Contents

List of Figures

(14 maggio 2020, versione molto provvisoria, considerate questo materiale un promemoria per alcune formule e dimostrazioni e un supporto alle lezioni, che quest'anno peraltro sono in buona parte registrate e restano in piattaforma)

Consultare il materiale dei notebook, degli esercizi e laboratorio

Per tenere aggiornato il software didattico del mio package MLANP, potete installarlo dal mio repository github. In Rstudio installate devtools e poi

devtools::install_github("marcellochiodi/packages")

1 Generalizzazioni del modello lineare (Regressione logistica, di Poisson e Generalized linear models)

1.1 Un approccio un po' antico: trasformazioni di variabili; riduzione dell'eteroscedasticità

Scopo:

- stabilizzazione della varianza
- o comunque riduzione dell'eteroscedasticità

Dall'analisi dei residui o dalla rappresentazione grafica delle varianze in funzione delle medie per dati raggruppati ci si potrebbe rendere conto che la varianza appare funzione della media:

Si veda per esempio la figura 1, ottenuta dai residui di un modello lineare adattato per descrivere i tempi di guarigione in funzione di alcuni fattori prognostici e diagnostici.

L'andamento dei residui in ordinata sembra variare di più al crescere dei valori stimati. Anche il grafico fornito da R (nella figura 2) sembra confermare tale indicazione.

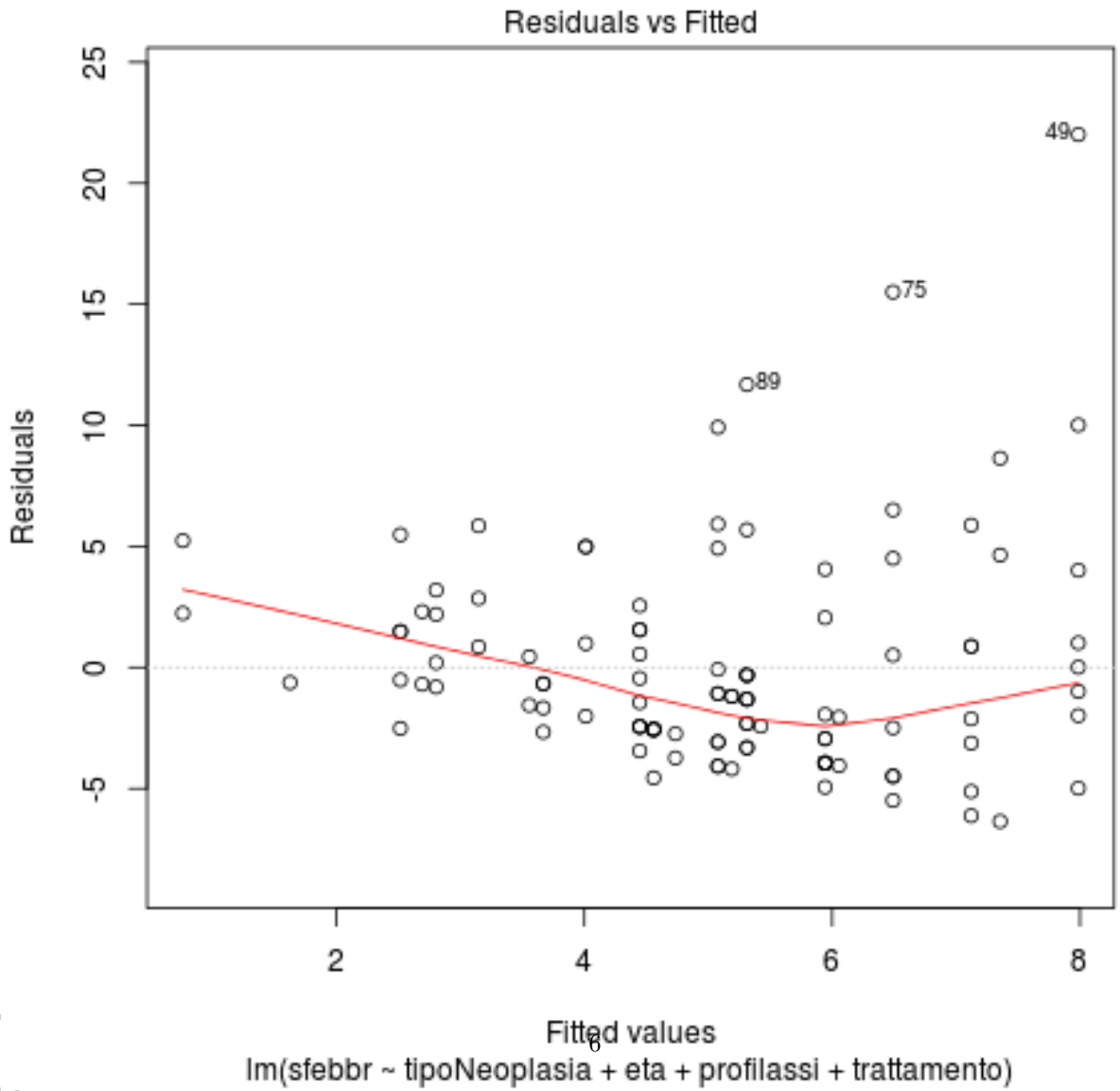

Figure 1: Un grafico di residui in funzione di valori stimati che mostra una possibile eteroscedasticità

