LM−88 SOCIOLOGIA E RICERCA SOCIALE

Metodi Statistici per la Ricerca Sociale

Regressione lineare e correlazione

1. Su un campione di individui sono rilevati i caratteri X (peso in Kg) e Y (altezza in cm), ottenendo i seguenti dati:

u_i	x_i	y_i
1	60	165
2	60	165
3	70	170
4	80	165
5	80	175

   (a) Calcolare il coefficiente di correlazione lineare tra peso e altezza

   (b) Sottoporre a test l'ipotesi che il coefficiente di correlazione lineare sia nullo al livello di significatività del 5%

   Soluzione:

   (a) $r_{XY} = 0.559$

   (b) Ipotesi: $H_0 : \rho_{XY} = 0$ versus $H_a : \rho_{XY} \neq 0$

   Regione critica: $RC_{0.05} = T < -3.18$ oppure $T > 3.18$.

   Valore osservato della statistica test: $T^{oss} = 1.168$

   Decisione: Non si può rifiutare H_0 al livello di significatività del 5%

2. Su un campione di $n = 10$ soggetti si osservano due caratteri X e Y. Si ha che

$$\bar{x} = 2.5 \qquad e \qquad \bar{y} = 6.2$$

   Inoltre

$$\sum_{i=1}^{10} x_i^2 = 81 \qquad \sum_{i=1}^{10} y_i^2 = 400 \qquad \sum_{i=1}^{10} x_i \cdot y_i = 166.6.$$

   (a) Calcolare il coefficiente di correlazione lineare tra X e Y.

   (b) Sottoporre a test l'ipotesi che il coefficiente di correlazione lineare sia nullo al livello di significatività del 10%

   Soluzione:

   (a) $r = 0.683$

   (b) Ipotesi: $H_0 : \rho_{XY} = 0$ versus $H_a : \rho_{XY} \neq 0$

   Regione critica: $RC_{0.1} = T < -1.86$ oppure $T > 1.86$.

   Valore osservato della statistica test: $T^{oss} = 2.644$

   Decisione: Si rifiuta H_0 al livello di significatività del 10%

3. Nella tabella sono riportati i tassi di attività lavorativa della popolazione, e i prodotti interni lordi per abitante (in migliaia di euro) per 8 regioni.

Regione	Tasso di attività	PIL per abitante
Piemonte	63	6.0
Lombardia	61	6.3
Liguria	55	6.2
Toscana	60	5.3
Emilia-Romagna	64	5.9
Lazio	53	4.6
Puglia	55	3.3
Sicilia	50	3.2

(a) Calcolare la covarianza e il coefficiente di correlazione lineare tra i due caratteri.

(b) Sottoporre a test l'ipotesi che il coefficiente di correlazione lineare sia nullo al livello di significatività del 5%

Soluzione:

(a) $s_{XY} = 4.729$ e $r_{XY} = 0.736$

(b) Ipotesi: $H_0 : \rho_{XY} = 0$ versus $H_a : \rho_{XY} \neq 0$

Regione critica: $RC_{0.05} = T < -2.447$ oppure $T > 2.447$.

Valore osservato della statistica test: $T^{oss} = 2.664$

Decisione: Si rifiuta H_0 al livello di significatività del 5%

---

4. L'archeopterix è un animale primitivo dotato di e piume, come gli uccelli, di denti e di una lunga coda ossea, come i rettili. Per questo esemplare sono stati ritrovati soltanto 5 reperti fossili. Dato che i fossili differiscono molto in grandezza, alcuni scienziati pensano che appartengano a specie diverse. Se i fossili appartenessero tutti alla stessa specie allora differirebbero tra di loro solo a causa delle diverse et' degli animali. In questo caso ci dovrebbe essere una relazione lineare positiva tra le lunghezze di una coppia di ossa rilevate nei vari reperti. L'assenza di relazione lineare potrebbe suggerire una specie diversa. Di seguito si riportano i dati delle lunghezze in centimetri del femore e dell'omero per i 5 reperti fossili di cui si sono ritrovate entrambe le ossa:

Femore	38	56	59	64	74
Omero	41	63	70	72	84

(a) Calcolare il coefficiente di correlazione

(b) Calcolare i valori standardizzati per il femore e l'omero. Calcolare la covarianza tra i valori standardizzati. Confrontare il risultato con il valore del coefficiente di correlazione calcolato al punto (a).

