LM−88 SOCIOLOGIA E RICERCA SOCIALE

Compito di
Metodi Statistici per la Ricerca Sociale
11 Settembre 2014

Studente: ____________________

Matricola: ____________________

Esercizi

1. Una ricerca in Svezia ha messo a confronto un gruppo di ex calciatori professionisti, un gruppo di calciatori dilettanti e delle persone della stessa età che non hanno mai giocato a calcio. La seguente tabella a doppia entrata classifica questi soggetti in base al fatto di avere sofferto o meno di artriti all'anca o al ginocchio entro i cinquant'anni di età.

	Professionisti	Dilettanti	Mai giocato
Artriti	10	9	24
Nessun tipo di artrite	61	206	548

(a) Sottoporre a test l'ipotesi che la pratica del gioco del calcio e l'aver sofferto o meno di artriti all'anca o al ginocchio entro i cinquant'anni di età siano statisticamente indipendenti al livello di significatività $\alpha = 0.05$.

(b) Si focalizzi l'attenzione sul gruppo degli ex calciatori professionisti, e sul gruppo delle persone della stessa età che non hanno mai giocato a calcio, ossia sulla sotto-tabella:

	Professionisti	Mai giocato
Artriti	10	24
Nessun tipo di artrite	61	548

Calcolare gli odds a favore della presenza di artriti per il gruppo degli ex calciatori professionisti e per gruppo delle persone che non hanno mai giocato a calcio. Calcolare quindi il rapporto degli odds e costruire un intervallo di confidenza per il rapporto degli odds al livello di confidenza del 95%.

Soluzione:

(a) Sistema di ipotesi

H_0 : La pratica del gioco del calcio e l'aver sofferto o meno di artriti sono statisticamente indipendenti

H_1 : La pratica del gioco del calcio e l'aver sofferto o meno di artriti NON sono statisticamente indipendenti

Regione di rifiuto

$$\chi^2 \geq X^2_{(2-1)(3-1)}(0.05) = X^2_2(0.05) = 5.9915$$

Si calcola il valore osservato della statistica test.

1. Distribuzioni marginali

	Professionisti	Dilettanti	Mai giocato	Totale
Artrite	10	9	24	43
Nessun tipo di artrite	61	206	548	815
Totale	71	215	572	858

2. Frequenze teoriche di indipendenza

	Professionisti	Dilettanti	Mai giocato
Artriti	3.558	10.775	28.667
Nessun tipo di artrite	67.442	204.225	543.333

3. Valore osservato della statistica test

$$\begin{aligned}
\chi^{2,oss} &= \frac{(10-3.558)^2}{3.558} + \frac{(9-10.775)^2}{10.775} + \frac{(24-28.667)^2}{28.667} + \\
&\quad \frac{(61-67.442)^2}{67.442} + \frac{(206-204.225)^2}{204.225} + \frac{(548-543.333)^2}{543.333} \\
&= \frac{(6.442)^2}{3.558} + \frac{(-1.775)^2}{10.775} + \frac{(-4.667)^2}{28.667} + \frac{(-6.442)^2}{67.442} + \frac{(1.775)^2}{204.225} + \frac{(4.667)^2}{543.333} \\
&= \frac{41.496}{3.558} + \frac{3.151}{10.775} + \frac{21.778}{28.667} + \frac{41.496}{67.442} + \frac{3.151}{204.225} + \frac{21.778}{543.333} \\
&= 11.662 + 0.292 + 0.760 + 0.615 + 0.015 + 0.040 \\
&= 13.385
\end{aligned}$$

Il valore osservato della statistica test appartiene alla regione di rifiuto quindi si rifiuta l'ipotesi nulla di indipendenza tra i due caratteri al livello di significatività $\alpha = 0.05$.

