Modelli per dati binari

Antonio Lucadamo

antonio.lucadamo@unisannio.it

- Innumerevoli contesti applicativi
- Biostatistica: modellare l'effetto dell'attitudine al fumo, del colesterolo, della pressione (e altri fattori di rischio) sull'insorgere di problemi cardiaci
- Scienze sociali: modellizzazione di opinioni, comportamenti
- Marketing: modellizzazione del comportamento dei clienti
- Finanza: modellizzazione di risposte legate al credito

- Dati binari: non raggruppati o raggruppati
- Non raggruppati: ciascuna osservazione è la realizzazione di un processo dicotomico

$$y_i = \{0, 1\}, i = 1, 2, \ldots, n$$

- Raggruppati: sottoinsiemi di osservazioni presentano lo stesso valore delle covariate (studi *dose-response*)

$$y_i = \{0, 1, 2 \ldots, m_i\}, i = 1, 2, \ldots, n$$

- Dati raggruppati possono essere convertiti in dati non raggruppati.
- Viceversa, dati non raggruppati possono essere convertiti in dati raggruppati solo quando più unità presentano lo stesso valore delle covariate
- Le SMV e i corrispondenti errori standard sono gli stessi ma cambiano altre quantità come ad esempio la devianza
- Nel caso di dati raggruppati, la teoria asintotica richiede che $m_i \to \infty$, in quanto n è fissato (corrispondente alle combinazioni dei livelli/valori delle covariate)

- Specificazione del modello

$$Y_i \sim Bin(m_i, \mu_i),\ \mu_i = F^{-1}(\eta_i),\ \eta_i = \beta^T x_i$$

- Quando $m_i = 1, \forall i \Rightarrow$ Dati binari
- La funzione di log-verosimiglianza è

$$\ell(\beta) = \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + m_i \log(1-\mu_i) \right\}$$

- Il parametro canonico (naturale) è il log-odds (logaritmo della quota, *logit*)
- La devianza è

$$D(y, \hat{\mu}) = 2 \sum_{i=1}^{n} \left\{ y_i \log \frac{y_i}{\hat{y}_i} + (m_i - y_i) \log \frac{m_i - y_i}{m_i - \hat{y}_i} \right\},\ \hat{y}_i = m_i \hat{\mu}_i$$

- La statistica di Pearson è

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - m_i \hat{\mu}_i)^2}{m_i \hat{\mu}_i (1 - \hat{\mu}_i)}$$

Modello con variabile latente

- La risposta y_i^* è continua ma non osservabile

$$y_i^* = \eta_i + \epsilon_i$$

- $\epsilon_i \sim F(\cdot)$, $E(\epsilon_i) = 0$, i=1, 2, . . . , n
- Esiste una soglia τ tale che osserviamo

$$\begin{cases} y_i = 0, & y_i^* \leq \tau \\ y_i = 1, & y_i^* > \tau \end{cases}$$

- Quindi

$$Pr(Y_i = 0) = Pr(Y_i^* \leq \tau) = Pr(\eta_i + \epsilon_i \leq \tau) = F(\tau - \eta_i)$$
$$Pr(Y_i = 1) = Pr(Y_i^* > \tau) = Pr(\eta_i + \epsilon_i > \tau) = 1 - F(\tau - \eta_i)$$

Modello con variabile latente

- Supponiamo che alcuni studenti rispondano ai quesiti di cui è costituito un esame
- Lo studente possiede un'abilità T
- Una particolare domanda è caratterizzata da un livello di difficoltà d
- Lo studente risponde correttamente $(Y = 1)$ solo se $T > d$
- T è una variabile latente che non osserviamo direttamente mentre osserviamo la variabile dicotomica Y

$$\begin{cases} Y = 0, & T \leq d \\ Y = 1, & T > d \end{cases}$$

Distribuzione di tolleranza

- Consideriamo d fissato ed assumiamo che la distribuzione della variabile latente sia $T \sim N(\mu, \sigma^2)$
- La probabilità che uno studente scelto a caso risponda correttamente alla domanda con difficoltà d è

$$\pi = Prob(T > d) = 1 - \Phi\left(\frac{d-\mu}{\sigma}\right) = \Phi\left(\frac{\mu-d}{\sigma}\right)$$

- Inoltre, possiamo scrivere

$$\Phi^{-1}(\pi) = \frac{\mu}{\sigma} - \frac{d}{\sigma} = \beta_0 + \beta_1 d$$

  che definisce un modello di regressione di tipo *probit*
- Modello probit: distribuzione di tolleranza normale
- Modello logit: distribuzione di tolleranza di tipo logistico
- Linear probability model: la distribuzione di tolleranza è Uniforme, il legame è, quindi, il legame identità
- Il termine tolleranza deriva da studi di tossicità

Interpretazione del modello

- Regressione logistica

$$\mu_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}, \quad \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

- μ_i è monotona rispetto al valore di ciascuna esplicativa, secondo il segno del coefficiente
- Quando $\beta_j = 0$ allora Y è condizionalmente indipendente da X_j
- L'incremento relativo nella probabilità, per una covariata quantitativa

$$\frac{\partial}{\partial x_{ij}} \mu_i = \beta_j \mu_i (1 - \mu_i)$$

è massimo in corrispondenza di quel valore x_{ij} per il quale $\mu_i = 0.5$ e decresce verso zero per μ_i che diventa zero o uno, con la stessa velocità

L'interpretazione degli odds

- La quota si definisce come

$$odds = o = \frac{\mu}{1-\mu}$$

- La quota moltiplicata per 100 può leggersi come il numero atteso di successi ogni 100 insuccessi
- Chiaramente la quota può essere espressa *contro* il verificarsi dell'evento piuttosto che *a favore*
- Vale la relazione inversa $\mu = \frac{o}{1+o}$
- Un vantaggio di natura matematico-statistica, che si evidenzia soprattutto nella fase di modellizzazione, della quota sulla probabilità di successo è che gli odds sono superiormente non limitati

