Principi di Econometria

lezione 12

AA 2015-2016

Paolo Brunori

5 minacce alla validità interna

1. variabili omesse
2. incorretta specificazione della forma funzionale
3. misura imprecisa delle variabili
4. selezione del campione
5. causalità simultanea

errore di misura nel regressore

- la variabile esplicativa X è misurata imprecisamente $\tilde{X}_i = X_i + w_i$
- dove w_i è l'errore di misurazione
- la regressione stimata diventa:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + [\beta_1 (X_i - \tilde{X}_i) + u_i]$$

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + v_i$$

- il termine errore v_i è una funzione dell'errore di misura

errore di misura nel regressore classico

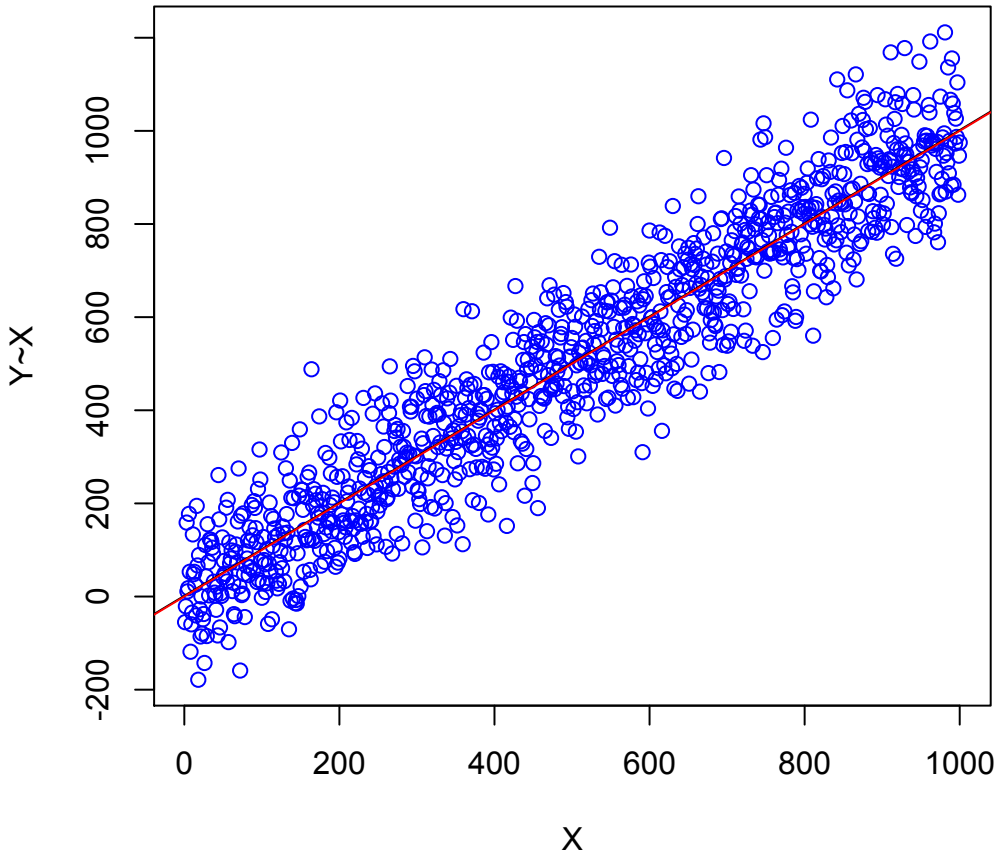
- l'errrore di misura si dice 'classico' nel caso in cui:

$$\tilde{X}_i = X_i + w_i$$

con $corr(w_i, X_i) = 0$ e $corr(w_i, u_i) = 0$

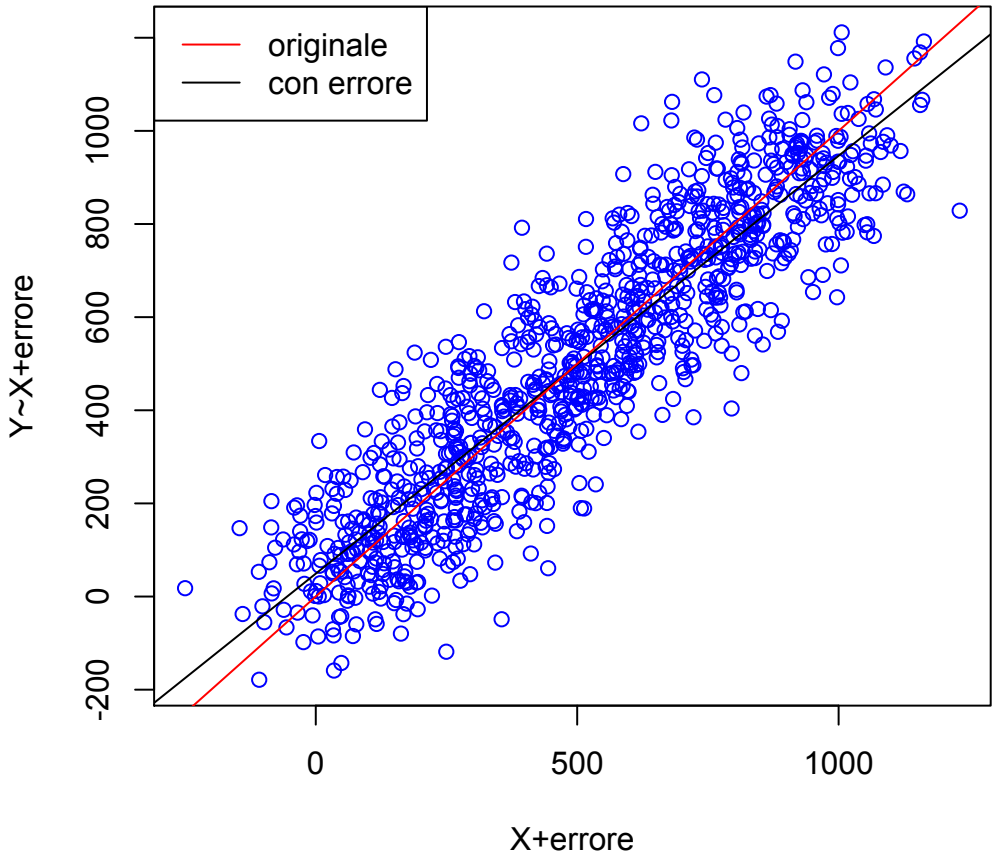
- si tratta di un errore di misurazione che non è correlato né con il regressore né con l'errore

Relazione fra X e Y senza errore



NB: nelle slide mostrate a lezione erano erroneamente invertite l'asse delle X e quella delle Y!

Relazione fra X e Y con errore



Errore di misura nel regressore classico

- in questo caso $\hat{\beta}_1$ è inconsistente:

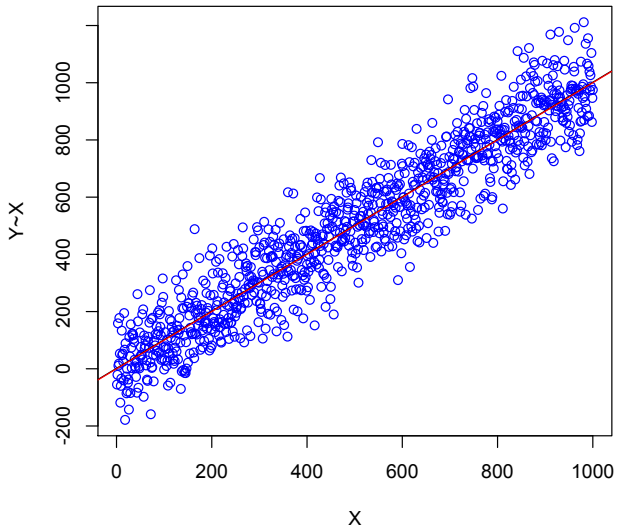
$$\hat{\beta}_1 \to \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\beta_1$$

