Generalized Linear Models

Vers. 1.1.2

Gianluca Mastrantonio

gianluca.mastrantonio@polito.it

"AN INTRODUCTION TO GLM (GENERALIZED LINEAR MODELS)"

ANNETTE J. DOBSON

(II EDIZIONE)

I modelli lineari generalizzati I

Cerchiamo di estendere il modello lineare per poter considerare dati che provengano da distribuzioni diverse. Per fare questo consideriamo una famiglia di distribuzioni, chiamata "esponenziale".

La densità di una variabile aleatoria y_i si dice appartenere alla famiglia exponenziale se può essere scritta come

$$f(y_i; \theta_i, \phi_i) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right) \quad (1)$$

dove

- θ_i è chiamato parametro *naturale* e
- ϕ_i è il parametro di *dispersione*

- $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ funzioni generiche
- θ_i, ϕ_i sono dipendenti dalla distribuzione considerata

I modelli lineari generalizzati II

Per le distribuioni della famiglia esponenziale, media e varianza si calcolano facilmente
Assumiamo che sia possibile cambiare l'ordine di integrazione e derivazione nella seguente formula

$$\frac{\partial}{\partial \theta_i} \int f(y_i; \theta_i, \phi_i) dy_i = \int \frac{\partial f(y_i; \theta_i, \phi_i)}{\partial \theta_i} dy_i$$

ASSUNZIONE

che si può dimostrare essere possibile per la famiglia esponenziale, allora

$$\int \frac{\partial f(y_i; \theta_i, \phi_i)}{\partial \theta_i} dy_i = 0$$

Se lo scriviamo in forma estesa per la famiglia esponenziale abbiamo

$$\int \frac{\partial f(y_i; \theta_i, \phi_i)}{\partial \theta_i} dy_i = \int \frac{y_i - b'(\theta_i)}{a(\phi_i)} f(y_i; \theta_i, \phi_i) dy_i =$$

DERIVATA DI 1

$$\frac{1}{a(\phi_i)} \int y_i f(y_i; \theta_i, \phi_i) dy_i - \frac{b'(\theta_i)}{a(\phi_i)} \int f(y_i; \theta_i, \phi_i) dy_i = \frac{E(y_i)}{a(\phi_i)} - \frac{b'(\theta_i)}{a(\phi_i)} = 0$$

$E[y_i]$!!

$= 1$

I modelli lineari generalizzati III

Quindi

$$E(y_i) = \mu_i = b'(\theta_i) \quad (2)$$

che ci da la formula per la media.

Assumendo di poter cambiare l'ordine di integrazione ederivazione anche per le derivate seconde

$$\frac{\partial^2}{\partial \theta_i^2} \overbrace{\int f(y_i; \theta_i, \phi_i) dy_i}^{1} = \int \frac{\partial^2 f(y_i; \theta_i, \phi_i)}{\partial \theta_i^2} dy_i \quad \text{ASSUNZIONE}$$

che è possibile per la famiglia esponenziale, abbiamo che

$$\int \frac{\partial^2 f(y_i; \theta_i, \phi_i)}{\partial \theta_i^2} dy_i = 0$$

I modelli lineari generalizzati IV

e possiamo scrivere

$$\int \frac{\partial^2 f(y_i;\theta_i,\phi_i)}{\partial \theta_i^2} dy_i = \int \frac{-b''(\theta_i)}{a(\phi_i)} f(y_i;\theta_i,\phi_i) + \frac{(y_i - b'(\theta_i))^2}{a(\phi_i)^2} f(y_i;\theta_i,\phi_i) dy_i =$$

$$= -\frac{b''(\theta_i)}{a(\phi_i)} \underbrace{\int f(y_i;\theta_i,\phi_i) dy_i}_{=1} + \frac{1}{a(\phi_i)^2} \int (y_i - \underbrace{b'(\theta_i)}_{\mu_i})^2 f(y_i;\theta_i,\phi_i) dy_i =$$

$$- \frac{b''(\theta_i)}{a(\phi_i)} + \frac{\mathrm{Var}(y_i)}{a(\phi_i)^2} = 0$$

$\mathbb{E}[y_i] = b'(\theta_i)$!!

μ_i

μ_i''

che ci porta ad avere

$$\mathrm{Var}(y_i) = b''(\theta_i) a(\phi_i) \quad (3)$$

vediamo adesso come possiamo passare dal modelli lineare a quello generalizzato.

I modelli lineari generalizzati V

Abbiamo sempre scritto il modello lineare come

x_{ij} : j-esima covariata associata all'i-esima osservazione

$$\left.\begin{aligned} y_i &= \overbrace{\sum_{j=1}^{p} \beta_j x_{ij}}^{\mu_i} + \epsilon_i \\ \epsilon &\sim N(0, \sigma^2) \end{aligned}\right\}$$

con tutte le varie ipotesi. Ma un modo equivalente per scriverlo è il seguente

$$\left.\begin{aligned} y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \sum_{j=1}^{p} \beta_j x_{ij} \end{aligned}\right\}$$

RAPPRESENTAZIONE GERARCHICA DEL MODELLO LINEARE

dove è chiaro che stiamo modellizzando il parametro media della normale.