L'nterpretazione degli odds

- Gli odds rappresentano la base per l'attribuzione soggettiva di probabilità
- Supponiamo di non essere capaci di fare una valutazione probabilistica su base frequentista (oggettiva)
- In queste circostanze, invece di valutare direttamente la probabilità associata al verificarsi di un evento, un individuo può esprimere quanto sarebbe disposto a pagare (ricevere) al verificarsi (non verificarsi) dell'evento
- Quando $100 \times o = k$ significa che siamo disposti a pagare 100 euro ogni k euro scommessi sul verificarsi dell'evento

Regressione logistica in tabelle 2×2

Un'unica covariata di tipo dicotomico

Response Y	Covariate X $x = 1$	$x = 0$
$y = 1$	$\mu(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\mu(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y = 0$	$1-\mu(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\mu(0) = \frac{1}{1+e^{\beta_0}}$

- e^{β_1} esprime il rapporto tra le quote

$$OR(X) = e^{\beta_1} = \frac{odds(X=1)}{odds(X=0)} = \frac{\frac{\mu(1)}{1-\mu(1)}}{\frac{\mu(0)}{1-\mu(0)}}$$

- In presenza di più variabili esplicative, la quota è funzione esponenziale di x_j: la quota si moltiplica per $\exp(\beta_j)$ in corrispondenza di un incremento unitario di x_j, a parità delle altre variabili
- $\exp(\beta_j)$ è un OR condizionale, a parità delle altre variabili

Crying of babies

- Consideriamo i seguenti dati provenienti da uno studio condotto nel reparto maternità di un ospedale
- Ogni giorno, per 18 giorni, un solo bambino tra quelli presenti nel nido viene cullato.
- La variabile risposta è il bambino non piange (Y=1), piange (Y=0).
- La variabile esplicativa è il bambino viene cullato (X=1),non viene cullato (X=0)
- Si vuole verificare in che modo il trattamento essere cullato agisca sulla probabilità che i bambini smettano di piangere $\mu(X), X = 0, 1$

	Cullati	
Piange	NO	SI
NO	77	15
SI	48	3

- 15 giorni su 18 il bambino cullato non piange

Crying of babies. Stima del modello

- Le stime delle probabilità che i bambini non piangano nei due gruppi sono

$$\hat{\mu}(1) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1+e^{\hat{\beta}_0+\hat{\beta}_1}} = \frac{15}{18} = 0.833$$

$$\hat{\mu}(0) = \frac{e^{\hat{\beta}_0}}{1+e^{\hat{\beta}_0}} = \frac{77}{125} = 0.616$$

- Gli errori standard sono

$$se\left(\hat{\mu}(1)\right) = \sqrt{\frac{\hat{\mu}(1)\left[1-\hat{\mu}(1)\right]}{18}} = 0.088$$

$$se\left(\hat{\mu}(0)\right) = \sqrt{\frac{\hat{\mu}(0)\left[1-\hat{\mu}(0)\right]}{125}} = 0.044$$

Crying of babies. Stima del modello

- La stima dell'OR a favore dell'evento $Y = 1$ (*il neonato non piange*) è

$$\widehat{OR} = \frac{\widehat{odds}(X = 1)}{\widehat{odds}(X = 0)} = \frac{15 \times 48}{3 \times 77} = 3.117$$

- Quando i bambini vengono cullati la quota attesa di neonati che non piangono diventa circa il triplo.
- Verificare $H_0 : \beta_1 = 0$ equivale a verificare l'ipotesi di indipendenza tra le variabili Y e X
- Verificare $H_0 : \beta_1 = 0$ equivale a verificare l'ipotesi di uguaglianza delle probabilità di successo nei due gruppi, $H_0 : \mu(1) = \mu(0)$
- Il test di indipendenza può essere condotto mediante il test esatto di Fisher o il test asintotico X^2 di Pearson (con correzione di continuità di Yates)

Il rischio relativo e il rapporto delle quote

- Il rischio relativo è definito come

$$r = \frac{\mu(X = 1)}{\mu(X = 0)}$$

  il rapporto fra le probabilità di successo nei due gruppi
- Si verifica che

$$OR = \text{relative risk} \times \frac{1 - \mu(X = 0)}{1 - \mu(X = 1)}$$

- L'OR e il rischio relativo assumono valori molto simili solo quando la probabilità del verificarsi dell'evento in esame è bassa in entrambi i gruppi individuati dalla presenza/assenza di trattamento
- Crying of babies: $\hat{r} = \frac{15/18}{77/125} = 1.353$
- La probabilità attesa di non piangere quando si è cullati è maggiore di circa il 35% della stessa probabilità quando non si è cullati

Studi prospettivi e retrospettivi

- *Prospective sampling*: le covariate sono fissate e la risposta è osservata nel tempo. **Studi per coorte**. Si seleziona un campione di individui con certe caratteristiche di cui si rileva il valore della risposta entro un certo orizzonte temporale
- *Retrospective sampling*: la risposta è fissata e si rilevano i valori delle covariate guardando indietro nel passato. **Studi caso-controllo**.

Studi prospettivi e retrospettivi

Consideriamo uno studio sui problemi respiratori dei neonati. La tabella riporta la proporzione di neonati che manifestano bronchite o pneunoma nel primo anno di vita, classificati in base al sesso e al tipo di nutrimento

	Solo bottiglia	Al seno con supplemento	Solo al seno
M	77/458	19/147	47/494
F	48/384	16/127	31/464

Obiettivo $\Rightarrow$ Capire se ed in che misura il tipo di nutrimento e il sesso dei neonati agiscono sulla probabilità del manifestarsi di complicazioni respiratorie durante il primo anno di vita.