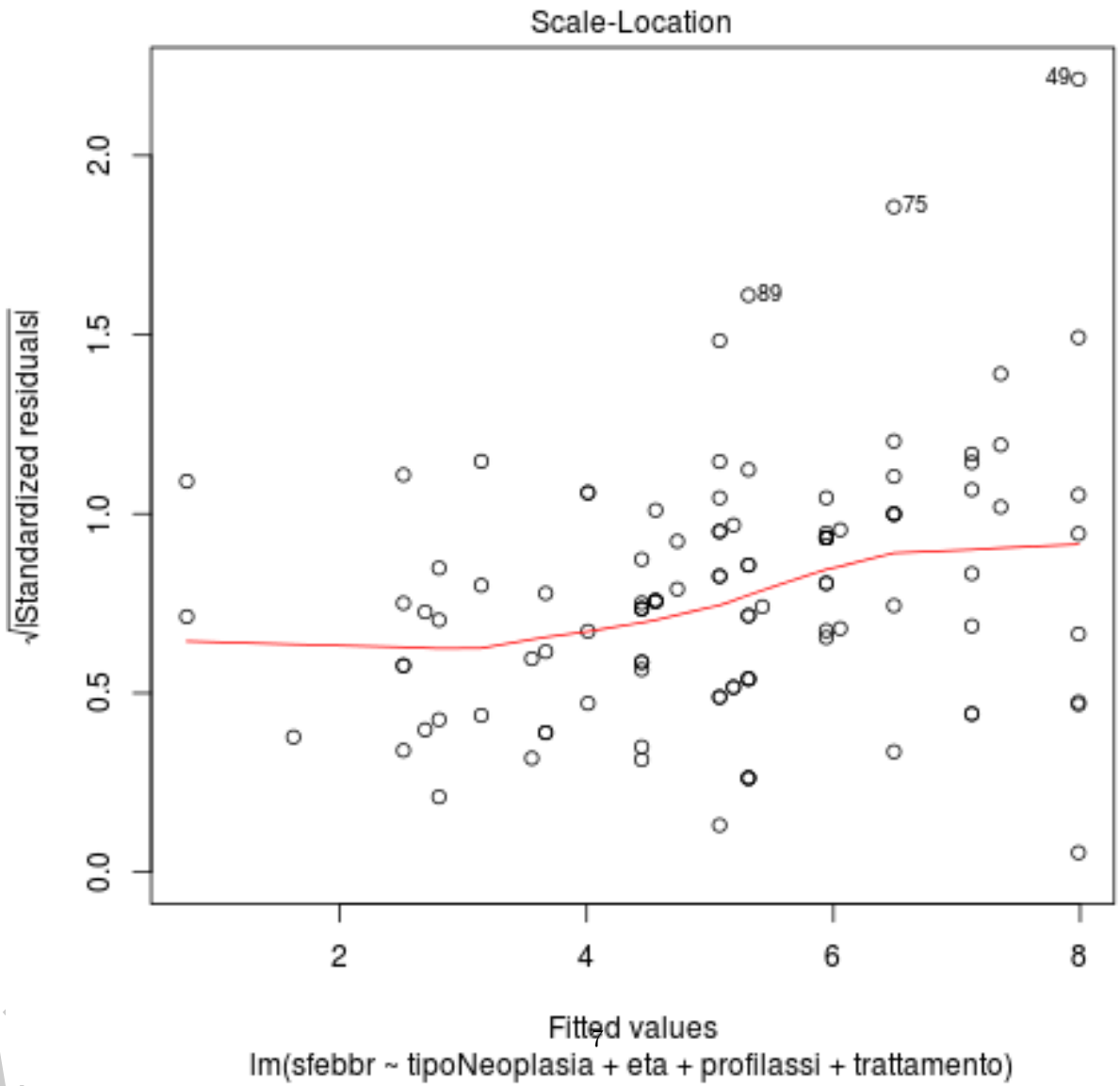

Figure 2: Un grafico location scale che mostra una possibile eteroscedasticità

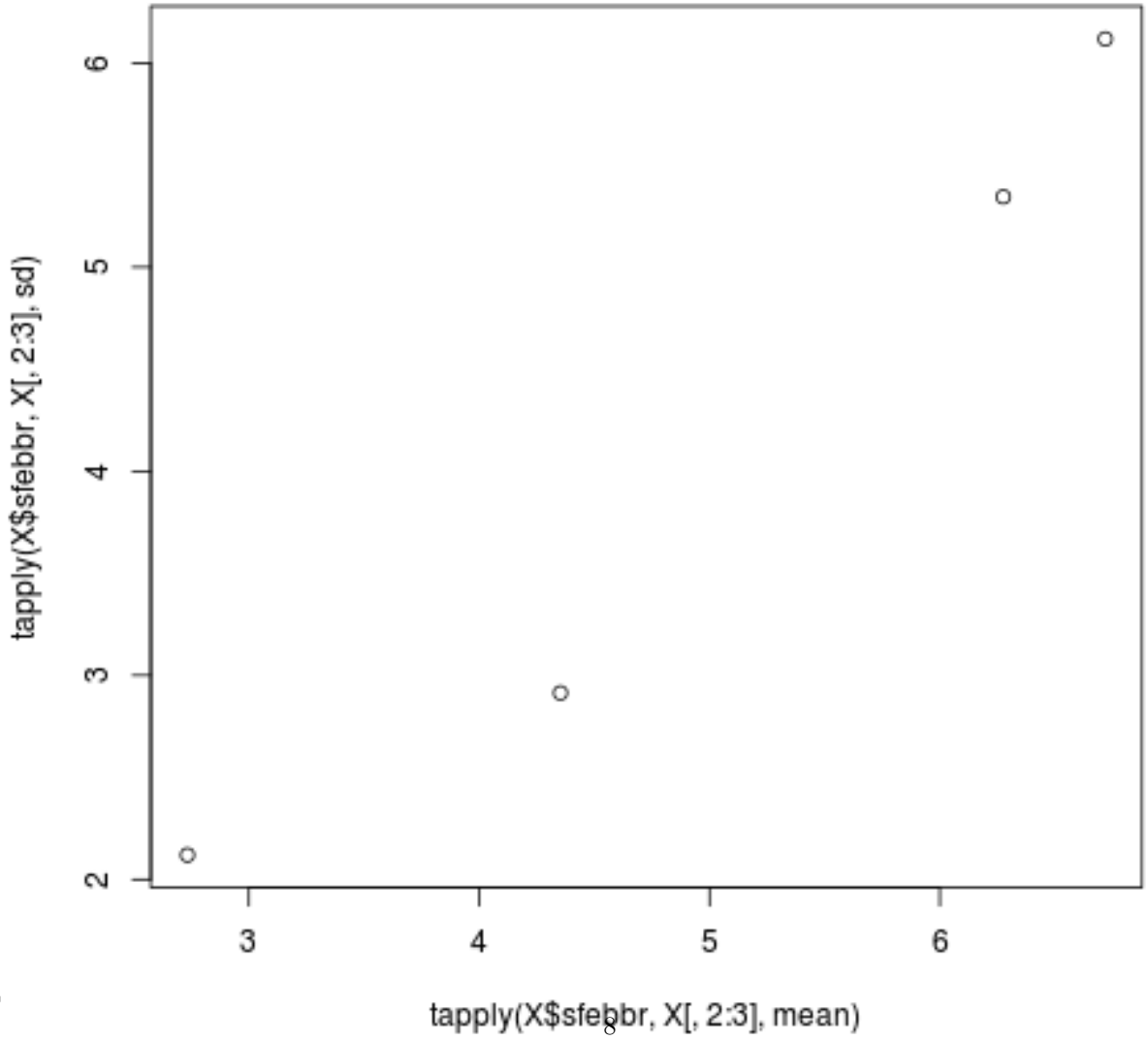

Figure 3: Un grafico deviazione standard vs medie calcolate in 4 gruppi

Nella figura 3 sono stati invece rappresentati le deviazioni standard in funzione delle medie per la stessa variabile in 4 diversi gruppi di pazienti (esempio visto a lezione)

In pratica stiamo vedendo che ciò potrebbe essere un indizio del fatto che da:

$$\mathrm{V}[Y_i] = g(\mathrm{E}[Y_i]),$$

mentre usualmente assumiamo $\mathrm{V}[Y_i] = \sigma^2$

(p.e. $\mathrm{V}[Y_i] = \sigma^2(\mathrm{E}[Y_i])^\theta$)

si può in prima battuta cercare una trasformazione $Z_i = f(Y_i)$ che renda le nuove variabili Z *approssimativamente* con varianza uguale.

Sappiamo che se

$$\mathrm{E}\left[Y_i\right] = \eta_i \qquad \text{e} \qquad \mathrm{V}\left[Y_i\right] = g(\eta_i),$$

possiamo approssimare i primi due momenti di $Z_i = f(Y_i)$ mediante le note relazioni (ricavate sviluppando in serie in η_i e troncando al primo termine):

$$\begin{aligned} \mathrm{E}\left[f(Y_i)\right] &\approx f(\eta_i); \\ \mathrm{V}\left[f(Y_i)\right] &\approx \left(f'(y_i)_{y_i=\eta_i}\right)^2 \mathrm{V}\left[Y_i\right] = \\ &= \left(f'(y_i)_{y_i=\eta_i}\right)^2 g(\eta_i) \end{aligned}$$

per cui occorre *(se la funzione $g()$ è nota!)* trovare una $f(Y_i)$ tale che:

$$\mathrm{V}\left[f(Y_i)\right] \approx k^2 \qquad \text{(costante)}$$

e quindi:

$$f'(y_i)_{y_i=\eta_i} = \frac{k}{\sqrt{g(\eta_i)}} \Longrightarrow f(y_i) = \int \frac{k}{\sqrt{g(y)}}\, d\, y$$

1.2 Trasformazioni classiche

1. Distribuzione di Poisson:

$$\mathrm{V}\left[f(Y_i)\right] = \mathrm{E}\left[f(Y_i)\right] = \eta_i$$

   per cui :

$$f(y_i) = \sqrt{y_i}$$

2. Distribuzione di una frequenza relativa di successo:

$$\mathrm{E}\left[f(Y_i)\right] = p_i$$

$$\mathrm{V}\left[f(Y_i)\right] = \frac{p_i(1-p_i)}{n_i}$$

   per cui:

$$f(y_i) = \sqrt{n_i} arcsin(\sqrt{Y_i})$$

3. Dispersione proporzionale alla media:

$$\sqrt{\mathrm{V}\left[f(Y_i)\right]} = k\mathrm{E}\left[f(Y_i)\right]$$

   per cui:

$$f(y_i) = \log Y_i$$

2 Regressione logistica

2.1 Risposte singole

Supponiamo di avere una situazione osservata diversa da quelle affrontate nei capitoli precedenti, con una matrice $\mathbf{X}$ di regressori analoga, ma con una variabile di risposta qualitativa, in particolare dicotoma, diciamo senza perdita di generalità di tipo 0-1. Potrebbe indicare insuccesso-successo, guarigione, ingresso nel mondo del lavoro, successo di un evento sportivo, etc.

Fin qui abbiamo sempre considerato come variabile di risposta una variabile quantitativa: abbiamo cercato di capire se e come il valore atteso di tale variabile dipende da altre variabili osservabili qualitative o quantitative, possibilmente secondo una relazione lineare.

Adesso ci poniamo un problema leggermente diverso e per semplificare le cose supponiamo di avere come variabile esplicativa un solo regressore quantitativo X, e come variabile di risposta una variabile Y che può assumere solo i valori 0 e 1. Ad esempio Y potrebbe indicare la condizione di guarigione da una certa patologia e X potrebbe essere la dose somministrata. Oppure Y potrebbe indicare

la condizione lavorativa (0 per i disoccupati e 1 per gli occupati) e X il numero di anni di studio, o il fatto che una squadra abbia vinto o meno un confronto sportivo in funzione della quotazione fornita dagli allibratori.

Nella figura 4 è mostrato un esempio di questa situazione: *l'evento* $Y = 1$ *è più frequente per valori di* X *alti, mentre per valori di* X *bassi l'evento positivo si presenta rararmente.*

Vi è anche una zona intermedia di valori di X in cui sembra essere maggiore l'incertezza rispetto al verificarsi o meno dell'evento $Y = 1$.

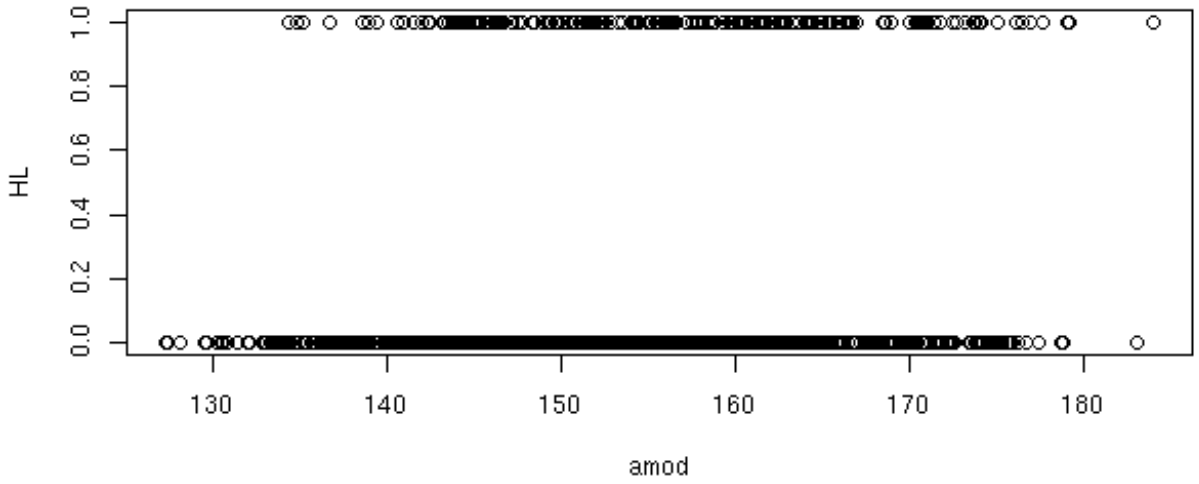

Figure 4: una variabile binaria in funzione di una var. quantitativa

Nella figura 5 è mostrato un esempio in cui la frequenza dell'evento $Y = 1$ *non sembra dipendere in modo marcato dal valore assunto da* X