I modelli lineari generalizzati VI

I GLM sono una generalizzazione del modello lineare, e si possono scrivere come

$$y_i \sim H(\theta_i, \phi_i)$$

FAMIGLIA ESPONENZIALE

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$$

!!!

$g: \mathcal{D} \longrightarrow \mathbb{R}$

- H è una distribuzione membro della famiglia esponenziale;
- $f(y_i; \theta_i, \phi_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right)$
- $E(y_i) = \mu_i = b'(\theta_i)$;
- $\text{Var}(y_i) = b''(\theta_i)a(\phi_i)$;
- $\mu_i \in \mathcal{D}$, dove $\mathcal{D}$ può essere $\mathbb{R}^+$, oppure $[0, 1]$, etc...;
- $g : \mathcal{D} \to \mathbb{R}$ è la funzione link;

I modelli lineari generalizzati VII

Tra tutte le funzioni link, quella che soddisfa $\theta_i = g(\mu_i)$ è chiamata **funzione link canonica**

USERETO SEMPRE QUESTA !!

vediamo alcuni esempi

Distribuzione normale $y \sim N(\alpha_i, \sigma^2)$ La densità è

$$f(y_i; \alpha_i, \sigma^2) = (2\pi\sigma^2) \exp\left(-\frac{(y_i - \alpha_i)^2}{2\sigma^2}\right) \quad (*)$$

$$= e^{\log(2\pi\sigma^2)^{-\frac{1}{2}}} = \exp\left(-\frac{1}{2}\log(2\pi\sigma^2)\right)$$

$$(*) \quad \frac{-y_i^2 + 2y_i\alpha_i - \alpha_i^2}{2\sigma^2}$$

$$= \exp\left(\frac{y_i\alpha_i - \frac{1}{2}\alpha_i^2}{\sigma^2 a(\phi)} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2}\right)$$

$c(y_i, \phi)$

Con

- $\theta_i = \alpha_i$ e $\phi = \sigma^2$;
- $b(\theta_i) = \frac{\theta_i^2}{2}$ $\quad a(\phi_i) = \phi_i$ $\quad c(y_i; \phi_i) = -\frac{1}{2}\log(2\pi\phi_i) - \frac{y_i^2}{2\phi_i}$;
- $E(y_i) = \theta_i = \alpha_i$ $\quad \mathrm{Var}(y_i) = \phi_i = \sigma^2$

$E[y_i] = b'(\theta) = \theta_i$ $\quad Var[y_i] = + b''(\theta) a(\phi) = 1 \cdot \sigma^2$

I modelli lineari generalizzati VIII

- la funzione link canonica è $g(\mu_i) = \mu_i$

Distribuzione Poisson $y_i \sim \text{Pois}(\lambda_i)$ La densità è

$$f(y_i;\lambda) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

$$= \exp\left(y_i \log(\lambda_i) - \lambda_i - \log(y!)\right)$$

$$\log(\lambda_i) = 0$$

$$a: \phi \longrightarrow 1$$

- $\theta = \log(\lambda_i)$ e $\phi = \emptyset$;
- $b(\theta_i) = \exp(\theta_i) \qquad a(\phi_i) = 1 \qquad c(y_i;\phi_i) = -\log(y_i!)$;
- $E(y_i) = \exp(\theta_i) = \lambda_i \qquad \text{Var}(y_i) = \exp(\theta_i) = \lambda_i$
- la funzione link canonica è $g(\mu_i) = \log(\mu_i)$

I modelli lineari generalizzati IX

Distribuzione Binomial $y_i \sim \text{Bin}(n_i, \pi_i)$ (n_i si considera noto e non parametro)

$$f(y_i; n_i, \pi_i) = \binom{n_i}{y_i} \cdot \pi_i^{y_i} \cdot (1-\pi_i)^{n_i - y_i}$$

$$= \exp\left(\frac{y_i \log(\pi_i/(1-\pi_i)) + n_i \log(1-\pi_i)}{1} - \log\binom{n_i}{y_i} \right)$$

$\pi_i = g(\theta)$ INVERSO (*)

$\log(\pi_i)$

- $\theta = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ e $\phi = \emptyset$; (*)
- $b(\theta_i) = n_i \log(1+\exp(\theta_i))$ $\quad a(\phi_i) = 1$ $\quad c(y_i; \phi_i) = \log\binom{n_i}{y_i}$;
- $E(y_i) = n_i \frac{\exp(\theta_i)}{1+\exp(\theta_i)} = n_i \pi_i$ $\quad \text{Var}(y_i) = \frac{\exp(\theta_i) n_i}{(1+\exp(\theta_i))^2} = n_i \pi_i (1-\pi_i)$
- la funzione link canonica è $g(\mu_i) = \log\left(\frac{\mu_i}{n_i - \mu_i}\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

LOGIT !!!

