Esercitazione 10

ESERCIZIO 1

Marco e Giulio hanno due negozi in viale dei Giardini. Marco vende libri, Giulio vende elettronica, tra cui tablet. Marco e Giulio, avendo a disposizione il numero di libri venduti ed il numero di tablet venduti per anno dal 2008 al 2015, si sono chiesti se esiste una relazione tra le due quantità.

Libri	*Tablet*
1284	47
971	62
1123	75
1047	69
921	103
874	113
889	136

1. Calcolare la covarianza con il metodo diretto e con il metodo indiretto. Confrontare i due risultati.
2. Stimare la retta di regressione, considerando X numero di libri e Y numero di tablet, commentare i coefficienti ottenuti.
3. Calcolare i residui di regressione.
4. Calcolare l'indice di determinazione lineare.
5. Calcolare l'indice di correlazione lineare, direttamente ed utilizzando l'indice di determinazione lineare. Confrontare i risultati ottenuti.

SOLUZIONE

x_i	y_i	x_i^2	y_i^2	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$x_i y_i$	r_i	r_i^2
1284	47	1648656	2209	268,429	-39,429	-10583,755	60348	7,236	52,363
971	62	942841	3844	-44,571	-24,429	1088,816	60202	-32,177	1035,364
1123	75	1261129	5625	107,429	-11,429	-1227,755	84225	7,247	52,523
1047	69	1096209	4761	31,429	-17,429	-547,755	72243	-11,965	143,159
921	103	848241	10609	-94,571	16,571	-1567,184	94863	0,131	0,017
874	113	763876	12769	-141,571	26,571	-3761,755	98762	1,960	3,842
889	136	790321	18496	-126,571	49,571	-6274,327	120904	27,568	759,977

Per la variabile X abbiamo

$$M_1^X = \frac{1}{n}\sum_{i=1}^{n} x_i = 1015,571$$

$$M_2^X = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 1050181,857$$

$$\sigma_X^2 = M_2^X - \left[M_1^X\right]^2 = 18796,531$$

$$\sigma_X = \sqrt{\sigma_X^2} = 137,100$$

Per la variabile Y abbiamo

$$M_1^Y = \frac{1}{n}\sum_{i=1}^{n} y_i = 86,429$$

$$M_2^Y = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = 8330,429$$

$$\sigma_Y^2 = M_2^Y - \left[M_1^Y\right]^2 = 860,531$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = 29,335$$

Tramite il metodo diretto abbiamo che

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) = \frac{1}{7}\,(-22873,714) = -3267,673$$

Tramite il metodo indiretto

$$M_1^{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1}{7}\,(591547) = 84507$$

$$cov(X,Y) = M_1^{XY} - M_1^X M_1^Y = 84507 - (1015,571 \times 86,429) = -3267,673$$

La covarianza è negativa, quindi le due variabili sono inversamente proporzionali. Al diminuire di X ci aspettiamo un aumento di Y.
La retta hai minimi quadrati è la retta della forma

$$y = \alpha_0 + \alpha_1 \times x$$

dove α_0 e α_1 minimizzano

$$\sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 \times x_i)$$

Possiamo ottenere quindi la stima ai minimi quadrati per α_0 e α_1 come

$$\hat{\alpha}_1 = \frac{cov(X,Y)}{Var(X)} = \frac{-3267,673}{18796,531} = -0,174$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \times \bar{x} = 86,429 + 0,174 \times 1015,571 = 262,980$$

e quindi il modello stimato è

$$y = 262,980 - 0,174 \times x$$

Per ogni libro in più venduto da Marco, Giulio vende 0,174 tablet in meno. Se Marco chiudesse il negozio, e non vende più libri, Giulio si aspetterebbe di vendere circa 263 tablet.
Possiamo calcolare i residui di regressione come

$$r_i = y_i - \hat{y}_i$$

dove $\hat{y}_i = \hat{\alpha}_0 - \hat{\alpha}_1 x_i$, i risultati sono nella tabella precedente.
Per valutare la bontà del modello possiamo utilizzare l'indice di determinazione

$$I_d^2 = \frac{Dev\,Spiegata}{Dev\,Totale} = 1 - \frac{Dev\,Residua}{Dev\,Totale} = 1 - \frac{Var\,Residua}{Var\,Totale}$$

La varianza totale è la varianza di Y calcolata precedentemente. La varianza residua è la varianza dei residui

$$\sigma^2_{RES} = M_2^{RES} - \left[M_1^{RES}\right]^2 = M_2^{RES} = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{1}{7}\left(2047,245\right) = 292,464$$

quindi

$$I_d^2 = 1 - \frac{Var\,Residua}{Var\,Totale} = 1 - \frac{292,464}{860,531} = 0,660$$

