Dati multi-fonte e analisi territoriali

Marco Tosi, Irene Barbiera, e Federico Gianoli

Dipartimento di Scienze Statistiche

Matching

Il matching è l'abbinamento di record con caratteristiche simili su più covariate.

- Nel record linkage utilizziamo tecniche di matching per massimizzare gli abbinamenti veri, ossia record appartenenti alla stessa unità.
- L'abbinamento statistico viene utilizzato anche quando abbiamo solo match falsi, ossia record che possono essere accoppiati sulla base delle loro caratteristiche ma *fanno parte di entità/unità diverse*.
- La finalità è spesso quella di identificare e stimare un «effetto causale» di un «trattamento» su un outcome. [in questo Corso non affrontiamo i metodi di stima dell'effetto ma i metodi di *abbinamento dei record*]

Il contesto

Il contesto in cui viene applicato il matching/ abbinamento statistico è quello dell'inferenza causale.

- Spesso è impossibile ricorrere ad un esperimento randomizzato o randomizzare il trattamento.
- Spesso è utilizzato nella scelta del controfattuale.

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

ID	D	Y(0)	Y(1)	X1	X2	…	Xn
1	0	21	?	..	..		
2	1	?	31	..	..		
n							

- In altre parole: come sarebbe l'outcome di interesse (Y0) se i trattati non avessero partecipato al trattamento (D)?

Randomizzazione

Gli studi randomizzati sono considerati il *gold standard* per avere un gruppo di controllo (controfattuale) le cui caratteristiche (distribuzione di Xn) sono statisticamente identiche a quelle del gruppo dei trattati.

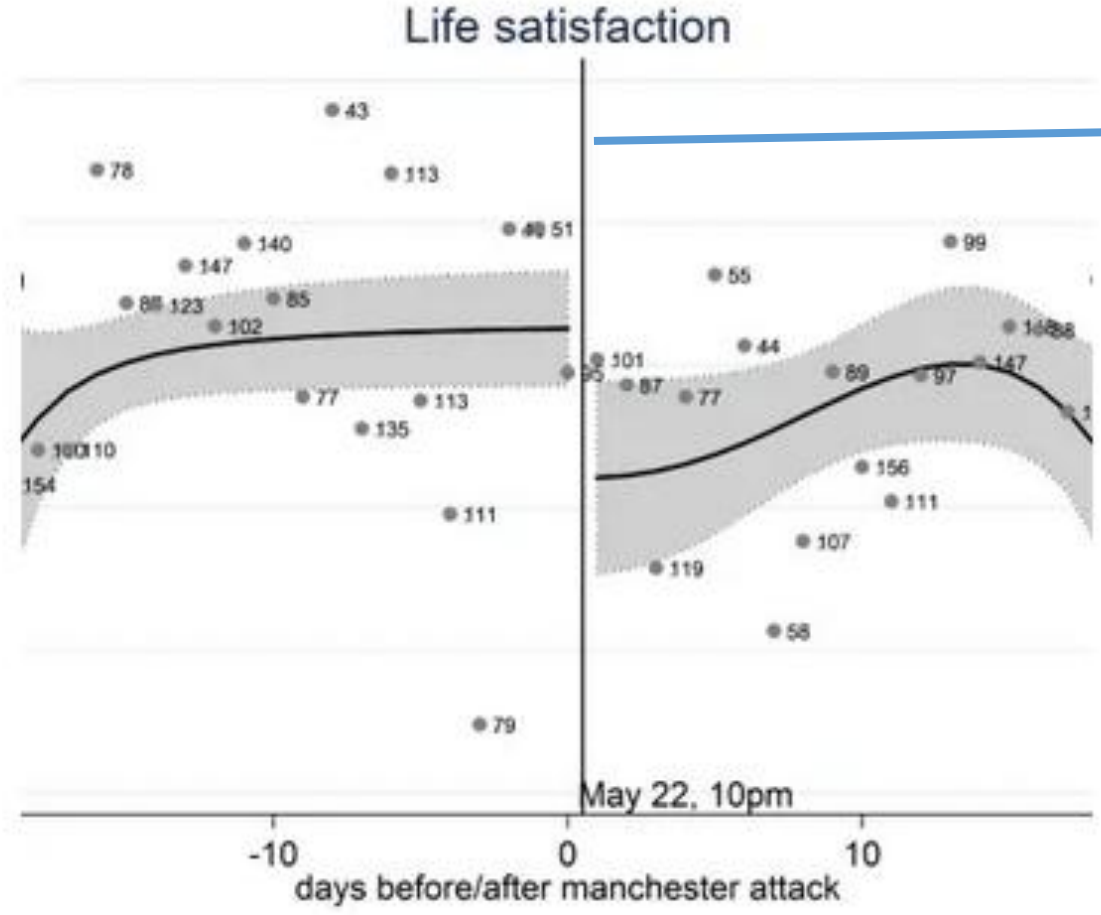
- I dati mancanti e la mancata partecipazione allo studio rimangono problematici nella randomizzazione, dato che le probabilità di assegnazione di un trattamento deve rimanere uguale per tutti i partecipanti.
- Problemi etici alla randomizzazione quando si tratta di dati sociali: e.g., sostegno al reddito (D) ad un gruppo random di famiglie in povertà.

Negli studi osservazionali sono spesso affetti dai fattori di confondimento.

Randomizzazione?

Gli studi randomizzati sono spesso non applicabili nelle scienze sociali.

Evento inaspettato che **non** avviene in luoghi scelti casualmente: es, chi vive in grandi città ha più probabilità di sperimentarlo.



Evento inaspettato che crea due gruppi selezionati *casualmente:*
-> Il timing delle interviste non è casuale (es., coprire quote) quindi prendiamo solo pochi giorni prima e dopo l'evento.

Propensity score

- Dobbiamo costruire un gruppo di controllo.
  - Assumiamo che tutte le caratteristiche che distinguono i trattati e i controlli siano catturate dal set di covariate X
  - Selezioniamo e abbiniamo quei record nel gruppo dei controlli che hanno una distribuzione delle variabili X simile a quelle dei trattati.

- **Propensity score** come probabilità di essere trattati e sintesi univariata di C, dove p è P(D = 1|C) (Rosenbaum and Rubin 1983).
  - Stimata da un modello probit o logit.

Esempio: studi osservazionali (SHARE)

- Dobbiamo ‘controllare’ per le variabili confondenti per ottenere stime non distorte.
  - Maledizione della multidimensionalità.

Se **C -> X** positivamente & **C-> Y** positivamente: ***sovrastima*** (C non- osservato)

Se **C -> X** positivamente & **C->Y** negativamente: ***sottostima*** (C non- osservato)

- **Propensity score**
  - Riduce la multidimensionalità di C in una dimensione, ossia nella probabilità di sperimentare il trattamento secondo le C rilevanti (Rosenbaum/Rubin, 1983). Esaustività di C per esaurire i ‘path’ che portano al trattamento (***solo confondenti!*** *Se no: overcontrol & endogeneity bias*).

Esempio

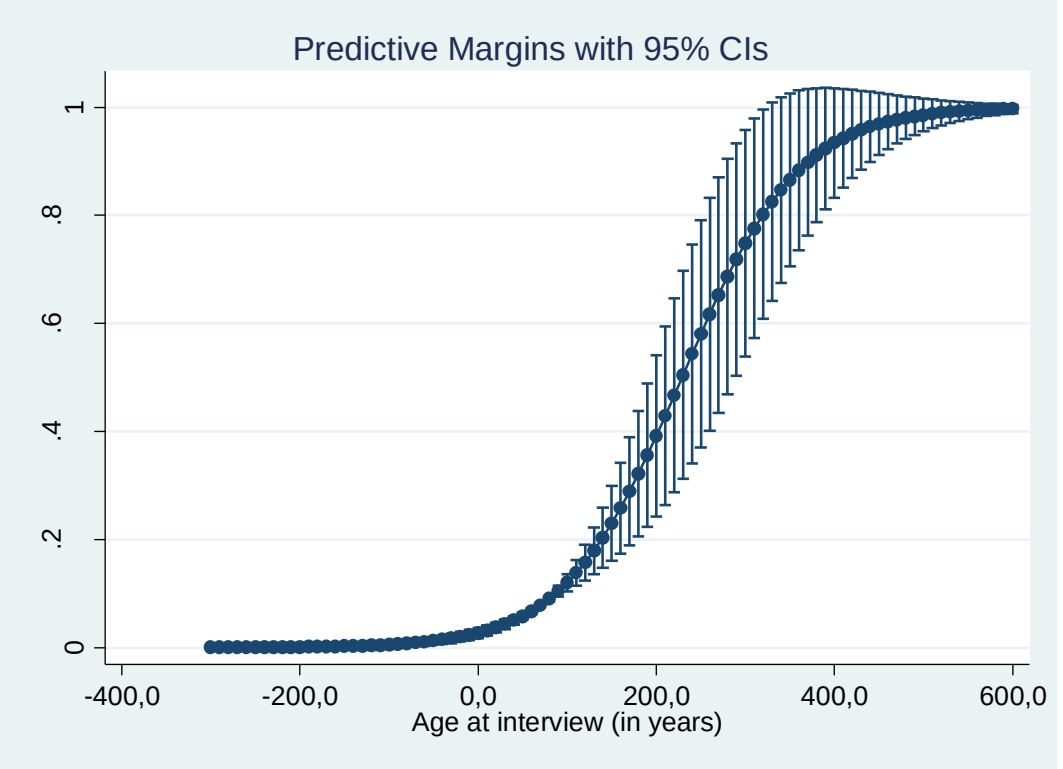
- Vogliamo studiare l'effetto di rimanere senza figli sulla salute psicologica (Eurod = sintomi di depressione) nell'ultima wave di SHARE.
  - Problema 1: *Sample selection bias.* Consideriamo solo quelli che hanno concluso le loro storie riproduttive, e.g. > 50 (più facile concentrarsi sulle donne per motivi biologici).
  - Problema 2: *Confondimento o selezione nel trattamento.* Ad esempio in alcune coorti di nascita ed in alcuni paesi la propensione a rimanere senza figli è più alta e anche i tassi di depressione sono più alti.
  - *Quali altri fattori confondenti?*