Stima dei parametri I

Per ognuno di queste distribuzioni abbiamo un valore medio e siamo interessati a modellizzarlo con una funzione lineare

$$g(\mu_i) = \eta_i = \beta_1 + \sum_{j=2}^{p} \beta_j x_{ij} = \mathbf{X}_i \boldsymbol{\beta}$$

e come nel modello lineare, vogliamo trovare le stime dei parametri trovando il massimo della verosimiglianza (o log laverosimiglianza), che può essere trovato imponendo la derivata uguale a 0.

Indichiamo con $L(\vec{\boldsymbol{\beta}}; \mathbf{y})$ la log-verosimiglianza di $\vec{\boldsymbol{\beta}}$

$$\boxed{L(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} c(y_i, \phi_i)}$$

e con $\mathbf{U}$ il vettore di derivate rispetto a $\boldsymbol{\beta}$, chiamato anche **statistica score**, con elementi:

$$U_j = \frac{\partial L(\boldsymbol{\beta}; \mathbf{y})}{\partial \beta_j}$$

Stima dei parametri II

Per calcolare le derivate utilizziamo la chain rule:

$$U_j = \frac{\partial L(\boldsymbol{\beta}; y_i)}{\partial \beta_j} = \underbrace{\frac{\partial L(\boldsymbol{\beta}; y_i)}{\partial \theta_i}}_{①} \underbrace{\frac{\partial \theta_i}{\partial \mu_i}}_{②} \underbrace{\frac{\partial \mu_i}{\partial \eta_i}}_{③} \underbrace{\frac{\partial \eta_i}{\partial \beta_j}}_{④}$$

dove

$$① \quad \frac{\partial L(\boldsymbol{\beta}; y_i)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi_i)} = \frac{y_i - \mu_i}{a(\phi_i)} \qquad b'(\theta_i) = \mu_i$$

$$② \quad \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\mathrm{Var}(y_i)}{a(\phi_i)} \qquad ②^{-1}$$

$$\mu_i = \mathbb{E}[y_i] = b'(\theta) \qquad \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta)$$

e sfruttc $\mathrm{Var}(y_i) = b'' \cdot a$

$$\rightarrow b'' = \frac{\mathrm{Var}(y_i)}{a(\phi_i)}$$

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} \qquad ④ \quad \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

Abbiamo quindi

$$U_j = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\cancel{a(\phi_i)}} \frac{\cancel{a(\phi_i)}}{\mathrm{Var}(y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{\mathrm{Var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} \qquad ④$$

$\mathbb{E}[U_j] = \ldots$

sfruttc $\mathbb{E}[y_i] = \mu_i$

DIPENDENTA DA MODELLO E DA $g(\cdot)$!!!

Stima dei parametri III

dove $\frac{\partial \mu_i}{\partial \eta_i}$ dipende dal particolare modello che stiamo usando.
La soluzione $\mathbf{U} = \mathbf{0}$ rispetto a $\boldsymbol{\beta}$ generalmente non si riesce a trovare, e algoritmi numerici vengono usati (Newton-Raphson o Fisher Scoring) ITERATIVI !!!.

Ipotizziamo di aver trovato la soluzione $\hat{\boldsymbol{\beta}}$, abbiamo adesso bisogno di trovare la sua distribuzione per poter fare inferenza: test, intervalli di confidenza etc.. Visto che $\hat{\boldsymbol{\beta}}$ non si può trovare in forma chiusa, non sappiamo neanche la sua distribuzione, e abbiamo bisogno do un qualche metodo per trovarla (almeno approssimativamente)

A questo scopo, possiamo utilizzare la statistica score $\mathbf{U}$. Indichiamo con $\mathbf{V}$ e $\mathbf{D}$ due matrici diagonali aventi come elemento i-esimo $\text{Var}(y_i)$ e $\partial \mu_i / \partial \eta_i$, rispettivamente.
Allora

$$\boxed{\mathbf{U} = \mathbf{X}^T \mathbf{D} \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu})}$$

RISCRITTURA COMPATTA DI (4)

La statistica $\mathbf{U}$ è una variabile aleatoria con media 0 $(E(y_i) = \mu_i)$

Stima dei parametri IV

I METODO

Definiamo $\mathbf{J}$ come la matrice di varianza e covarianze di $\mathbf{U}$, e chiamiamola **matrice d'informazione**, l'elemento (j, k) di $\mathbf{J}$ si calcola come

$$\text{Cov}(U_j, U_k) = E(U_j U_k) = E\left(\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} \sum_{h=1}^{n} \frac{(y_h - \mu_h)x_{hk}}{\text{Var}(y_h)} \frac{\partial \mu_h}{\partial \eta_h}\right)$$

Dato che y_i è indipendente da y_j abbiamo che $E((y_i - \mu_i)(y_h - \mu_h)) = 0$, quindi

$$(*) \quad J_{jk} = E(U_j U_k) = E\left(\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2 x_{ij} x_{ik}}{\text{Var}(y_i)^2} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right) = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