Studi prospettivi e retrospettivi

- Studio prospettivo: si selezionano neonati, maschi e femmine, i cui genitori hanno optato per un metodo di nutrimento e si registra l'eventuale verificarsi di complicazioni respiratorie
- Studio casi-controlli: i neonati sono condotti dal medico; alcuni manifesteranno i problemi respiratori di interesse (gruppo dei *casi*), altri non li manifesteranno (gruppo dei *controlli*). Per tutti i neonati si registra il sesso e il tipo di nutrimento scelto dai genitori

Studi prospettivi e retrospettivi

- Uno studio prospettivo è la scelta ideale.
- Consideriamo le variazioni della quota di neonati con problemi respiratori al variare delle modalità delle esplicative
- Concentriamoci solo sui neonati di sesso maschile.
- Condizionatamente al nutrimento solo al seno

$$\log \frac{\pi}{1-\pi} = \log \frac{47}{494-47} = -2.25$$

- Condizionatamente al nutrimento solo con bottiglia

$$\log \frac{\pi}{1-\pi} = \log \frac{77}{458-77} = -1.60$$

- Il log-OR campionario è $\Delta = -2.25 - (-1.60) = -0.65$
- Quando il il nutrimento al seno sostituisce il nutrimento con bottiglia la quota di neonati con problemi respiratori nel primo anno di vita si dimezza (OR campionario è $e^{-0.65} = 0.52$ circa)

Studi prospettivi e retrospettivi

- Supponiamo che lo studio sia avvenuto in modo retrospettivo $\Rightarrow$ condizioniamoci alla presenza/assenza dei problemi respiratori
- Consideriamo, quindi, la quota di neonati nutriti solo al seno rispetto a quelli nutriti solo con bottiglia, rispettivamente, per i casi e i controlli.
- Condizionatamente al verificarsi dei problemi respiratori

$$\log odds(Y = 1) = \log \frac{47}{77}$$

- Condizionatamente al non verificarsi dei problemi respiratori

$$\log odds(Y = 0) = \log \frac{494 - 47}{458 - 77}$$

- Il log OR campionario

$$\Delta = \log \frac{odds(Y = 1)}{\log odds(Y = 0)} = \log \frac{47}{494 - 47} - \log \frac{77}{458 - 77} = -0.65$$

Studi prospettivi e retrospettivi

- Studi prospettivi e retrospettivi conducono alle stesse stime degli OR
- Il calcolo dell'OR è simmetrico rispetto alle variabili coinvolte
- Per una tabella 2×2, per il Th. di Bayes

$$
\begin{aligned}
e^{\beta_1} &= \frac{Pr(y=1|x=1)/Pr(y=0|x=1)}{Pr(y=1|x=0)/Pr(y=0|x=0)} \\
&= \frac{Pr(x=1|y=1)/Pr(x=0|y=1)}{Pr(x=1|y=0)/Pr(x=0|y=0)}
\end{aligned}
$$

- Il risultato è vero solo quando si usa la funzione del legame canonico
- In generale, con la regressione logistica si possono stimare gli effetti anche invertendo il ruolo di y e x dato che i coefficienti sono interpretabili in termini di variazione nel logit.
- Gli studi retrospettivi sono più veloci ed economici, quindi più convenienti ma meno accurati nel senso della qualità dei dati

Studi prospettivi e retrospettivi

Analizziamo più in dettaglio le differenze tra i due schemi di campionamento in esame

- Sia Z una variabile casuale dicotomica che descrive l'inclusione (non inclusione) nello studio
- Assumiamo che la selezione di casi e controlli sia indipendente dalle covariate X
- Sia ξ_1 la probabilità di inclusione nello studio nel gruppo dei casi $\xi_1 = Prob(Z = 1|Y = 1, X) = Prob(Z = 1|Y = 1)$
- Sia ξ_0 la probabilità di inclusione nello studio nel gruppo di controllo $\xi_0 = Prob(Z = 1|Y = 0, X) = Prob(Z = 1|Y = 0)$
- ξ_0 e ξ_1 non dipendono da x
- In uno studio prospettivo, $\xi_0 = \xi_1$
- In uno studio retrospettivo, $\xi_0 << \xi_1$, ci sono più casi che controlli

Studi prospettivi e retrospettivi

- Sia $\mu(x) = Prob(Y = 1|X = x)$ la probabilità del manifestarsi dei problemi respiratori
- Sia $\mu^*(x) = Prob(Y = 1|X = x, Z = 1)$ la probabilità del manifestarsi dei problemi respiratori, **condizionata** al fatto di essere incluso nello studio.
- Applicando il Teorema di Bayes,

$$\begin{aligned}
\mu^*(x) &= \frac{Pr(Z = 1|Y = 1, X)Pr(Y = 1|X)}{Pr(Z = 1|Y = 0, X)Pr(Y = 0|X) + Pr(Z = 1|Y = 1, X)Pr(Y = 1|X)} \\
&= \frac{\xi_1 \mu(x)}{\xi_0(1 - \mu(x)) + \xi_1 \mu(x)} \\
&= \frac{\xi_1 e^{\eta}}{\xi_0 + \xi_1 e^{\eta}}
\end{aligned}$$

