Regressione lineare semplice

Esercizio 1

Si vuole stimare la relazione esistente tra il reddito mensile pro-capite X (in migliaia di Euro) e la spesa mensile per il tempo libero Y (in Euro). Da un campione casuale di 8 individui si sono ottenuti i seguenti risultati:

X	1.23	1.50	3.25	2.50	1.80	2.22	0.55	4.55
Y	120	270	330	230	130	220	100	500

Supponendo che le variabili X e Y siano legate tra loro da una relazione lineare del tipo

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

determinare:

- la varianza residua
- l'intervallo di confidenza al 95% per β_1
- un test d'ipotesi con $\alpha = 0.05$ per verificare l'ipotesi nulla $\beta_1 = 0$ stabilendo accettazione o rifiuto
- il coefficiente di determinazione lineare.

Soluzione

Le quantità che saranno utilizzate sono:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3}$$

$$\sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{4}$$

$$\sigma_{X,Y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) \tag{5}$$

$$= \mu_{x,y} - \bar{x} \cdot \bar{y} \quad \text{con} \quad \mu_{x,y} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i \tag{6}$$

$$\beta_1 = \frac{\sigma_{X,Y}}{\sigma_X^2} \tag{7}$$

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x} \tag{8}$$

Da cui si ottengono,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X \tag{9}$$

$$D_{res} = \sum_{i=1}^{n} e_i^2 \quad \text{con} \quad e_i = y_i - \hat{y}_i \tag{10}$$

$$\sigma_\epsilon^2 = S^2 = \frac{D_{res}}{n-2} \tag{11}$$

$$R^2 = \left(\frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y}\right)^2 = \frac{D_{sp}}{D_{tot}} = 1 - \frac{D_{res}}{D_{tot}} \tag{12}$$

In virtù delle formule da (1) a (12) si ricava $\bar{x} = 2.2$, $\bar{y} = 237.5$, $\sigma_X^2 = 1.3786$, $\sigma_Y^2 = 15393.75$, $\sigma_{X,Y} = 134.0626$, $\beta_1 = 97.2454$, $\beta_0 = 23.5601$, $\sigma_\epsilon^2 = 3142.3858$.

Per ricavare l'intervallo di confidenza per l'ignoto parametro β_1 occorre

ricordare che la statisitca

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{n \cdot \sigma_x^2}}} \sim T \; di \; Student(n-k-1)$$

e pertanto, il quantile di riferimento è $t_* = |t_{\frac{\alpha}{2},(n-k-1)}|$. Dai dati a disposizione, per $\alpha = 0.05$ si ricava che $t_* = 2.4469$. Ne consegue:

$$IC^- = \hat{\beta}_1 - t_* \cdot \sqrt{\frac{S^2}{n \cdot \sigma_x^2}} = 55.9422$$

$$IC^+ = \hat{\beta}_1 + t_* \cdot \sqrt{\frac{S^2}{n \cdot \sigma_x^2}} = 138.5486$$

In relazione al test d'ipotesi richiesto, $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, la statistica test diventa:

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{S^2}{n \cdot \sigma_x^2}}} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{S^2}{n \cdot \sigma_x^2}}} = \frac{\hat{\beta}_1}{\sqrt{\frac{S^2}{n \cdot \sigma_x^2}}} = 5.7611 = t$$

Dato che $|t| > t_*$ rifiuto H_0. Alla stessa conclusione si sarebbe giunti utilizzando il p-value. Dato che si tratta di un test bilaterale:

$$p - value = 2 \cdot P(T > t) = 0.0012 < \alpha$$

Infine, in relazione al coefficiente di determinazione lineare, in virtù della (12) si ricava $R^2 = 0.8469$. Una volta presa pratica:

- Matlab: scatter(x,y); tbl=table(x,y); mod=fitlm(tbl); plot(mod)

- Excel: REGR.LIN()

Esercizio 2

Un allevatore è interessato a sapere se l'età (espressa in mesi) delle proprie galline influisce sul peso (espresso in grammi) delle uova deposte dalle stesse. A tal fine, egli sceglie casualmente 6 galline e registra i seguenti risultati:

Età	31	30	35	40	45	50
Peso	54	58	60	63	63	65

Calcolare il coefficiente di correlazione.

Soluzione

In base a quanto riportato dal testo, si deduce; Y = Peso (variabile dipendente) e X = Età (variabile esplicativa). In virtù della (3) e della (6) e dai dati a disposizione si ottiene: $\sigma_X^2 = 52.9167$, $\sigma_Y^2 = 13.5833$ e $\sigma_{X,Y} = 23.9167$. In virtù della (12) si ricava infine $R^2 = 0.7958$.