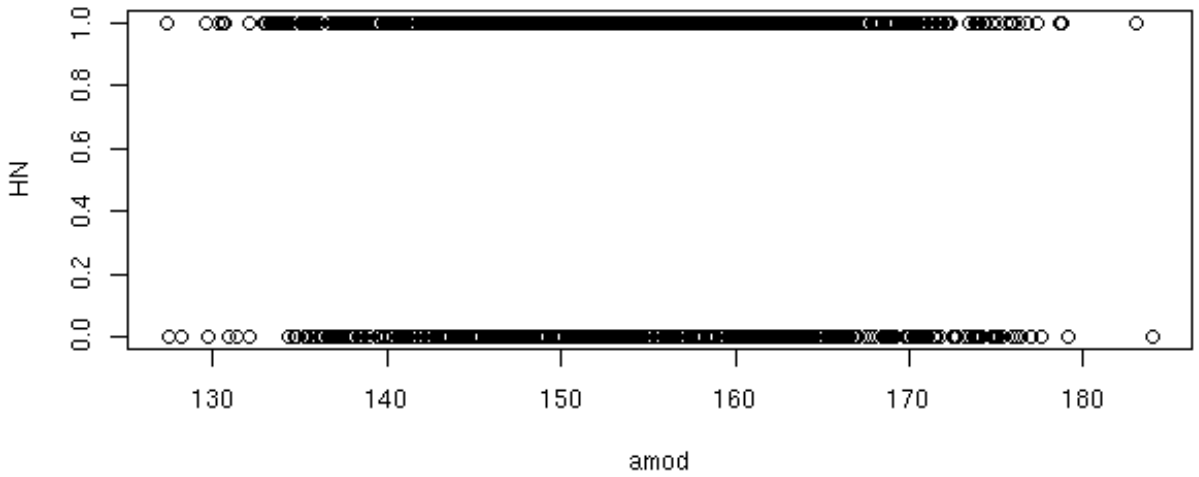

Figure 5: una variabile binaria in funzione di una var. quantitativa

2.2 Esempi visti in aula

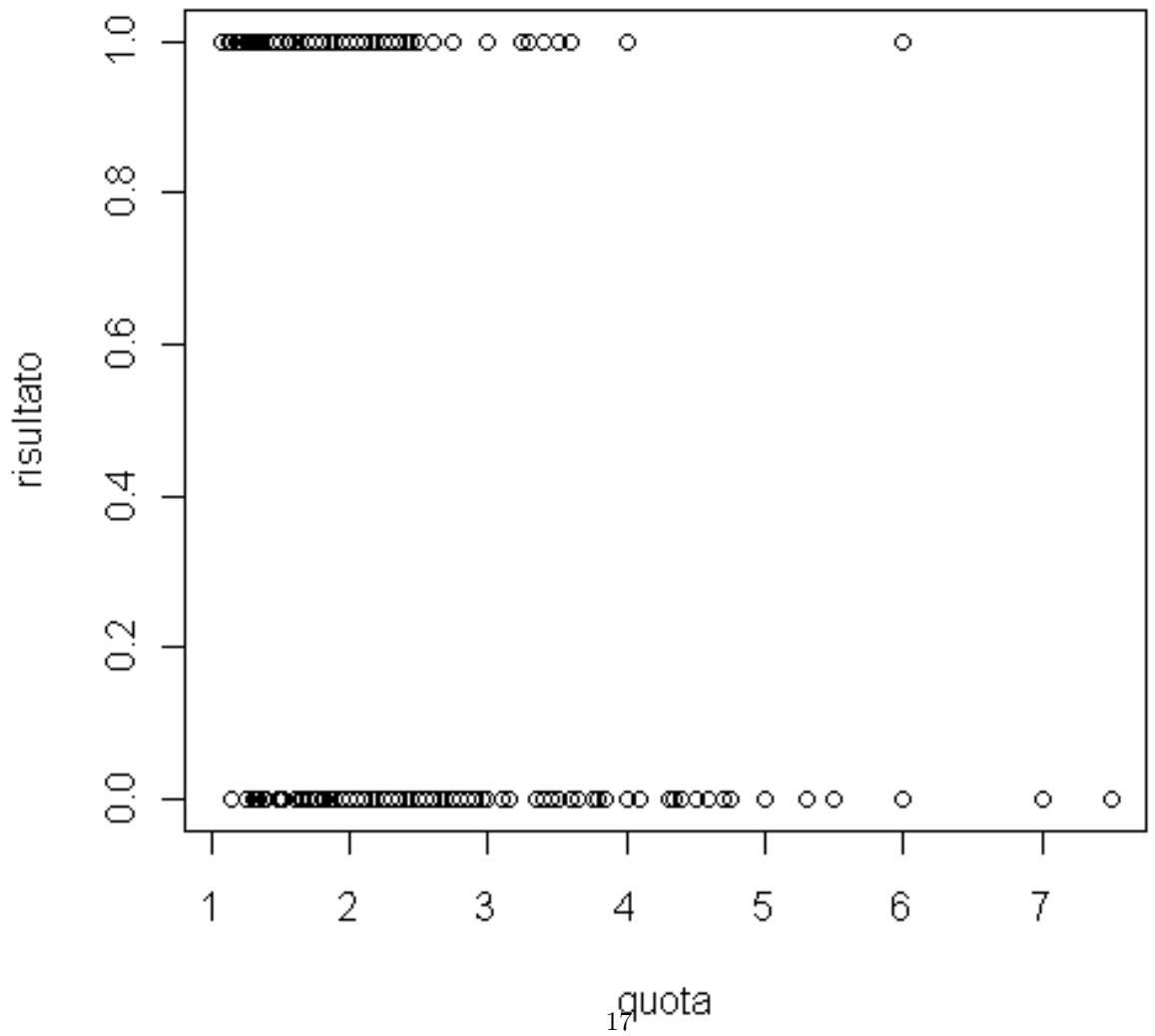

Figure 6: Risultati deille partite di un campionato di calcio in funzione della quota data dagli allibratori per la vittoria della squadra di casa; $Y = 1$ indica la vittoria della squadra di casa, $Y = 0$ pareggio o sconfitta

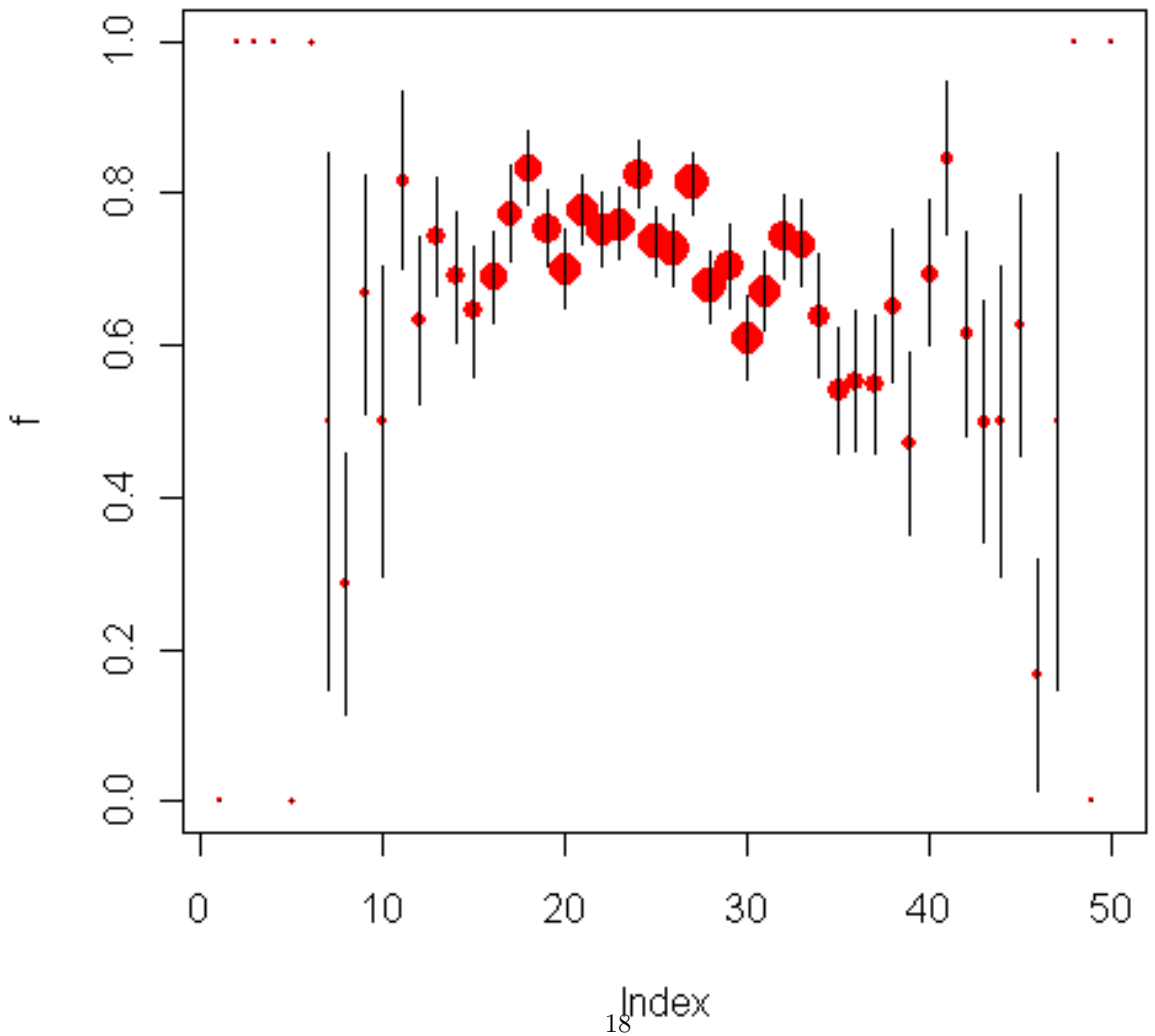

Figure 7: Dati raggruppati: frequenza di immatricolazione alla nostra facoltà nel 2009-2010 in funzione del risultato nel test d'ingresso

Ma potremmo ovviamente essere interessati ad analizzare la dipendenza dell'esito di un esperimento (indicata dalla variabile `esito`) da una **molteplicità di fattori** , quantitativi e non.

```
> str(trial1)

'data.frame':   110 obs. of  11 variables:
 $ id            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ codiceCentro  : int  1501 1501 1501 603 1501 1308 1501 501 501 1308 ...
 $ sesso         : Factor w/ 2 levels "f","m": 2 1 2 2 1 2 1 1 1 1 ...
 $ livelloRischio: Factor w/ 2 levels "alto","basso": 2 2 1 2 2 2 2 1 1 2 ...
 $ tipoNeoplasia : int  0 1 1 0 0 1 1 1 1 0 ...
 $ eta           : int  0 1 0 1 1 0 1 1 1 0 ...
 $ profilassi    : int  1 0 1 0 1 1 1 0 1 1 ...
 $ faseMalattia.1: int  0 0 0 2 2 0 0 0 0 0 ...
 $ trattamento   : Factor w/ 3 levels "A","B","C": 2 3 2 1 3 1 2 3 2 1 ...
 $ sfebbr        : int  2 6 7 3 5 2 5 12 5 3 ...
 $ esito         : int  1 1 1 1 1 1 0 1 1 1 ...
```

Ipotizzata l'indipendenza fra le osservazioni, possiamo supporre che Y, essendo un indicatore di evento, abbia una distribuzione bernoulliana.