- Si ricava

$$logit(\mu^*(x) = \log \frac{\xi_1}{\xi_0} + \eta$$

Studi prospettivi e retrospettivi

- Il modello di regressione logistica per $\mu(x)$ prevede che

$$logit(\mu(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p$$

- Il modello di regressione logistica per $\mu^*(x)$ prevede che

$$\begin{aligned} logit(\mu^*(x)) &= \log \frac{\xi_1}{\xi_0} + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p \\ &= \beta_0^* + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p \end{aligned}$$

con $\beta_0^* = \beta_0 + \log \frac{\xi_1}{\xi_0}$

Studi prospettivi e retrospettivi

- Ad eccezione del termine d'intercetta, gli altri coefficienti non cambiano, per cui le stime degli OR sono affidabili indipendentemente dalla natura dello schema di campionamento
- La quantità $\log \frac{\xi_1}{\xi_0}$ solitamente non è nota, per cui non siamo in grado di stimare β_0^*
- In uno studio casi-controlli è possibile stimare l'effetto delle esplicative in termini di OR anche se la stima del termine d'intercetta è distorta
- La stima dei $\beta_j, j \neq 0$ non è influenzata dal fatto che i dati siano raccolti retrospettivamente, purchè il modello includa un termine d'intercetta
- In uno studio prospettivo $\log \frac{\xi_1}{\xi_0} = 0$
- Un modello *probit* non si presta al trattamento di dati raccolti retrospettivamente

Dati sui problemi respiratori

- Consideriamo il modello con i soli **effetti principali**

$$logit(\mu) = \beta_0 + \beta_1 M + \beta_2 BS + \beta_3 S$$

  caratterizzato da esplicative dicotomiche, con $M = 1$ per i neonati di sesso maschile, $BS = 1$ per neonati allattati al seno con aggiunta e $S = 1$ per i neonati allattati solo al seno
- Il gruppo di riferimento è costituito da neonati di sesso femminile allattati con bottiglia
- Abbiamo facoltà di modificare la composizione del gruppo di riferimento in base alle nostre esigenze

Dati sui problemi respiratori

- Il modello specificato postula indipendenza tra sesso e nutrimento (possiamo verificare tale ipotesi)
- Il TRV conduce a preferire il modello **semplificato** senza interazione (p-valore: 0.697)
- Le frequenze attese

	Solo bottiglia	Al seno con supplemento	Solo al seno
M	76/458	21/147	46/494
F	49/384	14/127	32/464

- Gli OR attesi (con IC al 95% di tipo Wald)

OR(M)	1.367 (1.037, 1.802)
OR(BS|B)	0.842 (0.562,1.259)
OR(S|B)	0.512 (0.379,0.691)
OR(S|BS)	0.609 (0.398, 0.930)

LD50 (ED50)

- In quei casi in cui disponiamo di un'unica covariata X continua, o possiamo considerare fisse le altre covariate, a volte è utile stimare la quantità x che corrisponde ad un assegnato valore di μ
- Si definisce LD50 *lethal dose* o ED50 *effective dose* il valore x che corrisponde a $\mu = 0.50$
- Parleremo di LD50 quando il successo consiste nella morte di insetti, ad esempio come accade negli studi di tossicità, altrimenti parleremo di ED50
- Nel caso in cui c'è una sola esplicativa, si verifica che

$$\widehat{ED50} = -\frac{\hat{\beta}_0}{\hat{\beta}_1}$$

- In generale, la dose efettiva x_μ, per la probabilità μ è

$$x_\mu = \frac{logit(\mu) - \hat{\beta}_0}{\hat{\beta}_1}$$

Classificazione

- Classificazione delle unità statistiche sulla base dei valori osservati della risposta $y = 0$ o $y = 1$ e dei valori previsti della risposta $\hat{y} = 0$ o $\hat{y} = 1$
- Tabella 2×2 di classificazione (*matrice di confusione*)
- Si può porre $\hat{y}_i = 1$ quando $\hat{\pi}_i > \hat{\pi}_0$, zero altrimenti
- In genere il valore utilizzato è 0.5
- A volte si preferisce ottenere le previsioni mediante **leave-one-out cross validation**: $\hat{\pi}_i$ è ottenuta dal modello che esclude y_i

Tabella di classificazione

Osservati	Classificati $\hat{Y}$=1	Classificati $\hat{Y}$=0	
Y=1	a Veri positivi	b Falsi negativi	a+b
Y=0	c Falsi positivi	d Veri negativi	c+d
	a+c	b+d	n

- Tasso complessivo di corretta classificazione: $TCC = \frac{a+d}{n}$
- TCC delle unità per le quali $Y = 1$: *sensitività* $\theta = \frac{a}{a+b}$
- TCC delle unità per le quali $Y = 0$: *specificità*= $\gamma = \frac{d}{c+d}$
- Sensitività è una stima di $Pr(\hat{Y} = 1|Y = 1)$
- Specificità è una stima di $Pr(\hat{Y} = 0|Y = 0)$

Curva ROC

- Sensitività e specificità dipendono dal valore soglia k
- Uno strumento che consente di misurare la qualità della classificazione e quindi la capacità previsiva del modello stimato è la curva ROC (Receiver Operating Charcteristic)
- Costruzione della ROC: per diversi valori k, si calcolano sensitività $\theta(k)$ e specificità $\gamma(k)$ e si rappresentano i punti di coordinate $(\theta(k), 1-\gamma(k))$
- $\theta(k) = Pr(\hat{Y} = 1|y = 1)$ è il tasso di veri positivi
- $1-\gamma(k) = Pr(\hat{Y} = 1|y = 0)$ è il tasso di falsi positivi
- $\theta(0) = 1, \gamma(0) = 0$
- $\theta(1) = 0, \gamma(1) = 1$

Curva ROC

- Per ciascuna specificità si richiede che il modello abbia un'elevata sensitività, che si traduce in maggiore capacità predittiva, intesa come capacità di **discriminare** tra casi e controlli sulla base del valore delle esplicative
- Maggiore l'area sotto la curva ROC (AUC), migliore la capacità predittiva del modello stimato
- Sensitività e specificità dipendono dalla dimensione del gruppo dei casi e dei controlli, rispettivamente
- È maggiore il tasso di classificazione nel gruppo al quale appartengono più unità, indipendentemente dalle stime ottenute

