Principi di Econometria

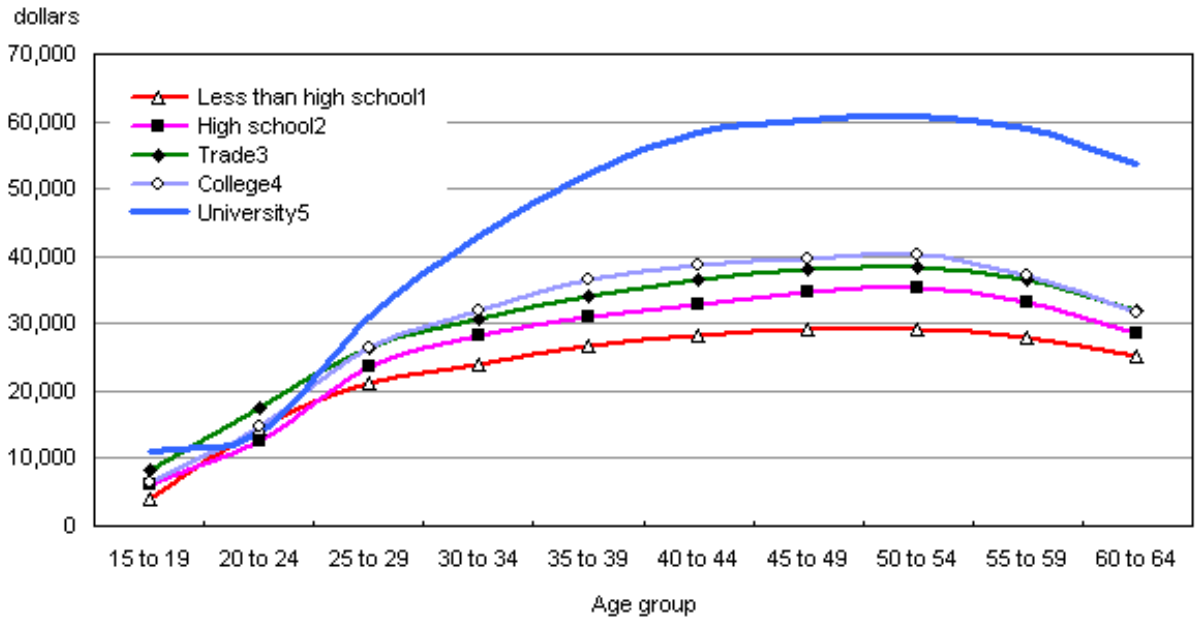
lezione 12

AA 2016-2017

Paolo Brunori

Interazione di variabili

- potrebbe essere che l'effetto dell'esperienza dipenda dal livello di istruzione



stime su dati Canadesi: www.statcan.gc.ca

Interazione di variabili

- in questo caso vorremmo poter tener conto del fatto che la distanza fra un istruito e un non istruito aumenta all'aumentare dell'esperienza
- ci serve una variabile che a parità di età sia più elevata se il livello di istruzione è alto
- analogamente: sia più alta quando l'età è elevata a parità di istruzione
- in un certo senso è una situazione simile a quella vista per l'interazione fra variabili dicotomiche e non dicotomiche (o fra due dicotomiche)
- anche in questo caso il termine $età \times edu$ può essere usato per risolvere il problema

interazione fra variabili continue: formula generale

- in generale per capire quale relazione coglie una trasformazione non lineare di un regressore si deve svolgere l'equazione
  $\Delta Y = f(X_1, ..., X_j + \Delta X_j, ..., X_k) - f(X_1, ..., X_j, ..., X_k)$
- e calcolare la variazione di Y dovuta alla variazione di X_j
- se $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2)$ abbiamo:

$$\Delta Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_2 + \Delta X_2) + \beta_3 [X_1 \times (X_2 + \Delta X_2)] + \\ - [\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2)]$$

$$\Delta Y = \beta_2 \Delta X_2 + \beta_3 (X_1 \times \Delta X_2)$$

interazione fra variabili continue: formula generale

$$\Delta Y = \beta_2 \Delta X_2 + \beta_3 (X_1 \times \Delta X_2)$$

$$\frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1$$

Per cui la variazione di Y dovuta ad una variazione di X_2 dipende dal valore di X_1

