Modelli Lineari Generalizzati (GLM): parte I

Leonardo Egidi

A.A. 2021/2022

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

Indice

Testi consigliati e di riferimento

1. capitolo 6 sui GLM, in italiano
   Azzalini, A. (2001), *Inferenza Statistica: una Presentazione basata sul Concetto di Verosimiglianza*, Springer-Italia, Milano.
2. cap. 1-7, contiene esemplificazioni in R
   Faraway J. (2006) *Extending the linear model with R*, Chapman & Hall, Londra.

3 il principale e più classico manuale sui GLM
McCullagh, P., Nelder, J.A. (1989), *Generalized Linear Models*, Chapman & Hall, London.

4 testo sui GLM a livello intermedio
Dobson, A.J. and Barnett A. (2008), *An Introduction to Generalized Linear Models*,Third Edition, Chapman & Hall, London.

5 testo con applicazioni a dati attuariali
P. de Jong and G.Z. Heller (2008) *Generalized Linear Models for Insurance Data*, Cambridge University Press.

Modelli lineari generalizzati (GLM)

GLM

- Ampia classe di modelli di regressione con l'obiettivo di estendere il modello lineare (LM) classico con errori normali, per spiegare la relazione fra una variabile risposta e una o più variabili esplicative.
- La classe dei GLM include quelli lineari ma consente anche di introdurre relazioni non lineari per spiegare il comportamento della media della variabile risposta e assunzioni distributive coerenti con la natura della variabile risposta stessa.
- Struttura - sia di costruzione che di analisi - simile a quella dei LM.
- Le applicazioni includono: regressione binomiale, regressione di Poisson, modelli per l'analisi di tabelle di frequenza a più entrate, analisi dei dati di sopravvivenza.
- Funzioni per stimare ed analizzare modelli della forma GLM esistono in `R` (ma anche in altri pacchetti statistici: SAS, Stata, etc.).

Introduzione: il modello lineare (LM) Normale

Riconsideriamo gli elementi che definiscono il LM Normale.

La scrittura del LM classico $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
equivale a dire che $Y_i \sim N(\mu_i, \sigma^2)$ con Y_i e Y_j incorrelati (e quindi in questo caso indipendenti) se $i \neq j$.
È necessario in esso specificare quanto segue:

1. **Il modello distributivo per la variabile dipendente**: $Y_i \sim N(\mu_i, \sigma^2)$, ove la media di Y_i è $\mu_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ e $\boldsymbol{x}_i^{\mathrm{T}}$ $i-$esima riga della matrice X (in cui al solito il primo vettore colonna è un vettore unitario e $\boldsymbol{\beta}$ è un vettore che contiene p parametri ignoti), $i = 1, 2, \ldots, n$;
2. **La struttura di una componente sistematica (espressa come predittore lineare)**: $\eta_i = \sum_{j=0}^{p-1} x_{ij}\beta_j = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$;
3. **Il legame** tra valor medio e predittore lineare: $\mu_i = \eta_i$.

I parametri ignoti sono quindi il vettore $\boldsymbol{\beta}$ e σ^2. Nei LM questi parametri hanno variazioni indipendenti e appartengono a parti separate del modello.

I GLM: definizione

La classe dei GLM si ottiene generalizzando le componenti 1. e 3., mantenendo l'ipotesi di indipendenza e la forma lineare del predittore. In particolare:

- $\mapsto$ Consideriamo Y_i determinazioni di una variabile aleatoria la cui funzione di densità (o probabilità) è un membro di una famiglia di distribuzioni, la *famiglia esponenziale* che ha media μ_i;
- $\mapsto$ Consideriamo altre forme di legame tra predittore lineare $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ e valor medio μ_i del tipo:

$$g(E(Y_i)) = g(\mu_i) = \eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \ , \tag{1}$$

con $g(\cdot)$ funzione monotona e derivabile (che verrà detta *funzione di legame*).

La famiglia di dispersione esponenziale

- Introduciamo una famiglia da usare nella 1. che comprenda non solo la distribuzione normale, ma anche altre distribuzioni: ad esempio, Poisson, binomiale, esponenziale, gamma.
- La variabile aleatoria Y appartiene alla famiglia (di dispersione) esponenziale se ha funzione di densità (o probabilità) del tipo

$$f(Y;\theta,\phi) = \exp\left\{\frac{Y\theta - b(\theta)}{\phi} + c(Y,\phi)\right\}\ , \qquad (2)$$

dove θ e ϕ sono parametri scalari ignoti, $b(\cdot)$ e $c(\cdot)$ sono funzioni note la cui scelta individua una particolare distribuzione e il dominio di Y non dipende da θ o ϕ. Scriveremo $Y \sim EF(b(\theta),\phi)$.

La famiglia di dispersione esponenziale

- θ è il *parametro naturale* della famiglia esponenziale.
- ϕ è un parametro di dispersione o di scala. In alcuni casi questo è un valore noto. In realtà è quando il valore è ignoto che si parla più propriamente di famiglia di dispersione esponenziale.
- Come detto, numerose fra le più comuni distribuzioni discrete e continue appartengono alla famiglia (cioè, ad esempio, Normale, Gamma, Poisson, Binomiale)
- In alcune occasioni si può scrivere la funzione di densità (probabilità) come

$$f(Y;\theta,\phi) = \exp\left\{\frac{Y\theta - b(\theta)}{\phi/\omega} + c(Y,\phi)\right\} \ , \tag{3}$$

ove ω è un peso supposto noto.

Esempio: Poisson

- È la distribuzione di base per descrivere fenomeni di tipo "conteggio".
- Se $Y \sim \text{Poisson}(\lambda)$, la sua funzione di probabilità è

$$\begin{aligned} f(Y;\lambda) &= \frac{e^{-\lambda}\lambda^{Y}}{Y!} \\ &= \exp\{Y \log \lambda - \lambda - \log Y!\} \ , \end{aligned}$$

per $Y = 0, 1, \ldots$

- Questa ha la forma richiesta per la famiglia esponenziale con $\theta = \log \lambda$ parametro naturale, $\phi = 1$, $b(\theta) = \lambda = e^{\theta}$ e $c(Y, \phi) = -\log Y!$.
- Scriveremo $Y \sim EF(e^{\theta}, 1)$.