Possiamo allora definire

$$\mathbf{J} = \mathbf{X}^T \mathbf{W} \mathbf{X} \quad (5)$$

con $W_{ii} = \frac{1}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$

RIMANGONO SOLO I TERMINI IN CUI $i = h$ (*)

Stima dei parametri V

II METODO

Abbiamo anche un diverso modo di calcolare la matrice di informazione. Calcoliamo prima il valore di $E(U_j)$

$$E(U_j) = \int U_j f(\mathbf{y}; \boldsymbol{\beta}, \phi_i) dy = \int \frac{\partial \log f(\mathbf{y}; \boldsymbol{\beta}, \phi_i)}{\partial \beta_j} f(\mathbf{y}; \boldsymbol{\beta}, \phi_i) dy = 0.$$

Se differenziamo rispetto a β_k, e scambiamo l'ordine di integrale con derivata, otteniamo

$$0 = \int \frac{\partial^2 \log f(\mathbf{y}; \boldsymbol{\beta}, \phi_i)}{\partial \beta_j \partial \beta_k} f(\mathbf{y}; \boldsymbol{\beta}, \phi_i) dy + \int \frac{\partial \log f(\mathbf{y}; \boldsymbol{\beta}, \phi_i)}{\partial \beta_j} \frac{\partial f(\mathbf{y}; \boldsymbol{\beta}, \phi_i)}{\partial \beta_k} dy =$$

usando la derivata del logaritmo, otteniamo

$$\int \frac{\partial^2 \log f(\mathbf{y}; \boldsymbol{\beta}, \phi_i)}{\partial \beta_j \partial \beta_k} f(\mathbf{y}; \boldsymbol{\beta}, \phi_i) dy +$$

$$\int \frac{\partial \log f(\mathbf{y}; \boldsymbol{\beta}, \phi_i)}{\partial \beta_j} \frac{\partial \log f(\mathbf{y}; \boldsymbol{\beta}, \phi_i)}{\partial \beta_k} f(\mathbf{y}; \boldsymbol{\beta}, \phi_i) dy =$$

Stima dei parametri VI

Si vede facilmente che entrambi gli integrali sono valori attesi, e precisamente

$$\int \frac{\partial^2 \log f(\mathbf{y};\boldsymbol{\beta},\phi_i)}{\partial\beta_j \partial\beta_k} f(\mathbf{y};\boldsymbol{\beta},\phi_i)dy = E\left(\frac{\partial U_j}{\partial\beta_k}\right)$$

e

$$\int \underbrace{\frac{\partial \log f(\mathbf{y};\boldsymbol{\beta},\phi_i)}{\partial\beta_j}}_{U_j} \underbrace{\frac{\partial \log f(\mathbf{y};\boldsymbol{\beta},\phi_i)}{\partial\beta_k}}_{U_k} f(\mathbf{y};\boldsymbol{\beta},\phi_i)dy = E\left(U_j U_k\right)$$

Abbiamo allora che

$$J_{jk} = E(U_j U_k) = -E\left(\frac{\partial U_j}{\partial\beta_k}\right)$$

Oppure, in forma matriciale

$$\boxed{\mathbf{J} = E(\mathbf{U}\mathbf{U}^T) = -E(\mathbf{U}')}$$

Let matrici $\mathbf{U}$ e $\mathbf{J}$ sono funzioni di $\boldsymbol{\beta}$, e quindi le indichiamo, in generale, $\mathbf{U}(\mathbf{b})$ se abbiamo bisogno di sapere dove sono calcolare.

Stima dei parametri VII

Queste matrici ci serviranno per fare i testi di ipotesi su $\boldsymbol{\beta}$. Ma per fare questo abbiamo bisogno delle distribuzione di $\mathbf{U}$

Calcoliamo l'approssimazione di Taylor per la log verosimiglianza, calcolata in $\hat{\boldsymbol{\beta}}$

- $L(\boldsymbol{\beta};\mathbf{y}) \approx L(\hat{\boldsymbol{\beta}};\mathbf{y}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \frac{\partial L(\hat{\boldsymbol{\beta}};\mathbf{y})}{\partial \boldsymbol{\beta}} + \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \frac{\partial^2 L(\hat{\boldsymbol{\beta}};\mathbf{y})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) =$

DEFINIZIONE DI U

$$L(\hat{\boldsymbol{\beta}};\mathbf{y}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \mathbf{U}(\hat{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \mathbf{U}(\hat{\boldsymbol{\beta}})'(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

Approssimiamo $\mathbf{U}'$ con $E(\mathbf{U}') = -\mathbf{J}$ e ricordiamo che $\mathbf{U}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, allora

$U(\hat{\beta}) = 0$!!! xke' $\hat{\beta}$ MLE

$$L(\boldsymbol{\beta};\mathbf{y}) \approx L(\hat{\boldsymbol{\beta}};\mathbf{y}) - \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \mathbf{J}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

Calcoliamo la derivata rispetto a $\boldsymbol{\beta}$ ottenendo

$\frac{\partial L(\beta|y)}{\partial \beta} = U$

$$\mathbf{U}(\boldsymbol{\beta}) = -(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T \mathbf{J}(\hat{\boldsymbol{\beta}}) = -\mathbf{J}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