Calcolare la propensione al trattamento

```
. *PROBIT
. probit childless c.age##c.age i.country i.educ if female==1 & age>49

Iteration 0:   log likelihood = -7019.8585
Iteration 1:   log likelihood = -6813.5279
Iteration 2:   log likelihood =  -6809.802
Iteration 3:   log likelihood = -6809.7947
Iteration 4:   log likelihood = -6809.7947

Probit regression                                 Number of obs   =      25464
                                                  LR chi2(30)     =     420.13
                                                  Prob > chi2     =     0.0000
Log likelihood = -6809.7947                       Pseudo R2       =     0.0299
```

childless	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0972985	.0150638	-6.46	0.000	-.126823	-.0677739
c.age#c.age	.0007279	.0001038	7.01	0.000	.0005244	.0009313
country						
12. Germany	-.0760307	.0732734	-1.04	0.299	-.219644	.0675825
13. Sweden	-.3356912	.0815046	-4.12	0.000	-.4954372	-.1759452
14. Netherlands	.0180876	.0789444	0.23	0.819	-.1366406	.1728157
15. Spain	-.10758	.0817216	-1.32	0.188	-.2677514	.0525914
16. Italy	.1142278	.0767578	1.49	0.137	-.0362147	.2646703
17. France	-.1532845	.0760804	-2.01	0.044	-.3023993	-.0041697
18. Denmark	-.3902609	.0841628	-4.64	0.000	-.555217	-.2253047
19. Greece	.0880355	.0709832	1.24	0.215	-.051089	.2271601

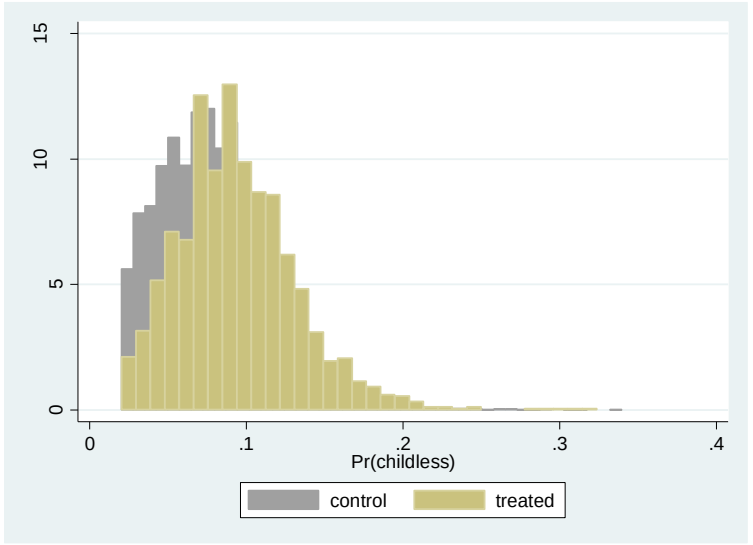
$$p(Y) = \frac{1}{1+e^{-(B_0+B_1X_1+\ldots BzXz)}}$$



Confrontare propensione al trattamento

```
. predict pr1_prob if e(sample)
(option pr assumed; Pr(childless))
(19014 missing values generated)
```

Definizione di un'area comune in cui i due gruppi hanno una simile propensione al trattamento



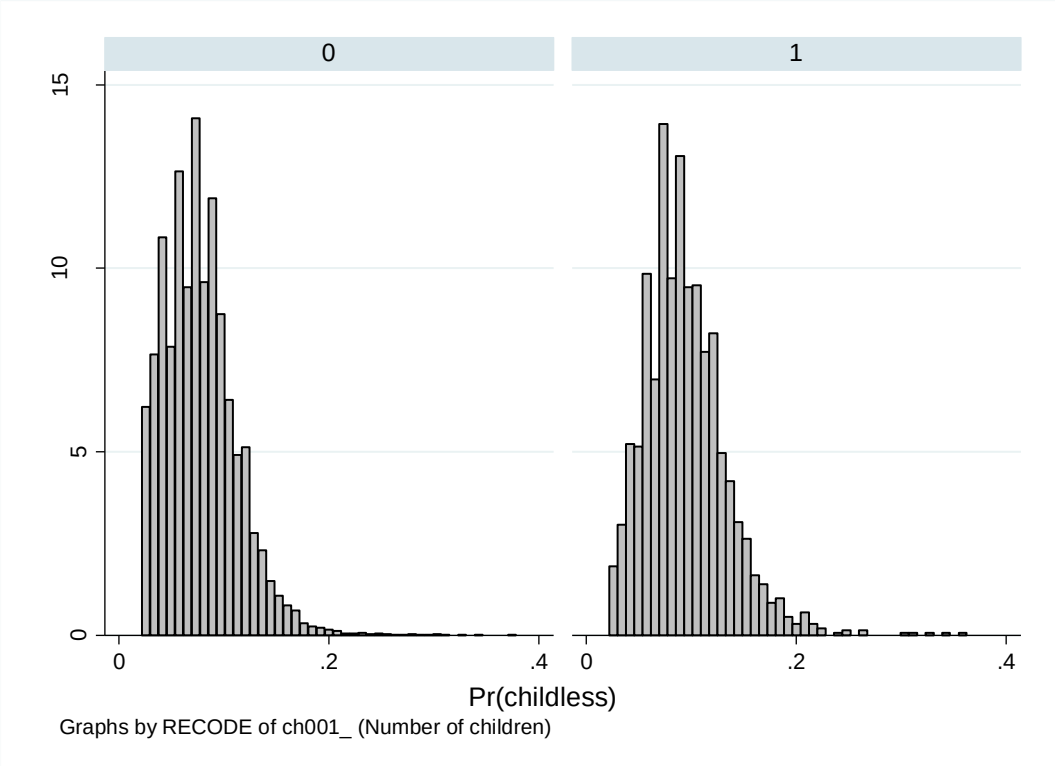
	mergeid	pr1_prob	pr1_logit	childless	age
1	CZ-840436-02	.0203449	.0218053	0	66,8
2	CZ-095057-01	.0203449	.0218053	0	66,8
3	CZ-753276-01	.0203449	.0218053	0	66,8
4	CZ-059855-01	.0203449	.0218053	0	66,8
5	CZ-005707-02	.0203452	.0218068	0	66,9
6	CZ-195759-01	.0203452	.0218068	0	66,9
7	CZ-517252-02	.0203454	.0218045	0	66,7
8	CZ-772502-01	.0203462	.0218088	0	67,0
9	CZ-722833-04	.0203462	.0218088	0	67,0
10	CZ-161259-01	.0203462	.0218088	0	67,0
11	CZ-135113-01	.0203465	.0218042	0	66,6
12	CZ-350463-01	.0203465	.0218042	0	66,6
13	CZ-055467-02	.0203465	.0218042	0	66,6

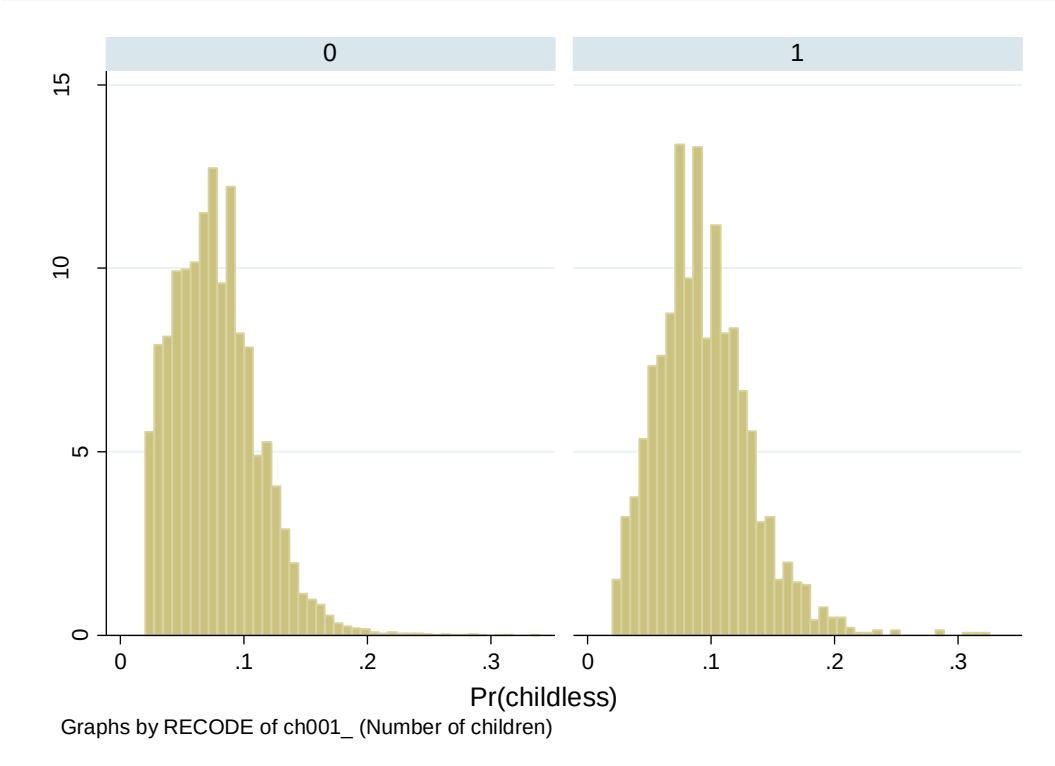
```
. tabstat pr1_prob, by(childless) st(min max n mean)

Summary for variables: pr1_prob
     by categories of: childless (RECODE of ch001_ (Number of children))
```

childless	min	max	N	mean
0	.0203178	.3407631	23459	.0774094
1	.0204022	.3241942	2005	.0942742
Total	.0203178	.3407631	25464	.0787373

