Principi di Econometria

lezione 13

AA 2016-2017

Paolo Brunori

popolazione studiata e popolazione di interesse

- popolazione studiata: popolazione da cui è stato estratto il campione
- popolazione di interesse: popolazione per la quale ci interessa che siano valide le nostre conclusioni
- le rivelazioni sull'orientamento di voto servono per prevedere il risultato elettorale ma sono rivolte a poche centinaia di intervistati

validità interna ed esterna

- validità interna: le inferenze statistiche sugli effetti casuali sono validi per la popolazione studiata
- validità esterna: le inferenze statistiche sugli effetti casuali possono essere generalizzate alla popolazione di interesse

validità interna

- β_i devono essere corretti
- $SE(\hat{\beta}_i)$ devono essere stimati precisamente
- questi requisiti vengono meno quando c'è violazione di una o piú assunzioni OLS:
  1. $E(u_i|X_i) = 0 \; \forall X_i$
  2. X, Y sono i.i.d.
  3. gli outlier sono improbabili
  4. non vi è collinearità perfetta

minacce alla validità esterna: differenze nelle popolazioni

- cavie animali
- immigrati in periodi diversi
- a chi possiamo generalizzare i risultati sui redditi svedesi o sulle sigarette turche?
- più simili le popolazioni più facilmente generalizzabili le stime della regressione

minacce alla validità esterna: differenze di contesto

- differenze nelle istituzioni (esempio sistema fiscale)
- più simile l'ambiente più facilmente generalizzabili le stime

valutare la validità esterna di uno studio

- esistono studi simili? (redditi finalandesi? italiani?)
- risultati coerenti su popolazioni simili rafforzano la validità esterna di uno studio
- questi sono tutti problemi che dovrebbero essere pensati prima
- quando si valutano le stime ci si deve sempre chiedere in che misura il campione utilizzato assomiglia alla popolazione di interesse

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

1. variabili omesse

- emerge se escludiamo una variabile Z che determina Y ed è correlata con un regressore X_1 si ha $corr(u_i, X_{1i}) \neq 0$
- distorsione che persiste anche nei grandi campioni
- risolvibile introducendo Z nel modello
- o introducendo una variabile di controllo che non ha un effetto causale su Y ma risolve la distorsione perché correlata con Z

variabili di controllo

- immaginiamo ad esempio di voler stimare quanto spenderanno gli studenti per le loro vacanze
- costruiamo un modello in cui la spesa Y è spiegata da: anno di corso, sesso, città di provenienza
- c'è una variabile omessa: il reddito
- quindi avremo una distorsione dei coefficienti
- se però osserviamo variabili correlate con il reddito possiamo mitigare il problema
- se osserviamo il mezzo di trasporto usato per venire in università è probabile che si tratti di una variabile correlata con quella omessa
- il coefficiente non è interpretabile: il mezzo di trasporto non è la causa della spesa per le vacanze
- la variabile di controllo però risolve o attenua il problema perché cattura buona parte della variabilità che sarebbe catturata dal reddito

inclusione di variabili di controllo

- questa inclusione va soppesata con la perdita di precisione delle stime
- dopo aver specificato la relazione di interesse la scelta si basa su 4 passaggi:
  a. fare un elenco delle fonti possibili di distorsione da variabili omesse
  b. inserire le variabili in elenco e testare l'ipotesi che abbiano coefficienti nulli
  c. controllare che i coefficienti iniziali non siano significativamente alterati

2. incorretta specifica della forma funzionale

- rende distorte le stime OLS
- in pratica si tratta dell'omissione di una variabile (trasformazione non lineare di una inserita)
- per cui introdurre la variabile omessa dovrebbe essere possibile

3. errori di misura nelle variabili

- qualsiasi indagine campionaria potrebbe contenere degli errori di misurazione
- nei regressori così come nella variabile dipendente

errore di misura nel regressore

- il reddito X è misurato imprecisamente $\tilde{X} \neq X$
- 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 'classico' compare nel caso in cui:

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

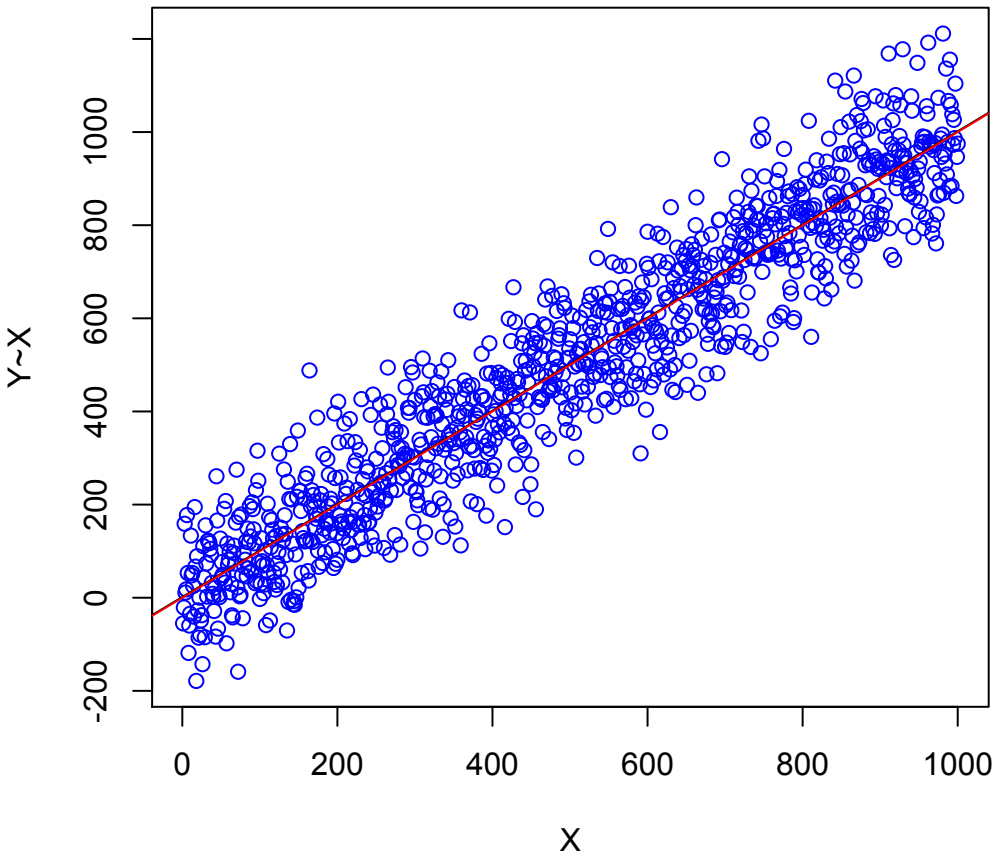
con $corr(w_i, X_i) = 0$ e $corr(w_i, u_i) = 0$

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

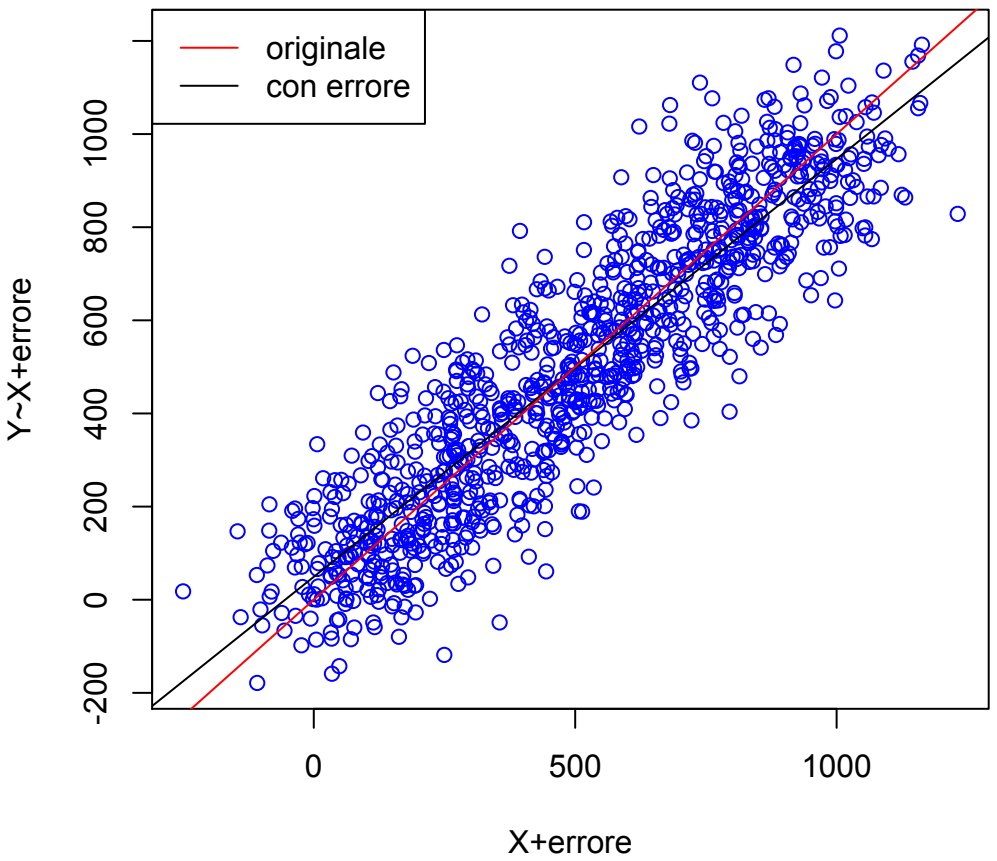
$$\hat{\beta}_1 \rightarrow \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1$$

- β_1 è sempre sottostimato e l'entità di questa distorsione dipende dalla varianza relativa di w rispetto a X
- conoscendo σ_X^2 e σ_w^2 si potrebbe calcolare il vero coefficiente β_1

