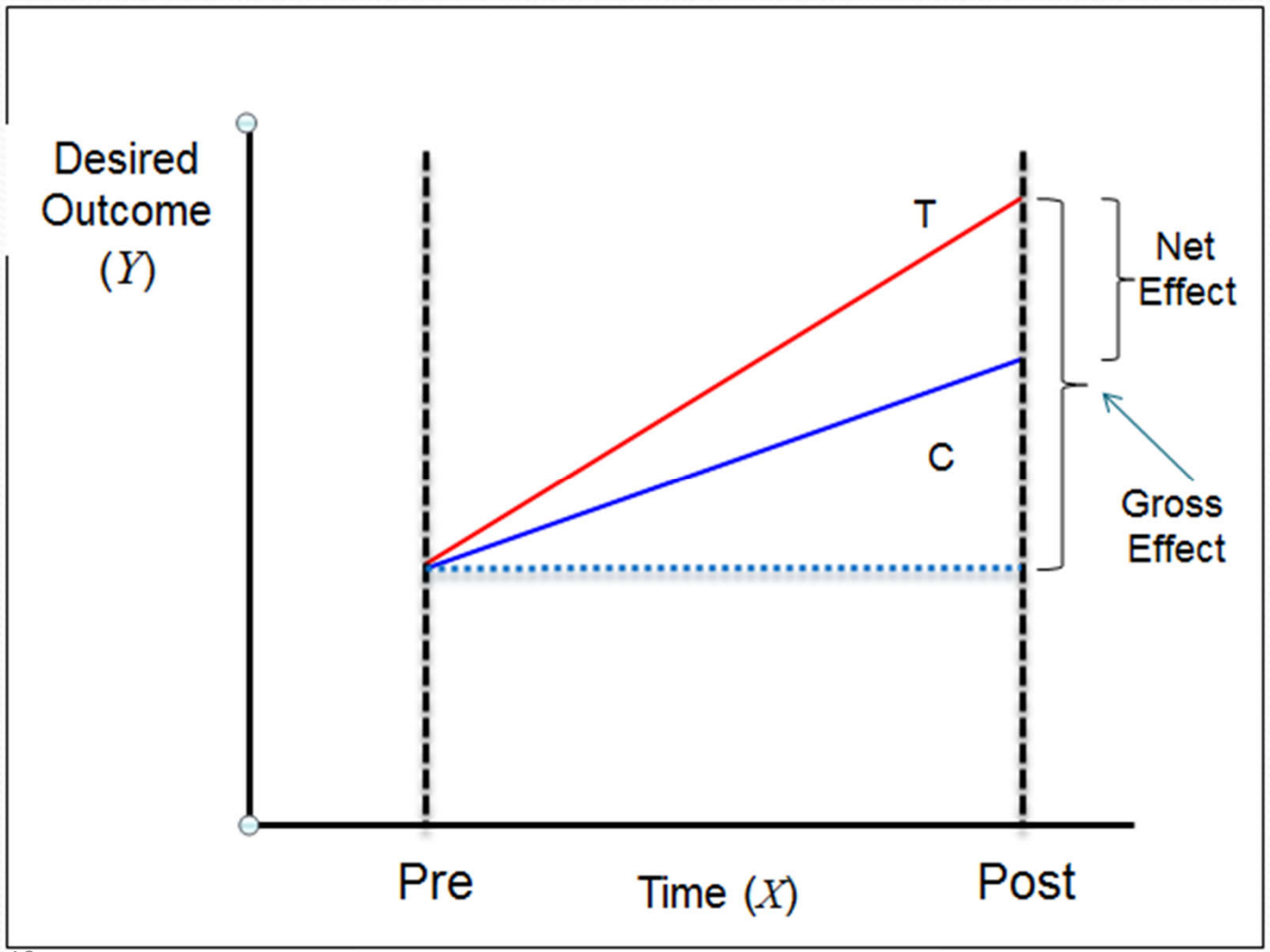
1. Why and when propensity score analysis is needed
2. Conceptual frameworks and assumptions
3. Overview of corrective methods
4. Propensity score matching

1. Why and when PSM is needed?

The Purpose of Evaluation

- The field of **project evaluation** is distinguished principally by **cause-effect** studies that aim to answer a key question:
  - *To what extent can the net difference observed in outcomes between treated and non-treated groups be attributed to an intervention, given that all other things are held constant?*

How much of the effect is due to the project?

