

Inferenza sulla retta di regressione

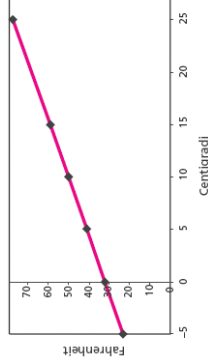
Date due variabili, X e Y, si è interessati a comprendere come la variabile Y (**dipendente** o **risposta**) sia influenzata dalla X (**esplicativa** o **indipendente**).

Y è funzione di X se ad ogni valore di X corrisponde un solo valore di Y. La **relazione funzionale** è **lineare**, se possiamo scrivere:

$$Y = \beta_0 + \beta_1 X$$

β_0 = **intercetta**

β_1 = **coefficiente angolare**



Introducendo opportune assunzioni si ottiene il **modello di regressione lineare semplice**.

Assunzione 1:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{per ogni osservazione } i=1, \dots, n$$

Assunzione 2:

Le ε_i sono variabili casuali indipendenti con valore atteso $E(\varepsilon_i) = 0$ e varianza costante $V(\varepsilon_i) = \sigma^2$ per ogni $i=1, \dots, n$

Assunzione 3:

I valori x_i della variabile esplicativa X sono noti senza errore

Negli studi empirici, la relazione tra Y e X non è mai funzionale (a un valore X corrispondono più valori di Y).

Una **relazione statistica** tra la Y e la X può essere descritta da:

$$Y = f(X) + \varepsilon$$

$f(X)$ definisce il contributo della X

ε rappresenta il contributo di tutti i fattori non osservati

→ $f(X)$ è una **componente deterministica**

→ ε è una **componente stocastica**

→ Y è una **variabile casuale**.

Assunzione 1: implica che la funzione $f(X)$ è **lineare**.

Assunzione 2: implica che per ogni valore fissato di X, la Y possiede sempre lo stesso grado di variabilità (**ipotesi di omoschedasticità**). Inoltre, poiché la ε_i è una variabile casuale, anche Y è una variabile casuale.

Pertanto, le osservazioni y_i sono realizzazioni di variabili casuali

- **indipendenti**
- **con valore atteso** $E(Y_i | X = x_i) = \beta_0 + \beta_1 x_i$
- **con varianza** $V(Y_i | X = x_i) = \sigma^2$

Stima puntuale dei coefficienti di regressione

Indicheremo con: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ il valore di Y fornito dalla retta stimata dove $\hat{\beta}_0$ e $\hat{\beta}_1$ sono le stime dei coefficienti di regressione.

Metodo di stima $\Rightarrow$ Metodo dei minimi quadrati

Consiste nel ricercare le stime di β_0 e β_1 , che rendono minima la funzione di perdita:

$$G(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Procedimento:

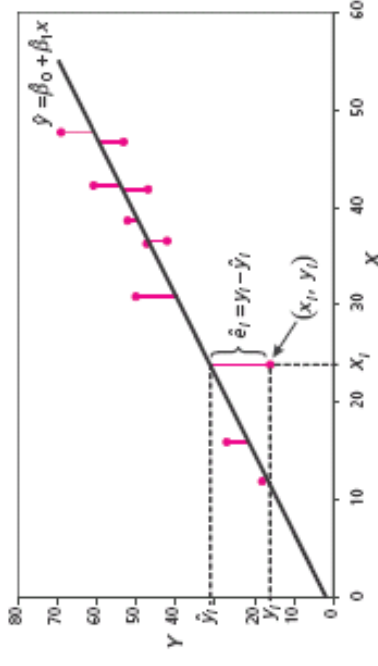
1) Porre uguali a zero le derivate prime rispetto ai parametri

$$\begin{cases} \frac{\partial G(\beta_0, \beta_1)}{\partial \beta_0} = 0 \\ \frac{\partial G(\beta_0, \beta_1)}{\partial \beta_1} = 0 \end{cases}$$

2) Risolvendo il sistema si ottengono le stime dei minimi quadrati dei coefficienti di regressione

$$\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Chiameremo **residuo** i-esimo la differenza tra il valore osservato y_i e quello fornito dalla retta stimata, $\hat{y}_i$



Proprietà degli stimatori dei coefficienti

Proprietà degli stimatori dei minimi quadrati

1. B_0 e B_1 sono stimatori **corretti** di β_0 e β_1
2. Nella classe degli stimatori corretti di β_0 e β_1 che sono funzioni lineari delle Y_i , gli stimatori dei minimi quadrati sono i **più efficienti**. (Gauss-Markov)
3. La varianza e covarianza degli stimatori dei minimi quadrati sono:

$$V(B_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad V(B_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\text{Cov}(B_0, B_1) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Proprietà dello stimatore della risposta media

