



SEMINARI INFERENZA CAUSALE

Dipartimento di Sanità Pubblica e Malattie Infettive

Aula Scuola di Specializzazione in Statistica Sanitaria e Biometria

AULA C1 ore 12:00 - 14:00

Martedì 29 Novembre: Introduzione alla causalità

Dr.ssa MARGHERITA MORETTI

Dottorato in Scuola di Scienze Statistiche
Sapienza Università di Roma

Dr.ssa LAURA MONTELISCIANI

Dottorato in Sanità Pubblica Biostatistica ed Epidemiologia
Università degli Studi di Milano-Bicocca

Martedì 6 Dicembre: Metodi per l'inferenza causale

Metodi per l'inferenza causale

Margherita Moretti

margherita.moretti@uniroma1.it

Laura Montelisciani

l.montelisciani@campus.unimib.it

PER SCARICARE LE SLIDE:
https://github.com/MMargherita/seminari_DSPMI

Ricetta per la causalità

1. Porre una domanda di ricerca causale

→ Tradurre l'intuizione in **domande di ricerca causali** e in **quantità misurabili**

2. Rappresentare le conoscenze a priori

→ Disegnare un **DAG** e chiarire quali sono i ruoli delle variabili considerate.

3. Identificare le minacce alla causalità

4. Scegliere un metodo per affrontare queste minacce

I metodi principali possono essere distinti in:

1. *Selection on observables*
2. *Selection on unobservables*

Selection on observables

H_p: siamo in grado di osservare tutti i possibili confondenti della relazione fra trattamento T e outcome Y

→ Assunzione 3: Exchangeability (o No Unmeasured Confounding o Ignorability)



Negli studi **osservazionali** si osserva, per ogni unità, la **presenza** o **l'assenza** del trattamento. Dato che il trattamento **non è assegnato in maniera casuale**, i due gruppi di trattati e non trattati potrebbero essere **sistematicamente diversi** rispetto ad alcune caratteristiche (osservate e non).

Per rendere i gruppo dei trattati e non trattati confrontabili

In fase di progettazione: randomizzazione (la probabilità di essere assegnati al trattamento è la stessa per ogni unità del campione)

In fase di analisi: correggere le differenze tra i due gruppi controllando le variabili pre-trattamento

Propensity Score

Propensity score PS(X)

E' un numero che indica la **probabilità**, per ogni individuo, **di essere trattato**, a fronte delle **caratteristiche individuali di base (pre-trattamento)**

$$PS(X) = \Pr(T = 1 | X = x) = E[T|X]$$

E' una sintesi delle caratteristiche di base X che confondono la relazione fra T e Y

In caso di **trattamento binario**, il **PS(X)** può essere stimato tramite una **regressione**:

-**Risposta binaria: trattamento T**

-Un serie di **covariate X: confondenti** della relazione fra trattamento e outcome

$$P(T = 1 | X) = PS(X) = \frac{\exp(\alpha + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)}{1 + \exp(\alpha + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)}$$

$$\text{logit}[PS(X)] = \alpha + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k$$

I **valori predetti dal modello** rappresentano la **probabilità**, per ogni individuo, **di essere trattato, dati i confondenti**

Il propensity score che abbiamo ottenuto, per ogni unità, ha una proprietà di bilanciamento:

Condizionatamente al propensity score, il gruppo dei trattati e quello dei non trattati hanno la stessa distribuzione dei confondenti

Condizionatamente al propensity score:

1. L'assegnazione al trattamento è indipendente dai confondenti
2. Il potential outcome è indipendente dall'assegnazione al trattamento

2. Il potential outcome è indipendente dall'assegnazione al trattamento, condizionatamente a X (CIA) → condizionatamente a PS(X)

$$Y_1, Y_0 \perp T | X \quad \text{allora} \quad Y_1, Y_0 \perp T | PS(X)$$

$$E(Y | X, T=1) = E(Y_1 | X) = E(Y_1 | PS(X))$$

$$E(Y | X, T=0) = E(Y_0 | X) = E(Y_0 | PS(X))$$

Possiamo stimare l'effetto causale condizionatamente a PS!

Diversi metodi basati sul Propensity Score (PS)

-Matching (PSM): utilizzare il PS come metrica della distanza, più le unità sono vicine più sono simili in termini di confondenti

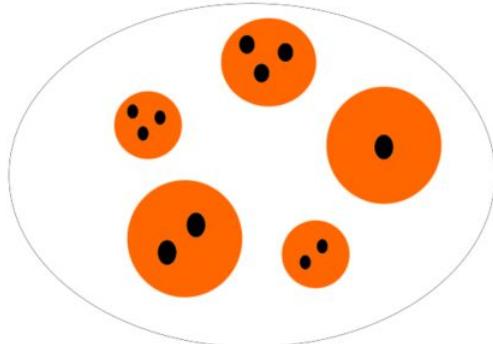
-Weighting (IPTW): calcolare un peso da assegnare a ogni unità, in funzione del PS, utilizzato per bilanciare trattati e non trattati rispetto alla distribuzione dei confondenti

→ **ridurre/eliminare le differenze nella distribuzione dei confondenti
fra trattati e non trattati, per renderli confrontabili**

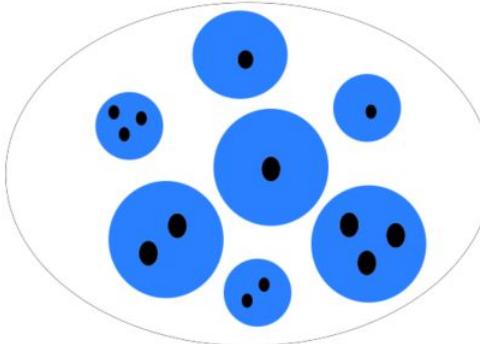
Matching

The matching game

TRATTATI

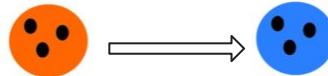


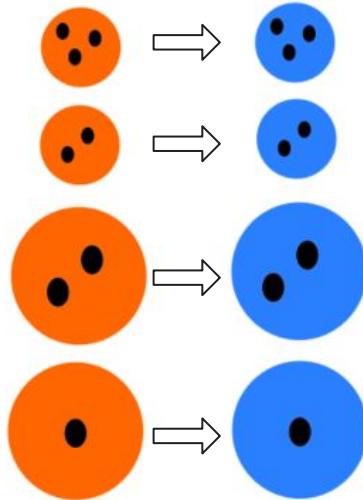
NON TRATTATI



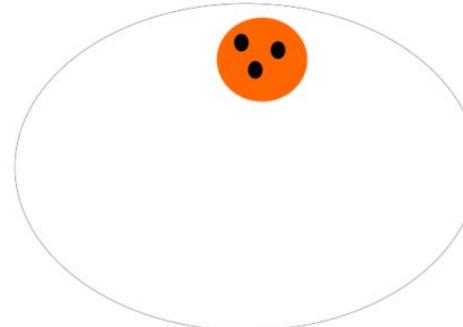
Le unità sono
distinte per:
colore (T),
dimensione (X1)
pallini neri (X2)

Per ogni unità del gruppo dei trattati, cerchiamo l'unità più simile nel gruppo dei non trattati

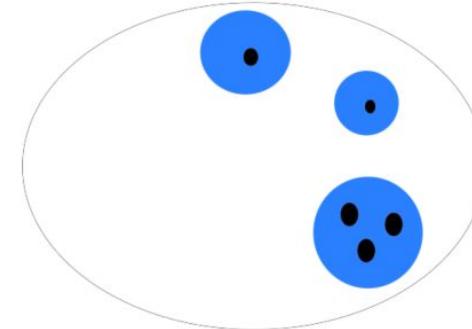




TRATTATI



NON TRATTATI



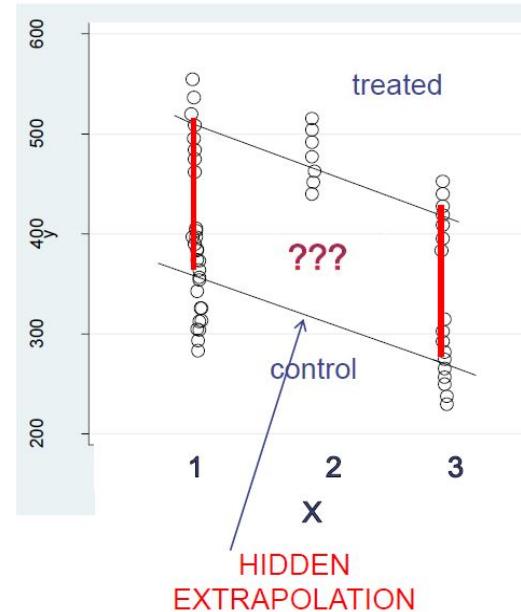
**Le unità rimaste nei cerchi saranno scartate dall'analisi
(non si possono confrontare unità con caratteristiche diverse)**
Cioè, **condizioniamo per il common support!**

Costruiamo una **serie di coppie di unità** costituite dal **match** fra **trattati e non trattati**, a fronte delle caratteristiche di base.

Otteniamo delle **coppie simili** a quelle che avremmo in un **random trials** → l'unica **caratteristica** per cui **differiscono** è il **trattamento**

L'idea alla base del **matching** è quella di scegliere, secondo delle **regole** ben precise, **una (o più) unità non trattata per ogni unità trattata**

In un **modello di regressione**, se manca il common support rischiamo di **estrapolare unità non presenti nel campione!**



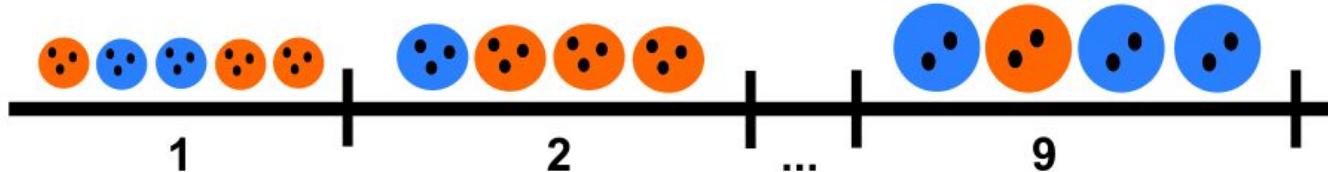
Tipi di matching: esatto, PSM, NN, K-NN, NN-C, NN-R, KM

Matching esatto

Si associa ad ogni unità trattata un'unità non trattata con le stesse identiche caratteristiche X

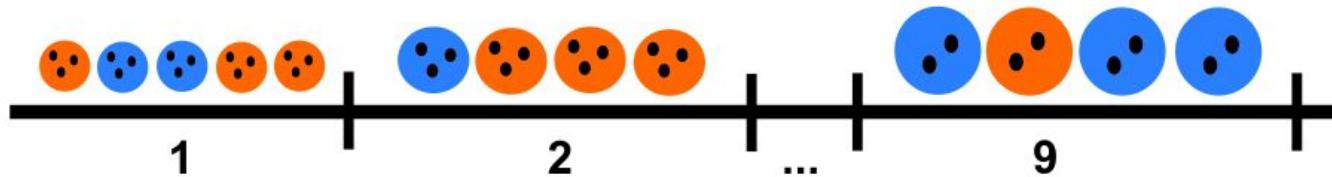
Le **caratteristiche X** vengono utilizzate per formare dei “**blocchi**”; le **unità** che appartengono allo **stesso blocco** sono **identiche (su X)**.

Immaginando di creare **un blocco per ogni combinazione di dimensione (X1) e # pallini neri (X2)**, otteniamo **9 blocchi distinti**; in ogni blocco, le unità differiscono solamente rispetto al trattamento



**In ogni blocco, confrontiamo
il mean outcome fra trattati e non trattati.**

Possiamo calcolare l'ATT come
media pesata del mean outcome in ogni blocco,
con **peso** uguale alla **proporzione di trattati in**
ogni blocco sul totale delle unità trattate



Se avessimo **20 unità trattate**, in totale, potremmo stimare l'ATT come:

$$\text{ATT} = \frac{3}{20} [\bar{Y}_1 - \bar{Y}_0]^{\text{blocco } 1} + \frac{3}{20} [\bar{Y}_1 - \bar{Y}_0]^{\text{blocco } 2} + \dots + \frac{1}{20} [\bar{Y}_1 - \bar{Y}_0]^{\text{blocco } 9}$$



Curse of dimensionality

- il **matching esatto non è perseguiibile** quando disponiamo di **confondenti continui** (soluzione: discretizzare in classi, se possibile)
- quando disponiamo di **molti confondenti** il **matching esatto** diventa **infattibile / non pratico** (es. con 10 variabili binarie arriviamo a 1024 blocchi)



Soluzione: utilizzare una metodologia per sintetizzare i confondenti → **Propensity Score!**

Propensity Score Matching (PSM)

Implementare il **matching** fra le **unità trattate e non trattate** **sulla base del Propensity Score (PS)**



il **PS** è una **variabile unidimensionale**,
in grado di **sintetizzare le caratteristiche**
per cui trattati e non trattati potrebbero differire
rispetto al **set di confondenti X**

Per stimare l'ATT:
confrontiamo l'outcome, per ogni unità trattata, con l'outcome di una (o più) unità simile, fra i non trattati

La somiglianza è definita rispetto al valore del propensity score

L'exchangeability e la positivity permettono di scrivere:

$$\begin{aligned}ATT &= E(Y_1 - Y_0 \mid T = 1) = E(Y_1 \mid T = 1) - E(Y_0 \mid T = 1) \\&= E_X[E(Y_1 \mid X = x, T = 1) - E(Y_0 \mid X = x, T = 1)] \\&= E_X[E(Y_1 \mid X = x, T = 1) - E(Y_0 \mid X = x, T = 0)]\end{aligned}$$

possiamo anche **sostituire X con il PS**

$$= E_{PS(x)}[E(Y_1 \mid PS(x) = ps(x), T = 1) - E(Y_0 \mid PS(x) = ps(x), T = 0)]$$

Tipi di PSM

Nearest Neighbor (NN)