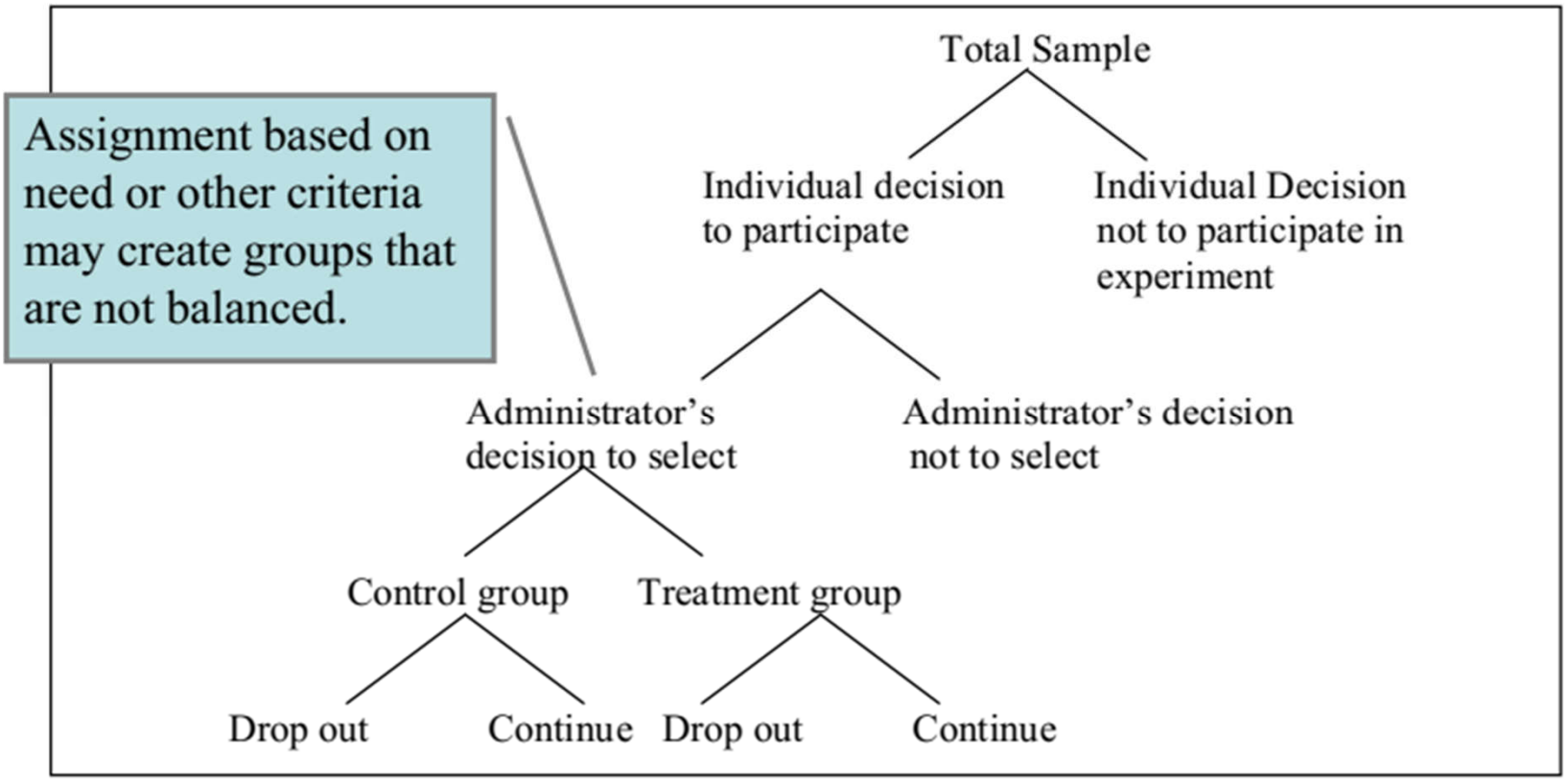


Internal Validity and Threats

- **Internal validity** – the validity of inferences about whether the relationship between two variables is causal.
  - In program evaluation and observational studies in general, researchers are concerned about threats to internal validity.
  - These threats are factors affecting outcomes other than intervention or the focal stimuli.
- **Selection bias** is the most problematic one!

Internal Validity and Threats…

Example of Selection Bias



Source: Maddala, 1983, p. 266

Why and when propensity score matching (PSM) is needed?

Need 1: Remove Selection Bias

- The randomized experimental trial is the **"gold standard"** in outcome evaluation.
- However, in social and health research, **RCTs** are not always practical, ethical, or even desirable.
- Under such conditions, evaluators often use **quasi-experimental** designs, which-in most instances – are vulnerable to **selection bias**.
- Propensity score models help to remove **selection bias.**

Example: In an evaluation of the effect of Catholic versus public school on learning, Morgan (2001) found that the Catholic school effect is strongest among Catholic school students who are less likely to attend Catholic schools.

Why and when PSM is needed?...

Need 2: Analyse causal effects in observational studies

- **Observational data -** those that are not generated by mechanisms of **randomized experiments**, such as surveys, administrative records, census data...
- To analyse such data, **OLS regression model** using a dichotomous indicator of **treatment does not work**, because in such model the **error term is correlated with treatment/explanatory** variables.
- The violation of **OLS assumption** will cause an **inflated and asymptotically biased estimate of treatment effect.**

The Problem of Contemporaneous Correlation in Regression Analysis

- Consider a routine regression equation for the outcome, $\boldsymbol{Y_i}$:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \varepsilon_i}$$

- Where $\boldsymbol{T_i}$ is a dichotomous variable indicating intervention, and *Xi* is the vector of covariates for case *i*.
- We wish to estimate the effect $\boldsymbol{\beta_1}$ of treatment $\boldsymbol{T_i}$ on $\boldsymbol{Y_i}$ by controlling for observed confounding variables $\boldsymbol{X_i}$.

When randomization is compromised or not used, the correlation between $\boldsymbol{T}$ and $\boldsymbol{\varepsilon}$ may not be equal to zero. As a result, OLS estimator of the effect of intervention ($\boldsymbol{\beta}$) may be biased and inconsistent. $\boldsymbol{T}$ is not exogenous.

How Big Is This Problem?

Very big!