Per lo **stimatore della risposta media** $\hat{Y}_i$ valgono le seguenti proprietà:

- 1. Lo stimatore $\hat{Y}_i$ è **corretto**, ossia $E(\hat{Y}_i) = \beta_0 + \beta_1 x_i$
- 2. La varianza è:
$$V(\hat{Y}_i) = \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{h=1}^n (x_h - \bar{x})^2} \right]$$

Una misura della variabilità degli stimatori dei coefficienti di regressione e della risposta media è data dagli **errori standard**, ossia le radici quadrate delle varianze:

$$\sigma(B_0) = \sqrt{V(B_0)} \quad \sigma(B_1) = \sqrt{V(B_1)} \quad \sigma(\hat{Y}_i) = \sqrt{V(\hat{Y}_i)}$$

Assunzione di normalità degli errori

Assunzione:

Le variabili casuali ε_i hanno distribuzione Normale.

Tenendo conto di questa ulteriore assunzione il modello di regressione lineare può essere definito nel seguente modo:

Le osservazioni Y_i sono realizzazioni di variabili casuali Normali con valore atteso $\beta_0 + \beta_1 x_i$ e varianza σ^2 ossia:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Errore standard

Gli errori standard dipendono dalla quantità ignota:

$$\sigma^2 = V(Y_i) = V(\varepsilon_i)$$

per tanto la si sostituisce con una sua stima s^2 ottenendo gli stimatori

$$s(B_0) \quad s(B_1) \quad s(\hat{Y}_i)$$

Lo stimatore che si utilizza per ottenere la stima della varianza è dato da:

$$s^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-2}$$

La radice quadrata di s^2 è una misura della variabilità degli scostamenti dei valori osservati da quelli previsti dal modello e per tale ragione viene usualmente chiamato **errore standard di regressione**.

Assunzione di normalità degli errori

L'introduzione di tale assunzione ha alcune importanti implicazioni:

- Gli stimatori dei m.q. B_0 e B_1 , (combinazioni lineari della v.c. Y_i) si distribuiscono come una Normale bivariata.
- Si può applicare il metodo di massima verosimiglianza per trovare gli **stimatori di massima verosimiglianza** dei parametri del modello. Questi stimatori risultano uguali a quelli dei minimi quadrati.

→ Si ha che:
$$\frac{B_1 - \beta_1}{s(B_1)} \sim t_{n-2} \quad \frac{B_0 - \beta_0}{s(B_0)} \sim t_{n-2}$$

dove t_{n-2} indica una v.c. t-Student con n-2 gradi di libertà e $s(B_0)$ e $s(B_1)$ indicano gli errori standard di B_0 e B_1 .

Intervalli di confidenza per i parametri

Gli intervalli di confidenza per i parametri β_0 e β_1 a un livello di confidenza $1 - \alpha$ sono dati da:

$B_0 \pm t_{\alpha/2} s(B_0)$ e $B_1 \pm t_{\alpha/2} s(B_1)$

Sotto l'ipotesi nulla $\beta_1 = b_1$ la statistica test è:

$t = \frac{B_1 - b_1}{s(B_1)} \sim t_{n-2}$

In corrispondenza del sistema d'ipotesi:

$H_0 : \beta_1 = b_1$ contro $H_1 : \beta_1 \neq b_1$

A un livello di significatività α la regione di rifiuto è data dai valori della statistica test superiori in valore assoluto a $t_{\alpha/2}$. La verifica d'ipotesi più frequente è:

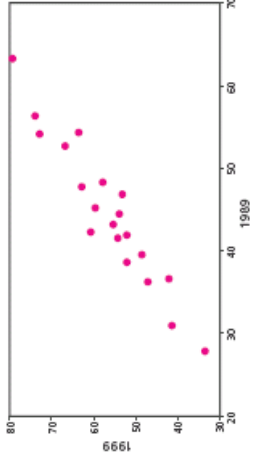
$H_0 : \beta_1 = 0$ contro $H_1 : \beta_1 \neq 0$

con $t = \frac{B_1}{s(B_1)} \sim t_{n-2}$

➡ ossia che la Y sia indipendente in media dalla X.

Esempio

Su un campione di 20 aree amministrative si osserva il **reddito pro-capite** nel 1989 (X) e 1999 (Y).