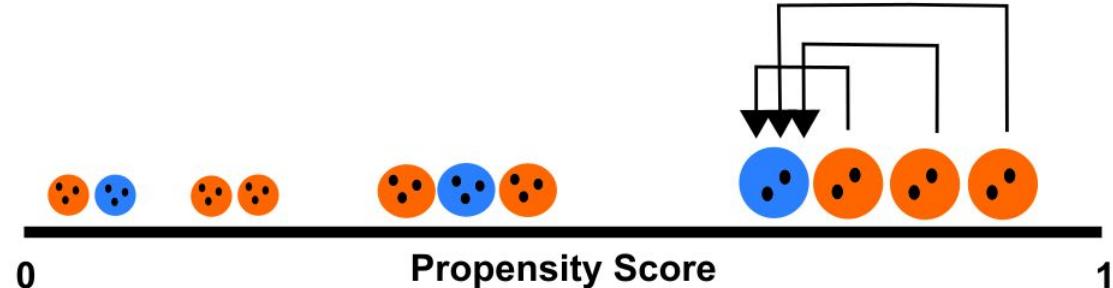
Per ogni unità trattata, cerchiamo l'unità non trattata più vicina sulla base del valore più simile del Propensity Score

Due varianti:

-senza ripetizione: ogni unità non trattata viene considerata una sola volta

-con ripetizione: un'unità non trattata può essere usata più di una volta come match per le unità trattate

Nearest Neighbor con ripetizione



Il NN matching con ripetizione, se confrontato con il NN senza ripetizione, comporta un trade-off tra bias e varianza

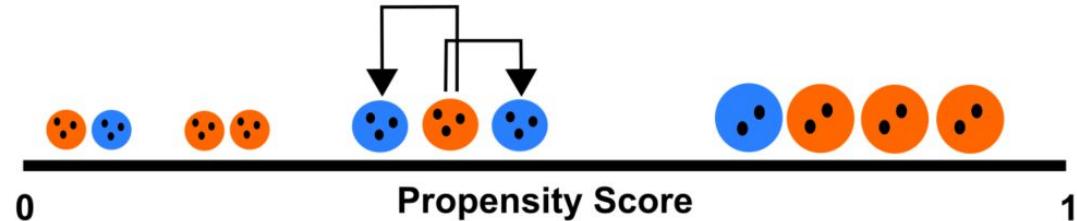


La qualità media del matching aumenta e il bias diminuisce



L'utilizzo della stessa unità non trattata per più di un'unità trattata aumenta la varianza degli stimatori dell'ATT

K-Nearest Neighbors (K-NN)

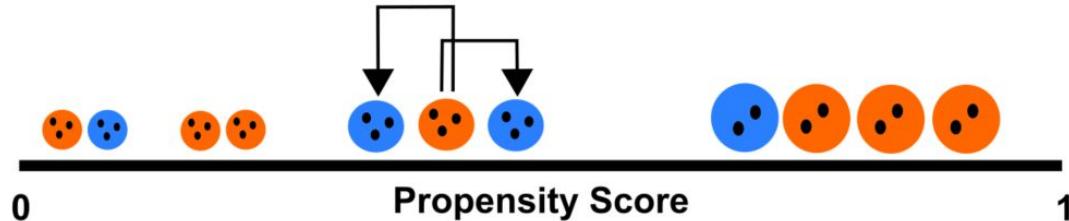


Per ogni unità trattata, scegliamo più di un'unità non trattata nearest neighbor in termini di PS

Anche il K-NN matching comporta un trade-off:  

L'utilizzo di un **maggior numero di informazioni** (unità non trattate) per costruire il controfattuale di ciascun trattato implica una **riduzione della varianza** ma corrisponde ad un **aumento del bias** che deriva dall'utilizzo di **match meno precisi** (il secondo, il terzo... più vicino)

K-Nearest Neighbors (K-NN)

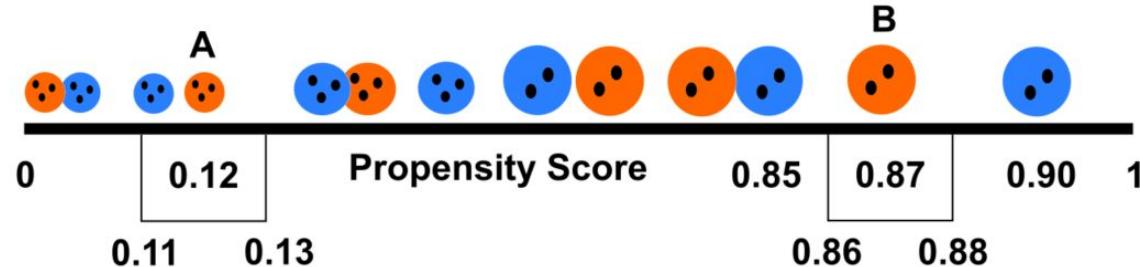


L'outcome di ogni unità trattata viene confrontato con la media ponderata degli outcome dei K-Nearest Neighbors

Ciò comporta scegliere i pesi da attribuire a ciascuno dei k vicini:

1. attribuire loro **pesi uguali**
2. **pesi proporzionali alla distanza** dall'unità trattata abbinata in termini di **PS**

Nearest Neighbors with CALIPER (NN-C)

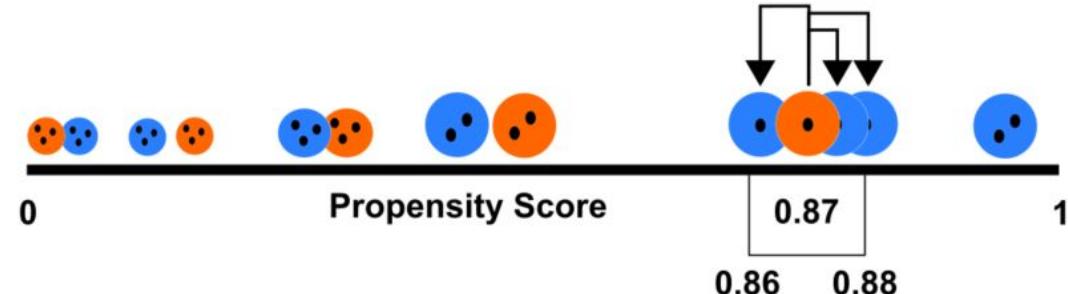


Ogni **unità trattata** è abbinata ad una **unità non trattata** se e solo se la **differenza** tra il **PS** dell'**unità trattata** e dell'**unità non trattata NN**, in **valore assoluto**, non è superiore al valore assegnato al **caliper**

Es. con **caliper pari 0.01**: l'**unità trattata A** verrà **abbinata** alla sua **NN** non trattata poiché la distanza del loro **PS** è entro lo $|0.01|$; l'**unità trattata B**, invece, non verrà **abbinata** poiché non ha **NN** entro lo **0.01**

(impostare il caliper a $0.25 \times \text{SD}$ del PS?)

Nearest Neighbors with RADIUS (NN-R)



Una **variante** del **caliper matching** è il cosiddetto **radius matching**

L'idea di base di questo metodo è che ad ogni **unità trattata** si abbina, non solo l'**unità non trattata** più vicina all'interno di ogni caliper, ma **tutte le unità non trattate all'interno del caliper**

NN-C e NN-R



si controlla direttamente il common support
difficile sapere a priori quale range sia ragionevole

Kernel Matching (KM)

Tutti i trattati sono abbinati a una media ponderata di tutti i non trattati, con pesi che dipendono dalla distanza tra i PS dei trattati e dei non trattati

Il modo in cui vengono calcolati i pesi dipende dalla **specifica funzione kernel** adottata (es. Gaussian Kernel)



utilizzando tutte le informazioni del set delle unità non trattate per ogni unità trattata si riduce la varianza dello stimatore



utilizzando anche unità non trattate lontane dalle unità trattate, può aumentare il bias

Decision	Bias	Variance
Nearest neighbour matching: multiple neighbours/single neighbour with caliper/without caliper	(+)/(-) (-)/(+)	(-)/(+) (+)/(-)
Use of control individuals: with replacement/without replacement	(-)/(+)	(+)/(-)
Choosing method: NN matching/Radius matching KM	(-)/(+) (+)/(-)	(+)/(-) (-)/(+)

Caliendo, M. and Kopeinig, S. (2008), SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING. Journal of Economic Surveys, 22: 31-72.

<https://doi.org/10.1111/j.1467-6419.2007.00527.x>

STEP per il PSM

1. Capire se l'**assunzione di exchangeability** è plausibile sulla base della teoria conoscenza a priori (e ai dati a disposizione)
2. Per ogni unità, **stimare la probabilità di ricevere il trattamento in funzione dei confondenti misurabili X** (es. modello logistico). Utilizzare i **valori predetti** per generare il **PS(X)**
3. Restringere il campione al **common support**
4. Realizzare il **matching fra le unità**: per ogni unità dei trattati, trovare una (o più) unità fra i non trattati sulla base del PS(X)
5. Controllare il **bilanciamento**. Se non è stato raggiunto, ripetere 2., 3., 4., 5. Es., includere interazioni e termini di ordine superiore nel modello per il PS(X) (passo 2.) oppure modificare l'algoritmo di matching (passo 5.)

Inverse Probability of Treatment Weighting (IPTW)

Idea simile a quella dei pesi campionari

I **pesi campionari**, utilizzati nelle indagini statistiche, hanno l'obiettivo di **correggere le differenze**, in termini di caratteristiche osservate, che ci sono fra **campione intervistato** e la **popolazione di riferimento**.

Infatti, i pesi di campionamento vengono utilizzati per correggere il sovra(sotto)campionamento di unità con caratteristiche specifiche ed arrivare ad un campione che riflette la popolazione di interesse

Utilizzando i pesi campionari è possibile generalizzare all'intera popolazione i risultati di ricerca osservati nel campione

Allo stesso modo, i propensity score, se utilizzati come pesi (IPTW)

permettono al nostro **campione in studio** di riprodurre una
“pseudo-popolazione” composta di unità trattate e non trattate
che **non differiscono nelle distribuzioni dei confondenti**



**I pesi eliminano le differenze di composizione
nei gruppi di trattati e non trattati,
modificando la popolazione di riferimento**

**Ogni individuo contribuisce alle stime
in funzione del suo PS(X)**

Nella stima dell'**ATE**: $w_i = \frac{T_i}{PS_i(X)} + \frac{(1 - T_i)}{1 - PS_i(X)}$

Nella stima dell'**ATT**: $w_i = T_i + (1 - T_i) * \frac{PS_i(X)}{1 - PS_i(X)}$

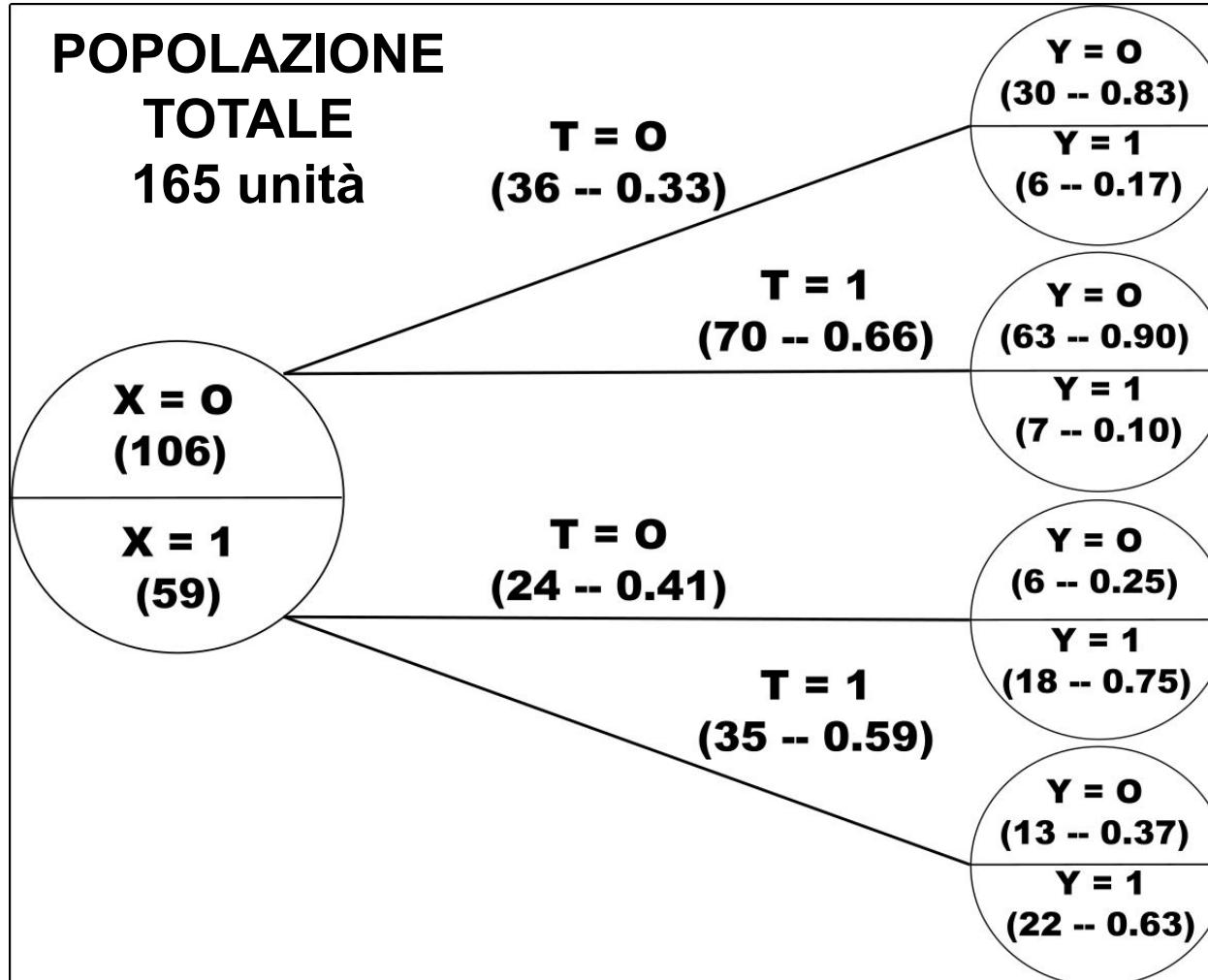
Pesi (w_i) estremi, soluzioni:

1. *Trimming*: troncare i pesi ad un definito percentile (95th, 99th?)
2. *Stabilized weights*

Ad es., per l'ATE:

$$w_i = P(T = 1) * \frac{T_i}{PS_i(X)} + P(T = 0) * \frac{(1 - T_i)}{1 - PS_i(X)}$$

