Esercitazione 6 del corso di Statistica (parte 1)

Dott.ssa Paola Costantini

8 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 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

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

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

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

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

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

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

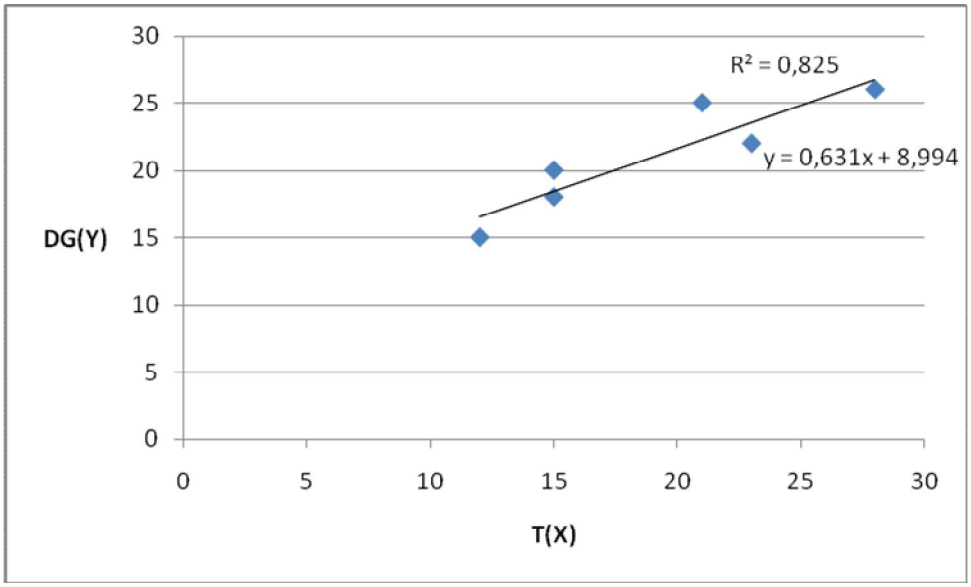
STIMA DELLA RETTA DI REGRESSIONE

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

$\hat{\beta}_1 = \frac{Cov_{x,y}}{\sigma^2{}_x} = \frac{19{,}167}{30{,}34} = 0{,}631$

$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 21 - (0{,}631 \cdot 19) = 8{,}9$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i = 8{,}9 + 0{,}632 \cdot x_i$



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 R2, che varia tra 0 e 1. Esso è dato dal rapporto tra devianza spiegata e devianza totale del modello.

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

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

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

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

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

Dal risultato di può notare come R2 = ρ2 .

Inoltre, è possibile ottenere R2 = $1 - \frac{D(E)}{D(Y)} = 1 - \frac{15{,}132}{88} = 0{,}826$

Esercizio 2

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

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

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:

$\hat{x}_i \hat{y}_j n_{ij}$	**162**	**167**	**174**	**182**
51	41.310	8.517	8.874	0
61	0	20.374	0	33.306
71	0	23.714	12.354	25.844
81	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
Totali	**20**		**20**	**3.441**	**1.260**		**593.361**		**81.700**

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

$$\mu_Y = \frac{1}{n}\sum_{j=1}^{h} y_j n_{.j} = \frac{1260}{20} = 63$$ peso medio

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

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

$$\sigma^2{}_y = VAR(YX) = E[Y^2] - \mu^2 = \frac{81.700}{20} - 63^2 = 116 \rightarrow \sigma = \sqrt{18,06} = 10,78$$

Sostituendo i valori ottenuti nella formula:

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

$$Corr_{x,y} = \rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \cdot \sigma_y} = \frac{54,85}{8,18 \cdot 10,78} = \frac{54,85}{88,18} = 0,62$$ **Correlazione positiva**

$$\rho_{x,y}^2 = R^2 = 0,63^2 = 0,38$$

Tale valore va confrontato con l'intervallo [-1, 1], quindi indica una correlazione lineare positiva abbastanza forte.

Esercizio 3

Considerando gli eventi elemenatri associati alla seguente distribuzione doppia:

Attività sportiva / Corso laurea	Nulla (N)	Media (M)	Alta (A)	**Totale**
Biologia (B)	0	1	0	**1**
Informatica (I)	4	7	1	**12**
Matematica (Mat)	2	5	0	**7**
Totale	**6**	**13**	**1**	**20**

determinare:

a) le probabilità elementari

Inoltre si determinino le probabilità che, scegliendo uno studente a caso:

b) sia iscritto in matematica e pratichi attività sportiva media;
c) sia iscritto in biologia e pratichi attività sportiva alta;
d) sia iscritto in informatica o pratichi attività sportiva nulla;
e) essendo iscritto in informatica, pratichi attività sportiva nulla;
f) praticando attività sportiva media, sia iscritto in biologia.

a)
gli eventi elementari sono riferiti in entrambi gli spazi alla selezione casuale di uno dei 20 studenti.