AUC

- L'area sotto la curva ROC fornisce una misura della capacità del modello stimato di discriminare tra i soggetti per i quali $Y = 1$ e $Y = 0$
- Area ROC $\in (0.5, 1)$
- Regola generale
  - $Area = 0.5 \Rightarrow$ non c'è discriminazione
  - $0.7 \leq Area < 0.8 \Rightarrow$ discriminazione accettabile
  - $0.8 \leq Area < 0.9 \Rightarrow$ discriminazione ottima
  - $Area \geq 0.9 \Rightarrow$ discriminazione eccellente
- Osserviamo che un'area maggiore di 0.9 equivale ad una situazione di quasi completa separazione

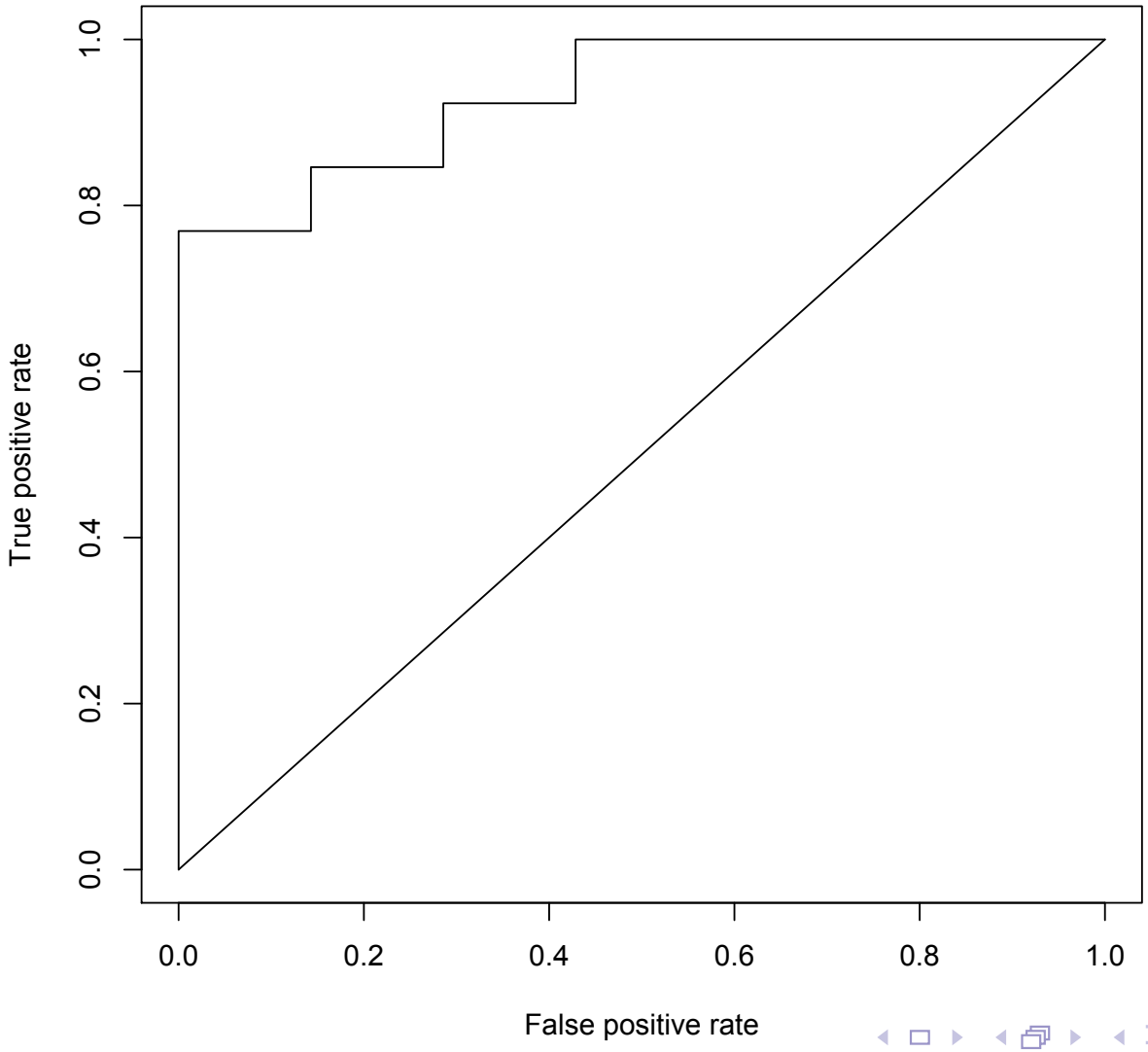
Esempio numerico

- Alcuni valori

k	θ	γ	$1-\gamma$
0.15	1.000	0.429	0.572
0.30	0.923	0.571	0.429
0.50	0.846	0.714	0.286
0.70	0.769	0.857	0.143
0.85	0.769	1.000	0.000
0.90	0.692	1.000	0.000

- $AUC = 0.934$

Esempio numerico



Procedure d'inferenza

- Per verificare $H_0 : \beta_j = 0$ possiamo utilizzare il test di Wald (versione con segno) $\hat{\beta}_j / se(\hat{\beta}_j)$ o il TRV (versione quadratica) dato dalla differenza tra la devianza del modello ridotto e del modello completo.
- Risultati analoghi per numerosità campionarie sufficientemente elevate.
- Il test di Wald presenta due tipi di problemi: non è invariante rispetto alla parametrizzazione, è meno potente del TRV e può assumere un comportamento aberrante quando $\hat{\beta}_j$ assume valori molto grandi (in valore assoluto)