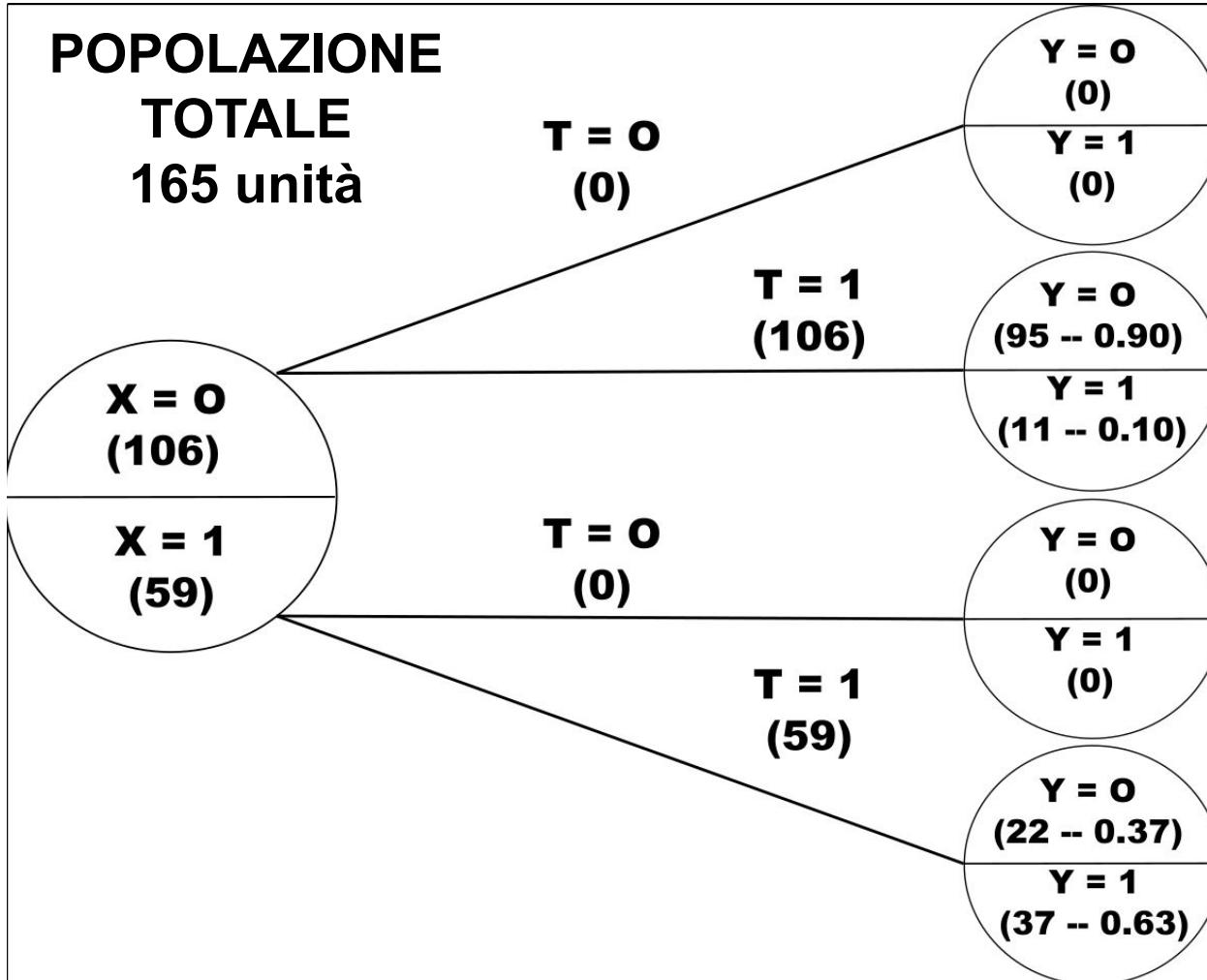
POPOLAZIONE TOTALE 165 unità



$$P(T=1|X=0) = 0.66$$

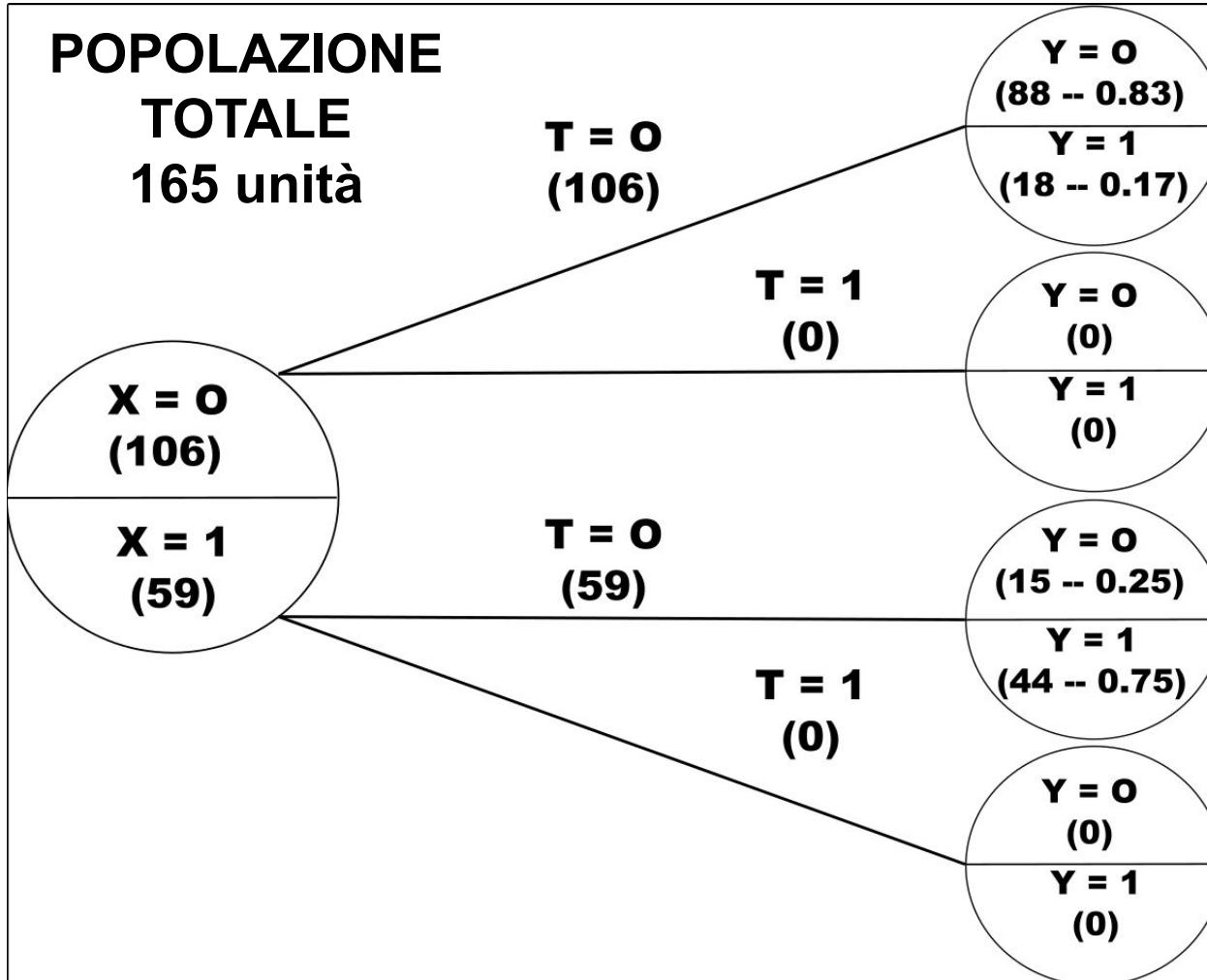
$$P(T=1|X=1) = 0.59$$

POPOLAZIONE TOTALE 165 unità



TUTTI
TRATTATI

POPOLAZIONE TOTALE 165 unità



**NESSUN
TRATTATO**

PSEUDO-POPOLAZIONE

330 unità

$T = 0$ (36 -- 0.33)
(106)

(70 -- 0.66) $T = 1$ (106)

$X = 0$
(106)

$X = 1$
(59)

$T = 0$ (59) (24 -- 0.41)

(35 -- 0.59) $T = 1$ (59)

$Y = 0$
0.83

$Y = 1$
0.17

$Y = 0$
0.90

$Y = 1$
(0.10)

$Y = 0$
0.25

$Y = 1$
0.75

$Y = 0$
0.37

$Y = 1$
0.63

$$1 / 0.33 = 3$$

$$3 * 36 = 108$$

$$108 * 0.17 = 18$$

$$1 / 0.66 = 1.5$$

$$1.5 * 70 = 105$$

$$105 * 0.1 = 11$$

$$1 / 0.41 = 2.4$$

$$2.4 * 24 = 58$$

$$58 * 0.75 = 44$$

$$1 / 0.59 = 1.7$$

$$1.7 * 35 = 60$$

$$60 * 0.63 = 37$$

$$1 / 0.33 = 3$$
$$3 * 36 = 108$$
$$108 * 0.17 = 18$$

Post IPTW:

$$\mathbb{E}[Y_1] = (11 + 37) / 165 = 0.29$$

$$\mathbb{E}[Y_0] = (18 + 44) / 165 = 0.38$$

$$\mathbb{E}[Y_1 - Y_0] = 0.29 - 0.38 = -0.09$$

$$1 / 0.66 = 1.5$$
$$1.5 * 70 = 105$$
$$105 * 0.1 = 11$$

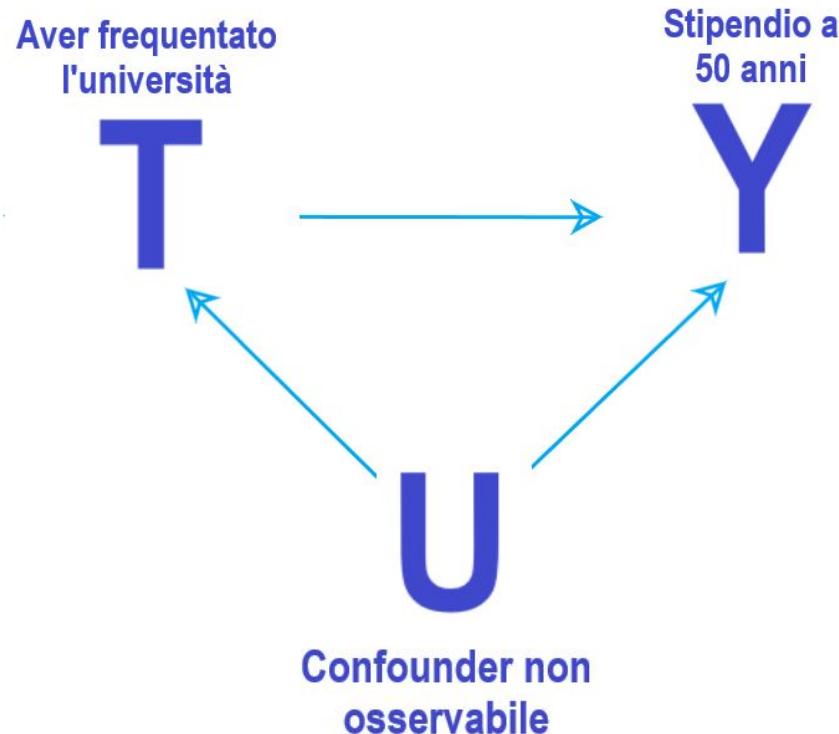
$$1 / 0.41 = 2.4$$
$$2.4 * 24 = 58$$
$$58 * 0.75 = 44$$

$$1 / 0.59 = 1.7$$
$$1.7 * 35 = 60$$
$$60 * 0.63 = 37$$

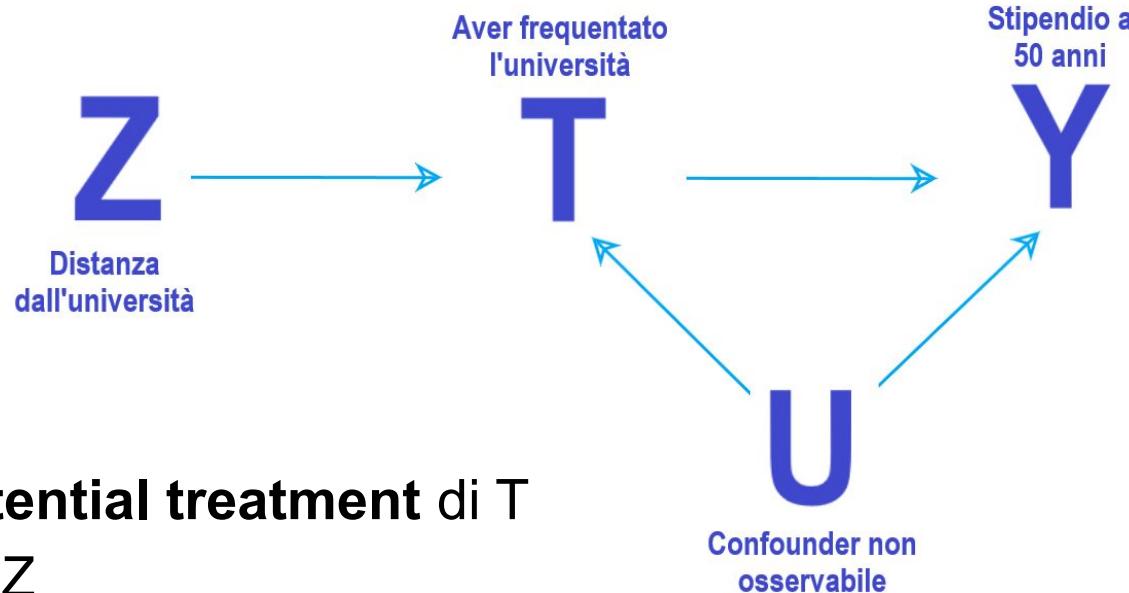
Selection on unobservables

Instrumental Variable

Le **Variabili Strumentali** permettono di aggiustare per tutti i confondenti, osservabili e non, presenti tra il trattamento e l'outcome.



1. La variabile Z ha un effetto causale diretto su T;
2. La variabile Z ha un effetto su Y solo attraverso T, ossia non c'è effetto causale diretto da Z a Y;
3. La variabile Z non condivide cause comuni con Y.