Esempio: binomiale

- Se $Y \sim \text{Bin}(n, \pi)$, la sua funzione di probabilità è

$$\begin{aligned} f(Y;\pi) &= \binom{n}{Y} \pi^Y (1-\pi)^{n-Y} \\ &= \exp\{\log \binom{n}{Y} + Y \log \pi + (n-Y)\log(1-\pi)\} \\ &= \exp\left\{ Y \log \frac{\pi}{1-\pi} + n \log(1-\pi) + \log \binom{n}{Y} \right\}, \quad \text{per } Y = 0, 1, \ldots, n. \end{aligned}$$

- Questa ha la forma richiesta per la famiglia esponenziale con $\theta = \log \frac{\pi}{1-\pi}$ parametro naturale, $\phi = 1$,

$$b(\theta) = -n \log(1-\pi)\big|_{\pi = \frac{e^\theta}{1+e^\theta}} = n \log(1+e^\theta)$$

e $c(Y, \phi) = \log \binom{n}{Y}$.

- Scriveremo $Y \sim EF(n \log(1+e^\theta), 1)$.

I momenti dell'EF

• Le funzioni $b(\cdot)$ e $c(\cdot)$ sono importanti nella valutazione e interpretazione dei momenti della distribuzione.

• Ricordiamo due risultati di base sulle derivate della funzione di log-verosimiglianza.

Siano $L(\theta) = L(\theta; Y)$ la verosimiglianza, $\ell(\theta) = \ell(\theta; Y) = \log L(\theta)$ la log-verosimiglianza e $\ell_*(\theta)$ la funzione score per θ a partire da una singola realizzazione campionaria Y. Allora, sotto condizioni di regolarità

$$E\left(\ell_*(\theta)\right) = E\left(\frac{d}{d\theta}\ell(\theta; Y)\right) = 0$$

e

$$I(\theta) = var\left(\ell_*(\theta)\right) = E\left(-\ell_{**}(\theta)\right) = E\left(-\frac{d^2}{d\theta^2}\ell(\theta; Y)\right).$$

I momenti dell'EF

Se Y è una realizzazione di una distribuzione appartenente alla famiglia esponenziale, la log-verosimiglianza per θ risulta:

$$\ell(\theta) = \frac{Y\theta - b(\theta)}{\phi} + c(Y, \phi) \ .$$

ϕ è una costante per cui si ha:

$$\ell_*(\theta) = \frac{d\ell}{d\theta} = \frac{Y - b'(\theta)}{\phi}$$

$$\ell_{**}(\theta) = \frac{d^2\ell}{d\theta^2} = -\frac{b''(\theta)}{\phi} \ .$$

Da queste espressioni segue che:

$$\boxed{E(Y) = \mu = b'(\theta)}$$

e

$$var\left(\frac{Y - b'(\theta)}{\phi}\right) = \frac{b''(\theta)}{\phi} \Rightarrow$$

$$\boxed{var(Y) = \phi b''(\theta)} \ .$$

I momenti dell'EF

- La funzione $b(\theta)$ è detta anche **funzione dei cumulanti**
- Di solito, si pone $V(\mu) = b''(\theta)$, quindi
  $\boxed{var(Y) = \phi V(\mu)}$
- La funzione $V(\mu)$ è detta *funzione di varianza* ed è intesa come funzione del valor medio μ anche se appare scritta come funzione di θ (basta invertire la relazione tra μ e θ data da $\mu = b'(\theta)$).
- Poiché la varianza della variabile risposta è legata al valor medio μ tramite una funzione di varianza $V(\mu)$, nei GLM sono presenti forme di eteroschedasticità.

Esempio: Poisson

Riprendendo l'esempio della Poisson(λ), abbiamo

$$b(\theta) = e^{\theta} \quad \text{e} \quad \phi = 1 \ .$$

Di conseguenza

$$E(Y) = b'(\theta) = e^{\theta} = \lambda$$

e

$$var(Y) = b''(\theta) = e^{\theta} = \lambda \ .$$

Allora, $V(\mu) = \mu$.

Esempio: binomiale

Nell'esempio della Bin(n, π), abbiamo

$$b(\theta) = n \log(1 + e^{\theta}) \quad \text{e} \quad \phi = 1 \ .$$

Di conseguenza

$$E(Y) = \mu = b'(\theta) = n \frac{e^{\theta}}{1 + e^{\theta}} = n\pi$$

e

$$var(Y) = b''(\theta) = n \frac{e^{\theta}}{(1 + e^{\theta})^2} = n\pi(1 - \pi) \ .$$

Allora, $V(\mu) = \mu(1 - \mu/n)$.

Tabella riassuntiva EF

Altre importanti distribuzioni del tipo $EF(b(\theta), \phi)$ sono indicate nella Tabella che segue, assieme agli ulteriori elementi caratteristici.

Distribuzione	Normale $N(\mu, \sigma^2)$	Poisson $Po(\mu)$	Binomiale $/m$ $Bin(m, \mu)/m$	Gamma $Ga(\omega, \omega/\mu)$
Supporto	$(-\infty, \infty)$	$\{0, 1, 2, \ldots\}$	$\{0, 1/m, 2/m, \ldots, 1\}$	$(0, \infty)$
ϕ	σ^2	1	1	ω^{-1}
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1 + e^\theta)$	$-\log(-\theta)$
$c(Y; \phi)$	$-\left(\frac{Y^2}{2\phi} + \frac{\log(2\pi\phi)}{2}\right)$	$-\log Y!$	$\log\binom{m}{mY}$	$\omega \log(\omega Y) - \log Y$ $-\log\Gamma(\omega)$
$\mu(\theta)$	θ	$\exp(\theta)$	$e^\theta/(1 + e^\theta)$	$-1/\theta$
$V(\mu)$	1	μ	$\mu(1 - \mu)$	μ^2
legame canonico	identità	logaritmo	logit	reciproco

La funzione legame

- La specificazione di un GLM viene completata attraverso la scelta della relazione fra il valor medio μ_i e le variabili esplicative (attraverso il predittore lineare η_i). In particolare, assumiamo che il legame fra μ_i, la media di Y_i, e $\boldsymbol{x}_i$, il vettore di covariate, abbia la forma

$$g(\mu_i) = \eta_i, \quad \text{con } \eta_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}$$

  e $g(\cdot)$ funzione nota monotona e derivabile. La funzione $g(\cdot)$ è detta *funzione legame*, o *collegamento*, tra μ_i e η_i.
- In questa maniera l'effetto delle covariate è lineare su una funzione (in generale) non lineare del valore atteso. Il legame fra μ_i e η_i è spesso non-lineare. La funzione g è invertibile per cui si può anche scrivere $\mu_i = g^{-1}(\eta_i)$.
- L'inversa g^{-1} è anche detta *funzione risposta*.