Stima dei parametri VIII

da cui abbiamo che

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{J}^{-1}\mathbf{U}$$!!!

oss: $\mathbf{J}^{-1} \cdot \mathbf{U} = \mathbf{J}^{-\frac{1}{2}} \cdot (\mathbf{J}^{-\frac{1}{2}})^T \cdot \mathbf{U}$

Vogliamo trovare la distribuzione di $\mathbf{U}$ con il teorema del limite centrale. Ricordiamo che l'elemento $j-$esimo della statistica score è

$$U_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

e

$$\mathbf{J}^{-1/2}\mathbf{U} = \sum_{i=1}^{n} a_i \frac{(y_i - \mu_i)}{\text{Var}(y_i)}$$ → Var = 1

è una combinazione lineare di $(y_i - \mu_i)$, con a_i specifici e dove $\mathbf{J}^{-1/2}(\mathbf{J}^{-1/2})^T = \mathbf{J}^{-1}$. Il teorema classico del limite centrale non si riece ad applicare nella sua versione base, ma possiamo usare la condizione di Lindeberg che ci assicura che (sotto opportune

Stima dei parametri IX

condizioni) si può applicare il TLC anche in questo caso, dato che $a_i(y_i - \mu_i)$ è una variabili aleatoria con media 0 e varianza a_i^2, con $\sum_{i=1}^{n} a_i^2 = 1$ perchè $Var(\mathbf{J}^{-1/2}\mathbf{U}) = \mathbf{I}$

! ! !

$$\hat{\beta} = \beta + J^{-\frac{1}{2}}(J^{-\frac{1}{2}})U$$

$$N(\vec{0}, I)$$

quind. (*)

Abbiamo allora che (approssimativamente, per $n \Rightarrow \infty$)

(*) → $\hat{\beta} \sim N(\beta, \mathbf{J}^{-1})$ (7)

Notate come la stima di β è più precisa se aumenta la curvatura della funzione di verosimiglianza $(-\mathbf{J})$

$$c = (0, 0, 0 \ldots 1, 0 \ldots)$$

$$c\hat{\beta} = \hat{\beta}_J$$

$$c\beta = \beta_J$$

$$cJ^{-1} = J_{JJ}^{-1}$$

Stima dei parametri X

Come nel caso lineare, possiamo usare la distribuzione di $\hat{\boldsymbol{\beta}}$ per fare test d'ipotesi, intervalli di confidenza, etc ...
Per esempio, potremmo testare

$$H_0 : \beta_j = b_j \qquad H_1 : \beta_j \neq b_j$$

con la statistica

$$z = (\hat{\beta}_j - b_j)\sqrt{J_{jj}} \sim N(0,1)$$

se si assume nota la varianza, oppure

$$z = (\hat{\beta}_j - b_j)\sqrt{\hat{J}_{jj}} \sim N(0,1)$$

se viene stimata, dove $\hat{\mathbf{J}} = \mathbf{X}^T\hat{\mathbf{W}}\mathbf{X}$ e $\hat{W}_{ii}$ è una stima di $W_{ii} = \frac{1}{\mathrm{Var}(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$.
Nel secondo caso dovremmo usare una T di student, ma la distribuzione di $\hat{\beta}$ è approssimativamente normale, come la T.

Stima dei parametri XI

Possimao testare un'ipotesi sull'intero modello

$$H_0 : \boldsymbol{\beta} = \mathbf{b} \qquad H_1 : \boldsymbol{\beta} \neq \mathbf{b}$$

possono essere testate utilizzando la statistica di Wald

$$(\boldsymbol{\beta} - \mathbf{b})^T \mathbf{J} (\boldsymbol{\beta} - \mathbf{b}) \sim \chi^2_p$$

sfrutta il fatto che $(\beta - b) J^{-\frac{1}{2}} \sim N(\vec{0}, I)$

o nella forma

$$\mathbf{U}^T \mathbf{J}^{-1} \mathbf{U} \sim \chi^2_p$$

visto che $\mathbf{U} \sim N(\mathbf{0}, \mathbf{J})$ $\Rightarrow$ $U^T J^{-\frac{1}{2}} \sim N(\vec{0}, I)$

Stima dei parametri XII

Attenzione: Sebbene abbiamo derivato solo rispetto a $\boldsymbol{\beta}$ la log-verosimiglianza, delle volte dobbiamo stimare anche il parametro di dispersione (per esempio σ^2 nella regresisone o la dispersione della binomiale negativa). Il parametro da trovare è ϕ.

I modello vengono scelti in modo tale che la matrice di informazione ha elementi che connettono i $\boldsymbol{\beta}$ e ϕ asintoticamente pari a 0. Rendendo i parametri indipendenti e possono essere stimati indipendentemente.

Stima dei parametri XIII

La Devianza:

La verosimiglianza può essere usata per fare dei test sul modello. Una possibile idea è di confrontare il modello in esame con il modello **saturato** (o **saturo**), i.e. il modello con il maggior numero di parametri possibile (1 per ogni osservazione univoca).