Area	X:1989	Y:1999
1	47,8	63,0
2	27,9	33,4
3	36,6	42,0
4	54,2	72,8
5	41,9	52,0
6	44,4	54,0
7	54,3	63,4
8	42,3	60,7
9	48,2	58,0
10	41,5	54,4
11	43,2	55,5
12	56,3	74,0
13	63,3	79,2
14	46,8	53,1
15	45,2	59,6
16	38,7	52,0
17	36,3	47,2
18	39,5	48,7
19	30,9	41,4
20	52,6	66,9

Si ipotizza il seguente modello:

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Si ottengono le seguente stime dei coefficienti del modello:

$\hat{\beta}_1 = 1,255$ $\hat{\beta}_0 = 0,595$

ossia la retta di regressione: $\hat{y}_i = 0,595 + 1,255x_i$

Il coefficiente di correlazione è $\rho_{XY} = 0,956$

SQT=2497,6 da cui: $R^2_{XY} = (0,956)^2 = 0,914$

ossia circa il 91% della variabilità totale di Y è spiegata dal modello di regressione.

Gli errori standard sono: $s(B_1) = 0,091$ $s(B_0) = 4,111$

Fissato un livello di confidenza $1 - \alpha = 0,95$ il valore della t-Student con 10 g.d.l. è $t_{0,025} = 2,101$. Pertanto, si trovano gli intervalli di confidenza:

int.conf. (β_0) : $0,595 \pm 8,637$ int.conf. (β_1) : $1,255 \pm 0,191$
ossia $- 8,042 \leq \beta_0 \leq 9,232$ $1,064 \leq \beta_1 \leq 1,446$

Si vuole verificare il sistema d'ipotesi:

$$H_0 : \beta_1 = 0 \quad \text{contro} \quad H_1 : \beta_1 \neq 0$$

La statistica test, sotto l'ipotesi nulla, è: $t = \frac{1,255}{0,091} = 13,79$

A un livello di significatività $\alpha = 0,99$ corrisponde un valore della t-Student con 18 g.d.l. pari a: $t_{0,005} = 2,8784$

pertanto $t = 13,79 > 2,8784 = t_{0,005}$  **rifiuto** H_0

Osservato il valore del p-value=0,000, possiamo rifiutare l'ipotesi nulla per qualsiasi ragionevole livello di α .
Esiste pertanto una relazione lineare tra il reddito pro-capite del 1999 e quello del 1989.

$$\hat{Y}_i \pm t_{\alpha/2} s(\hat{Y}_i)$$

$$s(\hat{Y}_i) = \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{h=1}^n (x_h - \bar{x})^2} \right]}$$

Dall'espressione dell'intervallo di confidenza si osserva che:

- ➔ l'intervallo aumenta al crescere dell'errore standard $\sqrt{s^2}$
- ➔ l'intervallo diminuisce al crescere della dimensione campionaria n
- ➔ l'intervallo aumenta al crescere della distanza di x_i da $\bar{x}$

Inferenza per la risposta media

La funzione di regressione per un dato valore $X=x_i$, restituisce la stima del valor medio $E(Y_i|x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{Y}_i$

Sotto l'assunzione di Normalità, lo stimatore $\hat{Y}_i$ è distribuito come una Normale e inoltre:

$$\frac{\hat{Y}_i - E(\hat{Y}_i|x_i)}{s(\hat{Y}_i)} \sim t - Student(n-2 \text{ g.d.l.})$$

segue che l'intervallo di confidenza per il **valore medio della** Y_i a un livello di confidenza $1-\alpha$ è

$$\hat{Y}_i \pm t_{\alpha/2} s(\hat{Y}_i)$$

dove

$$s(\hat{Y}_i) = \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{h=1}^n (x_h - \bar{x})^2} \right]}$$

Inferenza per la previsione

Quando il modello di regressione lineare è utilizzato a **fini previsivi**, bisogna distinguere tra:

- ➔ previsione del **valor medio** di Y_i , ossia $E(Y_i|x_i)$
- ➔ previsione del **valore singolo** della Y_i

La stima puntuale delle due quantità è la stessa e pari a $\hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{Y}_i$ ma lo stimatore del valore singolo presenta un errore standard maggiore dello stimatore per il valore medio.

L'errore standard dello stimatore del singolo valore è:

$$s(Y_i - \hat{Y}_i) = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{h=1}^n (x_h - \bar{x})^2}}$$

la cui stima si ottiene sostituendo a σ la stima $\sqrt{s^2}$

Si dimostra che: $\frac{Y_i - \hat{Y}_i}{s(Y_i - \hat{Y}_i)} \sim t_{n-2}$

e ne consegue che l'intervallo di confidenza per la previsione del singolo valore Y_i a livello $1 - \alpha$ è dato da:

$$\hat{Y}_i \pm t_{\alpha/2} s(Y_i - \hat{Y}_i)$$

Intervalli di confidenza al 95% per il valore medio e il valore previsto.

