Tecniche di indagine statistica

Lezione 28

Modello di regressione logistica - logit

$$logit(p_i) = \sum b_k X_{ik}$$

modello lineare generalizzato con risposta binomiale e trasformazione (link) logit

Interpretazione dei coefficienti b_k :

come nel modello di regressione lineare, ma su scala logit:

b_k rappresenta il cambiamento nel logit di $P(Y_i = 1)$ associato ad un cambiamento unitario di X_k , tenendo costanti le altre variabili (covariate o variabili indipendenti)

Modello di regressione logistica per $P(Y_i = 1)$

Effetti marginali

$$P(Y_i = 1) = \frac{\exp(\sum b_k X_{ik})}{1 + \exp(\sum b_k X_{ik})}$$

relazione non lineare dei predittori su $P(Y_i = 1)$

Interpretazione dei coefficienti b_k - X_k continua – si ricorre a:

derivata di $P(Y_i = 1)$ rispetto a X_k

$$\frac{\partial P(Y_i = 1)}{\partial X_{ik}} = \frac{\exp(\sum b_k X_{ik})}{[1+\exp(\sum b_k X_{ik})]} \frac{1}{[1+\exp(\sum b_k X_{ik})]} b_k = \frac{\exp(\sum b_k X_{ik})}{[1+\exp(\sum b_k X_{ik})]^2} b_k$$

b_k indica la direzione dell'effetto (positivo o negativo) della variabile X_k

ma l'entità dipende dalla "grandezza" di $\sum b_k X_{ik}$ (e quindi dai valori di tutte le X_{ik}):

variazione valutata in corrispondenza di particolari valori delle covariate (spesso sono scelti i valori medi $\bar{x}_k$, con $k = 1, \ldots, p$) per ottenere l'effetto marginale della variabile X_k.

Interpretazione dei coefficienti b_k - X_k dicotomica 0/1

Effetti marginali

Se X_k è una variabile dicotomica,

l'effetto della variazione di X_{ik} da (modalità) 0 a (modalità) 1, mantenendo tutte le altre variabili esplicative costanti (per esempio, pari al valore medio) è dato da:

$$P(Y_i = 1|x_{i1}, x_{i2}, \dots, x_{ik} = 1) - P(Y_i = 1|x_{i1}, x_{i2}, \dots, x_{ik} = 0)$$

Modello di regressione logistica – interpretazione con *odds ratio* (*OR*) /1

Odds per unità *i*-sima $\frac{p_i}{(1-p_i)}$

Con

$$p_i = \frac{\exp(\sum b_k X_{ik})}{1+\exp(\sum b_k X_{ik})}$$

$$(1 - p_i) = 1 - \frac{\exp(\sum b_k X_{ik})}{1+\exp(\sum b_k X_{ik})}$$

$$= \frac{[1+\exp(\sum b_k X_{ik})]-\exp(\sum b_k X_{ik})}{1+\exp(\sum b_k X_{ik})} = \frac{1}{1+\exp(\sum b_k X_{ik})}$$

$$\frac{p_i}{(1-p_i)} = \exp(\sum b_k X_{ik})$$

Modello di regressione logistica – interpretazione con *odds ratio* (*OR*) /2

Modello con due variabili esplicative, X_1 continua e X_2 dicotomica:

$$p_i = \frac{\exp(b_0 + b_1 X_1 + b_2 X_2)}{1 + \exp(b_0 + b_1 X_1 + b_2 X_2)}$$

$$Odds\,(x_2 = 1) = \frac{P(y=1|x_1, x_2=1)}{1 - P(y=1|x_1, x_2=1)} = \exp(b_0 + b_1 x_1 + b_2)$$

$$Odds\,(x_2 = 0) = \frac{P(y=1|x_1, x_2=0)}{1 - P(y=1|x_1, x_2=0)} = \exp(b_0 + b_1 x_1)$$

Allora $OR = \frac{Odds(x_2 = 1)}{Odds(x_2 = 0)} = \frac{\exp(b_0 + b_1 x_1 + b_2)}{\exp(b_0 + b_1 x_1)} = \exp(b_2)$

Se, per es., $\exp(b_2) = 2$: unità caratterizzate da $x_2 = 1$ hanno una propensione al successo $(y = 1)$ pari al doppio rispetto alle unità con $x_2 = 0$

Modello di regressione logistica – interpretazione con *odds ratio* (*OR*) /3

Modello con due variabili esplicative, X_1 continua e X_2 dicotomica:

$$OR = \frac{Odds(x_1 = x^* + 1)}{Odds(x_1 = x^*)} = \frac{\exp(b_0 + b_1(x^* + 1) + b_2 x_2)}{\exp(b_0 + b_1(x^*) + b_2 x_2)} = \exp(b_1)$$