Specificazione di un GLM

• Siano Y_i le osservazioni e $\boldsymbol{x}_i$ le variabili esplicative, $i = 1, 2, \ldots, n$. Un GLM è caraterizzato dalle seguenti componenti:

1. **Componente casuale**: $Y_i \sim EF(b(\theta_i), \phi)$, indipendenti, con $E(Y_i) = \mu_i = b'(\theta_i)$, $i = 1, 2, \ldots, n$;
2. **Componente sistematica (predittore lineare)**: $\eta_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}$, ove $\boldsymbol{x}_i$ è un vettore di costanti e $\boldsymbol{\beta}$ un vettore di parametri;
3. **Legame**: esiste una funzione legame $g(\cdot)$ tale per cui $g(\mu_i) = \eta_i \Leftrightarrow \mu_i = g^{-1}(\eta_i)$, $i = 1, 2, \ldots, n$.

• I parametri ignoti sono $\boldsymbol{\beta}$ e talvolta ϕ.

• A differenza del LM tradizionale, nel caso dei GLM la precisa separazione della variabile risposta Y_i in due componenti, sistematica e casuale, non è in generale più possibile.

Specificazione di un GLM

In forma schematica:

componente casuale	predittore lineare	legame
$Y_i \sim EF(b(\theta_i), \phi)$ con $b'(\theta_i) = \mu_i$	$\eta_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}$	$g(\mu_i) = \eta_i$

- Riassumendo, le quantità:

  $\mapsto$ *parametro naturale* θ_i

  $\mapsto$ *media* μ_i

  $\mapsto$ *predittore lineare* η_i

sono legate tra loro attraverso le relazioni:

$$\eta_i = g(\mu_i) \qquad \text{e} \qquad \mu_i = b'(\theta_i)$$

Regressione Normale: LM

La regressione specificata da un modello lineare con errori gaussiani è un GLM.

Infatti

$$Y_i \sim N(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2) \ , \quad i = 1, 2, \ldots, n,$$

indipendenti.

In questo caso si specifica:

$\theta_i = \eta_i = \mu_i$ e quindi $g(\cdot)$ è la funzione identità ed è facile verificare che il parametro di dispersione è σ^2.

Nel caso di normalità è infine equivalente scrivere $Y_i \sim N(\mu_i, \sigma^2)$ oppure $Y_i = \mu_i + \epsilon_i$, con $\epsilon_i \sim N(0, \sigma^2)$.

Si ha una separazione completa tra componente sistematica μ_i (che dipende da $\boldsymbol{x}_i$ e $\boldsymbol{\beta}$) e componente casuale ϵ_i (che dipende solo da σ^2).

Regressione di Poisson

Siano

$$Y_i \sim \text{Poisson}(\mu_i) \ , \quad i = 1, 2, \ldots, n,$$

indipendenti.

Supponiamo che per una qualche funzione di legame $g(\cdot)$ valga la relazione $g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}$. Ad esempio, potrebbe essere la funzione logaritmo, nel qual caso la positività di $\mu_i = e^{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}}$ è assicurata.

La scelta:

$$\mu_i = e^{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}} = e^{\eta_i} \ ,$$

ossia $\eta_i = \log \mu_i$ è detta *regressione poissoniana* e il modello ha legame logaritmico, con $\theta_i = \eta_i$ $(= \log \mu_i)$.

Regressione binomiale

- Abbiamo visto i legami più naturali per il modello normale ed il modello di Poisson.
- Consideriamo modelli per dati binari: situazione in cui la variabile risposta è dicotomica. Ciò avviene in numerosi contesti applicativi.
- Quale legame usare per la regressione Binomiale?
  È un modello che rientra nei GLM, con

$$Z_i \sim \text{Bin}(m_i, \pi_i) \ .$$

  Si vuole modellizzare la probabilità π_i, e quindi la probabilità che la variabile risposta assuma un valore piuttosto che un altro, in funzione di certi valori delle variabili esplicative.

Regressione binomiale

Poiché, in genere, il valore di Z_i dipende anche da m_i e il nostro obiettivo è studiare la relazione tra le variabili esplicative e la probabilità di successo, è più naturale utilizzare come variabile risposta non il numero di successi ma la frequenza relativa di successi:

$$Y_i = Z_i / m_i,$$

che rappresenta appunto la proporzione di successi.

In questo modo, $E(Y_i) = \mu_i = \pi_i$ è la probabilità in questione. Il parametro μ_i deve essere contenuto nell'intervallo $[0, 1]$. Ha allora senso scegliere una funzione legame che rispetta questo vincolo. Una scelta ragionevole si ottiene ponendo

$$\mu_i = \Psi(\eta_i) \ ,$$

ove $\eta_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}$ e $\Psi(\cdot)$ è una funzione di ripartizione. La funzione legame è allora $g(\mu) = \Psi^{-1}(\mu)$.

Regressione binomiale

Scelte standard per $\Psi(\cdot)$ sono:

1. $\Psi(\eta) = \Phi(\eta)$, per cui $g(\mu) = \Phi^{-1}(\mu)$ *regressione probit*.
2. $\Psi(\eta) = \frac{e^\eta}{1+e^\eta}$, per cui

$$g(\mu) = \Psi^{-1}(\mu) = \log \frac{\mu}{1-\mu} \ ,$$

   regressione logit o *logistica*.

3. $\Psi(\eta) = 1 - \exp(-e^\eta)$, per cui
   $g(\mu) = \Psi^{-1}(\mu) = \log\{-\log(1-\mu)\}$, *complementary log-log*.