interazione fra variabili continue

- il nuovo modello sarà:

$$Y_I = \beta_0 + \beta_1 age + \beta_2 edu + \beta_3 sex + \beta_4 eta^2 + \beta_5 (eta \times edu)$$

	coefficiente	errore standard	t	$p - value$
β_0	-1694.34	7644.765	-0.222	0.8246
β_{age}	847.69	217.519	3.897	0.0001
β_{age^2}	-9.81	1.805	-5.43	0.0000
β_{edu}	158.68	459.22	0.346	0.7297
$\beta_{age \times edu}$	16.798	8.30	2.023	0.0433
β_{sex}	1673.6	856.5	1.954	0.0510

- $R^2 = 0.1423$, $R^2 - \text{corretto} = 0.1371$
- errore standard di regressione $= 12340$

funzioni di logaritmi

- spesso si esprimono relazioni economiche in termini percentuali: l'elasticità è l'esempio più comune:

$$\eta_{Y,X} = \frac{\Delta Y/Y}{\Delta X/X} = \frac{\Delta \% Y}{\Delta \% X}$$

- l'uso della funzione logaritmica di un regressore ci permette di cogliere una relazione non lineare e di interpretare il coefficiente in termini di variazioni %
- la funzione logaritmica è la funzione inversa della funzione esponenziale

$$f(x) = e^x \rightarrow f^{-1}(x) = log_e(x) = ln(x)$$

- ovvero

$$x = ln(e^x)$$

proprietà di ln

- $ln(1/x) = -ln(x)$
- $ln(ax) = ln(a) + ln(x)$
- $ln(x/a) = ln(x) - ln(a)$
- $ln(x^a) = aln(x)$

ln e variazioni %

- la proprietà che ci interessa di più (valida se ΔX è piccolo)

$$ln(x + \Delta x) - ln(x) \cong \frac{\Delta x}{x}$$

- dove il simbolo $\cong$ indica 'approssimativamente uguale a'

regressione di funzioni logaritmiche

- 3 modelli possibile
  1. X è espressa in logaritmi Y no
  2. Y è espressa in logaritmi X no
  3. X e Y sono entrambe espresse in logaritmi

modello lineare-logaritmico

- se X è logaritmica e Y no

$$Y_i = \beta_0 + \beta_1 ln(X_i) + u_i$$

- ad una variazione dell'1% di X si associa una variazione di $0.01\beta_1$ di Y
- infatti

$$\Delta X \to \Delta Y = [\beta_0 + \beta_1 ln(X + \Delta X)] - [\beta_0 + \beta_1 ln(X)] =$$

$$\Delta Y = \beta_1 ln(X + \Delta X) - \beta_1 ln(X) \cong \beta_1 \frac{\Delta X}{X}$$

modello log-lineare

- se è la Y ad essere espressa in forma logaritmica ma la X no
- $ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$
- a seguito di ΔX il valore di Y è $ln(Y + \Delta Y)$:

$$ln(Y + \Delta Y) - ln(Y) = [\beta_0 + \beta_1 (X + \Delta X)] - [\beta_0 + \beta_1 X]$$

$$\frac{\Delta Y}{Y} \cong \beta_1 \Delta X$$

modello log-log

- nel caso in cui entrambe le variabili siano espresse in termini logaritmici

$$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + u_i$$

- β è la variazione percentuale di Y dovuta ad una variazione percentuale di X: $(\eta_{Y,X}!)$

Modello log-log

$$ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + u_i$$

$$ln(Y + \Delta Y) - ln(Y) = \beta_0 + \beta_1 ln(X + \Delta X) - [\beta_0 + \beta_1 ln(X)]$$

$$\frac{\Delta Y}{Y} \cong \beta_1 \frac{\Delta X}{X}$$

$$\beta_1 = \eta_{X,Y}$$

Elasticità del reddito all'età (Svezia)

	coefficiente	errore standard	t	$p-value$
β_0	9.4536	0.1983	47.655	0.000
$\beta_{ln(age)}$	0.16609	0.05163	3.217	0.0013

