Ripasso regressione e STATA

Valutazione delle politiche

(capitolo 1 Mastering Metrics)

Simona Comi

AA 2021/2022

Outline

A. Introduzione
B. Il problema fondamentale dell'inferenza causale (breve richiamo)
C. Dagli RCT alle regressioni: Avere una polizza assicurativa sanitaria fa stare meglio?

A. Introduzione

L’inferenza causale

- Spesso nelle scienze sociali si cercano risposte a domande come:
  - Quale è l’effetto della riduzione della classe sull’apprendimento dei bambini?
  - Una laurea aumenta il reddito?
  - Un corso di formazione aumenta la probabilità di trovare un’occupazione?
  - Avere un’assicurazione sanitaria migliora la salute?
  - Arrestare i criminali riduce il crimine?
  - Il vaccino causa la trombosi?
  - «L’helicopter money» è in grado di ridurre i danni causati dalla crisi economica causata dal Covid 19?

Domande come queste sono domande «What if», ovvero per rispondere vorremmo calcolare la differenza tra ciò che osserviamo essere successo dopo il trattamento e ciò che avremmo osservato in assenza del trattamento, che chiamiamo **situazione controfattuale**.

Avete visto come si identifica un nesso causale usando i RCT, oggi vediamo come possiamo farlo con una regressione

B. Il problema fondamentale dell’inferenza causale

La situazione controfattuale

- Le relazioni causali implicano un confront tra situazioni controfattuali del mondo, che sono come le due strade tra cui dobbiamo scegliere ad un bivio.
- Quando si sceglie una delle due strade, possiamo chiederci cosa sarebbe successo se avessimo scelto l'altra strada (il controfattuale), ma non saremo mai in grado di osservarlo.
- **La situazione controfattuale non è OSSERVABILE per definizione.**

Un primo esempio: l'assicurazione sanitaria negli Stati Uniti

- Abbiamo sentito tutti parlare del sistema sanitario americano e di come funziona…….
- Consideriamo i dati del NHIS, una survey che contiene informazioni sulla polizza sanitaria e anche informazioni sullo stato di salute degli individui, raccolte attraverso la domanda:
- Diresti che la tua salute in generale è: eccellente/molto buona/ buona/passabile/scarsa.
- Utilizzando le risposte a questa domanda possiamo calcolare un indice che assegna 5 allo stato di salute «eccellente» e 1 allo stato «scarsa»: questo indice sarà il nostro outcome, mentre la copertura assicurativa sarà il nostro trattamento

L'assicurazione sanitaria

In questo contesto gli individui con l'assicurazione sanitaria appartengono al gruppo dei trattati, gli individui senza assicurazione sanitaria apparterranno al gruppo di controllo.

➔ Chi ha l'assicurazione ha una salute migliore

TABLE 1.1

Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

E' corretto….

…..concludere che l'assicurazione sanitaria migliora la salute?

- NO: potremmo concludere che la salute migliora solo se assicurati e non fossero uguali in tutto e per tutto, ma nell'esempio chiaramente non è così.
- Si consideri l'istruzione degli individui, e ricordiamo che l'istruzione influenza la salute e influenza la probabilità di avere un assicurazione (come?)

TABLE 1.1

Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

Un po’ di nomenclatura

- Chiamiamo Y l’outcome, ovvero la salute
- Con un primo pedice indichiamo la persona che stiamo considerando, in generale indichiamo con *i* la generica persona i-esima, Y_i sarà il suo outcome.
- Ogni individuo ha però 2 outcome potenziali (pensate al bivio) e quindi a questa notazione dobbiamo aggiungere un secondo pedice, pari a 0 se non si è assicurati, 1 se si è assicurati.
- Y_{0i} Indicherà l’outcome nel caso in cui l’i-esima persona non è assicurata
- Y_{1i} Indicherà L’outcome nel caso in cui l’i-esima persona è assicurata
- Nell’esempio del bivio, i due outcome si trovano al termine delle due strade divergenti.
- L’effetto causale dell’assicurazione può essere scritto come:

$$Y_{1,i} – Y_{0,i}$$

Specifichiamo ulteriormente: Khuzdar

- Consideriamo lo studente in visita (visiting student) ad MIT Khuzdar Khalat dal Kazakhistan, paese in cui i cittadini hanno una copertura sanitaria gratuita e universale (come l'Italia) che li segue anche all'estero.
- All'arrivo al MIT a Khuzdar viene proposta l'assicurazione riservata agli studenti del college, che è abbastanza costosa.
- Temendo infezioni alle vie aree a causa del freddo clima del Massachusetts, Khuzdar accetta.
- Immaginiamo che $Y_{1,Khuzdar} = 4$ e $Y_{0,\ Khuzdar} = 3$ ➔ ovviamente non potremo mai osservare entrambi gli outcome, perché o Khuzdar sottoscrive l'assicurazione o non la sottoscrive.
- Per lui l'effetto causale dell'assicurazione è pari a :

$$Y_{1,Khuzdar} - Y_{0,\ Khuzdar} = 1$$

Specifichiamo ulteriormente: Maria

- Consideriamo la studentessa Maria Moreno in visita ad MIT dal Chile, abituata a climi ben più aspri, Maria non teme il clima del Massachusetts e non sottoscrive l'assicurazione.
- Siccome Maria si ammala raramente, avremo che $Y_{1,Maria} = 5$ e $Y_{0,Maria} = 5$, con un effetto causale dell'assicurazione che per Maria è pari a

$$Y_{1,Maria} - Y_{0,Maria} = 0$$

Il confronto tra Kuhzdar e Maria

- Kuhzdar e Maria fanno scelte assicurative diverse, hanno una salute diversa, e risponderanno rispettivamente 4 e 5 alla domanda sulla salute.
- Se calcolassimo la loro differenza di salute avremmo:

$Y_{Kuhzdar} - Y_{Maria} = -1$

Questa differenza (che osserviamo) ci porterebbe a conclusioni sbagliate, perché la salute di Kuhzdar è peggiore di quella di Maria, indipendentemente dall'assicurazione

Non osservabili

TABLE 1.2

Outcomes and treatments for Khuzdar and Maria

	Khuzdar Khalat	Maria Moreño
Potential outcome without insurance: Y_{0i}	3	5
Potential outcome with insurance: Y_{1i}	4	5
Treatment (insurance status chosen): D_i	1	0
Actual health outcome: Y_i	4	5
Treatment effect: $Y_{1i} - Y_{0i}$	1	0

Il confronto tra Kuhzdar e Maria

- Il confronto da i due studenti in realtà ci dice poco in merito all'effetto dell'assicurazione:

$$Y_{Kuhzdar} - Y_{Maria} = Y_{1,\,Kuhzdar} - Y_{0,\,Maria} = -1$$

- Aggiungiamo e sottraiamo $Y_{0,\,Kuhzdar}$ all'equazione sopra, otteniamo:

$$Y_{Kuhzdar} - Y_{Maria} = Y_{1,\,Kuhzdar} - Y_{0,\,Kuhzdar} + Y_{0,\,Kuhzdar} - Y_{0,\,Maria} = -1$$

1 -2

Con questa aggiunta, abbiamo mostrato 2 confronti nascosti: il primo termine, è l'effetto causale dell'assicurazione su Kuhzdar, ed è pari a 1;

Il secondo è la differenza nella salute dei due studenti quando non sono assicurati e riflette le loro differenze di salute, che chiameremo effetto distorsivo della selezione (**selection bias**).

E se utilizzassimo più individui?

- Si potrebbe sperare che queste differenze tra individui spariscano (annullandosi a vicenda?) quando consideriamo più individui invece che solo due.
- Ma non è così. L'**effetto causale** in un gruppo di n persone può essere scritto come (nb: avg=average=media):
- $Avg_n[Y_{1i} - Y_{0i}] = \frac{1}{n}\sum_{1=1}^{n}[Y_{1i} - Y_{0i}] = \frac{1}{n}\sum_{1=1}^{n} Y_{1i} - \frac{1}{n}\sum_{1=1}^{n} Y_{0i}$
- Nel caso dei nostri due individui, questo calcolo equivale a sommare (e poi divedere per n)$Y_{1,Khuzdar}$ - $Y_{0,\ Khuzdar}$ e $Y_{1,Maria}$ - $Y_{0,\ Maria}$ per tutti gli n studenti.