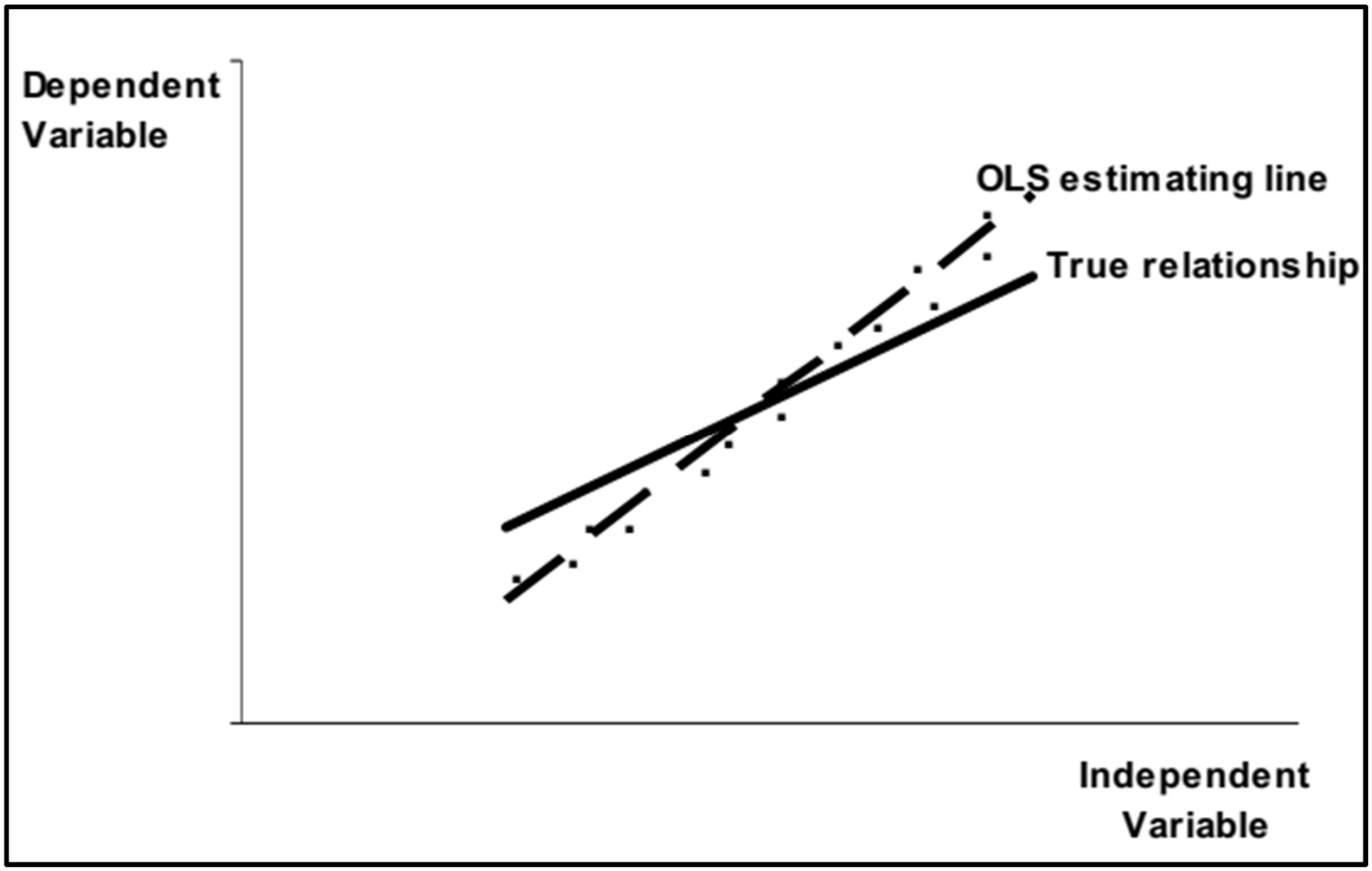
- The majority of **nonrandomized studies** that have used statistical controls to balance treatment and non-treatment groups may have produced **erroneous findings**.

Note

- The amount of error in findings will be related to the degree to which the error term is **NOT** independent of explanatory/treatment indicator measures. This problem applies to any statistical model in which the independence of the error term is assumed.

Consequence of Contemporaneous Correlation:

Inflated Slope and Asymptotical Bias



2. Conceptual frameworks and assumptions

The Neyman-Rubin Counterfactual Framework (1)

- **Counterfactual**: what would have happened to the treated subjects, had they not received treatment?
- One of the seminal developments in the conceptualization of program evaluation is the Neyman (1923) – Rubin (1978) **counterfactual framework**. The key assumption of this framework is that individuals selected into treatment and nontreatment groups have potential outcomes in both states: the one in which they are observed and the one in which they are not observed. This framework is expressed as:

$$Y_i = W_i Y_{1i} + (1 - W_i) Y_{0i}$$

- The key message conveyed in this equation is that to infer a causal relationship between W_i (the cause) and Y_i (the outcome) the analyst cannot directly link Y_{1i} to W_i under the condition $W_i = 1$; instead, the analyst must check the outcome of Y_{0i} under the condition of $W_i = 0$, and compare Y_{0i} with Y_{1i}.

The Neyman-Rubin Counterfactual Framework (2)

- There is a crucial problem in the above formulation: Y_{0i} is not observed. Holland (1986, p. 947) called this issue the "fundamental problem of causal inference."
- The Neyman-Rubin counterfactual framework holds that a researcher can estimate the counterfactual by examining the average outcome of the treatment participants (i.e., $E(Y_1|W=1)$]) and the average outcome of the nontreatment participants (i.e., $E(Y_0|W=0)$)in the population. Because both outcomes are observable, we can then define the treatment effect as a mean difference:

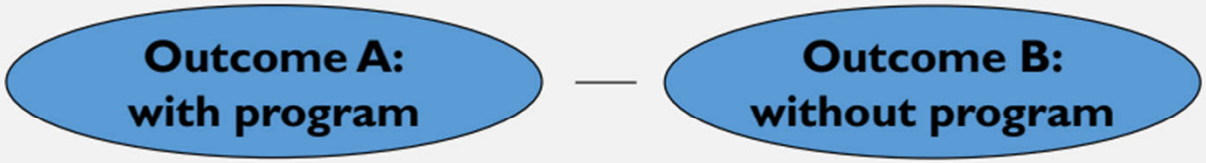
$$\tau = E(Y_1|W=1) - E(Y_0|W=0)$$

- Under this framework, the evaluation of $E(Y_1|W=1) - E(Y_0|W=0)$ can be thought as an effort that uses $E(Y_0|W=0)$ to estimate the counterfactual $E(Y_0|W=1)$. The central interest of the evaluation is not in $E(Y_0|W=0)$, but in $E(Y_0|W=1)$.