(c) Supponiamo che uno scienziato pazzo misuri il femore in metri e l'omero in millimetri. I dati allora diventerebbero:

Femore	0.38	0.56	0.59	0.64	0.74
Omero	410	630	700	720	840

Calcolare coefficiente di correlazione. Confrontare il risultato con il valore del coefficiente di correlazione calcolato al punto (a).

Soluzione:

(a) Coefficiente di correlazione: $r_{XY} = 0.994$.

(b) Covarianza tra valori standardizzati: $s_{Z_X,Z_Y} = 0.994 = r_{XY}$

(c) Il coefficiente di correlazione è ancora $r_{XY} = 0.994$ (il coefficiente di correlazione è un numero puro)

5. I seguenti dati si riferiscono alla lunghezza media (Y) e alla larghezza media (X) dei petali di 6 esemplari di un fiore:

Larghezza (X)	8	4	18	22	3	8
Lunghezza (Y)	4	2	8	6	5	6

(a) Stimare la retta di regressione che pone la lunghezza in funzione della larghezza

(b) La relazione lineare trovata spiega più del 70% della variabilità totale?

(c) Per un petalo di larghezza uguale a 7, qual è la lunghezza media prevista?

Soluzione:

(a) $\hat{\beta}_0 = 3.29104$ e $\hat{\beta}_1 = 0.17863$

(b) No, $R^2 = 0.4587$

(c) $\hat{Y}_{x=7} = 4.54$

6. Il direttore di un call center vuole verificare se la durata della telefonata è linearmente dipendente dal numero di persone coinvolte nel servizio richiesto. Vengono osservate 4 telefonate e per ognuna è registrata la durata in minuti (Y) e il numero di persone coinvolte (X):

Persone coinvolte (X)	1	2	3	6
Durata in minuti (Y)	40	70	90	200

(a) Disegnare il grafico a dispersione

(b) Determinare la retta di regressione e calcolare i valori teorici.

(c) Calcolare SSE (somma dei quadrati degli errori) e il coefficiente di determinazione lineare.

(d) Sottoporre a test l'ipotesi nulla che il coefficiente di regressione β_1 sia nullo al livello di significatività del 1%.

(e) In media, quanto durerebbe in media una telefonata nel caso in cui fossero coinvolte 4 persone? Calcolare un intervallo di confidenza (al livello di confidenza $1-\alpha = 0.99$) per la durata media della telefonata con 4 persone coinvolte.

Soluzione:

(a) Diagramma a dispersione.

(b) $\hat{\beta}_0 = 3.571$ e $\hat{\beta}_1 = 32.143$
Valori preditti: $\hat{y}_1 = 35.714, \hat{y}_2 = 67.857, \hat{y}_3 = 100.000, \hat{y}_4 = 196.429$

(c) $SSE = 135.7143$ e $R^2 = 0.9907$

(d) Ipotesi: $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$

Regione critica: $RC_{0.01} = T < -9.9248$ oppure $T > 9.9248$.

Valore osservato della statistica test: $T^{oss} = 14.60$

Decisione: Si rifiuta H_0 al livello di significatività del 1%

(e) $\hat{Y}_{x=4} = 132.1429$, $IC_{0.99}(Y_{=4}) = (85.7914; 178.4943)$

7. Da uno studio psicologico sulle caratteristiche della personalità si è trovato, attraverso specifici test psicologici, che la correlazione tra avarizia (X) e arroganza (Y) è pari a $-0,4$.

(a) La retta di regressione che mette Y in funzione di X (stimata con il metodo dei minimi quadrati) avrebbe pendenza positiva?

(b) La retta si adatterebbe bene ai dati osservati?

(c) Sapendo che per il campione osservato si ha $\bar{x} = 4$, $s_x^2 = 0.81$, $\bar{y} = 6$, e $s_x^2 = 1.44$, stimare i coefficienti di regressione della retta di regressione che pone Y in funzione di X.

Soluzione:

(a) No, la pendenza della retta di regressione, $\hat{\beta}_1$ sarebbe negativa perchè $r_{XY} = -0.4 < 0$.

(b) No, $R^2 = 0.16$: Solo il 16% della variabilità totale è dell'arroganza è spiegata dall'avarizia attraverso il modello di regressione lineare semplice.

