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 $x_1, x_2, ..., x_n$ 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 $F_{Y_1,...,Y_n}(y_1, ..., y_n)$ la distribuzione congiunta delle n variabili $Y_1, ..., Y_n$. Per le ipotesi fatte abbiamo

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

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

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

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

$$\frac{\partial}{\partial\alpha}\sum_{i=1}^{n}\frac{(y_i - \alpha - \beta\, x_i)^2}{\sigma^2} = 0 \tag{3}$$
$$\frac{\partial}{\partial\beta}\sum_{i=1}^{n}\frac{(y_i - \alpha - \beta\, x_i)^2}{\sigma^2} = 0. \tag{4}$$

Sviluppando i dettagli otteniamo:

$$\sum_{i=1}^{n}(y_i - \alpha - \beta\, x_i) = 0 \tag{5}$$
$$\sum_{i=1}^{n}x_i\,(y_i - \alpha - \beta\, x_i) = 0 \tag{6}$$

Introduciamo ora

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{7}$$

$$\overline{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{8}$$

Le equazioni (5) e (6) diventano così

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

cioè il sistema

$$\alpha + \overline{x}_n\,\beta = \overline{y}_n \tag{9}$$

$$n\,\overline{x}_n\,\alpha + \sum_{i=1}^{n} x_i^2\,\beta = \sum_{i=1}^{n} x_i\,y_i \tag{10}$$

L'equazione (9) fornisce

$$\alpha = \overline{y}_n - \overline{x}_n\,\beta; \tag{11}$$

introducendo per comodità la notazione

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

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

$$\begin{aligned} & n\,\overline{x}_n\,(\overline{y}_n - \overline{x}_n\,\beta) + \sum_{i=1}^{n} x_i^2\,\beta = \sum_{i=1}^{n} x_i\,y_i \\ & \Delta\,\beta = \sum_{i=1}^{n} x_i\,y_i - n\,\overline{x}_n\,\overline{y}_n \\ & \beta = \frac{\sum_{i=1}^{n} x_i\,y_i - n\,\overline{x}_n\,\overline{y}_n}{\Delta} \end{aligned}$$

Gli stimatori per α e β sono pertanto

$$\widehat{\beta} = B \equiv \frac{1}{\Delta} \left(\sum_{i=1}^{n} x_i\,Y_i - n\,\overline{x}_n\,\overline{Y}_n \right) = \frac{1}{\Delta} \sum_{i=1}^{n} (x_i - \overline{x}_n)\,Y_i \tag{13}$$

$$\widehat{\alpha} = A \equiv \overline{Y}_n - \overline{x}_n\,B \tag{14}$$

dove

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{15}$$

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

$$E[Y_i] = \alpha + \beta\, x_i \tag{16}$$

e quindi

$$E[\overline{Y}_n] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] = \alpha + \beta\, \overline{x}_n \tag{17}$$

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

$$\begin{aligned}
E[\widehat{\beta}] &= \frac{1}{\Delta}\left(\sum_{i=1}^{n} x_i\, E[Y_i] - n\,\overline{x}_n\, E[\overline{Y}_n]\right) = \\
&= \frac{1}{\Delta}\left[\sum_{i=1}^{n} x_i\,(\alpha + \beta\, x_i) - n\,\overline{x}_n\,(\alpha + \beta\,\overline{x}_n)\right] = \frac{1}{\Delta}\,\beta\left(\sum_{i=1}^{n} x_i^2 - n\,\overline{x}_n^2\right) = \beta \qquad (18)\\
E[\widehat{\alpha}] &= E[\overline{Y}_n] - \overline{x}_n\, E[\widehat{\beta}] = \alpha + \beta\,\overline{x}_n - \overline{x}_n\,\beta = \alpha \qquad (19)
\end{aligned}$$

Gli stimatori A e B dati dalle equazioni (13) e (14) sono dunque corretti.
Calcoliamo ora le varianze. Notiamo, preliminarmente, che $\overline{Y}_n$ e $\widehat{\beta}$ sono scorrelati:

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

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

Passando alle varianze:

$$\begin{aligned}
Var(\widehat{\beta}) &= Var\left(\frac{1}{\Delta}\sum_{i=1}^{n}(x_i - \overline{x}_n)\, Y_i\right) = \frac{1}{\Delta^2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\sigma^2 = \frac{\sigma^2}{\Delta} \qquad (21)\\
Var(\widehat{\alpha}) &= Var(\overline{Y}_n - \overline{x}_n\,\widehat{\beta}) = Var(\overline{Y}_n) + \overline{x}_n^2\, Var(\widehat{\beta}) = \frac{\sigma^2}{n} + \overline{x}_n^2\,\frac{\sigma^2}{\Delta} = \\
&= \frac{\sigma^2}{n\,\Delta}\,(\Delta + n\,\overline{x}_n^2) = \frac{\sum_{i=1}^{n} x_i^2}{n\,\Delta}\,\sigma^2 \qquad (22)
\end{aligned}$$