Relazione fra X e Y senza errore



Relazione fra X e Y con errore



Errore di misura nel regressore classico

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

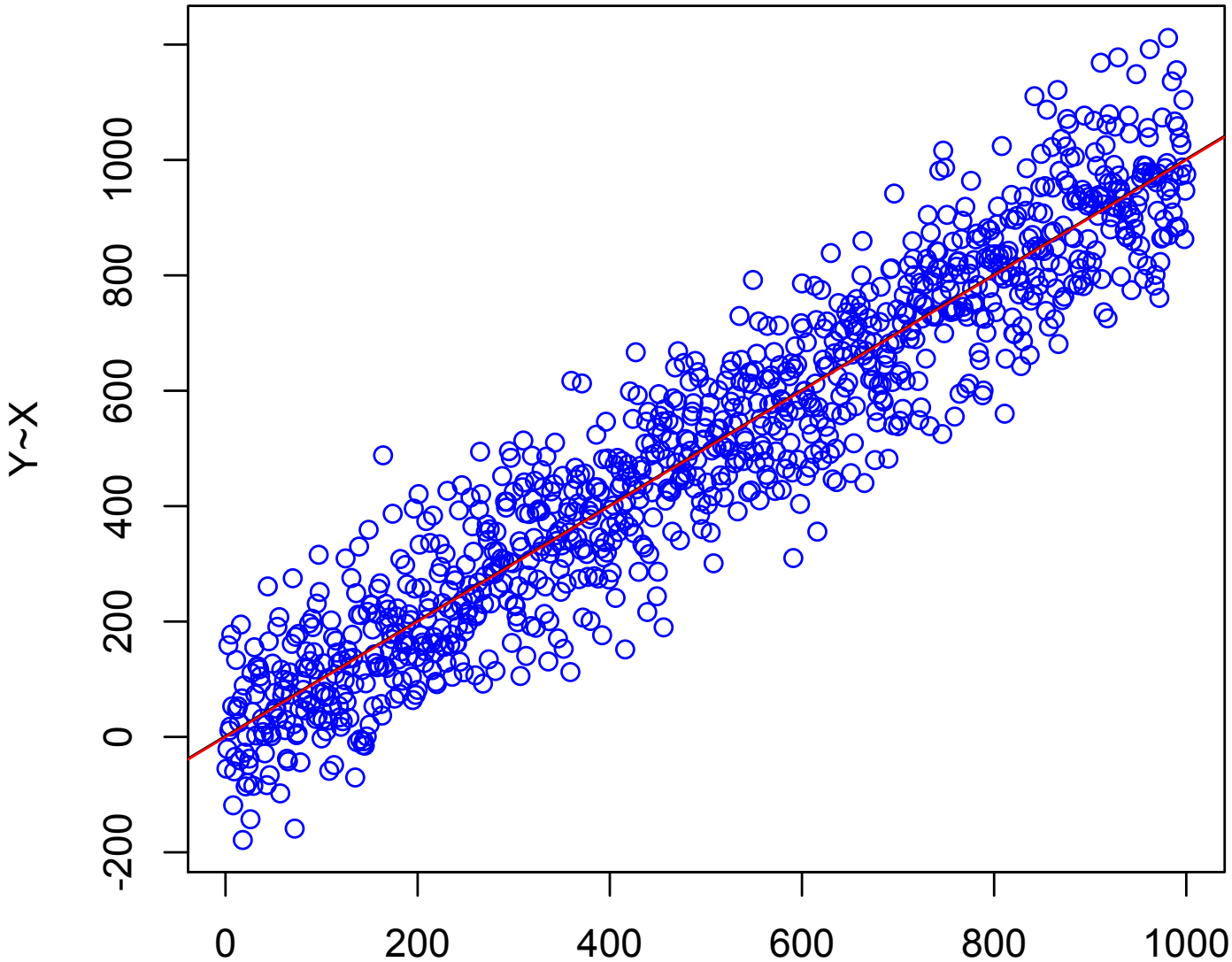
$$\hat{\beta}_1 \to \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}\beta_1$$

