BREVI CENNI SUI

MODELLI LINEARI GENERALIZZATI

- I modelli lineari generalizzati (GLM) sono una generalizzazione del *modello lineare* che abbiamo definito nell'ambito dell'analisi di regressione lineare.
- Nel modello lineare classico si ipotizza che la variabile dipendente, in particolare il termine di errore del modello, sia distribuito come una normale, invece nell'ambito dei modelli lineari generalizzati la variabile dipendente può essere distribuita come una qualsiasi variabile casuale della famiglia esponenziale e dunque, oltre alla v.c. normale anche le variabili casuali *binomiale, poissoniana, gamma, normale inversa* e altre.
- Ciascuna distribuzione (binomiale, poissoniana, gamma, normale inversa) corrisponde ad un tipo diverso di variabile dipendente.

Corrispondenza tra distribuzioni di probabilità e tipo di variabile

Variabile	Distribuzione	Espressione
Reale $(-\infty, +\infty)$	Normale	$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y-\mu)^2}{2\sigma^2}\}$
Intera $[0, \infty]$	Poisson	$f(y) = \frac{\mu^y \exp(-\mu)}{y!}$
Dicotomica $\{0, 1\}$	Bernoulli	$f(y) = p^y(1-p)^{(1-p)}$
Intera $[0, N]$	Binomiale	$f(y) = \binom{N}{y} p^y (1-p)^{N-y}$

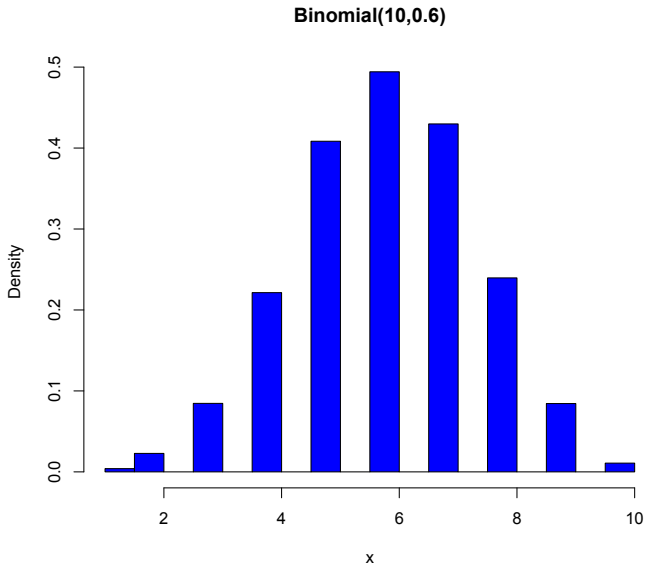
- L'idea sottostante i GLM è: cercare di ricondursi ad una situazione il più vicina possibile a quella della regressione lineare. Ovvero alla situazione in cui

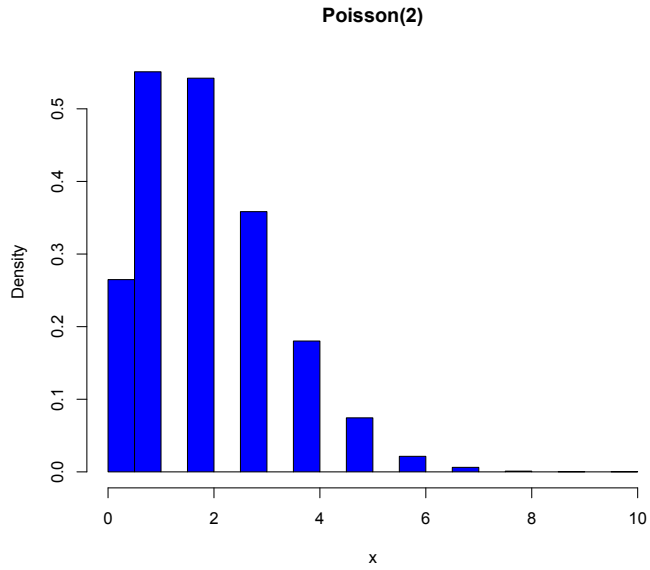
$$\mu = E(Y|X) = \alpha + \beta X$$

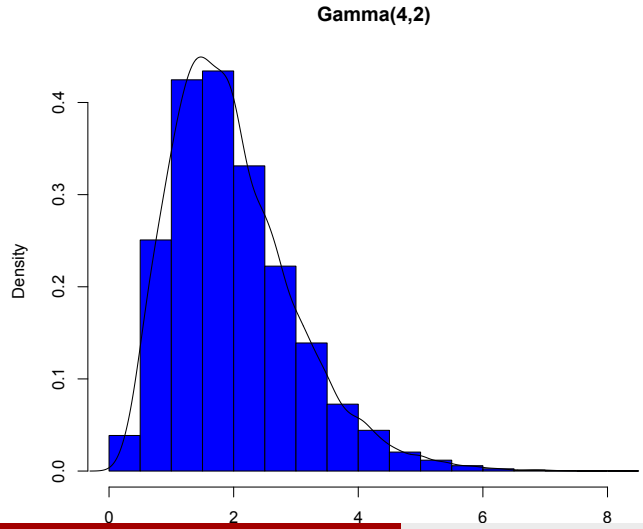
- *Quindi si cerca su quale scala la media di una variabile anche non Gaussiana è esprimibile come funzione lineare di altre variabili.*
  Vediamo i vari casi