(b) Probabilità di soffrire di artriti:

$$\hat{\pi}_{\text{Calciatori}} = \frac{10}{71} = 0.141 \qquad \hat{\pi}_{\text{Non Calciatori}} = \frac{24}{572} = 0.041$$

Odds a favore della presenza di artriti:

$$\hat{Odds}_{\text{Calciatori}} = \frac{0.141}{1-0.141} = \frac{0.141}{0.859} = 0.164$$

$$\hat{Odds}_{\text{Non Calciatori}} = \frac{0.041}{1-0.041} = \frac{0.041}{1-0.958} = 0.044$$

Rapporto degli Odds:

$$\hat{OR} = \frac{0.164}{0.044} = 3.743$$

Logaritmo del rapporto degli Odds e suo errore standard:

$$\log(\hat{OR}) = \log(3.743) = 1.320 \qquad \hat{e.s.}(\log(\hat{OR})) = \sqrt{\frac{1}{10} + \frac{1}{24} + \frac{1}{61} + \frac{1}{548}} = \sqrt{0.160} = 0.400$$

Intervallo di confidenza per il logaritmo dell'odds ratio:

$$IC_{0.95}(\log(OR)) = [1.320 \pm 1.96 \cdot 0.400] = [1.320 \pm 0.784] = [0.536; 2.104]$$

Intervallo di confidenza per l'odds ratio:

$$IC_{0.95}(OR) = [\exp(0.536); \exp(2.104)] = [1.710; 8.196]$$

2. In uno studio finalizzato a valutare se è possibile prevedere il livello di alcol presente nel suo sangue in base al numero di birre bevute da uno studente, 5 studenti volontari di una certa Università hanno bevuto un determinato numero, assegnato loro casualmente, di lattine di birra. Dopo 30 minuti un ufficiale di polizia ha misurato il tasso alcolico del loro sangue (TAS).

Studente	1	2	3	4	5
Birre	6	2	7	8	3
TAS·100	10	3	9.5	12	4

(a) Stimare la retta di regressione che pone il tasso alcolico nel sangue (Y) in funzione del numero di lattine di birra bevute (X)

(b) Calcolare l'indice di determinazione lineare (R^2) sapendo che la somma dei quadrati dei residui è $SQE = 2.2$. Interpretare il risultato.

(c) Stimare il livello medio di alcol nel sangue pari per soggetti che bevono 5 lattine di birra e costruire il relativo intervallo di confidenza al livello di confidenza del 99%.

Soluzione:

u_i	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	6	10.0	36	100.00	60.0	0.64	5.29	1.84
2	2	3.0	4	9.00	6.0	10.24	22.09	15.04
3	7	9.5	49	90.25	66.5	3.24	3.24	3.24
4	8	12.0	64	144.00	96.0	7.84	18.49	12.04
5	3	4.0	9	16.00	12.0	4.84	13.69	8.14
Totale	26	38.5	162	359.25	240.5	26.80	62.80	40.30

(a) Coefficienti di regressione

$$\bar{x} = \frac{26}{5} = 5.2 \qquad \bar{y} = \frac{38.5}{5} = 7.7$$

$$\begin{aligned} \hat{\beta}_1 &= \frac{240.5 - 5 \cdot 5.2 \cdot 7.7}{162 - 5 \cdot 5.2^2} = \frac{240.5 - 5 \cdot 40.04}{162 - 5 \cdot 27.04} = \frac{240.5 - 200.2}{162 - 135.2} = \frac{40.3}{26.80} = 1.504 \\ \hat{\beta}_0 &= 7.7 - 1.504 \cdot 5.2 = 7.7 - 7.819 = -0.119 \end{aligned}$$

(b) Indice di determinazione lineare (R^2)

$$SQT = 359.25 - 5 \cdot 7.7^2 = 359.25 - 5 \cdot 59.29 = 359.25 - 296.45 = 62.8$$

Quindi

$$R^2 = 1 - \frac{2.2}{62.8} = 1 - 0.035 = 0.965$$

Poichè l'indice di determinazione lineare è pari a 0.965, abbiamo che oltre il 96% della variabilità totale nel livello di alcol tra gli studenti è determinata dal numero di lattine di birra bevute

