Modelli Lineari Generalizzati (GLM): parte II

Leonardo Egidi

A.A. 2021/2022

Università di Trieste
Corso di laurea magistrale in Scienze Statistiche ed Attuariali

Indice

Modelli per tabelle di dati di frequenze

Schemi di campionamento

- Per l'inferenza si deve individuare il *meccanismo probabilistico generatore dei dati*, che può essere di tipi diversi a seconda del modo in cui i dati sono stati raccolti. I modelli adatti a descrivere dati di conteggio sono differenti a seconda dello schema di campionamento, della presenza o meno di variabili esplicative e della loro natura.
- I dati riassunti in una tabella di frequenze possono quindi essere generati da diversi schemi di campionamento. In dettaglio, la numerosità complessiva può essere prefissata, oppure può ritenersi realizzazione di una variabile casuale.

Tabella di frequenze

	B_1	B_2	...	B_J	Totale
A_1	y_{11}	y_{12}	...	y_{1J}	$y_{1\cdot}$
A_2	y_{21}	y_{22}	...	y_{2J}	$y_{2\cdot}$
...	...	...	...	...	...
A_I	y_{I1}	y_{I2}	...	y_{IJ}	$y_{I\cdot}$
Totale	$y_{\cdot 1}$	$y_{\cdot 2}$	...	$y_{\cdot J}$	$y_{\cdot\cdot} = n$

Tabella 1: Tabella di frequenze in base ai fattori A e B, avente $(I+1)$ righe e $(J+1)$ colonne. Riporta le frequenze osservate relativamente a ciascun incrocio dei due fattori in n prove indipendenti. Sono anche riportati i totali marginali, con la convenzione che il segno '$\cdot$' denota una somma delle frequenze rispetto all'indice corrispondente.

Osservazione diretta del fenomeno (Poisson)

Per i dati di conteggio con numerosità non prefissata, noi osserviamo il fenomeno per un dato periodo e classifichiamo gli eventi. Allora, è ragionevole assumere che le y_{ij} della tabella siano realizzazioni di variabili casuali indipendenti con distribuzione di Poisson con media μ_{ij}, $i = 1, \ldots, I$ e $j = 1, \ldots, J$. Ossia

$$Y_{ij} \sim \text{Poisson}(\mu_{ij}) \ ,$$

perciò

$$Pr(\boldsymbol{Y} = \boldsymbol{y}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \ , \tag{1}$$

con $\{\mu_{ij}\}$ parametri incogniti da stimare.

Osservazione per un numero fissato di eventi (Multinomiale)

Si decide preliminarmente di raccogliere i dati relativi a n unità, invece di fissare il tempo di osservazione del fenomeno. La distribuzione dei dati cambia. Per $n = \sum_i \sum_j y_{ij}$ fissato, il modello statistico appropriato è la distribuzione multinomiale $Mn_d(n, \pi)$, con $d = I \times J$, $\pi = (\pi_{11}, \ldots, \pi_{IJ})$ e funzione di probabilità:

$$Pr(\boldsymbol{Y} = \boldsymbol{y}) = \frac{n!}{\prod_{i=1}^{I} \prod_{j=1}^{J} y_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}} \ , \tag{2}$$

con $0 < \pi_{ij} < 1$ e $\sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = 1$.

Un totale di riga/colonna prefissato (Prodotto-Multinomiale)

Si può non vincolare solo il numero totale n di osservazioni, ma si impongono valori prefissati a tutta una riga o una colonna di frequenze marginali. Il modello statistico sarà allora il prodotto delle funzioni di probabilità multinomiali relative a ciascuna riga (o colonna).

Esistono definizioni simili degli schemi di campionamento per tabelle con più di due dimensioni.

Rapporto tra schemi di campionamento

- *Osservazione.* Vi è un importante collegamento tra i modelli dei due schemi (1) e (2). Assunto il modello di Poisson, la distribuzione della statistica $n = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}$ è una Poisson con parametro $\mu = \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}$. Allora la distribuzione di $\boldsymbol{y}$ condizionata a n ha funzione di probabilità multinomiale $Mn_d(n, \pi)$, con $\pi_{ij} = \mu_{ij}/\mu$, $i = 1, \ldots, I$ e $j = 1, \ldots, J$, in accordo con il modello (2).

$$
\begin{aligned}
Pr(\boldsymbol{Y} = \boldsymbol{y}|n) &= \frac{\prod_{i=1}^{I} \prod_{j=1}^{J} e^{-\mu_{ij}} \mu_{ij}^{y_{ij}} / y_{ij}!}{e^{-(\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij})} (\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij})^n / n!} \\
&= \frac{n!}{\prod_{i=1}^{I} \prod_{j=1}^{J} y_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{\mu_{ij}}{\mu} \right)^{y_{ij}} .
\end{aligned}
$$

- Il semplice esame dei dati in una tabella di frequenze congiunte non consente di individuare il modello giusto: si deve sapere il modo in cui sono stati raccolti i dati. In base a questo abbiamo forme diverse della verosimiglianza.

Rapporto tra schemi di campionamento

$\Longrightarrow$ **Risultato fondamentale.** Si può mostrare che l'inferenza basata sulla verosimiglianza è essenzialmente la stessa in ciascun caso di raccolta dei dati. Ciò consente di stimare modelli con distribuzione dei dati di tipo multinomiale, usando la verosimiglianza di tipo Poisson.

- In particolare, i valori dei parametri che massimizzano le verosimiglianze sono gli stessi, così come le derivate seconde delle log-verosimiglianze (e quindi anche gli errori standard delle stime coincidono).
- Anche per il caso di una marginale fissata, è possibile sviluppare considerazioni analoghe alle precedenti e constatare l'equivalenza delle verosimiglianze.
- Per i dettagli si veda Azzalini sez. 6.4.1—6.4.3

Modelli log-lineari

- Per il modello di Poisson, le frequenze attese nelle classi sono:

$$E(Y_{ij}) = \mu_{ij} \ .$$

  Nell'ipotesi di indipendenza tra le due variabili della tabella, avremmo

$$\mu_{ij} = \frac{\mu_{i\cdot}\mu_{\cdot j}}{\mu} \ .$$

- Per il modello multinomiale le frequenze attese nelle classi sono:

$$E(Y_{ij}) = n\pi_{ij} \ .$$

  Se le variabili fossero indipendenti, avremmo $\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$.

- Un'analoga argomentazione può essere sviluppata anche per il prodotto di multinomiali.

Modelli log-lineari

- Nelle tabelle di contingenza tutte le ipotesi usuali possono essere formulate come modelli moltiplicativi per le frequenze attese nelle classi.
- Obiettivo è lo studio della relazione tra $E(Y_{ij})$ e il predittore lineare η_{ij}, del tipo

$$\eta_{ij} = \theta + \alpha_i + \beta_j \ ,$$

ove gli α_i e i β_j sono gli *effetti principali*.

- In ciascun caso, una struttura *moltiplicativa* per la media μ_{ij} fornisce una semplificazione naturale. Questo corrisponde a un modello additivo per $\log \mu_{ij}$. Perciò, quello che ci interessa è la possibilità che il *modello log-lineare*

$$\log \mu_{ij} = \theta + \alpha_i + \beta_j$$

(dove $\sum_{i=1}^{I} \alpha_i = 0 = \sum_{j=1}^{J} \beta_j$) fornisca una buona descrizione dei dati.

Modelli log-lineari

- Per questi modelli la funzione logaritmo è il legame naturale tra μ_{ij} e la combinazione lineare dei parametri. Da qui il nome *modello log-lineare*. Inoltre, la funzione logaritmo è il legame canonico per la distribuzione di Poisson. Si comprende quindi che i modelli log-lineari sono quelli più comuni per l'analisi delle tabelle di frequenza. Di fatto, i modelli log-lineari sono rilevanti anche negli altri schemi di campionamento.
- Il modello più ricco di parametri (che è un modello saturo se ho due fattori) può essere scritto come

$$\log \mu_{ij} = \theta + \alpha_i + \beta_j + \delta_{ij}$$

  e quindi l'ipotesi di indipendenza corrisponde all'ipotesi di assenza di interazione, ossia $\delta_{ij} = 0$.

