Econometrics of Evaluation
Matching Methods

Jean-Noël Senne

Université Paris-Saclay

Problématiques Contemporaines du Développement
M2EIED
Université de Pau et des Pays de l'Adour
2023 - 2024

You said matching ?

- You have data on treated and untreated individuals but treatment comes out of a choice

→ **Endogenous selection**

- Very basic idea of matching : find a "match" for each treated individual in the untreated group

→ **Good match ?** A "twin" with similar observable characteristics (apart from being treated)

⇒ Conditional on observables, treatment is *as-good-as-random*

⇒ Matched "twins" mimic the **unobserved counterfactual**

- Matching is a set of evaluation methods that rely on :
  - The assumption that **selection is only based on unobservables**

  → Conditionnal Independance Assumption (CIA) or "Treatment ignorability"
  - **Non-parametric** estimation

  → No assumptions on functional forms

- Much progress during the 1990's but very less since... Why ?

→ Property (2) is attractive but property (1) much less so...

⇒ **What can we learn from matching methods ?**

Plan

Estimation under the CIA

Matching methods

Propensity score matching

Principle of matching

- Matching corrects (at least partly) for selection bias by **controlling for observable differences** between treated and control groups

- (Very strong) assumption that observable differences between treated and controls capture **all the determinants of selection**

- Conditional on observable variables X, treatment assignment is independent of potential outcomes (as-good-as random) :

$$(Y_0, Y_1) \perp T | X$$

$\Rightarrow$ **Conditional Independence Assumption (CIA)**

$\Leftrightarrow$ "Treatment ignorability"

$\Leftrightarrow$ "Uncounfoundedness assumption"

$\Leftrightarrow$ "Selection on observables"

What is the CIA ?

- Let's take 2 individuals with **similar observable characteristics**, but one is treated and the other is not.
- The CIA assumption : if their characteristics are similar, being treated or not is not due to differences in potential outcomes.
- The outcome of the untreated individual is a **good counterfactual** for the treated individual *in the absence of treatment* (and vice versa).

⇒ Comparing the "twins" provides an **unbiased estimator of the average treatment effect** (conditional on these observables).

- Note : For the ATT, un (little) less strong assumption is needed :

$$Y_0 \perp T | X$$

What do we need for matching to work ?

- CIA : there are no other characteristics than observables that influence both potential outcomes and treatment selection
  → **No selection on unobservables**.

- Matching methods require the existence of a **common support** (i.e. for all values of the observables there are both treated and untreated individuals) :

$$0 < P(T_i = 1|X_i) < 1$$

⇒ You need the “twin” to exist !

Regression as matching

- Let's assume that $E(Y_{i0}|X_i)$ is **linear** with respect to observable characteristics X_i :

$$E(Y_{i0}|X_i) = \alpha + \beta X_i$$

- Let's also assume **constant treatment effect** (Δ).
  The observed outcome writes :

$$Y_i = \alpha + \Delta T_i + \beta X_i + \epsilon_i$$

⇒ We can estimate treatment effect Δ with **OLS, "controlling" for observables.**

Note – *The assumption of a constant treatment effect can be relaxed by adding interaction terms between the treatment and the observables(observable heterogeneity)*.

The pros and cons for linear regression

Pros

- Simplicity
- **Well-know theoretical fundations** both for estimation and for statistical inference (cf. simple assumptions for being BLUE)
- Even if the distribution of outcome is not exactly a linear function of the observables, linearity often provides a **(very) good approximation**.

Cons

- Simplicity can also be a limit...
- If the conditional distribution deviates too much from a linear function, even the best linear approximation may yield **biased estimates**.
- Particular issue when treatment and control groups have **very different observable characteristics** (no common support restriction)

The limits of regression

- Under CIA, the best linear estimate of the counterfactual outcome ($\hat{Y}_0$) is the mean outcome of the control group, **"corrected" for differences in observables** between the 2 groups :

$$\hat{E}(Y_{i0}|T_i = 1) = \bar{Y}_0 + (\bar{X}_1 - \bar{X}_0)\hat{\beta}$$

where $\bar{X}$ and $\bar{Y}$ denote empirical means of X and Y

- The estimator of the **ATT** is :

$$\begin{aligned}\hat{\Delta} &= E(Y_{i1}|T_i = 1) - \hat{E}(Y_{i0}|T_i = 1) \\ &= \bar{Y}_1 - \bar{Y}_0 - (\bar{X}_1 - \bar{X}_0)\hat{\beta}\end{aligned}$$

⇒ If the difference $\bar{X}_1 - \bar{X}_0$ is too large, so will be the correction (very sensitive to the specification).