Spesso, a fini interpretativi, è più interessante considerare un incremento di c unità ($c \neq 1$) piuttosto che un incremento unitario della variabile:

allora $OR = \exp(cb_1)$

Modello di regressione logistica – interpretazione con *odds ratio* (*OR*) /4

Modello con variabile esplicativa X_k categoriale (j modalità)

j modalità: inserite nel modello attraverso $j-1$ variabili indicatrici che indicano la presenza/assenza della modalità considerata:

Modalità A, B, C: modello include X_B e X_C

con X_B $(X_C) = 1$ se $X_{ik} = B(C)$ e $X_B(X_C) = 0$ altrimenti

$X_{ik} = A$ quando $X_B = X_C = 0$; A modalità di riferimento (baseline)

Coefficienti b_B e b_C associati a X_B e X_C

$\exp(b_B)$ propensione al successo delle unità con modalità B rispetto a unità con modalità A

$\exp(b_C)$ propensione al successo delle unità con modalità C rispetto a unità con modalità A

$\exp(b_C - b_B)$ = propensione al successo delle unità con modalità C rispetto a unità con modalità B

(dopo aver controllato per la significatività della differenza $[b_C - b_B]$!)

Modello di regressione logistica – interpretazione con *odds ratio* (*OR*) /5

In generale

(ricordando che il range di OR non è simmetrico: $(1, \infty)$ se > 1 e $(0,1)$ se < 1)

1. se X_k non influenza la probabilità che la variabile risposta Y assuma valore 1:
   $OR = 1, b_k = 0$

2. se X_k influenza positivamente *(aumenta)* la probabilità che la variabile risposta Y assuma valore 1:
   $OR > 1, b_k > 0$

3. se X_k influenza negativamente *(riduce)* la probabilità che la variabile risposta Y assuma valore 1:
   $0 < OR < 1, b_k < 0$

Assunzioni per stima e inferenza

1. $Y_i \in \{0,1\} \quad i = 1,\ldots,n$

2. $P(Y_i = 1 \mid X_i) = \dfrac{\exp(\sum b_k X_{ik})}{1 + \exp(\sum b_k X_{ik})}$

3. $Y_1, Y_2, \ldots Y_n$ statisticamente indipendenti

4. X_{ik} non esiste perfetta o quasi perfetta multicollinearità tra di loro

Per stimare i parametri del modello:
metodo della massima verosimiglianza

poiché le equazioni generate dalla massimizzazione della verosimiglianza sono non lineari nei parametri (non ammettono soluzione esplicita),
le stime dei coefficienti si ottengono utilizzando metodi (algoritmi) numerici iterativi

Stima di massima verosimiglianza

$$p_i = \mathrm{P}(Y_i = 1 \mid X_{ik})$$

$$1 - p_i = \mathrm{P}(Y_i = 0 \mid X_{ik})$$

$$\mathrm{P}(Y_i \mid X_{ik}) = p_i^{\,y_i} (1 - p_i)^{1 - y_i}$$

probabilità di osservare il risultato Y_i

$$\mathrm{P}(Y \mid X) = \prod_{i=1}^{n} p_i^{\,y_i} (1 - p_i)^{1 - y_i} \quad = L\,(Y \mid X, b)$$

⇩

probabilità (verosimiglianza) di osservare quel particolare campione di valori Y_i dato X_{ik}

funzione di verosimiglianza

$\hat{b}$ ⇨ che massimizza $= L\,(Y \mid X, b)$

⇨ **algoritmi iterativi** (metodo Newton - Raphson e metodo scoring)

Stimatori di massima verosimiglianza e inferenza /1

- per grandi campioni proprietà 'analoghe' a stimatori OLS (correttezza, efficienza, normalità)

- valutare i risultati della stima del modello, con riferimento a:

a. singoli coefficienti $\hat{b}_k$ $(H_0: b_k = 0)$

- statistica test $z = \frac{\hat{b}_k - 0}{SE(\hat{b}_k)}$ (per grandi campioni tende a $N(0,1)$)

- in alternativa, anche statistica test z^2 (per grandi campioni tende a χ_1^2)

- intervallo di confidenza (IC) per $\hat{b}_k$ a livello $(1-\alpha)$:

  $\hat{b}_k \pm z_{1-\frac{\alpha}{2}} SE(\hat{b}_k)$

- IC in scala logit traslati in IC per gli OR:

  $\exp\left(\hat{b}_k \pm z_{1-\frac{\alpha}{2}} SE(\hat{b}_k)\right)$

Stimatori di massima verosimiglianza e inferenza /2

b. *plausibilità* modello stimato

test basato sul rapporto delle funzioni di verosimiglianza (LR test)

$logL_1$ verosimiglianza per il modello completo

$logL_0$ verosimiglianza quando tutti i coefficienti sono pari a 0 (tranne intercetta)

statistica test: $-2(logL_0 - logL_1) \approx \chi^2_{k-1}$

c. insiemi di coefficienti

$logL_1$ verosimiglianza per il modello completo

$logL_2$ verosimiglianza modello in cui g variabili (e relativi coefficienti) sono assunte non influenti

statistica test: $-2(logL_2 - logL_1) \approx \chi^2_g$ g = numero di coefficienti pari a 0

d. bontà di adattamento

proposte diverse misure (% casi correttamente classificati e altro (pseudo R^2))

Modello di regressione logistica in R

mod = glm(formula = y ~ x, family = binomial(link = logit), data=Dati)

Argomenti (principali) di glm:

- *formula* specifica il modello: a sx di ~ var. risposta, a dx ~ var. esplicative (puo' essere omessa)

- *family* = tipo di modello che deve essere stimato in relazione a var. risposta (Binomiale generalizzazione di Beornulli - distribuzione per Y),

- *link* indica il legame (collegamento) tra $E(Y)$ e var. esplicative (in modello logistico il collegamento avviene tramite trasformazione logit)

data = nome del data frame

oggetto mod contiene: coefficienti stimati, valori stimati, …

Script R logit (Moodle2)

Esempio 1 – efficacia pubblicità su acquisto prodotto /1

Azienda X vuole valutare efficacia pubblicità sulle vendite di un suo prodotto

Indagine su un campione di consumatori per avere dati su:
Y = acquisto/non acquisto prodotto e X = vista/non vista pubblicità prodotto

Stima modello regressione logistica per $P(Y_i = 1\ (acquisto\ sì) | X_i)$

	coef.=$\widehat{\beta}$	s.e.	z value	Pr>|z|=p-value
intercetta	-0.9694	0.3441	-2.738	0.00619
x	0.9027	0.4383	2.059	0.03945

- $H_0: \beta = 0$ non accettata
- $\hat{\beta}$ positivo: aver visto la pubblicità favorisce l'acquisto
- qual è l'effetto marginale? E OR?

Esempio 1 – efficacia pubblicità su acquisto prodotto /2

	coef.=$\widehat{\beta}$	s.e.	z value	Pr>|z|=p-value
intercetta	-0.9694	0.3441	-2.738	0.00619
x	0.9027	0.4383	2.059	0.03945

1. Calcolo effetto marginale:

$$P(Y_i = 1|x_{i1}, x_{i2}, \ldots, x_{ik} = 1) - P(Y_i = 1|x_{i1}, x_{i2}, \ldots, x_{ik} = 0)$$

$$P(Y_i = 1|x) = \frac{exp(-0.9694 + 0.9027x)}{1 + exp(-0.9694 + 0.9027x)}$$

$$P(Y_i = 1|x = 1) = \frac{exp(-0.9694 + 0.9027 \times 1)}{1 + exp(-0.9694 + 0.9027 \times 1)} = 0.4839 \quad P(Y_i = 1|x = 0) = \frac{exp(-0.9694)}{1 + exp(-0.9694)} = 0.2750$$

$$P(Y_i = 1|x = 1) - P(Y_i = 1|x = 0) = 0.4839 - 0.2750 = 0.2089$$

$$IC(\beta_{95\%}) = [0.9027 \pm 1.96 \times 0.4383]$$

Esempio 1 – efficacia pubblicità su acquisto prodotto /3

2. In termini di OR $logit(p_i) = -0.9694 + 0.9027x$

Odds: $\frac{P(Y_i = 1|x)}{1 - P(Y_i = 1|x)} = exp(-0.9694 + 0.9027x)$

$x = 1 \quad exp(-0.9694 + 0.9027 \times 1) = 0.9361$:

$x = 0 \quad exp(-0.9694 + 0.9027 \times 0) = 0.3795$

$OR = 0.9361/0.3795 = 2.46 = \frac{\frac{P(Y_i=1|x=1)}{1-P(Y_i=1|x=1)}}{\frac{P(Y_i=1|x=0)}{1-P(Y_i=1|x=0)}} = exp(0.9027)$

$IC(OR) = exp[0.9027 \pm 1.96 \times 0.4383]$

Azienda decide di investire ancora in pubblicità poiché aumenta propensione acquisto

Esempio 2 – scelta di rinnovare la polizza assicurativa/1

Azienda assicurativa vuole aumentare il volume di polizze rinnovate.