Modelli per tabelle con dimensione maggiore di due

- Nelle tabelle a due entrate di dimensione $I \times J$ relativamente a due variabili X e Z, il termine di interazione conduce al modello saturo.
- Nelle tabelle di dimensione più elevata, quindi con tre variabili, diciamo X, Z e W, ciascuna rispettivamente con I, J e K modalità, vi possono essere più termini di interazione e possono essere interessanti modelli che hanno o meno interazioni significative fra le diverse variabili coinvolte. Ad esempio, per un modello per 3 variabili, si considera il modello log-lineare:

$$\log \mu_{ijk} = \theta + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \epsilon_{ik} + \phi_{jk} + \psi_{ijk}$$

con i vincoli richiesti, per esempio $\epsilon_{i\cdot} = 0 \ \forall i$ e $\psi_{\cdot jk} = 0 \ \forall j, k$. Questo modello è saturo, però un modello del tipo

$$\log \mu_{ijk} = \theta + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \epsilon_{ik}$$

assume una qualche struttura di indipendenza nel modello.

Modello di indipendenza mutua completa

- Può essere di qualche interesse, limitandoci al caso più semplice di tre sole variabili, esaminare quali forme di indipendenza siano implicate dalla presenza/assenza di effetti di interazione.
- Il modello log-lineare

$$\log \mu_{ijk} = \theta + \alpha_i + \beta_j + \gamma_k$$

ove $\delta_{ij} = \epsilon_{ik} = \phi_{jk} = \psi_{ijk} = 0 \; \forall i, j, k$ postulerebbe che le tre variabili coinvolte, diciamo rispettivamente X, Z, W, siano completamente indipendenti. Questo implica che la probabilità che una determinazione campionaria appartenga a una generica cella sia

$$\pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k}$$

- Di fatto è questo il modello log-lineare più semplice che abbia senso considerare.
- L'indipendenza mutua completa implica l'indipendenza di ogni altro tipo.

Modello di indipendenza congiunta

- Il modello log-lineare

$$\log \mu_{ijk} = \theta + \alpha_i + \beta_j + \gamma_k + \delta_{ij}$$

ove

$$\epsilon_{ik} = \phi_{jk} = \psi_{ijk} = 0 \ \forall i, j, k$$

implica che che W sia congiuntamente indipendente da X e Z. In questo caso $\pi_{ijk} = \pi_{ij.}\pi_{..k} \ \forall i, j, k$ e quindi è come postulare l'indipendenza fra due variabili: le probabilità congiunte per X e Z saranno le stesse per ogni livello di W. Si noti che possono esistere tre diversi modelli di indipendenza congiunta.

Modello di indipendenza condizionale

- Il modello log-lineare

$$\log \mu_{ijk} = \theta + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \epsilon_{ik}$$

  ove

$$\phi_{jk} = \psi_{ijk} = 0 \ \forall i, j, k$$

  implica che vi sia indipendenza fra Z e W condizionatamente a X. Se si denota con $\pi_{jk|i} = \frac{\pi_{ijk}}{\pi_{i..}}$ la distribuzione congiunta di W e Z al livello i di X allora $\pi_{jk|i} = \pi_{j\cdot|i}\pi_{\cdot k|i} \ \forall j, k$. Ovvero nelle tabelle a doppia entrata per W e Z per ogni livello di X vi sarà indipendenza.

- Anche l'indipendenza condizionale può presentarsi in tre forme diverse.

- È importante ricordare che l'indipendenza tra due variabili condizionatamente a una terza non implica che vi sia indipendenza marginale tra le due variabili. Ovvero se sommiamo rispetto alla variabile di condizionamento non è detto che la distribuzione di probabilità che si ottiene presenti indipendenza.

Osservazioni

- Si noti anche che l'indipendenza congiunta tra diciamo W e la coppia X, Z implica anche l'indipendenza condizionale tra W e X.
- Se in un modello sono presenti le tre interazioni di secondo ordine ovvero il modello ha la forma

$$\log \mu_{ijk} = \theta + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \epsilon_{ik} + \phi_{jk}$$

  esiste una relazione fra ogni coppia di variabili separatamente e queste relazioni non dipendono dalla terza variabile e infatti si pone $\psi_{ijk} = 0 \ \forall i, j, k$.
- Il passo successivo è il modello saturo che include anche interazioni ψ_{ijk} non nulle.
- Le stesse interpretazioni delle diverse forme di indipendenza possono essere convenientemente estese al caso in cui si analizzano più di tre variabili.

Stima dei parametri

- L'inferenza può essere basata sul modello di Poisson in quanto l'inferenza basata sulla verosimiglianza è essenzialmente la stessa in ciascun caso di raccolta dei dati.
- Ciò consente di stimare modelli con distribuzione dei dati di tipo multinomiale, usando la verosimiglianza di tipo Poisson. In particolare, i valori dei parametri che massimizzano le due verosimiglianze sono gli stessi, così come le derivate seconde delle log-verosimiglianze (e quindi anche gli errori standard delle stime coincidono).
- A condizione che i parametri che corrispondono alle frequenze marginali fissate siano sempre incluse nel modello, anche per il caso di una marginale fissata è possibile sviluppare considerazioni analoghe alle precedenti e constatare l'equivalenza delle verosimiglianze.

Quasi-verosimiglianza e modelli con sovradispersione

Quasi-verosimiglianza

- Nell'ambito del LM classico il criterio dei minimi quadrati permette di stimare i parametri di regressione senza specificare un vero e proprio modello probabilistico. Il criterio dei minimi quadrati richiede la specificazione della relazione tra valore medio della risposta e predittore lineare, e la separazione tra valor medio e varianza dell'errore (che non è legata al valor medio).
- Anche nell'ambito dei GLM si può proseguire in questa direzione, introducendo però l'eventuale relazione tra media e varianza. Infatti, le equazioni di verosimiglianza

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi V(\mu_i) g'(\mu_i)} x_{ij} = 0 \ , \quad j = 1, \ldots, p \ ,$$

continuano ad essere delle "buone" equazioni di stima purchè sia $E(Y_i) = \mu_i = g^{-1}(\eta_i)$.

Quasi-verosimiglianza

- In altre parole, l'assunzione parametrica

$$Y_i \sim EF$$

  potrebbe anche non essere soddisfatta. È essenziale solo l'ipotesi sulla media.
- Si noti inoltre che l'unica caratteristica della distribuzione necessaria per esplicitare l'equazione di stima è la funzione di varianza $V(\mu)$.

Il modello di quasi-verosimiglianza

- Sotto condizioni di regolarità, le equazioni di verosimiglianza per un GLM producono stime per i coefficienti $\boldsymbol{\beta}$ che mantengono le loro proprietà anche se l'ipotesi che le osservazioni Y_i provengano da una famiglia esponenziale è sostituita dalle più deboli ipotesi sui momenti sino al secondo ordine (*assunzioni del secondo ordine*):
  1. $g(E(Y_i)) = \eta_i, \qquad i = 1, \ldots, n,$
  2. $var(Y_i) = \phi V(\mu_i), \qquad i = 1, \ldots, n,$
  3. $cov(Y_i, Y_j) = 0$, se $i \neq j$.
- Il modello statistico specificato dalle assunzioni 1–3 è detto *modello di quasi-verosimiglianza*.
- Se $V(\mu) = 1$ e $g(\mu) = \mu$, le ipotesi 1–3 coincidono con le usuali ipotesi del secondo ordine del LM classico.

Il modello di quasi-verosimiglianza

- **Ottimalità**: le proprietà di ottimalità di Gauss-Markov (BLUE) per i minimi quadrati in un LM valgono in modo analogo per le stime di quasi-verosimiglianza.
- Se consideriamo equazioni di stima lineari e non distorte lo stimatore di quasi-verosimiglianza ha asintoticamente la maggiore precisione: per qualsiasi combinazione lineare $a^{\mathrm{T}}\boldsymbol{\beta}$, si dimostra inoltre che lo stimatore di quasi verosimiglianza ha minima varianza asintotica in questa classe.
- In realtà, l'equazione di quasi-verosimiglianza per $\boldsymbol{\beta}$

$$q(y;\boldsymbol{\beta}) = \sum_{i=1}^{n} q(y_i;\boldsymbol{\beta})$$

$$= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi V(\mu_i) g'(\mu_i)} x_{ij} = 0 \ , \quad j = 1, \ldots, p \ ,$$

  si comporta come un vettore score

$$E(q(Y;\boldsymbol{\beta})) = 0 \ ,$$

$$var(q(Y;\boldsymbol{\beta})) = -E(\partial q(Y;\boldsymbol{\beta})/\partial \boldsymbol{\beta}) \ .$$

Il modello di quasi-verosimiglianza

- *Osservazione*: se valgono le equazioni di quasi-verosimiglianza definite sopra allora l'integrale di $q(y_i; \boldsymbol{\beta})$ dovrebbe comportarsi come una *funzione di log-verosimiglianza* per $\boldsymbol{\beta}$. Quindi si può porre per definizione che

$$\ell_Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} \, dt \ .$$

  è una funzione di quasi (log-)verosimiglianza

- La funzione $\ell_Q(\boldsymbol{\beta})$ non è in generale una log-verosimiglianza, anche se possiede molte delle sue proprietà formali.