Note – *ATT = ATE due to constant treatment effect.*

Why is matching so different ?

- Matching methods get rid of the linearity assumption

→ **Non-parametric estimation**

- You just need to find in the data the **best possible counterfactual "twin"** for treated units based on observed characteristics X
  Then compare their outcomes.

- **But how do we find this "twin"** ?

⇒ Several matching methods for building this counterfactual !

Plan

Estimation under the CIA

Matching methods

Propensity score matching

Nearest-neighbor matching

- The simplest method is to **find a “twin”** for each treated individual.
- We then estimate the ATT by comparing the outcome Y_i of each treated individual with an untreated individual with exactly the same observable characteristics X_i.
- However, it might be hard to find an exactly identical individual in the control group (especially if X_i are continuous variables)
  → **The ”nearest-neighbor” is chosen.**

⇒ We need to define a **metric for the distance** between individuals

Metric for distance

Two main metrics are widely used :

- **Euclidean distance** : the distance between two individuals is the sum of the distance between all covariates.

$$d(I_i, I_j) = \sqrt{\sum_{k=1}^{P} (x_i^k - x_j^k)^2}$$

- **The Mahalanobis distance** is sometimes preferred, because it weights the distance by the variance-covariance matrix of the covariates X_i :

$$d(x_i, x_j) = (x_i - x_j)'\Sigma^{-1}(x_i - x_j)$$

⇒ You match each individual in the treated sample with his **"nearest-neighbor" in the untreated sample**

Estimating causal effects

- **Out of matching :**

$$\hat{Y}_{i0} = \begin{cases} Y_{i(j)0} & \text{if } T_i = 1 \\ Y_{i0} & \text{if } T_i = 0 \end{cases}$$

$$\hat{Y}_{i1} = \begin{cases} Y_{i1} & \text{if } T_i = 1 \\ Y_{i(j)1} & \text{if } T_i = 0 \end{cases}$$

- **The average treatment effect on the treated** $E(Y_{i1} - Y_{i0} | T_i = 1)$ is estimated by the average of matched differences in the treated sample :

$$\widehat{ATT} = \frac{1}{N_1} \sum_{E_1} (Y_{i1} - \hat{Y}_{i0})$$

- **The average treatment effect** $E(Y_{i1} - Y_{i0})$ is estimated by the average of matched differences in the whole sample :

$$\widehat{ATE} = \frac{1}{N} \sum_{i} (\hat{Y}_{i1} - \hat{Y}_{i0})$$

where N is the total sample size
N_1 is the size of the treated sample E_1,
Y is the observed outcome of individual i
$\hat{Y}$ is the outcome of i's nearest-neighbor

Types of nearest-neighbor matching

- Matching can be done :
  - **Without replacement** : an individual in the control group can only be matched once with an individual in the treatment group
  - **With replacement** : the whole sample is used each time, which allows several matches with the same individual.
- But matching without replacement has some limitations :
  - It is **very demanding** in terms of data (large sample)
  - Estimates are **sensitive to the order** in matching

Limitations of nearest-neighor matching

- Nearest-neighbor matching is one of the most widely used matching estimators.
  → Intuitive and does not require any parametric choice.

- **Two limitations :**
  - No control over the **quality of the match :** the concept of nearest-neighbor is relative by nature.
  - Uses very **few information :** each individual's counterfactual uses only 1 observation (looses the information brought by other individuals)

⇒ **Variants :** estimate $\hat{Y}_{i0}$ using more individuals from the control group

⇒ **Efficiency gains** if counterfactual for individual i is based on some average over several "similar" individual j

⇒ **Tradeoff between bias and variance**

M-closest matching

- Matching with a **fixed number M of nearest-neighbors.**

→ The counterfactual outcome of individual i is the **average outcome** of his M nearest-neighbors.

$$\hat{Y}_{i0} = \begin{cases} \frac{1}{M} \sum_M Y_{i(j)0} & \text{if } T_i = 1 \\ Y_{i0} & \text{if } T_i = 0 \end{cases}$$

$$\hat{Y}_{i1} = \begin{cases} Y_{i1} & \text{if } T_i = 1 \\ \frac{1}{M} \sum_M Y_{i(j)1} & \text{if } T_i = 0 \end{cases}$$

- **Refinement :** exclude individual for whom you cannot find a “twin” (or M “twin”) within a given distance d (to be fixed).

