# Acknowledgement of Country

I acknowledge the Traditional Owners and their custodianship of the lands on which we meet today.

On behalf of us all, I pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

# Lecture 9: Propensity score matching

Julie Moschion
University of Queensland

# Reading for Lecture 9

- Mostly Harmless Econometrics (Angrist and Pischke): Chapter 3
- Additional Reading: Microeconometrics: Methods and Applications (Cameron and Trivedi): Chapter 25.4

# Introduction

- Our goal is to evaluate the **causal effect of a treatment** (policy, program, behaviour...) on outcome(s) of interest.
- Ideally, we'd compare outcomes between treated and control units using a RCT or a natural experiment (random exogenous shock).
- Often, data generated by **random assignment** of subjects to the treatment and control groups is not available.
- So we revert to **non-experimental data** providing information on whether subjects were treated or not and other characteritics (outcomes, demographics...).
- These data can be administrative data (i.e. not collected for the purpose of evaluation), survey data or a combination.

# Introduction

- In the absence of random assignment, we worry about **unobservables which may affect the outcomes and be associated with the treatment status**.
- Examples:
  - Do individuals with higher aptitude/motivation self-select into a job training program?
    If so, the treated will be more motivated which may also affect the outcomes (e.g. job performance) and bias OLS estimates.
  - Are schools selected for programs supporting students' performance (e.g. small class sizes) selected based on students' performance?
    If so, the treated exhibit different prior outcomes which is probably correlated with future outcomes and bias OLS estimates.

# Introduction

- We have seen a number of methods that can be used to solve this issue: **IVs, DiD, fixed effects, RD**.
- But all of these require specific data features which are sometimes not available: a source of exogenous variation in the treatment, observation of outcomes pre-treatment with parallel trends, panel data, a discontinuity in treatment allocation.
- And sometimes **none of these are available**...
- An alternative are **matching methods**.
  - The general idea is to find the best possible match(es) for each treated observation from the pool of control observations based on the propensity score, i.e. the probability of treatment given observable characteristics.
  - After matching, we compare the outcomes of the (matched) treated and control observations to estimate the **average treatment effects (ATE)**.

# Jalan and Ravallion (2003): piped water & diarrhea

- The aim of the paper is to estimate the impact of piped water on children diarrhea in rural India.
- Children diarrhea is a critical problem in many developing countries and evaluating the impact of programs that can reduce it can have enormous benefits.
- Important aspects in evaluating the impact of piped water are:
  - Whether a child is less vulnerable to getting diarrhea if they live in a household with access to piped water
  - Whether children in poor households or households with low levels of education have smaller health gains from piped water (heterogeneity) i.e. whether the impact of public investments depends on specific parental inputs.
  - Whether income matters independently of parental education.

# Jalan and Ravallion (2003): piped water & diarrhea

- The issue is that the availability of piped water across villages and households might not be exogenous.
- Classic problem with infrastructure programs: deployment was not randomised (although sometimes it is possible to randomise over time).
- The challenge: observable and unobservable differences across households with piped water and those without.

# Jalan and Ravallion (2003): piped water & diarrhea

- Jalan and Ravallion use cross-sectional 1993-1994 data from a nationally representative survey on 33,000 rural households from 1765 villages.
- Using PSM they find that:
  - Children under 5 with access to piped water are less likely to get diarrhea and if they do get it, it will be shorter.
  - Children with mothers that have low levels of education do not benefit from access to piped water suggesting that public information campaigns need to be combined with investments in infrastructures.

# Methodology

- Step 1: assign the observations into two groups: the **treated** group that receives the treatment and the **control** that does not.
  - Di is a binary variable that indicates if the observation is treated or not: Di=1 for treated observations and Di=0 for control observations.
  - In the previous example: households that have piped water get Di=1 (for household i) and those that do not get Di=0.
- Step 2: estimate a **probit/logit** model of the propensity to be assigned to the treated group.
  - In this model, Di is the dependent variable and Xs are the independent variables.
  - In the previous example: having piped water is the dependent variable. Xs include controls at the household level: ethnicity / caste / religion, asset ownership (bicycle, radio, thresher), educational background of household members. And controls at the village level: village size, amount of irrigated land, schools, infrastructure (bus stop, railway station).

# Methodology

- Step 3: use predicted outcome values from the probit/logit to **generate a propensity score** $p(X_i)$ for all treatment and control units.
  - The propensity score is the predicted probability of being treated conditional on observed pre-treatment characteristics.