Distribuzione del propensity score

- Date da Logit (in grigio) e Probit (in giallo)





Propensity score

- Essendo probabilità predette di essere senza figli, lo stesso valore (%) può essere ottenuto da diverse combinazioni di variabili.

  Prob = 0.0316 per una 65enne di Cipro e una 62enne Danese

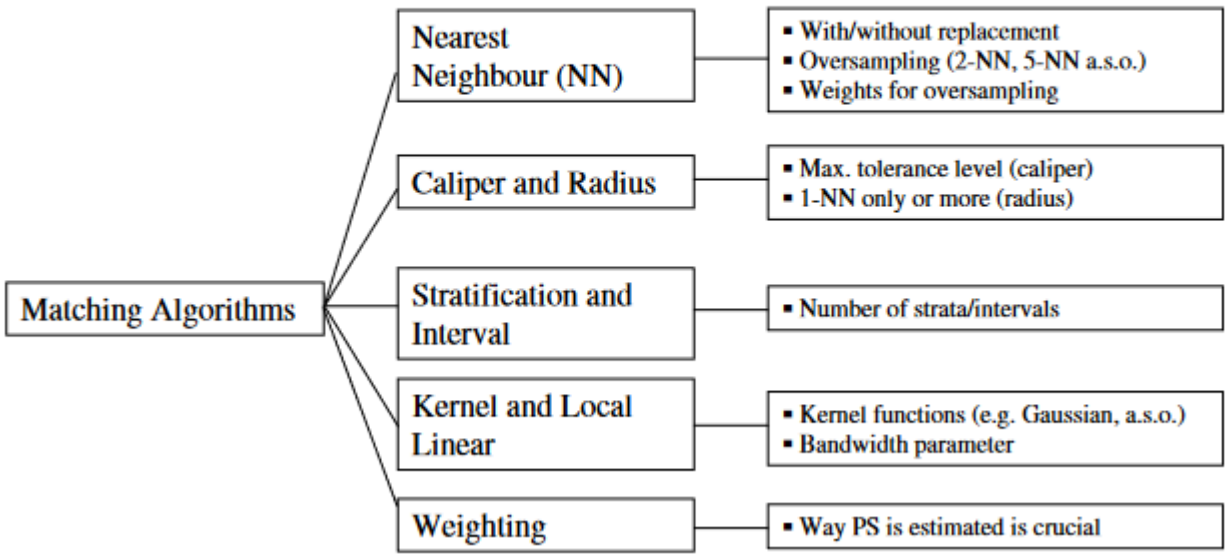
mergeid	pr1_prob	pr1_logit	childless	age	country	educ
SI-211266-02	.0315996	.0324294	0	58,9	34. Slovenia	low
CY-485751-01	.0316	.0326038	1	65,6	53. Cyprus	low
CY-387943-01	.0316	.0326038	0	65,6	53. Cyprus	low

mergeid	pr1_prob	pr1_logit	childless	age	country	educ
CY-882492-03	.03161	.032658	0	68,1	53. Cyprus	low
BG-789343-02	.0316117	.0325155	0	61,8	51. Bulgaria	med
BG-953386-01	.0316117	.0325155	0	61,8	51. Bulgaria	med
DK-194270-02	.0316206	.032988	0	61,9	18. Denmark	low
DK-106979-01	.0316206	.032988	1	61,9	18. Denmark	low

Abbinamento statistico

- Come abbinare due (o più) record con una propensione al trattamento simile? Molti algoritmi per l'abbinamento probabilistico.

Figure 2: Different Matching Algorithms



Caliendo & Kopeinig (2005)
Some Practical Guidance for the Implementation of Propensity Score Matching

1. Nearest Neighbor matching (NN)

- L'unità nel gruppo di controllo da abbinare è quella più vicina in termini di propensity score, ossia dove: $\min|p_d - p_j|$

$$C(P_i) = \min_j \|P_i - P_j\|, \; j \in I_0.$$

  - 'with replacement': le unità dei controlli possono essere abbinati con più di un trattato (one-to-many). Particolarmente utile quando la distribuzione del propensity score è diversa tra i due gruppi.
  - 'without replacement': un abbinamento one-to-one tra I gruppi. Dipende dall'ordine con cui le unità trattate vengono abbinate
  - Le osservazioni che non hanno un accoppiamento vengono scartate (fuori dal *common support*).
  - Di solito imponiamo che l'unità più vicina non superi una distanza definita, come grado di tolleranza ϵ (Caliper). Difficile definire a priori il grado di tolleranza (Cochrane & Rubin, 1973).

NN esempio

- *Risultati:* la differenza stimata tra i sintomi di depressione tra chi ha figli e no aumenta dopo l'abbinamento statistico da 0.16 a 0.25

```
. psmatch2 childless c.age##c.age i.country i.educ if female==1 & age>49,   ///
> out(eurod) n(1) common caliper(.002)
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
eurod	Unmatched	2.9032419	2.74014238	.163099519	.054776591	2.98
	ATT	2.90464304	2.65651523	.248127808	.0803451	3.09

Note: S.E. does not take into account that the propensity score is estimated.

Output di stima: differenza di medie prima e dopo l'abbinamento.

psmatch2: Treatment assignment	psmatch2: Common support		Total
	Off suppo	On suppor	
Untreated	0	23,459	23,459
Treated	2	2,003	2,005
Total	2	25,462	25,464

Benchè 23k di controlli siano *on support*. Quelli che utilizziamo sono 1773