$R^2 = 0.01232$, $R^2 - \text{corretto} = 0.01113$

errore standard di regressione $= 0.5324$

Elasticità del reddito all'età (Svezia)

	coefficiente	errore standard	t	$p-value$
β_0	8.1009	0.2893	27.99	0.0000
$\beta_{ln(age)}$	0.2956	0.0523	5.64	0.0000
β_{male}	0.0594	0.0237	3.724	0.0002
β_{edu}	0.0882	0.0353	1.683	0.0927
β_{edu^2}	-0.0017	0.0008	-1.964	0.0498

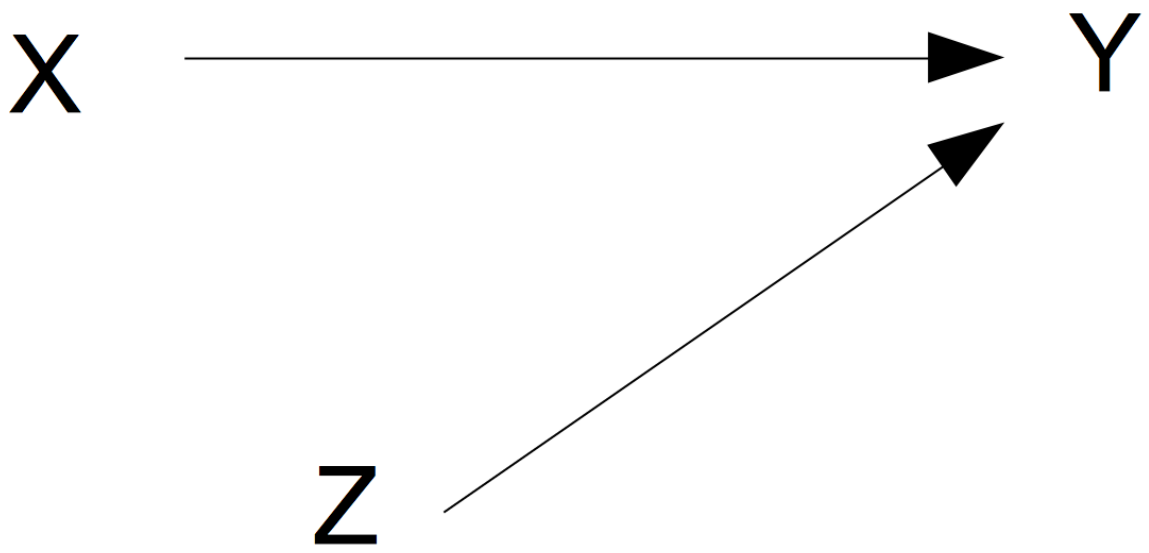
$R^2 = 0.1006$, $R^2 - \text{corretto} = 0.09627$