L'inferenza nei GLM

La funzione di (log) verosimiglianza

- L'assunzione distributiva consente di ricorrere agli stimatori di massima verosimiglianza .
- Per l'ipotesi di indipendenza tra le componenti, la log-verosimiglianza $\ell(\boldsymbol{\beta})$ ha la forma

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ \frac{Y_i \theta_i - b(\theta_i)}{\phi} + c(Y_i, \phi) \right\} = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \ ,$$

  ove θ_i è funzione di $\boldsymbol{\beta}$ attraverso la relazione

$$g(b'(\theta_i)) = \eta_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} \ .$$

- Per ottenere la SMV di $\boldsymbol{\beta}$ è necessario risolvere le *equazioni di verosimiglianza*:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0 \quad \text{per ogni } j = 0, 1, \ldots, p-1.$$

Stima di massima verosimiglianza

Si noti che utilizzando la regola a catena per la derivazione composta si può scrivere

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1}\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1}\frac{\partial \eta_i}{\partial \beta_j} \ ,$$

i cui termini possono essere riscritti come

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \theta_i} &= \frac{Y_i - b'(\theta_i)}{\phi} = \frac{Y_i - \mu_i}{\phi} \ , \\
\frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta_i) = \frac{var(Y_i)}{\phi} \ , \\
\frac{\partial \eta_i}{\partial \mu_i} &= g'(\mu_i) \ , \\
\frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} \ .
\end{aligned}$$

Stima di massima verosimiglianza

- Quindi abbiamo

$$\begin{aligned}\frac{\partial \ell_i}{\partial \beta_j} &= \frac{Y_i - \mu_i}{\phi} \frac{\phi}{var(Y_i)} \frac{1}{g'(\mu_i)} x_{ij} \\ &= \frac{(Y_i - \mu_i) x_{ij}}{\phi V(\mu_i) g'(\mu_i)} .\end{aligned}$$

- Le equazioni di verosimiglianza per $\boldsymbol{\beta}$ sono dunque

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\phi V(\mu_i) g'(\mu_i)} x_{ij} = 0,$$

$j = 0, 1, \ldots, p-1$, dove $\mu_i = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$.

Stima di massima verosimiglianza

- Il valore di $\hat{\boldsymbol{\beta}}$ è una soluzione del sistema per qualsiasi valore di ϕ.
- In forma matriciale le equazioni di verosimiglianza possono esser scritte come $\boldsymbol{X^T V(y - \mu) = 0}$ con $V = \text{diag}(v_1, \ldots, v_n)$ e

$$v_i = \frac{1}{\phi V(\mu_i)(g'(\mu_i))} \ .$$

- Quindi la funzione score è in generale il vettore di p elementi $\boldsymbol{X^T V(y - \mu)}$ il cui j-esimo elemento è

$$\sum_{i=1}^{n} (Y_i - \mu_i) v_i x_{ij} = 0.$$

Specificazione di un GLM in R

La specificazione in R è un'estensione naturale del modo in cui si specificano e si adattano i LM. Quindi si usa la funzione `glm()`, al posto della funzione `lm()`, nel seguente modo:

$$\texttt{glm}(\textit{formula}, \ \textit{family}, \ldots.)$$

dove

- *formula*: specifica la variabile risposta e descrive le variabili esplicative da includere nel predittore lineare (come in `lm()`).
- *family*: determina il modello probabilistico di riferimento e la funzione legame. Ad esempio, `normal(link=identity)` o `poisson(link=log)`.

$\Rightarrow$ Qual è la funzione legame di default in R?

Funzioni di legame canoniche

- In ogni GLM c'è una una funzione di legame che gode di particolari proprietà.
- Ricordiamo che $\eta = g(\mu)$ e $\mu = b'(\theta)$.
- Tra tutte le funzioni di legame si potrebbe privilegiare quella per cui $g(\mu) = \theta(\mu)$, secondo la quale il parametro naturale θ della famiglia esponenziale risulta combinazione lineare delle variabili esplicative. Formalmente,

$$\eta = g(\mu) = g(b'(\theta)) = \theta \ ,$$

  ossia $g(\cdot)$ è l'inversa di $b'(\cdot)$. Questa scelta di $g(\cdot)$ prende il nome di *legame canonico* per la famiglia esponenziale.
- Con questa scelta, il modello ha buone proprietà statistiche. Inoltre, per ciascuna distribuzione il legame canonico è la funzione di default in `R`.

Esempio: Poisson

Per la distribuzione di Poisson, si ha

$$b(\theta) = e^{\theta} \ .$$

Quindi, l'inversa di $b'(\cdot)$ è la funzione $\log(\cdot)$. Dunque,

$$\eta = g(\mu) = \log \mu$$

è il legame canonico per la distribuzione Poisson.

Segue che scrivere `poisson` per *family* nella funzione `glm()` di `R` è equivalente a scrivere `poisson(link=log)`.

Informazione di Fisher

- Consideriamo le derivate seconde di ℓ_i:

$$\begin{aligned} -E\left(\frac{\partial^2 \ell_i}{\partial\beta_j \partial\beta_k}\right) &= E\left(\frac{\partial \ell_i}{\partial\beta_j}\frac{\partial \ell_i}{\partial\beta_k}\right) \\ &= E\left(\left(\frac{(Y_i-\mu_i)x_{ij}}{\phi V(\mu_i) g'(\mu_i)}\right)\left(\frac{(Y_i-\mu_i)x_{ik}}{\phi V(\mu_i) g'(\mu_i)}\right)\right) \\ &= \frac{x_{ij}x_{ik}}{\phi V(\mu_i)(g'(\mu_i))^2}\ . \end{aligned}$$

La somma rispetto a i di tale quantità fornisce l'elemento (j,k)-esimo della matrice di informazione attesa. In notazione matriciale,

$$I(\beta) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

con $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_n)$ e

$$w_i = \frac{1}{\phi V(\mu_i)(g'(\mu_i))^2}\ .$$

Legame canonico

- L'adozione di una funzione di legame canonica ($\eta_i = g(\mu_i) = g(b'(\theta_i)) = \theta_i$) dà luogo ad alcune semplificazioni nell'inferenza basata sulla log-verosimiglianza $\ell(\boldsymbol{\beta})$.
- Per quanto riguarda la derivata prima, se la funzione legame è quella canonica si ha

$$g'(\mu_i)^{-1} = \frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = V(\mu_i)$$

e risulta quindi

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{(Y_i - \mu_i)x_{ij}}{\phi}$$

Legame canonico

- Questo implica che le equazioni di verosimiglianza si semplificano grandemente, risulta infatti

$$\sum_{i=1}^{n} Y_i x_{ij} = \sum_{i=1}^{n} x_{ij} \mu_i \ .$$

  In notazione matriciale, $\boldsymbol{X}^{\mathrm{T}} \boldsymbol{y} = \boldsymbol{X}^{\mathrm{T}} \hat{\boldsymbol{\mu}}$.
- I valori $\hat{\mu}_i$ del vettore $\hat{\boldsymbol{\mu}}$ sono i valori di μ_i corrispondenti ai $\hat{\boldsymbol{\beta}}_{ML}$.
- Queste equazioni sono coerenti con la struttura generale delle equazioni di verosimiglianza nelle famiglie esponenziali e implicano che (se è presente in $\boldsymbol{X}$ la colonna di 1 in corrispondenza dell'intercetta) il totale dei valori osservati y_i sia pari al totale dei valori previsti $\hat{\mu}_i$.