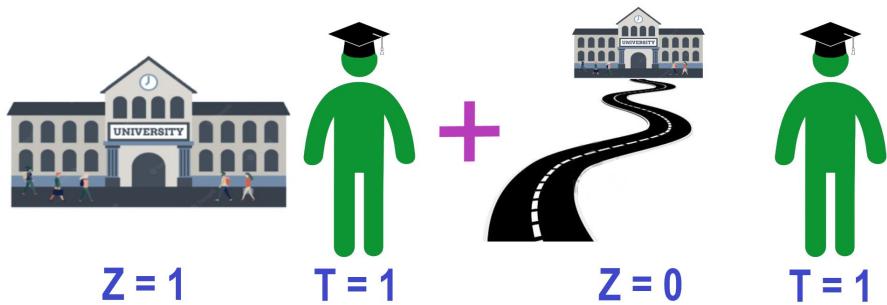
T_z è il potential treatment di T rispetto a Z

Se T e Z sono dicotomiche si generano 4 sottogruppi della popolazione:

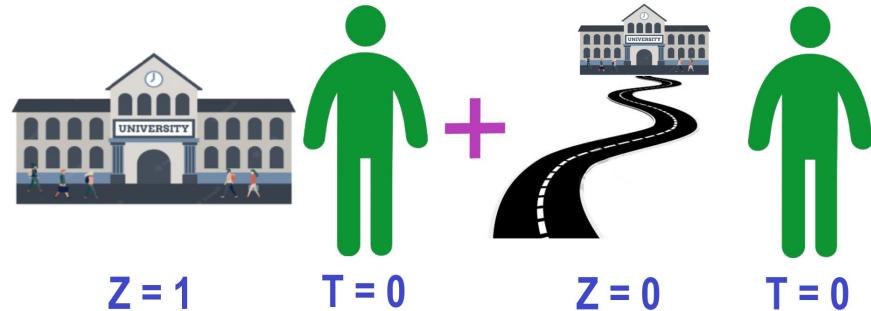
- **Always Takers** → $T_0 = 1$ & $T_1 = 1$, ossia qualunque sia il valore che assume lo strumento le unità saranno comunque trattate.
- **Never Takers** → $T_0 = 0$ & $T_1 = 0$, ossia qualunque sia il valore che assume lo strumento le unità non saranno trattate.
- **Compliers** → $T_0 = 0$ & $T_1 = 1$, ossia se ricevono lo strumento allora ricevono anche il trattamento, oppure, se non ricevono il trattamento non ricevono neanche il trattamento.
- **Defiers** → $T_0 = 1$ & $T_1 = 0$, ossia se ricevono lo strumento allora non ricevono anche il trattamento, oppure, se non ricevono lo strumento ricevono il trattamento.

Z = scuola vicino casa (1=si, 0=no), T = laurea (1=si; 0=no)

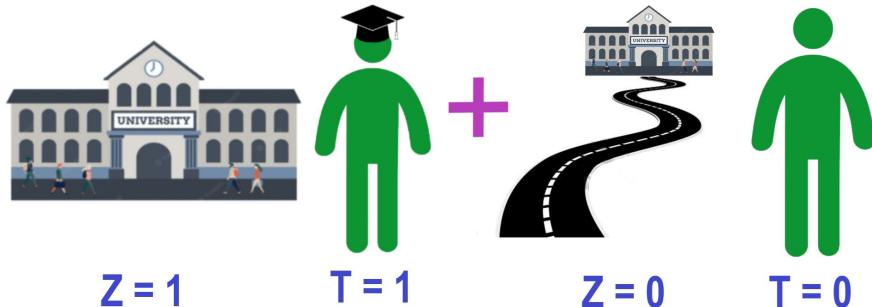
Always Takers



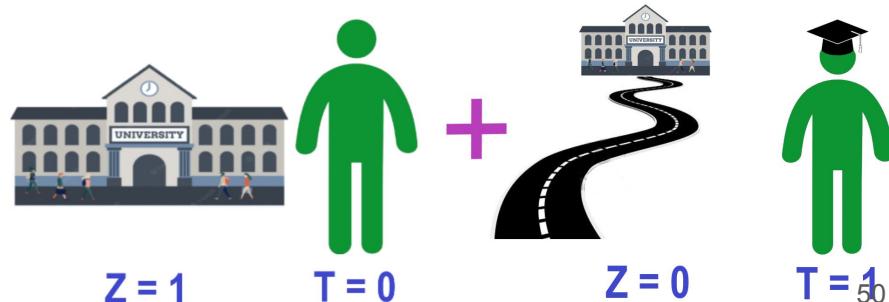
Never Takers



Compliers



Defiers



Ragioniamo sull'effetto causale in ogni strato:

- **Always Takers**
 - **Never Takers**
 - **Defiers**: assumo che non esistono perché hanno un comportamento controintuitivo;
 - **Compliers**: posso stimare l'**LATE**.
- } Non è possibile stimare alcun effetto causale perché il trattamento è insensibile allo strumento;



I **potential treatment T_z** non sono osservabili, ad esempio se osservo $T_1 = 1$ non so distinguere se si tratta di un Always Takers o di un Compliers.

Per stimare il **LATE** sono necessarie le seguenti assunzioni:

1. **Monotonicity** (non esistono i defiers):

$$P(T_0 = 1 \& T_1 = 0) = 0$$

2. **Esistenza dei compliers:**

$$P(T_0 = 0 \& T_1 = 1) > 0$$

3. **Uncounfunded Instrument:** (T_0, T_1) La distribuzione dello strumento Z negli strati del Trattamento T è sempre la stessa.

$$P(Z = z | T_0 = t_0) = P(Z = z | T_1 = t_1) = P(Z = z)$$

4. **Mean exclusion restrict:** In media i potential outcome, se condizionati allo strato g, non dipendono dallo strumento Z.

$$P(Y_T | Z=0, G=g) = P(Y_T | Z=1, G=g)$$

Se le assunzioni sono plausibili si può stimare il LATE utilizzando delle quantità osservabili:

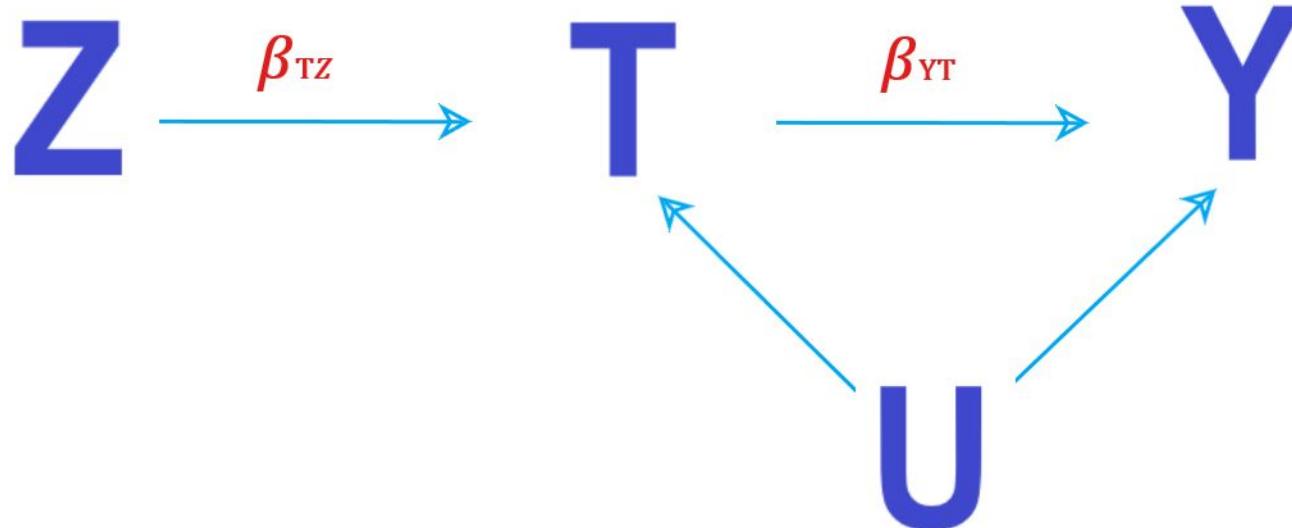
Effetto dello strumento Z sull'outcome Y

$$\text{LATE} = E[Y_1 - Y_0 \mid G = c] = \frac{E[Y \mid Z=1] - E[Y \mid Z=0]}{E[T \mid Z=1] - E[T \mid Z=0]}$$

Effetto dello strumento Z sul trattamento T



Stimare il LATE di T su Y → β_{YT}



⚠ $E[Y|T] = \mu_Y + \beta_{YT} * T$

→ Distorto da U (impossibile aggiustare perché U è non osservabile)

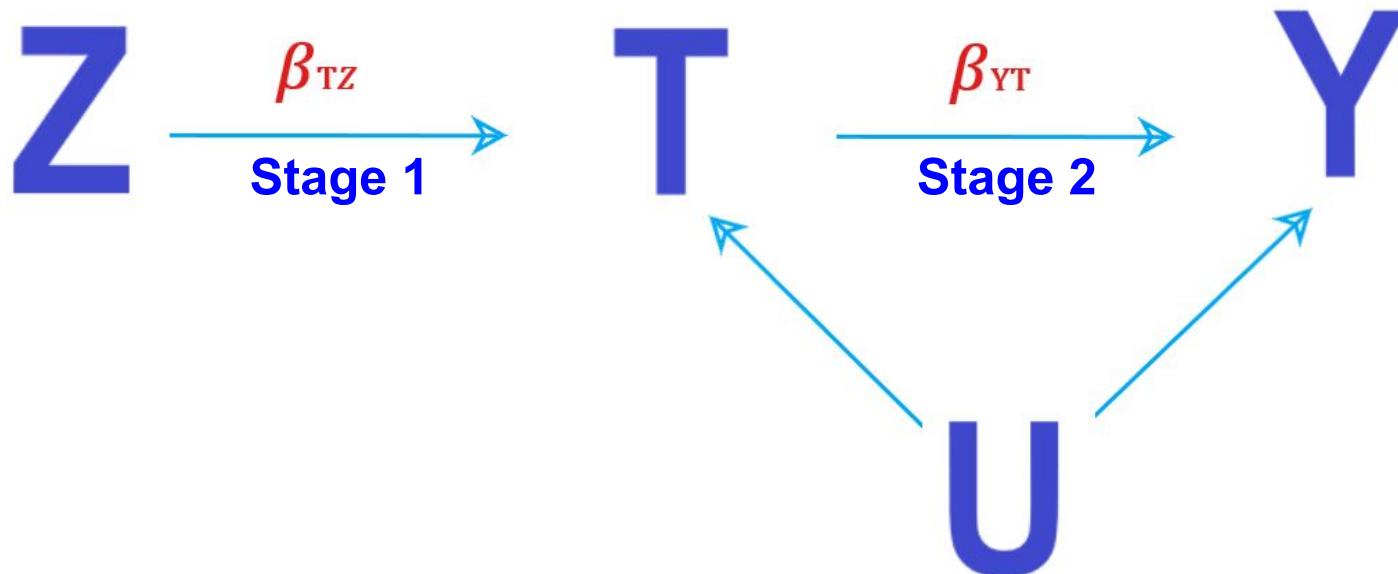


→ 2 Stage Least Square 2SLS

Obiettivo: Stima di β_{YT}

Stage 1: regressione di Z su T → Stima di β_{TZ}

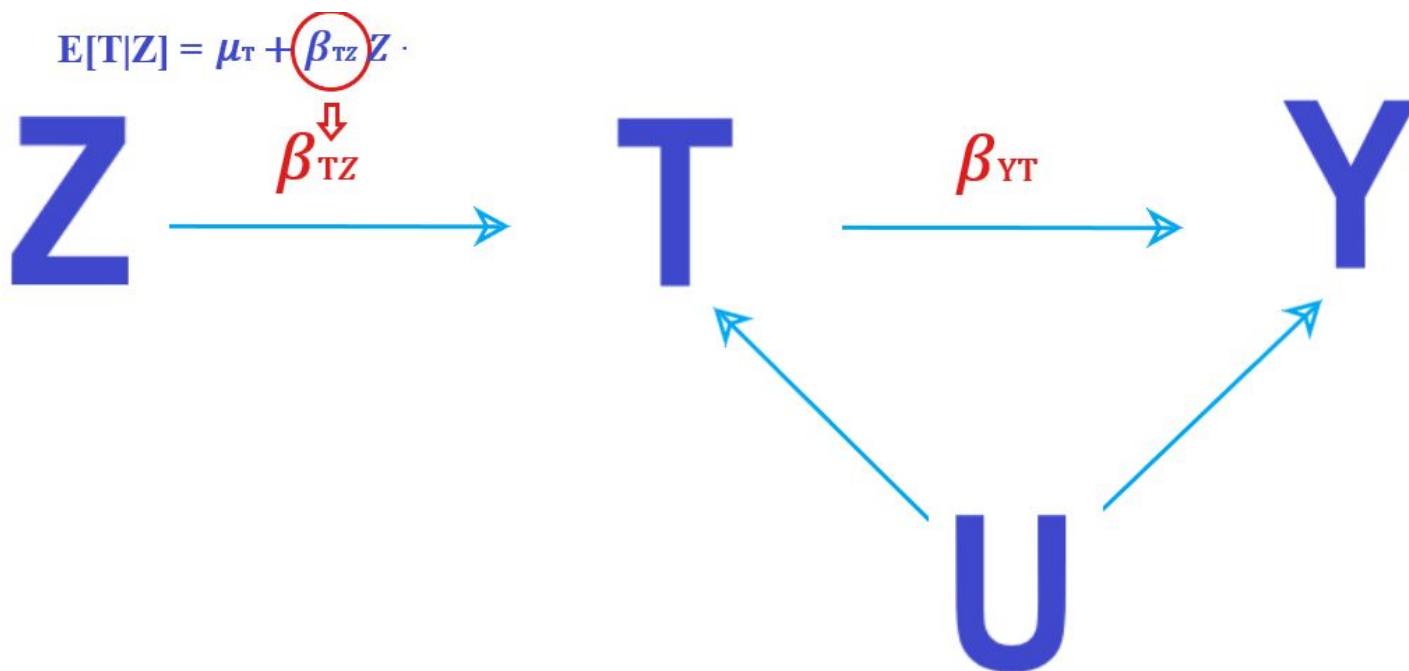
Stage 2: regressione di T_{stimato} su Y → Stima di $\beta_{YT_{\text{stimato}}}$



