Principi di Econometria

lezione 3

AA 2015-2016

Paolo Brunori

dove eravamo arrivati

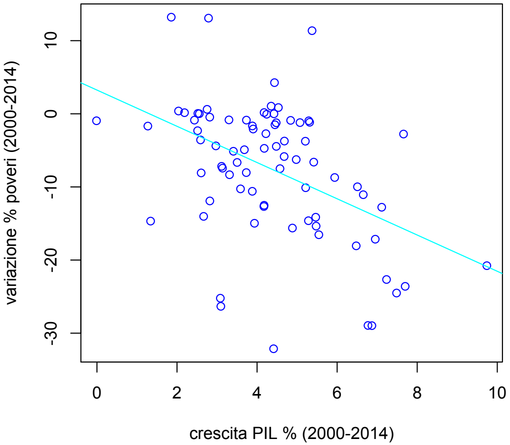
- se abbiamo dati per studiare una correlazione sistematica fra due variabili
- il modo più semplice di farlo è assumere una relazione lineare
- un criterio per individuarla è stimare la relazione lineare (retta) che minimizza gli errori di interpolazione
- la regressione lineare semplice fa questa cosa

variabile dipendente e indipendente

row.names	country	country_code	avg_growth	change_pov
2	Albania	ALB	5.313589	-1.16
6	Argentina	ARG	3.311352	-8.3325
7	Armenia	ARM	6.652428	-11.065
16	Belarus	BLR	4.478133	-1.2275
19	Benin	BEN	4.435067	4.26
21	Bhutan	BTN	7.233625	-22.66
22	Bolivia	BOL	4.175102	-12.51
25	Brazil	BRA	3.136631	-7.433333
27	Bulgaria	BGR	2.751186	0.61
31	Cambodia	KHM	7.699549	-23.59667
37	Chad	TCD	7.484236	-24.51
38	Chile	CHL	4.446929	-1.5
39	China	CHN	9.739079	-20.77
40	Colombia	COL	3.730497	-8.0425
42	Congo, De	COD	2.665066	-14.03
44	Costa Ric	CRI	4.479772	-4.455
46	Croatia	HRV	2.03731	0.38
49	Czech Rep	CZE	2.520561	0.05
51	Djibouti	DJI	2.514258	-2.31
53	Dominican	DOM	5.206205	-3.75
55	Ecuador	ECU	3.586102	-10.2675
57	El Salvad	SLV	2.6032	-8.085
60	Estonia	EST	4.422114	-0.0025

fonte: World Bank

retta dei minimi quadrati



fonte: elaborazione su dati World Bank

stimatori OLS β_0, β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

dove $\bar{X}, \bar{Y}$ sono le medie delle due variabili nel campione

assunzioni che rendono valido quanto detto fino ad ora

- questo metodo per stimare i parametri della retta $(\hat{\beta}_0, \hat{\beta}_1)$ non è sempre valido
- perché i parametri siano una buona stima di quelli veri, β_0, β_1, devono essere soddisfatte alcune condizioni

assunzioni che rendono valido quanto detto fino ad ora

0) il DGP deve effettivamente essere lineare
1) la distribuzione di u_i condizionata a X_i ha media nulla
2) X, Y sono indipendentemente e identicamente distribuite
3) valori estremi (outlier) devono essere improbabili

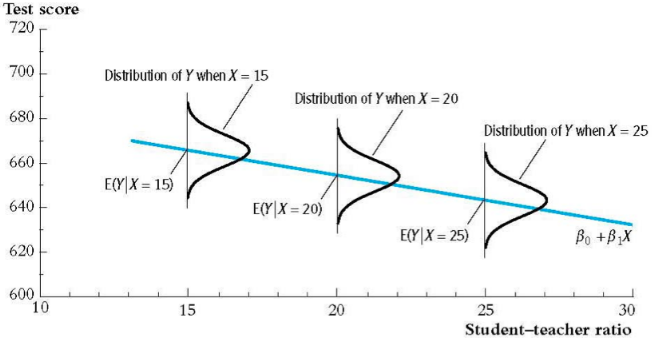
la distribuzione di u_i condizionata a X_i ha media nulla

1) $E(u_i|X_i) = 0 \ \forall i = 1, ..., n$

- gli 'altri fattori' che confluiscono in u e determinano Y sono non sono sistematicamente legati a X
- nell'esempio povertà e crescita economica occorre che altri fattori che influenzano la povertà non abbiano un impatto sistematico sulla crescita.
- esempio di violazione dell'assunzione: una guerra ha effetto sia sulla crescita che sulla povertà
- paesi in guerra avranno un valore basso di crescita e aumento della povertà ma questo non è dovuto al legame fra i due

la distribuzione di u_i condizionata a X_i ha media nulla

1) $E(u_i|X_i) = 0 \ \forall i = 1, ..., n$



X, Y i.i.d.