Alcuni problemi numerici

- Presenza di frequenze nulle

Y	x_1	x_2	x_3	
1	7	12	20	39
0	13	8	0	21
	20	20	20	60
$\widehat{OR}$	1	2.79	∞	

- Le stime degli OR sono state ottenute considerando il livello x_1 della covariata come gruppo di riferimento
- L' OR ottenuto in corrispondenza di $X = x_3$ diventa infinito
- Quando $X = x_3$ si realizza un adattamento perfetto (*perfect fit*), in quanto sappiamo che per tutte le unità statistiche per le quali $X = x_3$, la risposta è $Y = 1$
- Soluzioni: aggiungere $1/2$ a ciascuna frequenza o unire alcuni livelli in modo da eliminare le frequenze nulle

Alcuni problemi numerici

- Sia X una covariata (o una collezione di covariate, un iperpiano) che discrimina perfettamente, i.e. separa completamente le unità per le quali $Y = 1$ da quelle per le quali $Y = 0$: *complete separation*
- Ad esempio può accadere che per tutti gli individui di età inferiore a $k \Rightarrow Y = 1$, mentre per tutti quelli di età superiore a $k \Rightarrow Y = 0$. La conoscenza dell'età equivale a conoscere i valori della risposta $\Rightarrow$ *perfect fit, perfet discrimination*
- La SMV di β non esiste. La SMV esiste solo quando esiste sovrapposizione tra i dui gruppi nella distribuzione della covariata
- In presenza di stime infinite l'inferenza basata sul test di Wald non è affidabile ma è necessario ricorrere all'inferenza basata sul TRV
- Si verifica che gli errori standard crescono più rapidamente delle stime dei coefficienti, quando queste divergono

Alcuni problemi numerici

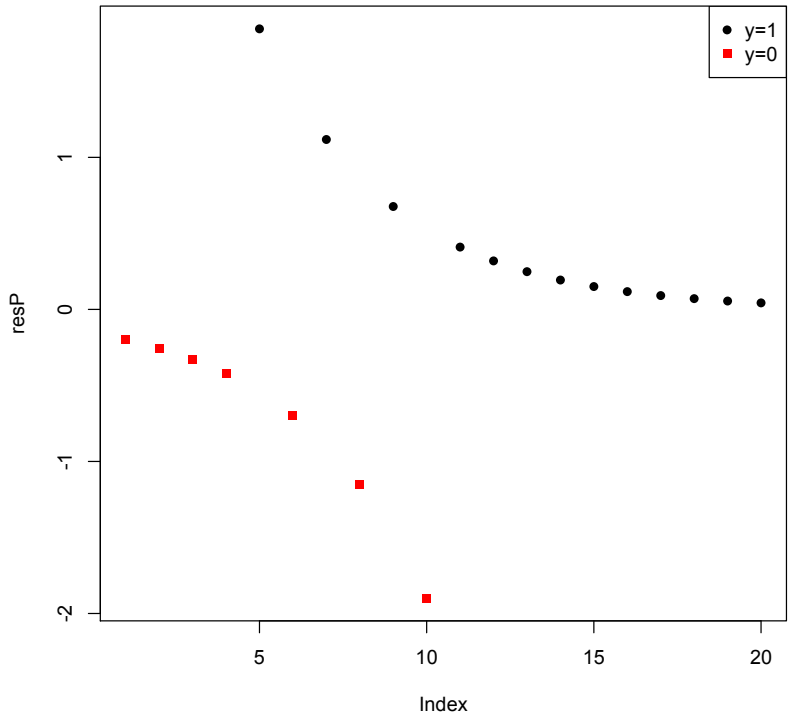
- In pratica, il software di stima può non accorgersi di SMV infinite e produrre risultati non affidabili
- Dopo un certo numero di iterazioni, l'algoritmo IWLS convergenze in quanto la log-verosimiglianza raggiunge un valore limite al crescere illimitato della stima di un coefficiente
- Una situazione analoga, più debole, che può condurre a sMV infinite è quella di *quasi complete separation*: un iperpiano separa i valori delle variabili esplicative per le quali $y = 1$ da quelle per cui $y = 0$ ma esistono casi con entrambe le risposte su quell'iperpiano.
- nonostante SMV infinite, è possibile procedere nell'applicazione del TRV
- Altri problemi di stima si verificano in situazioni di collinearità
- I problemi numerici legati alla presenza di frequenze nulle, perfetta separazione, collinearità si manifestano attraverso errori standard grandi in maniera irrealistica