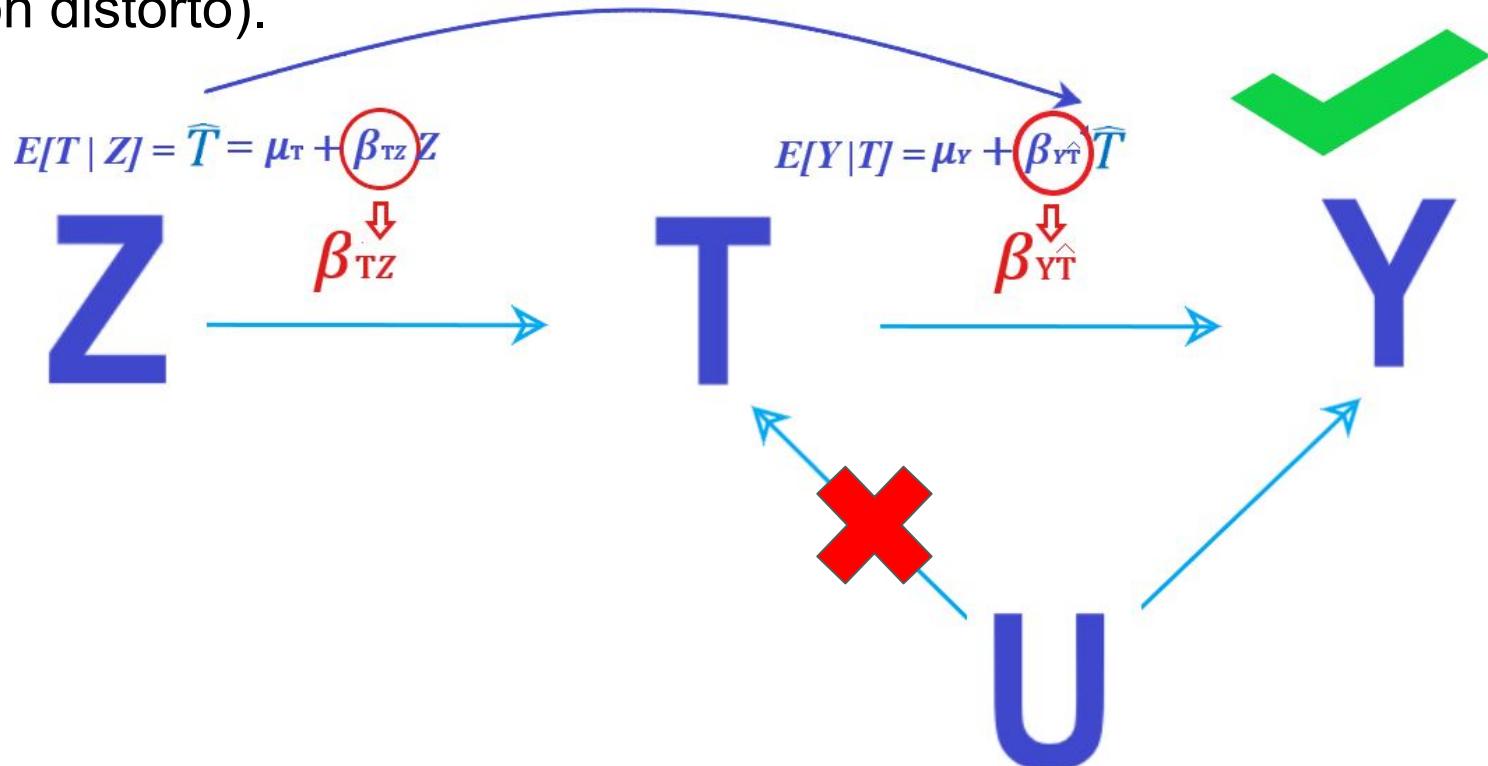


2 Stage Least Square

Stage 1: Stimo β_{TZ} con una prima regressione di Z su T



Stage 2: utilizzo $\widehat{\beta}_{TZ}$ stimato allo Stage 1 per stimare \widehat{T} (non distorto) che utilizzerò come variabile in una seconda regressione per stimare $\widehat{\beta}_{YT}$ (non distorto).



→ Stima del LATE ottenuta col 2SLS: effetto indiretto dello strumento

Effetto dello strumento Z sull'outcome: $\beta_{TZ} * \beta_{YT}$

$$\beta_{YT} = \text{LATE} = E[Y_1 - Y_0 \mid G=c] = \frac{E[Y \mid Z=1] - E[Y \mid Z=0]}{E[T \mid Z=1] - E[T \mid Z=0]} = \frac{\beta_{YZ}}{\beta_{TZ}}$$

2° stage (utilizzando la stima di T del 1° stage)

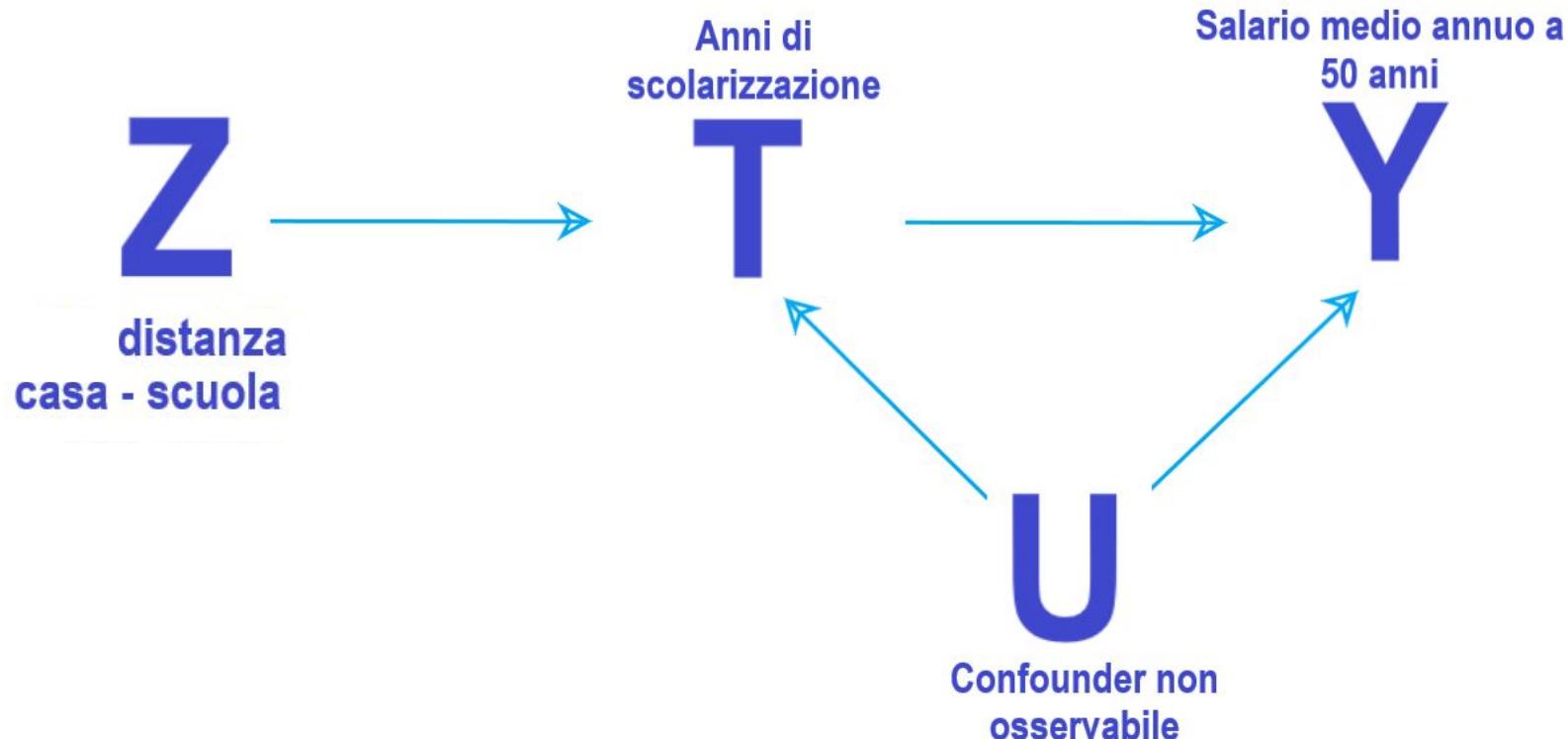
Effetto dello strumento Z sull'outcome Y

Effetto dello strumento Z sul trattamento T

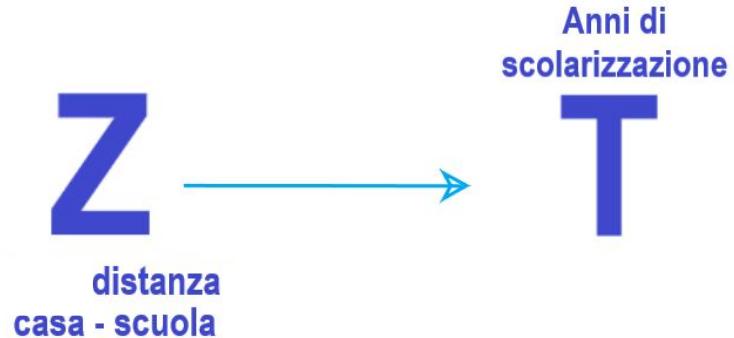
Instrumental Variable: esempio



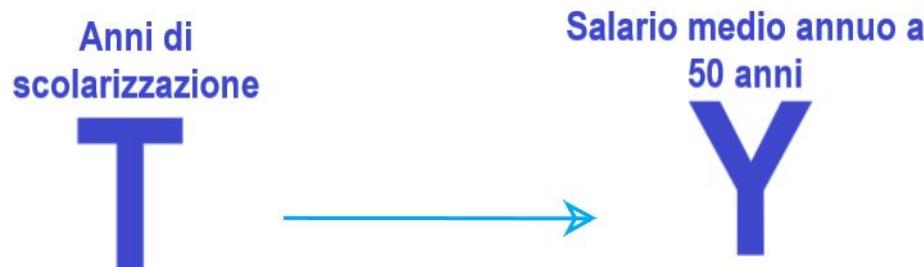
Vogliamo stimare l'effetto di un anno in più di scuola sul salario medio annuo percepito a 50 anni, utilizzando come strumento la distanza in Km dalla scuola.



1° stage: $\widehat{T} = E[T | Z] = \mu_T - \beta_{TZ} * Z$



2° stage: $\widehat{Y} = E[Y | \widehat{T}] = \mu_Y + \beta_{YT} \widehat{T} + \varepsilon_Y$



1° stage: $\widehat{T} = E[T | Z] = \mu_T - \beta_{Tz} * Z$

$$\widehat{AnniScuola} = \mu_{anniScuola} - \beta_{anniScuola, Distanza} * Distanza$$
$$\rightarrow \beta_{anniScuola, Distanza} = -0.2$$

Una riduzione unitaria in Z, ossia di un Km di distanza, provoca un aumento di 0.2 di T, ossia di anni di scolarizzazione (circa 2 mesi e mezzo).

2° stage:

$$\widehat{Y} = E[Y | T] = \mu_Y + \beta_{YT} \widehat{T}$$

$$\widehat{Y} = E[Y | T] = \mu_Y + \beta_{YT} (\mu_T - \beta_{TZ} * Z)$$

$$\widehat{\text{Salario}} = \mu_{\text{Salario}} + \beta_{\text{Salario, anniScuola}} \widehat{\text{AnniScuola}}$$

$$\rightarrow \beta_{\text{Salario, anniScuola}} = 2500\$ = \text{LATE}$$

Un aumento unitario di T, ossia di un anno della scolarizzazione, provoca un aumento del reddito di 2500\$.

Per far variare T di uno, Z dovrebbe aumentare di 5 volte (perchè $-0.2 \cdot 5 = -1$) ossia dovrei aumentare la distanza tra casa e scuola di 5 Km, ed avrei quindi una riduzione del reddito di 2500\$.



$$\beta_{YZ} = \text{Effetto indiretto di } Z \text{ su } Y: \beta_{TZ} * \beta_{YT} = 2500\$ * -0.2 = -500\$$$

Un variazione unitaria di Z, ossia di un Km di distanza, provoca una riduzione del salario medio annuo di 500 dollari.

Difference-in-difference

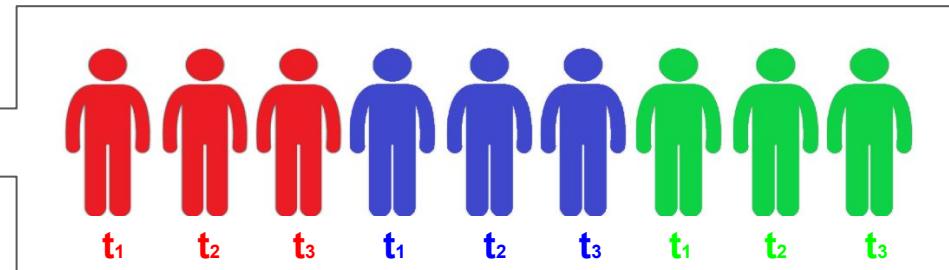
Fino ad ora abbiamo parlato di ...

... **Cross sectional**: l'outcome e le variabili sono state osservate nello stesso istante di tempo.

Ora parliamo di ...

... **Panel data o dati longitudinali**: ciascun individuo viene osservato in diversi istanti di tempo, avremo diverse rilevazioni dell'outcome e delle covariate per ciascun individuo.

*E' come se avessi 3 cross sectional,
posso stimare la differenza tra gli
individui in ogni istante, ma posso
anche vedere se il fenomeno ha un
certo andamento nel tempo.*



Cross sectional:



Quando stimo l'effetto di una covariata o di un trattamento sull'outcome c'è il problema delle **variabili omesse**, ossia di tutte quelle variabili o caratteristiche intrinseche delle unità che non introduco nel modello.

Panel data o dati longitudinali:



Quando stimo l'effetto di una covariata o di un trattamento sull'outcome posso stimare anche l'effetto delle **variabili omesse**. Ogni soggetto ha una serie di misurazioni che permettono di stimare un livello di partenza individuale dell'outcome (*random intercept*).

Cross sectional

$$Y_i = \mu_Y + \beta_1 X_i + \varepsilon_i$$

μ_Y è il valore atteso di Y per la popolazione generale quando $X = 0$ ed è **l'effetto di variabili omesse dal modello**, che definiscono un livello medio di partenza per tutto il campione.

Panel data o dati longitudinali

$$Y_{it} = \mu_{Yi} + \beta_{1i} X_{it} + \varepsilon_{it}$$

μ_{Yi} è **l'effetto di variabili omesse dal modello che variano con i**.

Avrà un'intercetta per ogni unità (*random intercept*).

Cross sectional: Sia Y il costo della casa ed X il numero di servizi di lusso di cui dispone (portineria, ascensore, vista, ecc).

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \forall \text{ unità } i$$

β_0 costo medio di una casa senza servizi di lusso;

β_1 incremento di costo medio se aggiungo un servizio di lusso;

Dati longitudinali: ipotizziamo che ci sia una sola variabile omessa, se la casa è collocata al centro città o se è in periferia.

$$Y_{it} = \beta_{0i} + \beta_{1i} X_{it} + \varepsilon_{it} \quad \forall \text{ unità } i \quad \forall \text{ tempo } t$$

Avrò due β_{0i} ossia due livelli di partenza possibili a seconda che la casa sia dentro o fuori dal centro: mi aspetto che, a parità di servizi, la casa in centro sia più cara;



Con la stessa covariata riesco a stimare anche l'effetto della variabile omessa

Supponiamo di avere due istanti di tempo: t e $t-1$. Al tempo $t-1$ nessuno è stato trattato, al tempo t alcuni riceveranno il trattamento mentre altri no.