$$p(X) = prob(D = 1|X) = E(D|X)$$

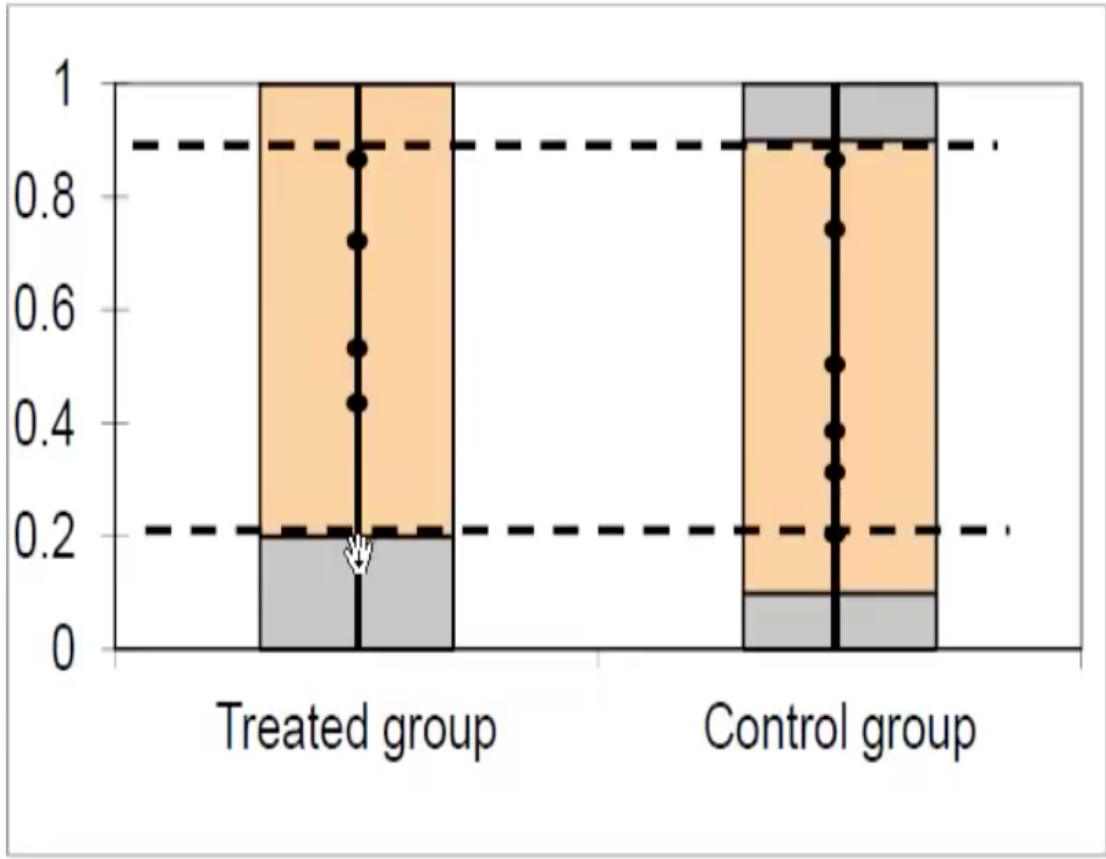
- Step 4: **match treated units to control units** based on their propensity scores and the chosen matching method.
  - Matching methods: kernel, nearest neighour, radius, stratification.
  - Check the common support (common range of propensity scores).
  - In the previous example: households with piped water (the treated group) are matched to those without (control group) on the basis of the propensity score p(x) using a modification of the nearest neighour.

# Methodology

- Step 5: **estimate the treatment effect** by comparing the outcomes y of the treated and control units after matching.
  - i.e by comparing: $y = y_1$ if $D_i = 1$ and $y = y_0$ if $D_i = 0$.
  - In the previous example: by comparing the incidence of diarrhea between the treated units and the matched controls.
- Step 6: assess the **matching quality**.
  - Run balancing tests to compare average characteristics between the matched treated and control observations (there should be no difference after matching).
  - Conduct sensitivity analysis to check the robustness of results to different ways of estimating the propensity score (using different variables...); different matching methods; describe the assumptions.
- But note that **PSM is based on observable characteristics ONLY**: it does not solve OVB! If an important driver of the outcome is correlated with the treatment but not with the observables, the mismatch between treated and controls will subsist after matching.