errore standard di regressione $= 0.5089$

ancora sulle variabili omesse

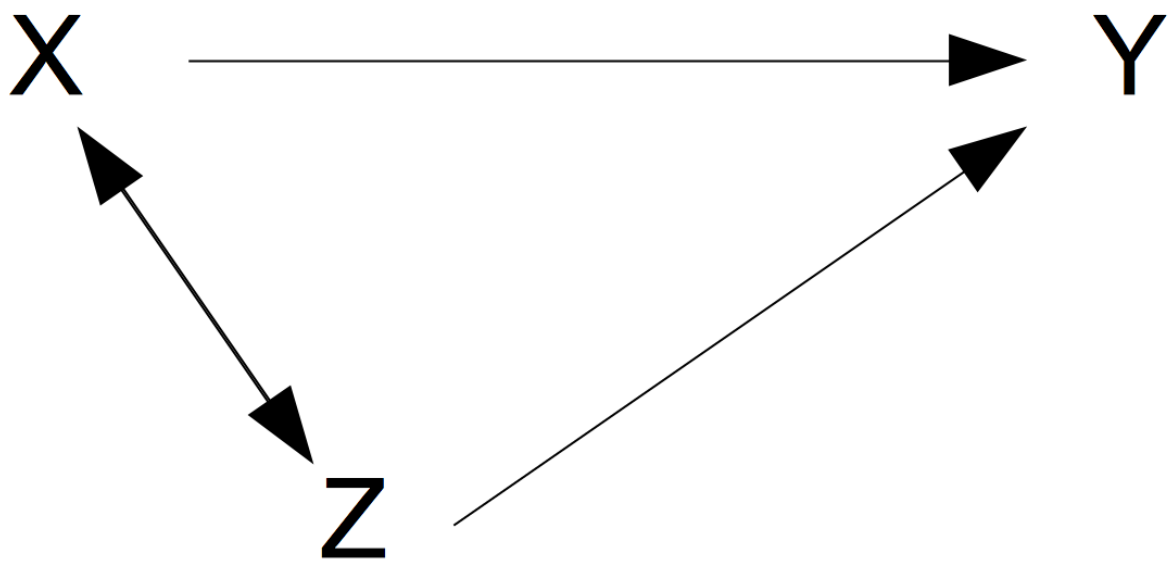
- quando il coefficiente di un regressore si modifica a seguito dell'introduzione di una variabile aggiuntiva che spiega la variabilità di Y
- la variabile aggiunta era una variabile omessa nel modello iniziale
- sappiamo che se ci sono variabili omesse i coefficienti sono distorti

variabili omesse



La variabile Z si dice omessa quando è una delle variabili che determina Y

variabili omesse

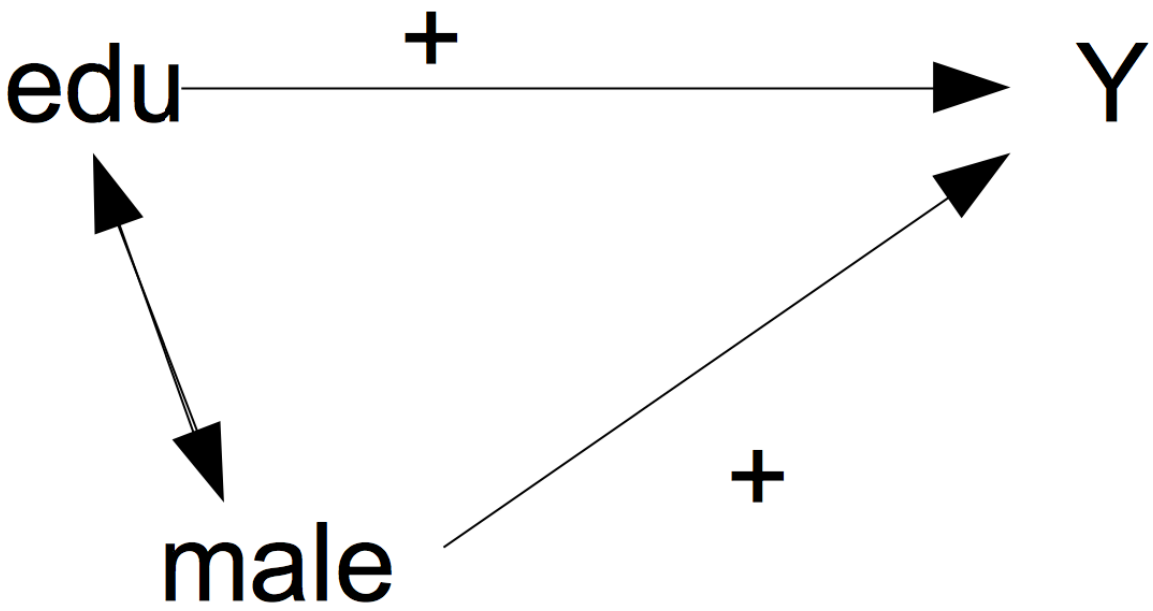


E quando è correlata con la variabile per la quale si sta cercando di identificare il coefficiente (la X)