- La funzione di quasi-verosimiglianza $\ell_Q(\boldsymbol{\beta})$ non è in generale una funzione di verosimglianza genuina ma ne condivide varie proprietà (gli stimatori di quasi MV $\hat{\boldsymbol{\beta}}$ sono asintoticamente Normali e il rapporto di quasi verosimiglianza ha una distribuzione chi quadrato sotto l'ipotesi nulla).

- Per dettagli si veda Azzalini sez. 6.5 e Faraway, Sez. 7.4

Perché la quasi-verosimiglianza?

- Il modello di quasi-verosimiglianza è adatto a trattare tanto dati continui quanto dati discreti. In particolare, per dati discreti che rappresentano conteggi o proporzioni e per dati continui che descrivono tempi di attesa, le ipotesi del secondo ordine 1–3 comportano un incremento di flessibilità rispetto alle specificazioni parametriche usuali del GLM con distribuzione Poisson, binomiale e esponenziale.
- Nella pratica vi sono applicazioni in cui il parametro di dispersione si trova in disaccordo con il modello ipotizzato.
- Ad esempio, mentre la famiglia binomiale ha parametro di dispersione ϕ noto ed uguale ad 1, l'analisi corrente dei dati può far ritenere ragionevole il valore $\phi > 1$.

Perché la quasi-verosimiglianza?

- Questo è un problema diffuso nell'analisi di dati discreti. Si parla di *sovradispersione* (*overdispersion*) quando la varianza della variabile Y è maggiore di quella teorica. Ad esempio, per dati binari si potrebbe ritenere che $var(Y) = \phi n\pi(1-\pi) > n\pi(1-\pi)$, con $\phi > 1$, ove $n\pi(1-\pi)$ è la varianza di una variabile binomiale.
- In conclusione, il metodo della quasi–verosimiglianza consente di affrontare i *problemi di sovradispersione*: si può infatti specificare $var(Y_i)$ in modo tale da consentire una maggiore variabilità rispetto a quella imposta dalla famiglia esponenziale di riferimento.
- Il caso della sottodispersione, cioè $\phi < 1$, è nelle applicazioni meno rilevante ma può essere anch'esso affrontato utilizzando un modello di quasi–verosimiglianza.

Motivi della sovradispersione nella binomiale i

- Il più semplice e comune meccanismo che dà origine alla sovradispersione è la presenza di raggruppamenti nella popolazione (clusters).
- Si consideri ad esempio il caso di una risposta binomiale. La variabile dipendente Y è sovradispersa se la sua varianza è più grande della varianza nominale di un modello binomiale di indice m, cioè $m\pi(1-\pi)$.
  Si assuma che i dati siano organizzati in gruppi (clusters) e che la dimensione dei clusters sia fissa e pari a k. Se m è il numero di unità nel campione vi sono quindi $l = m/k$ clusters. Ora si assuma che in ogni cluster il numero di successi Z_i segue una $Bi(k, \pi_i)$ che differisce nei vari clusters. Per cui la variabile dipendente è $Y = Z_1 + Z_2 + \cdots + Z_{m/k}$.

Motivi della sovradispersione nella binomiale ii

- Si supponga che π_i sia una variabile aleatoria con $E(\pi_i) = \pi$ e $var(\pi_i) = \tau^2\pi(1-\pi)$.
- La media di Y risulta pari a

$$E(Y) = \sum_i^{m/k} E(Z_i) = \sum_i^{m/k} kE(\pi_i) = \frac{m}{k} k\pi = m\pi$$

- la varianza di Y è

$$var(Y) = \sum Var(Z_i) = \sum \{E(var(Z_i|\pi_i)) + var(E(Z_i|\pi_i))\} = m\phi\pi(1-\pi).$$

ove $\phi = 1 + (k-1)\tau^2$ e inoltre risulta $\phi \geq 1$. Per dettagli si consulti Faraway sez 2.11 (esercizio!)

Motivi della sovradispersione nella binomiale iii

- Un meccanismo alternativo che origina la sovradispersione si ottiene se si assume che la $Y|P$ sia $Bi(m, p)$ e il valore p è un valore tratto da una variabile casuale P distribuita come una distribuzione beta $Be(\alpha, \beta)$. La variabile dipendente marginale Y ha una distribuzione beta-binomiale che è appunto sovradispersa rispetto la binomiale. In questo caso abbiamo:

$$E(Y) = m\pi$$

$$var(Y) = \phi m\pi(1 - \pi)$$

  dove $\pi = \alpha/(\alpha + \beta)$ è la media di una distribuzione Beta e $\phi > 1$ un opportuno parametro positivo (esercizio!).

- Nei modelli binomiali la sovradispersione (o sottodispersione) può verificarsi solo con modelli ove $m_i > 1$, quindi non nel caso di dati individuali.
- La sovradispersione in un modello binomiale è piuttosto comune.

Motivi della sovradispersione nella Poisson

- Schemi analoghi servono a giustificare la presenza di sovradispersione in modelli di Poisson. In questo caso quindi $Var(Y) > E(Y)$.
- Si assuma che ci sia variabilità entro le unità così che, ad esempio, il numero di incidenti per una unità è Poisson con media M. Se si assume che M sia un valore tratto da una distribuzione Gamma (con media μ e varianza μ/σ) la distribuzione marginale per Y è una binomiale negativa .
- Un GLM per una binomiale negativa è quindi adeguato in presenza di sovradispersione rispetto al modello di Poisson.
- In questi schemi i dati che sono osservati condizionatamente a un dato valore M sono indipendenti (ma le osservazioni marginali di Y sono dipendenti in quanto condividono il medesimo M).
- Si noti però che se il meccanismo che genera la sovradispersione fosse noto allora si potrebbe usare questo come modello generatore dei dati. Ma è ovvio che andremmo oltre i GLM. Quando non si è in grado di specificare il meccanismo con precisione allora una strategia appropriata è quella in cui si considera un GLM sovradisperso i cui parametri verranno stimati mediante quasi–verosimiglianza

Inferenza

- La procedura di stima dei coefficienti $\boldsymbol{\beta}$ non dipende da ϕ.
- Invece variano gli errori standard, essendo la matrice di varianze e covarianze asintotica proporzionale a ϕ, che va stimato.
- Anche sotto le ipotesi più deboli 1–3:
  1. Come detto, continua a valere l'identità dell'informazione: $E(\ell_* \ell_*^{\mathrm{T}}) = -E(\ell_{**})$ (argomentazione chiave delle proprietà asintotiche dello stimatore di massima verosimiglianza);
  2. Si mantiene la consistenza di $\hat{\boldsymbol{\beta}}$;
  3. Continua a valere l'approssimazione asintotica
  $$\hat{\boldsymbol{\beta}} \dot{\sim} N_p(\boldsymbol{\beta}, \phi(X^{\mathrm{T}} W X)^{-1}) \ ;$$
  4. Lo stimatore $\hat{\phi}$ è appropriato per stimare ϕ.
- Si può anche definire l'analogo della devianza, in termini di quasi-devianza.

I comandi in R i

- In R il metodo della quasi–verosimiglianza viene selezionato con l'opzione `quasi` della funzione `glm` che sostituisce il nome della famiglia esponenziale. Vanno specificati a quel punto solo la funzione legame e la funzione di varianza $V(\mu)$.
- La sintassi è:

  `glm(formula,family=quasi(link=`*legame*`,`
  `variance="`*funzionedivarianza*`"))`

I comandi in R ii

- Per le funzioni di varianza le opzioni disponibili sono:

`constant`	per 1;
`mu`	per μ;
`mu^2`	per μ^2;
`mu^3`	per μ^3;
`mu(1-mu)`	per $\mu(1-\mu)$.