- 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

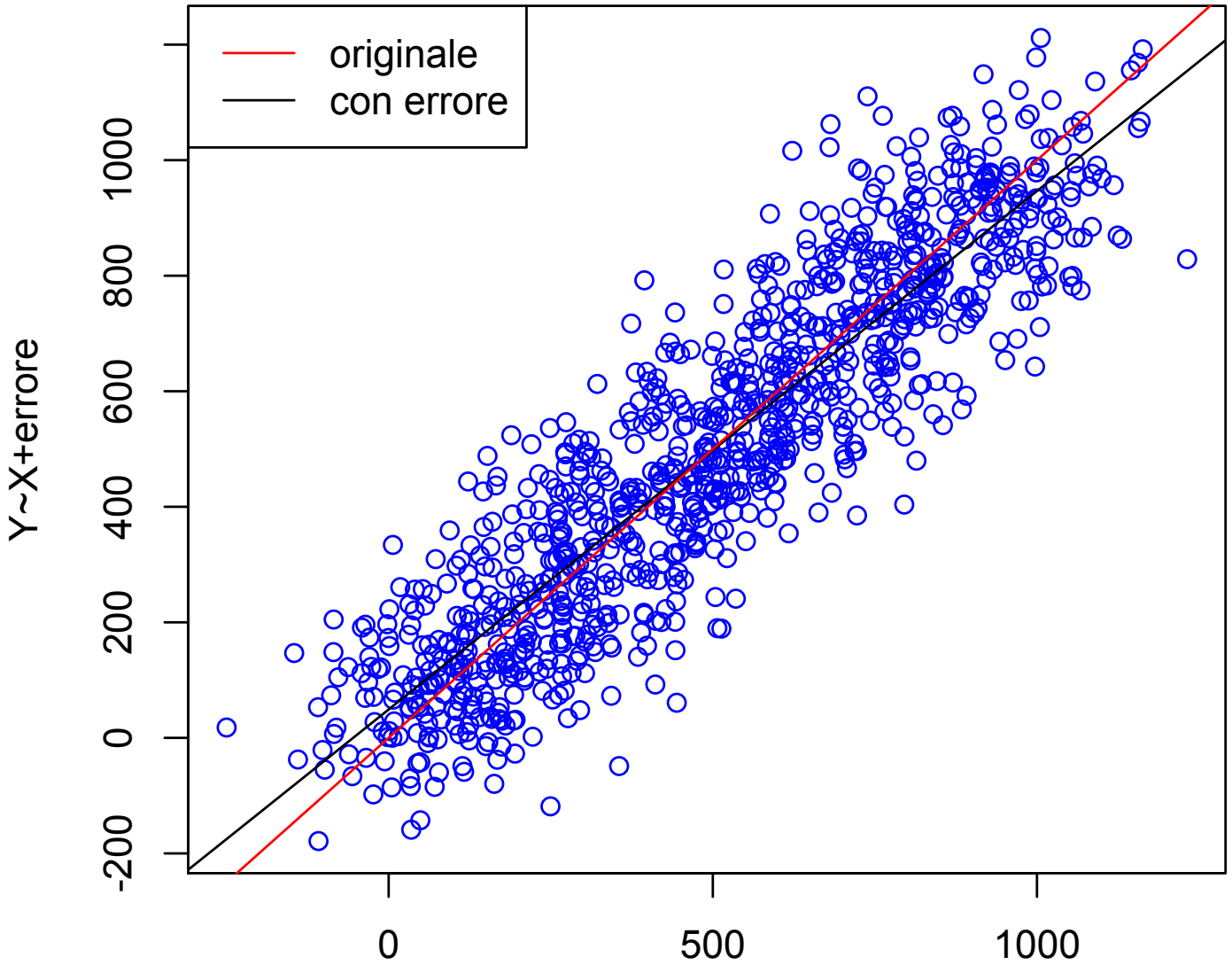
i coefficienti stimati nei 2 casi

- 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 ($w_i \sim N(0, 5000)$):
  $\beta_0 = 496.583, \;\; \beta_1 = 0.00393$

