LM−88 SOCIOLOGIA E RICERCA SOCIALE

Metodi Statistici per la Ricerca Sociale

Regressione multipla e correlazione

1. In uno studio sul consumo negli Stati Uniti si sono osservati tra il 1959 e il 1999 i valori delle seguenti variabili: Spesa annuale per consumi in bilioni di dollari (variabile risposta, Y), Prodotto interno lordo in bilioni di dollari (X_1), Popolazione in migliaia (X_2) e Tasso di disoccupazione in percentuale (X_3).

   Stimando il modello di regressione lineare multipla che fa dipendere la variabile Spesa annuale per consumi dalla Popolazione e dal Tasso di disoccupazione, si ottengono le seguenti tabelle di output:

Variabile	*Stima*	*ES*
Costante	−11478.9205	638.1019
Popolazione (migliaia)	0.0650	0.0027
Tasso di disoccupazione (%)	−161.0844	51.5013

$R^2 = 0.9373$ $SQT = 133208747$ Numero di osservazioni: $n = 41$

   (a) Dalla stima del coefficiente di regressione si può dire che la Spesa per consumi è legata inversamente al Tasso di disoccupazione?

   (b) Condurre test per l'uguaglianza a zero dei singoli coefficienti di regressione al livello di significatività del 1%.

   (c) Stimare la varianza degli errori di regressione

   (d) Sapendo che nel 1990 la Popolazione è di 249973 migliaia e il Tasso % di disoccupazione è di 5.6, qual è il valore atteso della Spesa annuale per consumi prevista dal modello?

   (e) Verificare l'ipotesi nulla $H_0 : \beta_1 = \beta_2 = 0$, al livello di significatività $\alpha = 0.01$ ($F_{critico} = 5.2112$)

   (f) Il $p-$value della statistica test F calcolata nel precedente punto è pari a 0.000. Cosa possiamo concludere?

Soluzione:

   (a) Si, poiché il segno negativo del coefficiente indica una relazione inversa tra Tasso % di disoccupazione e Spesa per consumi.

   (b) Ipotesi: $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$

   Regione critica: $RC_{0.01} = T < -2.7116$ *oppure* $T > 2.7116$

Variabile	*Stima*	*ES*	T^{oss}
Costante	−11478.9205	638.1019	
Popolazione (migliaia)	0.0650	0.0027	23.786
Tasso di disoccupazione (%)	−161.0844	51.5013	−3.128

   Evidenza contro l'ipotesi nulla per entrambi i coefficienti: Le due variabili esplicative sono significative al livello di significatività del 1%.

   (c) Stima della varianza degli errori di regressione: $s^2 = 219858.6$

   (d) Valore atteso della Spesa annuale per consumi prevista dal modello: $\hat{Y} = 3859.308$

(e) Ipotesi: $H_0 : \beta_1 = \beta_2 = 0$ versus $H_1 : \beta_1 \neq 0 \;\; o \;\; \beta_2 \neq 0$

Regione critica: $RC_{0.01} = F > 5.2112$.

Valore osservato della statistica test: $F^{oss} = 283.94$

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

(f) Il $p-$value della statistica test F calcolata nel precedente punto è pari a 0.000 quindi si ha forte evidenza contro l'ipotesi nulla $H_0 : \beta_1 = \beta_2 = 0$.

2. Su un campione di 38 paesi sono osservati dati relativi all'Aspettativa di vita, al Logaritmo del numero medio di persone per TV e al Logaritmo del numero di persone per medico (dati del 1993). Stimando modello di regressione lineare multipla che pone l'Aspettativa di vita in dipendenza delle variabili esplicative Logaritmo del numero medio di persone per TV e Logaritmo del numero di persone per medico si sono ottenuti i seguenti risultati:

Variabile	*Stima*	*ES*
Costante	90.6222	4.3557
Log n. medio di persone per TV	−6.7134	1.3601
Log n. medio di persone per medico	−5.2012	1.7211

$$SQE = 480.1128 \qquad SQT = 2252.368$$

(a) Valutare e commentare la significatività dei coefficienti di regressione ($\alpha = 0.05$). Commentare il tipo di relazione tenuto dalle variabili esplicative rispetto alla variabile dipendente.

(b) L'adattamento del modello ai dati si può ritenere soddisfacente?

(c) Il modello nel suo complesso risulta significativo al livello di significatività del 5% ($F_{critico} = 3.267$)?

Soluzione:

(a) Si, poiché il segno negativo del coefficiente indica una relazione inversa tra Tasso % di disoccupazione e Spesa per consumi.