Se non esistesse una dipendenza da X avremmo semplicemente:

$$\text{Prob}\{Y = 1\} = p.$$

Adesso vogliamo vedere se tale probabilità (che per la variabile Y è anche valore atteso) dipende in qualche modo da X. Possiamo ipotizzare allora che Y, indicatore di evento, condizionatamente a ciascun valore x di X, abbia una distribuzione bernoulliana con parametro funzione di x [1] :

$$\text{Prob}\{Y = 1|X = x\} = p(x)$$

[1] Va sottolineato che sebbene stia usando il termine *condizionatamente*, non si sta ipotizzando che X sia una variabile aleatoria, e quindi non ocorre neanche fare delle ipotesi sulla distribuzione congiunta di X, Y

Ci proponiamo di studiare la dipendenza da X di tale probabilità sulla base di un campione di osservazioni (indipendenti). Prima però vediamo di inquadrare il problema in un ambito familiare, magari collegabile poi ai modelli lineari.

Nella distribuzione di Bernoulli sappiamo che il valore atteso coincide con la probabilità del verificarsi dell'evento, per cui per la funzione di regressione si ha:

$$\mathrm{E}\left[Y|X=x\right] = p(x)$$

Dunque ci stiamo ponendo sempre la stessa domanda:

Il valore atteso di Y dipende da x?

Faccio subito notare che con le assunzioni che abbiamo fatto il modello è intrinsecamente eteroscedastico.

Infatti sappiamo che per tale variabile bernoulliana $\mathrm{V}[Y|X = x] = p(x)(1 - p(x))$, che risulta masima quando $p(x) = \frac{1}{2}$; d'altra parte nel grafico mostrato prima per valori di X intermedi avevamo una maggiore incertezza, mentre agli estremi della distribuzione di X sembra esserci meno incertezza, e quindi meno variabilità, sulla determinazione che assume Y.

Tornando comunque alla dipendenza di $\mathrm{E}[(]\, Y|X = x)$ da X, che forma può assumere questa dipendenza? Teniamo presente che $0 \leq p(x) \leq 1$, quindi non potremo assumere una semplice funzione lineare in x (perchè non potrebbe mai essere limitata, se non nel caso di una funzione costante)!

Come spesso accade, conviene ragionare prima sui casi estremi: *assenza di dipendenza in media e massima dipendenza.*

L'assenza di dipendenza in media è facilmente rappresentata da una relazione costante in x:

$p(x) = p_0$ (figura 8).

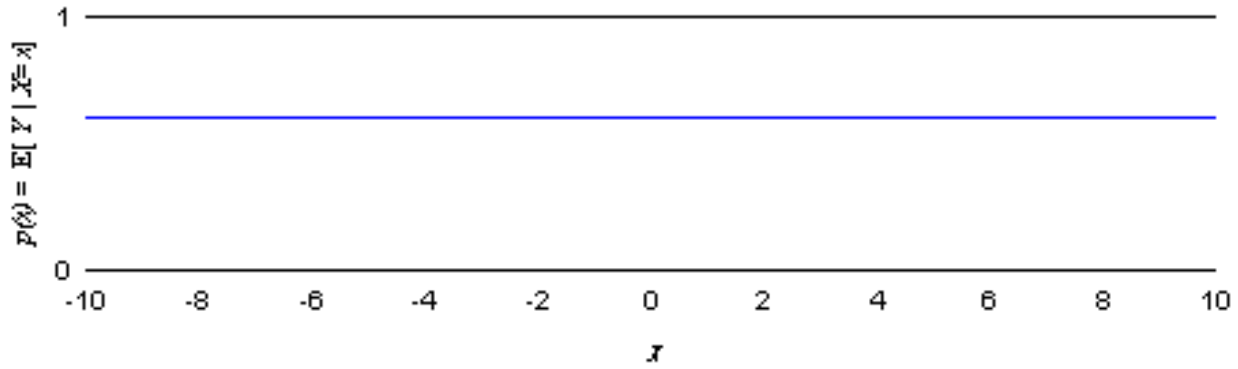

Figure 8: Probabilità costante al variare di x

Per la massima dipendenza possiamo pensare che esista un valore di soglia x_0 tale che:

$$p(x) = \begin{cases} 0 \;\; se \;\; x < x_0 \\ 1 \;\; se \;\; x \geq x_0 \end{cases};$$

Possiamo esprimere graficamente la situazione [2] (figura 9):

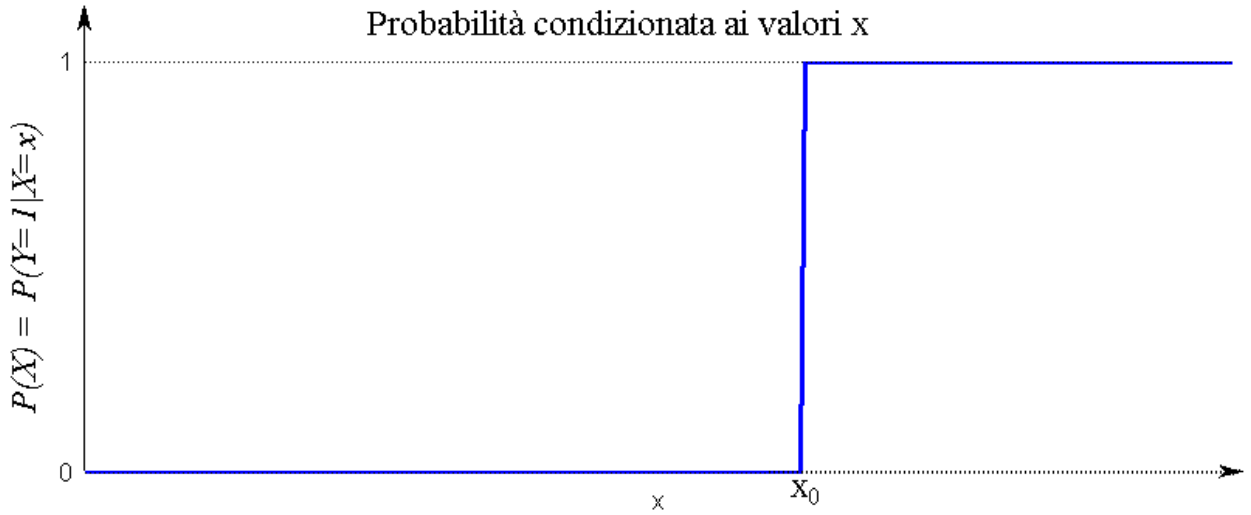

Figure 9: Massima dipendenza della probabilità da x

[2] Ovviamente non è l'unica possibile situazione di dipendenza massima, ma è l'unica in caso di dipendenza monotona. In altrenativa potremmo avere probabilità unitaria in un intervallo di valori di x e 0 fuori dall'intervallo, ma non sarebbe monotona

In pratica il verificarsi dell'evento è determinato con certezza dal valore assunto da x rispetto a x_0. Nel caso in cui si abbia

$$p(x) = \begin{cases} 1 \ \ se \ \ x < x_0 \\ 0 \ \ se \ \ x \geq x_0 \end{cases};$$

avremo allora una dipendenza massima di segno negativo.

Ovviamente nasce il problema di come descrivere le situazioni di dipendenza intermedie fra questi due estremi; proviamo a procedere per gradi successivi di approssimazione.

