UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche, aziendali, matematiche e statistiche "Bruno de Finetti"

Statistica (c.p.)

1. Modello statistico, verosimiglianza, SMV, esempi

Francesco Pauli

DEAMS

Università di Trieste

A.A. 2017/2018

Indice

Definizione e problema di specificazione

Un modello statistico parametrico consiste di una terna $(\mathcal{Y}, P_\theta, \Theta)$ dove

- $\mathcal{Y}$ è lo **spazio campionario** ossia l'insieme dei campioni che potrebbero essere osservati;
  - si noti che $\mathcal{Y} = \bigcup_{\theta \in \Theta} \text{supp}(P_\theta)$
- P_θ è una misura di probabilità su $\mathcal{Y}$, definire P_θ significa definire per ogni insieme $\mathcal{Y}_0 \in \mathcal{B}(\mathcal{Y})$ la probabilità $P_\theta(Y \in \mathcal{Y}_0)$. Le distribuzioni sono indicizzate da un parametro $\theta \in \mathbb{R}^d$

$$\mathcal{F} = \{p_\theta | \theta \in \Theta \subset \mathbb{R}^d\}$$

- $\Theta \subset \mathbb{R}^d$ è l'insieme dei possibili valori del parametro θ.

Sia poi

- P_0 la "vera" legge di probabilita che regola il meccanismo generatore dei dati.
- Il modello parametrico **specifica** una famiglia di distribuzioni $\mathcal{F} = \{P_\theta | \theta \in \Theta\}$ e si dice **correttamente specificato** se $P_0 \in \mathcal{F}$ e quindi $P_0 = P_{\theta_0}$ per un qualche θ_0.

Scelta del modello

La scelta del modello è un aspetto chiave della modellazione dei fenomeni e non vi è una regola generale, ma si possono fare una serie di considerazioni.

- rispettare il supporto delle variabili coinvolte, nel senso che il supporto del modello deve approssimare in maniera adeguata il supporto delle variabili
- Il modello da adottare è spesso suggerito da considerazioni di tipo asintotico, ad esempio il modello gaussiano nella teoria degli errori è in qualche modo suggerito dal teorema del limite centrale.
- Bilanciare le due esigenze contrapposte di elasticità del modello, per cui si vorrebbe una famiglia il più ampia possibile, e di precisione del modello (delle stime), per cui si vorrebbe un modello con un numero ridotto di parametri: non esiste una soluzione univoca al problema, è generalmente un processo iterativo.

Supporto di una v.a.

Sia X una v.a. e $p(x)$ la funzione di probabilità o di densità di X, si dice supporto di X l'insieme

$$\text{supp}\{p(X)\} = \{x \text{ t.c. } p(x) > 0\}$$

In altre parole

- se la v.a. è **discreta** il supporto è l'insieme di valori che X assume con **probabilità** positiva;
- se la v.a. è **continua** il supporto è l'insieme di valori che X assume con **densità** positiva;

Indice

pesci nel lago

Si osservano 20 pesci in un lago che ne contiene N di cui r rossi, sia $\theta = r/N$.

- modello per $y =$ numero di successi

$$\mathcal{Y} = \{0, 1, 2, \ldots, 20\}$$
$$p_\theta(y) = \text{Binom}(20, \theta)$$
$$\Theta = [0, 1]$$

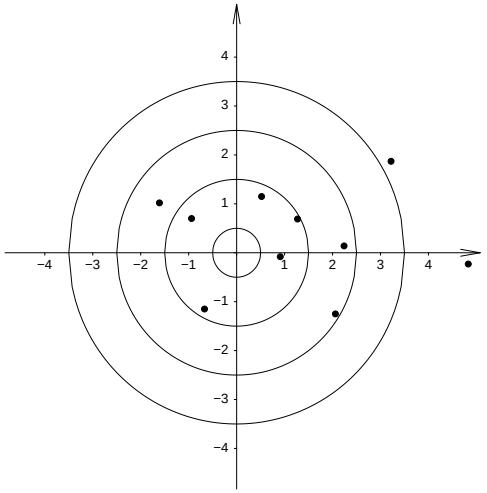
- modello per la sequenza di osservazioni $\mathbf{y} =$ 0011...0101

$$\mathcal{Y} = \{0, 1\}^{20}$$
$$p_\theta(\mathbf{y}) = \prod_{i=1}^{20} \theta^{y_i}(1-\theta)^{1-y_i}$$
$$\Theta = [0, 1]$$

nel primo non uso tutte le informazioni presenti nel campione.