(b) Ipotesi: $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$

Regione critica: $RC_{0.05} = T < -2.0301 \;\; oppure \;\; T > 2.0301$

Variabile	*Stima*	*ES*	T^{oss}
Costante	90.6222	4.3557	
Log n. medio di persone per TV	−6.7134	1.3601	−4.936
Log n. medio di persone per medico	−5.2012	1.7211	−3.022

Decisione: Evidenza contro l'ipotesi nulla per entrambi i coefficienti: Le due variabili esplicative sono significative al livello di significatività del 5%.

Entrambe le variabili possiedono una relazione inversa con l'Aspettativa di vita come mostrato dal segno negativo dei coefficienti.

(c) $R^2 = 0.7868$. Possiamo considerare l'adattamento soddisfacente visto che il modello spiega circa il 79% della variabilità complessiva.

(d) Ipotesi: $H_0 : \beta_1 = \beta_2 = 0$ versus $H_1 : \beta_1 \neq 0 \;\; o \;\; \beta_2 \neq 0$

Regione critica: $RC_{0.05} = F > 3.267$.

Valore osservato della statistica test: $F^{oss} = 64.598$

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

3. I seguenti dati si riferiscono a 46 province della svizzere con lingua madre francese. In particolare sono rilevate informazioni (relative al 1888 circa) su un opportuno indice di fecondità e su degli indicatori socio-economici: X_i = % di maschi occupati in attività agricole; X_2 = % di soldati di leva che hanno ricevuto un punteggio elevato all'esame militare; X_3 = % di soldati di leva con titolo di studio superiore alla scuola primaria; X_4 = % di cattolici (rispetto a protestanti); X_5 = numero di nati vivi che muoiono entro il primo anno di vita. Si sono stimati due modelli di regressione lineare multipla: (1) modello esteso che pone l'indice di fecondità in funzione di tutte le quattro variabili esplicative; (2) modello ridotto che pone l'indice di fecondità in funzione di X_3, X_4 e X_5. Nelle seguenti tabelle sono mostrati i risultati delle analisi:

Modello esteso: Tavola anova

Fonte di Variabilità	Simbolo	SQ	GdL	MQ	$F - value$
Regressione	SQR				
Residua	SQE	2068.053			
Totale	SQT	7074.604			

Modello ridotto: Tavola anova

Fonte di Variabilità	Simbolo	SQ	GdL	MQ	$F - value$
Regressione	SQR			1592.578	
Residua	SQE				
Totale	SQT	7074.604			

(a) Completare le tavole di analisi della varianza per il modello esteso e per il modello ridotto

(b) Si consideri il modello esteso: Testare l'ipotesi nulla che i coefficienti di regressioni siano tutti uguali a zero al livello di significatività $\alpha = 0.01$?

(c) Calcolare l'indice di determinazione multiplo standard e corretto per il modello esteso e per il modello ridotto.

(d) Sulla base del test F possiamo accettare il modello ridotto al livello di significatività $\alpha = 0.01$.

Soluzione:

(a) Tavole di analisi della varianza

Modello esteso: Tavola anova

Fonte di Variabilità	Simbolo	SQ	GdL	MQ	$F - value$
Regressione	SQR	5006.550	5	1001.310	19.367
Residua	SQE	2068.053	40	51.701	
Totale	SQT	7074.604	45		

Modello ridotto: Tavola anova

Fonte di Variabilità	Simbolo	SQ	GdL	MQ	$F - value$
Regressione	SQR	4777.735	3	1592.578	29.122
Residua	SQE	2296.868	42	54.687	
Totale	SQT	7074.604	45		

(b) Ipotesi: $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ versus H_1 : Almeno un coefficiente $\beta_j \neq 0$

Regione critica: $RC_{0.01} = F > 3.514$.

Valore osservato della statistica test: $F^{oss} = 19.36$

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

(c) Indici di determinazione

Modello	R^2	$R^2_{Adjusted}$
Ridotto	0.6753	0.6521
Esteso	0.7077	0.6711

(d) Ipotesi: $H_0 : \beta_1 = \beta_2 = 0$ versus $H_1 : \beta_1 \neq 0 \;\; o \;\; \beta_2 \neq 0$

Regione critica: $RC_{0.01} = F > 5.178$.

Valore osservato della statistica test: $F^{oss} = 2.2129$

Decisione: Non si può rifiutare H_0 al livello di significatività del 1%

Esercizi dal libro di testo: Capitolo 11

11.1 11.2 11.3 11.5 11.6 11.9 11.10 11.11(no punto d) 11.14 11.15 11.16 11.17
11.18 11.21 11.22 11.24 11.25 11.26 11.28 11.29 11.30 11.42 11.47 11.48
11.49 11.50 11.51 11.52 11.56