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 $R^2 = 0.7958$.

Esercizio 3

Sia $\rho(X, Y)$ il coefficiente di correlazione fra X ed Y. Dire come varia tale quantità se al posto di X si pone $u = \frac{x}{2}$ e si considera $\rho(U, Y)$.

Soluzione

Ricordando la (12) e dai dati a disposizione si ha:

$$
\begin{aligned}
E(U) &= E\left(\frac{X}{2}\right) = \frac{1}{2}E(X) = \frac{\bar{x}}{2} = \bar{u} \\
\sigma_U^2 &= V(U) = V\left(\frac{X}{2}\right) = \frac{1}{4}V(X) = \frac{1}{4}\sigma_X^2 \\
\sigma_{U,Y} &= \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(y_i - \bar{y}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{2} - \frac{\bar{x}}{2}\right)(y_i - \bar{y}) \\
&= \frac{1}{2}\,\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{2}\sigma_{X,Y}.
\end{aligned}
$$

Si può concludere dunque:

$$\rho(U,Y) = \frac{\sigma_{U,Y}}{\sigma_U \cdot \sigma_Y} = \frac{\frac{\sigma_{X,Y}}{2}}{\frac{\sigma_X \cdot \sigma_Y}{2}} = \rho(X,Y)$$

Esercizio 4

Un rivenditore di bibite registra per 9 giorni sia i valori delle vendite y che quelli della temperatura esterna x. Con questi dati calcola i coefficienti ai minimi quadrati del modello $\hat{y} = a_9 + b_9 x$ e il coefficiente di determinazione R_9^2. Il decimo giorno, la temperatura esterna risulta uguale alla media dei primi 9 giorni. Similmente le vendite del decimo giorno risultano essere uguali alla media del venduto nei primi 9 giorni. Considerando le quantità $S_{xx} = \sum(x - \bar{x})^2$, $S_{yy} = \sum(y - \bar{y})^2$, $S_{xy} = \sum(x - \bar{x})(y - \bar{y})$ esplicitare la relazione tra R_9^2 e R_{10}^2.

Soluzione

Sia $\bar{x}_9 = \frac{1}{9}(x_1 + \ldots + x_9)$ la media della variabile temperatura registrata nei 9 giorni. Dai dati a disposizione è noto che $x_{10} = \bar{x}_9$ e che $y_{10} = \bar{y}_9$. Dato che:

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{m} x_i + \sum_{i=m+1}^{n} x_i.$$

posto $m = 9$ e $n = 10$:

$$\begin{aligned}
\bar{x}_{10} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\
&= \frac{1}{n} \left(\sum_{i=1}^{m} x_i + \sum_{i=m+1}^{n} x_i \right) \\
&= \frac{1}{10} \left(\sum_{i=1}^{9} x_i + \sum_{i=10}^{10} x_i \right) = \frac{1}{10} \left(\sum_{i=1}^{9} x_i + x_{10} \right) \\
&= \frac{1}{10}(9 \cdot \bar{x}_9 + \bar{x}_9) = \bar{x}_9
\end{aligned}$$

In modo del tutto analogo si ricava $\bar{y}_{10} = \bar{y}_9$. Passando al calcolo della devianza D_x si ha:

$$\begin{aligned} D_x &= \sum_{i=1}^{10}(x_i - \bar{x})^2 \\ D_{x_{10}} &= \left[(x_1 - \bar{x}_9)^2 + \ldots + (x_9 - \bar{x}_9)^2 + (x_{10} - \bar{x}_9)^2\right] \\ &= \left[(x_1 - \bar{x}_9)^2 + \ldots + (x_9 - \bar{x}_9)^2 + (\bar{x}_9 - \bar{x}_9)^2\right] = D_{x_9} \end{aligned}$$

Anche in questo caso, in modo del tutto analogo, si ricava $D_{y_{10}} = D_{y_9}$. In relazione alla codevianza:

$$\begin{aligned} D_{xy_{10}} &= \sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) \\ &= (x_1 - \bar{x}_{10})(y_1 - \bar{y}_{10}) + \cdots + (x_{10} - \bar{x}_{10})(y_{10} - \bar{y}_{10}) \\ &= (x_1 - \bar{x}_{10})(y_1 - \bar{y}_{10}) + \cdots + 0 \cdot 0 = D_{xy_9} \end{aligned}$$

In virtù del risultato precedente segue che $a_9 = a_{10}$ e $b_9 = b_{10}$. Per valutare R^2_{10} occorre ora calcolare il decimo residuo e_{10}. In virtù dei risultati precedenti e per semplicità di notazione poniamo $a_{10} = a_9 = a$, $b_{10} = b_9 = b$, $\bar{x}_{10} = \bar{x}_9 = \bar{x}$ e $\bar{y}_{10} = \bar{y}_9 = \bar{y}$. Pertanto, ricordando che $y_{10} = a + bx_{10} + e_{10}$:

$$\begin{aligned} e_{10} &= y_{10} - (a + b\bar{x}) \\ &= \bar{y} - (a + b\bar{x}) \\ &= \bar{y} - a - b\bar{x} \\ &= -a + \bar{y} - b\bar{x} = a - a = 0, \end{aligned}$$

segue che $D_{res_9} = D_{res_{10}}$ e di conseguenza:

$$R^2_{10} = 1 - \frac{D_{res_{10}}}{D_{y_{10}}} = R^2_{10} = 1 - \frac{D_{res_9}}{D_{y_9}} = R^2_9$$

Esercizio 5

Dato un modello di regressione lineare del tipo $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, cosa succede ai coefficienti β_0 e β_1 se i valori Y_i raddoppiano?

Soluzione

Posto $Z = 2Y$ è noto che $\bar{z} = 2\bar{y}$. Pertanto:

$$\begin{aligned}\sigma_{XZ} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(2y_i - 2\bar{y}) \\ &= 2\sigma_{XY}\end{aligned}$$

Ne consegue che,

$$\begin{aligned}\beta_1^* &= \frac{\sigma_{XZ}}{\sigma_X^2} = \frac{2\sigma_{XY}}{\sigma_X^2} = 2\beta_1 \\ \beta_0^* &= \bar{z} - \beta_1^* x_i = 2\bar{y} - 2\beta_1 x_i = 2(\bar{y} - \beta_1 x_i) = 2\beta_0\end{aligned}$$

e pertanto si può concludere che entrambi i coefficienti raddoppiano.