Università degli Studi di Cassino, Anno accademico 2004-2005
Corso di Statistica 2, Prof. M. Furno

Esercitazione del 23/03/2005
dott. Claudio Conversano

Esercizio 1

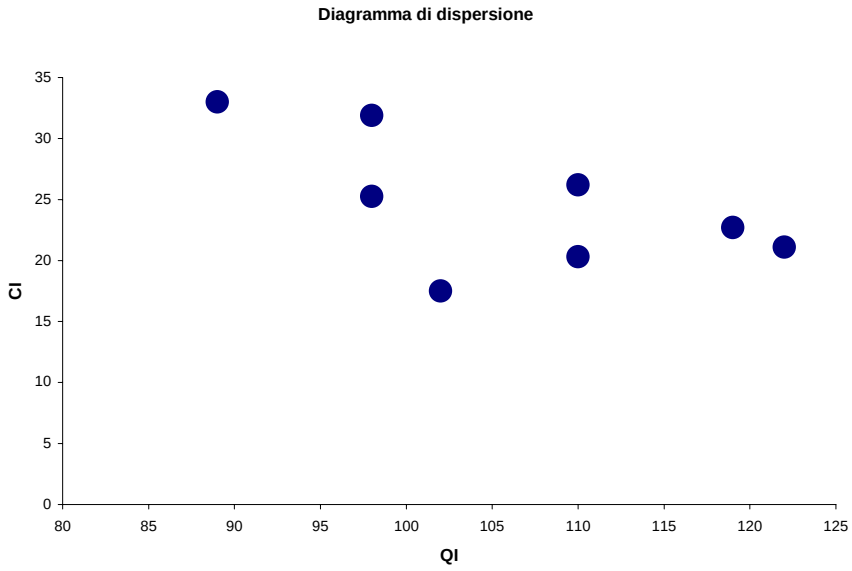
Un sociologo viene assunto in un carcere minorile per studiare se esiste relazione tra intelligenza e crimine. A tale scopo costruisce un indice di criminalità (IC) che tiene conto sia della gravità che della frequenza dei reati commessi dai giovani, mentre l'intelligenza viene misurata con un indice del quoziente di intelligenza (QI) ottenuto da alcuni test a cui si sottopone un campione di 8 detenuti. I risultati sono i seguenti:

(IC)	26.2	33.0	17.5	25.25	20.3	31.9	21.1	22.7
(QI)	110	89	102	98	110	98	122	119

a) Stimare i parametri della retta di regressione che misura la dipendenza dei crimini commessi dal quoziente intellettivo.
b) Dopo aver verificato la significatività dell'intercetta della retta, stimare l'intervallo di confidenza per α al 10%.
c) Si valuti l'ipotesi che non esista relazione tra le due variabili e si costruisca l'intervallo di confidenza per il coefficiente angolare fissando un livello di fiducia α=0.05.
d) Calcolare l'indice di determinazione lineare e verificare che esso è significativamente diverso da zero.
e) Calcolare il coefficiente di correlazione e verificare che esso è significativamente diverso da zero.

SVOLGIMENTO

a)



(QI) x_i	(IC) y_i	x_i^2	$x_i\, y_i$	y_i^2
110	26.20	12100	2882.0	686.44
89	33.00	7921	2937.0	1089.00
102	17.50	10404	1785.0	306.25
98	25.25	9604	2474.5	637.56
110	20.30	12100	2233.0	412.09
98	31.90	9604	3126.2	1017.61
122	21.10	14884	2574.2	445.21
119	22.70	14161	2701.3	515.29
848	**197.95**	**90778**	**20713.2**	**5109.45**

$$\hat{\beta} = \frac{Cov(X,Y)}{Var(X)} = \frac{\frac{20713.2}{8} - \left(\frac{848}{8} \cdot \frac{197.95}{8}\right)}{\frac{90778}{8} - \left(\frac{848}{8}\right)^2} = -0.30$$

$$\hat{\alpha} = E(Y) - \hat{\beta}_1 E(X) = \frac{197.95}{8} + 0.30(\frac{848}{8}) = 56.84$$

b)