In prima approssimazione magari potremmo pensare una funzione che, anzichè passare da 0 ad 1 con uno scalino in x_0, (o da 1 a 0), passi da 0 ad 1 con un segmento di retta, con inclinazione tanto maggiore, quanto maggiore è la dipendenza di $\mathrm{E}[Y|X = x]$ da X, e che sia costituita quindi, complessivamente, da una spezzata come quella della figura qua sotto (figura 10):

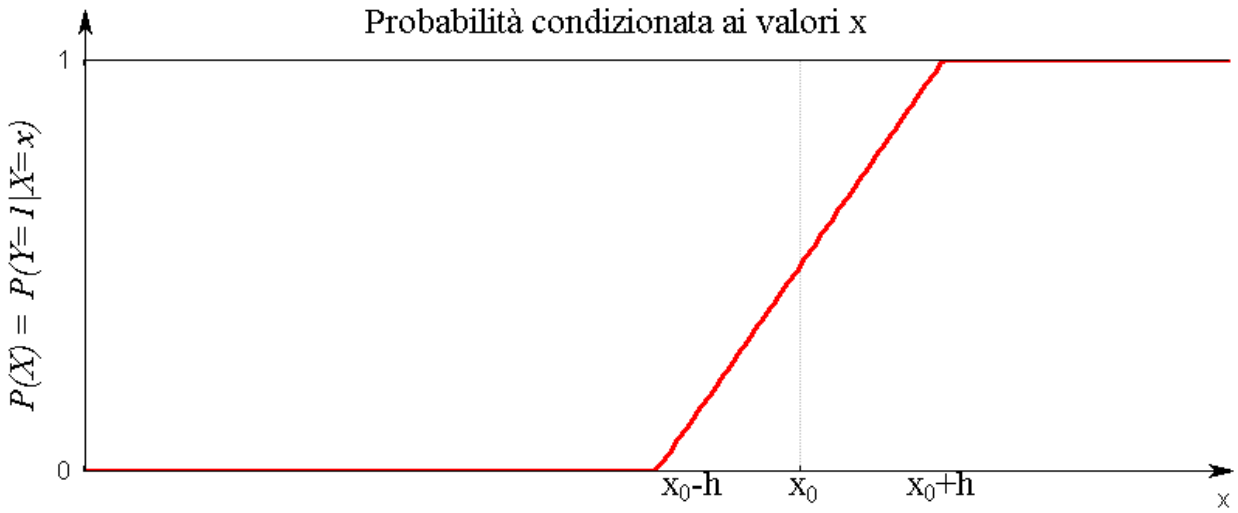

Figure 10: Probabilità funzione lineare a tratti di x

$$p(x) = \begin{cases} 0 & se \quad x < x_0 - h \\ \frac{x - x_0 + h}{2h} & se \quad x_0 - h \leq x < x_0 + h \\ 1 & se \quad x \geq x_0 + 1 \end{cases};$$

Abbiamo una probabilità nulla per valori di X inferiori a $x_0 - h$ e una probabilità unitaria per valori di X superiori a $x_0 + h$, e attorno a x_0 un segmento di retta di pendenza $\frac{1}{2h}$; una forte dipendenza di Prob $\{Y = 1 | X = x\}$ dal particolare valore di x corrisponderà ad una forte pendenza.

Una relazione di questo genere può essere sufficiente a modellizzare situazioni semplici, ma si presta poco ad estensioni a casi con più variabili esplicative, e non può essere trattata in modo analiticamente conveniente, dal momento che, per esempio, pur essendo una funzione continua, non possiede derivata continua (l'inclinazione infatti cambia in modo discontinuo nei due punti $x_0 - h$ e $x_0 + h$). Possiamo però partire da questa prima approssimazione per ricavare delle curve continue dotate di derivate continue.

Intanto stabiliamo alcune caratteristiche che dovrebbe avere $p(x)$: sembra ragionevole imporre una condizione di antisimmetria attorno ad un punto sull'asse delle x, diciamo x_0:

$$p(x_0 + \delta) = 1 - p(x_0 - \delta) \tag{1}$$

che rispecchia il fatto che ragionare sul verificarsi dell'evento $(Y = 1)$ o sul suo non verificarsi $(Y = 0)$ deve portare agli stessi risultati, dato che:

$$\text{Prob}\{Y = 1 | X = x\} = p(x) \qquad \text{e}$$

$$\text{Prob}\{Y = 0 | X = x\} = 1 - p(x)$$

Si deve dunque avere (dalla 1, per $\delta = 0$): $p(x_0) = \frac{1}{2}$. Potremmo anzi definire x_0 come il punto per il quale la probabilità di verificarsi dell'evento è $\frac{1}{2}$. In caso di associazione positiva si deve anche avere $p'(x) > 0$.

Inoltre appare naturale (ed analiticamente molto comodo), imporre delle condizioni sul comportamento di $p(x)$ al divergere di $|x|$:

$$\lim_{x \to -\infty} p(x) = 0 \qquad \text{e} \qquad \lim_{x \to +\infty} p(x) = 1$$

Attorno ad x_0 $p(x)$ dovrebbe essere approssimabile con una retta e dovrebbe esserci in x_0 un punto di flesso, in particolare per l'ipotesi di antisimmetria, con

$p''(x_0) = 0$ e quindi $p(x)$ dovrebbe avere il seguente sviluppo in serie di Taylor in x_0 al secondo ordine:

$$\begin{aligned} p(x) &= p(x_0) + p'(x_0)(x - x_0) + \\ &\quad \frac{1}{2}p''(x_0)(x - x_0)^2 + O((x - x_0)^3) = \\ &= \frac{1}{2} + \frac{1}{2h}(x - x_0) + O((x - x_0)^3) \end{aligned}$$

in quanto in x_0 si ha, per i ragionamenti fin qui fatti: $p(x_0) = \frac{1}{2}$, $p'(x_0) = \frac{1}{2h}$ e $p''(x_0) = 0$.

Per farla breve, una funzione [3] che risponde a tutte le condizioni ora viste è la seguente:

$$p(x) = \frac{\exp\left[\beta(x - x_0)\right]}{1 + \exp\left[\beta(x - x_0)\right]} \tag{2}$$

nota come *funzione logistica*.

[3] ma ovviamente non l'unica!

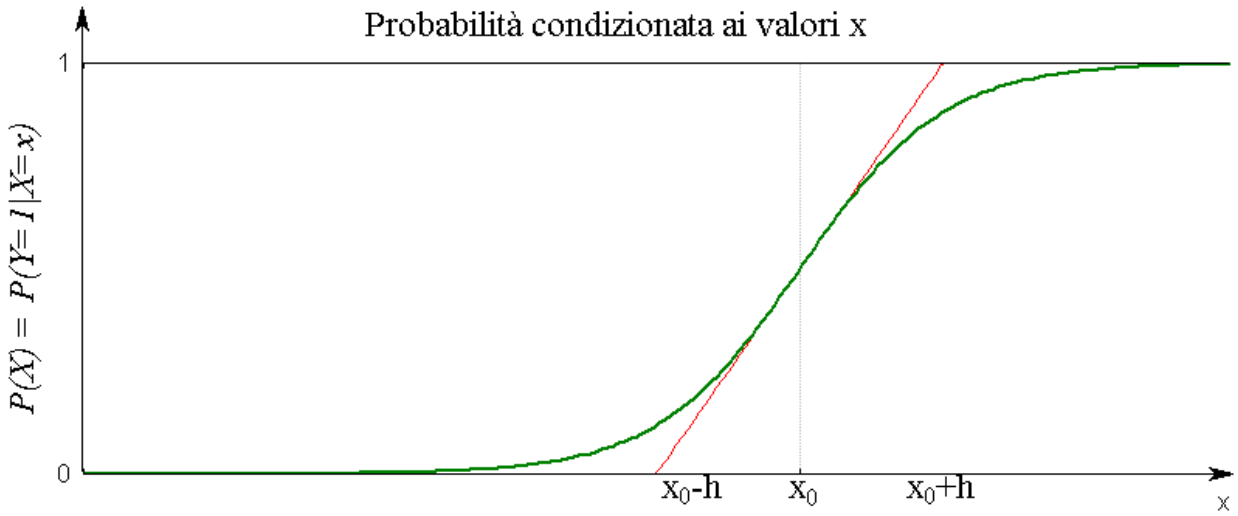

Figure 11: Funzione logistica e approssimazione lineare in x_0

Più avanti farò cenno ad altre possibili scelte di funzioni per $p(x)$; basta per ora segnalare che una qualsiasi funzione di ripartizione di una variabile aleatoria continua dotatà di densità, con distribuzione simmetrica, unimodale e di valore atteso x_0 soddisfa le nostre condizioni.

Approfondiamo lo studio della funzione logistica in (2); è facile vedere che:

$$p'(x) = \frac{\beta \exp\left[\beta(x - x_0)\right]}{(1 + \exp\left[\beta(x - x_0)\right])^2} \tag{3}$$

$$p''(x) = 0 \qquad \text{per} \qquad x = x_0;$$

$$p' x_0 = \frac{\beta}{4}$$

Intanto vediamo nella figura 12 l'influenza di un diverso valore di β sulla forza della dipendenza di $\mathrm{E}\left[Y|X = x\right]$ da x.

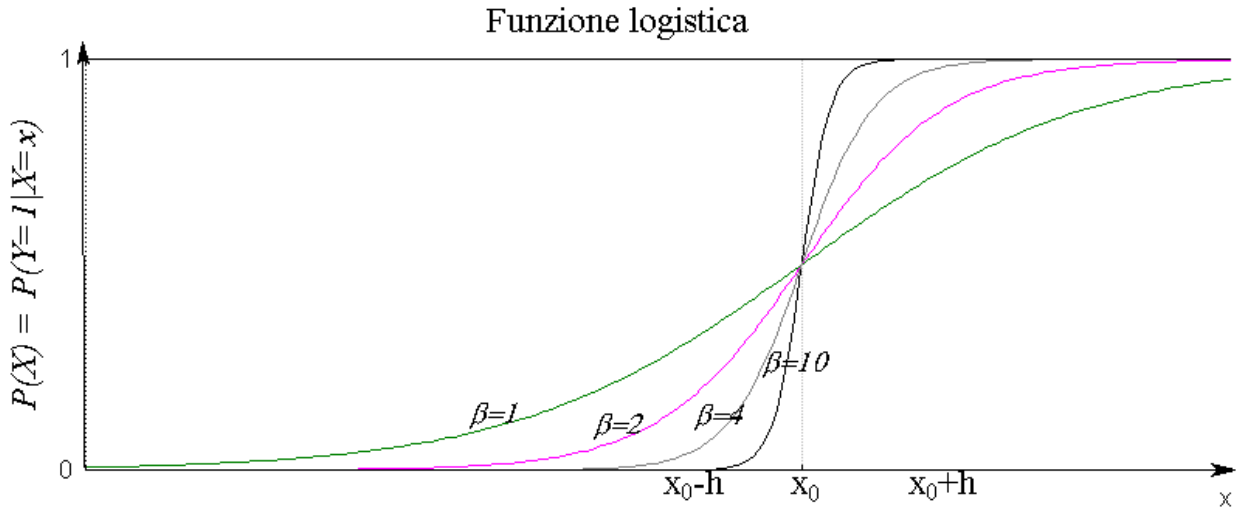

Figure 12: Funzione logistica (2) per quattro diversi valori di β: $\beta = 1; 2; 4 e 10$

Adesso cominciamo ad approfondire il significato di questa relazione e vediamo di dare un significato a β oltre quello geometrico: è facile vedere, con semplici passaggi, che dalla relazone (2) si ricava:

$$1 - p(x) = \frac{1}{1 + \exp\left[\beta(x - x_0)\right]}$$

e quindi, dividendo membro a membro la (2) e quest'ultima equazione, si ha l'importante relazione per gli *odds* (in corrispondenza di $X = x$):

$$\text{odds}(x) = \frac{p(x)}{1 - p(x)} = \exp\left[\beta(x - x_0)\right]$$

e quindi:

$$\beta(x - x_0) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \log \text{odds}(x)$$

Quindi il logaritmo degli odds è funzione lineare di x, adottando la funzione (2).