- campionamento semplice dalla stessa popolazione (età e altezza degli studenti di uniba) allora ogni osservazione si distribuisce alla stessa maniera
- se sono estratti in modo casuale sono anche indipendenti
- esistono casi di non-indipendenza: nel caso dei dati della Banca Mondiale ad esempio sono mancanti i valori di povertà per paesi molto arretrati
- in questi casi il campionamento non è casuale, il campione non rappresenta la popolazione

gli outlier sono improbabili

- outlier: misure fuori dall'intervallo 'normale'
- errori di imputazione dei dati
- si possono scovare guardando i dati
- nel caso ci siano conviene sempre ripetere l'analisi escludendoli ed inclucedendoli
- per valutare la sensibilità delle stime alla loro inclusione

$\hat{\beta}_0, \hat{\beta}_1$ sono stimatori non-distorti

- $\hat{\beta}_0, \hat{\beta}_1$ sono variabili casuali: il loro valore dipende dal campione selezionato
- se valgono le condizioni il loro valore si distribuisce attorno al vero valore (β_0, β_1)
- come accade per la media di un campione: la media campionaria è uno stimatore non distorto della vera media della popolazione
- così accade per i parametri $\hat{\beta}_0, \hat{\beta}_1$ se si verificano le condizioni

Inferenza

- se le assunzioni 1-3 si verificano sono non distorti: $E(\hat{\beta}_0) = \beta_0$ e $E(\hat{\beta}_1) = \beta_1$
- inoltre se n è abbastanza grande la loro distribuzione è ben approssimata da una normale bivariata
- si può applicare il teorema del limite centrale:

$$\hat{\beta}_0 \sim (\beta_0, \sigma_{\beta_0})$$

$$\hat{\beta}_1 \sim (\beta_1, \sigma_{\beta_1})$$

inferenza

- il teorema del limite centrale ci consente di sfruttare le informazioni contenute nel campione per quantificare l'incertezza delle stime ottenute
- quanto è probabile che la relazione che osservo fra Y e X sia veramente positiva?
- è possibile, ammettendo un certo grado di incertezza, individuare una stima minima e massima per il valore del coefficiente?

inferenza

- il grado di incertezza delle nostre stime è quantificabile guardando alla varianza dei due stimatori:

$$var(\hat{\beta}_0) = \sigma^2 \frac{\sum_i^n X_i^2}{\sum_i^n (X_i - \bar{X})^2}$$

$$var(\hat{\beta}_1) = \frac{\sigma^2}{n \sum_i^n (X_i - \bar{X})^2}$$

dove σ^2 è la varianza del termine errore u

inferenza

- il grado di incertezza delle nostre stime è quantificabile guardando alla varianza dei due stimatori:

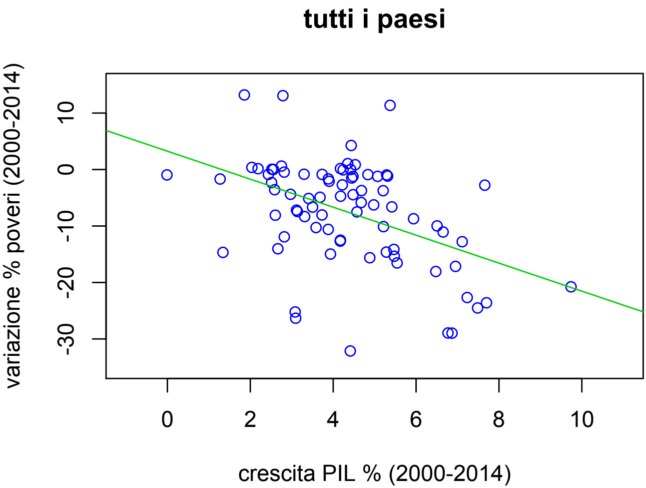
$$var(\hat{\beta}_0) = \sigma^2 \frac{\sum_i^n X_i^2}{\sum_i^n (X_i - \bar{X})^2}$$

$$var(\hat{\beta}_1) = \frac{\sigma^2}{n \sum_i^n (X_i - \bar{X})^2}$$

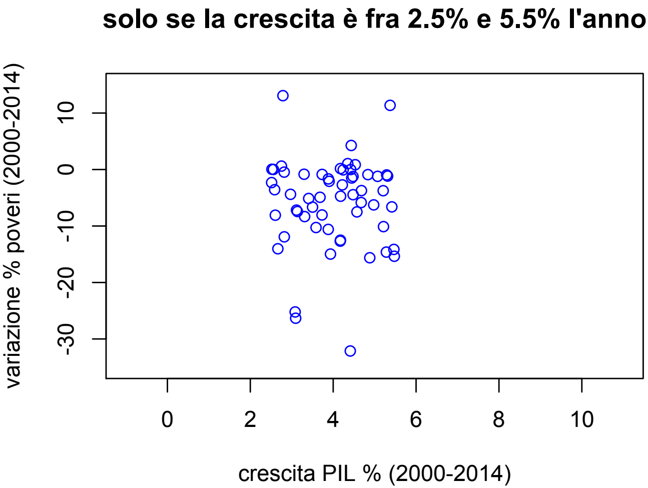
dove σ^2 è la varianza del termine errore u