Il confronto tra i gruppi

- Ma in realtà nel confrontare la salute di chi ha e chi non ha l'assicurazione, quello che tutti faremmo è costruire una variabile binaria
- $D_i = \begin{cases} 1 \; se \; i \text{ è } assicurato \\ 0 \qquad\quad altrimenti \end{cases}$
- Possiamo quindi scrivere le medie condizionate sullo stato assicurativo:

$Avg_n[Y_i|D_i = 1]$ media tra gli assicurati, in realtà $Avg_n[Y_{1i}|D_i = 1]$

$Avg_n[Y_i|D_i = 0]$ media tra i non assicurati, in realtà $Avg_n[Y_{0i}|D_i = 0]$

Nessuno dei due contiene informazioni sulla situazione controfattuale.

Queste medie equivalgono ai numeri contenuti nella tabella alla slide 8

Il confronto tra i gruppi

- E’ chiaro che

$$Avg_n[Y_{1i} - Y_{0i}] \neq Avg_n[Y_{1i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0]$$

- Perché:

$$Avg_n[Y_{1i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0] =$$
$$Avg_n[Y_{1i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 1] + Avg_n[Y_{0i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0]$$

Differenza tra le medie dei gruppi=

Effetto causale medio + distorsione da selezione

La distorsione da selezione è quindi calcolata come differenza tra la salute media che avrebbero gli assicurati senza assicurazione e che hanno i non assicurati senza assicurazione (differenza sistematica)

C. Dagli RCT alla regressione

La distorsione da selezione: soluzioni

- La distorsione da selezione o *selection bias* dipende da tutte le caratteristiche della persona *i osservabili e non* legate alla salute (ricordate le variabili omesse?), in primis l'istruzione.
- Quando le caratteristiche sono osservabili sappiamo come fare per eleminare la distorsione da selezione (prendendo la media delle medie condizionate a un certo livello di istruzione o facendo una regressione e includendo l'istruzione)
- Quando non sono osservabili, la soluzione non sempre esiste… ma laddove è possibile, la randomizzazione è la risposta.
- Randomizzare può risolvere il problema, perché sappiamo che gli esperimenti randomizzati controllati ideali permettono di risolvere la distorsione da selezione (selection bias).

La randomizzazione elimina il *selection bias*

Assegnando la polizza sanitaria in modo random all'interno di una popolazione, avrò che $Avg_n[Y_{0i}|D_i = 1] = Avg_n[Y_{0i}|D_i = 0]$

Quindi:

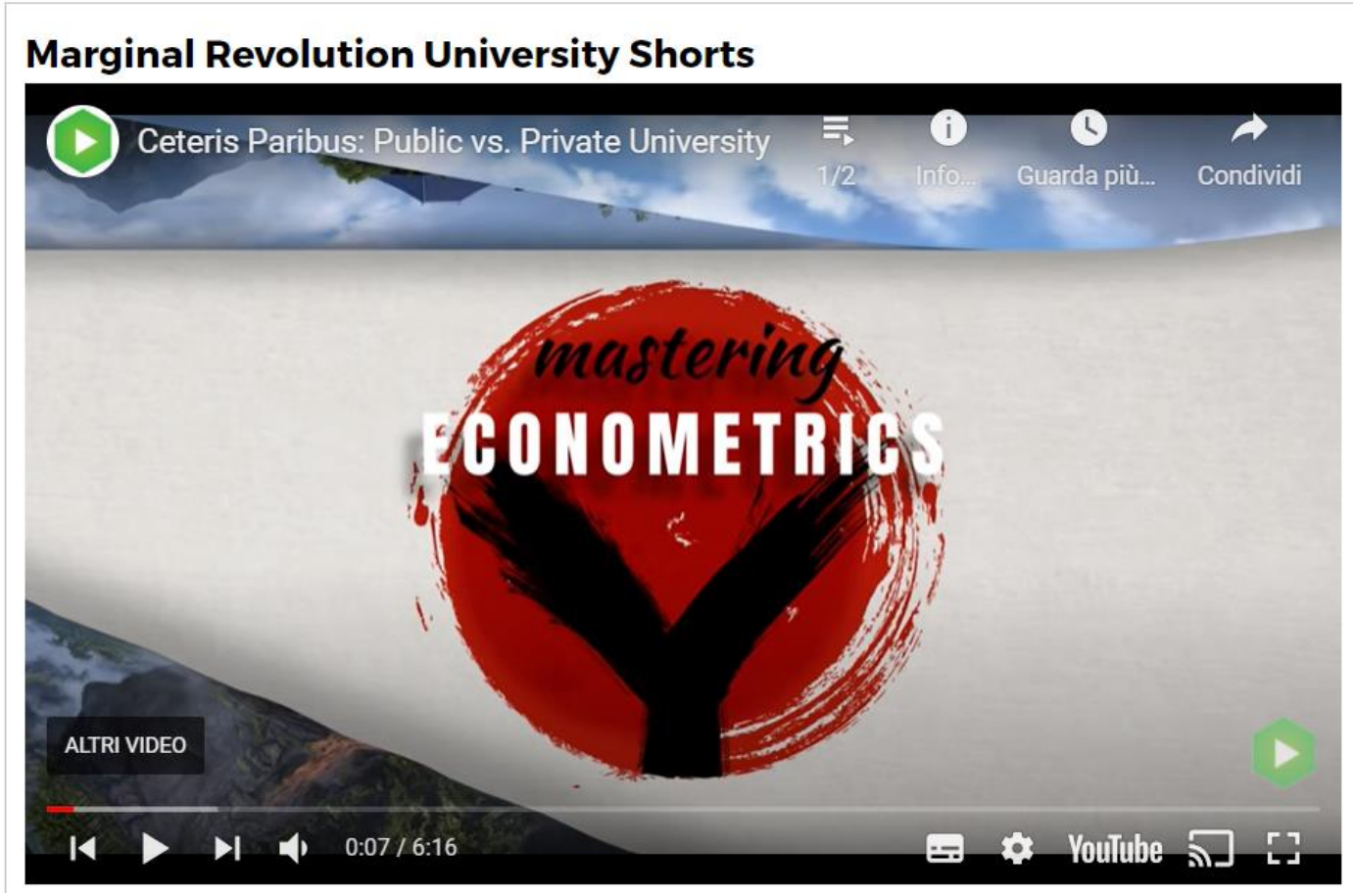
$$Avg_n[Y_{1i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0] =$$
$$Avg_n[Y_{1i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 1] + Avg_n[Y_{0i}|D_i = 1] - Avg_n[Y_{0i}|D_i = 0]$$

Differenza tra le medie dei gruppi = Effetto causale medio

Ma se non possiamo assegnare in modo randomico una polizza assicurativa, come possiamo identificarne l'effetto?

Uno strumento utile è la **regressione**

Video MM


Marginal Revolution University Shorts
Ceteris Paribus: Public vs. Private University
1/2
Info...
Guarda più...
Condividi
mastering
ECONOMETRICS
ALTRI VIDEO
0:07 / 6:16
YouTube

Matchmaker, Matchmaker

- La regressione con più regressori può essere utilizzata per controllare per i cosidetti «confoudning factors» e fare confronti a parità di tutto il resto (ceteris paribus).
- La regressione con più regressori è uno strumento automatico per accoppiare le osservazioni.
- Spesso siamo interessati alla relazione tra una variabile dipendente, Y_i, e un'altra variabile, X_{1i} , in uno scenario in cui X_{1i} è associata anche a X_{2i}, la quale predice anche Y_i.
- Considerate poer esempio, l'associazione tra l'assicurazione sanitaria e la salute nell'esempio NHIS, che può essere spiegata dal più elevato livello di istruzione degli assicurati.
- Questa ulteriore associazione crea quello che abbiamo chiamato il *selection bias,* e che nel contesto delle regressioni si chiama *omitted variables bias.*

Traduzione in italiano di http://www.masteringmetrics.com/wp-content/uploads/2020/07/lny20n07MRU_R1.pdf