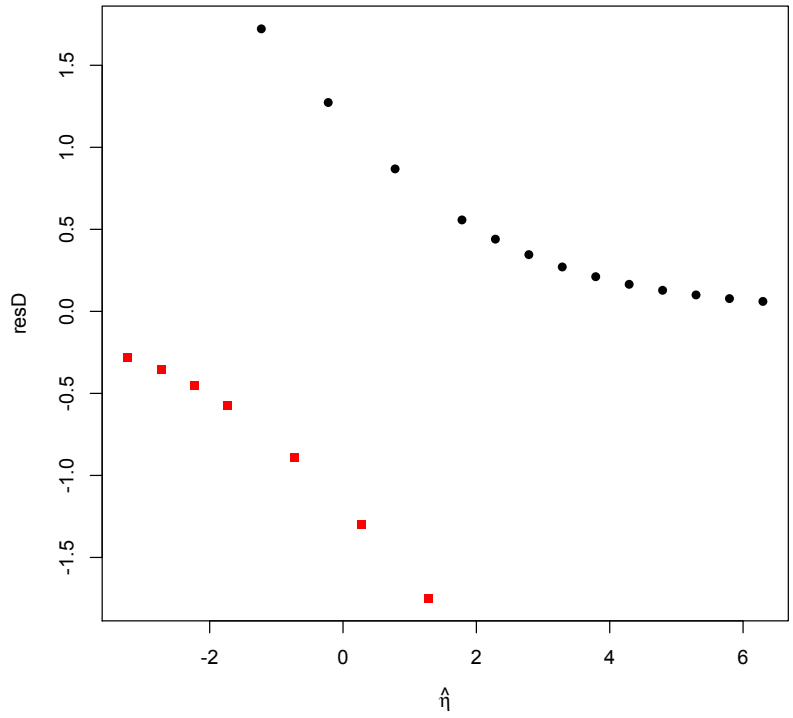
Bontà di adattamento

- Nel caso di dati raggruppati in classi di rischio, definite in corrispondenza di alcuni valori delle covariate, la bontà dell'approssimazione χ^2_{n-p}, dove n è il numero di classi, alla distribuzione della devianza o della statistica di Pearson, dipende dalla numerosità m_i delle singole classi.
- Si richiede che m_i, $\forall i$, sia sufficientemente grande (small-dispersion asymptotics)
- L'approssimazione non è soddisfacente quando il numero n di classi è molto grande, che significa disporre di numerose variabili esplicative (aumenta p) o che una variabile assume troppi valori distinti sulle unità (complicandone l'organizzazione in classi)
- Nel casi di dati dicotomici, l'approssimazione non vale
- In ogni caso, un valore grande della devianza o della statistica di Pearson indica mancanza di bontà di adattamento ma non è informativa circa la natura di questa carenza.

Residui

- Data la natura dicotomica della risposta, il residuo può assumere due valori (a seconda che y=0 o y=1)
- Quando possibile, è preferibile costruire i residui dopo aver raggruppato i dati





Modelli per dati dipendenti

- La risposta dicotomica viene osservata in occasioni diverse sulle stesse unità statistiche
- Le unità statistiche sono abbinate sulla base di certe variabili a formare dei *matched set*: ciascun caso è abbinato con uno (studi caso-controllo 1-1) o più (studi caso-controllo 1-M) controlli per i quali si osservano gli stessi valori (o anche solo simili) di certe variabili: otteniamo J insiemi abbinati ciascuno dei quali sarà costituito da 1 caso e M controlli (il numero di controlli può essere variabile)
- Non è più possibile valutare l'effetto sulla risposta di quelle variabili sulla base delle quali è stato realizzato l'abbinamento, in quanto la valutazione dipenderà necessariamente dal disegno adottato
- Nel confronto fra proporzioni bisogna tener conto del fatto che le risposte dello stesso soggetto o di soggetti abbinati sono statisticamente dipendenti tra loro

Giudizio sul primo ministro

	Second survey		
First Survey	Approve	Disapprove	
Approve	794	150	944
Disapprove	86	570	656
	880	720	**1600**

- 1600 cittadini canadesi sono stati intervistati in due occasioni successive e chiamati ad esprimere un giudizio sull'operato del Primo Ministro
- Ci sono 1600 dati appaiati (*matched pairs*)
- Ci interessa capire se ed in che modo il giudizio sull'operato del Primo Ministro è cambiato da una rilevazione a quella successiva
- I soggetti che non cambiano opinione sono 1364
- Chiaramente c'è forte associazione tra le opinioni registrate nelle due occasioni.

Giudizio sul primo ministro

- Verificare l'ipotesi che il giudizio non sia cambiato significa verificare l'uguaglianza delle distribuzioni marginali
- Le frequenze marginali di ottenere un giudizio favorevole sono, rispettivamente alla prima e seconda rilevazione, $\hat{\pi}_{1\cdot} = \frac{944}{1600} = 0.59$, $\hat{\pi}_{\cdot 1} = \frac{880}{1600} = 0.55$
- Indichiamo con $\hat{\pi}_{ij}, i, j = 1, 2$, le frequenze congiunte
- Consideriamo che $\hat{\pi}_{1\cdot} = \hat{\pi}_{11} + \hat{\pi}_{12}$ e $\hat{\pi}_{\cdot 1} = \hat{\pi}_{11} + \hat{\pi}_{21}$
- Ne segue che $\hat{\pi}_{1\cdot} - \hat{\pi}_{\cdot 1} = \hat{\pi}_{12} - \hat{\pi}_{21}$
- Verificare che il giudizio non sia cambiato equivale a verificare l'ipotesi nulla $H_0 : \pi_{12} - \pi_{21} = 0$

Test di omogeneità marginale

- Consideriamo le frequenze assolute, in particolare le frequenze sulla diagonale secondaria
- McNemar's Test

$$Z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \xrightarrow{d} N(0,1)$$

- $H_0 : \pi_{12} - \pi_{21} \leq 0$ vs $H_1 : \pi_{12} - \pi_{21} > 0$
- $z^{oss} = \frac{150-86}{\sqrt{150+86}} = 4.2$, $\alpha^{oss} = 1 - \Phi(4.2) = 1.3 \times 10^{-5}$
- Evidenza forte contro l'ipotesi nulla e a sostegno del fatto che il consenso verso l'operato del Primo Ministro è diminuito.
- Per costruire un intervallo di confidenza asintotico di tipo Wald per $(\pi_{.1} - \pi_{1.})$, abbiamo bisogno dell'errore standard $se(\hat{\pi}_{1.} - \hat{\pi}_{.1})$ associato alla sua stima

$$\left[\hat{\pi}_{1.}\left(1 - \hat{\pi}_{1.}\right) + \hat{\pi}_{.1}\left(1 - \hat{\pi}_{.1}\right) - 2\left(\hat{\pi}_{11}\hat{\pi}_{22} - \hat{\pi}_{12}\hat{\pi}_{21}\right)\right]^{1/2}$$