# Methodology

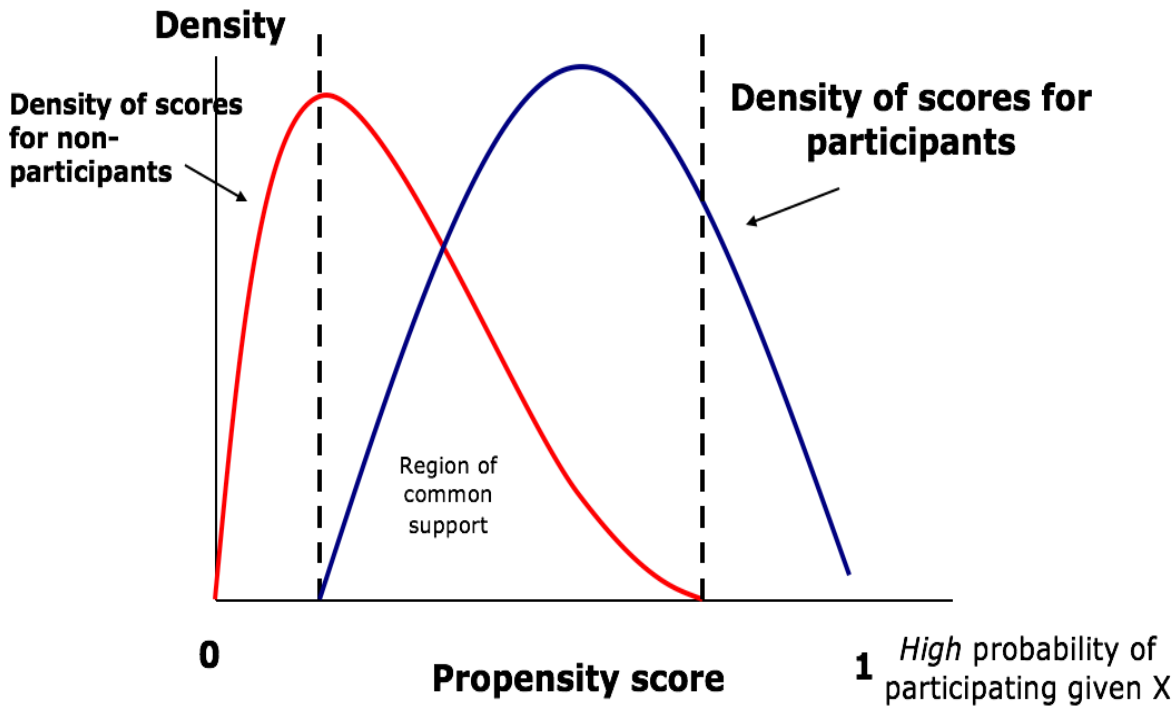
**Propensity score and common support**




- You have generated the propensity score (on the y-axis) for treated and control units. The figure shows a **common range of propensity scores**: propensity scores above 0.2 and below 0.9 overlap between treatment and control.

# Methodology

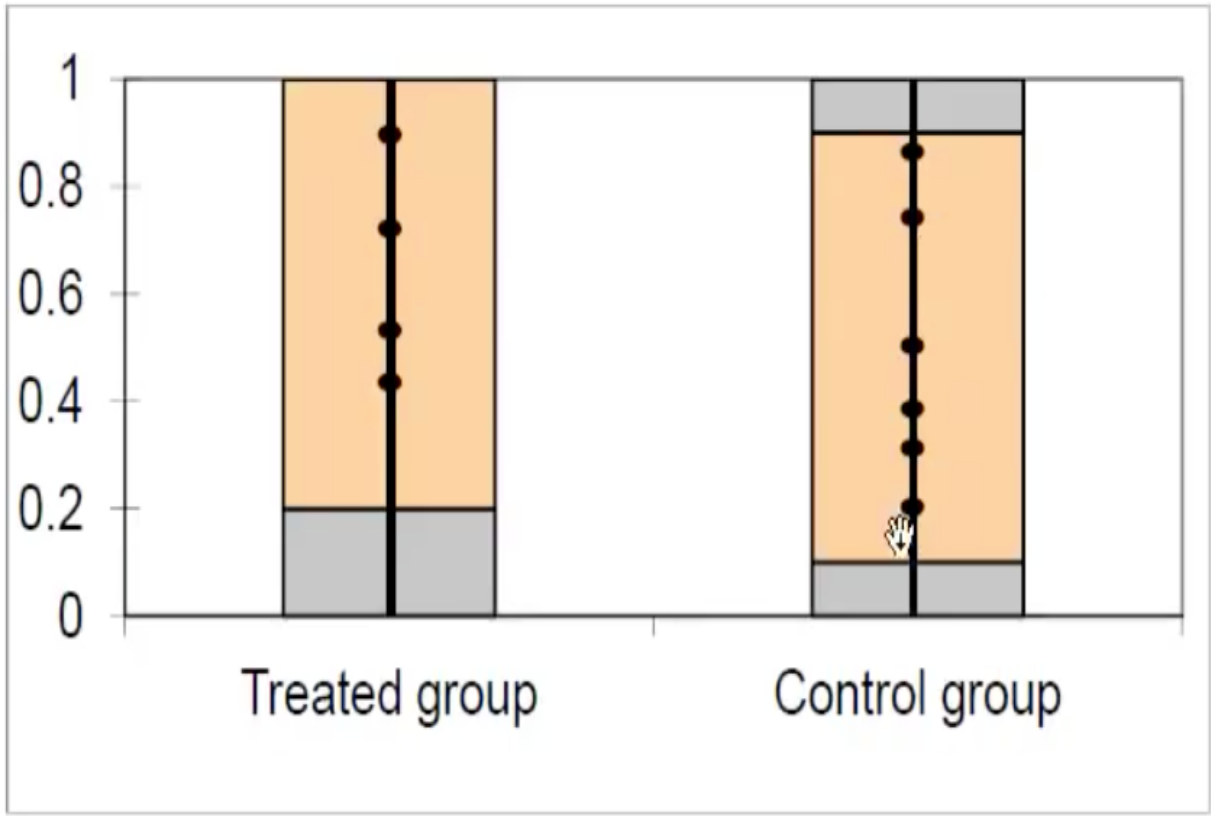
## Propensity score and common support




- This is a different way of looking at the propensity scores (on the x-axis) with the density. The propensity scores for which the density is 0 for one of the 2 groups is outside the common support.

# Methodology

## Matching




- For each treated observation we need to find match(es) in the control group, i.e. observation(s) with similar pre-treatment characteristics (which generate similar probabilities of treatment).

# Methodology

- There are a number of important features in a matching method:
  - Some methods yield a **single match** from the control group, others **multiple matches**.
  - Matching can be done **with or without replacement**:
    - Without replacement: each control observation is used only once as a match for a treated observation.
    - With replacement: each control observation can be used as a match to several treated observations.
- There are different matching methods: **kernel, nearest neighour, radius, stratification**.