Legame canonico

- Un'altra interessante proprietà riguarda l'informazione di Fisher. Riprendendo la derivata della log verosimiglianza semplificata con il legame canonico si ha

$$\left(\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\right) = -\frac{x_{ij}}{\phi}\frac{\partial \mu_i}{\partial \beta_k},$$

  questa quantità non dipende da y per cui

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = E\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right),$$

  informazione attesa e informazione osservata coincidono.
- *Osservazione*: Il risultato generale di normalità asintotica dello SMV fornisce l'approssimazione $\hat{\boldsymbol{\beta}} \dot{\sim} N_p(\boldsymbol{\beta}, I(\beta)^{-1})$ per n elevato. Una stima consistente per $I(\boldsymbol{\beta})$ è $I(\hat{\boldsymbol{\beta}})$. Se ϕ è ignoto, anche questo parametro va stimato.

Esempio : Normale

Abbiamo $g(\mu) = \mu$, per cui $g'(\mu) = 1$. Inoltre, $V(\mu) = 1$, $\phi = \sigma^2$ e $\mu_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. Le equazioni di verosimiglianza diventano

$$\sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})x_{ij}}{\sigma^2} = 0 \ .$$

Semplificando σ^2, le equazioni si riducono alle equazioni normali dei minimi quadrati:

$$\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = 0$$

o, equivalentemente,

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \ .$$

Esempio: Poisson

Abbiamo $g(\mu) = \log \mu$, per cui $g'(\mu) = 1/\mu$. Inoltre, $V(\mu) = \mu$, $\phi = 1$ e $\mu_i = \exp(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$. Le equazioni di verosimiglianza diventano

$$\sum_{i=1}^{n} \left(Y_i - e^{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}} \right) x_{ij} = 0 \ ,$$

che sono non lineari in $\boldsymbol{\beta}$. Dunque, una soluzione esplicita in genere non esiste.

Algoritmi iterativi

- Le equazioni di verosimiglianza per i GLM non ammettono, in genere, soluzione esplicita ed è necessario risolverle con metodi iterativi.
- Nei GLM c'è la possibilità di affrontare con un unico algoritmo il problema della soluzione delle equazioni di verosimiglianza: questo algoritmo agisce risolvendo una successione di problemi di stima di minimi quadrati.
- Un metodo iterativo prevede di partire con un valore iniziale $\boldsymbol{\beta}^{(0)}$ e ottenere una sequenza $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots$, secondo uno schema di aggiornamento della $\boldsymbol{\beta}^{(t+1)}$ tramite la $\boldsymbol{\beta}^{(t)}$, fino a quando, ad esempio, il valore

  $||\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}|| = \sum_{j=1}^{p} (\beta_j^{t+1} - \beta_j^t)^2$ è sufficientemente piccolo.

Newton-Raphson

- Indichiamo con

$$\ell_*(\boldsymbol{\beta}) = \left(\frac{\partial \ell}{\partial \beta_1}, \ldots, \frac{\partial \ell}{\partial \beta_p} \right)^{\mathrm{T}}$$

il vettore *score*. Si vuole risolvere l'equazione

$$\ell_*(\boldsymbol{\beta}) = 0 \ .$$

- Il metodo di Newton-Raphson (che viene giustificato tramite uno sviluppo di Taylor) è basato sulla regola di aggiornamento alla $(t+1)-$esima iterazione

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + I(\boldsymbol{\beta}^{(t)})^{-1} \ell_*^{(t)} \ , \tag{4}$$

con $\ell_*^{(t)} = \ell_*(\boldsymbol{\beta}^{(t)})$ e con la matrice Hessiana sostituita dall'informazione attesa e con segno cambiato e calcolata in $\boldsymbol{\beta}^{(t)}$.

Newton-Raphson

- Il generico elemento di $\boldsymbol{\beta}^{(t)}$ è

$$I(\beta^{(t)})_{jk} = E\left(\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right) \ ,$$

  per $j, k = 1, 2, \ldots, p$.

- Questa modificazione prende il nome di metodo *scoring* di Fisher. Così facendo le espressioni risultano semplificate (si ricordi che se la funzione legame è quella canonica le due espressioni coincidono).

Newton-Raphson

L'espressione (3) se si premoltiplica I e II membro per $I(\beta^{(t)})$è equivalente a

$$I(\beta^{(t)})\beta^{(t+1)} = I(\beta^{(t)})\beta^{(t)} + \ell_*^{(t)} \ . \tag{5}$$

Si ricorda che la funzione score in notazione matriciale è pari a $\boldsymbol{X}^T\boldsymbol{V}(\boldsymbol{y}-\boldsymbol{\mu})$, il membro di destra può quindi essere scritto come

$$I(\beta^{(t)})\beta^{(t)}+\ell_*^{(t)} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}\beta^{(t)}+\boldsymbol{X}^T\boldsymbol{V}^{(t)}(\boldsymbol{y}^{(t)}-\boldsymbol{\mu}^{(t)}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{s}^{(t)} \ ,$$

ove $\boldsymbol{s}^{(t)}$ è un vettore di *pseudo-dati* la cui componente i-esima $s_i^{(t)}$ è pari a

$$\boldsymbol{x}_i^{\mathrm{T}}\beta^{(t)} + (y_i - \mu_i^{(t)})g'(\mu_i^{(t)}) = g(\mu_i^{(t)}) + (y_i - \mu_i^{(t)})g'(\mu_i^{(t)})$$

e si noti che $\boldsymbol{V}^{(t)} = \boldsymbol{W}^{(t)}\boldsymbol{G}^{(t)}$ con $\boldsymbol{G}^{(t)} = \mathrm{diag}(g'(\mu_1^{(t)}), \ldots, g'(\mu_n^{(t)}))$.