- tanto maggiore σ^2 tanta più alta è l'incertezza riguardo ai coefficienti
- tanto minore la variabilità di X tanto più alta l'incertezza riguardo ai coefficienti

inferenza: ruolo della variabilità della X



inferenza: ruolo della variabilità della X



test delle ipotesi

- se è valido il test delle ipotesi sarà anche possibile effettuare test delle ipotesi
- " con quale grado di confidenza possiamo affermare che un aumento del PIL si traduce in una riduzione della povertà?"
- ciò equivale a testare l'ipotesi:

$$H_0 : \beta_1 = 0$$

contro l'ipotesi:

$$H_1 : \beta_1 \neq 0$$

test delle ipotesi

- per effettuare il test si deve costruire la statistica $\hat{t}$:

$$\hat{t} = \frac{\hat{\beta}_1 - \beta_1^{H_0}}{\sqrt{var(\hat{\beta}_1)}}$$

- rifiutiamo l'ipotesi nulla (H_0) se $\hat{t}$ è significativamente diversa da zero
- l'ipotesi H_0 può essere rigettata se:

$$|\hat{t}| > t^{\star}$$

- $t^{\star}$ è il valore critico che dipende dal numero di osservazioni (gradi di libertà) e dal livello di confidenza
- nel nostro caso 78 gradi di libertà e 95% $\rightarrow$ 1.99

intervalli di confidenza

- allo stesso modo possiamo individuare l'intervallo nel quale al 95% si trova il vero parametro β_1
- se l'intervallo vero al 95% include la nostra ipotesi nulla (in questo caso $\beta_1 = 0$) non possiamo rifiutarla
- l'intervallo si ottiene con la formula:

$$\left[\hat{\beta}_1 - \sqrt{var(\hat{\beta}_1)} \times |t^\star|, \quad \hat{\beta}_1 + \sqrt{var(\hat{\beta}_1)} \times |t^\star|\right]$$

valore $-$ *p*

- ci possiamo anche chiedere quale sia la probabilità che l'ipotesi nulla sia vera dato il valore $\hat{t}$ che osserviamo
- il *valore* $-$ *p* si ottiene confrontando la statistica $\hat{t}$ con la probabilità di osservare un valore così estremo nella distribuzione della t
- con 78 gradi di libertà un *valore* $-$ *p* pari a 0.05 si associa a una statistica $t^{\star} = 1.99$
- se osserviamo un $\hat{t} = 3.1$ la probabilità che β_1 sia pari a zero scende a 0.1%
- se n è maggiore di qualche decina possiamo abbandonare la t e usare i valori critici della normale standardizzata z

inferenza: sommario

- una volta stimato un modello di regressione OLS possiamo valutare il livello di certezza che si associa alle nostre stime che dipende:
  1. dalla numerosità campionaria
  2. dalla variabilità di X
- la stima della statistica $\hat{t}$ ci consente di:
  - testare ipotesi sul parametri
  - stimare intervalli di confidenza per i parametri
  - calcolare la probabilità che il vero parametro sia zero dato il campione che osserviamo

inferenza: crescita e povertà

- con i dati della Banca Mondiale calcoliamo:

i coefficiente: $\beta_1 = -2.4779$

l'errore standard: $\sqrt{var(\beta_1)} = 0.5533$

la statistica $\hat{t}$: $\frac{\beta_1}{\sqrt{var(\beta_1)}} = -4.478$

l'intervallo di confidenza al 95%:

$$-2.4779 - 1.99 \times 0.5533 \leq \beta_1 \leq -2.4779 + 1.99 \times 0.5533$$

$$-3.57896 \leq \beta_1 \leq -1.3768$$

il $valore - p = Pr(t < -4.478) = 0.000003$
(di conseguenza per un test delle ipotesi posso rigettare per qualsiasi livello di confidenza minore di 99.9997%)

esercizio:

- sapendo che $\beta_0 = 3.2459$, $\sqrt{var(\beta_0)} = 2.5765$
- calcolate gli stessi valori calcolati nella slide precedente
- i gradi di libertà rimangono 78

Il codice R

- nel file di testo che trovate caricato sulla pagina del corso con le slide:
  - le formule per ottenere la stima dei due coefficienti
  - la formula per ottenere i coefficienti con un solo comando
  - la formula per ottenere errori standard, t e $valore - p$ per i coefficienti