Corso di Statistica - Esercitazione 4

Dott. Davide Buttarazzi

✉ *d.buttarazzi@unicas.it*

Esercizio 1

Si vuole studiare la relazione tra le variabili quantitative X ed Y. I dati in classi sono riportati nella seguente tabella, espressi in frequenze assolute congiunte.

		Y			
		[50-60]	**(60-70]**	**(70-80]**	$\mathbf{n_{i.}}$
X	**[10-20]**	2	4	20	26
	(20-30]	5	10	3	18
	(30-40]	7	6	5	18
	$\mathbf{n_{.j}}$	14	20	28	$\mathbf{n_{..} = 62}$

1. Calcolare l'indice di correlazione lineare ρ_{XY}

Soluzioni esercizio 1

1. L'indice di correlazione è definito come:

   $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

   Pertanto è fondamentale calcolare la covarianza σ_{XY}. Dividendo tale quantità per il prodotto delle deviazioni standard σ_X e σ_Y, si ottiene un indice (di correlazione) normalizzato tra -1 ed 1.

   Per dati organizzati in classi, dopo aver individuato i centri (15,25,35 per la variabile X, e 55,65,75 per la variabile Y), è possibile operare con la seguente formulazione della covarianza.

   $\sigma_{XY} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{x_i y_j n_{ij}}{n} - \mu_x \mu_y$

   Per comodità si consiglia di costruire la seguente tabella, utile a calcolare le quantità $x_i y_j n_{ij}$

		Y			
		55	**65**	**75**	Tot.
X	**15**	$15 \times 55 \times 2 = 1650$	$15 \times 65 \times 4 = 3900$	$15 \times 75 \times 20 = 22500$	
	25	$25 \times 55 \times 5 = 6875$	$25 \times 65 \times 10 = 16250$	$25 \times 75 \times 3 = 5625$	
	35	$35 \times 55 \times 7 = 13475$	$35 \times 65 \times 10 = 13650$	$35 \times 75 \times 5 = 13125$	
	Tot.				97050

   La componente della covarianza $\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{x_i y_j n_{ij}}{n}$ si ottiene quindi dividendo per $n = 62$ la somma di tutti gli elementi della tabella appena costruita.

   Non resta quindi che calcolare le medie marginali di X e di Y:

   $\mu_X = \frac{1}{n} \sum_{i=1}^{r} x_i n_{i.} = \frac{15 \times 26 + 25 \times 18 + 35 \times 18}{62} = 23.7$

   $\mu_Y = \frac{1}{n} \sum_{j=1}^{c} y_i n_{.j} = \frac{55 \times 14 + 65 \times 20 + 75 \times 28}{62} = 67.3$

È ora possibile calcolare la covarianza:

$\sigma_{XY} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{x_i y_j n_{ij}}{n} - (\mu_x\mu_y) = \frac{97050}{62} - 23.7 \times 67.3 = -29.7$

Per calcolare l'indice di correlazione lineare occorre calcolare le deviazioni standard per le variabili X ed Y come segue:

$\sigma_X = \sqrt{\frac{\sum_{i=1}^{r}(x_i-\mu_X)^2 n_{i.}}{n}} = \sqrt{\frac{(15-23.7)^2\times 26+(25-23.7)^2\times 18+(35-23.7)^2\times 18}{62}} = 8.39$

$\sigma_Y = \sqrt{\frac{\sum_{j=1}^{c}(y_j-\mu_Y)^2 n_{.j}}{n}} = \sqrt{\frac{(55-67.3)^2\times 14+(65-67.3)^2\times 20+(75-67.3)^2\times 28}{62}} = 7.97$

Avendo a disposizione tutti gli elementi, è possibile calcolare l'indice di correlazione lineare:

$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{-29.7}{8.39\times 7.97} = -0.44$, il quale segnala una moderata discordanza.

Esercizio 2

Si vuole studiare la relazione tra le variabili quantitative X ed Y al seguito riportate.