Indichiamo con $L(\mathbf{y}; \hat{\boldsymbol{\beta}}_{\text{max}})$ la log-verosimiglianza per il modello saturo e come sempre $L(\mathbf{y}; \hat{\boldsymbol{\beta}})$ quella del modello in esame. Siamo interessati al rapporto

RAPPORTO FRA VEROSIMIGLIANZE

$$\lambda = \frac{\exp(L(\hat{\boldsymbol{\beta}}_{\text{max}}; \mathbf{y}))}{\exp(L(\hat{\boldsymbol{\beta}}; \mathbf{y}))}$$

o più precisamente al suo logaritmo moltiplicato per due, che chiamiamo **Devianza**

- $D = 2\log(\lambda) = 2\left(L(\hat{\boldsymbol{\beta}}_{\text{max}}; \mathbf{y}) - L(\hat{\boldsymbol{\beta}}; \mathbf{y})\right)$

Stima dei parametri XIV

Ricordiamo che l'approssimazione di Taylor della log-verosimiglianza ci dice che

$$L(\boldsymbol{\beta};\mathbf{y}) = L(\hat{\boldsymbol{\beta}};\mathbf{y}) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{U}(\hat{\boldsymbol{\beta}}) - \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{J}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

(approssimativamente) Se $\hat{\boldsymbol{\beta}}$ è lo stimatore di massima verosimiglianza, allora possiamo scrivere

$$L(\boldsymbol{\beta};\mathbf{y}) = L(\hat{\boldsymbol{\beta}};\mathbf{y}) - \frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{J}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})$$

e dato che $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{J}^{-1})$ allora

$$2\Big(L(\hat{\boldsymbol{\beta}};\mathbf{y}) - L(\boldsymbol{\beta};\mathbf{y})\Big) = (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\mathbf{J}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \sim \chi^2_p$$

Il risultato vale per ogni $\boldsymbol{\beta}$ di lunghezza p.

Stima dei parametri XV

Aggiungiamo e eliminiamo il valore "vero" dei $\boldsymbol{\beta}$, nel modello saturo e nel modello sotto esame, avendo

- $$D = 2\left(L(\hat{\boldsymbol{\beta}}_{\max}; \mathbf{y}) - L(\hat{\boldsymbol{\beta}}; \mathbf{y})\right) =$$
$$2\left(L(\hat{\boldsymbol{\beta}}_{max}; \mathbf{y}) - L(\boldsymbol{\beta}_{max}; \mathbf{y})\right) - 2\left(L(\hat{\boldsymbol{\beta}}; \mathbf{y}) - L(\boldsymbol{\beta}; \mathbf{y})\right) +$$
$$2\left(L(\boldsymbol{\beta}_{max}; \mathbf{y}) - L(\boldsymbol{\beta}; \mathbf{y})\right)$$

- $2\left(L(\hat{\boldsymbol{\beta}}_{max}; \mathbf{y}) - L(\boldsymbol{\beta}_{max}; \mathbf{y})\right) \sim \chi^2_n$, dove n è la lunghezza di $\boldsymbol{\beta}_{max}$;
- $2\left(L(\hat{\boldsymbol{\beta}}; \mathbf{y}) - L(\boldsymbol{\beta}; \mathbf{y})\right) \sim \chi^2_p$, dove p è la lunghezza di $\boldsymbol{\beta}$;
- $\nu = 2\left(L(\boldsymbol{\beta}_{max}; \mathbf{y}) - L(\boldsymbol{\beta}; \mathbf{y})\right) \geq 0$ è una costante che è uguale a 0, se i due modelli spiegano i dati nello stesso modo. Deve essere maggiore di zero perchè il saturo contiene il modello da testare (è un possibile caso particolare)

Stima dei parametri XVI

La differenza di due chi quadrato non è nota in generale, ma usando il teorema di Wilks, si può vedere che è

WILKS THEOREM

$$\boxed{D \sim \chi^2_{n-p,\nu}} \quad \cong \quad D = X + \nu \quad \text{con } X \sim \chi^2_{n-p}$$

Relazione simile vale per qualsiasi devianza, non solo calcolata con il saturo

La devianza può essere usata per confrontare modelli *annidati*. nello specifico ipotizziamo di voler testare

$$\boxed{H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \qquad H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1}$$

dove $\boldsymbol{\beta}_0$ ha q elementi (modello M_0) e $\boldsymbol{\beta}_1$ ha p elementi (modello M_1), con $q < p < n$.