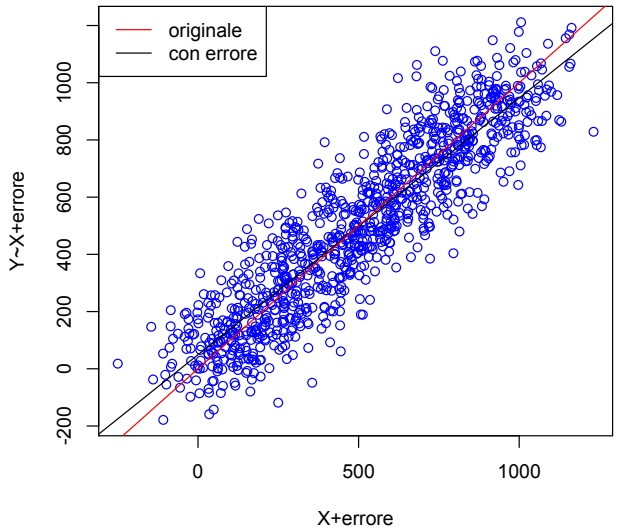
- β_1 è sempre sottostimato e l'entità di questa distorsione dipende dalla varianza relativa di w rispetto a X
- tanto maggiore la variabilità di w rispetto a quella di X tanto maggiore la distorsione

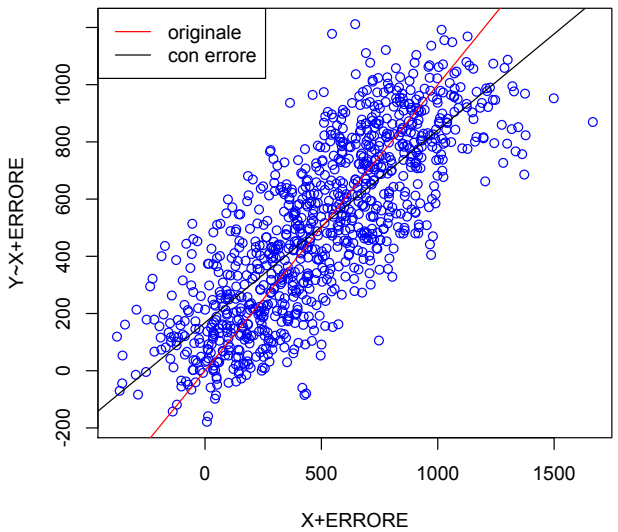
i coefficienti stimati nei 2 casi

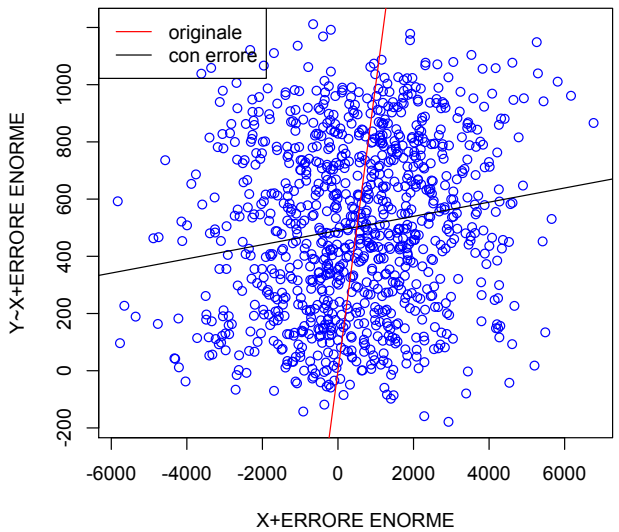
- senza errore: $\beta_0 = -5.406 \;\; \beta_1 = 1.007$
- con errore ($w_i \sim N(0, 100)$):
  $\beta_0 = 71.1574, \;\; \beta_1 = 0.8728$
- con errore grande ($w_i \sim N(0, 400)$):
  $\beta_0 = 317.909, \;\; \beta_1 = 0.364$
- con errore enorme ($w_i \sim N(0, 5000)$):
  $\beta_0 = 496.583, \;\; \beta_1 = 0.00393$