(IC) y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	$e_i^2 = \left(y_i - \hat{y}_i\right)^2$
26.20	23.53	2.67	7.12
33.00	29.89	3.11	9.66
17.50	25.95	-8.45	71.49
25.25	27.17	-1.92	3.67
20.30	23.53	-3.23	10.45
31.90	27.17	4.73	22.41
21.10	19.90	1.20	1.44
22.70	20.81	1.89	3.58
			129.82

$$S^2_{(e)} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{129.82}{6} = 21.64$$

$$H_0 : \alpha = 0; \qquad H_1 : \alpha \neq 0$$

$$\hat{\alpha} \approx N\left(\alpha, \frac{\sigma^2}{n} \frac{\mu_{X^2}}{Var(X)}\right)$$

Statistica Test:

$$t = \frac{|\hat{\alpha}|}{\sqrt{Var(\hat{\alpha})}} = \frac{|\hat{\alpha}|}{\sqrt{\frac{S^2_{(e)}}{n}\frac{E(X^2)}{Var(X)}}}$$

Regione di rifiuto:

$$t > t_{(n-2;1-\alpha/2)}$$

$$t = \frac{56.84}{\sqrt{\frac{21.64}{8}\cdot\frac{\left(\frac{90778}{8}\right)^2}{\left[\frac{90778}{8}-\left(\frac{848}{8}\right)^2\right]}}} = 3.42$$

$3.42 > t_{(6;0.975)} = 2.447$ Rifiuto H_0.

L'intervallo di confidenza al 10% per α risulta tale che:

$$P\left(-t_{(n-2;\alpha/2)} < \frac{\hat{\alpha}-\alpha}{\sqrt{Var(\hat{\alpha})}} < t_{(n-2;1-\alpha/2)}\right) = P\left(-t_{(n-2;\alpha/2)} < \frac{\hat{\alpha}-\alpha}{\sqrt{\frac{S^2_{(e)}}{n}\cdot\frac{E(X^2)}{Var(X)}}} < t_{(n-2;1-\alpha/2)}\right) = 1-\alpha$$

L'intervallo è dato da:

$$\left[\hat{\alpha} \pm t_{(n-2;1-\alpha/2)}\cdot\sqrt{\frac{S^2_{(e)}}{n}\cdot\frac{E(X^2)}{Var(X)}}\right] = \left[56.84 \pm 1.943\cdot\sqrt{\frac{21.64}{8}\cdot\frac{\left(\frac{90778}{8}\right)^2}{\left[\frac{90778}{8}-\left(\frac{848}{8}\right)^2\right]}}\right] = [24.57;89.11]$$

c)

$H_0 : \beta = 0$; $H_1 : \beta \neq 0$

$$\hat{\beta} \approx N\left(\beta, \frac{\sigma^2}{n}\frac{1}{Var(X)}\right)$$

Statistica Test:

$$t = \frac{\left|\hat{\beta}\right|}{\sqrt{Var\left(\hat{\beta}\right)}} = \frac{\left|\hat{\beta}\right|}{\sqrt{\frac{S^2_{(e)}}{n}\frac{1}{Var\left(X\right)}}}$$

Regione di rifiuto:

$$t > t_{(n-2;1-\alpha/2)}$$

$$t = \frac{0.30}{\sqrt{\frac{21.64}{8}\cdot\frac{1}{\left[\frac{90778}{8}-\left(\frac{848}{8}\right)^2\right]}}} = 1.942$$

$1.942 < t_{(6;0.975)} = 2.447$ Accetto H_0.