The missing Counterfactual

Evaluation question: what is the effect of a program?

Effect =



Problem: we only observe individuals that

- participate:

or

- do not participate :

... but never A *and* B for everyone!

The Neyman-Rubin Counterfactual Framework (3)

- With sample data, evaluators can estimate the average treatment effect as:

$$\hat{\tau} = E(\hat{y}_1 \mid w = 1) - E(\hat{y}_0 \mid w = 0)$$

- The real debate about the classical experimental approach centers on the question: whether $E(Y_0|W=0)$ really represents $E(Y_0|W=1)$?
- In a series of papers, Heckman and colleagues criticized this assumption.
- Consider $E(Y_1|W=1) - E(Y_0|W=0)$. Add and subtract $E(Y_0|W=1)$, we have

  $\{E(Y_1|W=1) - E(Y_0|W=1)\} + \{E(Y_0|W=1) - E(Y_0|W=0)\}$

  The standard estimator provides unbiased estimation if and only if **$E(Y_0|W=1) = E(Y_0|W=0)$.**

  In many empirical projects, $E(Y_0|W=1) \neq E(Y_0|W=0)$.

The Neyman-Rubin Counterfactual Framework (4)

Heckman & Smith (1995) - Four Important Questions:

- What are the effects of factors such as subsidies, advertising, local labor markets, family income, race, and sex on program application decision?
- What are the effects of bureaucratic performance standards, local labor markets and individual characteristics on administrative decisions to accept applicants and place them in specific programs?
- What are the effects of family background, subsidies and local market conditions on decisions to drop out from a program and on the length of time taken to complete a program?
- What are the costs of various alternative treatments?

The Fundamental Assumption: Strongly Ignorable Treatment Assignment

- **Rosenbaum and Rubin (1983)**

$$(Y_0, Y_1) \perp W \mid X$$

- **Different versions:**
  - "un-confoundedness" and "ignorable treatment assignment" (Rosenbaum and Robin 1983) "selection on observables" (Barnow, Cain, & Goldberger, 1980), "conditional independence" (Lechner 1999), and "exogeneity" (Imbens, 2004)

Estimating the Counterfactual

- On a conceptual level, solving the counterfactual problem requires the evaluator to identify a **"perfect clone"** for each program participant

- Although no perfect clone exists for a single individual, **statistical tools** exist

- These tools will be used to generate 2 groups of individuals that, ***if their numbers are large enough,*** are ***statistically indistinguishable*** from each other

Estimating the Counterfactual

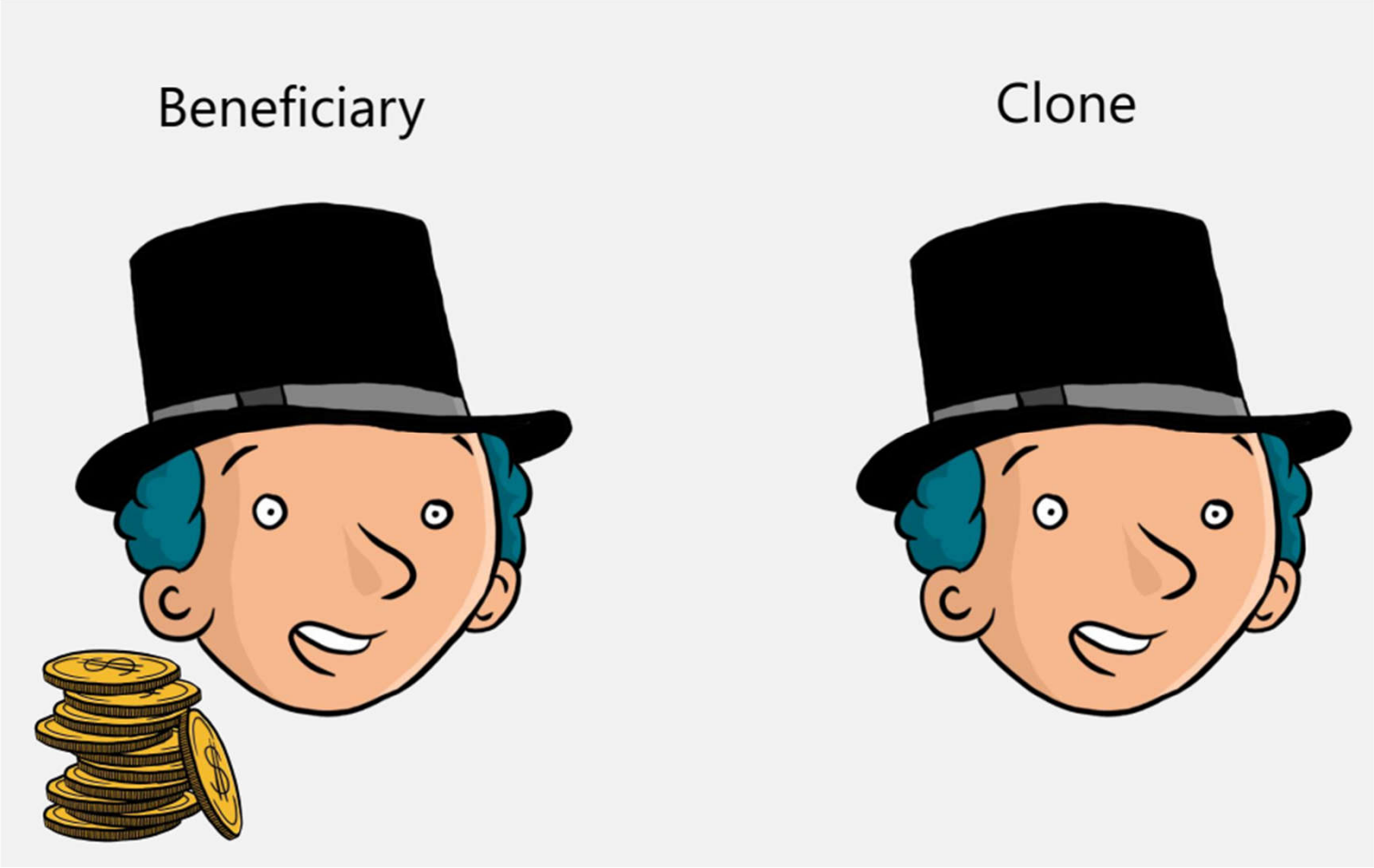
- Specifically, the treatment & comparison groups must be the same **in at least 3 ways:**

