LM−88 SOCIOLOGIA E RICERCA SOCIALE

Metodi Statistici per la Ricerca Sociale

Confronto tra gruppi: Modelli di regressione con variabili esplicative categoriali
Modelli di regressione con variabili esplicative categoriali e continue

1. Si registrano i punteggi ottenuti in un certo test da 15 studenti provenienti da 3 diverse scuole, e ci si chiede se le medie delle tre scuole siano uguali. Stabilire tramite l'analisi della varianza se, al 5% di significatività, i dati mostrano evidenza a favore o contraria l'ipotesi nulla.

Studente	Scuola 1	Scuola 2	Scuola 3
$1-2-3$	220	244	252
$4-5-6$	251	235	272
$7-8-9$	226	232	250
$10-11-12$	246	242	238
$13-14-15$	260	225	256

Soluzione: Ipotesi: $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$ versus $H_1 : \mu_1 \neq 0 \;\; o \;\; \mu_2 \neq 0 \;\; o \;\; \mu_3 \neq 0$

Regione critica: $RC_{0.05} = F > 3.885$.

Valore osservato della statistica test: $F^{oss} = 2.601$

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

2. In un certo studio, si sono osservati 1299 individui, suddivisi in 4 gruppi diversi. Sulla base dei dati raccolti, si vuole fare l'analisi della varianza. Completare la seguente tabella ANOVA e verificare se al livello $\alpha = 0.05$ si rifiuta l'ipotesi nulla di uguaglianza delle medie tra i gruppi.

Fonte di variabilità	Somma dei quadrati	GdL	Media dei quadrati	$F - value$
Tra gruppi				
Entro i gruppi	922.82			
Totale	934.54			

Soluzione:

Fonte di variabilità	Somma dei quadrati	GdL	Media dei quadrati	$F - value$
Tra gruppi	11.72	3	3.91	5.48
Entro i gruppi	922.82	1295	0.71	
Totale	934.54	1298	0.72	

Ipotesi: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ versus $H_1 : \mu_1 \neq 0 \;\; o \;\; \mu_2 \neq 0 \;\; o \;\; \mu_3 \neq 0 \;\; o \;\; \mu_4 \neq 0$

Regione critica: $RC_{0.05} = F > 2.61$.

Valore osservato della statistica test: $F^{oss} = 5.48$

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

3. In una indagine è stato chiesto ad un certo numero di persone quanti buoni amici hanno. Vengono confrontati i risultati per stato civile degli intervistati (sposati, vedovi, divorziati o separati, mai sposati). La tabella di analisi della varianza riporta $F^{oss} = 0.80$.

(a) Specificare H_0 e H_1.

(b) Basandosi sulla distribuzione F, ritenete che $F^{oss} = 0.80$ comporti una forte evidenza contro H_0? Spiegare.

(c) Il software usato per l'analisi dà un $p-$value di 0.53. Spiegare il risultato, anche alla luce del punto (b).

Soluzione: Si indichi con 1 = sposati, 2 = vedovi, 3 = divorziati o separati, 4 = mai sposati

(a) Ipotesi: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ versus $H_1 : \mu_1 \neq 0 \;\; o \;\; \mu_2 \neq 0 \;\; o \;\; \mu_3 \neq 0 \;\; o \;\; \mu_4 \neq 0$

(b) Il valore osservato della statistica test è piccolo (prossimo a zero) e dato che la distribuzione F è caratterizzata da una distribuzione asimmetrica a destra è ragionevole pensare che un valore osservato della statistica test $F^{oss} = 0.80$ non comporti una forte evidenza contro H_0.

(c) Il valore del $p-$value uguale a 0.53 supporta il ragionamento al punto (b), suggerendo che i dati non mostrano alcuna evidenza contro l'ipotesi nulla di uguaglianza delle medie.

4. In uno studio per il confronto della soddisfazione del cliente nei centri di servizio per il supporto tecnico ai PC di San Jose (California), Toronto (Canada) e Bangalore (India), ciascun centro ha estratto un campione di 100 persone tra coloro che hanno chiamato in una periodo di due settimane. Gli intervistati forniscono il loro grado di soddisfazione su una scala da 0 a 10. Le medie campionarie sono state 7.6 per San Jose, 7.8 per Toronto e 7.1 per Bangalore. Si effettua l'Anova per verificare se la soddisfazione media nelle 3 città è diversa.

Fonte di variabilità	Somma dei quadrati	GdL	Media dei quadrati	$F-value$	$p-value$
Tra gruppi					0.000
Entro i gruppi					
Totale	166.0				

(a) Sulla base dei dati forniti completare la seguente tabella Anova, interpretare e spiegare i risultati.

(b) Specificare un modello di analisi della varianza e riportare le stime dei minimi quadrati dei parametri. Interpretare quindi i coefficienti del modello.