Intuizione: errori nella misura di X









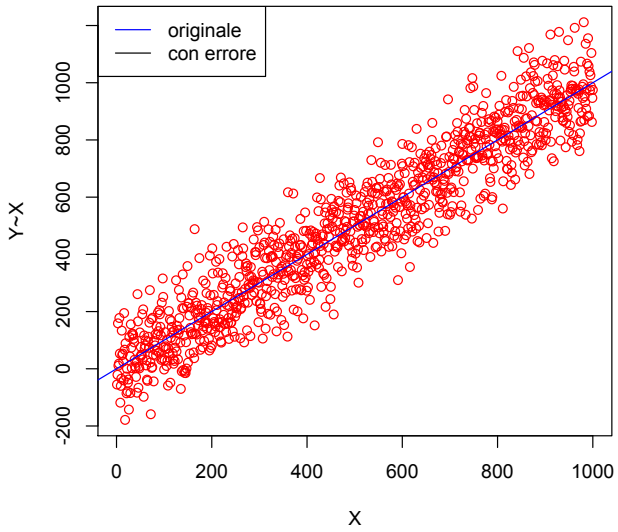
Errore di misura nel regressore non classico

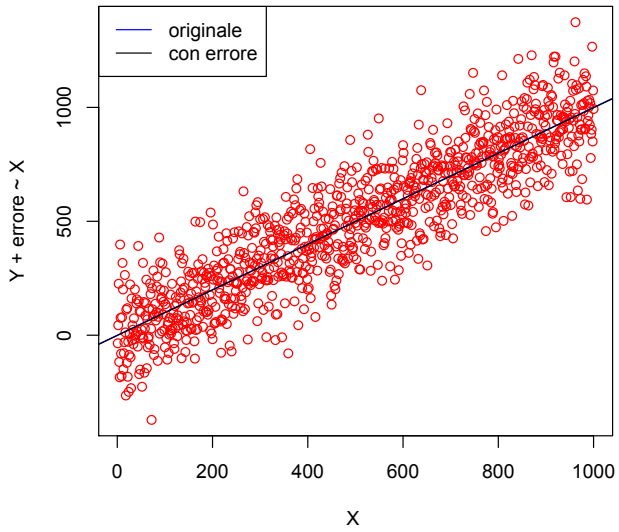
- anche in questo caso $\hat{\beta}_1$ è inconsistente
- il valore che mi aspetto di stimare non si avvicina a quello vero nemmeno avendo a disposizione campioni di dimensioni molto elevate
- β_1 è distorto, ma il segno di questa distorsione dipende dalla correlazione di w rispetto a X e u
- se l'errore di misurazione è correlato con l'errore siamo in una situazione simile a quella delle variabili omesse
- occorre applicare lo stesso ragionamento visto nella lezione 11 per capire quale sia il segno della distorsione

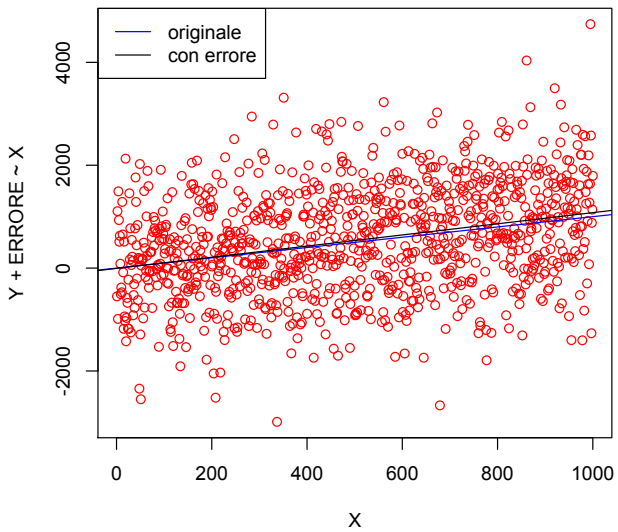
Errori di misura della variabile dipendente