Stimare L'ATE di T su Y: $E [Y_{1,t} - Y_{0,t}]$

Ma possiamo osservare l'effetto causale del trattamento solo nei trattati, dunque ...



Stimare il LATE di T su Y: $E [Y_{1,t} - Y_{0,t} | T=1]$

Ma l'effetto che vedo al tempo t , potrebbe essere costituito dall'effetto causale di T ma anche dall'effetto temporale!



Stimare il LATE di T su Y **al netto dell'effetto del tempo**, utilizzando l'effetto temporale osservato nei non trattati

$$\text{ATE} = E[Y_{1,t} - Y_{0,t}] = E[Y_{1,t} - \cancel{Y_{0,t-1}}] - E[Y_{0,t} - \cancel{Y_{0,t-1}}]$$

differenza
dell'outcome tra
t-1 e t nei trattati

differenza
dell'outcome tra t-1
e t nei non trattati

$$\tilde{Y}_1 \quad \tilde{Y}_0$$

Se $\tilde{Y}_1, \tilde{Y}_0 \perp T$

$$\text{ATE} = E[Y_t - Y_{t-1} | T = 1] - E[Y_t - Y_{t-1} | T = 0]$$

→ quantità osservabili

Se non si può assumere $\tilde{Y}_1, \tilde{Y}_0 \perp T$ allora utilizzo l'LATE

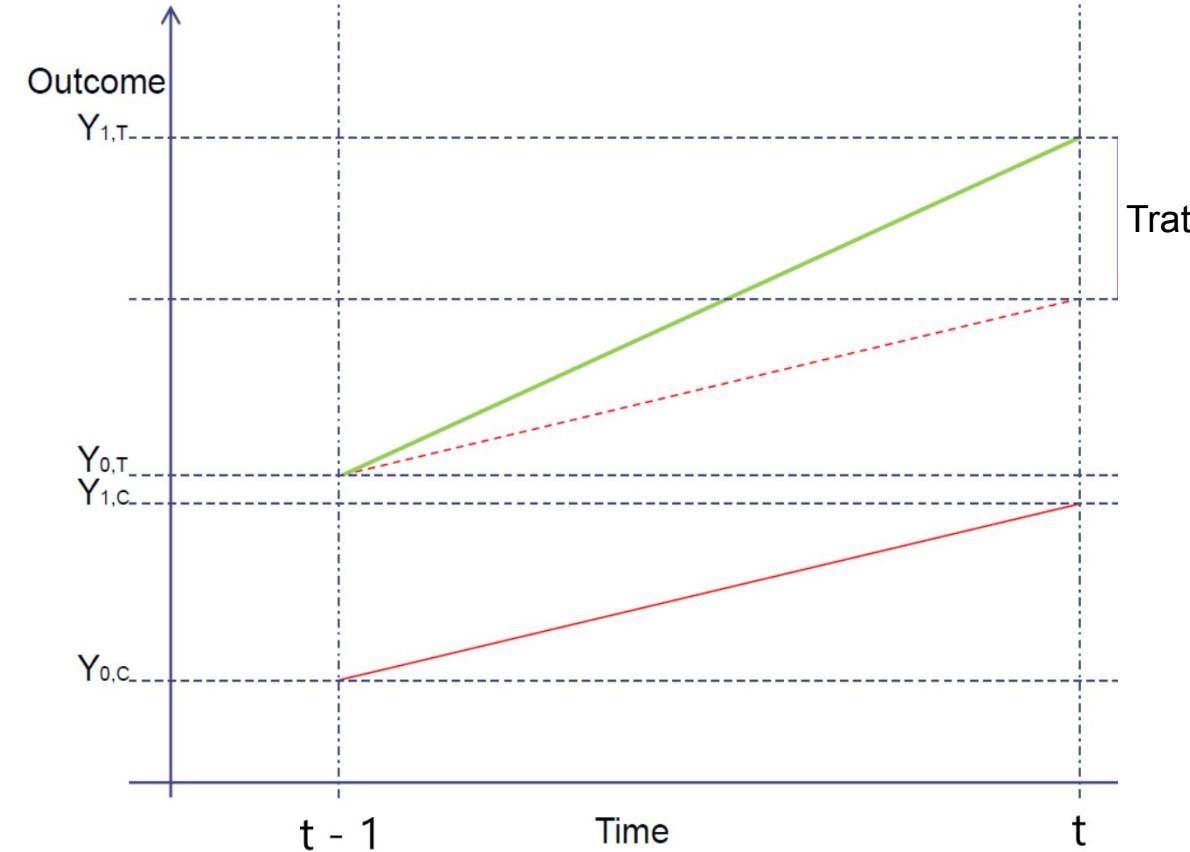
$$\text{LATE} = E[Y_{1,t} - Y_{0,t} | T = 1] = E[Y_{1,t} | T=1] - E[Y_{0,t} | T = 1]$$

↓ ↓

Osservato *Non osservato
Da stimare*

Per stimare $E[Y_{0,t} | T = 1]$ devo fare un'assunzione:

- **Common Trend o Parallel trend:** L'evoluzione del fenomeno nel tempo, senza essere intaccato dal trattamento, è lo stesso nei due gruppi.
$$E[Y_{0,t} - Y_{0,t-1} | T = 1] = E[Y_{0,t} - Y_{0,t-1} | T = 0]$$
- **Bias stability:** Il Bias tra trattati e non trattati è presente ma è stabile nel tempo. $E[Y_{0,t} | T=1] - E[Y_{0,t} | T=0] = E[Y_{0,t-1} | T = 1] - E[Y_{0,t-1} | T = 0]$



Common Trend o Parallel trend:

In assenza del trattamento, la differenza tra i trattati e i non trattati rimane la stessa sia al tempo $t - 1$ ed che al tempo t

Utilizzando l'assunzione di Common Trend ...

$$\mathbf{E}[Y_{0,t} | T = 1] = E[Y_{0,t-1} | T = 1] + E[Y_{0,t} - Y_{0,t-1} | T = 0]$$

... riesco a stimare l'elemento non osservato del LATE attraverso delle misure osservabili:

$$\begin{aligned}\text{LATE} &= E[Y_{1,t} - Y_{0,t} | T = 1] = E[Y_{1,t} | T=1] - \mathbf{E}[Y_{0,t} | T = 1] \\ &= E[Y_{1,t} | T=1] - E[Y_{0,t-1} | T = 1] - E[Y_{0,t} - Y_{0,t-1} | T = 0] \\ &= E[Y_{1,t} - Y_{0,t-1} | T=1] - E[Y_{0,t} - Y_{0,t-1} | T = 0] \\ &= \mathbf{E}[Y_t - Y_{t-1} | T = 1] - \mathbf{E}[Y_t - Y_{t-1} | T = 0]\end{aligned}$$

Se “Trat”(si=1, no=0) e “Tempo” (t-1=0, t=1) sono dicotomiche, il modello statistico per stimare il LATE è il seguente:

$$E[Y | \text{Trat}, \text{Tempo}] = \beta_0 + \beta_1 * \text{Trat} + \beta_2 * \text{Tempo} + \beta_3 * (\text{Trat} * \text{Tempo})$$

β_0 : outcome medio nel gruppo dei non trattati al tempo t-1;

β_1 : differenza tra l'outcome medio nel gruppo dei trattati e dei non trattati al tempo t-1;

β_2 : differenza tra l'outcome medio nel gruppo dei non trattati tra tempo t-1 e tempo t;

β_3 : Rappresenta quanto l'outcome medio nel gruppo dei trattati è cambiato dal tempo t-1 al tempo t, rispetto a ciò che sarebbe successo allo stesso gruppo se non fossero stati trattati.

Se non si può assumere **Bias Stability/Common trend**

$$E[Y_{0,t} | T = 1] = E[Y_{0,t} - Y_{0,t-1} | T = 0] + E[Y_{0,t-1} | T = 1] + \mathbf{B}_{t,t-1}$$

E' necessario stimare $B_{t,t-1}$: per farlo serviranno dei dati di un periodo precedente.

Ipotizzo $\rightarrow B_{t,t-1} = B_{t-1,t-2}$

E stimo il bias del periodo precedente come segue

$$B_{t-1,t-2} = E[Y_{0,t-1} - Y_{0,t-2} | T = 1] = E[Y_{0,t-1} - Y_{0,t-2} | T = 0]$$

In cui tutto è osservabile perché nessuno è ancora stato trattato, il trattamento sarà assegnato solo al tempo t.

Difference-in-difference: esempio

Outcome: punteggio che aumenta al migliorare della salute del paziente;

Trattamento: trattato = 1, non trattato = 0;

Tempo: tempo t-1 = 0, tempo t = 1.

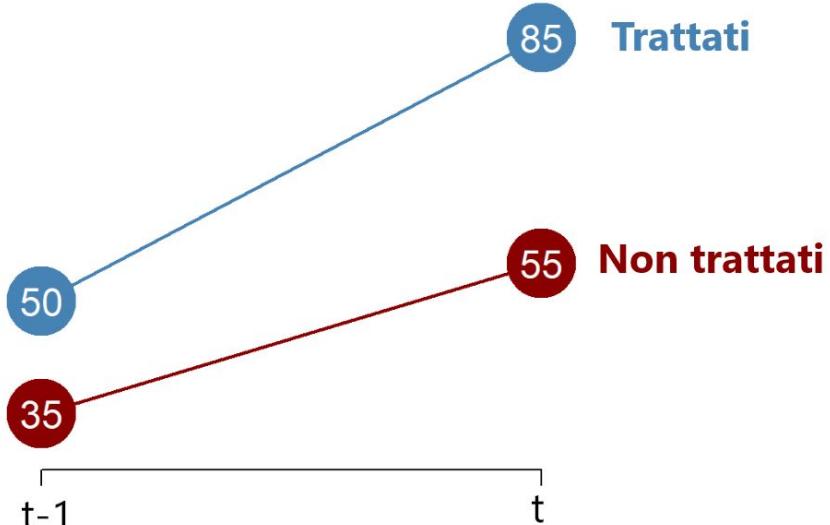
Soggetto	Outcome	Trattamento	Tempo
----------	---------	-------------	-------

1	74	1	1
1	46	0	0
2	96	1	1
2	54	0	0
3	50	0	1
3	30	0	0
4	60	0	1
4	40	0	0

1) Voglio osservare quanto differiscono le medie dell'outcome prima e dopo il trattamento nel gruppo dei trattati (effetto del trattamento da purificare);

2) Voglio osservare la differenza tra le medie nei non trattati nei due istanti di tempo (effetto temporale).

Concettualmente:



$$\text{Diff-in-Diff} = 35 - 20 = 15$$

il trattamento ha un effetto benefico.

1) Differenza tra le medie dell'outcome prima e dopo il trattamento nei trattati (effetto del trattamento da purificare);

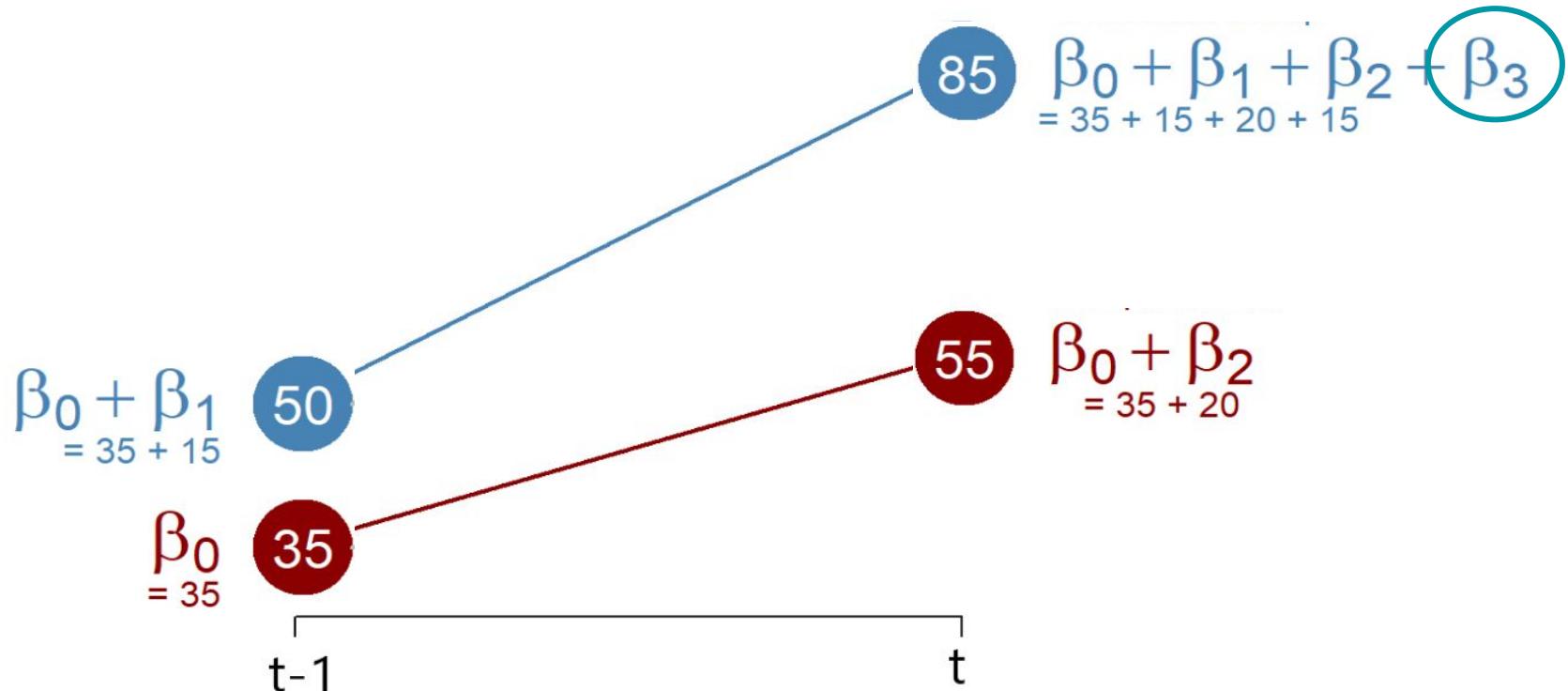
$$E[Y_{1,t} | T=1] - E[Y_{0,t-1} | T=1] = 85 - 50 = 35$$