```
. tab _treat if _weight!=.
```

psmatch2: Treatment assignment	Freq.	Percent	Cum.
Untreated	1,765	46.84	46.84
Treated	2,003	53.16	100.00
Total	3,768	100.00	

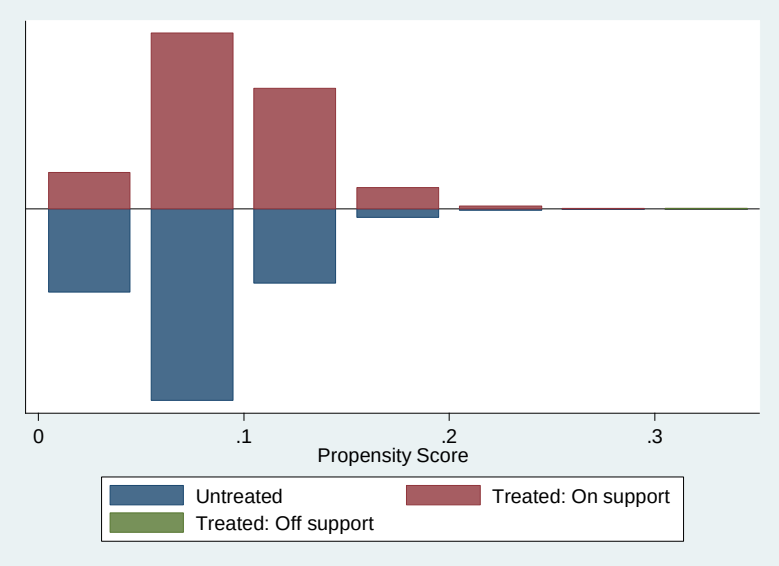
Nearest Neighbor matching (NN)

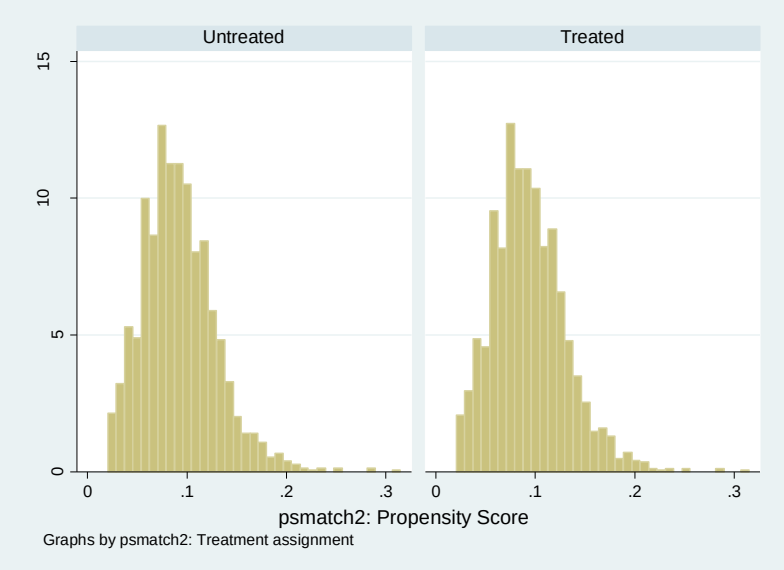
- Due unità fuori dal common support per una probabilità di essere senza figli molto alta (32.7% e 36%)
  - Nota 1: abbiamo imposto una differenza massima del propensity score di 0.2%. Ci sono possibili abbinamenti con un prob. alta ma che non rispetta la distanza massima.

mergeid	childless	educ	pr1_logit	pr1_prob	_pscore	_treated	_support	_weight	_eurod	_id	_n1	_nn	_pdif
Cg-644113-01	0	low	.306883	.2906801	.29068013	Untreated	On support	.	.	23492	.	0	.
IT-101327-01	0	low	.3135192	.2920178	.29201786	Untreated	On support	.	.	23493	.	0	.
Cf-271385-01	1	med	.3274284	.3044329	.30443293	Treated	Off support	.	.	25503	.	0	.
Cg-102212-02	0	low	.3263067	.3067937	.30679376	Untreated	On support	.	.	23494	.	0	.
Bf-041597-01	1	high	.3426791	.3133118	.31331184	Treated	On support	1	3	25504	23495	1	.00161951
Bf-850884-01	0	high	.3446701	.3149314	.31493135	Untreated	On support	1	.	23495	.	0	.
MT-265674-01	1	high	.3611447	.3236355	.32363548	Treated	Off support	.	.	25505	.	0	.
Cg-569670-01	0	med	.370741	.3397892	.33978915	Untreated	On support	.	.	23496	.	0	.

Diagnostica

- Se l'abbinamento statistico sta funzionando come vorremmo dovremmo trovare una simile distribuzione della propensione ad essere trattati tra i gruppi.
  - STATA: *psgraph* in cui consideriamo tutto il campione *on support* anche quelli non abbinati. L'istogramma presenta solo quelli abbinati.





Diagnostica

- Se l'abbinamento statistico sta funzionando come vorremmo, dovremmo trovare una simile distribuzione delle co-variate.
  - L'abbinamento ha creato una pseudo-popolazione selezionando e pesando le co-variate in modo che la loro distribuzione sia simile nei due gruppi. Nel NN i pesi sono uguali al n di unità abbinate (replacemente vs. no replacement)

Sample	Ps R2	LR chi2	p>chi2	MeanBias	MedBias	B	R	%Var
Unmatched	0.030	420.13	0.000	6.6	5.6	48.4*	0.90	100
Matched	0.004	22.80	0.823	2.3	2.4	15.1	1.23	0

* if B>25%, R outside [0.5; 2]

The standardised % bias rappresenta la differenza di medie come percentuale della radice quadrata delle varianze medie per I due gruppi (Rosenbaum and Rubin, 1985).

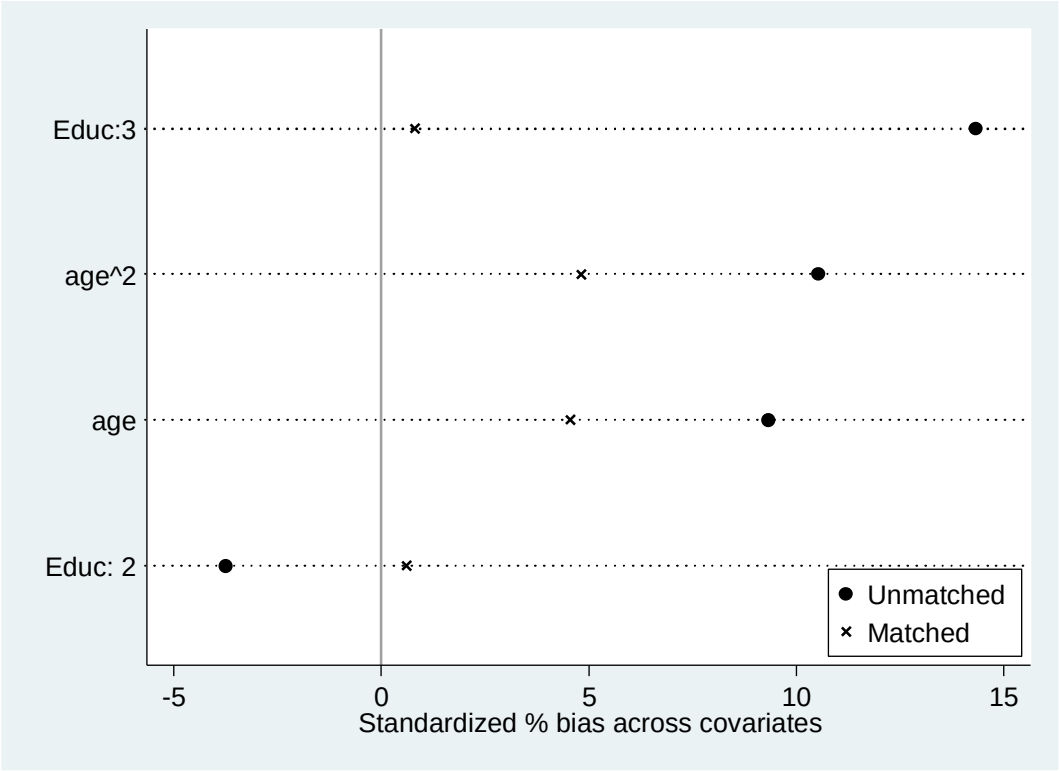
$$SB_{before} = 100 \cdot \frac{(\overline{X}_1 - \overline{X}_0)}{\sqrt{0.5 \cdot (V_1(X) + V_0(X))}}.$$

```
. pstest c.age##c.age i.country i.educ, both
```

Variable	Unmatched / Matched	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|	V(T)/V(C)
age	U	71.254	70.334	9.3		4.20	0.000	1.24*
	M	71.229	70.78	4.6	51.2	1.39	0.165	1.05
c.age#c.age	U	5184.8	5033.9	10.5		4.78	0.000	1.28*
	M	5180.7	5111.7	4.8	54.3	1.47	0.142	1.07
12.country	U	.06683	.05806	3.6		1.60	0.109	.
	M	.0669	.07339	-2.7	26.0	-0.80	0.421	.
13.country	U	.03491	.04906	-7.1		-2.85	0.004	.
	M	.03495	.03595	-0.5	92.9	-0.17	0.864	.
14.country	U	.04988	.03858	5.5		2.49	0.013	.
	M	.04993	.05192	-1.0	82.3	-0.29	0.774	.

Diagnostica

- La distorsione (o non bilanciamento) dei due gruppi dovrebbe diminuire e approssimarsi allo zero dopo l'abbinamento.

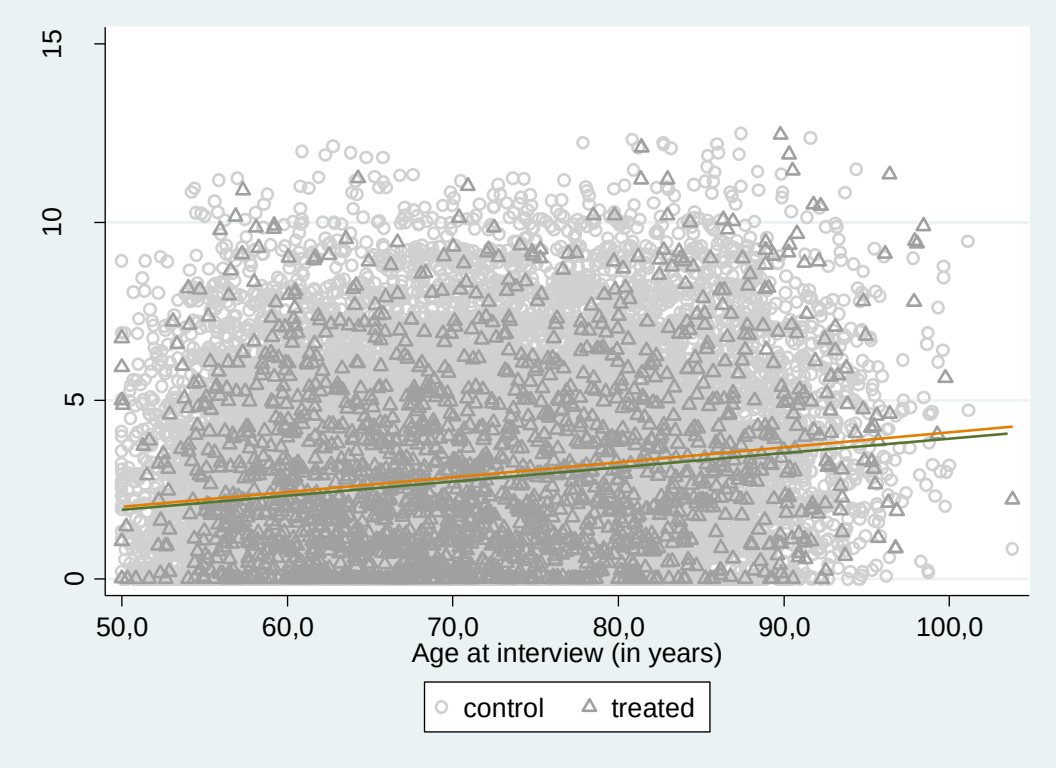


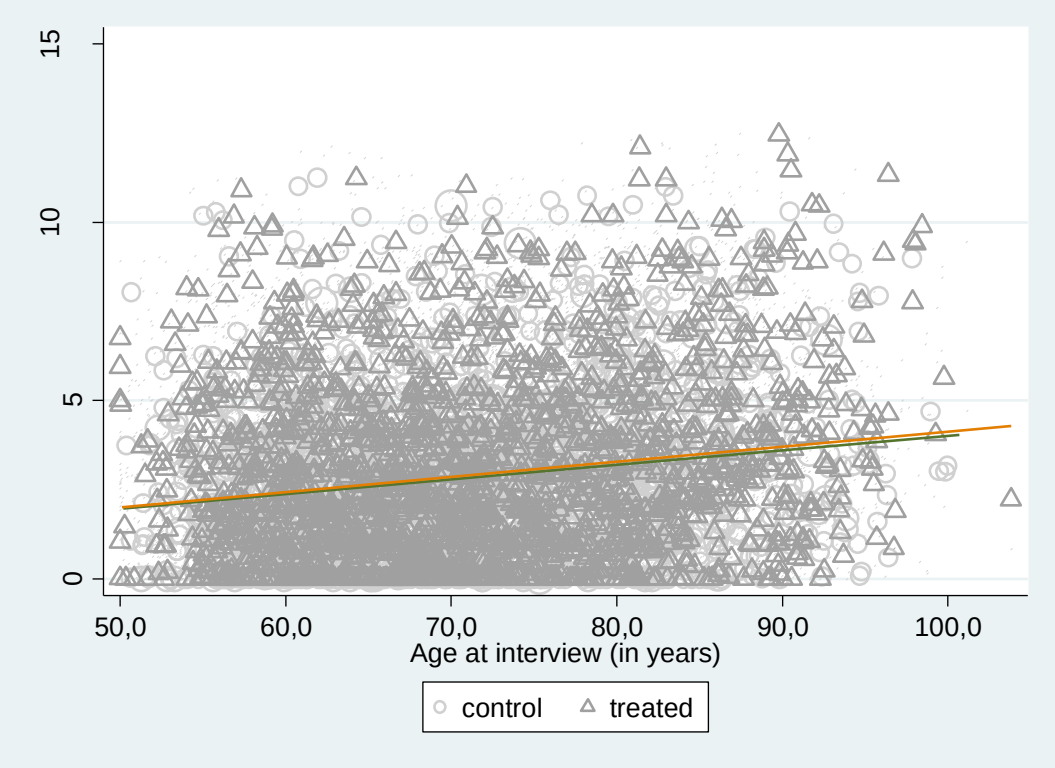