(c) Stima dell livello medio di alcol nel sangue pari per soggetti che bevono 5 lattine di birra:

$$\hat{y}_{x=5} = -0.119 + 1.504 \cdot 5 = -0.119 + 7.519 = 7.4$$

$$s^2 = \frac{SQE}{n-2} = \frac{2.2}{5-2} = 0.733$$

$$e\hat{.s}.(\hat{y}_{x=5}) = \sqrt{0.733 \cdot \left[\frac{1}{5} + \frac{(5-5.2)^2}{26.80}\right]} = \sqrt{0.733 \cdot \left[0.2 + \frac{0.04}{26.80}\right]} = \sqrt{0.1477} = 0.3844$$

Valore critico: $t_{0.005,3} = 5.84$

$$IC_{0.99}(\mathbb{E}[Y|X=5]) = (0.733 \pm 5.84 \cdot 0.3844) = (0.733 \pm 2.245) = (5.155; 9.644)$$

3. In uno studio in campo agrario un campione di 24 piante di un certo tipo sono esposte a quattro diversi dosaggi di un fertilizzante (Basso; Medio; Elevato; Molto Elevato). Dopo il trattamento è misurata l'altezza delle piante. La tabella che segue mostra i risultati ottenuti da un'analisi della varianza:

Fonte di variabilità	Somma dei Quadrati	GdL	Media dei Quadrati	$F-$value
Dosaggio			41.08	
Residua				
Totale	1185.96			

(a) Completare la tavola di analisi della varianza

(b) Verificare l'ipotesi che non ci sia differenza nelle le altezze medie di piante esposte a diversi dosaggi di fertilizzante al livello di significatività del 5%.

Soluzione:

(a) Tavola di analisi della varianza

Fonte di variabilità	Somma dei Quadrati	GdL	Media dei Quadrati	$F-$value
Dosaggio	123.24	3	41.08	0.773
Residua	1062.72	20	53.136	
Totale	1185.96	23	51.563	

dove

$$F - value = \frac{41.08}{53.136} = 0.773$$

(b) Regione critica al livello del 10%

$$RC_{0.5} = \{f : f > f_{3,20}(0.95) = 3.098\}$$

Il valore osservato della statistica test $f^{oss} = 0.773$ non appartiene alla regione di rifiuto quindi si conclude che non c'è evidenza contro l'ipotesi nulla che la crescita delle piante non dipenda in media dalla dose di fertilizzante.

Domande Teoriche

1. Di seguito si riportano i risulti di un'analisi in cui un modello di regressione logistica è stato stimato su un campione di 54 uomini anziani. La variabile risposta è una variabile binaria che assume valore 1 per soggetti che presentano sintomi di senilità. Come variabile esplicativa si considera il punteggio ottenuto a un particolare test (il punteggio WAIS = Wechsler Adult Intelligence Scale). Valori maggiori della variabile WAIS indicano capacità intellettive migliori.

Variabile	Coefficiente	SE	$p-value$
Costante	3.3728	1.6239	0.0378
WAIS	-0.4292	0.1621	0.0081

(a) Scrivere l'equazione che definisce i valori stimati delle probabilità: $\hat{\pi}_i$

(b) Interpretare il $p-$value relativo al coefficiente della variabile esplicatica WAIS.

(c) Interpretare il coefficiente relativo alla variabile WAIS.

Soluzione:

(a) Valori stimati delle probabilità

$$\hat{\pi}_i = \frac{\exp\{3.3728 - 0.4292 \cdot \text{WAIS}_i\}}{1 + \exp\{3.3728 - 0.4292 \cdot \text{WAIS}_i\}}$$

(b) Il $p-$value relativo al coefficiente della variabile esplicatica WAIS è piccolo (inferiore a 0.01). Quindi si ha forte evidenza che il coefficiente sia diverso da zero (l'ipotesi nulla $H_0 : \beta_1 = 0$ sarebbe rifiutata a ogni livello $\alpha > 0.0081$).

(c) Interpretazione in termini di odds: $e^{-0.4292} = 0.6510$. Per ogni incremento unitario della variabile WAIS, gli odds a favore della presenza di sintomi di senilità sono moltiplicati per 0.6510 (ossia si riducono del 35% circa).

2. Per ciascuno dei grafici dei residui sotto riportati, valutare se esiste evidenza contro le ipotesi del modello di regressione lineare. In caso affermativo specificare contro quale ipotesi il grafico mostra maggiore evidenza. Giustificare le risposte.