1. TG & CG **must be identical in the absence of the program**. **On average** the characteristics of T & C groups should be the same. E.g. the average age in the TG should be the same as the average age in the CG

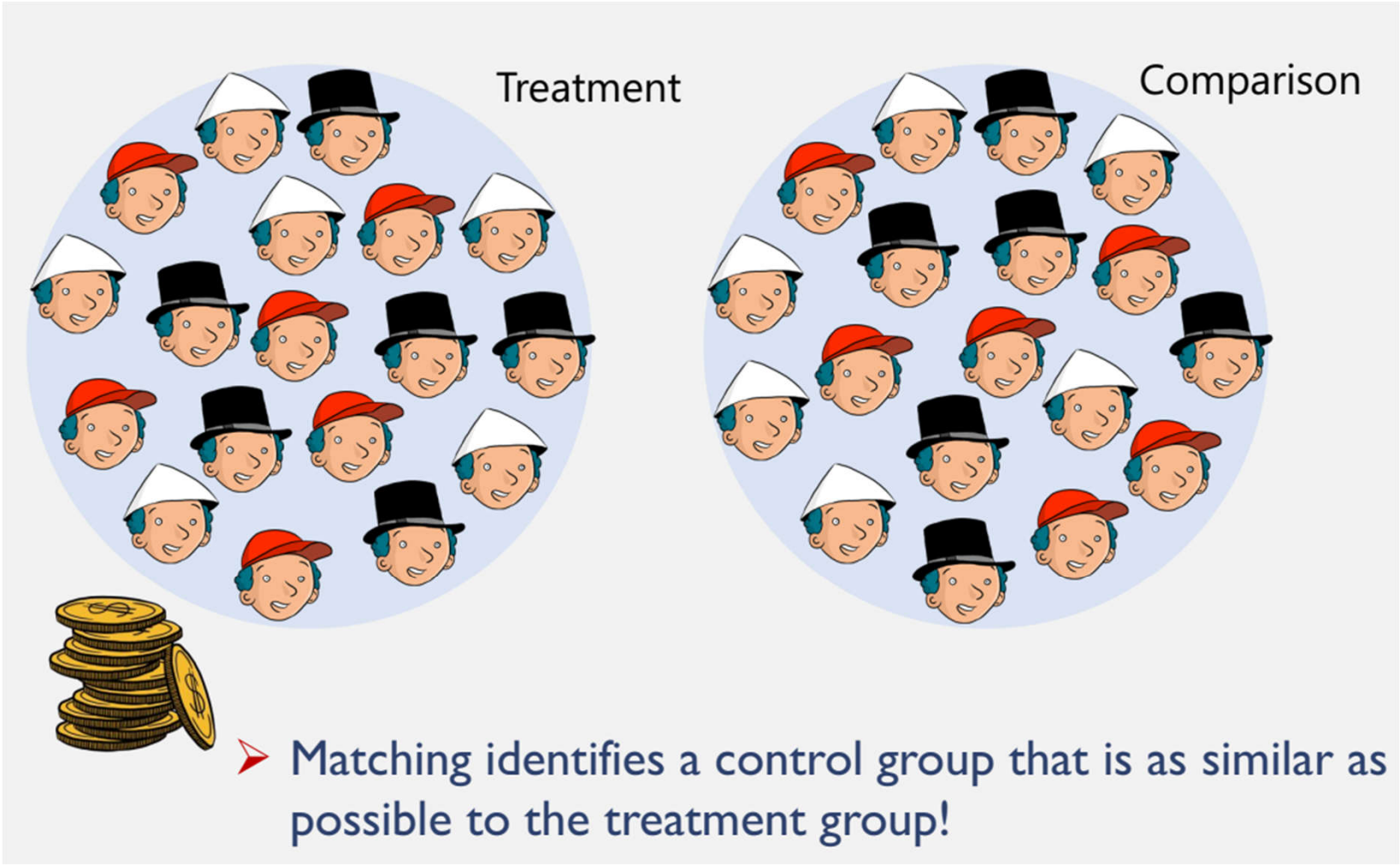
2. TG & CG **should react to the program in the same way**. E.g, the incomes of units in the TG should be as likely to benefit from training as the incomes of the CG