# Methodology

- **Nearest Neighbour** is the most intuitive technique.
  - For each treated observation i, we select the control observation j that has the closest propensity score:

$$min||p_i - p_j||$$

  - A control observation could be the closest neighbour for more than one treated observation.
  - Replacement is an optional feature of NN, this choice involves a trade-off between bias and efficiency:
    - With replacement produces better matches and decreases the bias;
    - But it can decrease precision if we only end up using a few control observations.
    - The trade-off is particularly strong if the distribution of propensity scores is very different between the treatment and control group.

# Methodology

- **Caliper matching** reduces the risk of bad matches, i.e. matching control observations that are far away.
  - The idea of this method is to restrict matches to control observations that are the closest if their propensity score is within a caliper of the treated observation.
  - In other words, each treated observation i is matched with the nearest control observations j if j falls within a specified caliper:
$$||p_i - p_j|| < c$$
  - Note that there is no rule to decide which is the right caliper.
  - A drawback is that while preventing bad matches, the standard errors of the treatment effect can increase if only few matches are performed.
- A variant is **radius matching** which extends the range of matches by using all control observations whose propensity score come from within the caliper of the treated observation.

# Methodology

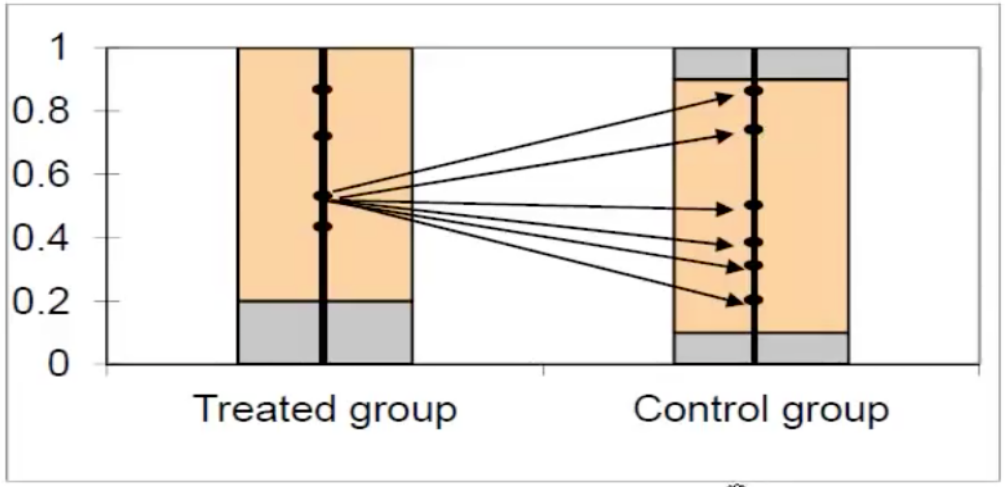
- **Stratification or interval matching** which consists in matching treated observations to control observations which fall in the same propensity score interval.
  - The common support is divided into a set of intervals, usually 5 (0 to 0.2; ...; 0.8 to 1). One way to justify the number of intervals is to check the balance of the propensity score and variables used to build the propensity score within intervals (if not you need more intervals or to adjust the specification of the propensity score equation).
  - Then treatment effects are estimated separately within each interval as the average difference in outcomes between treated and controls. Then weighting is based on the number of observations within an interval.

# Methodology

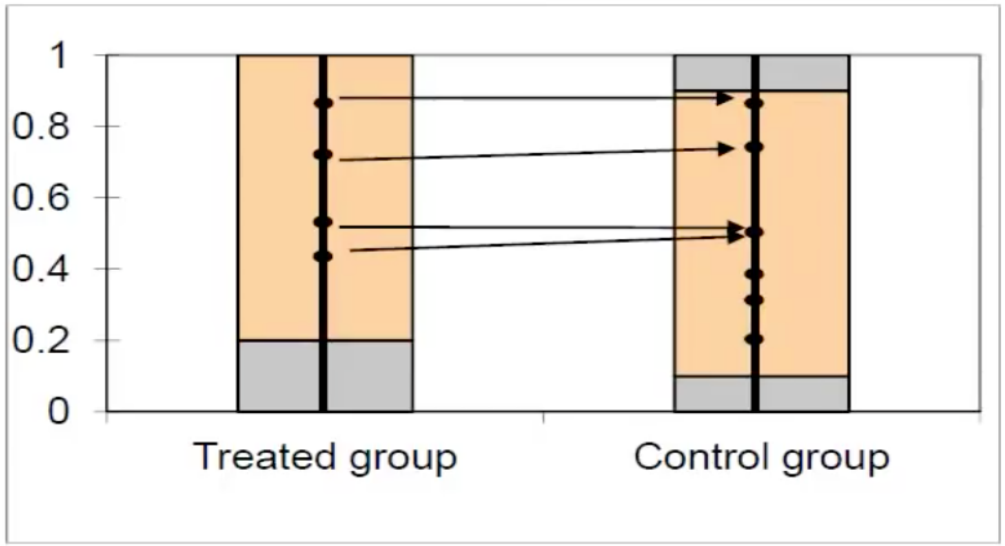
- **Kernel matching**
  - For each treated observation, we use **all control observations**.
  - The matching process will yield **weights** for control observations, such that:
    - controls with a propensity score which is close to the treatment observation will have a high weight;
    - and controls with a propensity score which is further away get a lower weight.
    - Weights are inversely proportional to the distance between treated and control observations.
  - One big advantage of this technique is that by using all controls, the estimates are more precise.
  - However bad matches can generate biases so the common support restriction is of particular importance here.
  - The choice of the kernel function is not important in practice.
  - The choice of the bandwidth comes back to a trade-off between bias and precision both of which increase with the bandwidth.

