Esercitazione 8 del corso di Statistica 2

Prof. Domenico Vistocco

Dott.ssa Paola Costantini

17 Giugno 2008

Esercizio

Si ha motivo di ritenere che un nuovo farmaco A abbia la proprietà di abbassare il livello di glicemia nel sangue. In ciascuno dei 6 pazienti diabetici osservati, con lo stesso livello di glicemia iniziale, in trattamento con A da diversi periodi di tempo, sono stai rilevati la diminuzione di glicemia conseguita DG e il periodo di tempo T, in giorni, di durata del trattamento.
Si vuole spiegare, mediante un modello di regressione lineare, l'abbassamento del livello di glicemia (Y) in funzione del tempo (X).

Soluzione

Tabella di calcolo:

DG (Y)	T (X)	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
15	12	-6	-7	42	49	36
20	15	-1	-4	4	16	1
18	15	-3	-4	12	16	9
25	21	4	2	8	4	16
22	23	1	4	4	16	1
26	28	5	9	45	81	25
126	**114**			**115**	**182**	**88**

$$\mu_x = \frac{114}{6} = 19 \qquad \mu_y = \frac{126}{6} = 21 \quad \sum_{1=1}^{n} x_i^2 = 2348 \qquad \sum_{1=1}^{n} y_i^2 = 2734$$

$$\sigma^2_x = VAR(X) = E[X^2] - \mu^2 = \frac{2348}{6} - 19^2 = 30{,}34 \rightarrow \sigma = \sqrt{30{,}34} = 5{,}5$$

$$\sigma^2_y = VAR(YX) = E[Y^2] - \mu^2 = \frac{2734}{6} - 21^2 = 14{,}66 \rightarrow \sigma = \sqrt{14{,}66} = 3{,}83$$

$$\mu(x \cdot y) = (15 \cdot 12 + 20 \cdot 15 + \ldots\ldots + 26 \cdot 28/6) = 2509/5 = 418{,}167$$

$$Cov_{x,y} = \mu(x \cdot y) - (\mu_x \cdot \mu_y) = 418{,}167 - (19 \cdot 21) = 19{,}167$$

$$Corr_{x,y} = \rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \cdot \sigma_y} = \frac{19{,}167}{5{,}5 \cdot 3{,}83} = 0{,}91 \quad \rho^2 = 0{,}91^2 = 0{,}826$$

STIMA DELLA RETTA DI REGRESSIONE

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{\beta}_1 = \frac{Cov_{x,y}}{\sigma^2_x} = \frac{19{,}167}{30{,}34} = 0{,}632$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 21 - (0{,}632 \cdot 19) = 9$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i = 9 + 0{,}632 \cdot x_i$$

Quando si costruisce un modello di regressione l'obiettivo è quello di spiegare le variazioni della variabile dipendente (Y) mediante le variazioni della variabile esplicativa (X). Maggiore è la percentuale della varianza della Y che si riesce a spiegare con la variabile X, più soddisfacente sarà il modello. L'informazione della percentuale della varianza di Y spiegata dal modello di regressione è fornita dall'indice di determinazione R^2, che varia tra 0 e 1. Esso è dato dal rapporto tra devianza spiegata e devianza totale del modello.

$$R^2 = \frac{DevSpiegata}{DevTotale} = \frac{D(R)}{D(Y)}$$

$$DevTotale = D(Y) = \sum (y_i - \overline{y})^2 = 88$$

$$Dev\,\mathrm{Re}sidua = D(E) = \sum (y_i - \overline{y})^2 - (1 - \rho^2) = 88(1 - 0{,}826) = 15{,}312$$

$$DevSpiegata = D(R) = D(Y) - D(E) = 88 - 15{,}312 = 72{,}688$$

$$R^2 = \frac{D(R)}{D(Y)} = \frac{72{,}688}{88} = 0{,}826$$

Dal risultato di può notare come $R^2 = \rho^2$.

Inoltre, è possibile ottenere $R^2 = 1 - \frac{D(E)}{D(Y)} = 1 - \frac{15{,}132}{88} = 0{,}826$

Costruire l'intervallo di confidenza per βcon un livello di confidenza al 95%.

In aggiunta a $\hat{\beta}_0$ e $\hat{\beta}_1$ nel modello di regressione è necessario stimare un altro parametro, s^2. Se gli errori fossero osservabili, sarebbe ragionevole stimare σ^2, mediante la media campionaria degli errori al quadrato. Ma siccome gli errori non sono osservabili, perché non conosciamo i parametri α e β, allora è possibile calcolare i residui. Uno stimatore non distorto di σ^2, è dato da:

$$s^2 = \frac{D(E)}{n-2} = \frac{15{,}132}{4} = 3{,}828 \rightarrow \quad s = 1{,}96$$

I.C.

$$P\left(\hat{\beta}_1 - t_{4,0,025} \cdot \frac{s}{\sqrt{(x_i - \bar{x})}} \leq \beta \leq \hat{\beta}_1 + t_{4,0,025} \cdot \frac{s}{\sqrt{(x_i - \bar{x})}} \right) = 0{,}95$$

$$P\left(0{,}632_1 - 2{,}776 \cdot \frac{1{,}96}{\sqrt{182}} \leq \beta \leq 0{,}632_1 + 2{,}776 \cdot \frac{1{,}96}{\sqrt{182}} \right) = 0{,}95$$

$$P(0{,}229 \leq \beta \leq 1{,}035) = 0{,}95$$

Costruire un test sulla pendenza (β), coefficiente che misura l'effetto che una variazione unitaria della variabile esplicativa X produce sulla variabile dipendente Y.

Di regola, dopo aver stimato il modello di regressione si sottopone ad ipotesi nulla:

$$H_0 = \beta = 0$$

In tal caso (β=0), il valore atteso della Y è costante e pari a $\hat{\beta}_0$ per qualsiasi valore di X. Ciò implica l'assenza di un legame in media tra Y ed X e di conseguenza il modello di regressione è inutile.

Supponiamo che la mia ipotesi alternativa sia:

$$H_0 = \beta \neq 0$$

La statistica test è data da:

$$t = \frac{\hat{\beta}_1}{s/\sqrt{(x_i - \bar{x})^2}} = \frac{0{,}632}{1{,}96/\sqrt{182}} = \frac{0{,}632}{0{,}145} = 4{,}35$$

Essendo $t_{n-2,\alpha/2} = t_{4,0,025} = 2{,}776$, respingiamo H_0 e concludiamo che la somministrazione del nuovo farmaco effettivamente incide sul livello di glicemia.

Ora costruiamo un test sul coefficiente di correlazione ρ.

Sottoponiamo a test le ipotesi:

$H_0 : \rho_{x,y} = 0$ vs $H_1 : \rho_{x,y} = 0$

Sotto l'ipotesi nulla la statistica test

$$T_n = \frac{R_{x,y}}{\sqrt{(1 - R^2)/(n - 2)}}$$

Ha una distribuzione *t* di Student con n-2 gradi di libertà e il valore osservato della statistica è dato da:

$$t_n = \frac{\rho_{x,y}}{\sqrt{(1 - \rho^2)/(n - 2)}} = \frac{0{,}91}{\sqrt{(1 - 0{,}826)/4}} = \frac{0{,}91}{0{,}208} = 4{,}375$$

Per n=6 la statistica test si distribuisce come una v.c t_4. Al livello di significatività α=0,05, il valore critico risulta $t_{4,0,025} = 2{,}776$

Anche in questo caso rifiutiamo H_0 e concludiamo che c'è correlazione tra il tempo di somministrazione del nuovo farmaco e il livello di glicemia nel sangue.