

La regressione lineare

Vogliamo studiare la dipendenza di una variabile da un'altra. Supponiamo che la relazione tra le due variabili possa essere scritta come

$$Y_i = a + bx_i + \varepsilon_i, \quad i = 1, 2, \dots, N$$

dove

- Y_i è una variabile casuale
- x_i è un osservazione di un'altra variabile
- a e b sono l'intercetta e il coefficiente lineare della retta di regressione. Sono costanti incognite
- ε_i è una variabile casuale t.c. $E(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$ e $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$

Abbiamo che

$$E(Y_i) = a + bx_i \quad \text{e} \quad \text{Var}(Y_i) = \sigma^2$$

Lo studio della relazione tra due variabili avviene in due passi:

- descrittivo: si tenta di spiegare il legame solo attraverso i dati. Nessuna ipotesi sulla distribuzione di probabilità degli errori è fatta
- inferenziale: si cerca di estendere i risultati ottenuti nel primo passo alla popolazione. Occorre fare delle ipotesi sulla distribuzione degli errori

Abbiamo visto che le stime di a e b , ottenute minimizzando la somma degli errori al quadrato, sono

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2} \quad \hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$$

Dove

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$$

$$\sigma_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}), \quad \sigma_x^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Il modello condizionatamente normale

Le osservazioni sono le N coppie (x_i, y_i) . I valori della variabile X , x_1, \dots, x_N , sono considerati fissati. I valori della variabile Y , y_1, \dots, y_N sono considerati le realizzazioni di N v.a. indipendenti Y_1, \dots, Y_N la cui distribuzione è

$$Y_i \sim N(a + bx_i, \sigma^2)$$

Questa ipotesi è equivalente a chiedere che

$$\varepsilon_i \sim N(0, \sigma^2)$$

Sotto queste ipotesi possiamo ricavare la distribuzione degli stimatori di a , b e σ . Ricordiamo infatti che anche la varianza degli errori (o delle Y_i) è incognita.

Infatti gli stimatori introdotti sono funzioni delle Y_i , e quindi sono v.a.

Per stimare la varianza dei residui ε_i , passiamo attraverso lo stimatore dei residui

$$\hat{\varepsilon}_i = Y_i - \hat{a} - \hat{b}x_i$$

e quindi definiamo

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N \hat{\varepsilon}_i^2$$

La distribuzione degli stimatori

Gli stimatori di a e b sono

$$\hat{b} = \frac{\sum_{i=1}^N (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^N (x_i - \bar{x})^2} \quad \text{e} \quad \hat{a} = \bar{Y} - \hat{b} \cdot \bar{x}$$

Si dimostra che, essendo funzioni lineari delle v.a. Y_i , essi sono ancora distribuiti come Normali di parametri rispettivamente

$$\hat{a} \sim N \left(a, \frac{\sigma^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} \sum_{i=1}^N x_i^2 \right)$$

$$\hat{b} \sim N \left(b, \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \right)$$

Quindi sono stimatori non distorti rispettivamente di a e b .

Per quanto riguarda lo stimatore di σ^2 esso è distorto, infatti si dimostra che

$$E(\hat{\sigma}^2) = \frac{N-2}{N} \sigma^2$$

Allora lo stimatore

$$S^2 = \frac{1}{N-2} \sum_{i=1}^N \varepsilon_i^2 \quad \text{è tale che} \quad \frac{N-2}{\sigma^2} S^2 \sim \chi^{N-2}$$

Inferenza per i parametri a e b

Poiché σ^2 non è nota abbiamo che

$$\frac{\hat{a} - a}{S \sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}}} \sim t^{N-2}$$

Infatti abbiamo standardizzato la v.a, \hat{a} utilizzando una stima della sua varianza

$$\frac{\hat{b} - b}{\frac{S}{\sqrt{\sum (x_i - \bar{x})^2}}} \sim t^{N-2}$$

Infatti, ancora, abbiamo standardizzato la v.a, \hat{b} utilizzando una stima della sua varianza

Sulla base di questi risultati possiamo calcolare gli intervalli di confidenza per i parametri a e b ed effettuare verifiche d'ipotesi sui valori di tali parametri.