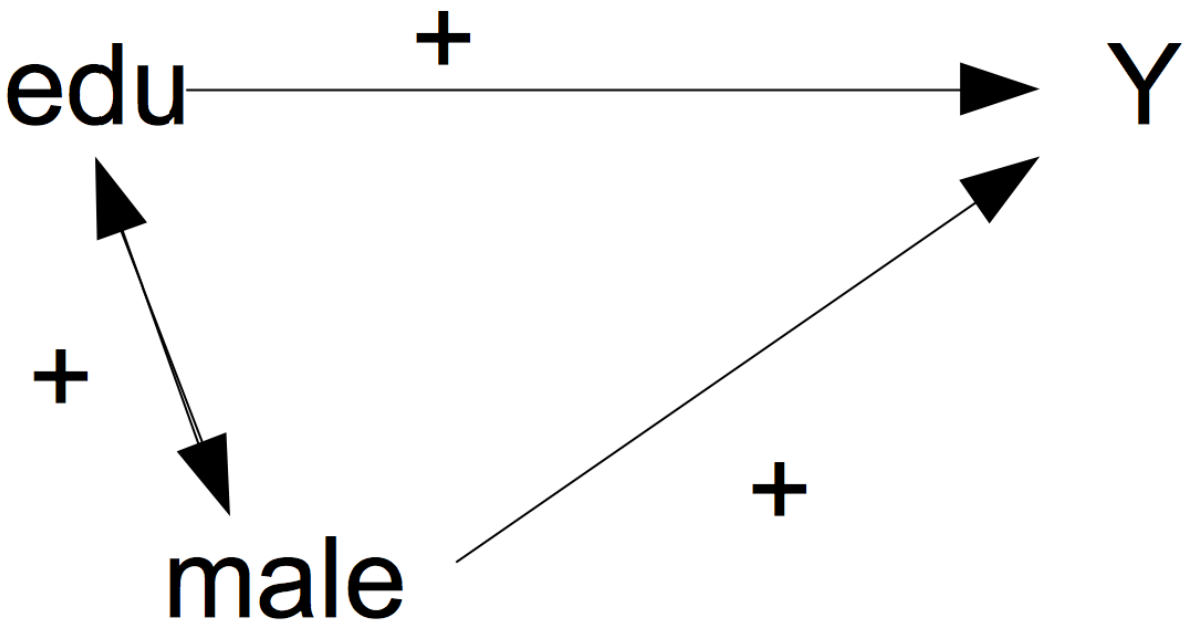
ancora sulle variabili omesse

- il segno della distorsione dipende dalla correlazioni X, Y; Z, Y e X, Z
- bisogna chiedersi come, la variabilità della Y spiegata dalla variabile omessa sarà in parte catturata dal coefficiente della X
- nel caso preso in considerazione tutte le correlazioni sono positive, quindi il coefficiente cattura un effetto positivo (risultato di effetto positivo di Z su Y e della correlazione positiva X, Y).
- quindi ad un coefficiente positivo si somma una distorsione positiva, il coefficiente risulta distorto verso l'alto

variabili omesse



variabili omesse



distorsione dovuta a variabili omesse

	coefficiente	errore standard	t	$p - value$
β_0	23615.96	1551.33	15.22	0.0000
β_{male}	1652.06	916.91	1.802	0.0719
β_{age}	54.47	29.54	1.844	0.0655

Cosa ci possiamo aspettare che avvenga al coefficiente β_{male} quando introduciamo la variabile omessa *edu*?

distorsione dovuta a variabili omesse

	coefficiente	errore standard	t	$p-value$
β_0	4886.97	2485.35	1.966	0.0496
β_{male}	1575.62	872.68	1.805	0.0714
β_{age}	133.09	29.34	4.536	0.0000
β_{edu}	1159.08	123.73	9.368	0.0000

la distorsione è quella attesa: verso l'alto

come interpretate la variazione del coefficiente dell'età?