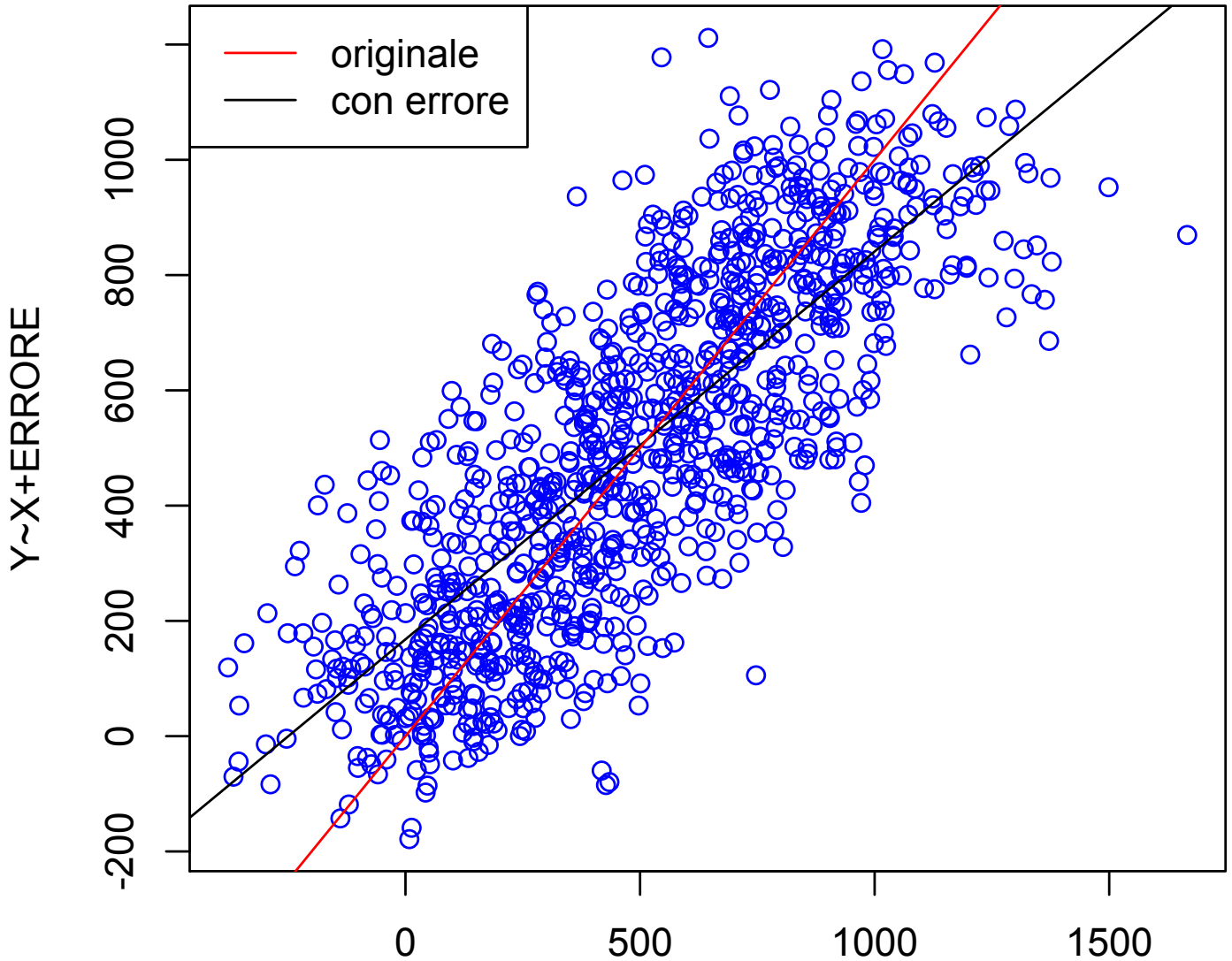
Intuizione: errori nella misura di X



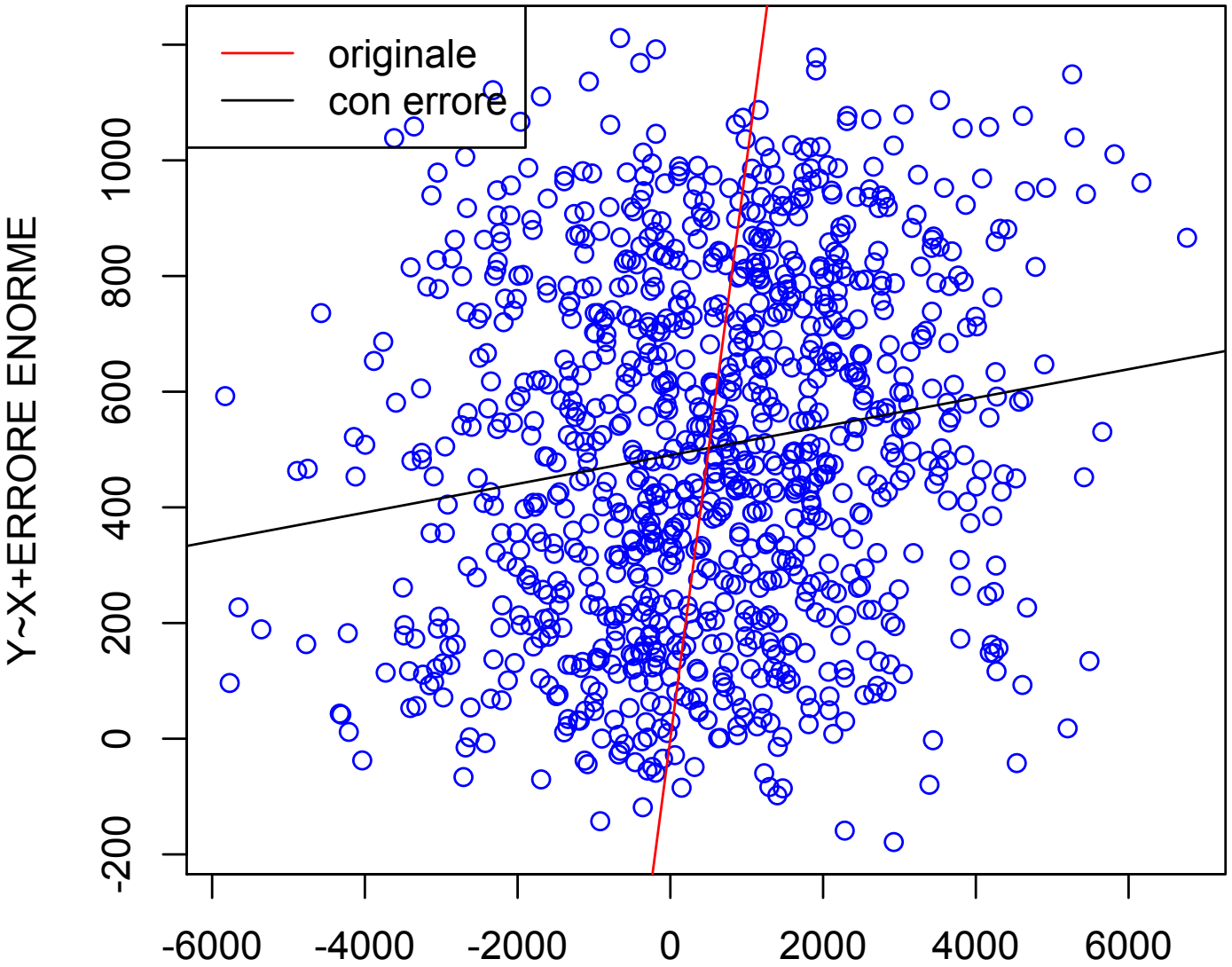
Intuizione: errori nella misura di X



Intuizione: errori nella misura di X



Intuizione: errori nella misura di X



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 ($w_i \sim N(0, 5000)$):
  $\beta_0 = 496.583, \;\; \beta_1 = 0.00393$