  Le parole chiave `quasibinomial` e `quasipoisson` permettono più sinteticamente di specificare che il link è quello canonico rispettivamente delle famiglie binomiale e Poisson e che la funzione di varianza è pari a $\mu(1-\mu)$ per la binomiale e μ per la Poisson.

- Le stime che si otterranno in R per un modello di quasi-verosimiglianza saranno esattamente uguali a quelle ottenute per un GLM con la stessa funzione legame e funzione di varianza. Infatti le equazioni di stima sono le stesse.

I comandi in R iii

- Tuttavia verrà fornita una stima del parametro ϕ:
  troverete ad esempio, per `quasipoisson`, invece che
  (Dispersion parameter for poisson family taken to be 1)
  la scritta
  (Dispersion parameter for poisson family taken to be xxx)

- I valori degli standard errors saranno invece diversi: se c'è sovradispersione, cioè se ϕ è maggiore di 1, saranno più grandi.

- I valori della devianza vanno invece adattati perché quelli prodotti non tengono conto del fatto che il valore di ϕ è diverso da 1. Basta quindi dividere la devianza per il valore di ϕ stimato. Questo valore si comporta sotto H_0 (ipotesi qui che il modello corrente sia 'corretto') come un chi quadrato con opportuni gradi di libertà $(n - p)$:

$$\frac{D(y;\hat{\theta})}{\hat{\phi}} \sim \chi^2_{n-p}, \quad \text{asintoticamente, sotto } H_0.$$

I comandi in R iv

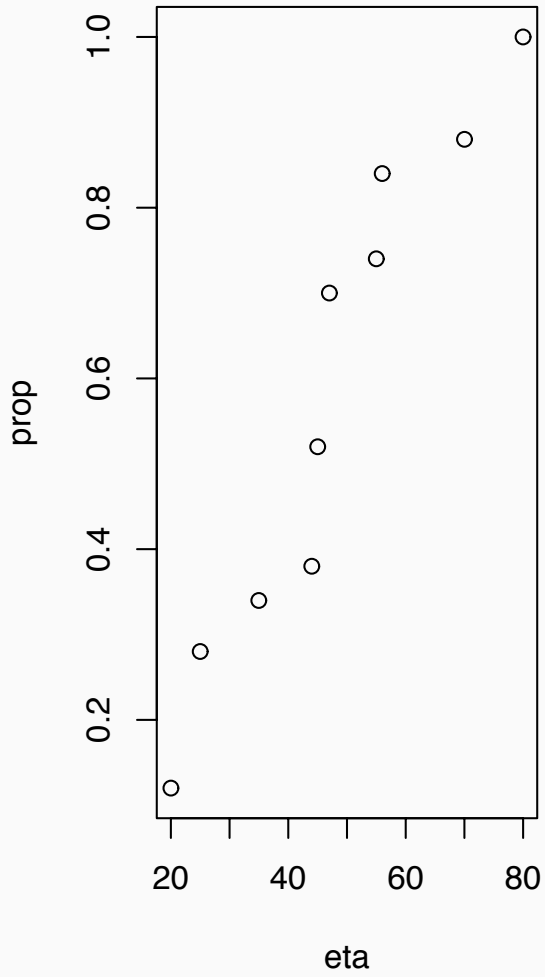
- **Attenzione**: quest'ultimo criterio, in generale, non è valido. Potrebbe funzionare quando *il numero di parametri è fisso*: questo è, per esempio, il caso di un modello binomiale per dati raggruppati o di un modello di Poisson con fattori come uniche covariate (come accade nei modelli log-lineari dalle tabelle di contingenza).
- Non è quindi in generale possibile ipotizzare un test per la devianza residua di un modello, a meno che non si ricada in alcuni casi particolari (vedi sopra).
- Nell'esempio che segue, tale test si può ipotizzare perché fissiamo le classi di età a priori e il numero di parametri del modello saturo non cresce con la numerosità campionaria n.

Esempio

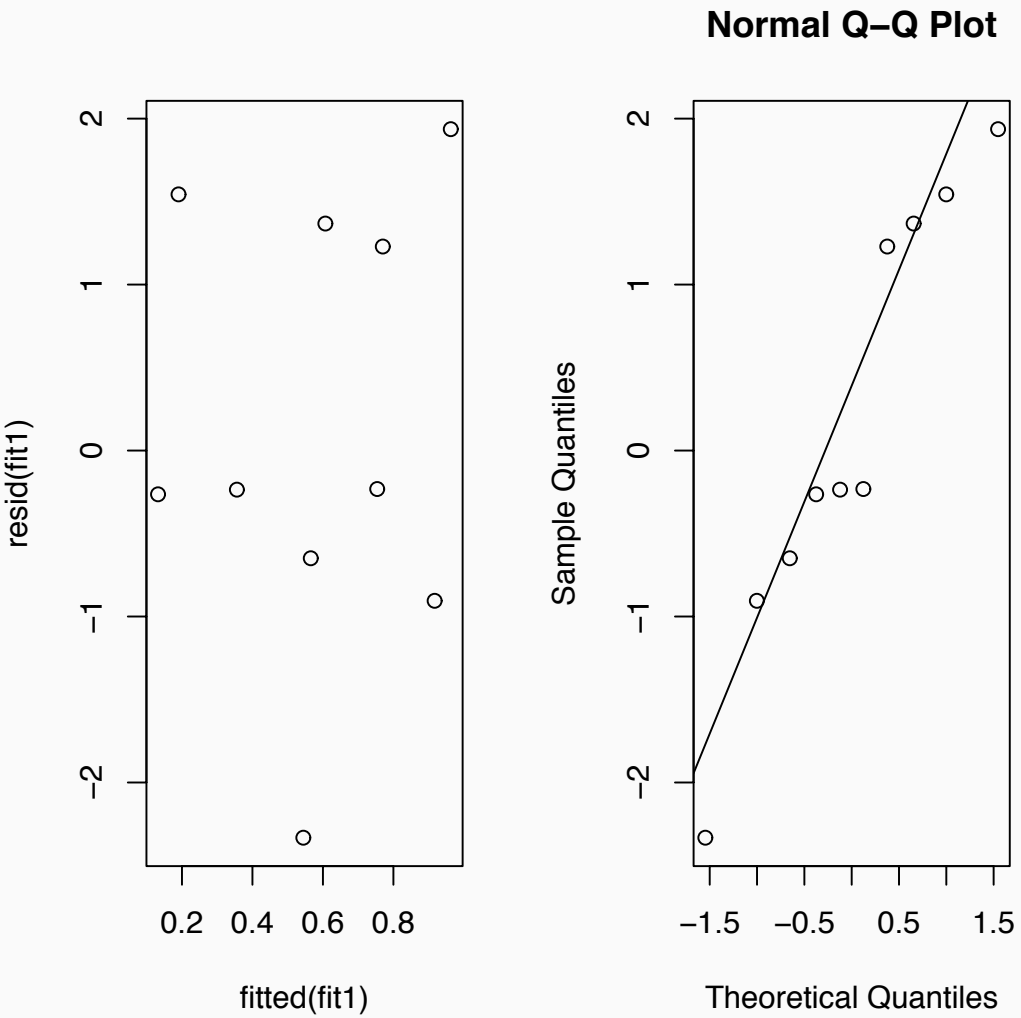
```
# Su un campione di persone di eta' diverse e' stato contato il numero
# di individui ciechi. I dati sono il numero di individui ciechi (nc), l'eta'
# (eta) e il numero di individui osservati per eta' (ni).
> nc <- c(6,14,17,19,26,35,37,42,44,50)
> ni <- c(50,50,50,50,50,50,50,50,50,50)
> eta <- c(20,25,35,44,45,47,55,56,68,70,80)
> prop <- nc/ni
> plot(eta,prop)

# Stima di un modello di regressione logistica:
> fit1 <- glm(cbind(nc,ni-nc)~eta,binomial)
> summary(fit1)

glm(formula = cbind(nc, ni - nc) ~ eta, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3331  -0.5521  -0.2336   1.3332   1.9363
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.59434    0.38557  -9.322   <2e-16 ***
eta          0.08574    0.00834  10.281   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 185.508  on 9  degrees of freedom
Residual deviance:  16.377  on 8  degrees of freedom
AIC: 56.635
Number of Fisher Scoring iterations: 4
```

```
# test sulla devianza e analisi dei residui: calcolo il p-value per la statistica
# della devianza scalata del modello secondo un chi-quadrato con n-p = 10-2 gdl.

> 1-pchisq(16.38,8)
[1] 0.03725306

> plot(fitted(fit1),resid(fit1))
> qqnorm(resid(fit1))
> qqline(resid(fit1))

# ampliamento del modello

> fit2 <- glm(cbind(nc,ni-nc)~eta+I(eta^2)+I(eta^3)+I(eta^4),binomial)
> anova(fit2)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(nc, ni - nc)

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev
NULL                          9    185.508
eta        1  169.131         8     16.377
I(eta^2)   1    1.492         7     14.885
I(eta^3)   1    0.044         6     14.841
I(eta^4)   1    0.073         5     14.768
```