Newton-Raphson

- Si noti peraltro che se si considera lo sviluppo in serie della funzione $g(y_i)$ a partire da μ_i si ottiene

$$g(y_i) \cong g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$$

- Allora, ricordando che

$$I(\beta) = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{X},$$

e sostituendo quest'ultima nella (4) si ha

$$(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}^{(t)} \boldsymbol{X}) \boldsymbol{\beta}^{(t+1)} = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W}^{(t)} \boldsymbol{s}^{(t)}$$

Newton-Raphson

- L'iterazione di Newton-Raphson è

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(t)}\boldsymbol{s}^{(t)} \ . \tag{6}$$

- Questo argomento mostra che ogni passo dell'algoritmo è equivalente a una stima ai minimi quadrati ponderati, sebbene i valori s_i e i pesi cambino ad ogni passo. Per questo prende il nome di *Algoritmo dei Minimi Quadrati Pesati Iterati*.

La stima del parametro di dispersione

- Nel caso del LM, la stima di $\boldsymbol{\beta}$ avviene indipendentemente dal valore della varianza σ^2. C'è un fenomeno identico per il parametro di dispersione ϕ nei GLM.
- La soluzione delle equazioni di verosimiglianza per $\boldsymbol{\beta}$, date da

$$\sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{\phi V(\mu_i) g'(\mu_i)} = 0 \ ,$$

  è la stessa per qualsiasi scelta di ϕ. Di conseguenza, la stima di $\boldsymbol{\beta}$ ha la stessa forma se ϕ è noto oppure no.
- Nelle situazioni che richiedono una stima di ϕ, si potrebbe ricorrere alla SMV. Nella pratica è più comune utilizzare uno stimatore alternativo, numericamente più stabile della SMV, e più robusto rispetto a scostamenti dal modello.

La stima del parametro di dispersione

- Ricordiamo che $var(Y_i) = \phi V(\mu_i)$. In altre parole,

$$\frac{E((Y_i - \mu_i)^2)}{V(\mu_i)} = \phi.$$

  Questo suggerisce la seguente stima per ϕ:

$$\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} \ ,$$

  ove

$$\widehat{\mu}_i = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}})$$

  sono i valori di μ_i stimati. Questi vengono messi a disposizione dall'algoritmo iterativo per $\boldsymbol{\beta}$.
- In generale, $\widehat{\phi}$ è consistente.
- Nel caso del LM con legame identità, $\widehat{\phi}$ coincide con l'usuale espressione di S^2; questa connessione spiega il termine $n - p$ al denominatore di $\widehat{\phi}$.

Risultato di `glm` in `R`

Il risultato della funzione `glm` è una lista che contiene varie informazioni sul modello stimato. Queste sono estraibili attraverso opportune funzioni. Le principali sono:

- `coef`: valori stimati dei parametri;
- `summary`: sintesi dei risultati della stima;
- `deviance`: per la devianza;
- `resid`: residui del modello;
- `predict`: valori previsti dal modello;
- `anova`: analisi della devianza;
- `plot`: analisi grafiche.

Stime dei parametri

Con `glm` i parametri $\boldsymbol{\beta}$ vengono stimati con il metodo della massima verosimiglianza. Per n moderatamente grande, la distribuzione stimata approssimata dello SMV $\hat{\boldsymbol{\beta}}$ è

$$\hat{\boldsymbol{\beta}} \dot{\sim} N_p(\boldsymbol{\beta}, [I(\hat{\boldsymbol{\beta}})]^{-1}) \ ,$$

con

$$I(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}^{\mathrm{T}} \hat{\boldsymbol{W}} \boldsymbol{X} \ ,$$

con $\hat{W}$ calcolato nella SMV $\hat{\beta}$. Con questo si ottengono le varianze delle componenti di $\hat{\beta}$: sono gli elementi della diagonale della matrice $(\boldsymbol{X}^{\mathrm{T}} \hat{\boldsymbol{W}} \boldsymbol{X})^{-1}$. Inoltre, la matrice $(\boldsymbol{X}^{\mathrm{T}} \hat{\boldsymbol{W}} \boldsymbol{X})^{-1}$ serve come matrice di varianza e covarianza per $\hat{\boldsymbol{\beta}}$. Cioè, i termini fuori dalla diagonale contegono le covarianze.

Adeguatezza dei modelli

Confronto fra modelli

- Consideriamo il problema di confrontare due GLM annidati, indicati con M_C e M_R, tali che $M_R \subset M_C$. In particolare, il modello corrente M_C contiene p parametri, e il modello ridotto M_R contiene p_0 parametri, con $p > p_0$.
- Si consideri la partizione di $\boldsymbol{\beta}$ in $\boldsymbol{\beta} = (\boldsymbol{\beta}_{MR}, \boldsymbol{\beta}_{MC})$, con $\boldsymbol{\beta}_{MR} = (\beta_1, \ldots, \beta_{p_0})$ e $\boldsymbol{\beta}_{MC} = (\beta_{p_0+1}, \ldots, \beta_p)$. Si supponga di voler verificare

$$H_0 : \boldsymbol{\beta}_{MC} = 0 \quad \text{contro} \quad H_1 : \boldsymbol{\beta}_{MC} \neq 0 \ .$$

- Come criterio per confrontare M_C e M_R si vuole usare il rapporto di verosimiglianza

$$W = 2\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{MR})\} \ .$$

Devianza nel LM

- Nei LM con errori normali, quando σ^2 è noto, il rapporto di verosimiglianza è funzione della devianza (somma dei quadrati dei residui) $D = SSE = \sum_i (y_i - \hat{\mu}_i)^2$ associata a ciascuno dei due modelli. In particolare per confrontare due modelli annidati ($M_R \subset M_C$), il rapporto di verosimiglianza suggerisce di rifiutare H_0 per valori elevati della statistica

$$W = 2\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{MR})\} = \frac{D_{MR} - D}{\sigma^2} \ ,$$

  ove $D_{MR} = SSE_{H_0}$ e $D = SSE$ rappresentano la somma dei quadrati dei residui con riferimento al modello ridotto e corrente, rispettivamente.
- Sotto H_0 tale statistica ha distribuzione $\chi^2_{p-p_0}$.

Confronto fra modelli con LR test

- In analogia ai LM con errori normali, per i GLM si vuole esprimere il rapporto di verosimiglianza in modo che mantenga chiara la connessione tra le due classi di modelli. A tale scopo abbiamo bisogno di estendere il concetto di devianza all'ambito dei GLM.
- La log-verosimiglianza per un GLM è

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \ ,$$

con

$$\ell_i(\boldsymbol{\beta}) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \ .$$

- Nel caso di modelli annidati, risulta che la statistica

$$W = 2\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{MR})\}$$

ha sotto H_0 distribuzione asintotica $\chi^2_{p-p_0}$.