Regressione logistica per matched pairs

- $n = 1600$ soggetti intervistati in 2 occasioni
- Classifichiamo le risposte per ciascun individuo singolarmente
- Otteniamo $n_{11} = 794$ tabelle parziali 2×2

	Response	
Survey	Approve	Disapprove
First	1	0
Second	1	0

- In maniera analoga si costruiscono le rimanenti tabelle parziali

Regressione logistica per matched pairs

- Modelliamo la probabilità di approvare l'operato del Primo Ministro per ciascun individuo
- Consideriamo un modello con un parametro α_i specifico per ciascun individuo

$$logit(\mu_i) = \alpha_i + \beta^T x_i, i = 1, 2, \ldots, 1600, \mu_i = Prob(Y_i = 1|x)$$

con $x = 0$ in corrispondenza della prima rilevazione, $x = 1$, in corrispondenza della seconda

- Alla prima rilevazione $\mu_i = \frac{e^{\alpha_i}}{1+e^{\alpha_i}}$
- Alla seconda rilevazione $\mu_i = \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}$
- Per ciascun soggetto, la variazione della quota di voti favorevoli è la stessa

$$OR = \frac{odds(x = 1)}{odds(x = 0)} = e^{\beta}$$

Regressione logistica per matched pairs

- Il valore $\beta = 0$ implica omogeneità marginale. In tal caso la probabilità di successo (in questo caso di approvare l'operato del Primo Ministro) per ciascun soggetto è la stessa nelle due occasioni
- Inferenza sul parametro β. Gli α_i sono parametri di disturbo
- Problema: al crescere del numero di unità statistiche n aumenta anche il numero di parametri da stimare $\Rightarrow$ Problemi con il metodo della massima verosimiglianza
- Soluzione: eliminare i parametri di disturbo α_i mediante condizionamento

Regressione logistica per matched pairs

- Consideriamo le probabilità condizionate

$$\mu_i = Prob\,(Y_i = 1|S_i = s, x)$$

dove S è il numero di volte che ciascun individuo esprime un giudizio favorevole

- $S_i \sim Binom(2, \mu_i)$, $S_i = \{0, 1, 2\}$
- Si verifica che

$$P(Y_{i1} = 0|S_i = 1) = \frac{e^{\beta}}{1 + e^{\beta}}$$

è il contributo delle coppie per le quali $(Y_{i1} = 0, Y_{i2} = 1)$

$$P(Y_{i1} = 1|S_i = 1) = P(Y_{i2} = 1|S_i = 1) = \frac{1}{1 + e^{\beta}}$$

è il contributo delle coppie per le quali $(Y_{i1} = 1, Y_{i2} = 0)$

- Il condizionamento rispetto a $S = 0, 2$ non contribuisce all'inferenza su β

Regressione logistica per matched pairs

- Verosimiglianza condizionata

$$L_C(\beta) = \prod_{i=1}^{n} P(Y_{i1} = y_{i1} | S_i = 1), y_{i1} = 0, 1$$

- Si verifica agevolmente che

$$\log L_C(\beta) = \ell_c(\beta) = 86 \log \left(\frac{e^\beta}{1 + e^\beta} \right) - 150 \log \left(1 + e^\beta \right)$$

da cui

$$\widehat{OR} = e^{\hat{\beta}} = \frac{n_{21}}{n_{12}} = \frac{86}{150} \approx 0.57$$

- La quota di giudizi positivi alla seconda rilevazione è diminuita del 43% rispetto alla prima rilevazione

Myocardial infarction among Navajo Indians

Diabetes	Myocardial Cases	Myocardial Controls	
Yes	46	25	71
No	98	119	217
	144	144	**288**

- 144 vittime (tra gli indiani Navajo) di MI sono state abbinate con 144 persone che non soffrono di MI sulla base del sesso e dell'età
- Ai soggetti è stato poi chiesto se hanno sofferto di diabete o meno
- A ciascun caso è abbinato un controllo
- Attenzione: la distribuzione marginale della risposta è fissata: un mezzo del campione soffre di infarto al miocardio

Myocardial infarction among Navajo Indians

- Nell'esempio disponiamo di $I = 144$ *matched set*

Matched Set	MI case	MI control
1	YES	NO
2	YES	YES
3	NO	YES
4	NO	YES
...	...	...

- 144 tabelle parziali 2×2 per il singolo insieme abbinato
- Ad esempio per il primo abbinamento

	Diabetes	
Infarction	Yes	No
Case	1	0
Control	0	1

- Ci sono 4 possibili modi con cui possono combinarsi le risposte che danno luogo ad altrettante tabelle parziali

Myocardial infarction among Navajo Indians

Riportiamo le frequenze osservate per ciascuna tabella parziale

	MI controls		
MI cases	Diabetes	No Diabetes	Total
Diabetes	9	37	46
No Diabetes	16	82	98
Total	25	119	**144**

Myocardial infarction among Navajo Indians

- Consideriamo il modello

$$logit(\mu_{ij}) = \alpha_i + \beta^T x_{ij}, i = 1, 2, \ldots, 144, \ j = 0, 1$$

  $j = 0$ per il caso, $j = 1$ per il controllo
- $x_{ij} = 1$ per i diabetici, zero altrimenti
- Le probabilità di infarto μ_{ij} variano tra gli insiemi abbinati.
- I parametri α_i modellano l'effetto delle variabili secondo le quali è stato realizzato il matching
- La stima del rapporto tra la quota di vittime di MI tra i diabetici e la quota di vittime di MI tra i non diabetici

$$OR = e^{\hat{\beta}} = \frac{37}{16} = 2.3$$