Principi di Econometria

lezione 6

AA 2015-2016

Paolo Brunori

Dove siamo arrivati?

- la regressione lineare multipla ci permette di stimare l'effetto della variabile X sulla Y **tenendo ferme tutte le altre** variabili osservabili che hanno un impatto su Y
- le stime si ottengono in modo analogo alla regressione univariata, minimizzando la somma degli errori compiuti nell'interpolare i dati
- in questo caso non usiamo una retta ma un **(iper)piano di regressione** di k dimensioni, tante quanti sono le variabili di controllo X
- ogni volta che aggiungiamo un regressore spieghiamo un po' della variabilità di Y
- per questo usiamo $R^2 - corretto$ invece che l'R^2

Regressione multipla perfetta: $n = k$

prezzo	rosso	grado alcolico
10,5	0	10
8,6	1	12

Regredire vuol dire risolvere un sistema in due incognite e due equazioni

$$10,5 = 10\beta_2 \rightarrow \beta_2 = \frac{10,5}{10} = 1,05$$

$$8,6 = \beta_1 + 12\beta_2 \rightarrow \beta_1 = 8,6 - 12 \times 1,05 = -4$$

Malgrado l'$R^2 = 1$ non ci fideremmo mai di questi coefficienti!

assunzioni necessarie per OLS

- le assunzioni sono le stesse già viste + 1
  1. $E(u_i|X_i) = 0 \ \forall i = 1, ..., k$
  2. $(X_1, ..., X_n)$ sono i.i.d
  3. gli outlier sono improbabili

$\Rightarrow$ assenza di collinearità perfetta

collninearità perfetta

- i regressori mostrano collinearità perfetta se uno dei regressori è funzione lineare degli altri
- se ad esempio $X_2 = a - bX_1$ o $X_2 = \frac{X_1}{100}$
- in presenza di collinearità perfetta non è possibile stimare la regressione OLS
- ma si tratta generalmente di un errore (materiale o logico)
- esempio: se X_1= età e X_2=anno di nascita, come posso stimare l'effetto della variazione di uno tenendo fermo l'altro?

casi frequenti di collinearità perfetta

- X_1= tasso di diplomati, X_2 tasso di drop-out o fuori-corso
- X_1= residenti al nord, X_2 residenti al centro, X_3 residenti al sud
- i software generalmente danno un messaggio di errore in questi casi o rimuovono una delle variabili collineari
- in casi come quello della regione di residenza è sufficiente eliminare uno dei regressori
- avremo coefficienti ad esempio per sud e centro e li interpretiamo per l'effetto di risiedere al sud e al centro rispetto ad abitare né al sud né al centro

nel caso del consumo di tabacco in Turchia

- immaginate di introdurre una variabile PIL ottenuta moltiplicando il reddito per 70milioni
- le due quantità PIL pro capite e PIL sono una funzione lineare dell'altra
- $PIL = 70.000.000 \times PIL_{PRO\ CAPITE}$
- non sarà quindi possibile calcolare i coefficienti
- ma in generale i software semplicemente elimineranno la variabile collineare ed effettueranno i conti come se non fosse parte del modello
- R ad esempio restituisce "NA" al posto del coefficiente

collinearità non perfetta

- se il legame fra una o più variabili è vicino ad un legame lineare la regressione OLS può essere stimata
- immaginate ad esempio di inserire una variabile 'stipendio medio' insieme alla variabile 'PIL pro capite'
- in questo caso la stima dei coefficienti è corretta
- ma per almeno uno di questi è imprecisa
- in effetti per stimare l'effetto della variazione di una variabile tenendo ferma l'altra si può sfruttare solo quella piccola parte della variabilità non identica per i due regressori

Regressione multipla del consumo di tabacco

	coefficiente	errore standard	t	$valore - p$
β_0	1.6572	0.1237	13.394	0.0000
β_Y	0.0003	0.0000	6.518	0.0000
β_P	-0 .4231	0.096	-4.3662	0.0001

- come posso interpretare questi coefficienti?