Proporzione di ciechi e età

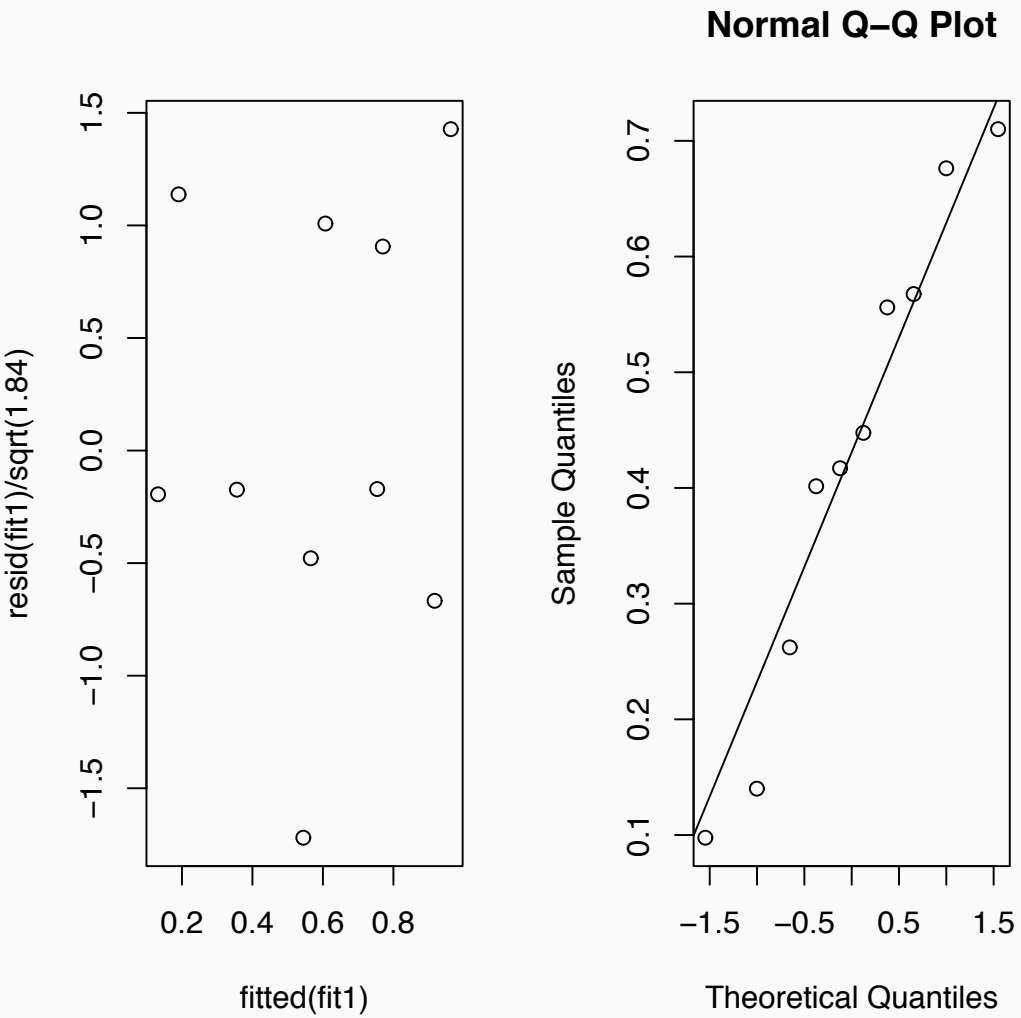


Residui e qqnorm del modello logistico



```
# modello di quasi-verosimiglianza: stimiamo attraverso i dati il
# parametro di dispersione

> fitq <- glm(cbind(nc,ni-nc)~eta,quasibinomial)
> summary(fitq)
glm(formula = cbind(nc, ni - nc) ~ eta, family = quasibinomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3331  -0.5521  -0.2336   1.3332   1.9363
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.59434    0.52304  -6.872 0.000128 ***
eta          0.08574    0.01131   7.579 6.43e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for quasibinomial family taken to be 1.840208)

    Null deviance: 185.508  on 9  degrees of freedom
Residual deviance:  16.377  on 8  degrees of freedom
AIC: NA
Number of Fisher Scoring iterations: 4

# sembra esserci una maggiore variabilita’ rispetto a quella indicata
# dal modello binomiale (phi stimato circa 1.84)

> 16.37/1.84    # devianza divisa par. di dispersione stimato
[1] 8.89674
> qqnorm(fitted(fitq)/sqrt(1.84))
> qqline(fitted(fitq)/sqrt(1.84))

# il modello sembra essere piu’ adeguato, se si tiene conto della
sovradispersione
```

Residui e qqnorm del modello con sovradispersione



Modelli per l'analisi di dati di sopravvivenza (durata)

Aspetti introduttivi

In vari contesti applicativi la variabile di principale interesse è rappresentata da una durata (e nelle applicazioni in medicina si tratta spesso del tempo di sopravvivenza).

Per esempio, potremmo essere interessati:

- a quanto tempo trascorre prima che una macchina si rompa,
- al tempo che trascorre fino alla guarigione (o alla morte) di un paziente;
- al tempo che trascorre prima che uno studente si laurei;
- alla durata della disoccupazione (ovvero al tempo che trascorre affinché un disoccupato trovi una nuova occupazione).

Dati censurati

- Dati di questo tipo presentano due particolarità salienti:
  1. sono non-negativi;
  2. possono essere censurati (e anzi spesso lo sono).
- Se ad esempio, decidiamo di studiare il tempo di sopravvivenza dopo una cura, e selezioniamo un campione di unità, di solito non potremo aspettare il tempo necessario affinché tutti i pazienti muoiano. L'osservazione terminerà prima: per alcuni soggetti sapremo la durata completa della sopravvivenza, per altri, e sono questi i dati censurati, sapremo solo che la sopravvivenza risulterà superiore ad un dato valore - il tempo di censura).
- In questo contesto, comunque, il nostro interesse è nella specificazione di un opportuno modello per la variabile dipendente costituita dalla durata, evidenziando anche l'effetto di potenziali variabili esplicative.

Notazione

- Siano T_i i tempi *completi* e c_i i tempi censurati per le osservazioni campionarie $i = 1, 2, \ldots, n$. Ciò che osserviamo può esser denotato con

$$Y_i = \min(T_i, c_i),$$

e l'indicatore

$$\delta_i = \begin{cases} 1 & T_i \leq c_i \text{ (non censurato)} \\ 0 & T_i > c_i \text{ (censurato)} \end{cases}$$

- Il primo problema è quello di individuare un opportuno modello distributivo per T_i (che dipenda da un ridotto numero di parametri); il secondo riguarda il modo con cui specificare la sua dipendenza da altre variabili (specificando il solito predittore lineare $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$). Ovviamente ci aspettiamo di poter usare le osservazioni (Y_i, δ_i) per fare inferenza sui parametri.