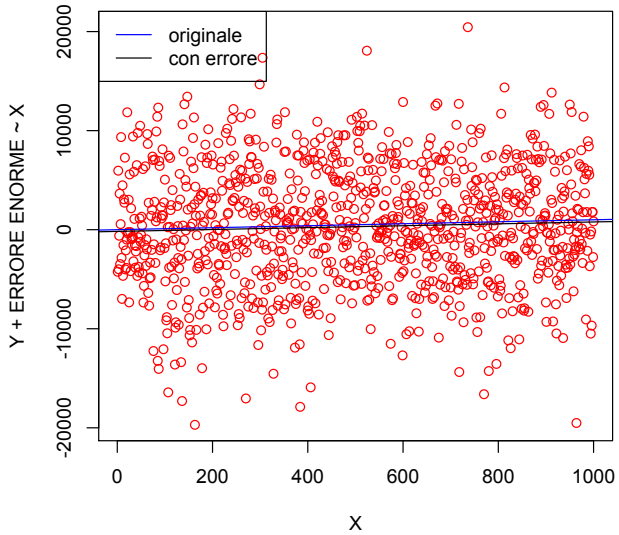
- se si tratta di errore classico $\tilde{Y}_i = Y_i + w_i$, con $corr(w_i, Y_i) = 0$ e $corr(w_i, u_i) = 0$
- la stima della relazione fra X e Y non è distorta ma più incerta
- la varianza di $\hat{\beta}_1$ sarà maggiore ma β_1 è consistente

Intuizione: errori nella misura di Y









Errori nella misura di Y

	coefficiente	errore standard	t	$p-value$
Y misurato correttamente				
β_0	1.562	6.279	0.249	0.804
β_1	1.0001	0.0108	92.01	0.0000
Y + ERRORE				
β_0	-21.46	188.17	-0.11	0.909
β_1	1.2528	0.3257	3.847	0.0001
Y + ERRORE ENORME				
β_0	-172.80	377.61	-0.458	0.647
β_1	0.9540	0.6536	1.460	0.145

Errori nella misura di Y

- ovviamente se l'errore di misura di Y non è classico ma correlato con X o u il coefficiente è distorto
- si tratta di una violazione delle assunzioni del modello OLS
- quando abbiamo il sospetto che la misura di una delle variabili che usiamo possa imprecisa dobbiamo sempre chiederci:
  1. quanto sarà rilevante l'errore rispetto alla variabilità osservata
  2. quale sarà il segno della eventuale distorsione

4. dati mancanti e selezione del campione

- il problema di fondo è lo stesso
- i dati mancanti non sono tutti uguali:
  - mancanti completamente a caso
  - mancanti in base a X
  - mancanti in base a X e Y

dati mancanti completamente a caso

- unico effetto: ridurre la dimensione del campione

$\rightarrow$ stime meno precise ma consistenti

dati mancanti in base a X

- effetto 1: riduzione della dimensione del campione
- effetto 2: riduzione della variabilità (o intervallo di variazione) di X

$\rightarrow$ stime meno precise ma consistenti

dati mancanti in base a X e Y

- effetto 1: riduzione della dimensione del campione
- effetto 2: riduzione della variabilità (o intervallo di variazione) di X
- effetto 3: possibile correlazione fra gli errori e i regressori (vince London)

$\rightarrow$ stime inconsistenti

5. causalità simultanea

- perché le classi sono formate da circa 30 studenti?
- cosa si valuta per determinare questo numero?
- immaginate di voler valutare se un numero minore di studenti per insegnante favorisca l'apprendimento
- come procedete?