Soluzione:

(a) Tavola ANOVA

Fonte di variabilità	Somma dei quadrati	GdL	Media dei quadrati	$F-value$	$p-value$
Tra gruppi	26	2	13	27.58	0.000
Entro i gruppi	140	297	0.47		
Totale	166.0	299	0.56		

Il $p-value$ suggerisce una forte evidenza contro l'ipotesi nulla Ipotesi: $H_0 : \mu_1 = \mu_2 = \mu_3 = 0$

(b) Considerando come livello di riferimento 'Bangalore' si ha

$$Y_i = \beta_0 + \beta_1 \cdot A_i^{San-Jose} + \beta_2 \cdot A_i^{Toronto} + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2) \;\; \text{indipendenti.}$$

Stime dei minimi quadrati: $\hat{\beta}_0 = 7.1$ (soddisfazione media dei clienti a 'Bangalore'); $\hat{\beta}_1 = 0.5$ (differenza tra la soddisfazione media dei clienti a 'San Jose' e a 'Bangalore'); $\hat{\beta}_2 = 0.7$ (differenza tra la soddisfazione media dei clienti a 'Toronto' e a 'Bangalore').

5. In una indagine sulla relazione tra salute e nutrizione sono stati selezionati due campioni di donne in due stati del nord America. Su ciascun soggetto sono rilevate le seguenti variabili: X = l'età (in anni); A = la regione ($A = 0$ se Iowa; $A = 1$ se Nebrasca); e Y = il tasso di colesterolo nel sangue (mg/100ml). Interessa studiare la relazione tra tasso di colesterolo ed eta' tenendo conto della regione. Si hanno le seguenti informazioni

	Iowa	Nebrasca
n_ℓ	11	19
$\bar{x}_\ell$	53.09	45.95
$\bar{y}_\ell$	207.73	217.11
$\sum_{i=1}^{n_\ell}(x_{i\ell} - \bar{x}_\ell)^2$	1828.91	5564.95
$\sum_{i=1}^{n_\ell}(x_{i\ell} - \bar{x}_\ell)(y_{i\ell} - \bar{y}_\ell)$	5922.27	14026.11

(a) Specificare un modello di regressione additivo che pone il tasso di colesterolo in funzione dell'età e della regione. Stimare i coefficienti del modello. Interpretare i coefficienti relativi alla variabile età e alla variabile regione.

(b) Si stima un modello di regressione con interazione ottenendo i seguenti risultati

Costante	*Stima*	*ES*
Variabili	35.81	55.12
Età	3.24	1.01
Regione	65.49	61.98
Interazione	−0.72	1.16

Specificare un modello di regressione con interazione. Valutare la significatività dell'interazione attraverso un test t al livello di significatività $\alpha = 0.05$

(c) Sapendo che la somma dei quadrati dei residui del modello additivo è 49103.91 e la somma dei quadrati dei residui del modello con interazione è 48394.86 valutare la significatività dell'interazione attraverso un test t al livello di significatività $\alpha = 0.05$. Confrontare il risultato ottenuto con quello ottenuto al punto (b)

Soluzione:

(a) Modello additivo: $Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot A_i^{Nebrasca} + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ indipendenti.

Stime dei coefficienti: $\hat{\beta}_0 = 64.49$; $\hat{\beta}_1 = 2.698$; $\hat{\beta}_2 = 28.651$.

$\hat{\beta}_1 = 2.698$: In Nebrasca e in Iowa, ogni anno di età in più comporta un incremento nel tasso di colesterolo medio di 2.698. $\hat{\beta}_2 = 28.651$: Per ogni livello di età fissato, la differenza tra il tasso di colesterolo medio in Nebrasca e in Iowa è pari a 28.651.

(b) Modello con interazione: $Y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot A_i^{Nebrasca} + \beta_3 \cdot A_i^{Nebrasca} \cdot x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ indipendenti.

Ipotesi: $H_0 : \beta_3 = 0$ versus $H_0 : \beta_3 \neq 0$

Regione critica: $RC_{0.05} = T < -2.056 \;\; o \;\; T > 2.056$.

Valore osservato della statistica test: $T^{oss} = -0.62$

Decisione: Non si può rifiuta H_0 al livello di significatività del 5% (Evidenza a favore del modello additivo)

(c) Ipotesi: $H_0 : \beta_3 = 0$ versus $H_0 : \beta_3 \neq 0$

Regione critica: $RC_{0.05} = F > 4.225$.

Valore osservato della statistica test: $F^{oss} = 0.38$

Decisione: Non si può rifiuta H_0 al livello di significatività del 5% (Evidenza a favore del modello additivo)

Esercizi dal libro di testo: Capitolo 12

12.1 12.2 12.4 12.5 12.6 12.8 12.11 12.13 12.16 12.17 12.18 12.19 12.20 12.21
12.22 12.24 12.27 12.38 12.39 12.40 12.42 12.47 12.49 12.50 12.51 12.52 12.53

Nota bene: La somma dei quadrati relativa a una variabile è la differenza tra la somma dei quadrati dei residui del modello ridotto in cui *non è inclusa* tale variabile e la somma dei quadrati dei residui del modello esteso in cui *è inclusa* tale variabile.

Esercizi dal libro di testo: Capitolo 13

13.1 13.2 13.3 13.5 13.6 13.7 13.8 13.10 13.16 13.18 13.19 13.24 13.25 13.29
13.30 13.32