```
. pstest c.age##c.age i.country i.educ, both
```

Variable	Unmatched / Matched	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|	V(T)/ V(C)
age	U	71.254	70.334	9.3		4.20	0.000	1.24*
	M	71.229	70.78	4.6	51.2	1.39	0.165	1.05
c.age#c.age	U	5184.8	5033.9	10.5		4.78	0.000	1.28*
	M	5180.7	5111.7	4.8	54.3	1.47	0.142	1.07
12.country	U	.06683	.05806	3.6		1.60	0.109	.
	M	.0669	.07339	-2.7	26.0	-0.80	0.421	.
13.country	U	.03491	.04906	-7.1		-2.85	0.004	.
	M	.03495	.03595	-0.5	92.9	-0.17	0.864	.
14.country	U	.04988	.03858	5.5		2.49	0.013	.
	M	.04993	.05192	-1.0	82.3	-0.29	0.774	.

Stima

- Stima prima (grafico a sinistra) e dopo l'abbinamento (grafico a destra): l'associazione tra età e depressione diventa più simile tra i gruppi (casi più selezionati).





Stima

- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*,15(3), 199-236.

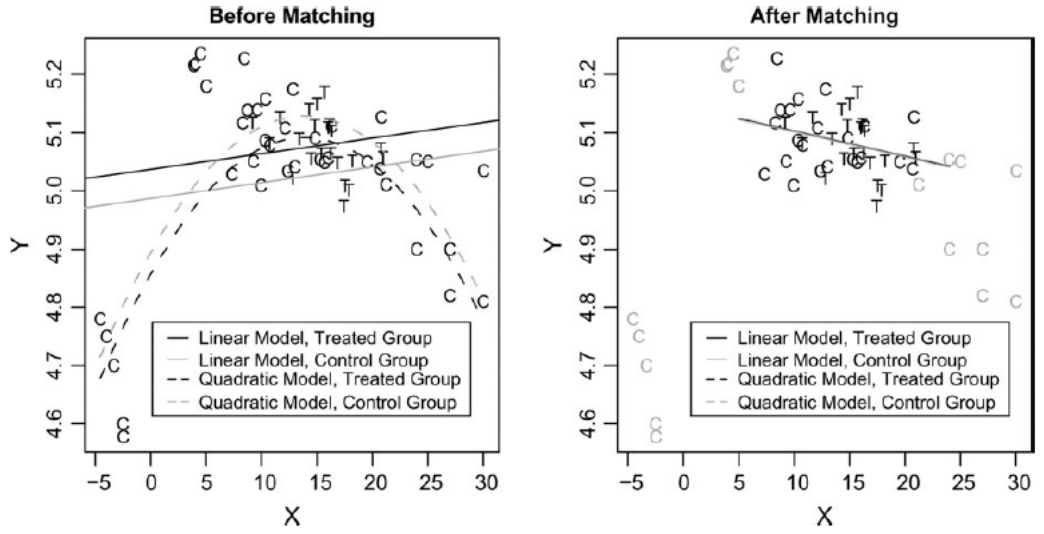


Fig. 1 Model sensitivity of ATE estimates for imbalanced raw and balanced matched data. This figure presents an artificial data set of treated units represented by "T" and control units represented by "C." The vertical axis plots Y_i and the horizontal axis plots X_i. The panels depict estimates of the ATE for a linear and quadratic specification, represented by the difference between parallel lines and parabolas, respectively. Dark lines are fitted to the treated points and gray to the controls. In the raw data, plotted in the left panel, some of the control units are far outside the range of the treated units, and these outlying control units are influential in the parametric models. In the matched data, plotted in the right panel, treated units are matched with control units that are close in X_i (gray units are discarded), and as a result treatment effect estimates are similar regardless of model specification. The two linear and two quadratic lines also appear on the right graph (on top of one another), truncated to the location of the matched data.

2. *NN* senza replacement

- Campione bilanciato tra trattati e controlli (tra coloro che rispettano il *common support* e differenze nel *propensity score* <0.2%)
  - Nota: la stima cambia: la differenza tra chi ha figli e no si riduce a 0.21 sintomi di depressione.

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
eurod	Unmatched	2.9032419	2.74014238	.163099519	.054776591	2.98
	ATT	2.90464304	2.69795307	.206689965	.076257559	2.71

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support		
	Off suppo	On suppor	Total
Untreated	0	23,459	23,459
Treated	2	2,003	2,005
Total	2	25,462	25,464

. tab _weight _treat

psmatch2: weight of matched controls	psmatch2: Treatment assignment		
	Untreated	Treated	Total
1	2,003	2,003	4,006
Total	2,003	2,003	4,006

3. *NN: abbinamento con le 5 unità più vicine*

- Campione bilanciato tra trattati e controlli (tra coloro che rispettano il *common support* e differenze nel *propensity score* <0.2%)
  - Nota: la stima cambia: la differenza è ora 0.20 sintomi di depressione.

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
eurod	Unmatched	2.9032419	2.74014238	.163099519	.054776591	2.98
	ATT	2.90464304	2.70534199	.199301048	.063450855	3.14

Note: S.E. does not take into account that the propensity score is estimated.

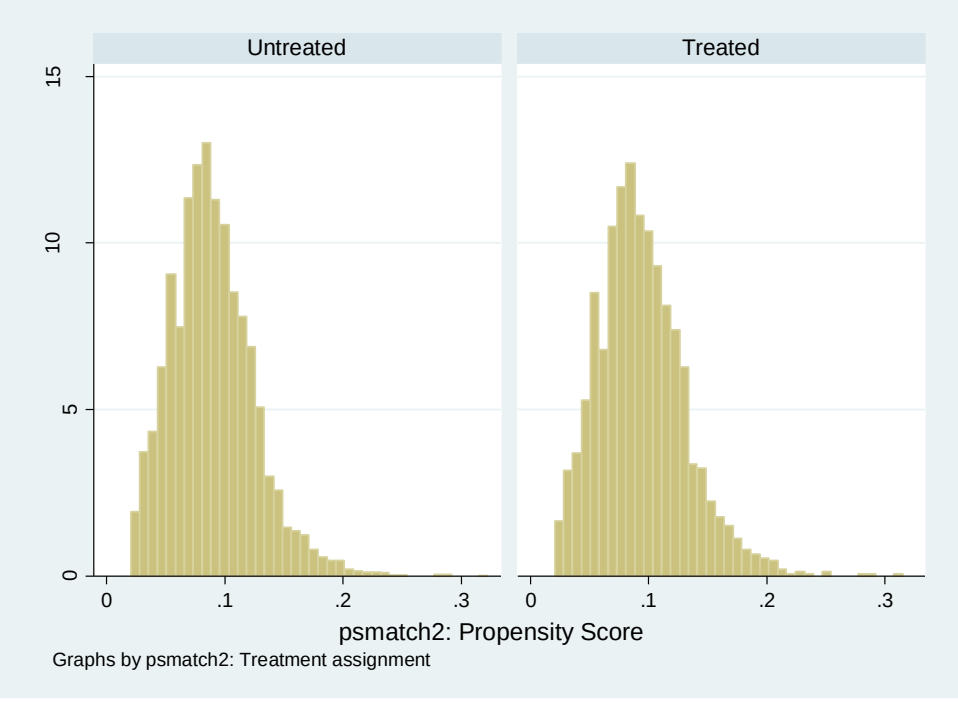
psmatch2: Treatment assignment	psmatch2: Common support		
	Off suppo	On suppor	Total
Untreated	0	23,459	23,459
Treated	2	2,003	2,005
Total	2	25,462	25,464