(c) $\hat{\beta}_1 = -0.533$, $\hat{\beta}_0 = 8.133$

8. La seguente tabella riporta i dati di reddito e di consumo medio mensile misurati in migliaia di Euro:

Reddito (X)	7	6	8	3	6
Consumo (Y)	4	3	5	2	4

(a) Calcolare i coefficienti della retta di regressione.

(b) Dalla relazione trovata, si può dire che mediamente circa la metà del reddito di una famiglia finisce in consumi?

(c) La retta spiega più del 80% della variabilità totale del consumo?

(d) Costruire un intervallo di confidenza per il coefficiente di regressione β_1 al livello di confidenza del 95%. Interpretare il risultato.

(e) Quale sarebbe il consumo di una famiglia che guadagna 4000 euro? Determinare un intervallo di previsione al livello di confidenza del 95%.

Soluzione:

(a) $\hat{\beta}_0 = 0.1714$ e $\hat{\beta}_1 = 0.5714$

(b) Si, essendo $\hat{\beta}_1 = 0.5714$.

(c) Si, $R^2 = 0.8791$

(d) $IC_{0.95}(\beta_1) = (0.1821; 0.9607)$

(e) $\hat{Y}_{x=4} = 2.457$ (Consumo previsto: 2457 euro), $IC_{0.95}(Y_{x=4}) = (0.6815; 4.2327)$

9. È stata misurata l'altezza, X, in cm e il peso, Y, in Kg di 1000 individui, e dai dati sono state calcolate le seguenti statistiche di sintesi:

$$\bar{x} = 168 \qquad s_x^2 = 7.6 \qquad \bar{y} = 68 \qquad s_y^2 = 9 \qquad r_{XY} = 0.6$$

(a) Determinare la stima dei coefficienti della retta di regressione che pone il peso Y in funzione dell'altezza X

(b) Determinare R^2 e una stima della varianza degli errori del modello di regressione.

(c) Stimare il peso medio per soggetti alti 160 cm. Costruire un intervallo di confidenza al livello di confidenza del 95% per il peso medio di soggetti alti 160 cm.

Soluzione:

(a) $\hat{\beta}_0 = -41.692$ e $\hat{\beta}_1 = 0.653$

(b) $R^2 = 0.36$, $s^2 = 5.7658$

(c) $\hat{Y}_{x=160} = 62.78$; $IC_{0.95}(Y_{x=160}) = (62.32; 63.23)$

10. Utilizzando i dati dell'esercizio 5 determinare la stima dei coefficienti della retta di regressione che pone l'altezza X in funzione del peso Y. L'indice di determinazione lineare cambia?

Soluzione: $\hat{\beta}_0 = 130.507$ e $\hat{\beta}_1 = 0.551$. L'indice di determinazione lineare non cambia: $R^2 = 0.36$.

11. Si sono osservati 8 valori per le due variabili X e Y:

x_i	-2	-5	4	5	8	10	-7	12
y_i	2	-3	10	8	20	60	-18	24

(a) Determinare la stima dei coefficienti della retta di regressione che pone Y in funzione di X e calcolare R^2.

(b) Tra i dati osservati, individuare l'osservazione che presenta il residuo più elevato.

(c) Stimare la retta di regressione eliminando l'osservazione individuata al punto precedente. In questo caso, l'indice di determinazione R^2 spiega più del 90% della variabilità totale?

Soluzione:

(a) $\hat{\beta}_0 = 4.4271$ e $\hat{\beta}_1 = 2.7033$; $R^2 = 0.6797$

(b) L'osservazione che presenta il residuo più elevato è l'osservazione sei per cui il residuo è $\hat{e}_6 = 28.54$

(c) $\hat{\beta}_0 = 1.9797$ e $\hat{\beta}_1 = 1.9428$. In questo caso, l'indice di determinazione R^2 spiega più del 90% della variabilità totale: $R^2 = 0.9176$

Esercizi dal libro di testo: Capitolo 9

9.1	9.2	9.3	9.4	9.6	9.7	9.11	9.12	9.13	9.15	9.17	9.19	9.20	9.29	9.30
9.42	9.43	9.47	9.48	9.55	9.58	9.59	9.60	9.61	9.62					