⇒ **But how do we select the M-neighbors ?**

Radius or Caliper matching

- You select all individuals in the control group who are **located in a fixed neighborhood** of individual i, for a given **neighborhood radius** h, i.e. such that :

$$\|X_i - X_j\| < h$$

- Need to define a metric (Euclidian or Mahalanobis)

Kernel matching

- The counterfactual of individual i is computed with a **kernel estimation.**
- All individuals in the control group are used, but **weighted by their distance** from the treated group :

$$\hat{Y}_{i0} = \frac{\sum_{E_0}^{k} K\left(\frac{\|X_i - X_k\|}{h}\right) Y_k}{\sum_{E_0}^{k} K\left(\frac{\|X_i - X_k\|}{h}\right)}$$

where $K(.)$ is the kernel function used (most often the gaussian density)

i is a treated individual of the treatment group E_1

k is an untreated individual of the control group E_0.

h is the window (*bandwidth*) of the kernel.

More on Kernel matching

- The bandwidth *h* gives the size of the neighborhood outside of which **weights are very small.**
- **The smaller the bandwidth**, the more likely the counterfactual will be estimated only for individuals in the control group whose observable characteristics are **very close**.
- **No set rule** for choosing the bandwidth (in practice, *ad hoc* choice or classic "rules of thumb".)
- Need to **check the sensitivity** of the results to various bandwidth *h*.

Note – *Have a look at the data to identify a "natural" threshold, but also test the sensitivity of the results to the chosen threshold, to finally choose the threshold that best arbitrates between precision and bias*.

How do we choose between matching methods ?

- Each of matching methods has **pros and cons**.
- The opposition between the simplest (nearest-neighbor) and the most complex (Kernel) method illustrates the standard **tradeoff between bias and precision**.
  - → Nearest-neighbor matching does not use all available information and thus reduces precision.
  - → Kernel function estimates are always more accurate but might generate mismatches, and thus bias.

⇒ **Check the sensitivity of your results to the method used !**

Can we be more simple ?

- For the CIA to hold, we want to use as much information as possible. You want to match on as many variables as possible
  → Difficult to find (a) close neighbor(s)
  → **High-dimensionality comparisons**

- Imagine that you have to cut the population into boxes according to the whole set of observable characteristics you have, and in each of these boxes you have to find a treatment and its control...

- It is shown that at finite distance, the estimators are **all the more biased** when the number of conditioning variables X is high (and even more when the conditioning variables are continuous).

⇒ Matching on the **propensity score**

⇒ **But what is the propensity score ?**

Plan

Estimation under the CIA

Matching methods

Propensity score matching

A revised CIA

- An important property shown by Rosenbaum and Rubin (1983).
- If the CIA assumption holds for the X variables, then the potential outcomes are also independent of the treatment **conditional on any function of** X
- **The propensity score** is such a function : it gives the probability of being treated conditional on the X observable characteristics :

$$p(X) = P(T = 1|X)$$

- We can then **revise the CIA assumption :**

$$(Y_0, Y_1) \perp T | X \quad \Longrightarrow \quad (Y_0, Y_1) \perp T | p(X)$$

Why is the propensity score useful ?

- This property helps to **reduce the dimensionality** of the comparisons.
- But the propensity score is unknown
  → **Need to be estimated**
- We often use **logit or probit specifications** to estimate the propensity score
- **Previous matching methods** (nearest neighbor matching, radius, or kernel) can all be applied on the estimated propensity score $\widehat{p}(X)$ to measure the distance between two observations.

Estimating the propensity score

- To take into account the fact that the score is bounded between $[0, 1]$, we often use a logistic form (or a probit) :

$$\widehat{p}(X) = \frac{exp(f(X))}{1 + exp(f(X))}$$

where $f(X)$ is a function of observable characteristics X. The simplest function has a linear form $f(X) = X\beta$

Note – *It is however recommended to use a polynomial approximation to get closer to the true distribution (see Hirano, Imbens & Ridder, 2003)*

Restriction to common support

- Remember that the CIA assumption is inherently based on the **existence of a common support**
  → We can find “twins”, i.e. individuals with similar values of observable characteristics

$$0 < P(T_i = 1 | X_i) < 1$$

- If not, impossible to find comparable treated and controls.