PICCOLO

GRANDE

Possiamo calcolare

$$\Delta D = D_0 - D_1 = 2\left(L(\hat{\boldsymbol{\beta}}_{\max}; \mathbf{y}) - L(\hat{\boldsymbol{\beta}}_0; \mathbf{y})\right) - 2\left(L(\hat{\boldsymbol{\beta}}_{\max}; \mathbf{y}) - L(\hat{\boldsymbol{\beta}}_1; \mathbf{y})\right) =$$

$\geq D_1$

Stima dei parametri XVII

e abbiamo quindi che

$$\Delta D = 2\left(L(\hat{\boldsymbol{\beta}}_1; \mathbf{y}) - L(\hat{\boldsymbol{\beta}}_0; \mathbf{y})\right) \sim \chi^2_{p-q, v_0 - v_1}$$

dove

- $\nu_0 = 2\left(L(\boldsymbol{\beta}_{max}; \mathbf{y}) - L(\boldsymbol{\beta}_0; \mathbf{y})\right)$;
- $\nu_1 = 2\left(L(\boldsymbol{\beta}_{max}; \mathbf{y}) - L(\boldsymbol{\beta}_1; \mathbf{y})\right)$;

SOTTO H_0 : $L(\beta_0; y) = L(\beta_1; y)$

CHE' I DATI SONO SPIEGATI UGUALMENTE DA M_0 E M_1

Se entrambi i modelli sono equivalenti in termine di spiegazione dei dati, abbiamo

$$\nu_0 = \nu_1.$$

e $\Delta D \sim \chi^2_{p-q}$.

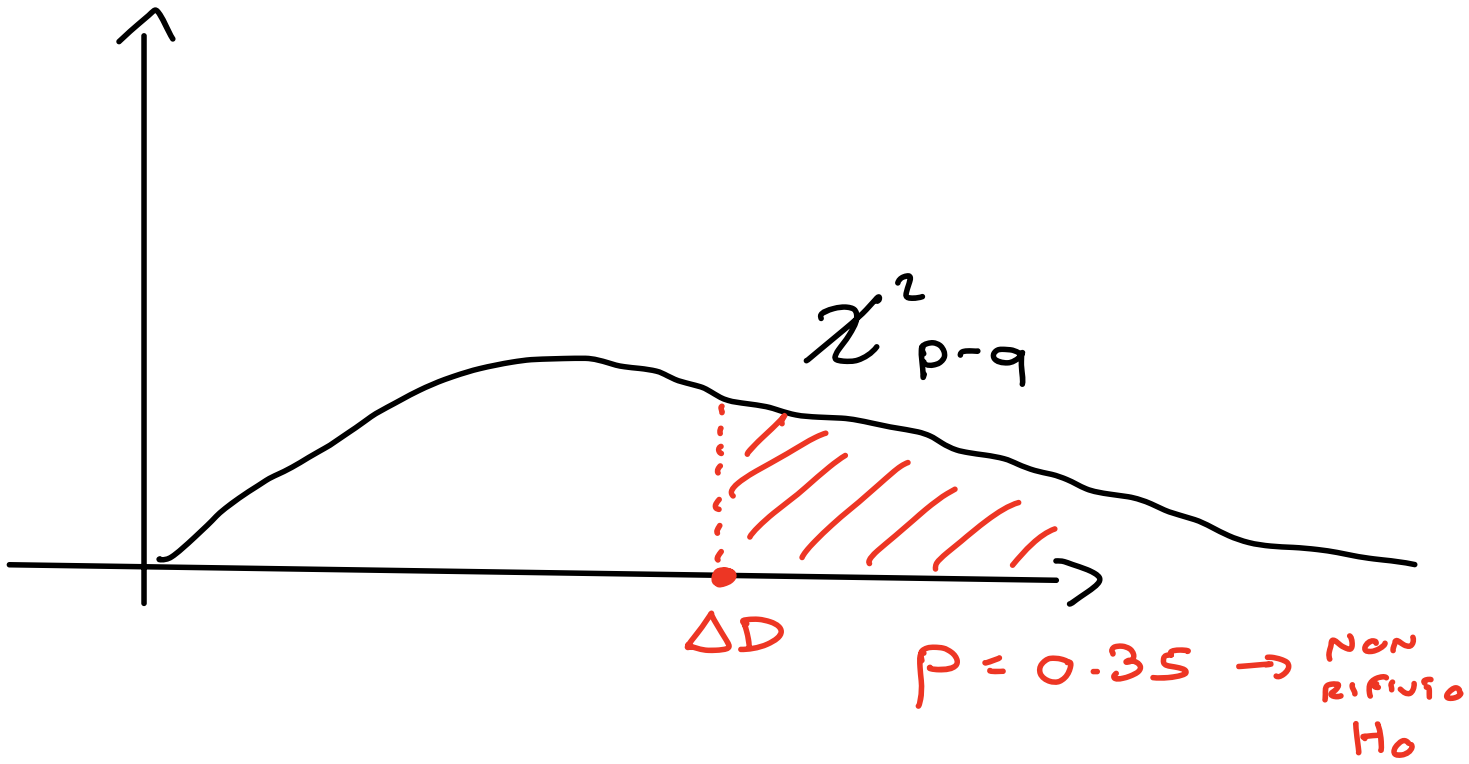
Se M_0 spiega in maniera peggiore i dati, rispetto a M_1, abbiamo che

$$\nu_0 > \nu_1$$

Stima dei parametri XVIII

e $\Delta D \sim \chi^2_{p-q,\nu_0-\nu_1}$.