Regressione multipla del consumo di tabacco: PIL e reddito pro capite fortemente correlati

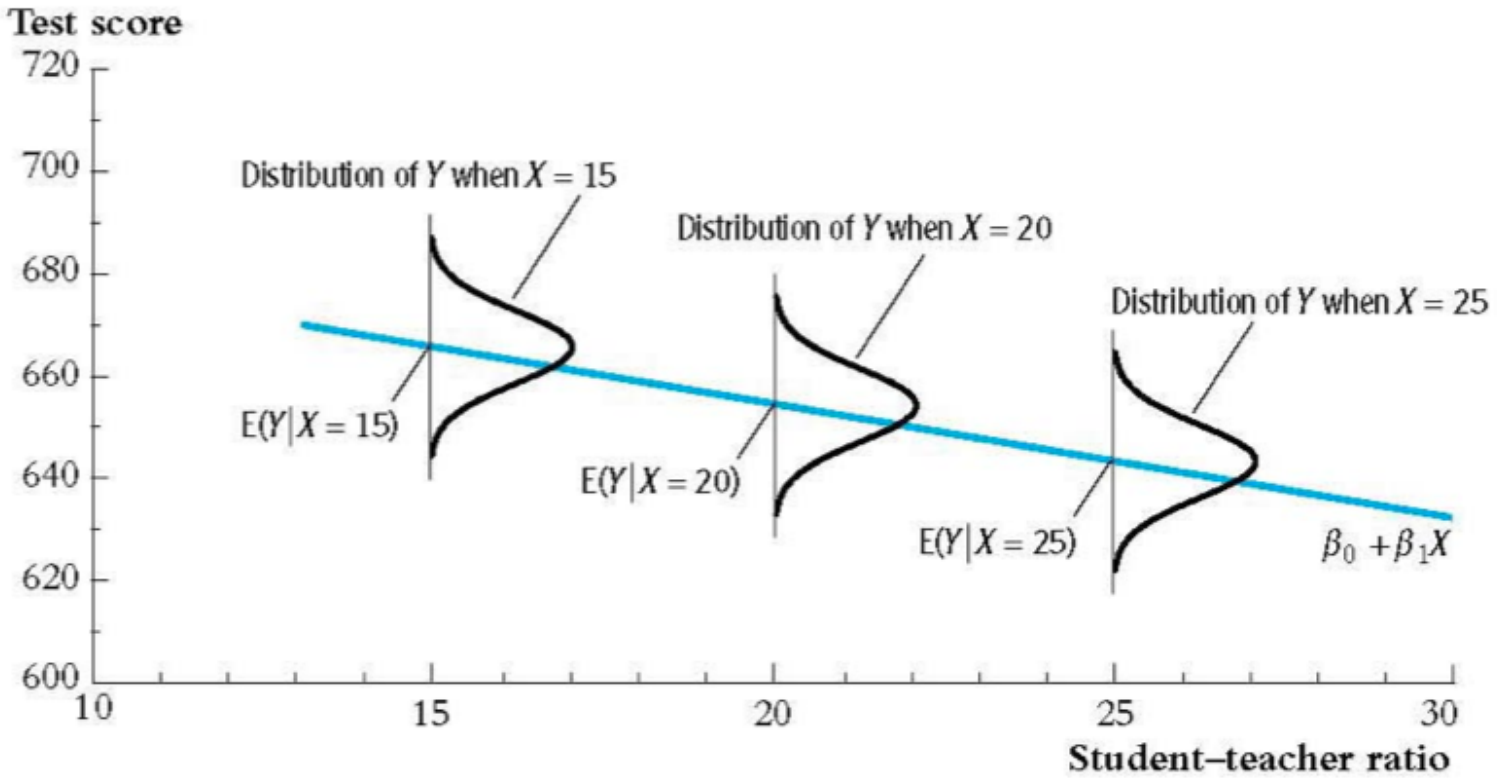
	coefficiente	errore standard	t	$valore - p$
β_0	1.5440	0.1339	11.53	0.0000
β_P	-0.4299	0.0930	-4.621	0.0001
β_Y	0.0003	0.0000	-1.043	0.3069
β_{PIL}	0.0008	0.0004	1.809	0.0825

- come posso interpretare questi coefficienti?