Modelli distributivi per le durate T_i

- È utile esprimere modelli per le v.a. introducendo rappresentazioni alternative alla usuale funzione di densità o di ripartizione.
- *Funzione di Sopravvivenza*

$$S(t) = P\{T \geq t\}$$

- *Funzione di Rischio*

$$\begin{aligned} h(t) &= \lim_{\Delta \to 0} \frac{P\{t \leq T < t + \Delta | T \geq t\}}{\Delta} \\ &= \lim_{\Delta \to 0} \frac{P\{t \leq T < t + \Delta\}}{\Delta P\{T \geq t\}} \\ &= \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t) \end{aligned}$$

La funzione di rischio integrato

Funzione di rischio integrato

Ora,

$$\log S(t) = \log S(t) - \log S(0) = -\int_0^t h(u)du.$$

Cioè,

$$S(t) = \exp(-H(t)),$$

dove $H(t)$ è la funzione di *rischio integrato*:

$$H(t) = \int_0^t h(u)du.$$

Dunque,

$$f(t) = h(t)\exp(-H(t)).$$

Il comportamento di una variabile viene definito equivalentemente attraverso la specificazione di una delle funzioni introdotte: la funzione di rischio, di sopravvivenza ecc.

Alcuni esempi rilevanti

- *Esponenziale* (ρ)

$$S(t) = e^{-\rho t},$$
$$f(t) = \rho e^{-\rho t},$$
$$h(t) = \rho \quad \text{e} \quad H(t) = \rho t,$$

per ogni $t > 0$.

In particolare la funzione di rischio è costante. Questo corrisponde a un'assenza di memoria.

- *Weibull* (κ, ρ)

$$S(t) = \exp\{-(\rho t)^{\kappa}\}$$
$$f(t) = \kappa\rho(\rho t)^{\kappa-1}\exp\{-(\rho t)^{\kappa}\}$$
$$h(t) = \kappa\rho(\rho t)^{\kappa-1} \quad \text{e} \quad H(t) = (\rho t)^{\kappa}$$

La funzione di rischio è un polinomio con esponente κ e ha andamento monotono. Segue che il caso $\kappa = 1$ corrisponde alla distribuzione esponenziale, ma in generale il modello ha una flessibilità più ampia.

Inferenza parametrica

- Assumiamo di disporre di un campione di dati (indipendenti e con distribuzione comune) $(Y_i, \delta_i), i = 1, 2, \ldots, n$ da un modello specificato da una funzione del tipo f, S, h o H, con parametro θ ignoto.

  La verosimiglianza è il metodo più naturale per la stima di θ.

- Se $\delta_i = 1$, allora $Y_i = T_i$, e il contributo alla verosimiglianza è $f(T_i; \theta) = f(Y_i; \theta)$.

- Se $\delta_i = 0$, allora $Y_i = c_i$, e il contributo alla verosimiglianza è $P\{T_i > c_i\} = S(c_i; \theta) = S(Y_i; \theta)$.

- Quindi, la funzione di verosimiglianza totale è

$$\prod_{i:\delta_i=1} f(Y_i; \theta) \prod_{i:\delta_i=0} S(Y_i; \theta)$$

$$= \prod_{i:\delta_i=1} h(Y_i; \theta) \prod_{i=1}^{n} S(Y_i; \theta).$$

Funzione di log-verosimiglianza: caso esponenziale

- Perciò, la funzione di log-verosimiglianza è

$$\ell(\theta) = \sum_{i:\delta_i=1} \log h(Y_i;\theta) - \sum_{i=1}^{n} H(Y_i;\theta)$$

- *Caso Esponenziale*
  Per semplificare la notazione, si ponga $\sum_{i:\delta_i=1}$ come $\sum_U$.

$$\ell(\rho) = \sum_U \log \rho - \sum_{i=1}^{n} \rho Y_i$$

  quindi $\ell'(\rho) = m/\rho - \sum Y_i$, che dà $\rho = m/\sum Y_i$, dove m è il numero di osservazioni non censurate.

Introdurre variabili esplicative

- In generale, la distribuzione di T_i potrebbe dipendere dalle covariate $\boldsymbol{x}$.
- Come solito, le variabili possono essere quantitative oppure fattori qualitativi. Per esempio:
  - variabili di trattamento (ad es., uso o meno di un certo tipo di farmaco)
  - caratteristiche individuali (ad es., età, sesso)
  - condizioni ambientali (ad es., la clinica o l'ospedale dove si viene curati)

Modelli per dati di durata

- Il criterio prevalente per esplicitare la dipendenza delle durate da un insieme di variabili esplicative è quello di considerare l'effetto di queste sulla funzione di rischio. In generale, si specifica quindi

$$T_i \sim h(t; \boldsymbol{x}_i, \boldsymbol{\beta})$$

  dove h è una funzione di rischio e $\boldsymbol{\beta}$ è un vettore di parametri che deve essere stimato.
- Ci sono due tipi di modelli che esplicitano in forma diversa l'impatto delle variabili esplicative sulla funzione di rischio:
  - modelli a rischi proporzionali;
  - modelli a tempi accelerati.
- Qui consideriamo solo il primo di questi (un testo per approfondire i modelli per l'analisi di dati di sopravvivenza è quello di Cox e Oakes, 1984 - Chapman and Hall).

Modello a rischi proporzionali

- In questo modello si assume che i tempi T_i siano determinazioni di una variabile aleatoria caratterizzata da una funzione di rischio di base $h_0(t)$ che è la stessa per ogni individuo ma su di essa vi è un effetto moltiplicativo di una componente positiva che dipende da $\boldsymbol{x}$.
- Assumiamo cioè che

$$h(t;\boldsymbol{x}) = \psi(\boldsymbol{x})h_0(t)$$

  dove $h_0(t)$ è appunto la funzione di rischio di base e non dipende dalle covariate $\boldsymbol{x}$ (si assume per l'identificabilità che essa soddisfi $h_0(t) = h(t;0)$) e $\psi(\boldsymbol{x})$ è una funzione di scala che esprime la variabilità del rischio con $\boldsymbol{x}$.
- Segue che

$$\begin{aligned} H(t;\boldsymbol{x}) &= \psi(\boldsymbol{x})H_0(t) \\ S(t;\boldsymbol{x}) &= (S_0(t))^{\psi(\boldsymbol{x})} \\ f(t;\boldsymbol{x}) &= \psi(\boldsymbol{x})(S_0(t))^{\psi(\boldsymbol{x})-1} f_0(t) \end{aligned}$$

Esempio: funzione di rischio esponenziale e Weibull

- *Esempio: esponenziale*

  Il modello esponenziale, $T_i \sim$ esponenziale(λ_i) ha rischi proporzionali:

  $$\frac{h(t;\lambda_1)}{h(t;\lambda_2)} = \frac{\lambda_1}{\lambda_2}.$$

  Si noti che il rapporto non dipende da t. Inoltre, per questo modello,

  $$h_0(t) = 1.$$

- *Esempio: Weibull*

  Il modello, $T_i \sim$ Weibull(κ, ρ_i) ha rischi proporzionali:

  $$\frac{h(t;\kappa,\rho_1)}{h(t;\kappa,\rho_2)} = \frac{\kappa\rho_1^\kappa t^{\kappa-1}}{\kappa\rho_2^\kappa t^{\kappa-1}} = \frac{\rho_1^\kappa}{\rho_2^\kappa},$$

  il rapporto non dipende da t. Per questo modello,

  $$h_0(t) = \rho t^{\kappa-1}.$$

Stima del modello con rischi proporzionali

Prima, dobbiamo scegliere un'opportuna specificazione per $\psi(\boldsymbol{x})$. La scelta

$$\psi(\boldsymbol{x}) = \exp(\boldsymbol{x}^T \boldsymbol{\beta})$$

è piuttosto naturale, infatti garantisce che:

1. ψ sia positiva per ogni $\boldsymbol{x}$,
2. come per i GLM, l'effetto delle covariate si espliciti attraverso una trasformazione di una previsore lineare,
3. si può sfruttare una somiglianza con i GLM che ci permetterà di utilizzare lo stesso algoritmo di stima.
4. La log-verosimiglianza è

$$\ell(\boldsymbol{\beta}) = \sum_U \log h - \sum H$$

Equazioni di verosimiglianza

- Sostituendo $h(t; \boldsymbol{x}) = e^{\boldsymbol{x}^T \boldsymbol{\beta}} h_0(t)$ e $H(t; \boldsymbol{x}) = e^{\boldsymbol{x}^T \boldsymbol{\beta}} H_0(t)$, abbiamo

$$\ell(\boldsymbol{\beta}) = \sum_U \boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_U \log h_0(Y_i) - \sum_{i=1}^n e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} H_0(Y_i).$$

- Dunque, le equazioni di verosimiglianza sono

$$\frac{\partial \ell}{\partial \beta_j} = \sum_U x_{ij} - \sum_{i=1}^n x_{ij} e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} H_0(Y_i) = 0$$

- In genere una soluzione esplicita non esiste. Però esistono alcuni casi speciali.