Quali fattori influenzano la scelta di rinnovo della polizza?

Dati da archivio clienti:

- variabile dipendente RINNOVO= y=(1 se si, 0 se no)
- $x_1$età del cliente
- x_2 reddito del cliente
- x_3 collocazione dell'ufficio in cui il cliente si serve (1 se in centro, 0 altrimenti)

che dire di x_3?

Esempio 2 – scelta di rinnovare la polizza assicurativa/2

	stime	errore standard	χ^2	p-value	OR
intercetta	-8.4349	0.0854	9760.72	<.0001	
x_1	0.0223	0.0004	2967.84	<.0001	1.023
x_2	0.7431	0.0191	1512.45	<.0001	2.102
x_3	0.8237	0.0186	1862.48	<.0001	2.279

exp(.8237)

Modello stimato

$logit(p_i) = -8.439+0.0223x_1+0.7431x_2+0.8237x_3$

$OR_{x_3} = 2.279$ i clienti che si servono di un ufficio in centro hanno una propensione a rinnovare la polizza 2.3 volte maggiore di chi va altrove (mantenendo costanti le altre variabili)

Per x_1 : per un aumento di 1 anno di età, l'incremento nel $logit(p_i)$ è pari a .0223

$ORx_1 = \exp(.0223) = 1.023$ per ogni anno di età in più, incremento in OR pari al 2.3% (per 5 anni in più: $\exp(5 * .0223) = 1.118$)

(analogo ragionamento per x_2 reddito (in migliaia di euro)

Esempio 2 – scelta di rinnovare la polizza assicurativa/3

Stima *P* (rinnovo polizza) per un cliente con:

x_1 = 58 anni
x_2 = 50 (reddito in migliaia di euro)
x_3 = ufficio non in centro

$$P(y = 1|x_1 = 58, x_2 = 50, x_3 = 0) = \frac{\exp(-8.439 + .0223*58 + .7431*50 + .8237*0)}{1 + \exp(-8.439 + .0223*58 + .7431*50 + .8237*0}$$

L'azienda conclude che:

- età e reddito influiscono sul rinnovo della polizza e decidono di promuoverlo ai clienti con disponibilità finanziarie e non troppo giovani
- anche l'ubicazione dell'ufficio ha un ruolo importante, quindi in una politica di espansione delle sedi, sarà preferibile considerare zone centrali

Effetti di interazione tra var esplicative

Parameter	Estimate (S.E)	Number of Cases
Constant	-3.56 (0.15)	6119
Preceding Birth Interval		
19 months or longer	0.00	5530
13-18 months	0.35 (0.21)	447
Less than 13	0.74 (0.30)	142
Preceding Child		
Alive	0.00	5626
Dead	0.96 (0.17)	493
Region		
Other	0.00	5423
Central	0.82 (0.16)	696
Parity		
Less than 6	0.00	3351
6 or more	0.45 (0.17)	2768
Sex		
Male	0.00	3160
Female	-0.94 (0.23)	2959
Maternal education		
Less than 5 years	0.00	3732
5-8 years	0.42 (0.17)	970
9+	0.27 (0.17	1417
Interaction Term		
Parity.sex		
6+.Female	0.77 (0.28)	

Modello logit per la probabilità di morte neonatale (Ghana da DHS Program – Demographic and Health Surveys, United States Agency for International Development - USAID

Variabile interazione

$z = Parity * Female$

Rappresentazione grafica effetto di interazione /1

Calcolo delle probabilità relative a "parity" e "sex"
(con modalità baseline per le altre variabili)

$$P_{ij} = \frac{exp\,(constant + Parity(i) + Sex(j) + Parity.Sex(ij))}{1 + exp\,(constant + Parity(i) + Sex(j) + Parity.Sex(ij))}$$

For PARITY= less than 6, SEX= male, the probability of neonatal death is given by

P_{11} = EXP (-3.56 + 0 + 0 + 0)/[1+EXP(-3.56)]= 0.028

For PARITY= less than 6, SEX=female

P_{12} = EXP (-3.56 +0 -0.94 +0)/[1+EXP(-3.56 +0 -0.94 +0)]

= 0.011

For PARITY= 6+, SEX=male

P_{21} = EXP (-3.56 +0.45 +0 +0)/[1+EXP(-3.56 +0.45 + 0 +0)]