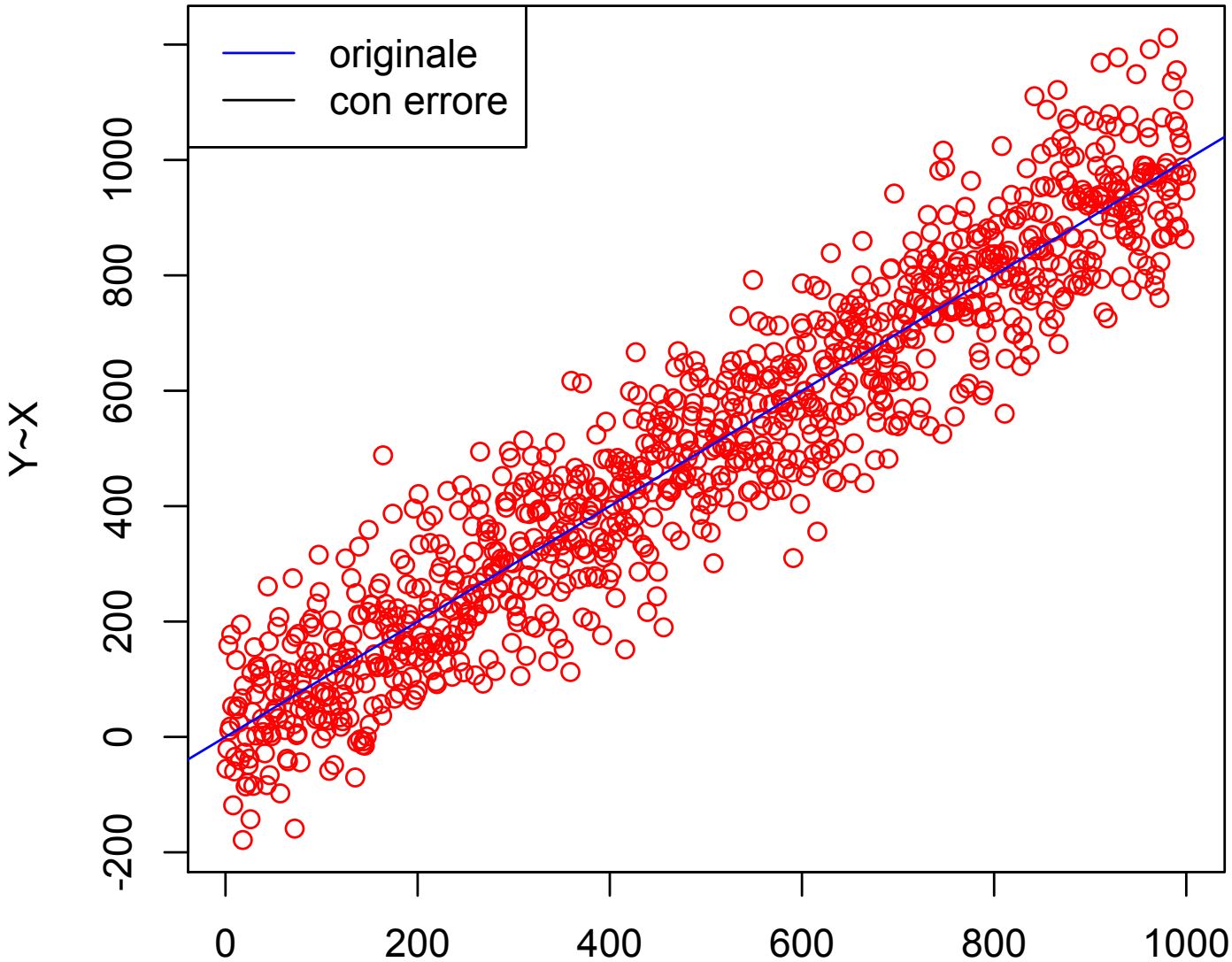
Errore di misura nel regressore non classico

- 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 per capire quale sia il segno della distorsione