Gli eventi elementari dello spazio Ω_1 sono:

B = “studente iscritto al CDL in biologia”
I = “studente iscritto al CDLin informatica”
Mat = “studente iscritto al CDL in matematica”

Gli eventi elementari dello spazio Ω_2 sono:

N = “attività sportiva nulla”
M = “attività sportiva media”
A = “attività sportiva alta”

Le probabilità elementari si ottengono come frequenze relative marginali della tabella.

$$P(B) = \frac{1}{20} = 0,05 \qquad P(N) = \frac{6}{20} = 0,3$$

$$P(I) = \frac{12}{20} = 0,6 \qquad P(M) = \frac{13}{20} = 0,65$$

$$P(Mat) = \frac{7}{20} = 0,35 \qquad P(Mat) = \frac{1}{20} = 0,05$$

Naturalmente la somma delle probabilità in ciascuno dei due spazi campionari è pari ad 1:

$$\sum_i P(E_i) = 0,05 + 0,6 + 0,35 =$$
$$= 0,3 + 0,65 + 0,05 =$$
$$= 1$$

b) sia "iscritto in matematica" *e* "pratichi attività sportiva media"

Si tratta della probabilità dell'intersezione dei due eventi:

$$P(\text{Mat} \cap \text{M}) = \frac{5}{20} = 0,25$$

c) sia "iscritto in biologia" *e* "pratichi attività sportiva alta"

Si tratta della probabilità dell'intersezione dei due eventi:

$$P(\text{B} \cap \text{A}) = \frac{0}{20} = 0$$

d) sia "iscritto in informatica" *o* "pratichi attività sportiva nulla"

Si tratta della probabilità dell'unione dei due eventi:

$$P(\text{I} \cup \text{N}) = P(\text{I}) + P(\text{N}) - P(\text{I} \cap \text{N}) = \frac{12}{20} + \frac{6}{20} - \frac{4}{20} = \frac{14}{20} = 0,7$$

e) *essendo iscritto in informatica*, pratichi attività sportiva nulla:

Si tratta di una probabilità condizionata, dove l'iscrizione in informatica è l'evento condizionante, quindi lo spazio campione di riferimento si riduce a quello costituito dai soli studenti iscritti in informatica.

$$P(\text{N} \mid \text{I}) = \frac{P(\text{N} \cap \text{I})}{P(\text{I})} = \frac{4/20}{12/20} = \frac{4}{12} = 0,\overline{3}$$

f) *praticando attività sportiva media*, sia iscritto in biologia:

Si tratta di una probabilità condizionata, dove il praticare attività sportiva media è l'evento condizionante, quindi lo spazio campione di riferimento si riduce a quello costituito dai soli che praticano attività sportiva media.

$$P(B \mid M) = \frac{P(B \cap M)}{P(M)} = \frac{1/20}{13/20} = \frac{1}{13} = 0,769$$

Esercizio n. 4

Estraendo due carte da un mazzo di carte napoletane con la reimmissione della carta nel mazzo dopo ciascuna prova (estrazione con ripetizione, gli eventi sono indipendenti in quanto, di prova in prova, il mazzo resta immutato), calcolare la probabilità che si presentino, nell'ordine,

1. Asso (A) e figura (F): $P(A \cap F) = P(A) \cdot P(F)$

2. Una carta di coppe (C) e una carta di denari (D): $P(C \cap D) = P(C) \cdot P(D)$

3. Un asso, una figura, un cinque ("5"): $P(A \cap F \cap "5") = P(A) \cdot P(F) \cdot P("5")$

Estraendo invece due carte dal mazzo, senza rimettere la prima carta estratta (estrazione senza ripetizione, il secondo evento non è indipendente dal primo), le stesse probabilità valgono:

4. Asso (A) e figura (F): $P(A \cap F) = P(A) * P(F|A)$

5. Una carta di coppe (C) e una carta di denari (D): $P(C \cap D) = P(C) \cdot P(D|C)$

6. Un asso, una figura, un cinque ("5"): $P(A \cap F \cap "5") = P(A) \cdot P(F|A) \cdot P("5"|A \cap F)$

Svolgimento

1. $P(A \cap F) = P(A) \cdot P(F) = \frac{4}{40} \cdot \frac{12}{40} = \frac{48}{1600}$

2. $P(C \cap D) = P(C) \cdot P(D) = \frac{10}{40} \cdot \frac{10}{40} = \frac{100}{1600}$

3. $P(A \cap F \cap "5") = P(A) \cdot P(F) \cdot P("5") = \frac{4}{40} \cdot \frac{12}{40} \cdot \frac{4}{40} = \frac{192}{64000}$

4. $P(A \cap F) = P(A) * P(F|A) = \frac{4}{40} \cdot \frac{12}{39} = \frac{48}{1560}$

5. $P(C \cap D) = P(C) \cdot P(D|C) = \frac{10}{40} \cdot \frac{10}{39} = \frac{100}{1560}$

6. $P(A \cap F \cap \text{"5"}) = P(A) \cdot P(F|A) \cdot P(\text{"5"}|A \cap F) = \frac{4}{40} \cdot \frac{12}{39} \cdot \frac{4}{38} = \frac{192}{59280}$