

Regressione lineare

Lucio Demeio

Dipartimento di Ingegneria Industriale e Scienze Matematiche
Università Politecnica delle Marche

Siano x ed y due variabili legate tra loro da una forma funzionale del tipo $y = f(x)$. Supponiamo di eseguire n misure (o n esperimenti) in corrispondenza ai valori x_1, x_2, \dots, x_n della variabile indipendente x . I valori Y_i , corrispondenti alle $x = x_i$, non seguiranno esattamente la forma funzionale $f(x)$ perchè sono soggetti ad errori; le Y_i sono quindi delle variabili casuali tali che $Y_i = f(x_i) + W_i$ dove W_i è un errore che si può rappresentare come una variabile casuale di media nulla e varianza σ^2 incognita. Facciamo l'ipotesi che $W_i \sim N(0, \sigma^2)$, quindi $Y_i \sim N(f(x_i), \sigma^2)$. Sia $F_{Y_1, \dots, Y_n}(y_1, \dots, y_n)$ la distribuzione congiunta delle n variabili Y_1, \dots, Y_n . Per le ipotesi fatte abbiamo

$$\begin{aligned} F_{Y_1, \dots, Y_n}(y_1, \dots, y_n) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_1 - f(x_1))^2}{\sigma^2}\right\} \dots \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_n - f(x_n))^2}{\sigma^2}\right\} = \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{-\sum_{i=1}^n \frac{(y_i - f(x_i))^2}{\sigma^2}\right\} \end{aligned} \quad (1)$$

Limitiamoci ad esaminare il caso semplice della dipendenza lineare, $f(x) = \alpha + \beta x$. Abbiamo allora

$$F_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{-\sum_{i=1}^n \frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2}\right\}, \quad (2)$$

con α e β parametri da stimare. Seguiamo il principio di massima verosimiglianza; la (2) è massima quando l'esponente è minimo. Gli stimatori per α e β corrispondono quindi alle soluzioni delle equazioni

$$\frac{\partial}{\partial \alpha} \sum_{i=1}^n \frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2} = 0 \quad (3)$$

$$\frac{\partial}{\partial \beta} \sum_{i=1}^n \frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2} = 0. \quad (4)$$

Sviluppando i dettagli otteniamo:

$$\sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0 \quad (5)$$

$$\sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) = 0 \quad (6)$$

Introduciamo ora

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \quad (7)$$

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i \quad (8)$$

Le equazioni (5) e (6) diventano così

$$\begin{aligned} n \bar{y}_n - n \alpha - n \beta \bar{x}_n &= 0 \\ \sum_{i=1}^n x_i y_i - n \alpha \bar{x}_n - \beta \sum_{i=1}^n x_i^2 &= 0, \end{aligned}$$

cioè il sistema

$$\alpha + \bar{x}_n \beta = \bar{y}_n \quad (9)$$

$$n \bar{x}_n \alpha + \sum_{i=1}^n x_i^2 \beta = \sum_{i=1}^n x_i y_i \quad (10)$$

L'equazione (9) fornisce

$$\alpha = \bar{y}_n - \bar{x}_n \beta; \quad (11)$$

introducendo per comodità la notazione

$$\Delta = \sum_{i=1}^n x_i^2 - n \bar{x}_n^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2, \quad (12)$$

sostituendo la (11) nella (10), otteniamo:

$$n \bar{x}_n (\bar{y}_n - \bar{x}_n \beta) + \sum_{i=1}^n x_i^2 \beta = \sum_{i=1}^n x_i y_i$$

$$\Delta \beta = \sum_{i=1}^n x_i y_i - n \bar{x}_n \bar{y}_n$$

$$\beta = \frac{\sum_{i=1}^n x_i y_i - n \bar{x}_n \bar{y}_n}{\Delta}$$

Gli stimatori per α e β sono pertanto

$$\hat{\beta} = B \equiv \frac{1}{\Delta} \left(\sum_{i=1}^n x_i Y_i - n \bar{x}_n \bar{Y}_n \right) = \frac{1}{\Delta} \sum_{i=1}^n (x_i - \bar{x}_n) Y_i \quad (13)$$

$$\hat{\alpha} = A \equiv \bar{Y}_n - \bar{x}_n B \quad (14)$$

dove

$$\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i. \quad (15)$$

Dimostriamo che gli stimatori (13) e (14) sono corretti. A tal proposito ricordiamo che

$$E[Y_i] = \alpha + \beta x_i \quad (16)$$

e quindi

$$E[\bar{Y}_n] = \frac{1}{n} \sum_{i=1}^n E[Y_i] = \alpha + \beta \bar{x}_n \quad (17)$$