Un esempio con dati raggruppati

Ad esempio, siano le osservazioni raggruppate in k gruppi distinti, con $x_{ij} = 1$ se i appartiene al gruppo j, e 0 altrimenti. Sia m_j il numero di osservazioni non censurati nel gruppo j. Cioè $m_j = \sum_U x_{ij}$. Le equazioni di verosimiglianza sono

$$\frac{\partial \ell}{\partial \beta_j} = m_j - \sum_{i:x_{ij}=1} e^{\beta_j} H_0(Y_i) = 0$$

perciò

$$\widehat{\beta}_j = \log \left[m_j / \sum_{i:x_{ij}=1} H_0(Y_i) \right].$$

Stima di un modello di durata esponenziale utilizzando i GLM

- Si ricorda che per un modello a rischi proporzionali è $h(t; \boldsymbol{x}) = e^{\boldsymbol{x}^T \boldsymbol{\beta}} h_0(t)$, e che

$$\ell(\boldsymbol{\beta}) = \sum_U \boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_U \log h_0(Y_i) - \sum_{i=1}^n e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} H_0(Y_i)$$

  Adesso si immagini di considerare una situazione artificiale. Supponiamo che ciascuna variabile $\delta_i, i = 1, 2, \ldots, n$ sia indipendente e possa assumere valore 0 o 1. Assumiamo che queste variabili seguano la distribuzione di Poisson con medie

$$\mu_i = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} H_0(Y_i).$$

- Per questo esperimento la log-verosimiglianza è

$$\begin{aligned}
\ell^\star(\boldsymbol{\beta}) &= \sum_{i=1}^n (\delta_i \log \mu_i - \mu_i - \log \delta_i!) \\
&= \sum_U \boldsymbol{x}_i^T \boldsymbol{\beta} + \sum_U \log H_0(Y_i) - \sum_{i=1}^n e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} H_0(Y_i) \\
&= \ell(\boldsymbol{\beta}) - \sum_U \log \frac{h_0(Y_i)}{H_0(Y_i)}
\end{aligned}$$

Collegamento con log-versosimiglianza di Poisson

- Quindi, ℓ e $\ell^\star$ sono uguali a parte un termine che non dipende di $\boldsymbol{\beta}$. Dunque,
$$\frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell^\star}{\partial \beta_j}$$
e le stime di SMV saranno inevitabilmente le stesse.
- Potremo quindi stimare il modello artificiale come un GLM.
- Le variabili seguono la distribuzione di Poisson con una media che soddisfa
$$\log \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \log H_0(Y_i).$$
- Questo è precisamente un GLM con legame logaritmico a parte un termine di offset (di esposizione quindi) $\log H_0(Y_i)$ che si aggiunge all'intercetta e il cui valore è noto.

L'uso di R

- La funzione `glm` in **R** permette l'inclusione di un termine offset che va specificato in relazione alle ipotesi distributive riguardo la funzione di rischio di base.
- Ad esempio, se avessimo un dataframe *data* con variabili `Eta`, `Sesso`, e le osservazioni (Y_i, δ_i) tenute in `Tempo` e `Delta`, potremmo scrivere

  `glm(Delta~Age+Sex+offset(log(`H_0`(Tempo))), poisson,`*data*`)`

  dove H_0 viene sostituita con un'espressione valida in **R**. In questa modo vengono stimati i parametri del modello artificiale, e quindi del modello a rischi proporzionali.
- Si noti che ciò è agevole se ipotizziamo che la funzione di rischio di base è esponenziale poiché questa non dipende da ulteriori parametri ignoti. Se invece avessimo, ad esempio, una Weibull la funzione di rischio di base dipenderebbe da κ.

Altri aspetti rilevanti per l'analisi di sopravvivenza

Alcuni degli aspetti più interessanti per l'analisi di dati di sopravvivenza vanno ben oltre gli scopi di questo corso. In particolare aspetti interessanti sono:

- La stima di modelli a rischi proporzionali quando la funzione h_0 include parametri ignoti (come nel caso del modello Weibull);
- La stima di modelli per cui la funzione h_0 rimane non specificata (la cosidetta regressione di Cox mediante la tecnica della verosimiglianza parziale);
- La stima di modelli che non hanno rischi proporzionali (dove non ci si può richiamare alla teoria dei GLM).

Stima della funzione di sopravvivenza

- Se si dispone di dati non censurati la stima (non parametrica) della funzione di siopravvivenza è semplice: basta prendere il complemento a 1 della funzione di ripartizione empirica

$$\hat{S}(t) = 1 - \hat{F}(t) = 1 - \frac{(\#\{t_i \leq t\})}{n} = 1 - \frac{n - r(t)}{n},$$

  ove $r(t)$ è il numero di unità la cui durata supera t, ovvero l'insieme di coloro ancora a rischio (ancora vivi quindi) a t.

- Si noti che per un dato valore di $\hat{S}(t)$ è possibile costruire anche delle bande di confidenza basate sulla distribuzione binomiale, ovvero

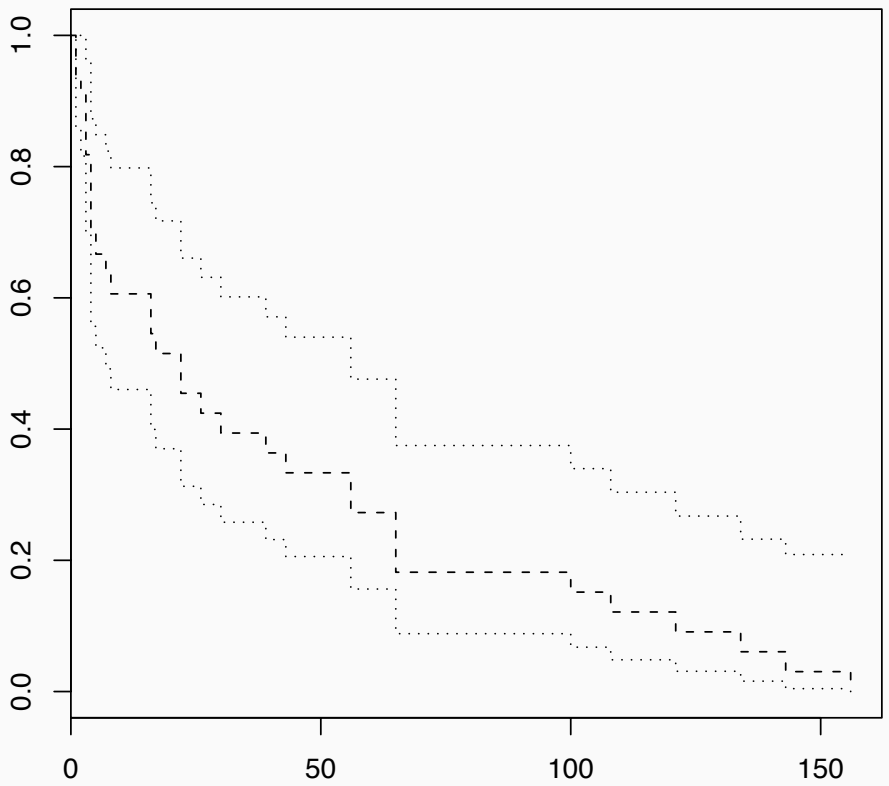
$$\hat{S}(t)[1 - \hat{S}(t)]/n = r(t)[n - r(t)]/n^3$$

- Si considerino ad esempio i seguenti dati:

```
65 156 100 134  16 108 121   4
39 143  56  26  22   1   1   5
65  56  65  17   7  16  22   3
4   2   3   8   4   3  30   4  43
```

Stima della funzione di sopravvivenza

La funzione di sopravvivenza empirica (con le bande di confidenza) risulta



Stima della funzione di sopravvivenza - dati censurati

- Se i dati sono censurati è necessario tenere conto dell'informazione parziale che forniscono i dati censurati: essi contribuiranno a definire l'insieme di rischio immediatamente prima di un tempo t ma non osserveremo per essi il tempo di sopravvivenza.
- Sia quindi $r(t)$ il numero di unità statistiche ancora vive immediatamente prima del tempo t, $r(t)$ è detto insieme di rischio al tempo t.
- Immaginiamo ora di considerare un insieme di intervalli $I_i = [t_i, t_{i+1}]$ che partizionano l'insieme $[0, \text{inf})$ e sia d_i il numero di tempi completi nell'intervallo (ovvero il numero di unità che muoiono nell'intervallo).
- In ciascun intervallo I_i la probabilità p_i di sopravvivere oltre t_{i+1} per coloro che sono sopravvissuti fino a t_i è stimata banalmente da $[r(t_i) - d_i]/r(t_i)$.