2) Differenza tra le medie nei due istanti di tempo nei non trattati (effetto temporale).

$$E[Y_{0,t} | T=0] - E[Y_{0,t-1} | T=0] = 55 - 35 = 20$$

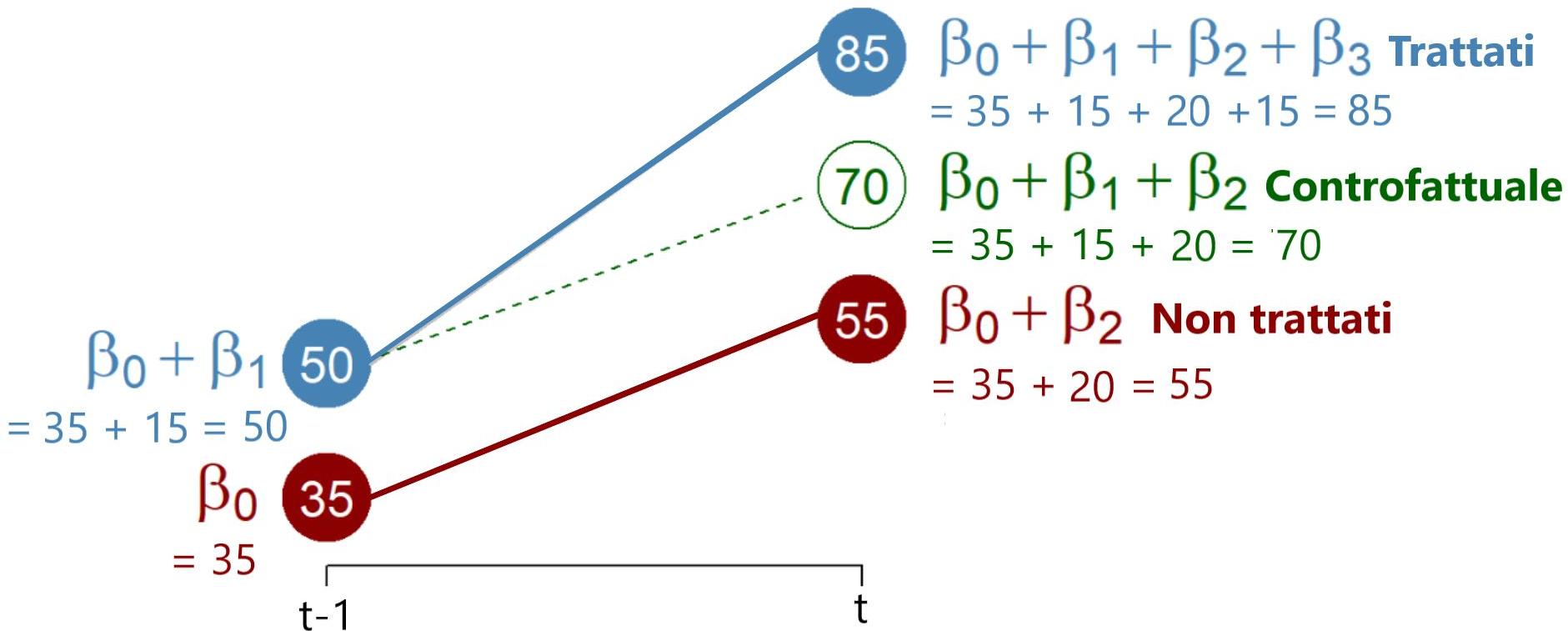
Modello statistico:

$$E[Y | Tr, Te] = 35 + 15 * Trat + 20 * Tempo + 15 * (Trat * Tempo)$$



Modello statistico:

$$E[Y | Tr, Te] = 35 + 15 * Trat + 20 * Tempo + 15 * (Trat * Tempo)$$



Regression Discontinuity (sharp)



Stimare l'effetto causale del trattamento sull'outcome

- **Y** outcome di interesse;
- **T** trattamento dicotomico (si=1, no=0);
- **Z** variabile in base alla quale si assegna il trattamento.

Esempio:

Si vuole valutare l'effetto l'assegnazione di una borsa di studio (T) sul rendimento scolastico (Y). Tale borsa viene assegnata solo ai ragazzi più meritevoli, che ottengono un punteggio ad un pre-test (Z) maggiore di 80/100 (z_0).

se $Z \geq z_0 \Rightarrow T = 1$

Consideriamo il valore atteso $E[Y | Z = z]$

- se $Z < z_0 \Rightarrow T = 0$ allora $E[Y | Z = z] = E[Y_0 | Z = z]$

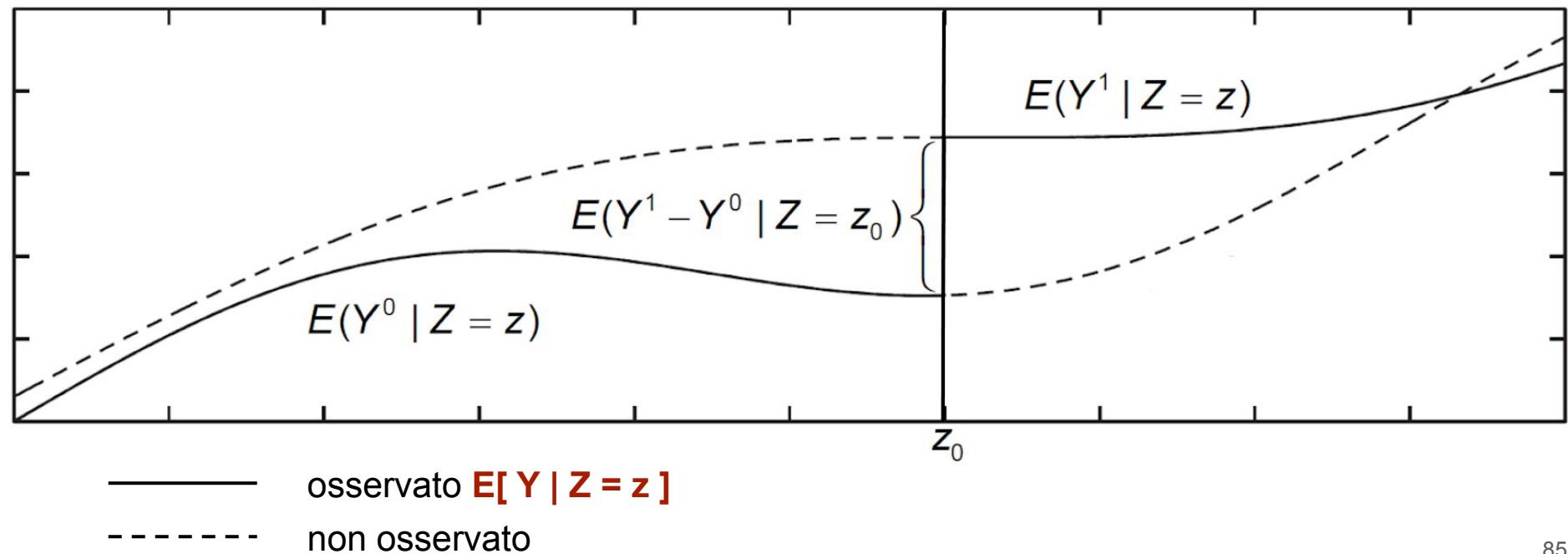
- se $Z \geq z_0 \Rightarrow T = 1$ allora $E[Y | Z = z] = E[Y_1 | Z = z]$

è una funzione discontinua, con punto di discontinuità in z_0 .

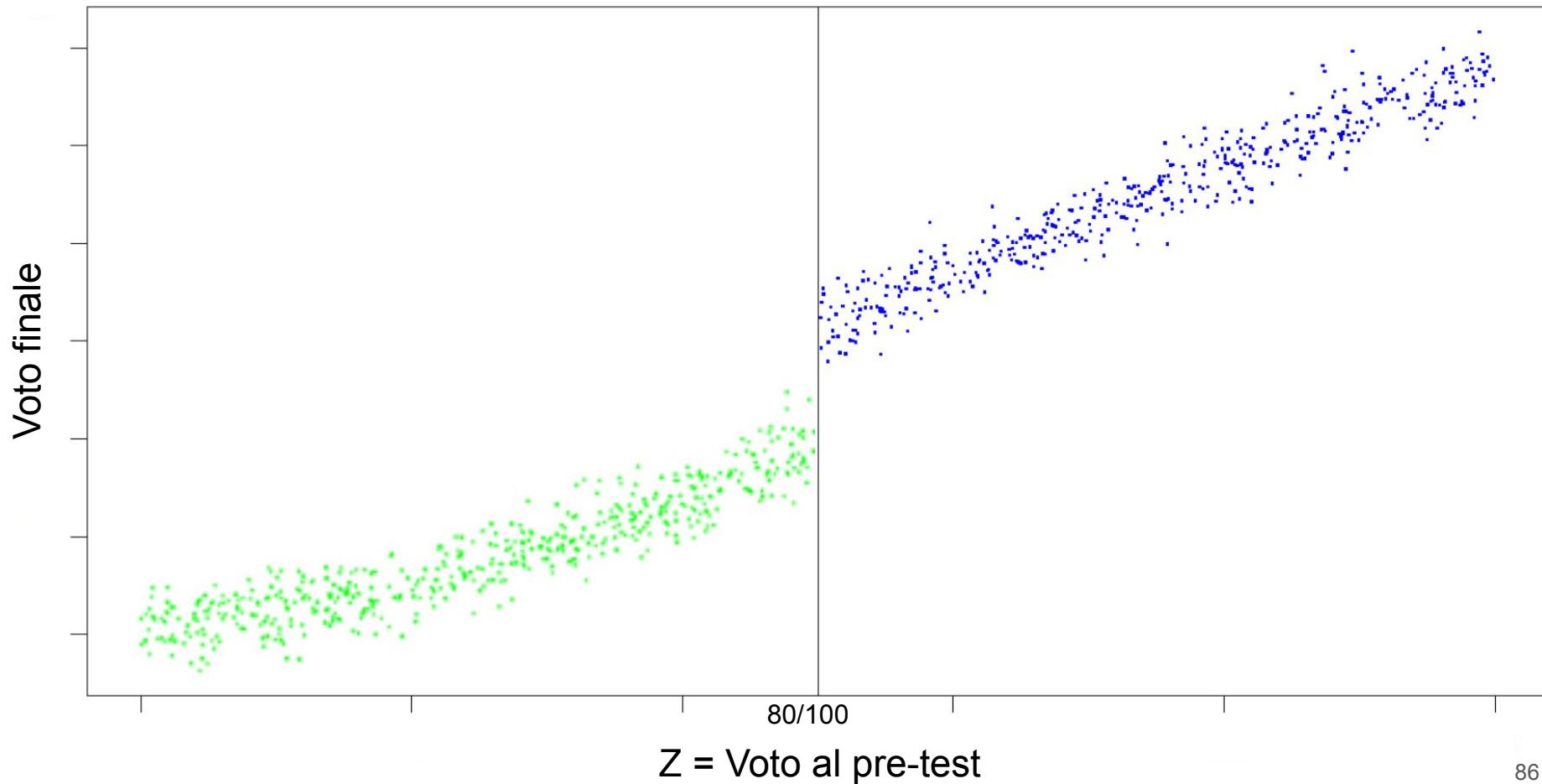


L'altezza del salto è l'**ATE condizionato a $Z=z_0$** , ed è ciò che vogliamo stimare.

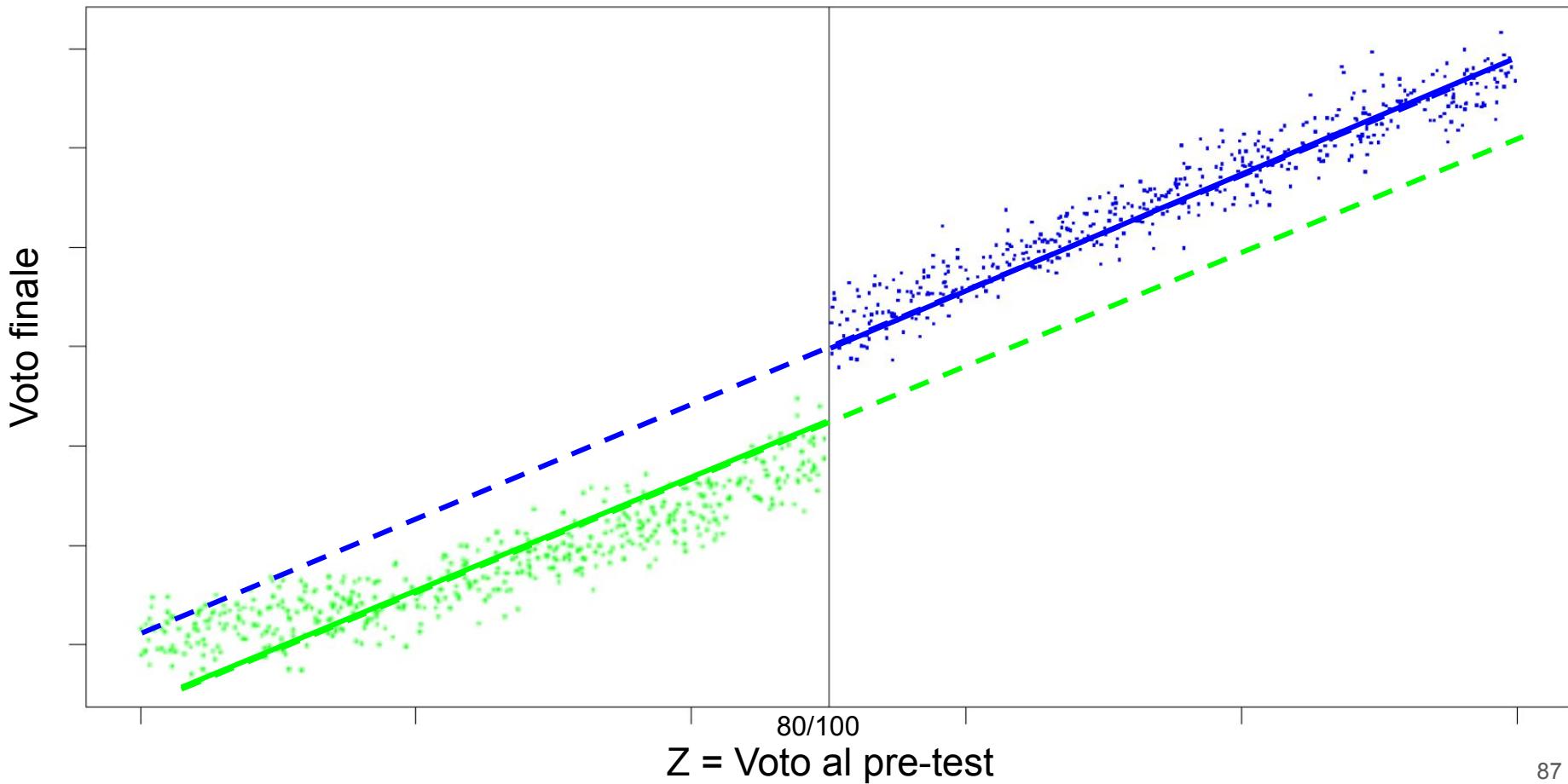
Assunzione: continuità di $E[Y_T | Z = z]$ in $Z=z_0$, per $T = 1, 0$.