Usando le equazioni (16) e (17) e le proprietà dell'operatore di media otteniamo quindi:

$$\begin{aligned} E[\hat{\beta}] &= \frac{1}{\Delta} \left(\sum_{i=1}^n x_i E[Y_i] - n \bar{x}_n E[\bar{Y}_n] \right) = \\ &= \frac{1}{\Delta} \left[\sum_{i=1}^n x_i (\alpha + \beta x_i) - n \bar{x}_n (\alpha + \beta \bar{x}_n) \right] = \frac{1}{\Delta} \beta \left(\sum_{i=1}^n x_i^2 - n \bar{x}_n^2 \right) = \beta \end{aligned} \quad (18)$$

$$E[\hat{\alpha}] = E[\bar{Y}_n] - \bar{x}_n E[\hat{\beta}] = \alpha + \beta \bar{x}_n - \bar{x}_n \beta = \alpha \quad (19)$$

Gli stimatori A e B dati dalle equazioni (13) e (14) sono dunque corretti.

Calcoliamo ora le varianze. Notiamo, preliminarmente, che \bar{Y}_n e $\hat{\beta}$ sono scorrelati:

$$\begin{aligned} Cov(\bar{Y}_n, \hat{\beta}) &= Cov \left(\frac{1}{n} \sum_{i=1}^n Y_i, \frac{1}{\Delta} \sum_{j=1}^n (x_j - \bar{x}_n) Y_j \right) = \\ &= \frac{1}{n \Delta} \sum_{i,j=1}^n (x_j - \bar{x}_n) Cov(Y_i, Y_j) = \frac{1}{n \Delta} \sum_{i=1}^n (x_i - \bar{x}_n) Var(Y_i) = \\ &= \frac{\sigma^2}{n \Delta} \sum_{i=1}^n (x_i - \bar{x}_n) = 0, \end{aligned} \quad (20)$$

dove abbiamo usato l'indipendenza di Y_i ed Y_j per $i \neq j$.

Passando alle varianze:

$$Var(\hat{\beta}) = Var \left(\frac{1}{\Delta} \sum_{i=1}^n (x_i - \bar{x}_n) Y_i \right) = \frac{1}{\Delta^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 \sigma^2 = \frac{\sigma^2}{\Delta} \quad (21)$$

$$\begin{aligned} Var(\hat{\alpha}) &= Var(\bar{Y}_n - \bar{x}_n \hat{\beta}) = Var(\bar{Y}_n) + \bar{x}_n^2 Var(\hat{\beta}) = \frac{\sigma^2}{n} + \bar{x}_n^2 \frac{\sigma^2}{\Delta} = \\ &= \frac{\sigma^2}{n \Delta} (\Delta + n \bar{x}_n^2) = \frac{\sum_{i=1}^n x_i^2}{n \Delta} \sigma^2 \end{aligned} \quad (22)$$

Possiamo pertanto concludere che

$$\hat{\alpha} \sim N \left(\alpha, \frac{\sum_{i=1}^n x_i^2}{n \Delta} \sigma^2 \right) \quad (23)$$

$$\hat{\beta} \sim N \left(\beta, \frac{\sigma^2}{\Delta} \right) \quad (24)$$

Per determinare gli intervalli di confidenza consideriamo inizialmente la variabile

$$S_R^2 = \sum_{i=1}^n \frac{(Y_i - \alpha - \beta x_i)^2}{\sigma^2}$$

che è la somma di n variabili normali standard, quindi $S_R^2 \sim \chi_n^2$. Se ai parametri α e β sostituiamo gli stimatori $\hat{\alpha}$ e $\hat{\beta}$ perdiamo due gradi di libertà, il che, in analogia con quanto succede per la varianza campionaria, rende plausibile l'affermazione che

$$S_R^2 = \sum_{i=1}^n \frac{(Y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{\sigma^2} \sim \chi_{n-2}^2. \quad (25)$$

Abbiamo quindi

$$\frac{\hat{\beta} - \beta}{\sigma} \sqrt{\Delta} \sqrt{\frac{n-2}{S_R^2}} = (\hat{\beta} - \beta) \sqrt{\Delta \frac{n-2}{\sigma^2 S_R^2}} = (\hat{\beta} - \beta) \sqrt{\Delta \frac{n-2}{S^2}} \sim t_{n-2} \quad (26)$$

$$\frac{\hat{\alpha} - \alpha}{\sigma} \sqrt{\frac{n \Delta}{\sum_{i=1}^n x_i^2}} \sqrt{\frac{n-2}{S_R^2}} = (\hat{\alpha} - \alpha) \sqrt{\frac{n(n-2) \Delta}{S^2 \sum_{i=1}^n x_i^2}} \sim t_{n-2}, \quad (27)$$