# Methodology

Kernel matching




Nearest neighbor matching



# Methodology

- The choice between the different methods come back down to a trade-off between bias and efficiency.
- Asymptotically, all PSM estimators should yield the same results. In small samples, the choice of matching method matters.
- No one-size fit all answer. Finite sample performance depends on different factors:
  - If the density of propensity scores differs a lot between the treatment and control groups, the risk of bias becomes important $\Rightarrow$ if using NN, do it with replacement; if using KM ensure common support.
  - If the control group is small $\Rightarrow$ if using NN, do it with replacement but in this case it may be better to use KM to increase the information used in the matching.
- In practice, try various methods and check how sensitive your results are to different methods (hopefully not too much).
- Other types of matching: local linear matching, coarsened exact matching,...

# Rubin causal model (RCM)

- The RCM is a mathematical framework for causal inference which uses the concept of "potential outcomes" to define the causal estimate to be estimated. This framework can be applied to all econometric methods we have seen in this class.
- $D_i$ is the treatment status: treated ($D_i = 1$), control ($D_i = 0$)
- $y_{D_i}$ is the outcome given $D_i$: with treatment $\mathbf{(}y_{1_i}\mathbf{)}$ vs without treatment $\mathbf{(}y_{0_i}\mathbf{)}$.
- These are the **potential outcomes** for individual $i$.

# Rubin causal model (RCM)

- For example, consider a job training program as treatment. If treated ($D_i = 1$) an individual earns a wage $y_{1_i}$, while if she is not treated she earns a wage $y_{0_i}$.
- The causal effect is:

$$Causal\ effect_i = y_{1_i} - y_{0_i}$$

- By definition $y_{1_i}$ and $y_{0_i}$ can never be observed simultaneously for the same individual. Each individual is either treated or not treated.
- Only one potential outcome is observed for each $i$. The **unobserved outcome is called the counterfactual outcome**.
- It is impossible to calculate individual treatment effects, so we estimate average treatment effects on the sample instead.

# Rubin causal model (RCM)

- The model also includes $x$: a vector of observed variables.
- The **average treatment effect (ATE)** is the difference between the outcomes of treated and control observations:

$$\begin{aligned} ATE &= E[y_1 - y_0] \\ &= E[y_1 | D = 1, x] - E[y_0 | D = 0, x] \end{aligned}$$

- The ATE yields the causal effect for random experiments, but may be biased in observational studies if the treated and control observations differ in ways that affect selection into treatment and also determine the outcome.
- In the training program example, certain people decide to participate in the program and this decision may be correlated with what they expect to gain from doing the program (which may differ by ability, effort...).

# Rubin causal model (RCM)

- In this context, we are often interested in recovering the **average treatment effect on the treated (ATET)**.
- The ATET is the difference between the outcomes of the treated with their potential outcomes if they had not been treated:

$$ATET = E[y_1 - y_0|D = 1, x] = E[y_1|D = 1, x] - E[y_0|D = 1, x]$$

- The counterfactual outcome $E[y_0|D = 1, x]$ is not observable and needs to be estimated.
- PSM tries to select a comparison group from all the control units for whom $y_0$ is observed and because they look a lot like the treated units, their $y_0$ can be taken as the counterfactual outcomes for the treated.

# Rubin causal model (RCM)

- Interestingly, the ATE can be expressed with respect to the ATET:

$$\begin{aligned} ATE =& E[y_1|D=1,x] - E[y_0|D=0,x] \\ =& E[y_1|D=1,x] - E[y_0|D=0,x] \\ & - E[y_0|D=1,x] + E[y_0|D=1,x] \\ =& ATET + E[y_0|D=1,x] - E[y_0|D=0,x] \end{aligned}$$

- The difference between the ATET and the ATE gives the condition under which the ATET can be identified in practice, i.e. under which the outcomes of the comparison group provide a good proxy for the counterfactual outcomes of the treated group:
$E[y_0|D=1,x] - E[y_0|D=0,x] = 0$

# PSM assumptions

1. **Conditional Independence Assumptions (CIA)**
   - The potential outcomes are independent of the treatment status conditional on observed variables x:

$$(y_0, y_1) \perp D|x$$
$$E(y|D = 1, x) = E(y|D = 0, x)$$

   - This means that conditional on observable characteristics, the assignment of units to treatment is 'as good as random'.
   - This assumes that selection only comes from observables: it is a very strong assumption which cannot be tested.
   - It is the equivalent of the exogeneity assumption adjusted to the potential outcome framework.