Possiamo confrontare ΔD con un χ^2_{p-q}.
Se il valore ΔD si trova nella regione critica $\int_D^\infty f^*(x)dx < \alpha$, dove $f^*(x)$ è la densità di una χ^2_{p-q}, rifiutiamo H_0, altrimenti, per il rasoio di Occam, scegliamo il modello più parsimonioso.



Stima dei parametri XIX

Prendiamo il caso normale.

Possiamo calcolare la devianza per un modello normale, con lo scopo di confrontare modelli nested. Nel modello saturo, il vettore dei parametri è della stessa dimensione delle osservazionie, e abbiamo che $\mathbf{y} = \hat{\mathbf{y}}$, quindi

$$L(\hat{\boldsymbol{\beta}}_{max}; \mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2)$$

mentre per il modello di interesse è

$$L(\hat{\boldsymbol{\beta}}; \mathbf{y}) = -\frac{\sum_{i=1}^{n}(y_i - \mathbf{X}_{i.}\hat{\boldsymbol{\beta}})^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$$

e

$$D = 2\left(L(\hat{\boldsymbol{\beta}}_{max}; \mathbf{y}) - L(\hat{\boldsymbol{\beta}}; \mathbf{y})\right) = \frac{\sum_{i=1}^{n}(y_i - \mathbf{X}_{i.}\hat{\boldsymbol{\beta}})^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Sfortunatamente, questa non può essere usata direttamente dato che dipende da σ^2, che non conosciamo.

Stima dei parametri XX

Indichiamo con

- $D_0 = \frac{\sum_{i=1}^{n}(y_i - \mathbf{X}_{i.}\hat{\boldsymbol{\beta}}_0)^2}{\sigma^2} \sim \chi^2_{n-q,\nu_0}$
- $D_1 = \frac{\sum_{i=1}^{n}(y_i - \mathbf{X}_{i.}\hat{\boldsymbol{\beta}}_1)^2}{\sigma^2} \sim \chi^2_{n-p,\nu_1}$

Assumendo che M_1 fitta bene i dati, e quindi la χ^2 ha parametro di non centralià pari a 0, possiamo utilizzare la seguente statistica

$$F = \frac{D_0 - D_1}{p - q} \Big/ \frac{D_1}{n - p} \sim F_{p-q,n-p,\nu_0-\nu_1}$$

Quindi, se assumiamo che M_1 e M_0 spiegano allo stesso modo i dati, $\nu_0 = \nu_1$ e possiamo confrontare F con una $F_{p-q,n-p}$. Valori elevati di F suggeriscono che M_1 spieghi meglio di M_0.

Stima dei parametri XXI

Rappresentazione grafica del test di Wald e rapporto di verosimiglianze

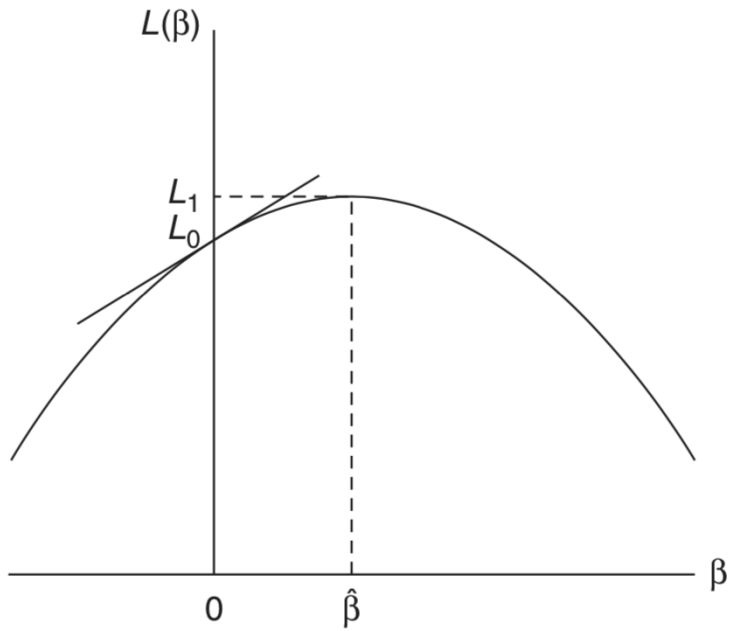


Figure 4.1 Log-likelihood function and information used in likelihood-ratio, score, and Wald tests of H_0: $\beta = 0$.

Stima dei parametri XXII

Concludiamo con dei residui

- Residui di Pearson

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mathsf{Var}}(y_i)}}$$

- deviance residuals

$$d_i = \mathsf{sign}(y_i - \hat{\mu}_i)\sqrt{2\left(L_i(\hat{\boldsymbol{\beta}}_{\mathsf{max}}; \mathbf{y}) - L_i(\hat{\boldsymbol{\beta}}); \mathbf{y}\right)}$$

dove L_i è il contributo dato alla log-ver. dalla i-esima osservazione