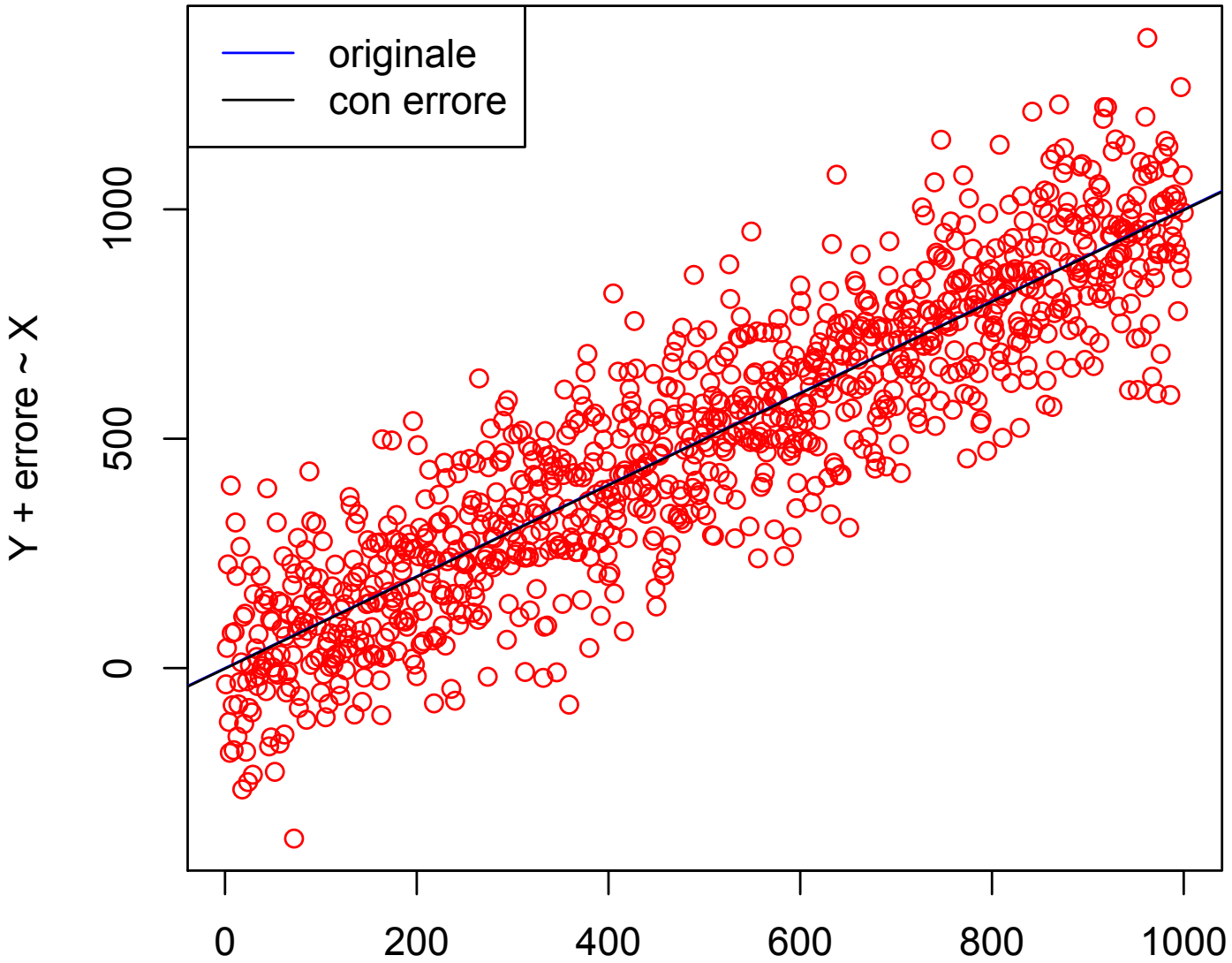
Errori di misura della variabile dipendente