# PSM assumptions

- With PSM, CIA implies that potential outcomes are independent of treatment conditional on the propensity score:

$$(y_0, y_1) \perp D | P(x)$$
$$E[y|D = 1, P(x)] = E[y|D = 0, P(x)]$$
$$E[y_0|D = 1, P(x)] = E[y_0|D = 0, P(x)]$$

- Under CIA, for a given $P(x)$, $ATET = ATE$, i.e. selection bias is zero conditioning on the propensity score.
- Intuitively the CIA holds if for every treated individual we can find a combination from the control group with $P(x)$ as close as possible such that we can compare the outcomes of treated and matched control units.

# PSM assumptions

## 2. Common support (or overlap) assumption

- This ensures that there the treatment D is not perfectly predictable given x, so for each value of x there is a positive probability of being both treated and untreated:

$$\forall x : 0 < P[D = 1|x] < 1$$

- This means that each combination of $x$ observed for the treated can also be observed for the controls $\Rightarrow$ there is sufficient overlap in the characteristics of treated and controls to find adequate matches (ATET).
- Note that for ATE we also need to identify a counterfactual for the treated, i.e. observe all combinations of $x$ from the controls in the treated.
- The results of the estimation are valid only for the region of common support.

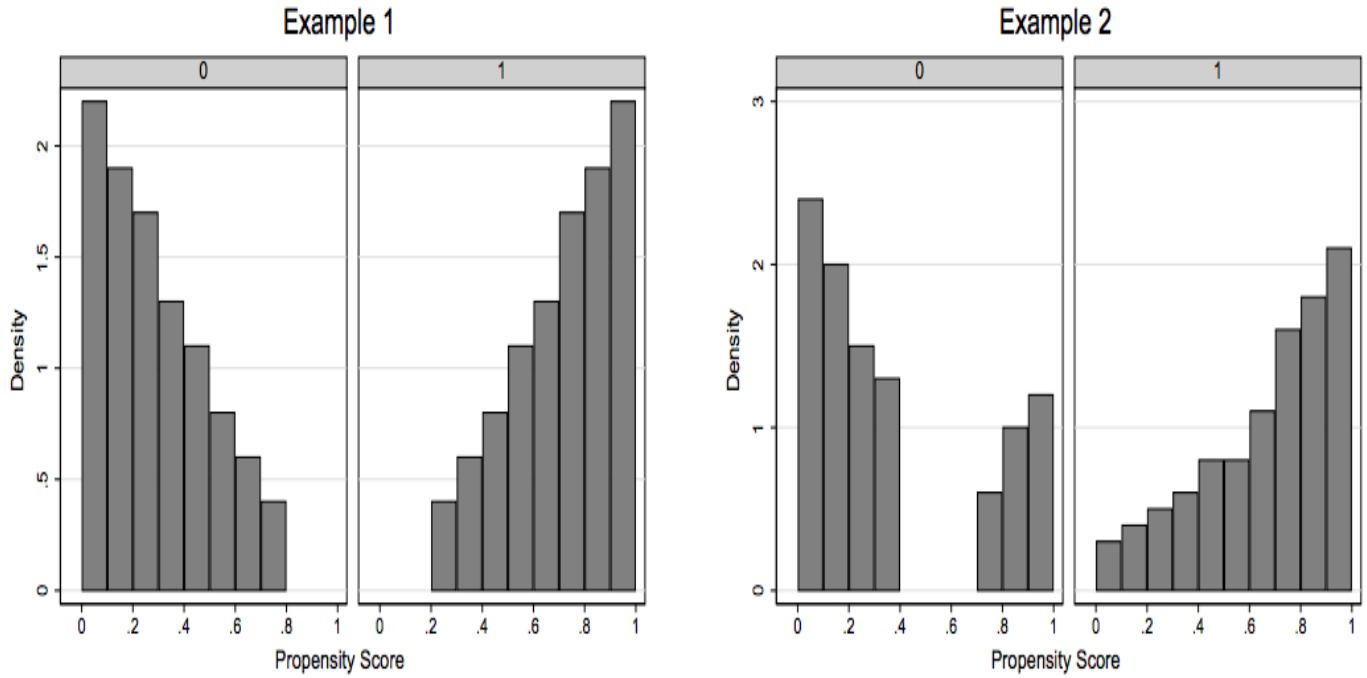
# PSM assumptions

- The common support can and should be tested: a violation of the common support assumption is a major source of evaluation bias (Heckman et al., 1997).
- Testing can be done by:
  - Checking the min and max of the propensity score in both groups.
  - Visually inspecting the density distribution of the propensity score in both groups (or more formally via a comparison test such as the Kolmogorov-Smirnov test).

# PSM assumptions

- If there are sizeable differences between the two groups, it is good practice to remove cases that lie outside the support of the other distribution.
- At the same time, estimating treatment effects only within the common support may be misleading if a large number of observations are discarded (Lechner, 2000).
- Depending on the case a strict rule can be applied or some observations very close to the boundaries may be kept.
- This is more important for KM which uses all observations than a radius approach which restricts bad matches.