dove $S^2 = \sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2$. Pertanto, fissato un livello di confidenza γ , in analogia con il problema degli intervalli di confidenza per la media, abbiamo

$$P \left(|B - \beta| \sqrt{\Delta \frac{n-2}{S^2}} \leq t_{n-2}(\gamma/2) \right) = 1 - \gamma$$

$$P \left(|A - \alpha| \sqrt{\frac{n(n-2) \Delta}{S^2 \sum_{i=1}^n x_i^2}} \leq t_{n-2}(\gamma/2) \right) = 1 - \gamma$$

ovvero

$$P \left(B - \sqrt{\frac{S^2}{(n-2) \Delta}} t_{n-2}(\gamma/2) \leq \beta \leq B + \sqrt{\frac{S^2}{(n-2) \Delta}} t_{n-2}(\gamma/2) \right) = 1 - \gamma$$

$$P \left(A - \sqrt{\frac{S^2 \sum_{i=1}^n x_i^2}{n(n-2) \Delta}} t_{n-2}(\gamma/2) \leq \alpha \leq A + \sqrt{\frac{S^2 \sum_{i=1}^n x_i^2}{n(n-2) \Delta}} t_{n-2}(\gamma/2) \right) = 1 - \gamma$$

Gli intervalli di confidenza di livello γ per i parametri α e β sono pertanto

$$\left(A - \sqrt{\frac{S^2 \sum_{i=1}^n x_i^2}{n(n-2) \Delta}} t_{n-2}(\gamma/2), A + \sqrt{\frac{S^2 \sum_{i=1}^n x_i^2}{n(n-2) \Delta}} t_{n-2}(\gamma/2) \right) \quad (28)$$

$$\left(B - \sqrt{\frac{S^2}{(n-2) \Delta}} t_{n-2}(\gamma/2), B + \sqrt{\frac{S^2}{(n-2) \Delta}} t_{n-2}(\gamma/2) \right) \quad (29)$$

Problema. L'ossigeno consumato da una persona che cammina è funzione della sua velocità. La seguente tabella riporta il volume di ossigeno consumato a varie velocità di cammino. Ipotezzando una relazione lineare, scrivere l'equazione della retta di regressione.

Velocità (km/h)	Ossigeno (l/h)
0	19.5
1	22.1
2	24.3
3	25.7
4	26.1
5	28.5
6	30.0
7	32.1
8	32.7
9	32.7
10	35.0

Soluzione. La retta di regressione ha equazione $y = A + Bx$ e le formule da applicare sono le (13) e

(14):

$$B = \frac{\sum_{i=1}^{10} x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^{10} x_i^2 - n \bar{x}^2}$$

$$A = \bar{y} - B \bar{x}$$

ottenendo $B = 1.47$ e $A = 20.7$. Per il calcolo degli intervalli di confidenza ricaviamo innanzitutto dalle tavole i quantili di Student a 9 gradi di libertà: $t_{n-2}(0.05) = 1.833$ per l'intervallo al 90%, $t_{n-2}(0.025) = 2.262$ per l'intervallo al 95% e $t_{n-2}(0.005) = 3.250$ per l'intervallo al 99%. Applicando le formule (29) e (29) otteniamo:

intervallo al 90%: $a \in (19.87, 21.54)$ $b \in (1.33, 1.61)$

intervallo al 95%: $a \in (19.68, 21.74)$ $b \in (1.30, 1.64)$

intervallo al 99%: $a \in (19.23, 22.19)$ $b \in (1.22, 1.72)$

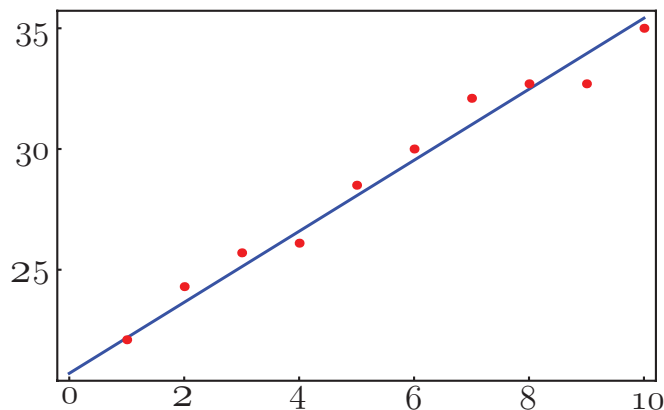


Figura 1: Retta di regressione per l'esempio nel testo.