L'intervallo di confidenza al 5% per β risulta tale che:

$$P\left(-t_{(n-2;\alpha/2)} < \frac{\hat{\beta}-\beta}{\sqrt{Var\left(\hat{\beta}\right)}} < t_{(n-2;1-\alpha/2)}\right) = P\left(-t_{(n-2;\alpha/2)} < \frac{\hat{\beta}-\beta}{\sqrt{\frac{S^2_{(e)}}{n}\cdot\frac{1}{Var\left(X\right)}}} < t_{(n-2;1-\alpha/2)}\right) = 1-\alpha$$

L'intervallo è dato da:

$$\left[\hat{\beta} \pm t_{(n-2;1-\alpha/2)}\cdot\sqrt{\frac{S^2_{(e)}}{n}\cdot\frac{1}{Var\left(X\right)}}\right] = \left[-0.30 \pm 2.447\cdot\sqrt{\frac{21.64}{8}\cdot\frac{1}{\left[\frac{90778}{8}-\left(\frac{848}{8}\right)^2\right]}}\right] = \left[-0.68; 0.08\right]$$

d)

$$R^2 = 1-\frac{Var\left(E\right)}{Var\left(Y\right)} = 1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \Big/ n}{\sum_{i=1}^{n}\left(y_i-E\left(Y\right)\right)^2 \Big/ n} = 1-\frac{\frac{129.82}{8}}{26.42} = 0.386$$

$H_0 : R^2 = 0; \quad H_1 : R^2 \neq 0$

Statistica Test

$$F = \frac{\hat{\beta}^2 n Var(X)}{S^2_{(e)}} \approx F_{[1,(n-2)]}$$

Regione di rifiuto:

$$F > F_{[1,(n-2);1-\alpha]}$$

$$F = \frac{-0.30^2 \cdot 8 \cdot \left[\frac{90778}{8} - \left(\frac{848}{8}\right)^2\right]}{21.64} = 3.77$$

$3.77 < F_{[1,(8-2);0.95]} = 5.99$ Accetto H_0.

e)

$$Corr(XY) = \frac{Cov(X,Y)}{Var(X)Var(Y)} = \frac{\frac{20713.2}{8} - \left(\frac{848}{8} \cdot \frac{197.95}{8}\right)}{\left[\frac{5109.45}{8} - \left(\frac{197.95}{8}\right)^2\right]\left[\frac{90778}{8} - \left(\frac{848}{8}\right)^2\right]} = -0.01$$

$H_0 : \rho = 0; \quad H_1 : \rho \neq 0$

Statistica Test

$$t = \frac{|Corr(X,Y)|}{\sqrt{\frac{1 - Corr(X,Y)^2}{n-2}}} \approx t_{(n-2;1-\alpha/2)}$$

Regione di rifiuto:

$$t > t_{(n-2;1-\alpha/2)}$$

$$t = \frac{0.01}{\sqrt{\frac{1-0.01^2}{6}}} = 0.02807$$

$0.02807 < t_{(6;0.95)} = 1.943$ Accetto H_0.

Esercizio 2
La distribuzione di un campione di appartamenti in affitto a Cassino, in base alla superficie in mq ed al canone mensile in Euro è la seguente:

	Canone mensile				
Mq	200|-300	300|-400	400|-500	500|-600	Totale
40|-60	20	16	4	0	40
60|-80	24	92	22	2	140
80|-120	0	32	50	18	100
120|-180	0	0	6	14	20
Totale	44	140	82	34	300

Determinare i parametri della retta di regressione (costi – superficie) e calcolare il coefficiente di correlazione lineare.