Modello saturo

- L'analogia formale con il LM normale può essere evidenziata introducendo la verosimiglianza associata al modello *saturo* o *massimale*. Il modello saturo è definito come:
  - $\mapsto$ un GLM basato sulla stessa distribuzione del modello corrente;
  - $\mapsto$ un GLM con la stessa funzione legame del modello corrente;
  - $\mapsto$ un GLM con numero di parametri pari a n;
- Le funzioni di log-verosimiglianza per il modello saturo e per il modello corrente possono essere valutate nelle rispettive SMV (date da $\tilde{\theta}$ e da $\hat{\theta}$). Se il modello corrente si adatta bene ai dati, $\ell(\tilde{\theta})$ dovrebbe essere equivalente a $\ell(\hat{\theta})$. Se invece il modello corrente si adatta male, allora $\ell(\hat{\theta})$ dovrebbe essere molto più piccola di $\ell(\tilde{\theta})$.

Devianza nei GLM

Formalizzando, la quantità

$$D(y;\hat{\theta}) \quad = \quad 2\phi\{\ell(\tilde{\theta}) - \ell(\hat{\theta})\} = \sum_{i=1}^{n} D_i$$

con

$$D_i = 2\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\} \ ,$$

è detta *funzione di devianza* del modello e la funzione

$$\frac{D(y;\hat{\theta})}{\phi} = \frac{1}{\phi}\sum_{i=1}^{n} D_i \tag{7}$$

è la *devianza scalata*, che è sempre non negativa. La quantità sopra è piccola quando il modello è ben stimato, mentre se la stessa è grande il modello corrente non si adatta bene ai dati. In questo senso la *devianza* è equivalente alla SSE nel LM.

Devianza nei GLM

- Si osservi che $\ell(\tilde{\theta})$ rappresenta la log-verosimiglianza ottenuta ponendo $\tilde{\mu}_i = b'(\theta_i) = y_i (\Leftrightarrow (\partial \ell_i / \partial \theta_i) = 0)$, ovvero adattando il modello di regressione saturo con $p = n$ parametri. Tale modello ha tanti parametri quante sono le osservazioni.
- La differenza delle log-verosimiglianze tra modello saturo e quello in esame rappresenta una misura della diminuzione della bontà di adattamento dovuta al passaggio dal modello saturo a quello corrente con $p < n$ variabili esplicative. Pertanto, la devianza $D(y; \hat{\theta})$ ha lo stesso ruolo della somma dei quadrati dei residui (devianza residua SSE) nel LM classico.
- Dunque, $\ell(\tilde{\theta})$ fornisce un valore di riferimento per la log-verosimiglianza. Tale modello non è di utilità pratica, dal momento che non sintetizza i dati, ma serve come termine di confronto per il modello effettivamente in esame.

Esempio: Normale

- $Y_i \sim N(\mu_i, \sigma^2)$, $b(\theta) = \frac{\theta^2}{2}$, $\theta = \mu = b'(\theta)$ e $\phi = \sigma^2$.
- $\ell(\theta) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2$
- Per il modello saturo si ha $\tilde{\mu}_i = y_i$, e quindi

$$\ell(\tilde{\theta}) = -\frac{n}{2} \log \sigma^2 \ .$$

- Per il modello corrente si ha $\hat{\mu}_i = x_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}$, e quindi

$$\ell(\hat{\theta}) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

- Allora, la devianza scalata è

$$\frac{D(y; \hat{\theta})}{\sigma^2} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2} \ ,$$

che coincide con la SSE del modello corrente, divisa per σ^2.

Esempio: Poisson

- $Y_i \sim Poisson(\mu_i)$, $b(\theta_i) = e^{\mu_i} = b'(\theta_i)$, $\phi = 1$, $\log \mu_i = x_i^{\mathrm{T}} \boldsymbol{\beta}$
- $\ell(\theta) = \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \mu_i$
- Per il modello saturo si ha $\tilde{\mu}_i = y_i$, e quindi

$$\ell(\tilde{\theta}) = \sum_{i=1}^{n} y_i \log y_i - \sum_{i=1}^{n} y_i \ .$$

- Per il modello corrente si ha $\log \hat{\mu}_i = x_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}$, e quindi

$$\ell(\hat{\theta}) = \sum_{i=1}^{n} y_i \log \hat{\mu}_i - \sum_{i=1}^{n} \hat{\mu}_i \ .$$

-

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

Esempio: Binomiale

- $Y_i \sim Bin(1, \pi_i)$, con $\pi_i = Pr(Y_i = 1) = E(Y_i) = \mu_i$
- $\ell(\theta) = \sum_{i=1}^{n} (y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i))$
- Per il modello saturo si ha $\tilde{\mu}_i = y_i$ e

$$\ell(\tilde{\theta}) = \sum_{i=1}^{n} (y_i \log y_i + (1 - y_i) \log(1 - y_i)) \ .$$

- Per il modello corrente si ha $\text{logit}(\hat{\mu}_i) = x_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}$ e

$$\ell(\hat{\theta}) = \sum_{i=1}^{n} (y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)) \ .$$

- Quindi

$$D(y; \hat{\theta}) \quad = \quad 2 \sum_{i=1}^{n} \left(y_i \log \frac{y_i}{\hat{\pi}_i} + (1 - y_i) \log \frac{1 - y_i}{1 - \hat{\pi}_i} \right) = 2 \sum_{i=1}^{n} \sum_{j=1}^{2} o_{ij} \log \frac{o_{ij}}{e_{ij}}$$

con $o_{i\cdot} = (y_i, 1 - y_i)$ e $e_{i\cdot} = (\hat{\pi}_i, 1 - \hat{\pi}_i)$.

Confronto fra modelli

- Nel caso di modelli annidati M_C e M_R, il test del rapporto di verosimiglianza è

$$W = 2\left\{\ell(\hat{\beta}) - \ell(\hat{\beta}_{MR})\right\} = \frac{D(Y, \hat{\theta}_{MR}) - D(Y, \hat{\theta})}{\phi} \ ,$$

  che per $n \to \infty$ segue la distribuzione $\chi^2_{p-p_0}$ sotto H_0.
- Il test della validità del modello ridotto viene dal confronto di

$$W = \frac{D(Y, \hat{\theta}_{MR}) - D(Y, \hat{\theta})}{\phi}$$

  con i quantili della distribuzione $\chi^2_{p-p_0}$. Si rifiuta H_0 se la statistica assume valori elevati (*p-value* piccolo).
- Si noti che nel caso in cui ϕ non sia noto, si può proporre una sua stima che in analogia a quanto avviene nel caso del LM stimi il parametro di dispersione attraverso la devianza dei residui diviso il numero di gradi di libertà, quindi

$$\widehat{\phi} = \frac{D(Y, \hat{\theta})}{(n-p)}$$

  e si usa la stessa procedura.