Bersaglio

Di un fucile si vuole stabilire se il mirino sia difettoso. Si sparano allora (nelle stesse condizioni) n colpi contro un bersaglio e si rilevano le coordinate dei punti colpiti avendo fissato l'origine al centro del bersaglio. Si assume che la distribuzione delle coordinate sia gaussiana.



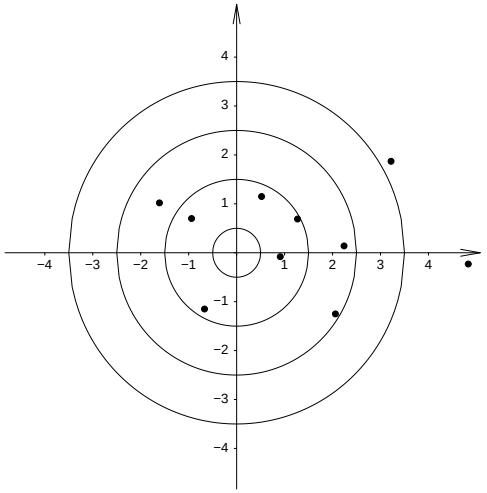
$$\mathcal{Y} = \mathbb{R}^2$$

$$p_\theta = \mathcal{N}\left(\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}, \begin{bmatrix}\sigma_x^2 & c \\ c & \sigma_y^2\end{bmatrix}\right)$$

$$\Theta =$$

Bersaglio

Di un fucile si vuole stabilire se il mirino sia difettoso. Si sparano allora (nelle stesse condizioni) n colpi contro un bersaglio e si rilevano le coordinate dei punti colpiti avendo fissato l'origine al centro del bersaglio. Si assume che la distribuzione delle coordinate sia gaussiana.



$$\mathcal{Y} = \mathbb{R}^2$$

$$p_\theta = \mathcal{N}\left(\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}, \begin{bmatrix}\sigma_x^2 & c \\ c & \sigma_y^2\end{bmatrix}\right)$$

$$\Theta = \left\{(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, c) \middle| \, \mu_x, \mu_y \in \mathbb{R}; \sigma_x^2, \sigma_y^2 \in \mathbb{R}_+, c \in \mathbb{R}, \begin{bmatrix}\sigma_x^2 & c \\ c & \sigma_y^2\end{bmatrix} \text{ def. pos.}\right\}$$

La definizione di Θ non è banale per via dell'ultima condizione.

Bersaglio: parametrizzazione

Essendo $\sigma_x^2 > 0$ la matrice è definita positiva sse

$$\begin{vmatrix} \sigma_x^2 & c \\ c & \sigma_y^2 \end{vmatrix} > 0 \quad \Leftrightarrow \quad \sigma_x^2\sigma_y^2 - c^2 > 0 \quad \Leftrightarrow \quad -1 < \frac{c}{\sigma_x\sigma_y} < 1$$

Conviene allora riscrivere il modello ponendo

$$\rho = \frac{c}{\sigma_x\sigma_y}$$

sicché **invece del parametro θ si ha il nuovo parametro**

$$\psi = (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

dove lo spazio parametrico è

$$\Psi = \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+ \times [-1, 1]$$

si noti che al cambiamento dei parametri non corrisponde un cambiamento del modello.

Riparametrizzazione

Sia $\mathcal{F} = \{p(\cdot;\theta)|\theta \in \Theta\}$ una famiglia parametrica e

$$h : \Theta \to \Psi = \{h(\theta)|\theta \in \Theta\}$$

un'applicazione biunivoca, se consideriamo allora la famiglia

$$\mathcal{F}' = \{p(\cdot;\psi)|\psi \in \Psi\}$$

questa coincide con $\mathcal{F}$ nel senso che nelle due si trovano le stesse distribuzioni, solo indicizzate in modi diversi.
Notiamo che, chiaramente, vorremmo che le conclusioni dell'inferenza fossero invarianti rispetto alla parametrizzazione dato il carattere arbitrario di questa.