Matchmaker, Matchmaker - cont

- Per illustrare, immaginate che la variabile X_{1i} sia una dummy, noi siamo interessati a

$$E[Y_i \mid X_{1i} = 1] - E[Y_i \mid X_{1i} = 0]$$

che nel caso di medie condizionate diventa:

$$E[Y_i \mid X_{1i} = 1, X_{2i} = x] - E[Y \mid X_{1i} = 0, X_{2i} = x]$$

Che è esattamente quello che la regressione lineare di Y_i su X_{1i} e X_{2i} stima.

$$E[Y_i \mid X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Asian and white under control

- In un campione di giovani uomini laureati nella American Community Survey, gli asiatici (75% nati all'estero) guadagnano più dei bianchi.

```
59 . summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
agep	57,696	44.632	2.834972	40	49
wagp	57,696	85197.99	88589.83	0	714000
wkhp	57,696	45.16632	10.0141	1	99
racasn	57,696	.0987243	.2982941	0	1
racpi	57,696	.0009706	.0311397	0	1
racwht	57,696	.9083299	.2885622	0	1
uhe	56,924	34.81603	29.37749	0	201.5789
loguhe	53,750	3.361481	.7235695	-6.437752	5.306181
immig	57,696	.1748475	.3798399	0	1
yearsEd	57,696	14.53616	2.42775	12	21
hsgrad	57,696	1	0	1	1
somecol	57,696	.5348551	.498788	0	1
colgrad	57,696	.4422664	.4966599	0	1
asianpac	57,696	.0987243	.2982941	0	1
white	57,289	.9076786	.2894816	0	1

```
60 . bys asianpac: summarize loguhe yearsEd colgrad immig
```

-> asianpac = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
loguhe	48,411	3.345458	.7155705	-6.437752	5.306181
yearsEd	52,000	14.40617	2.377595	12	21
colgrad	52,000	.4188462	.4933748	0	1
immig	52,000	.11025	.3132041	0	1

-> asianpac = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
loguhe	5,339	3.506776	.7775642	-3.912023	5.30231
yearsEd	5,696	15.72279	2.555968	12	21
colgrad	5,696	.6560744	.4750583	0	1
immig	5,696	.7645716	.4243035	0	1

Model	125.139748	1	125.139748	F(1, 53748)	= 240.08
Residual	28015.2987	53,748	.521234253	Prob > F	= 0.0000
				R-squared	= 0.0044
Total	28140.4384	53,749	.523552781	Adj R-squared	= 0.0044
				Root MSE	= .72197

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
asianpac	.1613188	.0104113	15.49	0.000	.1409126	.1817249
_cons	3.345458	.0032813	1019.56	0.000	3.339026	3.351889

```
3 . reg loguhe asianpac colgrad
```

Source	SS	df	MS	Number of obs	= 53,750
				F(2, 53747)	= 5623.46
Model	4869.57938	2	2434.78969	Prob > F	= 0.0000
Residual	23270.859	53,747	.43297038	R-squared	= 0.1730
				Adj R-squared	= 0.1730
Total	28140.4384	53,749	.523552781	Root MSE	= .658

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
asianpac	.016507	.0095892	1.72	0.085	-.002288	.0353019
colgrad	.6043304	.0057731	104.68	0.000	.593015	.6156457
_cons	3.091722	.0038496	803.14	0.000	3.084176	3.099267

