

# Esercitazione 7 del corso di Statistica 2

Dott.ssa Paola Costantini

13 Marzo 2012

## Esercizio 1

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 stati 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
<b>126</b>	<b>114</b>			<b>115</b>	<b>182</b>	<b>88</b>

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

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

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

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

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

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

## STIMA DELLA RETTA DI REGRESSIONE

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

$$\hat{\beta}_1 = \frac{\text{Cov}_{x,y}}{\sigma_x^2} = \frac{19,167}{30,34} = 0,632$$

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

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

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.

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

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

$$\text{DevResidua} = D(E) = \sum (y_i - \hat{y})^2 = (1 - \rho^2) D(Y) = 88(1 - 0,826) = 15,312$$

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

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

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

$$\text{Inoltre, è possibile ottenere } R^2 = 1 - \frac{D(E)}{D(Y)} = 1 - \frac{15,312}{88} = 0,826$$

Costruire l'intervallo di confidenza per  $\beta$  con 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  $\sigma^2$ , mediante la media campionaria degli

errori al quadrato. Ma siccome gli errori non sono osservabili, perché non conosciamo i parametri  $\alpha$  e  $\beta$ , allora è possibile calcolare i residui. Uno stimatore non distorto di  $\sigma^2$ , è dato da:

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

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 ( $\beta$ ), 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 ( $\beta=0$ ), 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 :**  $t = \frac{\hat{\beta}_1}{s / \sqrt{(x_i - \bar{x})^2}}$

$$\text{Valore test} = \frac{0,632}{1,96 / \sqrt{182}} = \frac{0,632}{0,145} = 4,35$$

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 regressione  $R^2$ .

Sottoponiamo a test le ipotesi:

$$H_0 : R^2 = 0 \quad \text{vs} \quad H_1 : R^2 > 0$$

Sotto l'ipotesi nulla la statistica test

$$X_{test} = \frac{R^2(n-2)}{(1-R^2)}$$

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

$$\frac{0,826(4)}{1-0,826} = \frac{3,304}{0,174} = 18,98$$

Per  $n=6$  la statistica test si distribuisce come una v.c  $F_{\alpha, 1, n-2}$ . Al livello di significatività  $\alpha=0,05$ , il valore critico risulta  $F_{0,05, 1, 4} = 7,71$

Anche in questo caso rifiutiamo  $H_0$  e concludiamo che esiste una relazione lineare tra il tempo di somministrazione del nuovo farmaco e il livello di glicemia nel sangue.

## Esercizio 2

Data la seguente tabella, determinare in quale misura i caratteri PESO e ALTEZZA della seguente distribuzione doppia sono tra loro correlati.

Altezza (x) \ Peso (y)	160 -  164	164 -  170	170 -  178	178 -  186	Totale
46 -  56	5	1	1	0	<b>7</b>
56 -  66	0	2	0	3	<b>5</b>
66 -  76	0	2	1	2	<b>5</b>
76 -  86	0	0	1	2	<b>3</b>
<b>Totale</b>	<b>5</b>	<b>5</b>	<b>3</b>	<b>7</b>	<b>20</b>

Valori centrali per l'altezza:  $x_1 = 162$ ;  $x_2 = 167$ ;  $x_3 = 174$ ;  $x_4 = 182$   
 Valori centrali per il peso:  $y_1 = 51$ ;  $y_2 = 61$ ;  $y_3 = 71$ ;  $y_4 = 81$

Le quantità  $\hat{x}_i \hat{y}_j n_{ij}$  sono raccolte nella tabella che segue:

$\hat{x}_i \hat{y}_j n_{ij}$	<b>162</b>	<b>167</b>	<b>174</b>	<b>182</b>
<b>51</b>	41.310	8.517	8.874	0
<b>61</b>	0	20.374	0	33.306
<b>71</b>	0	23.714	12.354	25.844
<b>81</b>	0	0	14.094	29.484

**Totale generale: 217.871**

Da cui deriva:

$$\mu_{XY} = \frac{\sum_{i=1}^k \sum_{j=1}^h x_i y_j n_{ij}}{n} = \frac{217.871}{20} = 10.894$$

Le altre quantità necessarie sono contenute nella seconda tabella:

$x_i$	$n_i$	$y_i$	$n_j$	$x_i n_i$	$y_j n_j$	$x_i^2$	$x_i^2 n_i$	$y_j^2$	$y_j^2 n_j$
162	5	51	7	810	357	2601	131220	26244	18207
167	5	61	5	835	305	3721	139445	27889	18605
174	3	71	5	522	355	5041	90828	30276	25205
182	7	81	3	1274	243	6561	231868	33124	19683
<b>Totali</b>	<b>20</b>		<b>20</b>	<b>3.441</b>	<b>1.260</b>		<b>593.361</b>		<b>81.700</b>

$$\mu_X = \frac{1}{n} \sum_{i=1}^k x_i n_i = \frac{3.441}{20} = 172,05 \quad \text{altezza media}$$

$$\mu_Y = \frac{1}{n} \sum_{j=1}^h y_j n_j = \frac{1260}{20} = 63 \quad \text{peso medio}$$

$$\sum_{i=1}^n x_i^2 \cdot n_i = 593.361 \quad \sum_{i=1}^n y_i^2 \cdot n_i = 81.700$$

$$\sigma^2_x = \text{VAR}(X) = E[X^2] - \mu^2 = \frac{593.361}{20} - 172,05^2 = 67,05 \rightarrow \sigma = \sqrt{67,05} = 8,18$$

$$\sigma^2_y = \text{VAR}(Y) = E[Y^2] - \mu^2 = \frac{81.700}{20} - 63^2 = 116 \rightarrow \sigma = \sqrt{116} = 10,78$$

Sostituendo i valori ottenuti nella formula:

$$\text{cov}(x, y) = \mu_{xy} - \mu_x \mu_y = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^h \hat{x}_i \hat{y}_j n_{ij} - \mu_x \mu_y = 10.894 - 172,05 \times 63 = 54,85$$

$$\text{Corr}_{x,y} = \rho_{x,y} = \frac{\text{Cov}_{x,y}}{\sigma_x \cdot \sigma_y} = \frac{54,85}{8,18 \cdot 10,78} = \frac{54,85}{88,18} = 0,62 \quad \text{Correlazione positiva}$$

$$\rho^2_{x,y} = R^2 = 0,62^2 = 0,38$$

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

$$\text{DevTotale} = D(Y) = \sum (y_i - \bar{y})^2 * n_j = 1008 + 20 + 320 + 972 = 2320$$

$$\text{DevResidua} = D(E) = \sum (y_i - \bar{y})^2 * n_j * (1 - \rho^2) = 2320(1 - 0,38) = 1438,4$$

$$\text{DevSpiegata} = D(R) = D(Y) - D(E) = 2320 - 1438,4 = 881,6$$

STIMA DELLA RETTA DI REGRESSIONE

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

$$\hat{\beta}_1 = \frac{\text{Cov}_{x,y}}{\sigma^2_x} = \frac{54,85}{67,05} = 0,818$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 63 - (0,818 \cdot 172,05) = -77,7$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i = -77,79 + 0,818 \cdot x_i$$

Costruire un test sulla pendenza ( $\beta$ ), 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 ( $\beta=0$ ), 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 :** 
$$t = \frac{\hat{\beta}_1}{s / \sqrt{(x_i - \bar{x})^2 * ni}}$$

**Valore test** 
$$= \frac{0,818}{8,9 / \sqrt{1336,95}} = \frac{0,818}{8,9 / 36,56} = \frac{0,818}{0,243} = 3,36$$

dove

$$s^2 = \frac{D(E)}{n-2} = \frac{1438,4}{18} = 79,9 \rightarrow s = 8,9$$

Essendo  $t_{n-2, \alpha/2} = t_{18, 0,025} = 2,101$ , rifiutiamo  $H_0$  e concludiamo il peso incide sull'altezza.

Ora costruiamo un test sul coefficiente di regressione  $R^2$ .

Sottoponiamo a test le ipotesi:

$$H_0 : R^2 = 0 \quad \text{vs} \quad H_1 : R^2 > 0$$

Sotto l'ipotesi nulla la statistica test

$$X_{test} = \frac{R^2(n-2)}{(1-R^2)}$$

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

$$\frac{0,38(18)}{1-0,38} = \frac{6,84}{0,174} = 11$$

Per  $n=20$  la statistica test si distribuisce come una v.c  $F_{\alpha, 1, n-2}$ . Al livello di significatività  $\alpha=0,05$ , il valore critico risulta  $F_{0,05, 1, 18} = 4,41$

In questo caso rifiutiamo  $H_0$  e concludiamo che esiste una relazione lineare tra il peso e l'altezza.