Causalità simultanea

- avendo dati riguardo alla numerosità delle classi e risultati scolastici si può cercare di isolare l'effetto della numerosità della classe
- ma è possibile che la dimensione delle classi sia in qualche modo decisa sulla base dei risultati?
- se così fosse saremmo in presenza di una situazione di causalità simultanea
- situazione molto complessa per riuscire ad ottenere stime affidabili

Causalità simultanea

- abbiamo sempre ipotizzato $X \rightarrow Y$
- in molti casi è vero anche che $Y \rightarrow X$
- in questo caso si parla di causalità simultanea
- la conseguenza è correlazione fra regressore ed errore:

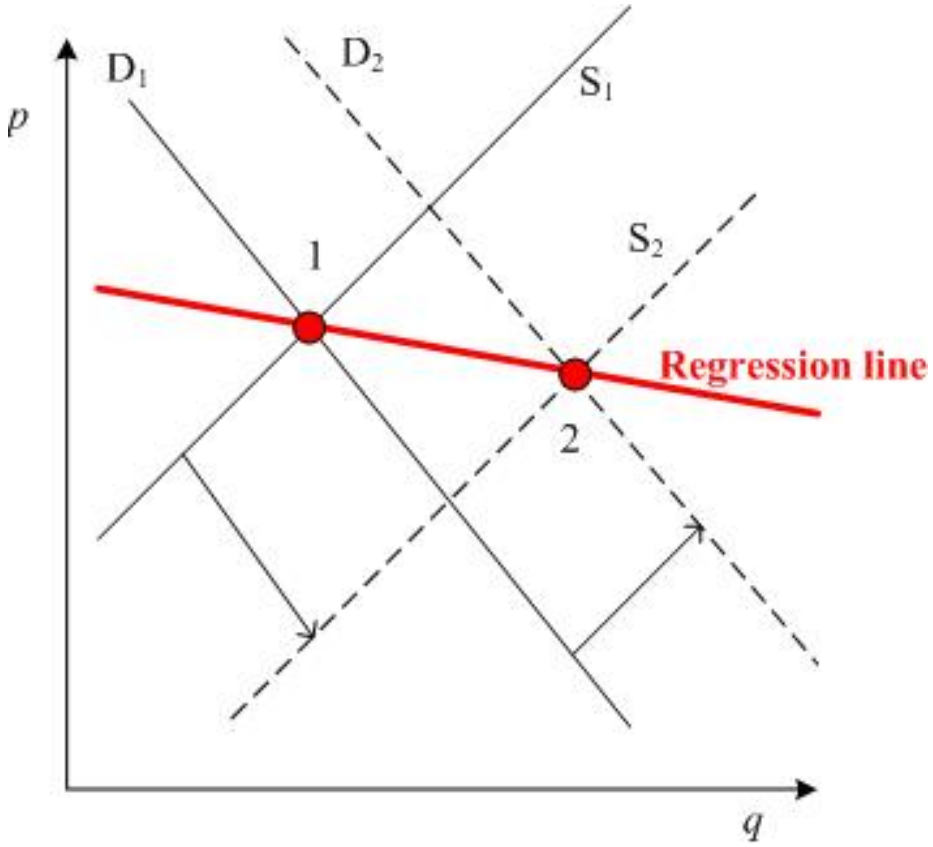
$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i$$

- esempio: $u_i < 0 \rightarrow Y_i < \hat{Y}_i$, ma Y_i piccolo influenza X_i attraverso γ_1
- le stime sono distorte e il problema non facilmente risolvibile

Endogeneità: il caso di prezzo e quantità di un bene

- $Q_D = a - bP$
- $Q_S = c + dP$



Soluzioni per la causalità simultanea

- le soluzioni non sono semplici
- è necessario osservare una variazione di X che avvenga indipendentemente da Y
- come avviene ad esempio negli esperimenti controllati in laboratorio per altre scienze
- ci sono casi fortunati in cui questo avviene, si tratta di situazioni quasi-sperimentali o sperimentali che studieremo in una delle ultime lezioin
- in alternativa ci sono tecniche statistiche che sfruttano la covarianza della variabile X con altre variabili non correlate con Y