I determinanti del voto di maturità

- introduciamo un altro dataset
- vogliamo studiare i determinati del voto di maturità
- disponiamo di un campione raccolto dall'Istat ogni 3 anni
- variabile dipendente: voto
- regressori: molti... occorre scegliere

diplomati Istat

	id	votomat	tscuola	privata	donna	votmed	area_istat	edu_padre	informatica	peso	uni
	1	77	3	0	1	suff	4	3	0	4	1
	2	72	2	0	0	buono	4	1	0	6	0
	3	85	3	0	1	suff	3	4	1	5	1
	4	65	3	0	0	suff	1	3	1	3	0
	5	81	2	0	1	suff	3	2	1	5	0
	6	84	3	0	0	buono	4	3	0	56	1
	7	84	3	1	0	buono	1	4	1	3	1
	8	69	2	0	0	suff	2	2	0	23	0
	9	68	2	0	0	suff	5	1	0	6	0
	10	83	3	0	1	suff	3	2	1	16	1
	11	60	2	1	0	suff	1	3	1	25	0
	12	81	3	1	1	suff	1	1	0	16	0
	13	61	3	0	0	buono	1	2	1	23	0
	14	65	2	0	0	suff	5	1	0	3	1
	15	84	3	1	0	buono	1	3	1	24	0
	16	64	3	1	0	buono	1	4	0	3	1
	17	66	1	0	1	suff	3	1	1	5	0
	18	60	1	0	0	suff	4	1	0	3	1
	19	61	3	1	0	suff	1	4	0	3	1
	20	64	2	1	0	buono	5	2	1	62	0
	21	60	2	0	0	buono	1	2	0	8	0
	22	82	3	0	1	ottimo	1	2	0	54	1
	23	91	2	0	0	buono	4	3	1	18	0
	24	75	4	0	1	buono	5	2	1	6	0

Variabile indipendente qualitativa

- Spesso capita di avere variabili fra i regressori che descrivono un fenomeno qualitativo
- abitare nel nord Italia, essere donna, essere laureato
- questi fenomeni possono essere inseriti come regressori soltanto utilizzando variabili dicotomiche
- un modello che spiega un risultato scolastico:

$$Y = \beta_0 + \beta_1 Nord + \beta_2 Donna$$

- in questo caso tutti i β colgono l'effetto di variabili dicotomiche

Attenzione alle variabili qualitative

- Ma nelle indagini statistiche alcune variabili qualitative possono indurci in errore
- L'Istat ad esempio tipicamente divide i residenti in macreoaree:
  1=Nord-ovest; 2=Nord-est; 3=Centro; 4=Sud;
  5=Isole
- Se regrediamo il voto finale alla scuola secondaria superiore così:

$$Y = \beta_0 + \beta_1 Donna + \beta_2 Area\ Istat$$

- commettiamo un serio errore

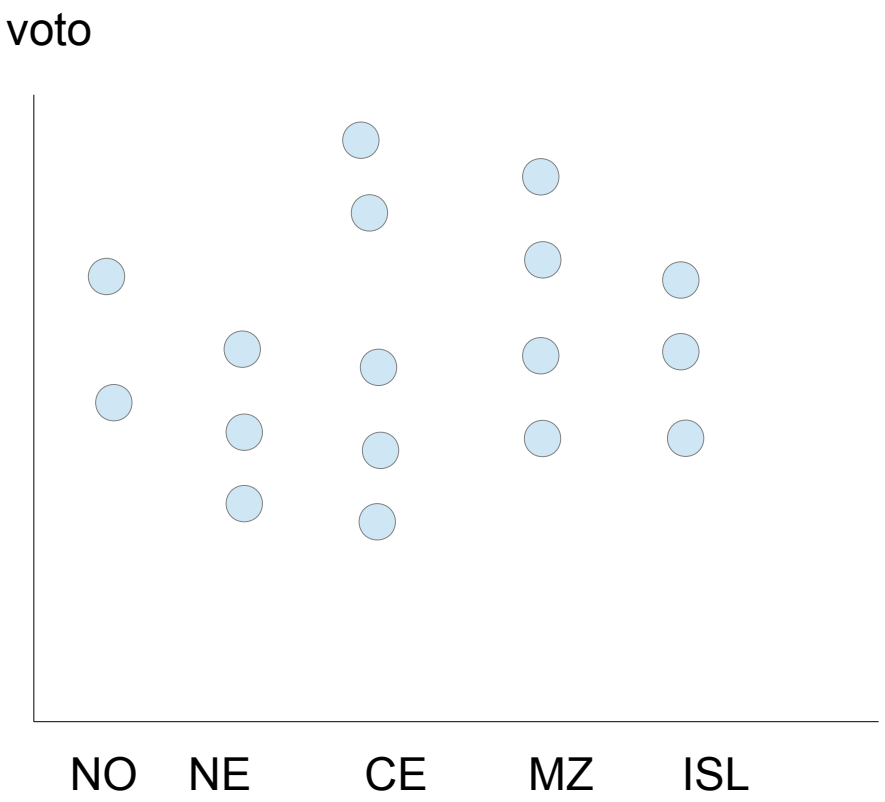
Attenzione alle variabili qualitative

- Stiamo dando un valore cardinale a una variabile qualitativa
- Immaginate che il risultato sia il seguente

	coefficiente	errore standard	t	$p-value$
β_0	72.488	11.0437	69.454	0.0000
β_{donna}	2.9786	0.8116	3.670	0.0002
$\beta_{arealstat}$	0.3409	0.3051	1.117	0.2642

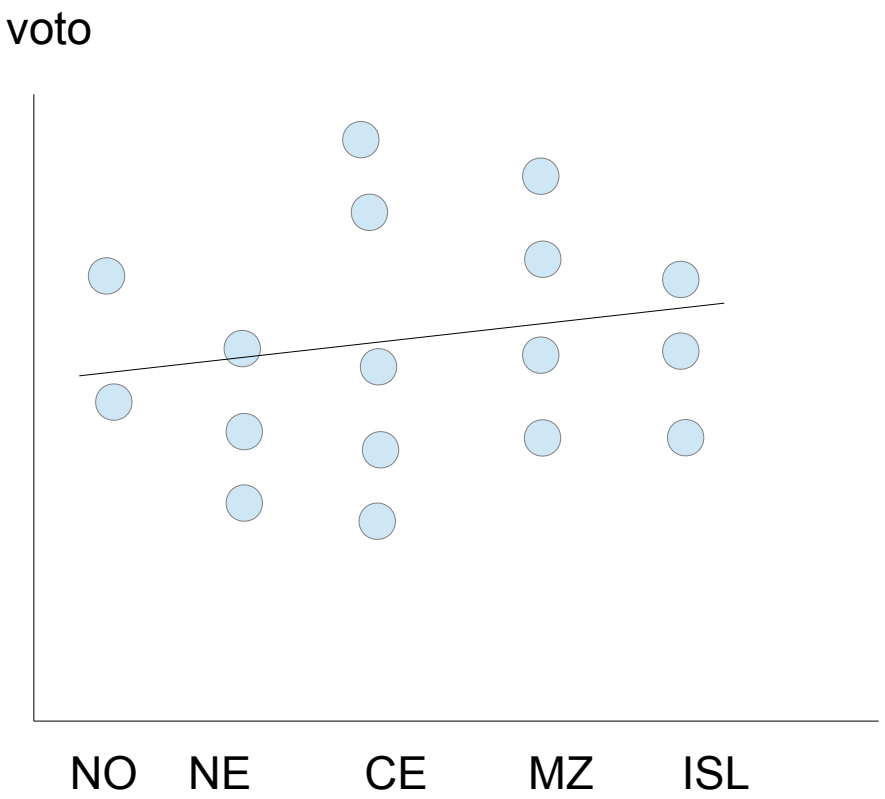
- come potete interpretare i coefficienti?

Area di residenza e voti



Area di residenza e voti

Utilizzando la variabile VOTOMAT come codificata dall'Istat:



Area di residenza e voti

Ma noi forse abbiamo in mente di identificare un effetto fisso di area geografica:

Ricodifica delle variabili qualitative

- per identificare l'effetto fisso dell'area geografica specifichiamo il modello:

$$Y = \beta_0 + \beta_1 Donna + \beta_2 Nord\text{-}Ovest + \beta_3 Nord\text{-}Est +$$

$$+\beta_4 Centro + \beta_5 Sud + \beta_6 Isole$$

- ognuna delle variabili è ottenuta creando una variabile dicotomica
- Sud ad esempio assume valore 1 quando Area Istat = 4 e assume valore 0 in ogni altro caso
- purtroppo anche questo modello è problematico. Perchè?