Ciò che osserviamo



Ciò che confrontiamo (due regressioni):



Per stimare l'ATE consideriamo l'equazione di regressione :

$$E[Y | Z = z] = \mu_Y + \beta_1 * T + \beta_2 * Z$$

$$E[Finale | Pretest] = \mu_{Finale} + \beta_{Borsa} * Borsa + \beta_{Pretest} * Pretest$$

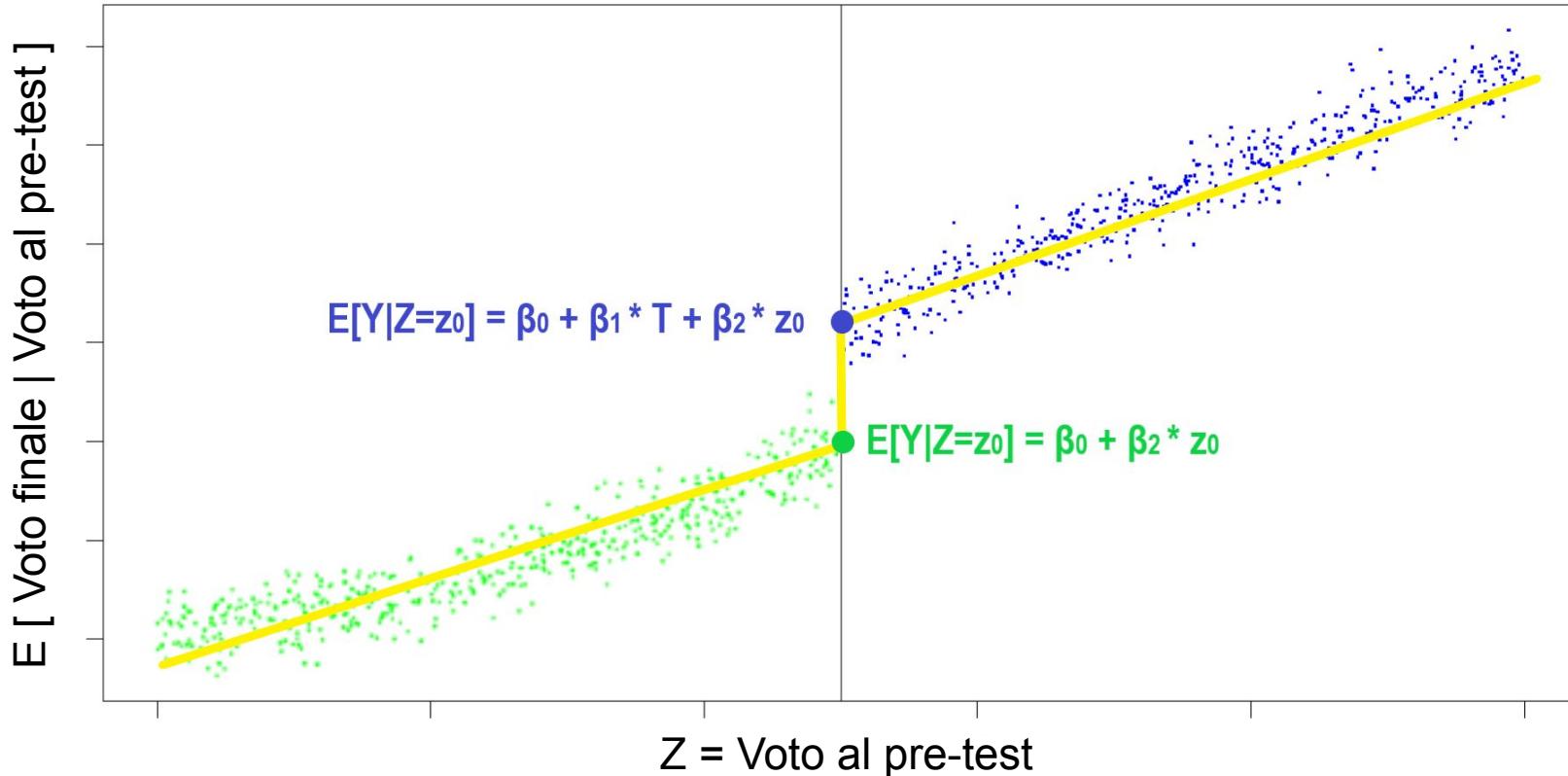
- Per $Z < z_0$ ossia con $T=0$ sarà: $E[Y_0 | Z=z] = \mu_Y + \beta_2 * z$

$$E[Finale_0 | Pretest] = \mu_{Finale} + \beta_{Pretest} * Pretest$$

- Per $Z > z_0$ ossia con $T=1$ sarà: $E[Y_1 | Z=z] = \mu_Y + \beta_1 + \beta_2 * z$

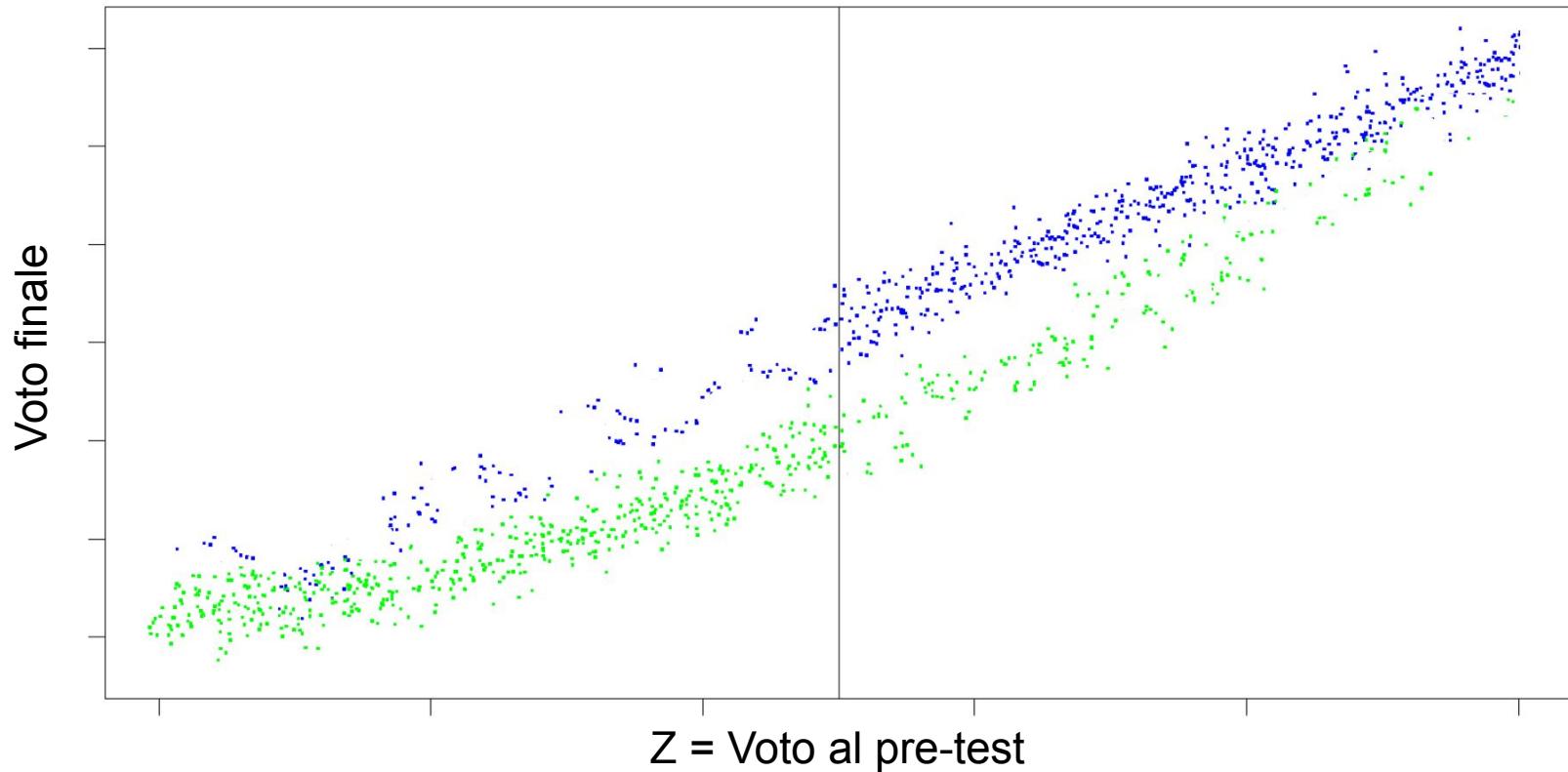
$$E[Finale_1 | Pretest] = \mu_{Finale} + \beta_{Borsa} + \beta_{Pretest} * Pretest$$

$$\text{ATE} = E(Y_1 - Y_0 \mid Z=z_0) = E(Y_1 \mid Z=z_0) - E(Y_0 \mid Z=z_0) =$$
$$\beta_0 + \beta_1 + \cancel{\beta_2 * z_0} - \cancel{\beta_0} - \cancel{\beta_2 * z_0} = \beta_1$$



Regression Discontinuity (fuzzy)

Il superamento della soglia z_0 , dunque il valore di $T=1$, non combacia con l'assegnazione vera del trattamento (e viceversa), ma crea comunque una discontinuità nella probabilità di essere trattati.



Possiamo fare un ragionamento affine all'Instrumental Variable: dato un $\varepsilon > 0$ piccolo a piacere, possiamo identificare quattro gruppi latenti distinti: AT, NT, C e D. Per gli stessi motivi che abbiamo visto nell'Instrumental Variable considereremo solo i compliers, ossia solo chi sarà trattato perché ha un valore di Z superiore alla soglia.



Il LATE condizionato a $Z=z_0$ è ciò che vogliamo stimare.

$$\text{LATE} = E(Y^1 - Y^0 | Z = z_0, G = c) = \frac{\lim_{\varepsilon \downarrow 0} [E(Y | Z = z_0 + \varepsilon) - E(Y | Z = z_0 - \varepsilon)]}{\lim_{\varepsilon \downarrow 0} [E(T | Z = z_0 + \varepsilon) - E(T | Z = z_0 - \varepsilon)]}$$

$$\text{LATE} = E(Y^1 - Y^0 | Z = z_0, G = c) = \frac{\lim_{\varepsilon \downarrow 0} [E(Y | Z = z_0 + \varepsilon) - E(Y | Z = z_0 - \varepsilon)]}{\lim_{\varepsilon \downarrow 0} [E(T | Z = z_0 + \varepsilon) - E(T | Z = z_0 - \varepsilon)]}$$

→ Il numeratore esprime il LATE moltiplicato per la probabilità di essere compliers.

$$E(Y^1 - Y^0 | Z = z_0, G = c) P(G = c | Z = z_0)$$

→ Per questo aggiungiamo il denominatore che esprime la probabilità di essere compliers condizionatamente a z_0 . Il denominatore serve a purificare il numeratore.

$$P(G = c | Z = z_0)$$

Nel caso della RD fuzzy non c'è una concordanza perfetta tra Z e T, dunque β_1 non esprime più l'ATE ma esprime l'intent-to-treat ossia l'intento iniziale di trattare l'unità ma non la ricezione effettiva del trattamento.

Il LATE può essere stimato con il 2SLS visto in ambito dell'IV:

$$\text{1° stage} \rightarrow \widehat{\mathbf{P}} = \beta_0 + \beta_1 * T + \beta_2 * Z$$

$$\text{2° stage} \rightarrow Y = \alpha_0 + \alpha_1 * \widehat{\mathbf{P}} + \alpha_2 * Z$$

dove $\widehat{\mathbf{P}}$ è una variabile che indica l'effettiva partecipazione dell'unità al programma / trattamento.

$$LATE = E(Y_1 - Y_0 | Z=z_0, G=c) = E(Y_1 | Z=z_0, G=c) - E(Y_0 | Z=z_0, G=c) =$$

$$\cancel{\alpha_0 + \alpha_1 + \alpha_2 * z_0} - \cancel{\alpha_0 + \alpha_2 * z_0} = \alpha_1$$

Parte delle slide presentate derivano dal materiale dei corsi:

- Roberto Rocci "[Inferenza Causale](#)"
- Miguel Hernan "[Causal Diagrams: Draw Your Assumptions Before Your Conclusions](#)"
- Bruno Arpino "[Course on Causal Inference](#)"
- Peter Eibich & Angelo Lorenti "[Advanced Methods for Causal Inference](#)"
- Linda Valeri "[Mediation Analysis and Causal Inference](#)"

Libri “divulgativi” sull’inferenza causale:

-Miguel Hernan “Causal Inference: What If (the book)”

-Scott Cunningham “Causal Inference The Mixtape”

-Judea Pearl “The book of why”

Altri: <https://www.bradyneal.com/which-causal-inference-book>



SEMINARI INFERENZA CAUSALE

Dipartimento di Sanità Pubblica e Malattie Infettive

Aula Scuola di Specializzazione in Statistica Sanitaria e Biometria

AULA C1 ore 12:00 - 14:00

**Martedì 29 Novembre:
Introduzione alla causalità**

Dr.ssa MARGHERITA MORETTI

Dottorato in Scuola di Scienze Statistiche
Sapienza Università di Roma

Dr.ssa LAURA MONTELISCIANI

Dottorato in Sanità Pubblica Biostatistica ed Epidemiologia
Università degli Studi di Milano-Bicocca

**Martedì 6 Dicembre:
Metodi per l'inferenza causale**

Margherita Moretti

margherita.moretti@uniroma1.it

Laura Montelisciani

l.montelisciani@campus.unimib.it

PER SCARICARE LE SLIDE:

https://github.com/MMargherita/seminari_DSPMI