```
. tab _support _treat if _weight!=.
```

psmatch2: Common support	psmatch2: Treatment assignment		
	Untreated	Treated	Total
On support	7,552	2,003	9,555
Total	7,552	2,003	9,555

NN: abbinamento con le 5 unità più vicine

- *Diagnostica:* si perde di qualità negli abbinamenti. Tuttavia le co-variate sono bilanciate.
- *Avendo più di un abbinamento per ogni trattato i pesi di match diventano importanti.*



Rapporto di Rubin = il rapport tra la varianza del pscore dei due gruppi

Variable	Unmatched / Matched	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|	V(T)/ V(C)
2.educ	U	.39671	.41471	-3.7		-1.57	0.116	.
	M	.39661	.40634	-2.0	46.0	-0.63	0.530	.
3.educ	U	.27974	.21846	14.2		6.33	0.000	.
	M	.27952	.26891	2.5	82.7	0.75	0.451	.

Pesi di match

- I pesi di match tengono in considerazione di quante volte una unità dei controlli è stata abbinata con i trattati e quante volte le stesse unità dei trattati sono state abbinate.

Es. il caso 22311 è abbinato a 6 trattati che hanno 5 abbinamenti a testa: peso = 1/5*6 = 1.2

	mergeid	_id	_n1	_n2	_n3	_n4	_n5	_nn	_weight	_treated	_pscore
24080	GR-830126-02	22308	.	.	.	.	.	0	.6	Untreated	.13617788
24081	AT-100124-02	22309	.	.	.	.	.	0	.2	Untreated	.13619995
24082	NL-286818-02	25269	22310	22309	22311	22308	22307	5	1	Treated	.13623741
24083	Bf-901745-02	22310	.	.	.	.	.	0	1	Untreated	.13624918
24084	Cg-829055-01	25272	22311	22310	22312	22313	22314	5	1	Treated	.13628446
24085	Cg-394173-01	25273	22311	22310	22312	22313	22314	5	1	Treated	.13628446
24086	Ci-189037-01	25270	22311	22310	22312	22313	22314	5	1	Treated	.13628446
24087	Cg-272095-01	22311	.	.	.	.	.	0	1.2	Untreated	.13628446
24088	Cf-318351-01	25271	22311	22310	22312	22313	22314	5	1	Treated	.13628446
24089	GR-886696-01	25274	22312	22313	22314	22315	22311	5	1	Treated	.13631229

Pesi di match

- I pesi di match per coloro che sono nel gruppo dei trattati sono uguali a 1. Es. = 0.2 per i controlli che sono abbinati ad un solo trattato che ha un solo match all'interno del caliper.

```
. tab educ _treat [aw=_weight],col
```

in ISCED-97 Coding)	psmatch2: Treatment assignment Untreated	Treated	Total
low	1,620.088	1,547.977	3,168.065
	33.91	32.40	33.16
med	1,835.986	1,891.442	3,727.428
	38.43	39.59	39.01
high	1,321.425	1,338.082	2,659.5068
	27.66	28.01	27.83
Total	4,777.5	4,777.5	9,555
	100.00	100.00	100.00

```
. tab _weight _treat
```

psmatch2: weight of matched controls	psmatch2: Treatment assignment Untreated	Treated	Total
.2	5,645	0	5,645
.25	4	0	4
.3333333	5	0	5
.4	1,454	0	1,454
.45	2	0	2
.5333333	1	0	1
.5833333	2	0	2
.6	342	0	342
.8	79	0	79
1	15	2,003	2,018
1.2	2	0	2
1.333333	1	0	1
Total	7,552	2,003	9,555

4. Abbinamento Caliper e Radius

- Caliper: grado di tolleranza («propensity range») nel confrontare le unità con propensity score più vicino.
  - Aumenta la qualità dell'abbinamento ma aumenta anche la varianza delle stime, se non ci sono abbinamenti vicini.
  - Difficile definite a priori il grado di tolleranza.
- Radius: usa tutti i potenziali abbinamenti all'interno della soglia definita dal caliper (Dehejia & Wahba, 2002).
  - Aumenta il numero di abbinamenti quando non esistono unità simili/ vicine, abbassa la qualità degli abbinamenti.

$$C(i) = \{p_j | \|p_i - p_j\| < r\}$$

Esempio di Caliper / Radius

- Utilizziamo quasi tutte le informazioni disponibili nei dati.
  - La stima è più simile a quella pre-abbinamento: effetti di confondimento sono irrilevanti?

```
.  psmatch2 childless c.age##c.age i.country i.educ if female==1 & age>49,   ///
> out(eurod) radius common caliper(.002)
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
eurod	Unmatched	2.9032419	2.74014238	.163099519	.054776591	2.98
	ATT	2.90464304	2.735744	.168899036	.058737418	2.88

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support		
	Off suppo	On suppor	Total
Untreated	0	23,459	23,459
Treated	2	2,003	2,005
Total	2	25,462	25,464

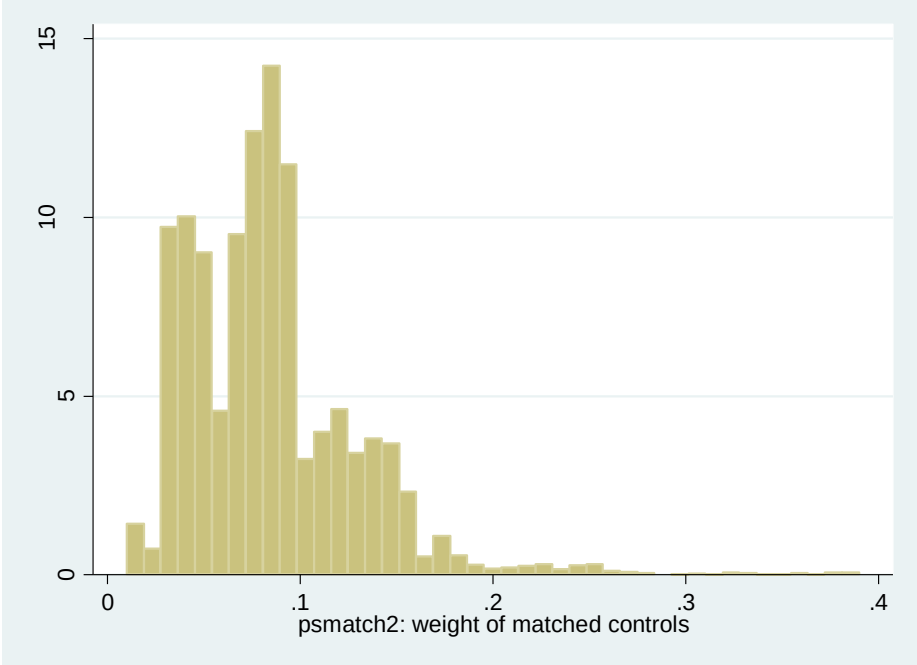
Esempio di Caliper / Radius

- Distribuzione dei pesi di match tra i controlli.
  - In questo caso la restrizione del Caliper è fondamentale per aumentare la qualità dell'abbinamento.

```
. tab _treat _support
```

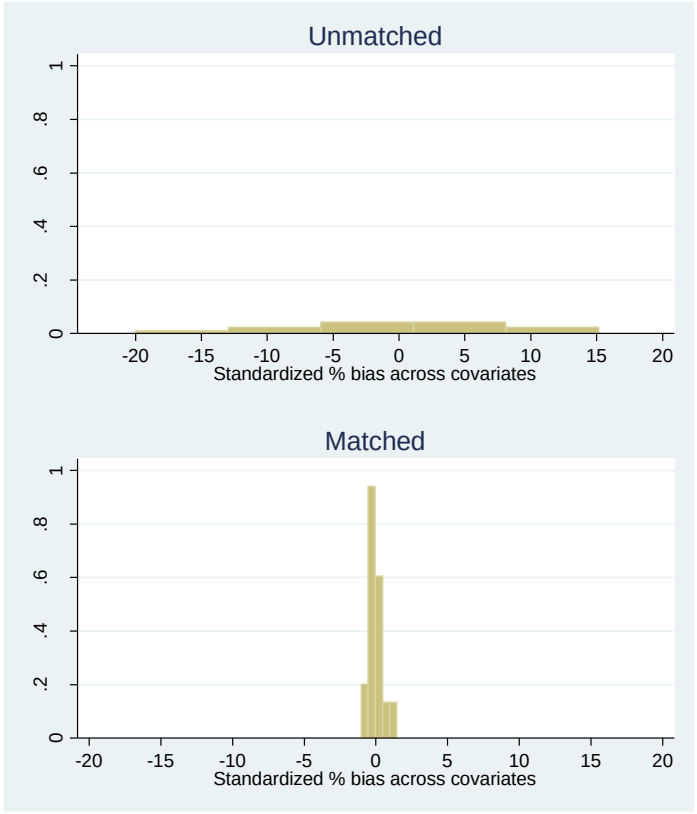
psmatch2: Treatment assignment	psmatch2: Common support		
	Off suppo	On suppor	Total
Untreated	0	23,459	23,459
Treated	2	2,003	2,005
Total	2	25,462	25,464