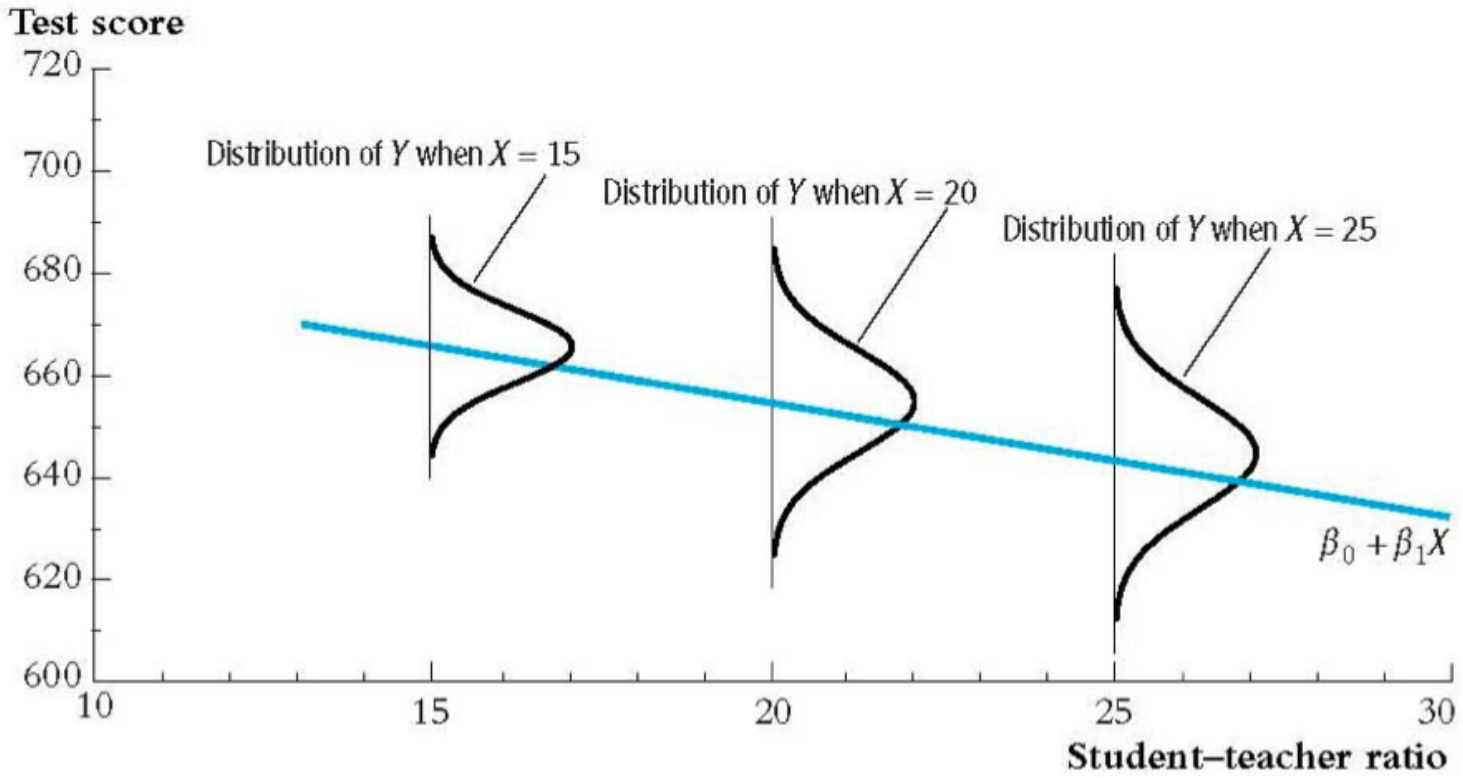
eteroschedasticità

- la prima assunzione per la validità degli OLS è che $E(u|X) = 0 \ \forall i = 1, .., n$
- se inoltre la varianza di u condizionata a X è costante allora gli errori sono **omoschedastici**