```
64 . reg loguhe asianpac yearsEd
```

Source	SS	df	MS		
Model	5332.56924	2	2666.28462	Number of obs	= 53,750
Residual	22807.8692	53,747	.424356134	F(2, 53747)	= 6283.13
				Prob > F	= 0.0000
				R-squared	= 0.1895
				Adj R-squared	= 0.1895
Total	28140.4384	53,749	.523552781	Root MSE	= .65143

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
asianpac	-.0109312	.0095219	-1.15	0.251	-.0295942	.0077317
yearsEd	.1301684	.0011751	110.78	0.000	.1278653	.1324715
_cons	1.469512	.0171914	85.48	0.000	1.435816	1.503207

Frequentare una università privata premia?

- Guardiamo ora alle differenze nel reddito associate all'essersi laureati in una università pubblica o privata (o selettiva e non selettiva).
- Che esista un premio salariale associato alla frequenza a college privati (che sono anche più selettivi) è indubbio.
- Ma questa differenza è davvero imputabile all'università frequentata? O piuttosto dipende dalle differenze nelle caratteristiche degli studenti che si iscrivono ad un corso piuttosto che all'altro?
- Consideriamo una ipotetica matrice di ammissione al college

La matrice di ammissione di 9 studenti

TABLE 2.1

The college matching matrix

Applicant group	Student	Private			Public			1996 earnings
		Ivy	Leafy	Smart	All State	Tall State	Altered State	
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Un possibile confronto

- Possiamo confrontare gli studenti 1 e 2 con lo studente 3 nel gruppo A, e lo studente 4 con lo studente 5 nel gruppo B…..
- Ma invece di fare un confronto così macchinoso, potremmo semplicemente regredire

$$Y_i = \alpha + \beta P_i + \gamma A_i + \varepsilon_i$$

- Dove P è la dummy privato e A è la variabile che indica gli studenti nel gruppo A.
- Dalla stima troviamo questi risultati

$$\alpha = 40{,}000$$
$$\beta = 10{,}000$$
$$\gamma = 60{,}000.$$

Uno studio più ampio

- Utilizzando i dati College and Beyond (C&B) che comprendono informazioni su 5583 laureati da 30 università diverse possono essere raggruppati in 151 diversi gruppi, definiti sulla base del grado di selettività dell'università, che comprende università sia pubbliche che private.
- Possiamo quindi stimare il seguente modello

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + \varepsilon_i$$

I risultati delle regressioni

	No Selection Controls			Selection Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private School	0.135	0.095	0.086	0.007	0.003	0.013
	(0.055)	(0.052)	(0.034)	(0.038)	(0.039)	(0.025)
Own SAT score/100		0.048	0.016		0.033	0.001
		(0.009)	(0.007)		(0.007)	(0.007)
Predicted log(Parental Income)			0.219			0.190
			(0.022)			(0.023)
Female			-0.403			-0.395
			(0.018)			(0.021)
Black			0.005			-0.040
			(0.041)			(0.042)
Hispanic			0.062			0.032
			(0.072)			(0.070)
Asian			0.170			0.145
			(0.074)			(0.068)
Other/Missing Race			-0.074			-0.079
			(0.157)			(0.156)
High School Top 10 Percent			0.095			0.082
			(0.027)			(0.028)
High School Rank Missing			0.019			0.015
			(0.033)			(0.037)
Athlete			0.123			0.115
			(0.025)			(0.027)
Selection Controls	N	N	N	Y	Y	Y

Notes: Columns (1)-(3) include no selection controls. Columns (4)-(6) include a dummy for each group formed by matching students according to schools at which they were accepted or rejected. Each model is estimated using only observations with Barron's matches for which different students attended both private and public schools. The sample size is 5,583. Standard errors are shown in parentheses.

E con ulteriori controlli

	No Selection Controls			Selection Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
School Avg. SAT Score/100	0.109	0.071	0.076	-0.021	-0.031	0.000
	(0.026)	(0.025)	(0.016)	(0.026)	(0.026)	(0.018)
Own SAT score/100		0.049	0.018		0.037	0.009
		(0.007)	(0.006)		(0.006)	(0.006)
Predicted log(Parental Income)			0.187			0.161
			(0.024)			(0.025)
Female			-0.403			-0.396
			(0.015)			(0.014)
Black			-0.023			-0.034
			(0.035)			(0.035)
Hispanic			0.015			0.006
			(0.052)			(0.053)
Asian			0.173			0.155
			(0.036)			(0.037)
Other/Missing Race			-0.188			-0.193
			(0.119)			(0.116)
High School Top 10 Percent			0.061			0.063
			(0.018)			(0.019)
High School Rank Missing			0.001			-0.009
			(0.024)			(0.022)
Athlete			0.102			0.094
			(0.025)			(0.024)
Average SAT Score of				0.138	0.116	0.089
Schools Applied To/100				(0.017)	(0.015)	(0.013)
Sent Two Application				0.082	0.075	0.063
				(0.015)	(0.014)	(0.011)
Sent Three Applications				0.107	0.096	0.074
				(0.026)	(0.024)	(0.022)
Sent Four or more Applications				0.153	0.143	0.106
				(0.031)	(0.030)	(0.025)

Note: Standard errors are shown in parentheses. The sample size is 14,238.