Y	**X**
32.1	24.3
33.2	25.7
33.6	27.4
37.1	29.3
42.1	31.4
41.1	33.7
44.7	35.6
44.6	36.8
47.8	38.1
48.2	40.1
51.3	42.5

1. Calcolare l'indice di correlazione lineare ρ_{XY}
2. Ipotizzando una dipendenza lineare di Y da X, calcolare i coefficienti di regressione β_0 e β_1 ed esplicitare l'equazione della retta di regressione.
3. Utilizzare i risultati appena ottenuti ai punti 1) e 2) per calcolare l'indice di bontà di adattamento R^2.

Soluzioni esercizio 2

1. L'indice di correlazione lineare è così definito:

   $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$

   Il numeratore di tale indice, la covarianza, può essere così definita:

   $\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)$

   oppure, equivalentemente:

   $\sigma_{XY} = \frac{\sum_{i=1}^{n} x_i y_j}{n} - \mu_x\mu_y$

   La seguente tabella fornisce tutte le quantità (colonne) necessarie per poter calcolare la covarianza utilizzando entrambe le definizioni riportate:

   $\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i = 33.17$

   $\mu_y = \frac{1}{n}\sum_{i=1}^{n} y_i = 41.44$

$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)(y_i - \mu_y)$	$x_i y_i$
-8.873	-9.336	82.839	780.03
-7.473	-8.236	61.548	853.24
-5.773	-7.836	45.237	920.64
-3.873	-4.336	16.794	1087.03
-1.773	0.664	-1.176	1321.94
0.527	-0.336	-0.177	1385.07
2.427	3.264	7.922	1591.32
3.627	3.164	11.475	1641.28
4.927	6.364	31.355	1821.18
6.927	6.764	46.854	1932.82
9.327	9.864	92.001	2180.25
		=394.67	=15514.8

Si avrà quindi:

$\sigma_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y) = \frac{394.67}{11} = 35.87$

oppure, equivalentemente:

$\sigma_{XY} = \frac{\sum_{i=1}^{n} x_i y_j}{n} - \mu_x \mu_y = \frac{15514.8}{11} - 33.17 \times 41.44 = 35.87$

Per poter calcolare l'indice di correlazione lineare, occorre individuare le deviazioni standard per le variabili X ed Y:

$(x_i - \mu_x)^2$	$(y_i - \mu_y)^2$
78.725	87.168
55.842	67.838
33.324	61.409
14.998	18.804
3.143	0.440
0.278	0.113
5.892	10.651
13.157	10.009
24.278	40.496
47.987	45.747
86.998	97.291
=364.62	=439.97

$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2}{n}} = 5.757$

$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \mu_y)^2}{n}} = 6.324$

Quindi:

$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{35.87}{5.757 \times 6.324} = 0.98$, il quale segnala una forte concordanza.

2. I coefficienti di regressione possono essere così calcolati:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_x)}{\sum_{i=1}^{n}(x_i-\mu_x)^2}$$

Il numeratore di β_1 è la codevianza, mentre il denominatore di β_1 è la devianza di X. Entrambe le quantità sono state calcolate al punto precedente. Si avrà quindi:

$\beta_1 = \frac{394.67}{364.62} = 1.08$

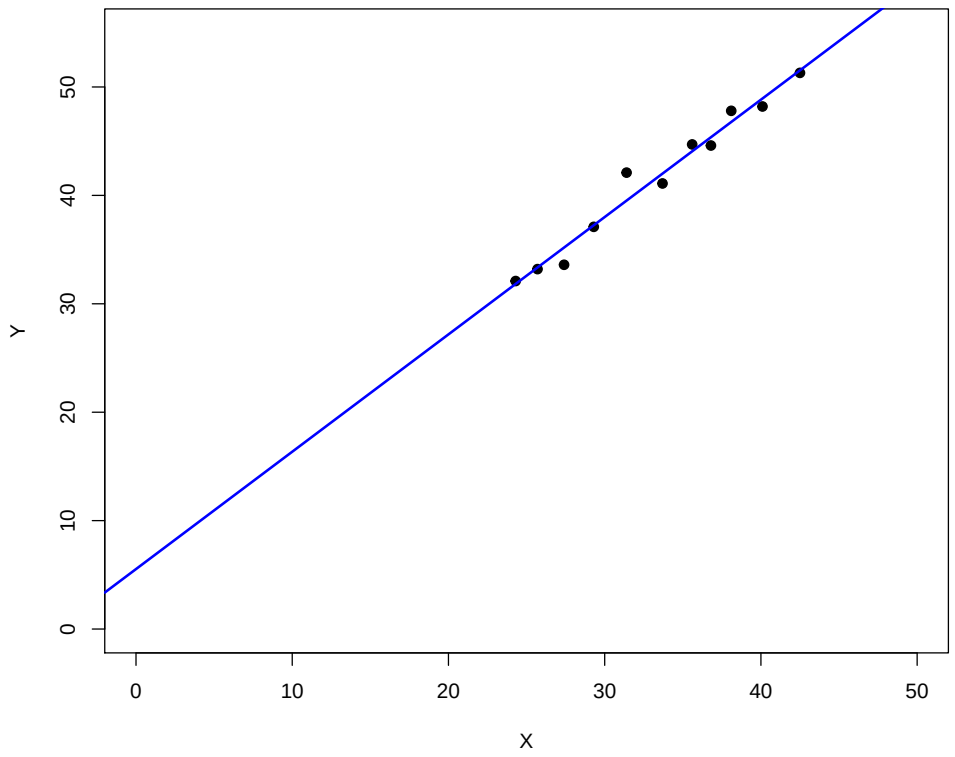
Avendo ottenuto β_1, è possibile ricavare facilmente β_0:

$\beta_0 = \mu_y - \beta_1\mu_x = 41.44 - (1.0824 \times 33.17) = 5.53$

L'equazione della retta di regressione è quindi:

$\hat{y_i} = 5.53 + 1.08x_i$

Graficamente:



3. Utilizzando i risultati ottenuti al punto 1), è possibile calcolare l'indice di bontà di adattamento sfruttando la relazione:

   $R^2 = \rho_{xy}^2 = 0.98^2 = 0.9604$

   Utilizzando invece i risultati ottenuti al punto 2), è possibile sfruttare le proprietà di decomposizione della devianza ed utilizzare, ad esempio, le seguenti relazioni:

   $R^2 = \frac{Dev(R)}{Dev(Y)}$

   oppure,

   $R^2 = \beta_1^2 \frac{Dev(X)}{Dev(Y)}$

$\hat{y_i} = 5.53 + 1.08x_i$	$Dev(R) = \sum_{i=1}^{n}(\hat{y_i} - \mu_y)^2$	$Dev(Y) = \sum_{i=1}^{n}(y_i - \mu_y)^2$
31.77	93.36	87.17
33.29	66.43	67.84
35.12	39.87	61.41
37.17	18.17	18.80
39.44	3.98	0.44
41.93	0.24	0.11
43.98	6.46	10.65
45.27	14.73	10.01
46.68	27.47	40.50
48.84	54.78	45.75
51.43	99.87	97.29
	=425.36	=439.97

Si avrà quindi:

$R^2 = \frac{Dev(R)}{Dev(Y)} = \frac{425.36}{439.97} = 0.96$

o, alternativamente, ricordando $Dev(X) = \sum_{i=1}^{n}(x_i - \mu_x)^2 = 364.62$ calcolata al punto 1):

$R^2 = \beta_1^2 \frac{Dev(X)}{Dev(Y)} = 1.08^2 \times \frac{364.62}{439.97} = 0.96$

Nota per assenti o non frequentanti: durante l'esercitazione 4 sono stati discusse le diverse formule a disposizione per il calcolo di covarianza/correlazione, indice di bontà di adattamento e loro interpretazioni. Nell'ambito dell'analisi di regressione è stata riesaminata con attenzione la relazione $DEV(Y) = DEV(Err) + DEV(Reg)$.