errori omoschedastici



errori eteroschedastici



efficienza dello stimatore OLS con omoschedasticità

- anche in caso di eteroschedasticità OLS è non distorto, consistente ed asintoticamente normale
- l'omoschedasticità aggiunge una proprietà: gli stimatori OLS sono stimatori lineari efficienti
- la prova è nel teorema di Gauss-Markov (appendice 5.2 del libro Stock & Watson)

Teorema di Gauss-Markov

- "lo stimatore OLS è il miglior stimatore lineare condizionatamente non distorto" (BLUE)
- uno stimatore lineare si può scrivere come:

$$\tilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i m$$

dove i pesi a_i possono dipendere da X_i ma non da Y_i

- la non distorsione condizionata implica:

$$E(\tilde{\beta}|X_1, ..., X_n) = \beta_1$$

- sotto le assunzioni dell'OLS (1-3) + l'omoschedacità + $u_i \sim N(0, \sigma_u^2)$: il teorema è valido

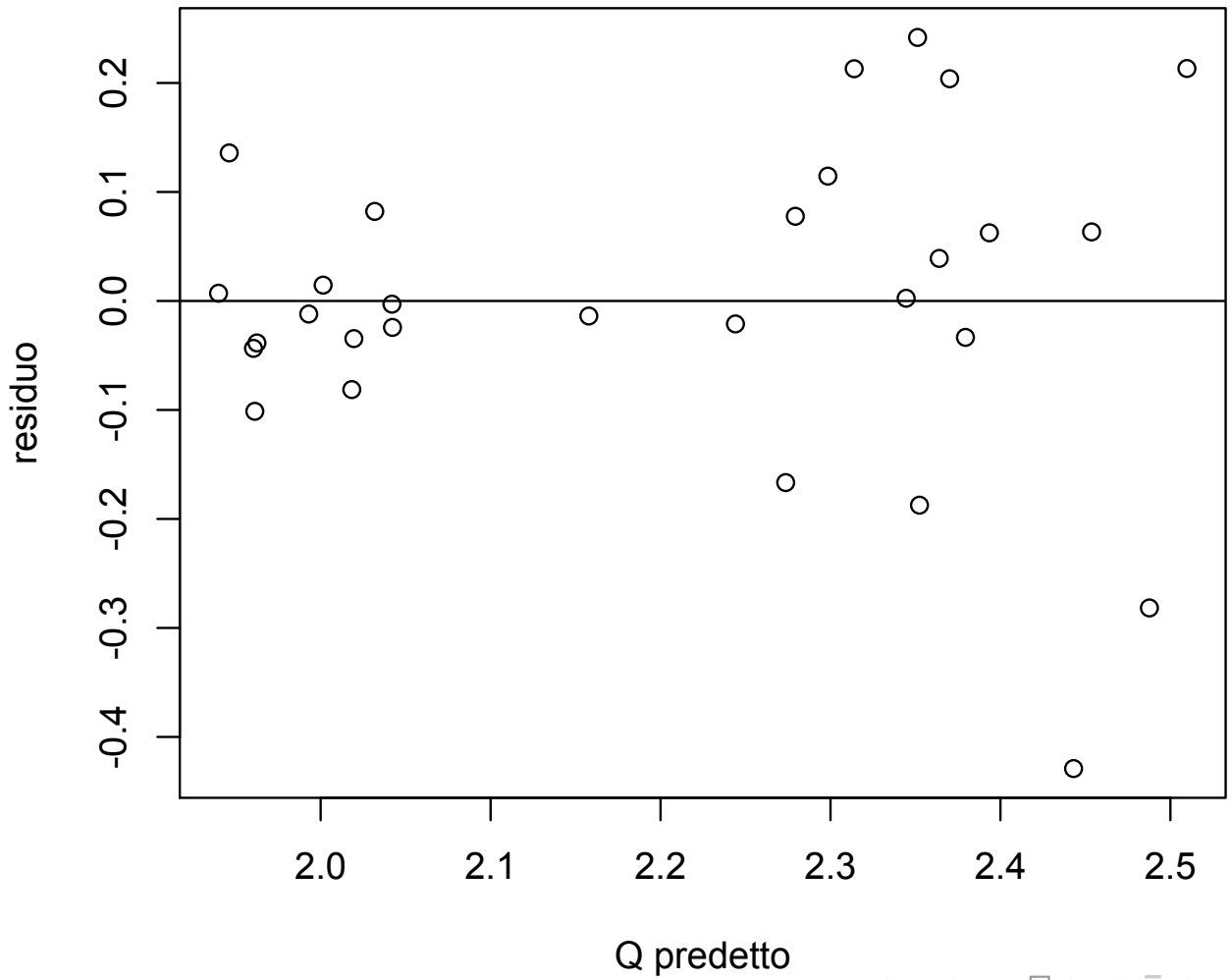
stimatore dei minimi quadrati ponderati

- se conosciamo l'eteroschedasticità (varianza di u condizionata a x)
- allora si può costruire uno stimatore più efficiente dell'OLS
- lo stimatore **dei minimi quadrati ponderati** utilizza un peso per ogni osservazione i:

$$w_i = \frac{1}{\sqrt{\sigma^2_{u|X_i}}}$$

- metodo elegante ma raramente si conosce $\sigma^2_{u|X_i}$
- gli errori standard robusti sono generalmente calcolati dai software (R non fa eccezione)

consumo di tabacco in Turchia: errori eteroschedastici