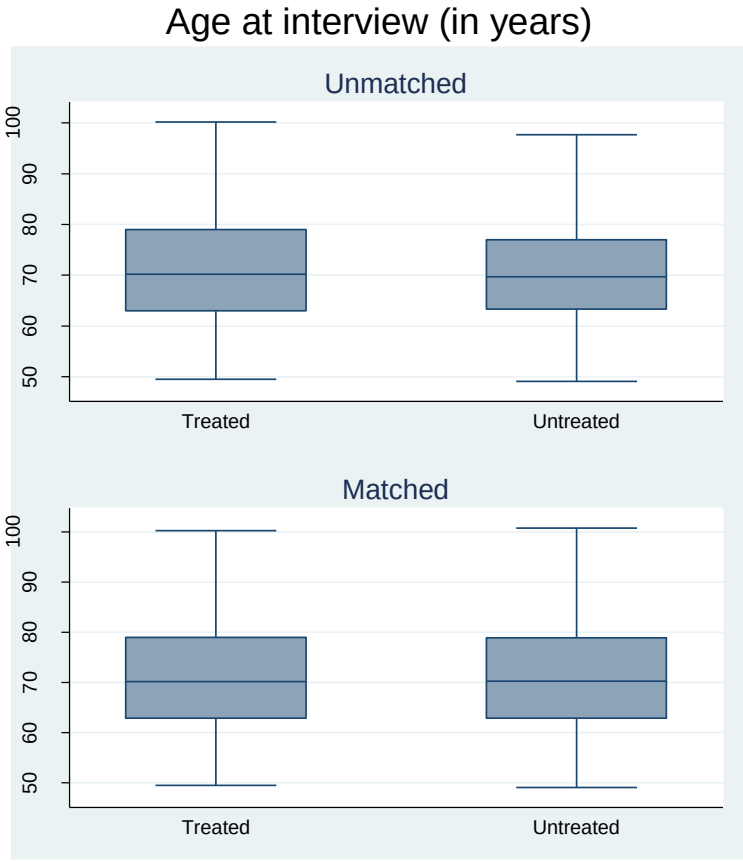
$$w_{ij} = \frac{1}{N_i^C} \text{ if } j \in C(i) \text{ and } w_{ij} = 0 \text{ otherwise.}$$

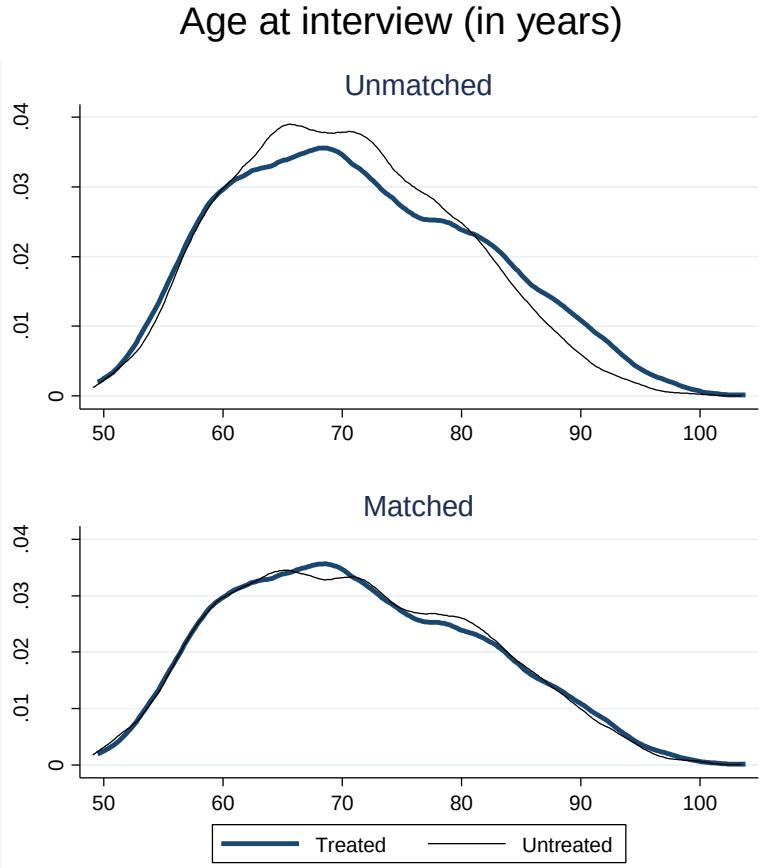


Diagnostica del Caliper / Radius

Differenza di medie tra gruppi come % della varianza







La direzione della distorsione

- Avere gruppi bilanciati per tenere in considerazione dell'effetto confondente delle variabili di controllo:

X->Y | C, dove X->Y *è sottostimato*

Quindi C è positivamente associato a X e negativamente a Y

RECODE of ch001_ (Number of children)	RECODE of isced1997_r (Education of respondent in ISCED-97 Coding)			
	low	med	high	Total
0	8,612	9,722	5,125	23,459
	92.99	92.45	90.12	92.13
1	649	794	562	2,005
	7.01	7.55	9.88	7.87
Total	9,261	10,516	5,687	25,464
	100.00	100.00	100.00	100.00

Mean estimation — Number of obs = 25464

low: educ = low
med: educ = med
high: educ = high

	Over	Mean	Std. Err.	[95% Conf.	Interval]
eurod					
	low	3.229025	.0271208	3.175867	3.282183
	med	2.572556	.0214102	2.530591	2.614521
	high	2.311412	.0271589	2.258179	2.364645

5. Kernel matching

- Il kernel matching è un metodo che usa i pesi di match per abbinare tutti i record del gruppo dei controlli con i trattati.
  - Specifichiamo la funzione (ad esempio, kernel normale) con la quale il propensity score si distribuisce attorno ai trattati (1) e specifichiamo il parametro bandwidth (valori alti la funzione è più smooth e consente stime migliori benchè si allontani dalla funzione realte (> bias)).
  - I pesi di match sono così stimati dalla funzione kernel scelta e la differenza tra il propensity score dei controlli e dei trattati. Il peso W dipende dalla distanza dei propensity score.

$$W_j(P(X_i)) = \frac{K(\frac{P(X_i)-P(X_k)}{h_n})}{\sum_{\{D_k=0\}}^{n_0}{}_{k=1} K(\frac{P(X_i)-P(X_k)}{h_n})}$$

Esempio di Kernel matching

```
.  psmatch2 childless c.age##c.age i.country i.educ if female==1 & age>49,   ///
> out(eurod) common kernel k(normal) bw(0.05)
```

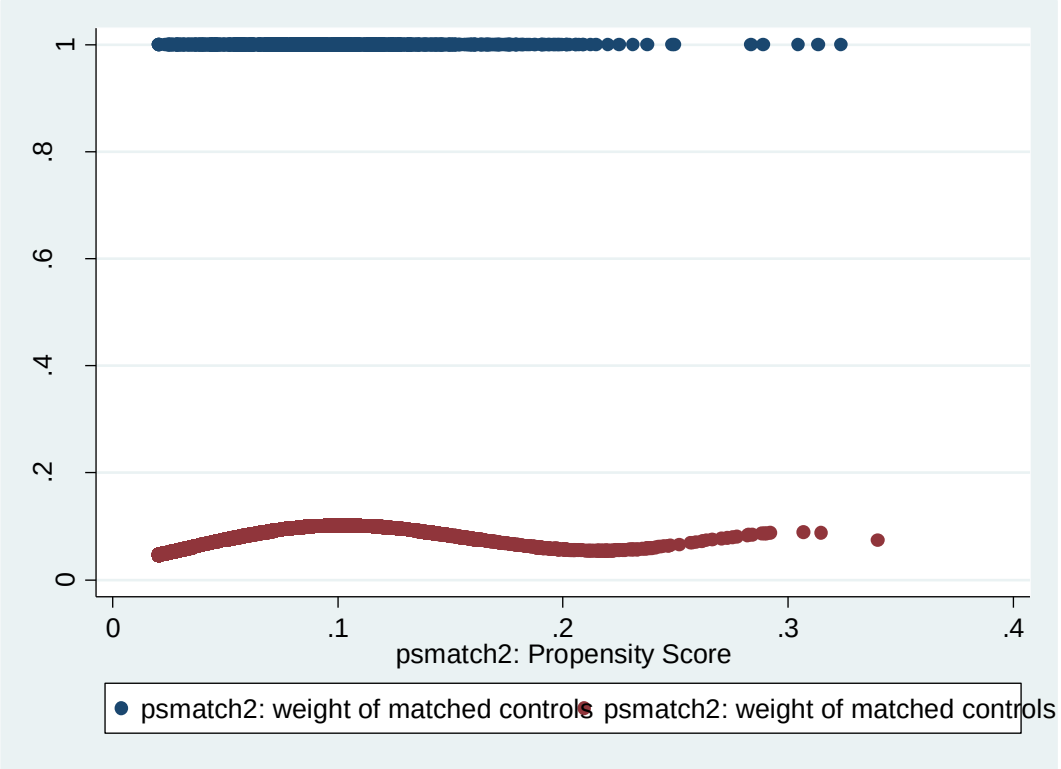
Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
eurod	Unmatched	2.9032419	2.74014238	.163099519	.054776591	2.98
	ATT	2.9032419	2.7438402	.159401691	.058122701	2.74