Infine se vogliamo misurare il rapporto fra gli odds per due valori x e $x + \delta$, si

ha facilmente:

$$\frac{\frac{p(x+\delta)}{1-p(x+\delta)}}{\frac{p(x)}{1-p(x)}} = \frac{\exp\left[\beta(x+\delta-x_0)\right]}{\exp\left[\beta(x-x_0)\right]} = \exp\left(\beta\delta\right)$$

che per $\delta = 1$ diventa:

$$\frac{\frac{p(x+1)}{1-p(x+1)}}{\frac{p(x)}{1-p(x)}} = e^{\beta} \quad \text{e} \quad \beta = \log\left(\frac{\frac{p(x+1)}{1-p(x+1)}}{\frac{p(x)}{1-p(x)}}\right) \tag{4}$$

Quindi beta è uguale al logaritmo del rapporto degli *odds* (*odds ratio*, indicato spesso con OR) in corrispondenza di un incremento unitario di x, qualsiasi sia il valore di x.[4]

$$\beta = \log \mathrm{OR} \tag{5}$$

3 Estensioni della regressione logistica semplice

Possiamo anche parametrizzare questa curva in modo leggermente diverso:

$$\beta(x - x_0) = \alpha + \beta x, \quad \text{con} \quad \alpha = -\beta x_0$$

[4] Si pensi al caso in cui X è una variabile che può assumere i valori 0/1, ossia è una variabile dummy associata ad una variabile qualitativa dicotoma. β è il logaritmo dell'odds ratio della tavola 2x2

Più in generale possiamo esprimere il legame fra $\mathrm{E}[Y|X=x]$, ossia $p(x)$, e x in modo generale come:

$$\log\left(\frac{\mathrm{E}[Y|X=x]}{1-\mathrm{E}[Y|X=x]}\right)=\eta(x) \qquad (=\alpha+\beta x)$$

Chiamiamo:

- $\eta(x)$ il *predittore lineare*
- e la funzione $g(\mathrm{E}[Y|X=x])$ la *funzione legame* (nel nostro caso $g(p(x))=\log\left(\frac{p(x)}{1-p(x)}\right)$.

La nostra relazione di dipendenza è allora formalizzabile molto schematicamente con:

$$g(\mathrm{E}[Y|X=x])=\eta(x)$$

Riprenderemo più avanti questa terminologia nel contesto dei modelli lineari generalizzati.(GLM)[5]

Adesso la cosa importante è notare che in effetti abbiamo ancora, come nel modello lineare generale fin qui trattato, un valore atteso condizionato della variabile di

[5] si noti che se $g(\cdot)$ fosse banalmente la funzione identità, avremmo la relazione del modello lineare

risposta, ossia $\mathrm{E}[Y|X = x]$, espresso come funzione di una trasformazione lineare della variabile esplicativa x, ossia $\eta(x)$; in modo molto generale:

$$\mathrm{E}[Y|X = x] = h(\cdot);$$

ove con $h(\cdot)$ ho indicato la funzione inversa di $g(\cdot)$.
$(h(\cdot) = g^{-1}(\cdot))$.

E' una relazione analoga a quella adottata nei modelli lineari generali ma con due differenze sostanziali:

- La distribuzione di Y non è normale (ma bernoulliana nel caso che stiamo esaminando)
- Il legame con la trasformata lineare non è diretto, ma avviene attraverso una funzione $h(\cdot)$, la cui inversa $g(\cdot)$ abbiamo chiamato *funzione legame*.

Non ci preoccupiamo invece del fatto che nel modello lineare generale abbiamo trattato con k variabili esplicative: anche nel modello di regressione logistica che stiamo generalizzando possiamo pensare ad una dipendenza da una funzione lineare di più regressori; indichiamo con $\mathbf{x}^{\mathsf{T}}$ il vettore di variabili esplicative a k

componenti e così il predittore lineare sarà semplicemente:

$$\eta(\mathbf{x}) = \mathbf{x}^\mathsf{T}\boldsymbol{\beta}$$

e avendo indicato al solito con $\boldsymbol{\beta}$ un vettore di parametri a k componenti.

Indicando con $\mathbf{X}$ la variabile a k componenti, costituita dalle k variabili esplicative X_j, possiamo modellizzare nel modo che segue la dipendenza del valore atteso della variabile bernoulliana Y condizionatamente ai valori assunti dal vettore di variabili esplicative $\mathbf{X}$:

$$\log\left(\frac{\mathrm{E}\left[Y|\mathbf{X}=\mathbf{x}\right]}{1-\mathrm{E}\left[Y|\mathbf{X}=\mathbf{x}\right]}\right) = \mathbf{x}^\mathsf{T}\boldsymbol{\beta} \quad (= \eta(\mathbf{x}))$$

A questo punto è semplice pensare di poter modellizzare la dipendenza del valore atteso di Y da più variabili esplicative sia di tipo quantitativo che di tipo qualitativo, in funzione delle particolari configurazioni di valori osservati $\mathbf{x}_i^\mathsf{T}$.

L'interpretazione dei singoli parametri β_j è nuovamente quella di logaritmo dell'odds ratio relativo ad un incremento unitario della variabile X_j.

Possiamo qundi impostare modelli più o meno complessi, in funzione delle variabili che riteniamo possano influenzare la variabile di risposta (che adesso è una

probabilità del verificarsi di un evento); potremo inserire delle componenti polinomiali, delle interazioni, potremo lavorare con variabili esplicative qualitative, quantitative o miste.

L'interpretazione dei risultati e del contributo delle singole variabili esplicative sarà del tutto simile a quella vista per i modelli lineari.

4 Verosimiglianza ed inferenza nella regressione logistica

Si parte dalla verosimiglianza per una osservazione da una variabile bernoulliana:

$$L(p; y) = p^y \ (1-p)^{(1-y)}$$

per cui, con n osservazioni indipendenti y_i, con:

$$p_i = \mathrm{E}\left[Y_i\right] = \frac{\exp\left\{\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}\right\}}{1+\exp\left\{\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}\right\}}$$

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n}\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right] =$$

$$= \sum_{i=1}^{n} y_i \log\frac{p_i}{1-p_i} + \sum_{i=1}^{n}\log(1-p_i) =$$

$$= \sum_{i=1}^{n} y_i \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} - \sum_{i=1}^{n}\log(1+\exp\left\{\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}\right\}) =$$

Decremento di verosimiglianza=
=incremento di devianza

`Il confronto fra modelli verrà ripreso con i GLM`

---

5 Conteggi: Regressione di Poisson

Analogamente a quanto visto per la regressione logistica, possiamo chiederci se il valore atteso di una variabile Y espressa da un conteggio dipenda da una variabile esplicativa x (o in generale, da un predittore lineare $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$). La nostra variabile di risposta è un conteggio, per la cui distribuzione viene spesso utilizzata la distribuzione di Poisson, di parametro μ (nella distribuzione di Poisson si ha: $\mathrm{E}[Y] = \mathrm{V}[Y] = \mu$)

Esempi:

- Numero di pazienti affetti da una malattia in una certa località
- Numero di incidenti in funzione dell'orario e del giorno della settimana
- modellizzazione delle frequenze osservate in una tavola a più entrate
- numero di gol segnati da una squadra in funzione della quota proposta