Esempio Poisson con R

```
# Numero di visite mediche (y) in funzione dell'eta' (x)

> fit <- glm(y~x,family=poisson)
> summary(fit)

Call:
glm(formula = y ~ x, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4869  -0.9250  -0.4152   0.4447   2.4897

Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -0.491953  0.365772  -1.345  0.179
x            0.068856  0.007205   9.557   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 136.494  on 15  degrees of freedom
Residual deviance:  19.265  on 14  degrees of freedom
AIC: 89.853

Number of Fisher Scoring iterations: 3
```

```
> anova(fit)
Analysis of Deviance Table
Model: poisson, link: log
Response: y
Terms added sequentially (first to last)
             Df Deviance Resid.  Df Resid. Dev
NULL                                    15    136.494
x            1  117.229                 14     19.265
> fit1 <- glm(y~x+I(x^2),family=poisson)
> anova(fit1)
Analysis of Deviance Table
Model: poisson, link: log
Response: y
Terms added sequentially (first to last)
            Df Deviance Resid.  Df Resid.  Dev
NULL                                15   136.494
x            1     117.229         14    19.265
I(x^2)       1       6.681         13    12.584
> summary(fit1)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0116  -0.3954   0.0421   0.3014   1.8836
Coefficients:
                   Estimate      Std. Error    z value Pr(>|z|)
(Intercept)         -3.8398836  1.4722613       -2.608  0.009103 **
x                    0.2256146  0.0653598        3.452  0.000557 ***
I(x^2)              -0.0017462  0.0007134       -2.447  0.014385 *
Null deviance: 136.494  on 15  degrees of freedom
Residual deviance:  12.584  on 13  degrees of freedom
AIC: 85.172
Number of Fisher Scoring iterations: 3
```

```
> fitp <- glm(y~x+I(x^2)+I(x^3)+I(x^4),family=poisson)

> anova(fitp)
Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)

          Df  Deviance Resid. Df Resid. Dev
NULL                         15     136.494
x          1    117.229        14      19.265
I(x^2)     1      6.681        13      12.584
I(x^3)     1      0.091        12      12.493
I(x^4)     1      0.338        11      12.154
```

Esempio Binomiale con R

```
# Proporzione di cavie decedute in base a dose veleno
> attach(rotenone)
> Kill <- cbind(Kill.1,Kill.2)
> Poison <- as.factor(Poison)
> fit <- glm(Kill ~ Poison + Logdose,binomial)
> summary(fit)

Call:
glm(formula = Kill ~ Poison + Logdose, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7725  -1.0948   0.5153   1.4039   2.2419

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.2654     0.2797   -11.673  < 2e-16 ***
Poison2            -1.6034     0.2656   -6.036    1.58e-09 ***
Poison3            -0.6911     0.2309   -2.994    0.00276 **
Logdose             4.8277     0.3395    14.222   < 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: 350.069  on 16  degrees of freedom
Residual deviance:  31.971  on 13  degrees of freedom
AIC: 98.955
Number of Fisher Scoring iterations: 4
```

```
> anova(fit)
Analysis of Deviance Table
Model: binomial, link: logit
Response: Kill
Terms added sequentially (first to last)
               Df  Deviance Resid.  Df Resid.  Dev
NULL                                   16       350.07
Poison          2    18.00             14       332.07
Logdose         1   300.10             13        31.97
> fit1 <- glm(Kill ~ Logdose,binomial)
> anova(fit1,fit)
Analysis of Deviance Table
Model 1: Kill ~ Logdose
Model 2: Kill ~ Poison + Logdose
       Resid. Df    Resid. Dev     Df   Deviance
1         15         71.782
2         13         31.971         2    39.811
```

Residui nei GLM: residui di Pearson

- Nel caso dei LM, i residui sono uno strumento per valutare l'adeguatezza di un modello stimato.
- Esistono diverse estensioni di residui per i GLM.
- L'estensione diretta del concetto di residuo standardizzato è data da

$$r_{Pi} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} V(\hat{\mu}_i)}} \,, \tag{8}$$

che è detto *residuo di Pearson*. La definizione è analoga alla definizione di residui nei LM, in cui viene stimato il termine di errore ϵ_i.

Residui nei GLM: residui di devianza

Poiché nei GLM tale ϵ_i non esiste, ha senso considerare il contributo alla devianza. Infatti nel LM classico la devianza dei residui (SSE) è pari a

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - x_i^{\mathrm{T}} \hat{\beta})^2 \ ,$$

mentre nel GLM l'analoga quantità è data dalla devianza. Ricordiamo che la devianza è

$$D(y, \hat{\theta}) = \sum_{i=1}^{n} D_i \ .$$

Valori grandi di D_i corrispondono a dati che non vanno bene per il modello. Definiamo allora

$$r_{Di} = \mathrm{sgn}(y_i - \hat{\mu}_i) \sqrt{D_i} \ ,$$

che prende il nome di *residuo di devianza* per il modello.

Residui nei GLM: residui di Anscombe

- Uno sviluppo di Taylor mostra che $r_{Pi} \approx r_{Di}$.
- Esistono, infine, i residui di Anscombe

$$r_{Ai} = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}} \ ,$$

  in cui $A(x)$ è una funzione opportuna che serve per rendere la loro distribuzione prossima alla normale.
- Se è valido il modello, (molto) approssimativamente i residui (di qualsiasi tipo), riscalati per $\hat{\phi}$ se necessario, dovrebbero seguire la distribuzione $N(0, 1)$. Pertanto,
  - $\mapsto$ grafici che controllano se i residui seguono una distribuzione normale (utile anche per identificare valori anomali);
  - $\mapsto$ grafici dei residui contro i valori stimati $\hat{Y}_i$;
  - $\mapsto$ grafici dei residui contro le variabili esplicative (corretta specificazione della formula del modello),

  possono essere usati per controllare la bontà del modello stimato.