

Esercizio 8.21

In 7 quartieri di una città sono state rilevate le seguenti variabili:

X=tasso di violenza sui minori [X= numero di casi segnalati/popolazione $\times$ 100].

Y=il numero medio di anni di scuola per adulto,

Z=la percentuale di famiglie con entrambi i genitori.

X	Y	Z
1.2	12.1	90
0.7	12.2	86
3.5	9.2	80
6.7	11.1	75
5.8	8.5	65
4.2	11.8	76
3.8	10.5	67

a) Calcolare una misura per valutare se esiste una relazione tra

i) violenza sui minori e istruzione

ii) violenza sui minori e percentuale di famiglie con 2 genitori.

senza stabilire se, nelle due coppie di caratteri, uno sia da considerare antecedente e l'altro conseguente.

b) Determinare la retta di regressione in cui l'istruzione è la variabile indipendente e la violenza sui minori la variabile dipendente.

c) Determinare la retta di regressione in cui la percentuale di famiglie con 2 genitori è la variabile indipendente e la violenza sui minori è la variabile dipendente.

d) Dovendo scegliere di spiegare il fenomeno della violenza sui minori con una delle due altre variabili, quale scegliereste, e in base a quale indicatore?

Scopo dell'esercizio: date due possibili variabili esplicative per una variabile dipendente, ricavare le due rette di regressione e commentare la differenza nei risultati.

Risoluzione

I calcoli intermedi sono:

X_i	X_i^2	$X_i Y_i$	Y_i	Y_i^2	$Y_i Z_i$	Z_i	Z_i^2	$X_i Z_i$
1.2	1.44	14.52	12.1	146.41	1089.0	90	8100	108.0
0.7	0.49	8.54	12.2	148.84	1049.2	86	7396	60.2
3.5	12.25	32.20	9.2	84.64	736.0	80	6400	280.0
6.7	44.89	74.37	11.1	123.21	832.5	75	5625	502.5
5.8	33.64	49.30	8.5	72.25	552.5	65	4225	377.0
4.2	17.64	49.56	11.8	139.24	896.8	76	5776	319.2
3.8	14.44	39.90	10.5	110.25	703.5	67	4489	254.6

$$\sum_i X_i = 25.9 \quad \bar{X} = 3.7$$

$$\sum_i X_i^2 = 124.75$$

$$V(X) = \frac{124.75}{7} - 3.2^2 = 17.82 - 13.69 = 4.13$$

$$\sigma(X) = 2.03$$

$$\sum_i Y_i = 75.4 \quad \bar{Y} = 10.77$$

$$\sum_i Y_i^2 = 824.84$$

$$V(Y) = \frac{824.84}{7} - 10.77^2 = 117.83 - 115.99 = 1.84$$

$$\sigma(Y) = 1.36$$

$$\sum_i Z_i = 539 \quad \bar{Z} = 77$$

$$\sum_i Z_i^2 = 42011$$

$$V(Z) = \frac{42011}{7} - 77^2 = 6001.57 - 5929 = 72.57$$

$$\sigma(Z) = 8.52$$

$$\sum_i X_i Y_i = 268.39$$

$$\sum_i \frac{X_i Y_i}{7} = 38.34$$

$$\sum_i X_i Z_i = 1901.5$$

$$\sum_i \frac{X_i Z_i}{7} = 271.64$$

$$\sum_i Y_i Z_i = 5859.5$$

$$\sum_i \frac{Y_i Z_i}{7} = 337.07$$

a) Per valutare se esiste una relazione tra i caratteri senza stabilire quale sia da considerarsi antecedente e quale conseguente, si dovrà calcolare il coefficiente di correlazione lineare tra i due. Si avrà quindi:

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = \frac{\sum_i (X_i Y_i)/7 - \bar{X}\bar{Y}}{\sigma(X)\sigma(Y)} = \frac{38.34 - (3.7)(10.77)}{(2.03)(1.36)} = \frac{-1.51}{2.7608} = -0.546$$

Tra il tasso di violenza sui minori e l'istruzione vi è una correlazione negativa di media entità

$$r_{X,Z} = \frac{Cov(X,Z)}{\sigma(X)\sigma(Z)} = \frac{\sum_i (X_i Z_i)/7 - \bar{X}\bar{Z}}{\sigma(X)\sigma(Z)} = \frac{271.64 - (3.7)(77)}{(2.03)(8.52)} = \frac{-13.26}{17.29} = 0.767$$

Tra il tasso di violenza sui minori e la percentuale di famiglie con due genitori vi è una correlazione negativa più forte della precedente.

b) Si determinano ora i parametri della retta in cui il tasso di violenza sui minori dipende dall'istruzione.

$$b_{X|Y} = \frac{Cov(X,Y)}{V(Y)} = \frac{-1.51}{1.84} = -0.821$$

$$a_{X|Y} = \bar{X} - b_{X|Y}\bar{Y} = 3.7 + 0.821(10.77) = 3.7 + 8.66 = 12.542$$

$$X = 12.542 - 0.821Y$$

$$r_{X|Y}^2 = 0.298$$

La parte di variabilità complessiva di X spiegata dalla variabile Y in modo lineare è il 29.8%

c) Si determinano ora i parametri della retta in cui il tasso di violenza sui minori dipende dalla percentuale di famiglie con due genitori.

$$b_{X|Z} = \frac{Cov(X,Z)}{V(Z)} = \frac{-13.26}{72.57} = -0.183$$

$$a_{X|Z} = \bar{X} - b_{X|Z}\bar{Z} = 3.7 + (0.183)(77) = 17.791$$

$$X = 17.791 - 0.183Z$$

$$r_{X|Z}^2 = 0.588$$

La parte di variabilità complessiva di X spiegata dalla variabile Z in modo lineare è il 58.8%

d) La variabile "tasso di violenza sui minori" è spiegata meglio, in modo lineare, dalla variabile "percentuale delle famiglie con due genitori" che dalla variabile "livello medio di istruzione", essendo nel primo caso più alto il coefficiente di determinazione lineare r^2 .