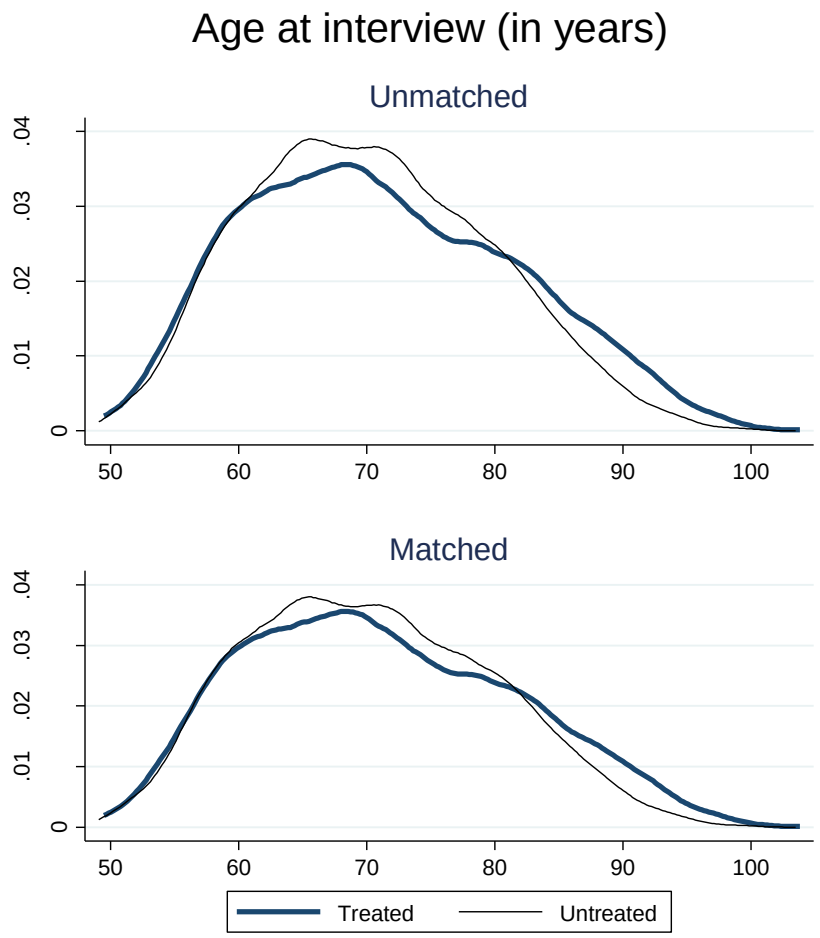
Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	23,459	23,459
Treated	2,005	2,005
Total	25,464	25,464

Esempio di Kernel matching

Pesi e propensity score





Diversi tipi di matching

Table 1: Trade-Offs in Terms of Bias and Efficiency

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 or LLM / NN-methods	(+)/(-)	(-)/(+)
Bandwidth choice with KM:		
small / large	(-)/(+)	(+)/(-)

KM: Kernel Matching, LLM: Local Linear Matching
NN: Nearest Neighbour
Increase: (+), Decrease: (-)

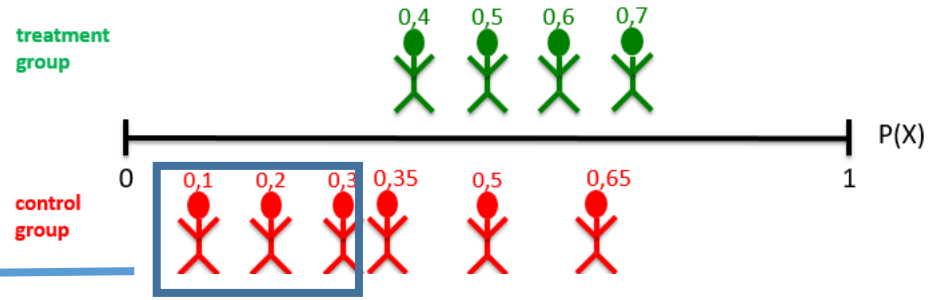
Caliendo et al. 2005

Diverse tecniche di abbinamento

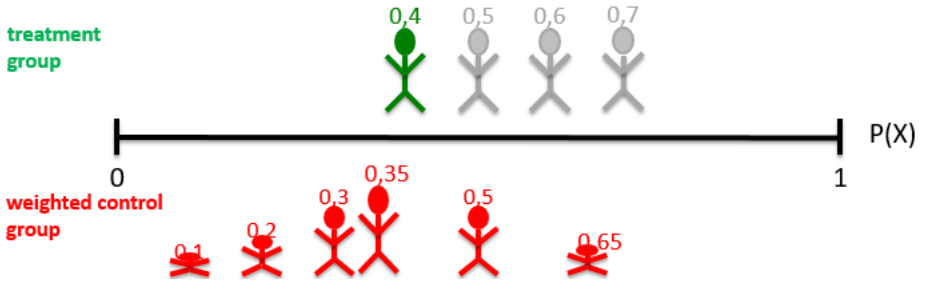
- **Nearest Neighbor matching & Radius / Caliper**

Fuori dalla distanza minima

$$C(P_i) = \min_j \|P_i - P_j\|, \; j \in I_0.$$



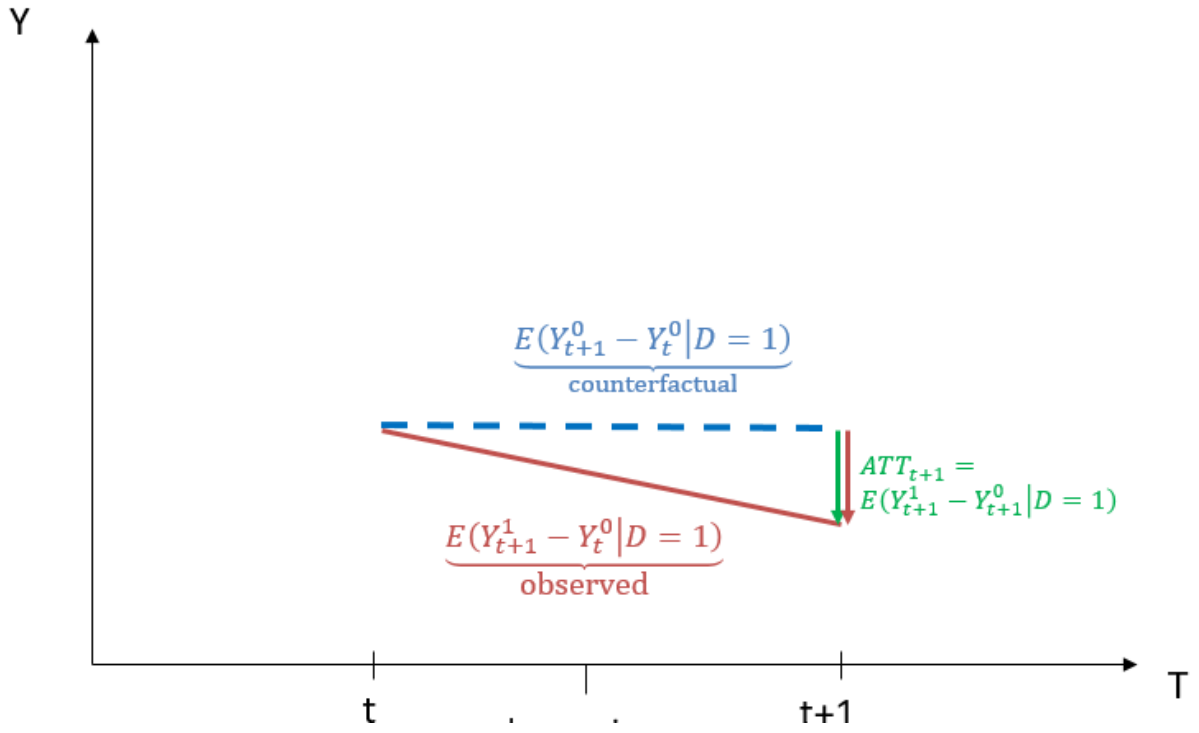
- **Kernel:** $W_{ij} = \dfrac{K\left[\dfrac{\left(P_j(X) - P_i(X)\right)}{h}\right]}{\Sigma_k K\left[\dfrac{\left(P_j(X) - P_i(X)\right)}{h}\right]}$ in funzione della distribuzione K (es, normale) e del parametro bandwith



Dati longitudinali

- Ipoteticamente vorremmo trovare due gruppi comparabili al tempo T il cui outcome Y non differisca tra i due gruppi (*trend paralleli*).

Comparazione prima-dopo evento/trattamento



Identificare un gruppo di controllo

Voßemer & Gebel (2021) SSM