SVOLGIMENTO

Per una distribuzione doppia i parametri della retta di regressione sono:

$$\hat{\beta} = \frac{Cov(X,Y)}{Var(X)} = \frac{E(X,Y) - E(X)E(Y)}{E(X^2) - [E(X)]^2} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{h} \hat{x}_i \hat{y}_j p(\hat{x}_i \hat{y}_j) - \left(\sum_{i=1}^{k} \hat{x}_i p(x_i) \cdot \sum_{j=1}^{h} \hat{y}_j p(\hat{y}_j)\right)}{\sum_{i=1}^{k} \hat{x}_i^2 p(\hat{x}_i) - \left(\sum_{i=1}^{k} \hat{x}_i p(\hat{x}_i)\right)^2}$$

$$\hat{\alpha} = E(Y) - \hat{\beta}E(X) = \sum_{j=1}^{h} \hat{y}_j p(\hat{y}_j) - \hat{\beta}\sum_{i=1}^{k} \hat{x}_i p(x_i)$$

Per calcolare il termine $\sum_{i=1}^{k}\sum_{j=1}^{h} \hat{x}_i \hat{y}_j p(\hat{x}_i \hat{y}_j)$ è consigliabile costruire la tabella delle $\hat{x}_i \hat{y}_j p(\hat{x}_i \hat{y}_j)$, ossia:

$\hat{x}_i$ \ $\hat{y}_i$	250	350	450	550
50	833	933	300	0
70	1400	7513	2310	257
100	0	3733	7500	3300
150	0	0	1350	3850

La somma degli elementi all'interno di tale tabella è pari a $\sum_{i=1}^{k}\sum_{j=1}^{h} \hat{x}_i \hat{y}_j p(\hat{x}_i \hat{y}_j) = 33280$ che corrisponde al momento misto $E(X,Y)$, primo termine della formula della covarianza.

Per il calcolo della covarianza, dei parametri della retta di regressione e della correlazione è utile considerare la seguente tabella:

$\hat{x}_i$	$p(\hat{x}_i)$	$\hat{y}_j$	$p(\hat{y}_j)$	$\hat{x}_i p(\hat{x}_i)$	$\hat{y}_j p(\hat{y}_j)$	$\hat{x}_i^2$	$\hat{x}_i^2 p(\hat{x}_i)$	$\hat{y}_j^2$	$\hat{y}_j^2 p(\hat{y}_j)$
50	0.13	250	0.15	6.67	36.67	2500	333.33	62500	9166.67
70	0.47	350	0.47	32.67	163.33	4900	2286.67	122500	57166.67
100	0.33	450	0.27	33.33	123.00	10000	3333.33	202500	55350.00
150	0.07	550	0.11	10.00	62.33	22500	1500.00	302500	34283.33
				82.67	**385.33**		**7453.33**		**155966.67**

$$Cov(XY) = 33280 - 82.67 \cdot 385.33 = 1425.78$$

$$Var(X) = 7453.33 - (82.67)^2 = 619.55$$

$$\hat{\beta} = \frac{1425.78}{619.55} = 2.30$$

$$\hat{\alpha} = 385.33 - 2.30 \cdot 82.67 = 195.09$$

Per una distribuzione doppia la correlazione è data da:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E(X,Y) - E(X)E(Y)}{\sqrt{E(X^2) - [E(X)]^2 \cdot E(Y^2) - [E(Y)]^2}} =$$

$$= \frac{\sum_{i=1}^{k}\sum_{j=1}^{h} \hat{x}_i \hat{y}_j p(\hat{x}_i \hat{y}_j) - \left(\sum_{i=1}^{k} \hat{x}_i p(x_i) \cdot \sum_{j=1}^{h} \hat{y}_j p(\hat{y}_j) \right)}{\sqrt{\left[\sum_{i=1}^{k} \hat{x}_i^2 p(\hat{x}_i) - \left(\sum_{i=1}^{k} \hat{x}_i p(\hat{x}_i) \right)^2 \right] \cdot \left[\sum_{i=1}^{h} \hat{y}_j^2 p(\hat{y}_j) - \left(\sum_{i=1}^{k} \hat{y}_j p(\hat{y}_j) \right)^2 \right]}}$$

Dai dati del campione risulta:

$$Corr(X,Y) = \frac{1425.78}{\sqrt{7453.33 \cdot (155966.67 - 385.33^2)}} = 0.66$$

Esiste correlazione diretta tra la superficie ed il canone mensile di locazione.