Stimatore di Kaplan-Meyer

- La probabilità complessiva di sopravvivere fino a t_i è quindi

$$P(T > t_i) = S(t_i) \approx \prod_{j=0}^{i-1} p_j \approx \prod_{j=0}^{i-1} \frac{r(t_j) - d_j}{r(t_j)}$$

  Si noti peraltro che il rapporto $d_i/r(t_i)$ è assimilabile a una approssimazione della funzione di rischio discretizzata h_i.
- Se immaginassimo di fare intervalli sempre più piccoli, il valore d_i sarà diverso da 1 solo per intervalli in cui c'è un tempo completo (una morte), per cui al limite otterremo

$$\hat{S}(t) = \prod_{t_i \leq t : \delta_i = 1} \frac{r(t_i) - d_i}{r(t_i)}$$

  .
- Lo stimatore introdotto è detto di Kaplan-Meyer (o stimatore limite-prodotto).

Stimatore di Nelson-Aalen

- Si noti che dopo l'ultimo dato non censurato t_i la funzione è costante, di solito infatti ci si ferma prima. Si dimostra che tale stimatore è anche stimatore di massima verosimiglianza e a partire da questo è possibile ottenere stima della varianza.
- Un ragionamento simile si può applicare per stimare la funzione di rischio cumulativa

$$H(t_i) \approx \sum_{j \leq i} h(t_j)(t_{j+1} - t_j) \approx \sum_{j \leq i} \frac{d_j}{r(t_j)}$$

  che al limite diventa $\hat{H}(t) = \sum \frac{d_j}{r(t_j)}$. La somma è estesa ai tempi completi minori di t. Questo stimatore è detto di Nelson-Aalen e da esso si ottiene la stima di $S(t)$ mediante la relazione $\hat{S}(t) = \exp(-\hat{H}(t))$.
- I due stimatori tendono a coincidere se l'insieme di rischio è ampio.

Stima della funzione di sopravvivenza: esempio

Lo stimatore della varianza (formula di Greenwood) è dato da

$$var(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j \leq i} \frac{d_j}{r(t_j)[r(t_j) - d_j]}$$

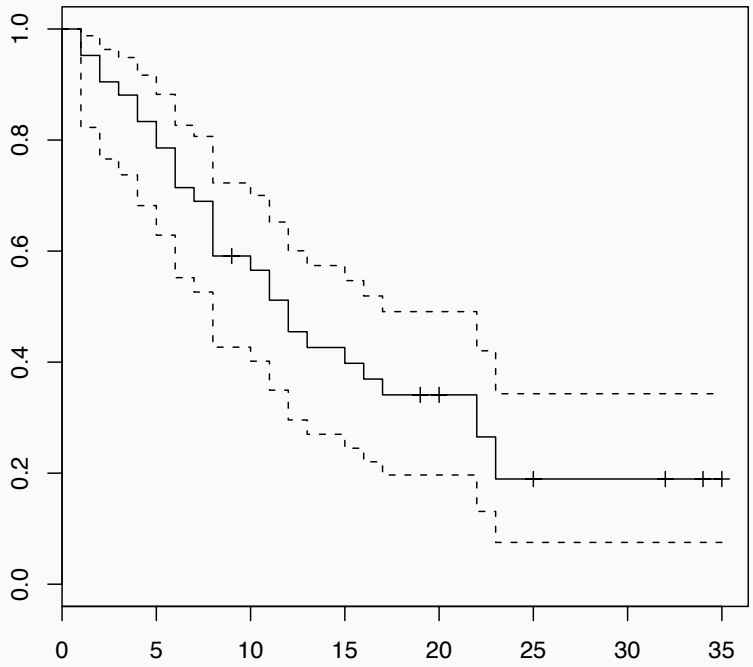
Come esempio si considerino i seguenti dati:

1 10 22 7 3 32+ 12 23 8 22
17 6 2 16 11 34+ 8 32+ 12
25+ 2 11+ 5 20+ 4 19+ 15 6 8
17+ 23 35+ 5 6 11 13 4 9+
1 6+ 8 10+

I dati con + sono censurati.

Stima della funzione di sopravvivenza: esempio

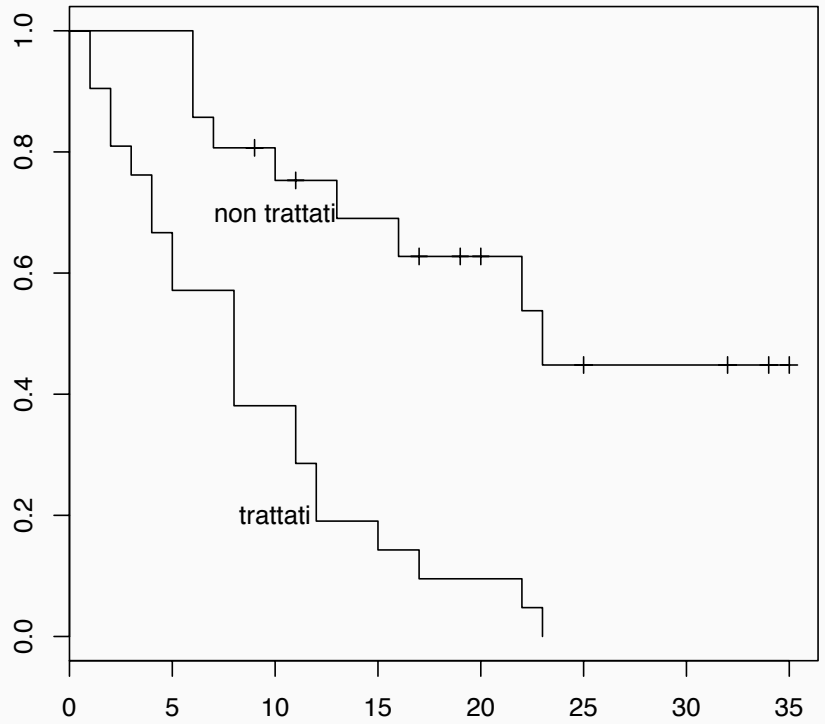
La stima di Kaplan-Meyer fornisce:



In generale è utile ottenere una stima della curva di sopravvivenza separatamente per i livelli di una variabile categoriale. Ad esempio, la curva di sopravvivenza per trattati con una terapia o con un placebo.

Confronto fra curve di sopravvivenza

Un esempio è nel grafico dove sono riportati i tempi di remissione dalla malattia per malati di leucemia. Alcuni sono trattati con il farmaco 6MP (mercaptopurina), altri non sono trattati.



Un test per la differenza fra curve di sopravvivenza

- Esiste la possibilità di condurre una verifica dell'ipotesi che le due curve siano significativamente diverse. Il test comunemente usato è detto log-rank test.
- Il test si basa sul confronto delle funzioni di rischio in corrispondenza di ciascun evento osservato. Per ciascuno dei j tempi completi osservati siano N_{1j} e N_{2j} il numero di soggetti a rischio nei due gruppi all'inizio del periodo t_j.
- Sia $N_j = N_{1j} + N_{2j}$ e siano O_{1j} e O_{2j} il numero di eventi osservati nei due gruppi fino al tempo t_j e $O_j = O_{1j} + O_{2j}$.
- Sotto l'ipotesi nulla di eguale curva di sopravvivenza il numero di eventi O_{1i} ha una distribuzione ipergeometrica con parametri N_j, N_{1j} e O_j con valore atteso $E_j = O_j \frac{N_{1j}}{N_j}$ e varianza $V_j = \frac{(O_j)(\frac{N_{1j}}{N_j})(1-\frac{N_{1j}}{N_j})(N_J - O_j)}{N_j - 1}$.
- Un test può essere ricavato considerando che $\sum \frac{(O_j - E_j)^2}{E_j}$ asintoticamente è approssimato, se i due gruppi hanno uguale funzione di sopravvivenza, da un χ^2_1.