3. T &G **groups cannot be differentially exposed** to **other interventions** during the evaluation period

The Perfect Clone



The Perfect Clone



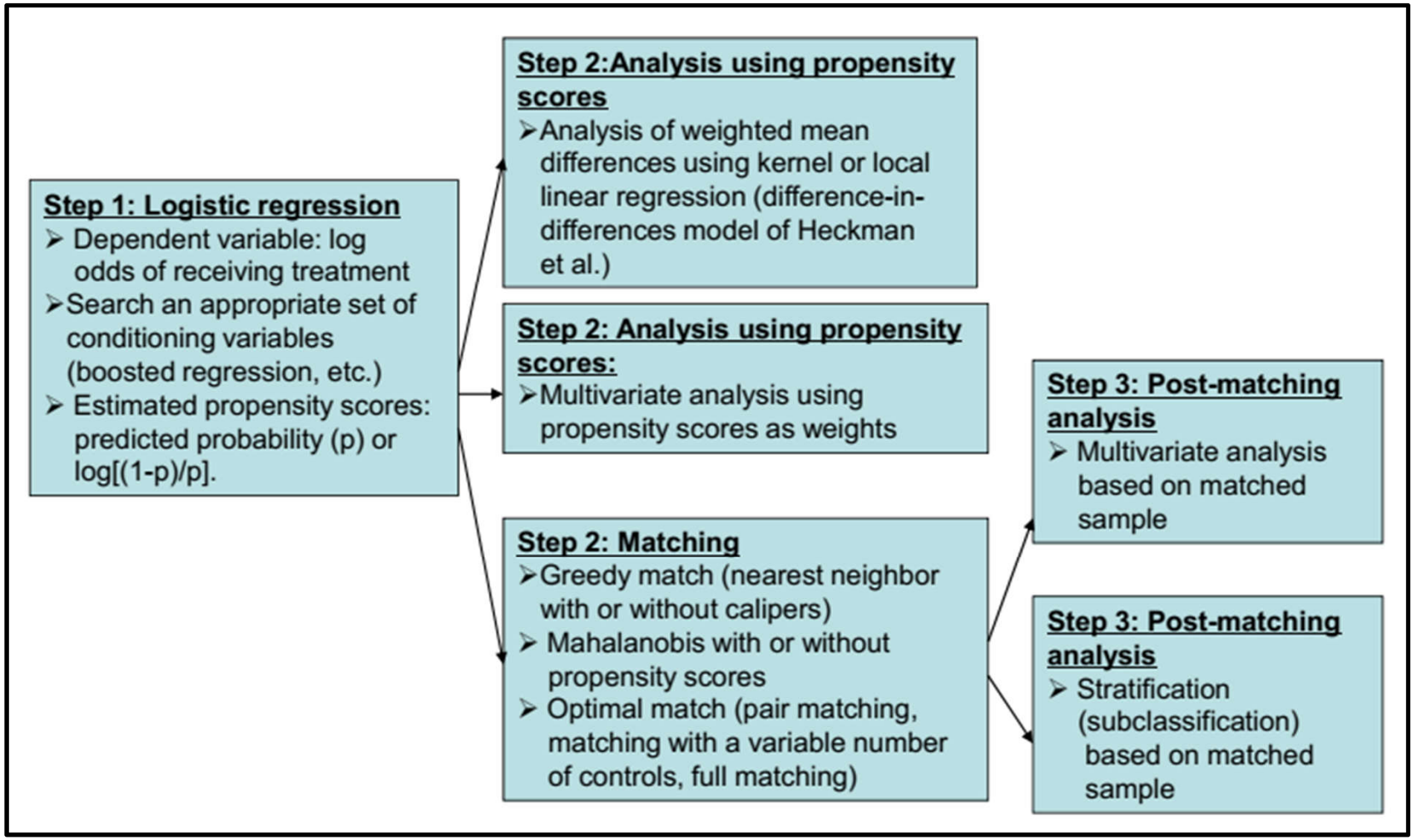
- Matching identifies a control group that is as similar as possible to the treatment group!

3. Overview of Corrective Methods

Four Models Described by Guo and Fraser (2010)

1. **Heckman's sample selection model** (Heckman, 1976, 1978, 1979) and its revised version estimating treatment effects (Maddala, 1983).
2. **Propensity score matching** (Rosenbaum and Rubin, 1983), optimal matching (Rosenbaum, 2002), **propensity score weighting, modeling treatment dosage, and related models.**
3. **Matching estimators** (Abadie and Imbens, 2002, 2006).
4. **Propensity score analysis** with Nonparametric regression (Heckman, Ichimura, and Todd, 1997, 1998)

General Procedure for PSM Summarized by Guo and Fraser (2010)



Other Corrective Models

- Regression discontinuity designs
- Instrumental variables approaches
- Interrupted time series designs
- Bayesian approaches to inference for average treatment effects

4. Propensity Score Matching (Rosenbaum and Rubin, 1983)

Overview to Matching Methods

- PSM constructs a counterfactual comparison group for the evaluation problem.

- Let D indicate whether the household receives the programme or "treatment":

  D = 1 if the household receives the program;

  D = 0 otherwise.

- The evaluation problem is to estimate the average impact of the program's intervention on those that receive it:

$$E(Y_1 - Y_0|D=1) = E(Y_1|D=1) - E(Y_0|D=1)$$

Overview to Matching Methods…

- PSM constructs a statistical comparison group that is based on a model of the probability of participating in the treatment, using observed characteristics (Heckman et al., 1998; Smith & Todd, 2001)
- Participants are then matched on the basis of this probability, or *propensity score*, to nonparticipants.
- The average treatment effect of the program is then calculated as the mean difference in outcomes across these two groups.

Propensity Score Matching (PSM)

- **Estimating the propensity score matching:** match treated and untreated observations on the estimated probability of being treated **(propensity score).**
- Match on the basis of the **propensity score**

$$P(X) = Pr(d=1|X)$$

  - D indicates participation in project
  - Instead of attempting to create a match for each participant with exactly the same value of X, we can instead match on the probability of participation.

PSM: Key Assumptions

- **Key assumption: participation is independent of outcomes conditional on X_i**

$$E[y_0 \mid X, d = 1] = E[y_0 \mid X, d = 0]$$

  - This is false if there are unobserved outcomes affecting participation.