Con calcoli analoghi

$$\sigma^2_{SP} = \left[\hat{\alpha}_1\right]^2 \times \sigma_X^2 = \left(-0,174\right)^2 \times 18796,531 = 569,084$$

quindi

$$I_d^2 = \frac{Var\,Spiegata}{Var\,Totale} = \frac{569,084}{860,531} = 0,660$$

Il modello spiega il 66% della variabilità di Y.
Partendo dall'indice precedente, possiamo ottenere l'indice di correlazione lineare come

$$r = sign\left(cov\left(X,Y\right)\right) \times \sqrt{I_d^2} = -0,812$$

Si poteva altresì ottenere come

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{-3267,673}{137,1 \times 29,335} = -0,812$$

i due risultati sono equivalenti.

ESERCIZIO 2

Un centro di ricerca sito in via Marco polo effettua, tra le varie attività, ricerche sui tempi di reazione dei soggetti in studio. Nella seguente vengono riportati i tempi di reazione (ms) e le età degli individui

Età	*Tempo reazione*
8	280
11	212
14	167
20	122
26	183
30	201
37	267

1. Stimare la retta di regressione, considerando X età e Y tempo di reazione.
2. Calcolare i residui di regressione.
3. Mostrare che i residui di regressione hanno media nulla.
4. Calcolare l'indice di determinazione lineare.
5. Calcolare l'indice di correlazione lineare.

SOLUZIONE

x_i	y_i	x_i^2	y_i^2	$x_i y_i$	r_i	r_i^2
8	280	64	78400	2240	78,253	6123,539
11	212	121	44944	2332	9,594	92,045
14	167	196	27889	2338	-36,065	1300,687
20	122	400	14884	2440	-82,383	6786,980
26	183	676	33489	4758	-22,701	515,345
30	201	900	40401	6030	-5,580	31,136
37	267	1369	71289	9879	58,882	3467,124

Per la variabile X abbiamo

$$M_1^X = \frac{1}{n}\sum_{i=1}^{n} x_i = 20,857$$

$$M_2^X = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 532,286$$

$$\sigma_X^2 = M_2^X - \left[M_1^X\right]^2 = 97,265$$

$$\sigma_X = \sqrt{\sigma_X^2} = 9,862$$

Per la variabile Y abbiamo

$$M_1^Y = \frac{1}{n}\sum_{i=1}^{n} y_i = 204,571$$

$$M_2^Y = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = 44470,857$$

$$\sigma_Y^2 = M_2^Y - \left[M_1^Y\right]^2 = 2621,388$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = 51,199$$

La covarianza possiamo calcolarla come

$$M_1^{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1}{7}\left(30017\right) = 4288$$

$$cov(X,Y) = M_1^{XY} - M_1^X M_1^Y = 4288 - (20,857 \times 204,571) = 21,367$$

La covarianza è positiva, quindi le due variabili sono direttamente proporzionali. All'aumentare di X ci aspettiamo un aumento di Y.
La retta hai minimi quadrati è la retta della forma

$$y = \alpha_0 + \alpha_1 \times x$$

dove α_0 e α_1 minimizzano

$$\sum_{i=1}^{n} (y_i - \alpha_0 - \alpha_1 \times x_i)$$

Possiamo ottenere quindi la stima ai minimi quadrati per α_0 e α_1 come

$$\hat{\alpha}_1 = \frac{cov(X,Y)}{Var(X)} = \frac{21,367}{97,265} = 0,220$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \times \bar{x} = 204,571 - 0,220 \times 20,857 = 199,990$$

e quindi il modello stimato è

$$y = 199,99 + 0,22 \times x$$

Possiamo calcolare i residui di regressione come

$$r_i = y_i - \hat{y}_i$$

dove $\hat{y}_i = \hat{\alpha}_0 - \hat{\alpha}_1 x_i$, i risultati sono nella tabella precedente.
Possiamo vedere facilmente che i residui hanno media nulla

$$M_1^{RES} = \frac{1}{n} \sum_{i=1}^{n} r_i = \frac{1}{7} (0) = 0$$

Per valutare la bontà del modello possiamo utilizzare l'indice di determinazione

$$I_d^2 = \frac{Dev\,Spiegata}{Dev\,Totale} = 1 - \frac{Dev\,Residua}{Dev\,Totale} = 1 - \frac{Var\,Residua}{Var\,Totale}$$