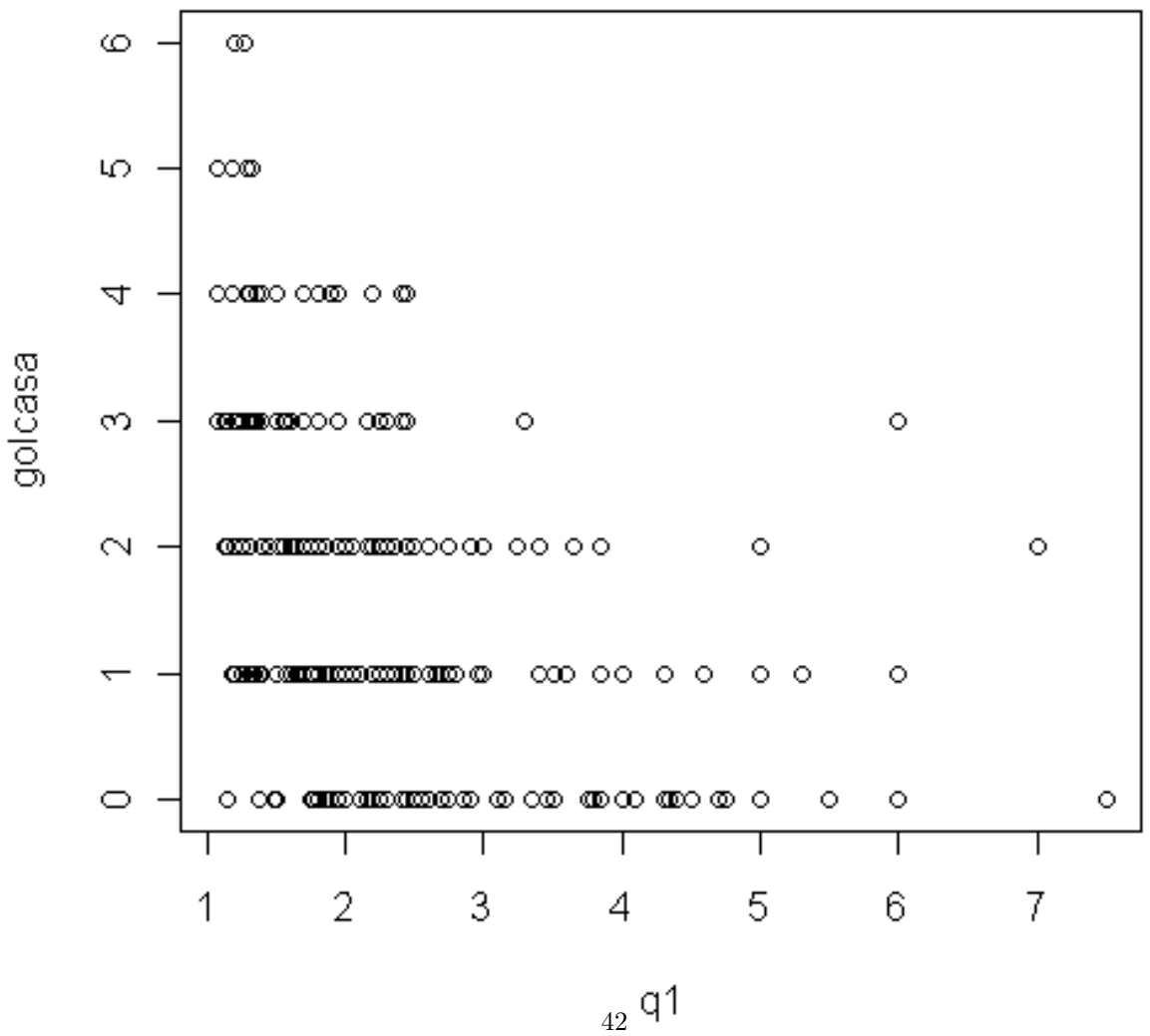

Figure 13: Risultati deille partite di un campionato di calcio: gol segnati dalla squadra di casa in funzione della quota data dagli allibratori per la vittoria della squadra di casa

Possiamo suppore che il valore atteso μ_i dipenda dalla variabile esplicativa x (o in generale, da un predittore lineare $\eta_i = \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}$) in modo simile a come abbiamo visto per la regressione logistica.

Per la ricerca della funzione legame possiamo partire dal legame della binomiale:

$$p_i = \frac{\mu_i}{n_i}$$

Se Y_i è distribuito secondo una Poisson, allora p_i sarà piccolo e possiamo approssimare:

$$\eta_i = \log \frac{p_i}{1-p_i} \approx \log p_i (= \log \mu_i - \log n_i)$$

per cui possiamo ottenere ancora una relazione fra *valore atteso e predittore lineare*:

$$\log \mu_i = \log n_i + \eta_i$$

($\log n_i$ è chiamato *offset*; se è costante andrà inglobato entro η_i)

In definitiva una soluzione soddisfacente è un legame logaritmico che porta a :

$$\mathrm{E}\left[Y_i\right] = \mu_i = \exp(\eta_i) = \exp\left\{\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}\right\}$$

Il significato dei singoli parametri β_j parametri è immediato considerando che:

$$\frac{\partial \mathrm{E}\left[Y_i\right]}{\partial x_{ij}} = \frac{\partial \exp\left\{\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right\}}{\partial x_{ij}} =$$

$$= \beta_j \ \exp\left\{\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right\}$$

(Il contributo del singolo regressore alla variazione del valore atteso dipende comunque anche dagli altri regressori)

Un interpretazione particolare dei parametri si ha per variabili esplicative dicotome:

$$\frac{\mathrm{E}\left[Y_i|x_{ij}=1\right]}{\mathrm{E}\left[Y_i|x_{ij}=0\right]} = \frac{\exp\left\{\mathbf{x}_{i(-j)}^{\mathsf{T}}\boldsymbol{\beta}_{(-j)} + \beta_j\right\}}{\exp\left\{\mathbf{x}_{i(-j)}^{\mathsf{T}}\boldsymbol{\beta}_{(-j)}\right\}} = \exp\left\{\beta_j\right\}$$

Dal momento che:

$$Prob(Y = y|\mathbf{x} = \mathbf{x}_i) \;=$$

$$Prob(Y_i = y_i) \;=\; \frac{\mu_i^{y_i}\, e^{-\mu_i}}{y_i!}$$

con $\mu_i = \exp\{\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}\}$.

La log-verosimiglianza è data da:

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \{y_i \log \mu_i - \mu_i\}$$

rispetto al modello saturo (devianza):

$$-2(l(\boldsymbol{\beta}) - l_{max}) \;=\; 2 \sum_{i=1}^{n} \left\{ y_i \left[\log \frac{y_i}{\mu_i} \right] + (\mu_i - y_i) \right\}$$

Tornando alla verosimiglianza:

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} \left\{ y_i \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} - \exp\left\{\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}\right\} \right\}$$

$$\frac{\partial l(\boldsymbol{\beta};\mathbf{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \left\{ y_i - \exp\left\{ \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} \right\} \right\}$$

Le condizioni del primo ordine di annullamento delle derivate rispetto a $\boldsymbol{\beta}$ portano ad un sistema di equazioni non lineari in $\boldsymbol{\beta}$ che andrà risolto con metodi numerici.

Converrà affrontare il problema (come pure quello della regressione logistica) direttamente nell'ambito dei modelli lineari generalizzati (GLM)

7 Modelli lineari generalizzati

(questi appunti sono in forma molto sommaria
e sono strettamente legati alle parti esposte
in aula verbalmente, alla lavagna,
e agli esempi in R)

7.1 La famiglia esponenziale

$$\log(f_Y(y;\theta\,\phi)) = \frac{y\,\theta - b(\theta)}{a(\phi)} + c(y,\phi) \tag{6}$$

ϕ è un parametro di disturbo. Se non è noto la (6) potrebbe non essere una famiglia esponenziale

Se ϕ è noto: la (6) è in **forma canonica**
e θ è il **parametro naturale**

esempio alla lavagna sulla binomiale e la poisson esempio sulla distr. g

E' facile vedere che la distribuzione di Bernoulli e la distribuzione di Poisson fanno parte della famiglia esponenziale di distribuzioni:

Bernoulli:

- $a(\cdot) = 1$
- $b(\theta) = \log(1 + \exp(\theta))$
- $\mu = \mathrm{E}[Y] = b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}$
- $b''(\theta) = \mu(1 - \mu)$ variance function

(Per la binomiale di parametro n occorre introdurre un termine $a = \frac{1}{n}$)

Poisson:

- $a(\cdot) = 1$
- $b(\theta) = \exp(\theta)$
- $\mu = \mathrm{E}[Y] = b'(\theta) = \exp(\theta)$
- $b''(\theta) = \mu$ variance function

Altre: Normale, Gamma, etc.

Per la distribuzione Gamma di parametri λ, α, e di densità:

$$f(y) = \frac{\lambda^{\alpha}\ y^{\alpha-1}\ e^{-\lambda y}}{\Gamma(\alpha)}$$

converrà porre:

$$\lambda = \alpha\theta \quad \Rightarrow \quad \theta = \frac{\lambda}{\alpha}\left(= \frac{1}{\mathrm{E}\,[Y]}\right),$$

in modo da avere (riparametrizzando rispetto a θ, α):

$$f(y) = \frac{(\alpha\theta)^{\alpha}\ y^{\alpha-1}\ e^{-\alpha\theta y}}{\Gamma(\alpha)}$$

e quindi

$$\log f(y) = \alpha \log \alpha\theta - \alpha\theta y + (\alpha - 1)\log y - \log\Gamma(\alpha) =$$

$$= \alpha \log \theta - \alpha\theta y + \alpha \log \alpha + (\alpha - 1)\log y - \log\Gamma(\alpha)$$

e raccogliendo in $c(y, \alpha)$ tutti i termini che non contengono θ, e mettendo poi in

evidenza il termine che contiene α si ha:

$$\log f(y) = \frac{\theta y - \log \theta}{-\frac{1}{\alpha}} + c(y, \alpha)$$

per cui finalmente vediamo che anche la distribuzione Gamma fa parte della famiglia esponenziale con:

- $\phi = \alpha$ e $a(\alpha) = -\frac{1}{\alpha}$
- $b(\theta) = \log \theta)$
- $\mathrm{E}[Y] = b'(\theta) = \frac{1}{\theta}$
- $\mathrm{V}[Y] = b''(\theta)a(\alpha) = \frac{1}{\alpha\theta^2}$

Anche la normale rientra nella famiglia, con $\theta = \mu$, come è facile vedere, ponendo :

- $\phi = \sigma^2$ e $a(\phi) = \sigma^2$
- $b(\theta) = \frac{\theta^2}{2}$
- $\mathrm{E}[Y] = b'(\theta) = \theta$
- $\mathrm{V}[Y] = b''(\theta)a(\phi) = \sigma^2$

7.2 Valore atteso e varianza

Quando la log-densità della nostra distribuzione appartenente alla famiglia esponenziale è espressa nella forma classica vista prima (6), l'elemento essenziale è la funzione $b(\theta)$, che, insieme con l'eventuale funzione del parametro di disturbo $a(\phi)$, determina i primi due momenti della distribuzione.