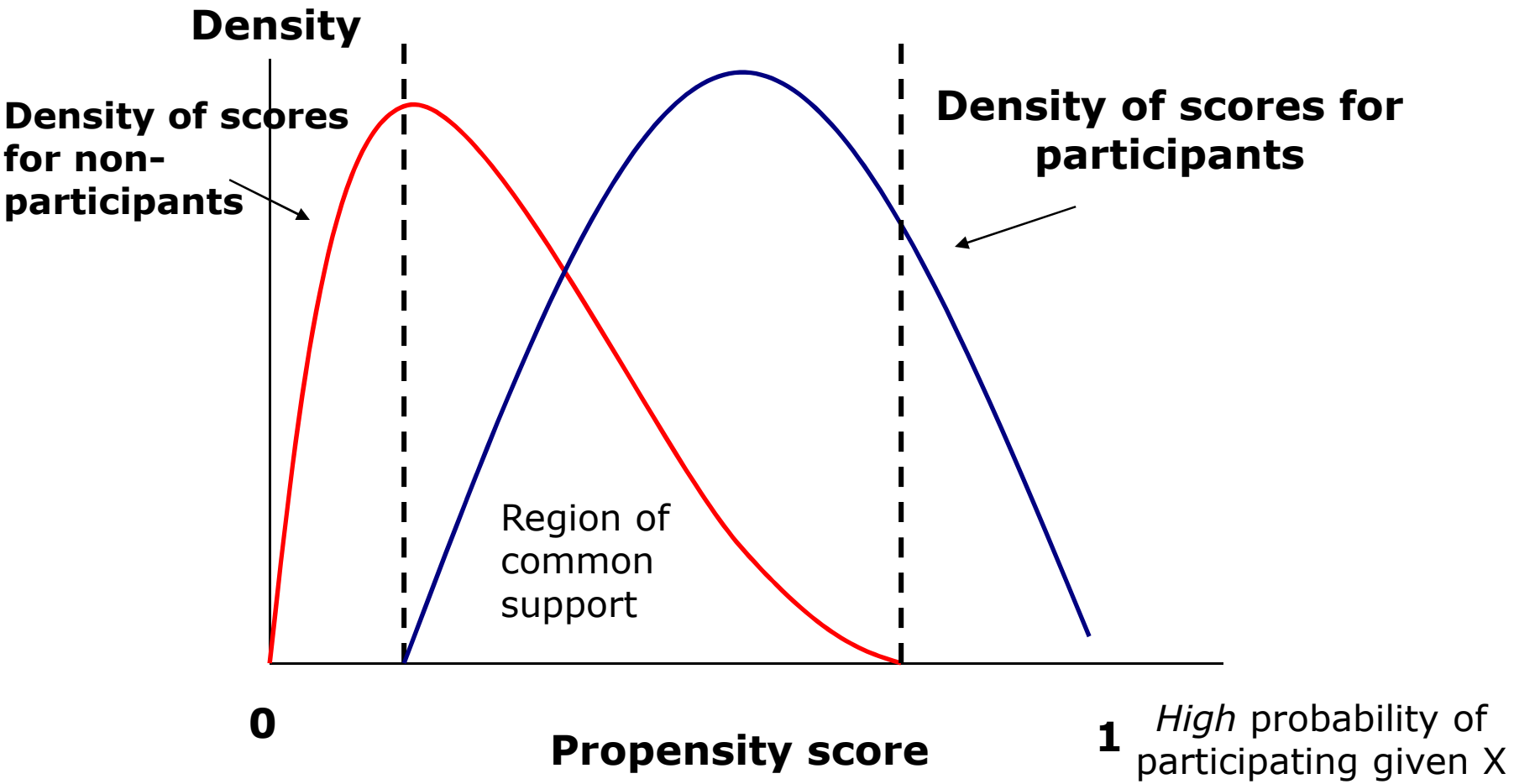
- **Enables matching not just at the mean but balances** the distribution of observed characteristics across treatment and control

Range of Common Support

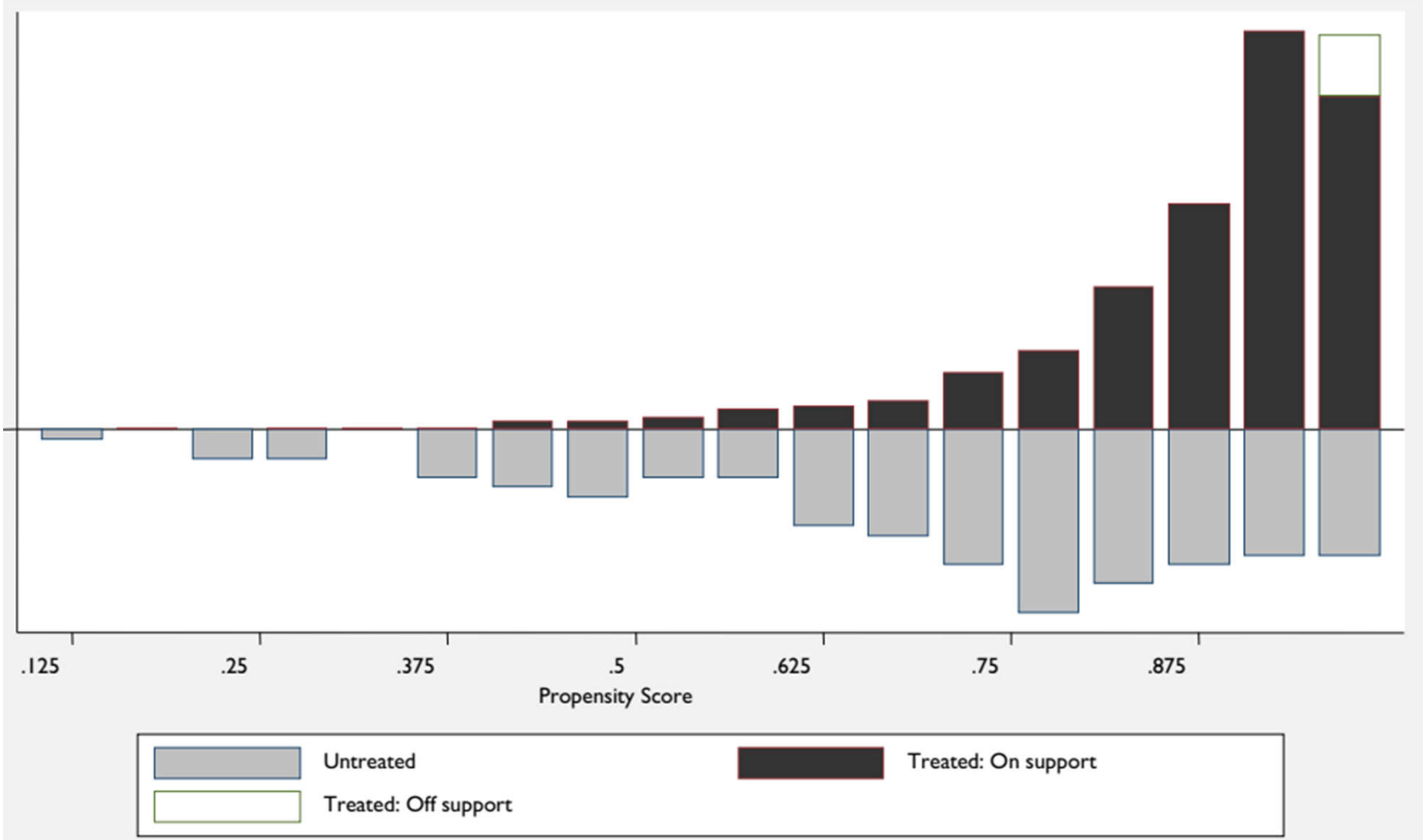
- Common support:
  - The distribution of $p(X)$ may differ for treated and controls
  - Especially for controls it can be hard to find high values of $p(X)$
  - Matching is only possible if there is a similar range of $p(X)$ for both treated and control units!