Possiamo pertanto concludere che

$$\begin{aligned}
\widehat{\alpha} &\sim N\left(\alpha, \frac{\sum_{i=1}^{n} x_i^2}{n\,\Delta}\,\sigma^2\right) \qquad (23)\\
\widehat{\beta} &\sim N\left(\beta, \frac{\sigma^2}{\Delta}\right) \qquad (24)
\end{aligned}$$

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 $\widehat{\alpha}$ e $\widehat{\beta}$ perdiamo due gradi di libertà, il che, in analogia con quanto succede per la varianza campionaria, rende plausibile l'affermazione che

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

Abbiamo quindi

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

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

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

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

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

ovvero

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

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

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

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

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

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. Ipotizzando 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 $y = A + B\,x$ e le formule da applicare sono le (13) e

(14):

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

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

Problema. I dati seguenti mettono in relazione la percentuale di acqua x, contenuta in un certo

materiale in una delle fasi di lavorazione, con la densità Y del prodotto finito:

acqua	**densità**
5	7.4
6	9.3
7	10.6
10	15.4
12	18.1
15	22.2
18	24.1
29	24.8

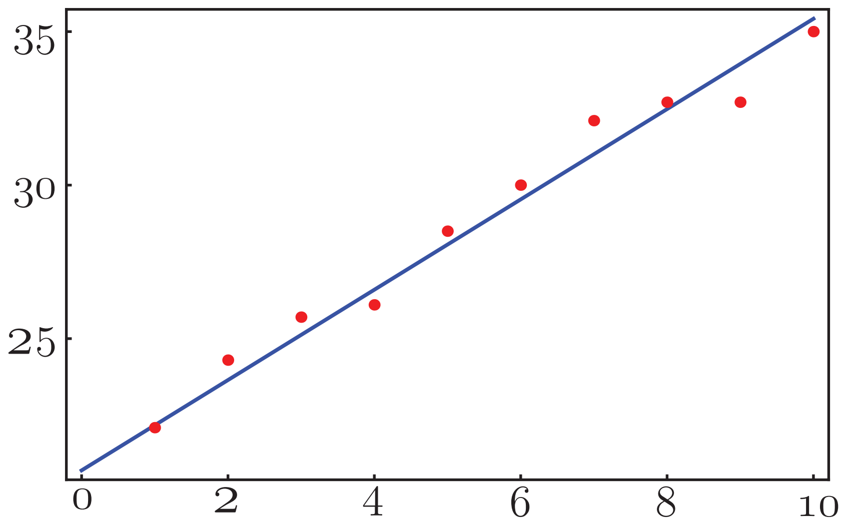

Figura 1: Retta di regressione per l'esempio nel testo.

Determinare gli intervalli di confidenza per i parametri della retta di regressione. noindent **Soluzione.** La retta di regressione ha equazione $y = A + B\,x$ e le formule da applicare sono le

(13) e (14):

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

ottenendo $B = 0.772$ e $A = 6.642$. Per il calcolo degli intervalli di confidenza ricaviamo innanzitutto dalle tavole i quantili di Student a 6 gradi di libertà: $t_{n-2}(0.05) = 1.943$ per l'intervallo al 90%, $t_{n-2}(0.025) = 2.447$ per l'intervallo al 95% e $t_{n-2}(0.005) = 3.707$ per l'intervallo al 99%. Applicando le formule (29) e (29) otteniamo:

intervallo al 90%: $a \in (2.06, 11.22)$ $b \in (0.46, 1.08)$
intervallo al 95%: $a \in (0.87, 12.41)$ $b \in (0.38, 1.16)$
intervallo al 99%: $a \in (-2.10, 15.39)$ $b \in (0.18, 1.36)$

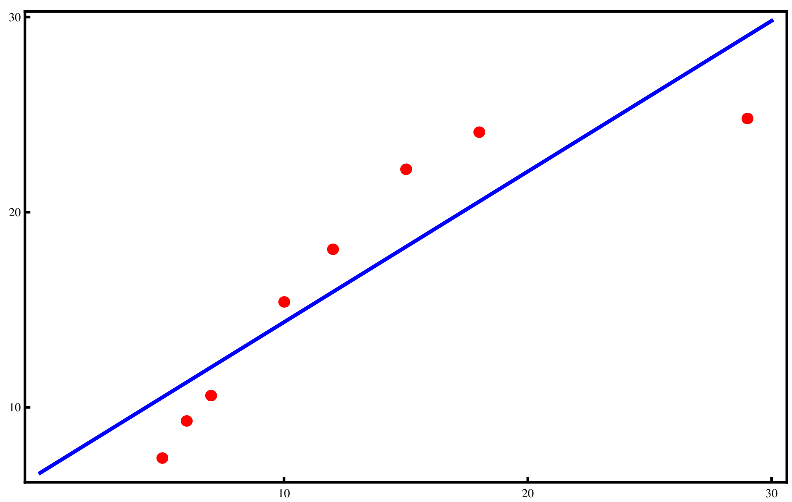

Figura 2: Retta di regressione per l'esempio nel testo.