Variabili qualitative e multicollinearità perfetta

- non è possibile stimare il coefficiente per tutte le variabili dicotomiche

	coefficiente	errore standard	t	$p-value$
β_0	72.7407	0.9376	77.582	0.0000
β_{donna}	2.9491	0.8152	3.618	0.0003
β_{NE}	0.8470	1.3494	0.628	0.5303
β_C	0.7013	1.1625	0.603	0.5465
β_{SUD}	0.8867	1.2148	0.730	0.4656
β_{ISL}	1.7455	1.4242	1.226	0.2207

- come potete interpretare i coefficienti?
- perché il coefficiente del Nord-Ovest è assente?

Variabili qualitative e multicollinearità perfetta

- in alternativa si possono tenere tutte le variabili che descrivono le macroaree ma è necessario eliminare l'intercetta

	coefficiente	errore standard	t	$p-value$
β_{donna}	2.949	0.8152	3.618	0.0003
β_{NO}	72.7407	0.9376	77.582	0.0000
β_{NE}	73.5877	1.1128	66.128	0.0000
β_{C}	73.4420	0.857	85.668	0.0000
β_{SUD}	73.6274	0.9084	81.050	0.0000
β_{ISL}	74.4862	1.1868	62.760	0.0000

- in questo caso i coefficienti delle macroaree sono l'intercetta di ogni macroarea
- mentre prima si interpretavano come distanza dell'intercetta dalla macroarea di base (NO)

Regressori ordinali

- per l'aerea di residenza non sapevamo esattamente cosa attenderci in quanto a voti
- altre variabili invece identificano variabili qualitative che identificano fenomeni intrinsecamente ordinali
- l'esempio nel dataset è l'istruzione del padre (variabile edu_padre) che assume valori: 1=elementare, 2=media inferiore, 3=media superiore, 4=universitaria

Regressori ordinali

- se si specifica il modello

$$Y = \beta_0 + \beta_1 Donna + \beta_2 Nord\text{-}Ovest + \beta_3 Nord\text{-}Est +$$

$$+\beta_4 Centro + \beta_5 Sud + \beta_6 Isole + \beta_7 edu\_padre$$

- questo implica imporre che la differenza di voto attesa fra avere un padre con licenza media invece che con la licenza elementare è la stessa differenza di voto attesa fra avere un padre con laurea invece che con licenza media superiore
- anche in questo caso è prudente stimare una variabile dicotomica per ogni titolo di studio

Determinanti del voto di laurea

- stessa cosa vale per la variabile *VOTOMED* che assume valori da 'sufficiente' a 'ottimo'
- aggiungiamo anche il tipo di scuola frequentata (liceo, istituto tecnico,...) e se si tratta di una scuola privata o pubblica
- il nostro output di regressione diventa lungo per cui mostriamo solo i coefficienti statisticamente significativi

Determinanti del voto di laurea

	coefficiente	errore standard	t	$p-value$
β_0	73.2902	1.5768	46.481	0.0000
β_{donna}	1.4040	0.6785	2.069	0.0387
β_{ISL}	1.7942	1.0649	1.685	0.0923
$\beta_{DISTINTO}$	4.1482	0.9575	4.332	0.0000
β_{OTTIMO}	7.148	1.283	5.569	0.0000
β_{SUFF}	-2.405	0.7351	-3.272	0.0011
$\beta_{EDUP=SECINF}$	-1.8663	0.9744	-1.915	0.0557
$\beta_{PRIVATA}$	-2.1789	0.7015	-3.106	0.0019
β_{LICEO}	3.0605	1.2547	2.439	0.0148

$$\text{RSE}= 10.33,\ R^2 = 0.1251,\ Adj - R^2 = 0.1132$$