Regressione multipla del consumo di tabacco

	coefficiente	errore standard	t	$valore - p$
β_0	1.6572	0.1237	13.394	0.0000
β_Y	0.0003	0.0000	6.518	0.0000
β_P	-0 .4231	0.096	-4.3662	0.0001

- come posso interpretare questi coefficienti?

Regressione multipla robusta per eteroschedasticità

	coefficiente	errore standard	t	$valore - p$
β_0	1.6572	0.1746	9.4859	0.0000
β_Y	0.0003	0.0000	3.599	0.0014
β_P	-0 .4231	-0.24	-1.7494	0.09202

- come posso interpretare questi coefficienti?

Il modello completo

- sfruttiamo tutti i dati che abbiamo a disposizione
- oltre a reddito e prezzo abbiamo una variabile dicotomica (dummy) che identifica gli anni posteriori alla campagna di sensibilizzazione
- il coefficiente di una variabile dicotomica ha un'interpretazione diversa da una misura con significato cardinale
- l'aumento possibile della variabile dicotomica è solo uno $0 \rightarrow 1$
- il coefficiente sarà quindi l'effetto sulla variabile dipendente dell'avere la caratteristica indicata dalla variabile dummy
- in questo caso: cosa accade al consumo di tabacco se siamo in un anno successivo all'81

Regressione multipla del consumo di tabacco

	coefficiente	errore standard	t	$valore - p$
β_0	1.3283	0.1237	10.445	0.0000
β_Y	0.0003	0.0000	7.229	0.0000
β_P	-0 .2862	0.1482	-1.9309	0.0649
β_{1982}	-0.3233	0.1156	-2.7958	0.0098

- $R^2 = 0.8061$, R^2-corretto$= 0.7828$