Intervalli di confidenza per i parametri a e b

Gli intervalli di confidenza a livello di fiducia $1 - \alpha$ sono rispettivamente dati da

$$a \in \left(\hat{a} - t_{1-\frac{\alpha}{2}}^{N-2} S \sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}}, \hat{a} + t_{1-\frac{\alpha}{2}}^{N-2} S \sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}} \right)$$

mentre per b abbiamo

$$b \in \left(\hat{b} - t_{1-\frac{\alpha}{2}}^{N-2} \frac{S}{\sqrt{\sum (x_i - \bar{x})^2}}, \hat{b} + t_{1-\frac{\alpha}{2}}^{N-2} \frac{S}{\sqrt{\sum (x_i - \bar{x})^2}} \right)$$

Verifica d'ipotesi per il parametro b

Di solito si vuole verificare se la retta stimata abbia senso. Questo equivale a verificare se il parametro b sia uguale a zero. Quindi effettuiamo il test

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

Rifiutiamo l'ipotesi H_0 a livello α se

$$\left| \frac{\frac{\hat{b} - 0}{S}}{\sqrt{\sum (x_i - \bar{x})^2}} \right| > t_{1-\frac{\alpha}{2}}^{N-2}$$

Stima del valore di Y per un fissato x_0

Supponiamo che (x_i, Y_i) , $i = 1, \dots, N$ soddisfino le ipotesi del modello condizionatamente Normale, e sulla base delle N osservazioni abbiamo le stime \hat{a} , \hat{b} e S . Sia x_0 il valore della variabile X per il quale ci interessa una stima della variabile risposta Y .

Si può dimostrare che lo stimatore $\hat{y}_0 = \hat{a} + \hat{b}x_0$ è tale che

$$\hat{y}_0 \sim N \left(a + bx_0, \sigma^2 \left(\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \right)$$

Quindi

$$\frac{y_0 - (a + bx_0)}{S \sqrt{\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}} \sim t^{N-2}$$

e l'intervallo di confidenza a livello di fiducia $1 - \alpha$ è

$$y_0 \in \left(\hat{a} + \hat{b}x_0 \pm t_{1-\frac{\alpha}{2}}^{N-2} S \sqrt{\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right)$$

Esempio: riprendiamo l'esempio: abbiamo rilevato i seguenti caratteri su $n = 25$ unità

Y : libbre di vapore utilizzate in un mese

X_1 : temperatura media mensile in gradi F

X_2 : numero di giorni di operatività in un mese

X_3 : numero di riavviamenti (startup) in un mese

Y	X_1	X_2	X_3
10.98	35.3	20	4
11.13	29.7	20	5
12.51	30.8	23	4
8.40	58.8	20	4
9.27	61.4	21	5
8.73	71.3	22	4
6.36	74.4	11	2
8.50	76.7	23	5
7.82	70.7	21	4
9.14	57.5	20	5
8.24	46.4	20	4
12.19	28.9	21	4
11.88	28.1	21	5
9.57	39.1	19	5
10.94	46.8	23	4
9.58	48.5	20	4
10.09	59.3	22	6
8.11	70.0	22	4
6.83	70.0	11	3
8.88	74.5	23	4
7.68	72.1	20	4
8.47	58.1	21	6
8.86	44.6	20	4
10.36	33.4	20	4
11.08	28.6	22	5

1. Dopo aver trovato i parametri della retta di regressione $Y_i = a + bx_i + \varepsilon_i$ stabilire se il parametro b è diverso da zero con un livello di significatività pari a $\alpha = 0.01$
2. Calcolare l'intervallo di fiducia per il parametro a e per il parametro b a livello $1 - \alpha = 0.95$.
3. Calcolare l'intervallo di fiducia per i valori $x_0 = \bar{x}$ e $x_0 = 30$ a livello $1 - \alpha = 0.90$

1. I valori delle stime dei parametri sono (le abbiamo già calcolate)

$$\hat{a} = 13.62, \quad \hat{b} = -0.08$$

Per verificare l'ipotesi $H_0 : b = 0$ contro l'alternativa $H_1 : b \neq 0$ dobbiamo innanzitutto calcolare la stima corretta di σ^2 . Questa si ottiene andando a calcolare tutti i residui $\hat{\varepsilon}_i = y_i - \hat{y}_i$. Calcoliamo i primi due

$$\hat{y}_1 = 13.62 - 0.08 * 35.3 = 10.81,$$

$$y_1 - \hat{y}_1 = 10.98 - 10.81 = 0.17$$

$$\hat{y}_2 = 13.62 - 0.08 * 29.7 = 11.25,$$

$$y_2 - \hat{y}_2 = 11.13 - 11.25 = -0.12$$

Quindi calcoliamo S

$$S^2 = \sum_{i=1}^{25} \hat{\varepsilon}_i^2 = 18.223$$

A questo punto calcoliamo la statistica test

$$\left| \frac{\frac{\hat{b} - 0}{S}}{\sqrt{\sum (x_i - \bar{x})^2}} \right| = |-0.08/0.011| = |-7.59| = 7.59$$

Il valore $t_{1-\frac{\alpha}{2}}^{N-2} = t_{0.995}^{23} = 2.807$. Quindi l'ipotesi nulla va rifiutata, cioè il parametro b è significativamente diverso 0.