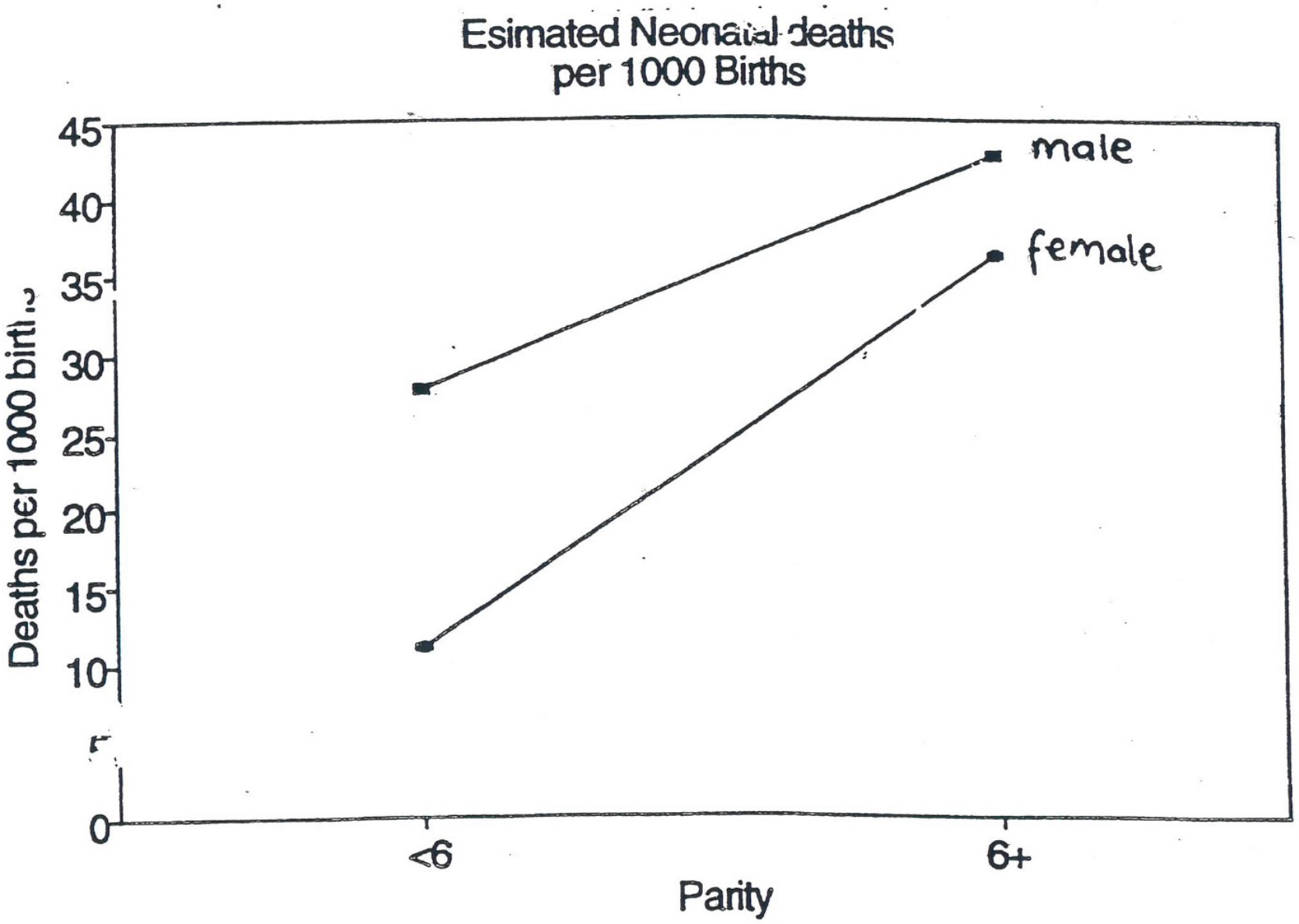
= 0.043

For PARITY= 6+, SEX=female

P_{22} = EXP (-3.56 +0.45 -0.94 +0.77) / [1+EXP(-3.56 +0.45 -0.94 +0.77)]

= 0.036

Rappresentazione grafica effetto di interazione /2



La condizione più favorevole alla sopravvivenza delle bambine si riduce (e diventa molto simile ai maschi) se l'ordine di parità è elevato

Costruzione modello e selezione variabili

Obiettivo: selezionare le variabili per il "migliore" modello nello specifico contesto applicativo

Ingredienti necessari:

- 'procedura' da seguire per la selezione (ragionata, metodi stepwise,…)
- metodo/i per valutare l'adeguatezza (singole variabili che modello globale)

Hosmer, Lemeshow (2000):

Successful modeling of a complex data set is part ***science****, part* ***statistical methods****, and part* ***experience*** *and* ***common sense***.