- 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

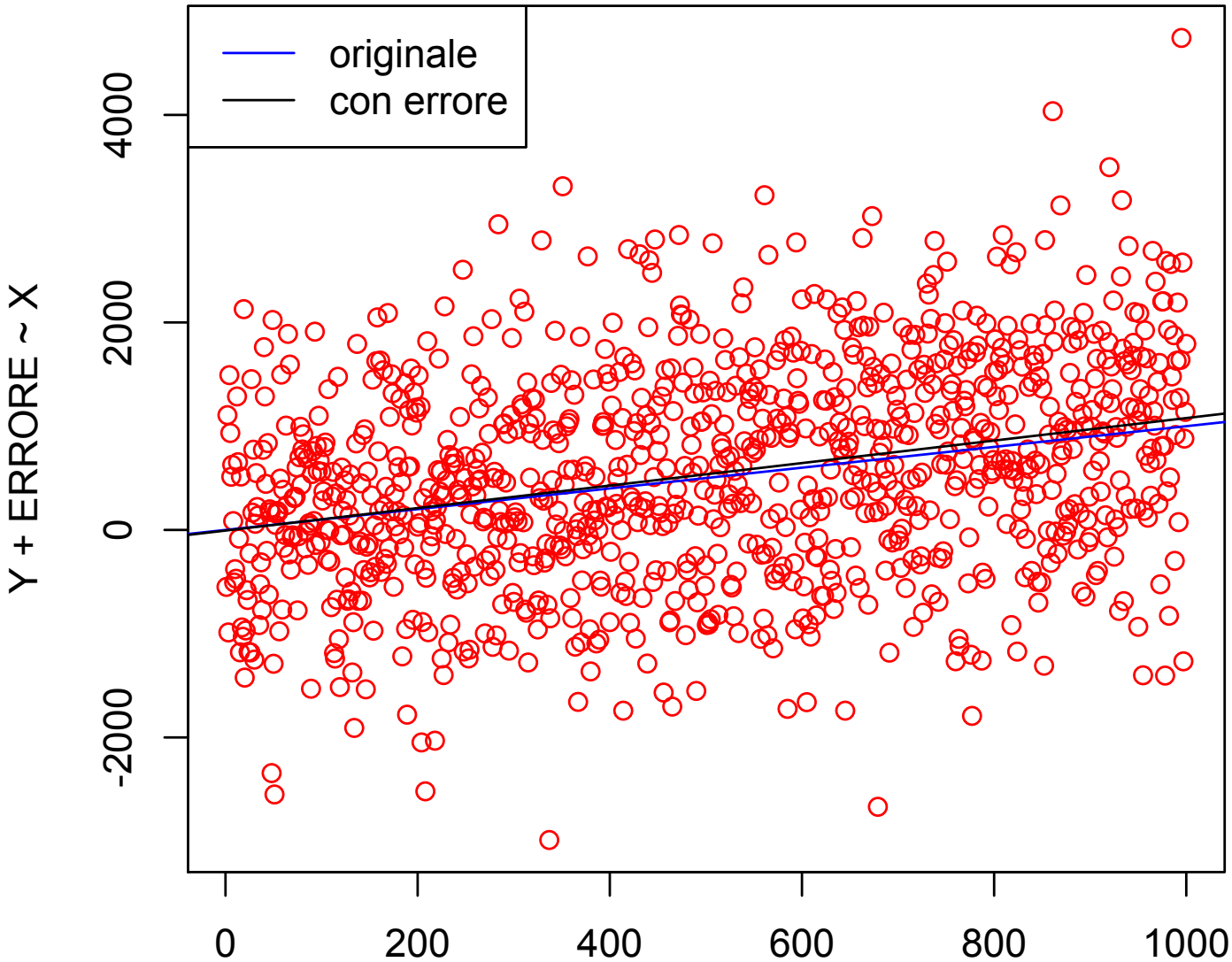
Intuizione: errori nella misura di Y



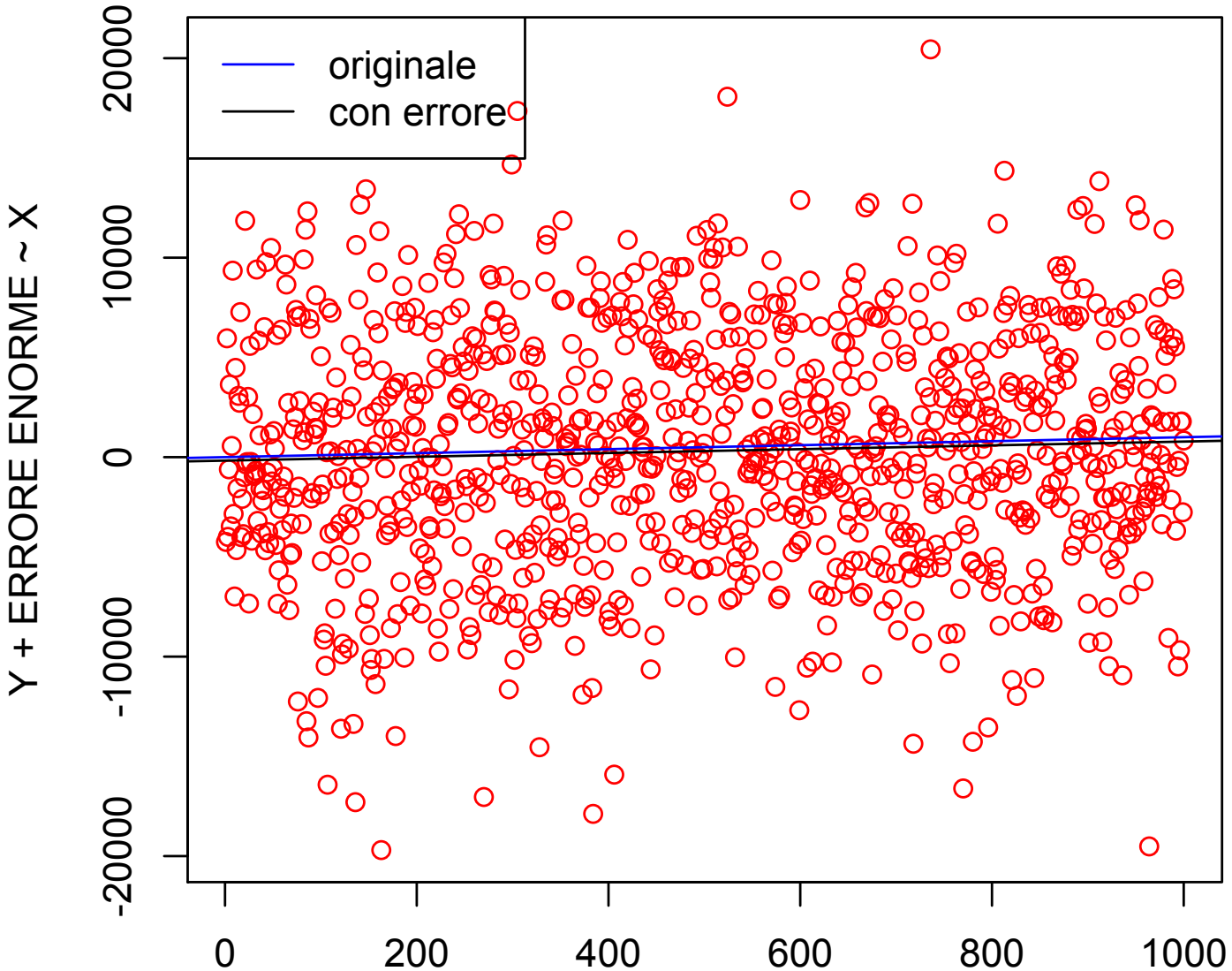
Intuizione: errori nella misura di Y



Intuizione: errori nella misura di Y



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

$\rightarrow$ stime inconsistenti