# PSM assumptions




The left side in each example refers to non-participants (D=0), the right side to participants (D=1).
*Source:* Hypothetical Example

- On the left, the min max method successfully identifies the common support: [0.2, 0.8].
- On the right it would fail and the density should be used as there are no controls in the region [0.4, 0.7].

# Summary

- There are two assumptions: CIA and common support.
- If both hold, it's like having a random sample at each propensity score.
  - Conditioning on the propensity score, each individual has the same probability of being assigned to the treatment (as in a RCT).
  - So individuals with the same value of P(x) but a different treatment status can act as counterfactuals for each other.
  - At any value of $P(x)$, the difference between the treatment and the control averages yields an unbiased estimate of the ATE.

# Summary

- PSM doesn't solve OVB arising from unobservables BUT it does 3 things that standard OLS doesn't:
  - The common support enables to exclude observations for which we cannot get the same likelihood of treatment given their observables (i.e for who the drivers of treatment have not been identified and possibly drive the outcomes as well);
  - The matching enables to optimise the comparison of treatment and controls by making the control group as similar as possible pre-estimation.
  - It also helps against the curse of dimensionality when there are many variables in x.

# Assessing the matching quality

- **Balancing Tests** are used to ensure that covariates are balanced between the treated and the matched controls.
- Balancing tests consist in comparing all observed Xs using t-tests (each one separately) and F-test (all together for joint significance).
- Differences are expected pre-matching but not post-matching.
- Differences post-matching indicate that the matching was not successful either because:
  - The propensity score is misspecified: try including higher order and interaction terms in the estimation of propensity scores.
  - There is selection bias which cannot be resolved with your observables (failure of the CIA).
- It can also be useful to incorporate information for **individuals who fail the common support assumption**: number, average characteristics.

# Caliendo and Tubbicke (2020): start-up subsidies

- The authors use PSM to evaluate a 2011 reform of the German program that offers start-up subsidies to unemployed people.
- They are interested in the impact of the subsidy on employment and wages up to 40 months after entering the program.
- The aim of these subsidies is to support exit from unemployment by incentivising entrepreneurship and securing people's livelihood in the first bumpy months.

# Caliendo and Tubbicke (2020): start-up subsidies

- PSM relies on the CIA and they include many variables in the estimation of the propensity score to alleviate OVB concerns:
  - Socio-demographics: age, gender, health status, German citizenship, marital status and single parent status, number of children, the presence of young children, highest schooling degree, professional education and qualification.
  - Detailed labour market history: short- and long-term unemployment history, short- to medium-term employment and treatment history, the employment status before unemployment, previous occupation, the size of unemployment benefits received as well as last labour earnings.
  - Regional characteristics: regional dummies, local macroeconomic conditions and self-employment activity.
  - And include polynomials and interaction terms.
- The treated are previously unemployed people who joined the program between February and June 2012.
- The controls were unemployed for at least one day, eligible for the program but did not apply for it in this period.

# Caliendo and Tubbicke (2020): start-up subsidies

- The propensity score is estimated using a probit regression on the pooled sample.