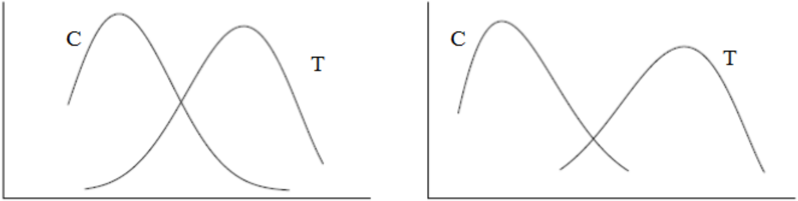
⇒ **What is a common support ?**
→ Area of $\widehat{p(X)}$ distribution over which this condition is verified (you can find both treated and controls)

How do we find the common support ?

- Important to check that this area is **large enough.**
- **Graphical analysis :** you can plot the distribution of the score over the two subsamples (treated and controls)

  → **Histograms** of the estimated probability or **density functions** of being treated for both treated and controls
- **Check that the overlap is large :** for each value of the score, there must be a sufficient number of individuals in both subsamples (treatment and controls)
  - If no overlap for some values of the observables, incorrect to use these individuals for estimation.
  - If you do not restrict to the common support, estimates may be biased !

Graphical analysis



Distributions du score de propension

How do we restrict to the common support ?

- Several methods for restricting to the common support.
- **Warning !** It changes the scope the estimation : the impact is now estimated on part of the population (i.e. whose observables are such that an overlap is observed for the two subsamples)
  → **Local treatment effect estimator**

Min/max method

- For the *ATT*, you can drop individuals from the control group whose propensity score is below the minimum observed in the treatment group.
- For the *ATE*, also do the reverse : drop individuals from the treatment group whose propensity score is higher than the maximum observed in the control group.

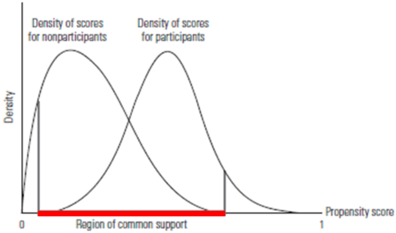
Trimming method

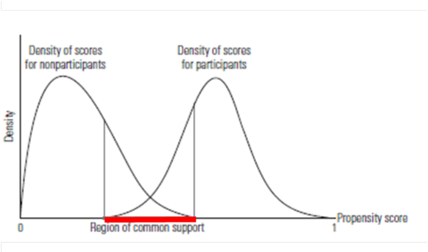
- Drop individuals whose propensity is too high or too low :

$$\alpha \leq p(X) \leq 1 - \alpha$$

- Imbens and Wooldridge (2008) suggest as a rule of thumb $\alpha = 0.01$ for the estimator to be efficient.

Graphical analysis


Density of scores for nonparticipants
Density of scores for participants
Density
Propensity score
0
1
Region of common support


Density of scores for nonparticipants
Density of scores for participants
Density
Propensity score
0
1
Region of common support

How do we choose the conditioning variables ?

- The CIA assumption requires a **sufficient number of observable characteristics**

- The choice of these variables is crucial
  → They must have an impact on the variable of interest **AND** on the probability of being treated

- However, **no precise rule** for selecting "good" variables
  - Do not use variables that are measured **"after" treatment** (they may also be affected by the treatment)
    → **Endogeneity issues**
  - For the common support to be large enough, conditioning variables must not explain the probability of being treated **"too much"**

To sump up

Steps for implementing matching methods :

1. Identify the **control group.**
2. Select a set of **conditioning variables.**
3. Choose an **estimation method** (a linear specification can be used first).
4. For propensity score matching, estimate the **propensity score.**
5. Check that the **common support** is large enough (and that matching does reduce differences between treated and controls).
6. Estimate the **average treatment effects** of interest
7. Check the **sensitivity** of your results to alternative matching methods

What can we learn from matching ?

- The robustness of matching methods relies on the validity of the **Conditional Independence Assumption (CIA)**
- This hypothesis is **very strong.**
- It is **very demanding** in terms of data (too little information won't eliminate the selection bias)
- **Standard errors** are hard to compute (bootstrap)

⇒ **Matching methods are often matched with other evaluation methods !**