  → Restrict the match to the ***range of common support***!
- Look where distributions of the propensity score overlap
- Plot $p(X)$ for the treated and non-treated
- Drop non-treated who fall outside of the region of common support

Range of Common Support



Range of Common Support



Matching Methods

- Once we have estimated $p(X)$ there are several methods for propensity core matching
  1. Nearest neighbour matching
  2. Calliper matching
  3. Kernel matching
  4. Weighting by the propensity score

Steps in Score Matching

1. Need **representative and comparable data** for both treatment and comparison groups
2. Use a logit (or other discrete choice model) to estimate program participations as a function of observable characteristics
3. Use predicted values from logit to generate propensity score $p(x_i)$ for all treatment and comparison group members

Steps in Score Matching…

4. **Match Pairs:**

   - Restrict sample to common support (as in Figure)
   - Need to determine a *tolerance limit*: how different can control individuals or villages be and still be a match?
     - 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 matched pairs

Final comments on PSM and OLS

In reality, PSM and OLS do not eliminate bias (unobservables), but may reduce it

The level of success critically depends on X and knowledge of the selection process

Basically, PSM methods recreate an experimental design *ex-post*, by re-weighting the data based on $p(X)$

Main difference PSM and OLS:

- PSM does not impose functional form (non-parametric estimation)
- OLS assumes constant treatment effects
- PSM only considers observations within range of common support
- PSM only considers control variables for the selection process (T), while OLS should control for variables affecting the outcome (Y)

PSM in Practice

To estimate the propensity score, authors used:

Village level characteristics

- Including: Village size, amount of irrigated land, schools,

Household variables

- Including: asset ownership, educational background of HH members

Are there variables which can not be included?

- Only using cross-section, so no variables influenced by project

Table 3
Impacts of piped water on diarrhea prevalence and duration for children under five

	Prevalence of diarrhea		Duration of illness	
	Mean for those with piped water (st. dev.)	Impact of piped water (st. error)	Mean for those with piped water (st. dev.)	Impact of piped water (st. error)
Full sample	0.0108	−0.0023*	0.3254	−0.0957*
	(0.046)	(0.001)	(1.650)	(0.021)
Stratified by household income per capita (quintiles)				
1 (poorest)	0.0155	0.0032*	0.4805	0.0713
	(0.055)	(0.001)	(2.030)	(0.053)
2	0.0136	0.0007	0.4170	0.0312
	(0.051)	(0.001)	(1.805)	(0.051)
3	0.0083	−0.0039*	0.2636	−0.1258*
	(0.038)	(0.001)	(1.418)	(0.042)
4	0.0100	−0.0036*	0.3195	−0.1392*
	(0.044)	(0.001)	(1.703)	(0.048)
5	0.0076	−0.0068*	0.1848	−0.2682*
	(0.042)	(0.001)	(1.254)	(0.036)

Design	When to use	Advantages	Disadvantages
Randomization	▫ Whenever feasible ▫ When there is variation at the individual or community level	▫ Gold standard ▫ Most powerful	▫ Not always feasible ▫ Not always ethical
Randomized Encouragement Design	▫ When an intervention is universally implemented	▫ Provides exogenous variation for a subset of beneficiaries	▫ Only looks at sub-group of sample ▫ Power of encouragement design only known ex post
Regression Discontinuity	▫ If an intervention has a clear, sharp assignment rule	▫ Project beneficiaries often must qualify through established criteria	▫ Only look at sub-group of sample ▫ Assignment rule in practice often not implemented strictly
Difference-in-Differences	▫ If two groups are growing at similar rates ▫ Baseline and follow-up data are available	▫ Eliminates fixed differences not related to treatment	▫ Can be biased if trends change ▫ Ideally have 2 pre-intervention periods of data
Matching	▫ When other methods are not possible	▫ Overcomes observed differences between treatment and comparison	▫ Assumes no unobserved differences (often implausible)