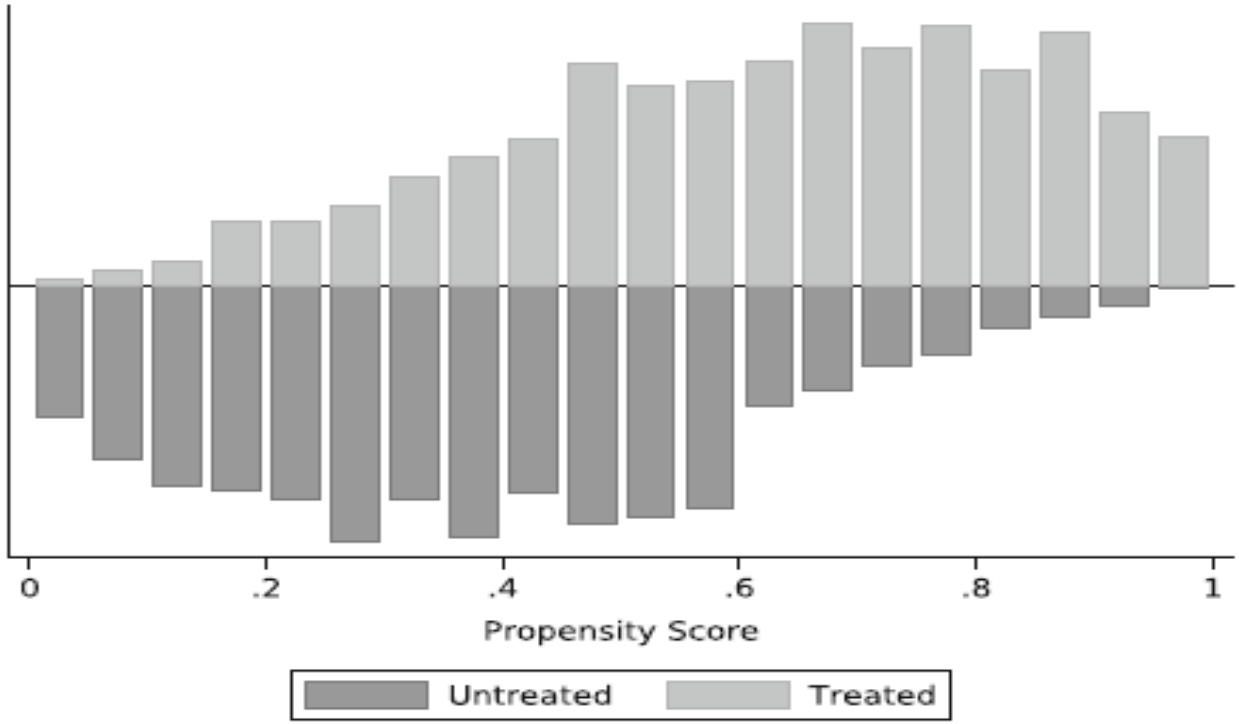



**Fig. 2** Propensity score distribution—Baseline specification. *Note* This graph shows the distribution of estimated propensity scores for the treated and comparison group using a probit regression based on the baseline specification including information on socio-demographics, human capital, labor market history, intergenerational transmission, and regional controls for local labor market conditions and self-employment activity. For details on the specification and estimated coefficients, see Table A.2

- There is extensive overlap between treated and control units. The authors impose common support.

# Caliendo and Tubbicke (2020): start-up subsidies

- They estimate treatment effects using kernel matching (to estimate the balancing weights).
- Provided that the CIA and common support assumptions hold, the average treatment effect on the treated can be estimated as the simple difference in the outcomes between the treated and the weighted control group:

$$T_{ATET} = \frac{1}{N_1} \sum_{i=1}^{N_1} Y_i(1) - \sum_{j=1}^{N_0} \hat{w}_j Y_j(0)$$

- Where $N_1$ and $N_0$ are the number of treated and control units and $\hat{w}_j$ are the balancing weights obtained via the matching.

# Caliendo and Tubbicke (2020): start-up subsidies

## Balancing test

**Table 3** Balancing quality

| | Before matching | After matching |
|---|---|---|
| Number of variables with significant differences in means[a] | | |
| At 1%-level | 38 | 0 |
| At 5%-level | 50 | 0 |
| At 10%-level | 58 | 0 |

- Variables are all balanced post-matching.

# Caliendo and Tubbicke (2020): start-up subsidies

## PSM results

**Table 4** Baseline results and sensitivity analyses with respect to implementation

| | Outcomes after 20 months | | Outcomes after 40 months | |
|---|---|---|---|---|
| | Self- or regular employment | Net monthly earned income | Self- or regular employment | Net monthly earned income |
| *A. Baseline results* | | | | |
| | 0.280*** | 764.2*** | 0.215*** | 980.2*** |
| | (0.000) | (0.000) | (0.000) | (0.000) |

- Treated people are 28 pp more likely to be employed 20 months after entering the program, and they earn 760 more euros per month than the comparison group. Results persist at least 40 months after the program.

# Caliendo and Tubbicke (2020): start-up subsidies

- However, if some unobserved characteristic has an impact on treatment assignment and the outcome, the CIA fails.
- The authors propose an IV strategy to test the robustness of their result:
  - The IV exploits the fact that the reform increased the discretionary power of local employment agencies in allocating active labor market policy funds to their program of choice. This creates regional variation in the likelihood of receiving treatment.
  - IV: regional application approval rates for the program, conditional on many local labor market conditions (incl the regional start-up rate out of unemployment and the overall self-employment rate using flexible categorical dummies).
  - BUT if local differences in the allocation of funds relate to different expectations about the impact of the treatment that are not unobserved, the exogeneity assumption doesn't hold...this is possibly why the IV strategy is used here as a back up to PSM rather than the main strategy.

# Caliendo and Tubbicke (2020): start-up subsidies

## IV results

| | First stage | Second stage | |
|---|---|---|---|
| | At entry | After 20 months | |
| | SUS receipt | Self- or regular employment | Net monthly earned income |
| *A. 2SLS (continuous)* | | | |
| | 0.003*** | 0.413*** | 792.7* |
| | (5.937) | (3.114) | (1.956) |
| | | [.195;.631] | [126.0; 1,459.3] |
| *B. 2SLS (dummy)* | | | |
| | 0.098*** | 0.342** | 880.0* |
| | (4.705) | (2.094) | (1.709) |
| | | [.073;.611] | [33.0;1727.0] |
| *C. IV-matching* | | | |
| | 0.086*** | 0.326 | 849.5 |
| | (3.728) | (1.476) | (1.162) |
| | | [−.037; .689] | [−353.1; 2052.1] |

- Results are consistent in magnitude albeit less precisely estimated.
- IV matching: ratio of 2 matching estimators, *RF*/*FS*.

# In Stata

- psmatch2 depvar indepvar (if) (in), [options] is the main PSM command and integrates the PS and estimation steps within the one command.
- But you can also do the propensity score by running a probit or logit to estimate the propensity score and generate the predicted PS for all treated and control units:
  - probit depvar indepvar...
  - margins, dydx(*)
  - predict pscore
- Then use psmatch2 for the matching and estimation.
- Check common support with: psgraph
- And assess the balance of covariates with: pstest
- Practical Tips: when using a matching methods picking a sub-group of matches (NN for example), the order of the observations could affect the estimates. It is advisable to sort the data randomly before calling psmatch2.