2. Calcoliamo l'intervallo di fiducia per il parametro b . Abbiamo $t_{1-\frac{\alpha}{2}}^{N-2} = t_{0.975}^{23} = 2.069$. L'estremo inferiore è dato da

$$\hat{b} - t_{1-\frac{\alpha}{2}}^{N-2} \frac{S}{\sqrt{\sum (x_i - \bar{x})^2}} = -0.08 - 2.069 * 0.011 = -0.102$$

L'estremo superiore è dato da

$$\hat{b} + t_{1-\frac{\alpha}{2}}^{N-2} \frac{S}{\sqrt{\sum (x_i - \bar{x})^2}} = -0.08 + 2.069 * 0.011 = -0.057$$

Come si vede l'intervallo a livello $1 - \alpha = 0.95$ non contiene lo 0, questo significa che rifiutiamo l'ipotesi nulla a livello $\alpha = 0.05$.

Calcoliamo l'intervallo di fiducia per a , il valore dello scarto quadratico medio di \hat{a} è

$$S\sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}} = 0.581$$

L'estremo inferiore è dato da

$$\begin{aligned}\hat{a} - t_{1-\frac{\alpha}{2}}^{N-2} S\sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}} &= 13.62 - 2.069 * 0.581 \\ &= 12.418\end{aligned}$$

L'estremo superiore è dato da

$$\begin{aligned}\hat{a} + t_{1-\frac{\alpha}{2}}^{N-2} S\sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}} &= 13.62 + 2.069 * 0.581 \\ &= 14.822\end{aligned}$$

3. Farlo per esercizio

Esercizio: Una azienda farmaceutica osserva un campione di 10 fumatori, onde verificare in che modo e con quale intensità il numero di sigarette fumate (X) possa influenzare l'andamento spirometrico (capacità polmonare espressa in volume d'aria che esce ad ogni atto respiratorio), Y .

Analisi spirometrica	sigarette fumate al dì
600	1
420	3
480	4
500	2
320	10
380	9
620	4
280	20
380	12
400	8

- Determinare i parametri della retta di regressione che spieghi la capacità spirometrica in funzione del numero di sigarette fumate.
- Verificare a livello $\alpha = 0.05$ se il coefficiente angolare della retta stimato può ritenersi significativo, cioè diverso da zero.

- c) Costruire l'intervallo di fiducia a livello $1 - \alpha = 0.99$ per l'intercetta della retta.

La stima coi minimi quadrati dei coefficienti della retta $y = a + bx$ ci da

$$\hat{a} = 552.393, \quad \hat{b} = -15.670.$$

Dobbiamo verificare l'ipotesi $H_0 : b = 0$ contro l'alternativa $H_1 : b \neq 0$. Sappiamo che rifiutiamo l'ipotesi H_0 a livello α se

$$\left| \frac{\frac{\hat{b}}{S}}{\sqrt{\sum (x_i - \bar{x})^2}} \right| > t_{1-\frac{\alpha}{2}}^{N-2}.$$

Abbiamo che $S = 69.08$, mentre $\sum (x_i - \bar{x})^2 = 302.1$ e quindi $\frac{S}{\sqrt{\sum (x_i - \bar{x})^2}} = 3.974$. Da cui

$$\left| \frac{\frac{\hat{b}}{S}}{\sqrt{\sum (x_i - \bar{x})^2}} \right| = |-3.94| = 3.94$$

Poiché $t_{1-\frac{\alpha}{2}}^{N-2} = t_{0.975}^8 = 2.3060$ rifiutiamo l'ipotesi nulla, il modello è significativo.

Gli estremi inferiore e superiore dell'intervallo di fiducia per il parametro σ sono rispettivamente

$$\hat{\sigma} \mp t_{1-\frac{\alpha}{2}}^{N-2} S \sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}}.$$

Abbiamo $\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2} = 0.2763$, mentre

$$S \sqrt{\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}} = 36.318$$

Inoltre $t_{1-\frac{\alpha}{2}}^{N-2} = t_{0.995}^8 = 3.355$ per cui l'estremo inferiore risulta 430.5321 mentre l'estremo superiore 674.2543.