Infatti dalle prime due identità di Bartlett si ha:

$$\begin{cases} \mathrm{E}\left[\frac{\partial \log f}{\partial \theta}\right] = 0 \\ \mathrm{E}\left[\frac{\partial^2 \log f}{\partial \theta^2}\right] = -\mathrm{E}\left[\left(\frac{\partial \log f}{\partial \theta}\right)^2\right] \end{cases}$$

e si ricavano i primi due momenti (derivando la 6 rispetto a θ prima una volta e poi due volte e prendendo i valori attesi):

`passaggi alla lavagna`

$$\begin{cases} \mathrm{E}\left[Y\right] = b'(\theta) \\ \mathrm{V}\left[Y\right] = b''(\theta)a(\phi) \text{ (funzione di varianza)} \end{cases} \tag{7}$$

Se indichiamo con μ_i il valore atteso dell'i-esima osservazione:

$$\mu_i = \mathrm{E}\left[Y_i\right] \text{ (ossia } = \mathrm{E}\left[Y|X = \mathbf{x}_i\right])$$

dalla 7 si ricava facilmente:

$$\mathrm{V}\left[Y_i\right] = b''(\theta_i)a(\phi) = a(\phi)\frac{\partial b'(\theta_i)}{\partial \theta_i} = a(\phi)\frac{\partial \mu_i}{\partial \theta_i}$$

e quindi anche:

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{a(\phi)}{\mathrm{V}\left[Y_i\right]}$$

8 Generalized linear models (GLM)

I cosiddetti modelli lineari generalizzati sono dei modelli di dipendenza abbastanza flessibili (di cui segnalo in questi appunti solo gli aspetti essenziali) caratterizzati da alcuni punti comuni:

- Abbiamo una variabile di risposta Y il cui valore atteso dipende da k regressori (che possono anche essere delle variabili indicatrici di eventi)
- Ciascuna Y_i segue una distribuzione appartenente alla **Famiglia esponenziale**
- Definiamo un **Predittore lineare:** $\eta_i = \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}$ combinazione lineare dei regressori attraverso dei parametri da stimare
- Il valore atteso di Y_i e il predittore lineare η_i sono collegati da una **funzione legame** (Link function) definita da: $\eta_i = g(\mu_i) = g(\mathrm{E}[Y_i])$ ($h(\cdot)$ è la funzione inversa di $g(\cdot)$, tale che: $\mu_i = h(\eta_i)$)
- Si ha un **legame canonico** se: $\eta_i = \theta_i$, essendo θ il parametro naturale della distribuzione della famiglia esponenziale di Y

- p.e.: $g(\mu_i) = \log \frac{\mu_i}{1-\mu_i}$

legami canonici:

- Normale: $\eta = \mu$
- Poisson: $\eta = \log \mu$
- Binomiale: $\eta = \log \frac{\mu}{1-\mu}$
- Gamma $\eta = \frac{1}{\mu}$

(per la distribuzione gamma spesso si usa il legame logaritmico $\eta = \log \mu$, sebbene non sia canonico, perchè in questo modo si rispetta la condizione di positività del valore atteso)

Sono soltanto dei cenni: dal punto di vista numerico per quanto riguarda la stima dei parametri utilizziamo in questo corso la funzione `glm` di `R` ed i relativi output

E' comunque importante vedere la base dei procedimenti iterativi di stima almeno per i casi più semplici (legame canonico)

Otteniamo intanto la funzione di verosimiglianza, prendendo le mosse dalla log-densità in 6, e supponendo di avere un campione di n osservazioni indipendenti y_i da una famiglia esponenziale di parametro naturale θ_i (ma noi poi dovremo stimare gli elementi di $\boldsymbol{\beta}$):

$$l(\cdot; \mathbf{y}) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi) \right]$$

Deriviamo la log-verosimiglianza $l(\cdot)$ (rispetto a $\boldsymbol{\beta}$), per ottenere la score-function; procediamo per passi successivi:

$$\frac{\partial l}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$$

Tratto più estesamente il caso di legame canonico ($\mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i = \theta_i$), che porta ad

alcune semplificazioni.

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} =$$

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} \mathbf{x}_i$$

Quindi per trovare $\hat{\boldsymbol{\beta}}$ occorre risolvere rispetto a $\boldsymbol{\beta}$ il sistema di k equazioni:

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)} \mathbf{x}_i = \mathbf{0} \tag{8}$$

Usualmente $a(\phi)$ è costante; inoltre possiamo porre il sistema in forma matriciale (sempre nel caso semplificato di legame canonico)

$$\mathbf{X}^{\mathsf{T}}(\mathbf{y} - \boldsymbol{\mu}) \;=\; \mathbf{0}$$

Per la soluzione di questo sistema di equazioni si può ricorrere al metodo IRLS (Iterated reweighted Least Squares; minimi quadrati iterati riponderati); un alternativa è il metodo di newton-raphson, asintoticamente equivalente.

traccia del metodo iterativo IRLS:

Sviluppiamo prima in serie di taylor gli scarti della formula di sopra in un intorno della soluzione η_i*:

$$y_i - \mu_i \approx \frac{\partial \mu_i}{\partial \eta_i}(\eta_i * -\eta_i)$$

e quindi (dato che nel caso del legame canonico $\eta_i = \theta_i$), esplicitando rispetto a $(\eta_i * -\eta_i)$:

$$\eta_i * -\eta_i = (y_i - \mu_i)\frac{\partial \theta_i}{\partial \mu_i}$$

e ricordando che

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\mathrm{V}[Y_i]}{a(\phi)}$$

si ottiene:

$$y_i - \mu_i \approx (\eta_i * -\eta_i)\mathrm{V}[Y_i]\frac{1}{a(\phi)}$$

$$(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{V}(\boldsymbol{\eta} * -\boldsymbol{\eta})\frac{1}{a(\phi)}$$

sostituendo questa approssimazione nel sistema di equazioni 8, (e considerando

$a(\phi)$ costante), otteniamo in termini matriciali:

$$\mathbf{X}^{\mathsf{T}}(\mathbf{y} - \boldsymbol{\mu}) \;=\; \mathbf{0} \;\Rightarrow\; \mathbf{X}^{\mathsf{T}}\mathbf{V}(\boldsymbol{\eta} * - \boldsymbol{\eta}) = \mathbf{0}$$

e quindi riesprimendo rispetto a $\boldsymbol{\beta}$

$$\mathbf{X}^{\mathsf{T}}\mathbf{V}(\boldsymbol{\eta} * - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

e quindi

$$\mathbf{b} = (\mathbf{X}^{\mathsf{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{V}\boldsymbol{\eta}*$$

In questa espressione $\boldsymbol{\eta}*$ non è noto e può essere approssimato con l'approssimazione già usata:

$$\eta_i* = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$$

e in termini matriciali (sempre ricordando che siamo nel caso di legame canonico e quindi $\boldsymbol{\eta} = \boldsymbol{\theta}$):

$$\boldsymbol{\eta}* = \boldsymbol{\eta} + (\mathbf{y} - \boldsymbol{\mu})\mathbf{V}^{-1}a(\phi)$$

da cui si imposta uno schema iterativo.

8.2 proprietà asintotiche di b (link canonico)

$$\mathrm{E}\left[\mathbf{b}\right] = \boldsymbol{\beta}$$
$$\mathrm{V}\left[\mathbf{b}\right] = a(\phi)^2(\mathbf{X}^\mathsf{T}\mathbf{V}\mathbf{X})^{-1}$$

8.3 legame non canonico

(se il legame non è canonico possiamo procedere così:)

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} =$$

$$\frac{\partial \mu_i}{\partial \eta_i}x_{ij}\frac{(y-\mu)}{\mathrm{V}[Y]}$$

$$\frac{\partial l}{\partial \mu} = \frac{\partial l}{\partial \theta}\frac{\partial \theta}{\partial \mu} = \frac{\frac{\partial l}{\partial \theta}}{\frac{\partial \mu}{\partial \theta}} =$$

$$= \frac{(y-\mu)\frac{w}{\phi}}{b''(\theta)} = \frac{(y-\mu)}{\mathrm{V}[Y]}$$

$$l(\mathbf{y};\mu(\boldsymbol{\theta});\phi) =$$

$$\sum_{i=1}^{n} w_i(y_i\theta_i - b(\theta_i))/\phi + c(y_i;\phi)$$

$$\text{DEVIANZA} = -2[l(\mathbf{y};\mu(\boldsymbol{\theta});\phi) - l(\mathbf{y};\mathbf{y};\phi)] =$$

$$-2[\underbrace{l(\mathbf{y};\mu(\boldsymbol{\theta});\phi)}_{\text{log. ver. con parametro } \boldsymbol{\theta} \text{ e } \mu(\boldsymbol{\theta})} - \underbrace{l(\mathbf{y};\mathbf{y};\phi)}_{\text{log. ver. } \textbf{massima } \mu = \mathbf{y}}] =$$

Confronto fra la verosimiglianza per il modello in esame e la verosimiglianza massima, relativa al **modello saturo**

---

si può esprimere anche come:

$$\text{Dev} = \sum_{i=1}^{n} 2w_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) - b(\hat{\theta}_i))/\phi$$

$$= D(y;\hat{\boldsymbol{\mu}})/\phi$$

$\tilde{\theta}_i$ stima col modello saturo $\hat{\theta}_i$ stima col modello in esame

Per esempio:

- Normale: $D(y;\hat{\boldsymbol{\mu}}) = \Sigma_{i=1}^{n}(y_i - \hat{\mu}_i)^2$
- Poisson: $D(y;\hat{\boldsymbol{\mu}}) = 2\,\Sigma_{i=1}^{n}\, y_i(\log(y_i/\hat{\mu}_i)) - (y_i - \hat{\mu}_i)$

I residui vengono definiti a partire degli addendi della devianza:
se $D(y;\hat{\boldsymbol{\mu}}) = \Sigma_{i=1}^{n}\, d_i$ allora:
Residuo=$sign((y_i - \hat{\mu}_i))\sqrt{d_i}$
Queste quantità vengono utilizzate per validare a posteriori il modello adattato

---

Per confrontare modelli, possibilmente nidificati, si usa il rapporto delle verosimiglianze, ovviamente collegato alla differenza fra devianze.
Ha una distribuzione asintotica χ^2_q