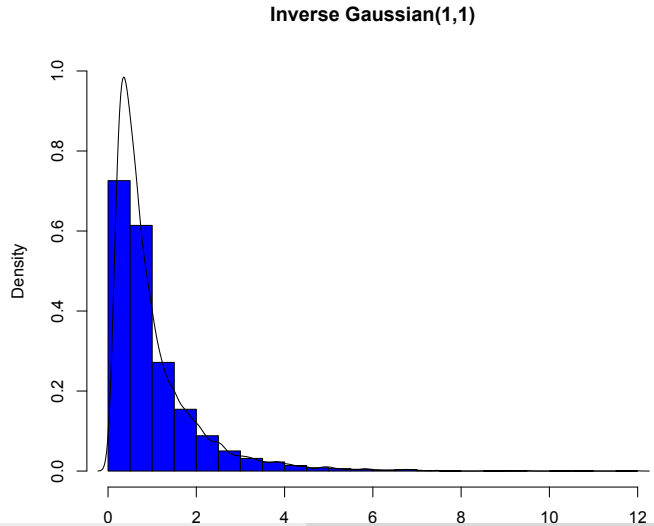
Variabile	Distribuzione	Trasformazione
Continua	Simmetrica	Potenza$=\begin{cases} \mu^{\alpha} & \alpha \neq 0 \\ \log \mu & \alpha = 0 \end{cases}$
Intera $[0, \infty]$	Poisson	Logaritmo naturale ($\log(\mu)$)
Dicotomica $\{0, 1\}$	Bernoulli	Logit ($\log \frac{\mu}{1-\mu}$), doppio logaritmica (cloglog $\log(-\log(1-\mu))$), funzione quantile della normale (probit $\Phi(\mu)$)
Intera $[0, N]$	Binomiale	Logit, logaritmo,

Esempi di distribuzioni:









Variabile 0/1

Per una variabile dicotomica y (dati di presenza assenza ad esempio), si modellizza la *probabilità di osservare un 1*. Infatti se prendiamo $y = 0, 1$ e calcoliamo $\bar{y} = \frac{1}{n}\sum_i y_i = \frac{n_1}{n}$ è uguale alla proporzione di 1 osservata nel campione.
Quindi cerco un modello in cui $\mu = Prob(y = 1)$, lo ottengo ponendo $z = \log \frac{\mu}{1-\mu}$ che è il *logit* di μ. Ora stimiamo un modello di regressione usuale per la nuova variabile z

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

In questa situazione si parla di **Regressione logistica**

Variabile Intera $[0, \infty]$

Questa situazione si presenta quando abbiamo a disposizione dei conteggi, tipicamente il numero di individui in diverse regioni. Allora assumiamo che $z = \log y$

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

in pratica diciamo che $E(Y|X) = \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p\}$.
Questo modello è anche detto **Regressione di Poisson**

Valutazione dei modelli

Per valutare la qualità di un modello lineare generalizzato utilizziamo due criteri:

% **di devianza spiegata** questa è semplicemente il rapporto tra la devianza del modello e la devianza totale, ovvero l'equivalente dell'R^2 visto nell'ambito dei modelli lineari

AIC ovvero **Akaike information criterion** è un criterio basato su di una misura di entropia come definita in teoria dell'informazione. La misura d'informazione è la *log-verosimiglianza*.

Valutazione dei modelli

- In gergo colloquiale spesso "verosimiglianza" è usato come sinonimo di "probabilità", ma in campo statistico vi è una distinzione tecnica precisa. Esempio: Una persona potrebbe chiedere "Se lanciassi una moneta non truccata 100 volte, qual è la probabilità che esca testa tutte le volte?" oppure "Dato che ho lanciato una moneta 100 volte ed è uscita testa 100 volte, qual è la verosimiglianza (likelihood) che la moneta sia truccata?". La logica delle due domande non è equivalente, esprimono due punti di vista diversi.
- Una distribuzione di probabilità che dipende da un parametro si può considerare secondo due diversi punti di vista: Il primo interpreta la distribuzione di probabilità come una funzione del risultato, dato un valore fissato del parametro mentre il secondo la interpreta come una funzione del parametro, dato un risultato fissato.
- Nell'ultimo caso la funzione è chiamata **funzione di verosimiglianza** del parametro, e indica quanto verosimilmente il valore di un parametro è plausibile rispetto al risultato osservato (campione).

Valutazione dei modelli

- In generale, poichè le funzioni di verosimiglianza hanno la forma di prodotti di funzioni, si preferisce misurare la plausibilità di un risultato su scala logaritmica, ovvero se $L(\theta, x)$ è la funzione di verosimiglianza, prenderemo come riferimento $\log L(\theta, x)$
- **La normale**: verosimiglianza Gaussiana date n osservazioni, $\theta = \{\mu, \sigma^2\}$

$$
\begin{aligned}
L(\theta, \{x_1, \ldots, x_n\}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x_i - \mu)^2}{2\sigma^2}\} \\
&= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\}
\end{aligned}
$$

passando ai logaritmi

$$
\log L(\theta, \{x_1, \ldots, x_n\}) = -\frac{n}{2} \log \pi\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
$$

Valutazione dei modelli

- L'AIC è costruito usando la log-verosimiglianza ovvero, se ho p parametri nel modello che sto stimando il valore di AIC del modello è calcolato come: $2p - 2 \log L$.
- Se sto confrontando più modelli dello stesso tipo, ad esempio regressioni logistiche con diverse variabili indipendenti, allora sceglierò il modello che mi da il valore di AIC più piccolo. Questo sarà il più *verosimile*.
- Attenzione l'AIC non mi dice quanto bene il modello si adatta ai dati. Indicazioni al riguardo le ottengo dalla percentuale di devianza spiegata. Tanto più questa è alta tanto migliore è il modello.

Esempio: Regressione logistica

File dati= exlogit.csv
File di script di R = esempiologit.R