La varianza totale è la varianza di Y calcolata precedentemente. La varianza residua è la varianza dei residui

$$\sigma_{RES}^2 = M_2^{RES} - \left[M_1^{RES}\right]^2 = M_2^{RES} = \frac{1}{n} \sum_{i=1}^{n} r_i^2 = \frac{1}{7} (18316,856) = 2616,694$$

quindi

$$I_d^2 = 1 - \frac{Var\,Residua}{Var\,Totale} = 1 - \frac{2616,694}{2621,338} = 0,002$$

Il modello spiega lo 0,2% della variabilità di Y.
Partendo dall'indice precedente, possiamo ottenere l'indice di correlazione lineare come

$$r = sign\left(cov\left(X,Y\right)\right) \times \sqrt{I_d^2} = 0,042$$

Si poteva altresì ottenere come

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{21,367}{9,862 \times 51,199} = 0,042$$

i due risultati sono equivalenti.

ESERCIZIO 3

Alessandra, un'allegra gelataia di una piccola gelateria in viale Traiano, ha deciso di mettere in relazione il numero di gelati venduti con la temperatura massima della giornata nel periodo estivo. Sia quindi X la temperatura massima della giornata, espressa in gradi gradi Celsius, ed Y il numero di gelati venduti.

Y\X	23 – |26	26 – |28	28 – |31
0 – 50	11	2	1
50 – 100	6	5	4
100 – 200	3	17	3
200 – 350	2	9	27

1. Stimare la retta di regressione, considerando X temperatura e Y numero di gelati.
2. Calcolare l'indice di determinazione lineare.
3. Calcolare l'indice di correlazione lineare.

SOLUZIONE

Utilizziamo i valori centrali delle classi. Per la variabile X abbiamo

i	x_i	n_i	$n_i x_i$	x_i^2	$n_i x_i^2$
1	24,5	22	539	600,25	13205,5
2	27	33	891	729	24057
3	29,5	35	1032,5	870,25	30458,75

Quindi

$$M_1^X = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = 27,361$$

$$M_2^X = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 = 752,458$$

$$\sigma_X^2 = M_2^X - \left[M_1^X\right]^2 = 3,828$$

$$\sigma_X = \sqrt{\sigma_X^2} = 1,957$$

Per la variabile Y abbiamo

j	y_j	n_j	$n_j y_j$	y_j^2	$n_j y_j^2$
1	25	14	350	625	8750
2	75	15	1125	5625	84375
3	150	23	3450	22500	517500
4	250	38	9500	62500	2375000

Quindi

$$M_1^Y = \frac{1}{n}\sum_{i=1}^{n} n_j y_i = 160,278$$

$$M_2^Y = \frac{1}{n}\sum_{j=1}^{n} n_j y_j^2 = 33173,611$$

$$\sigma_Y^2 = M_2^Y - \left[M_1^Y\right]^2 = 7484,645$$

$$\sigma_Y = \sqrt{\sigma_Y^2} = 86,514$$

Per il momento misto di ordine uno abbiamo

$n_{ji}x_iy_j$	24,5	27	29,5
25	6737,5	1350	737,5
75	11025	10125	8850
150	11025	68850	13275
250	12250	60750	199125

E possiamo calcolare

$$M_1^{XY} = \frac{1}{n}\sum_{i=1}^{k_X}\sum_{j=1}^{k_Y} n_{ji}x_iy_j = \frac{1}{90}\left(404110\right) = 4490$$

quindi

$$cov(X,Y) = M_1^{XY} - M_1^X M_1^Y = 4490 - \left(27,361 \times 160,278\right) = 104,622$$

La covarianza è positiva, quindi le due variabili sono direttamente proporzionali. All'aumentare di X ci aspettiamo un aumento di Y.
Possiamo ottenere quindi la stima ai minimi quadrati per α_0 e α_1 come

$$\hat{\alpha}_1 = \frac{cov(X,Y)}{Var(X)} = \frac{104,622}{3,828} = 27,331$$

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \times \bar{x} = 160,278 - 27,331 \times 27,361 = -587,534$$

La varianza spiegata può essere calcolata come

$$\sigma_{SP}^2 = \left[\hat{\alpha}_1\right]^2 \times \sigma_X^2 = 27,331^2 \times 3,828 = 2859,441$$

quindi l'indice di determinazione lineare diventa

$$I_d^2 = \frac{\sigma_{SP}^2}{\sigma_{TOT}^2} = \frac{28599,441}{7484,645} = 0,382$$

L'indice di correlazione lineare può essere ottenuto come

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{104,622}{1,957 \times 86,514} = 0,618$$