Scelta della parametrizzazione

La parametrizzazione è arbitraria.
La scelta è dettata principalmente da

- interpretabilità: il parametro dovrebbe avere un significato diretto nel problema che si va a considerare, tale interpretazione dovrebbe poi essere robusta rispetto a ragionevoli perturbazioni del modello;
- semplicità di inferire: la teoria statistica per la stima dovrebbe essere semplice, ad esempio si vorrebbe che le componenti di θ fossero il più possibile ortogonali e che la parametrizzazione favorisse la convergenza di eventuali metodi computazionali iterativi usati nella stima.

Notiamo che nel primo punto abbiamo raggruppato gli aspetti legati alla materia oggetto di analisi, mentre nel secondo punto consideriamo gli aspetti tecnici, legati al metodo di stima.
Se si hanno **parametri di disturbo**, per questi valgono solo le considerazioni tecniche, mentre per i parametri di interesse occorrerà bilanciare interpretabilità ed esigenze tecniche.

Marche di lampadine

- Confronto la durata di lampadine di due marche.
- Si assume che, per ciascuna marca, il tempo prima che si fulminino sia distribuito secondo un'esponenziale.

Si osservano i tempi di rottura di

- n lampadine di marca A: $y_1, \ldots, y_n$
- m di marca B: $z_1, \ldots, z_m$.

Si definisce il modello

$$\mathcal{Y} = \mathbb{R}_+^{n+m}$$

$$\begin{cases} Y_i \sim \text{espon}(a+b) \\ Z_i \sim \text{espon}(a+c) \end{cases}$$

$$\Theta =$$

Marche di lampadine

- Confronto la durata di lampadine di due marche.
- Si assume che, per ciascuna marca, il tempo prima che si fulminino sia distribuito secondo un'esponenziale.

Si osservano i tempi di rottura di

- n lampadine di marca A: $y_1, \ldots, y_n$
- m di marca B: $z_1, \ldots, z_m$.

Si definisce il modello

$$\mathcal{Y} = \mathbb{R}_+^{n+m}$$

$$\begin{cases} Y_i \sim \text{espon}(a+b) \\ Z_i \sim \text{espon}(a+c) \end{cases}$$

$$\Theta = \{(a,b,c) | \, a \in \mathbb{R}_+, b > -a, c > -a\}$$

Marche di lampadine

- Confronto la durata di lampadine di due marche.
- Si assume che, per ciascuna marca, il tempo prima che si fulminino sia distribuito secondo un'esponenziale.

Si osservano i tempi di rottura di

- n lampadine di marca A: $y_1, \ldots, y_n$
- m di marca B: $z_1, \ldots, z_m$.

Si definisce il modello

$$\mathcal{Y} = \mathbb{R}_+^{n+m}$$

$$\begin{cases} Y_i \sim \text{espon}(a+b) \\ Z_i \sim \text{espon}(a+c) \end{cases}$$

$$\Theta = \{(a,b,c) | \, a \in \mathbb{R}_+, b > -a, c > -a\}$$

Questo "modello" ha un difetto, quale che sia $d > 0$ i due elementi dello spazio parametrico (a, b, c) e $(a + d, b - d, c - d)$ identificano la stessa legge di probabilità del campione.

Marche di lampadine

- Confronto la durata di lampadine di due marche.
- Si assume che, per ciascuna marca, il tempo prima che si fulminino sia distribuito secondo un'esponenziale.

Si osservano i tempi di rottura di

- n lampadine di marca A: $y_1, \ldots, y_n$
- m di marca B: $z_1, \ldots, z_m$.

Un modello accettabile sarebbe invece

$$\begin{aligned} &\mathcal{Y} = \mathbb{R}_+^{n+m} \\ &\begin{cases} Y_i \sim \text{espon}(a) \\ Z_i \sim \text{espon}(a+b) \end{cases} \\ &\Theta = \{(a,b) | \, a \in \mathbb{R}_+, b > -a\} \end{aligned} \tag{1}$$

(non l'unico, proporne altri).

Identificabilità

Definizione

Un modello è **identificabile** se per ogni $\theta_1 \neq \theta_2$ esiste $B \subset \mathcal{Y}$ tale che $P_{\theta_1}(Y \in B) \neq P_{\theta_2}(Y \in B)$.

Per contro, un modello risulta non identificabile quando esiste una coppia θ_1, θ_2 con $\theta_1 \neq \theta_2$ tale per cui qualunque sia $B \subset \mathcal{Y}$ si ha $P_{\theta_1}(Y \in B) = P_{\theta_2}(Y \in B)$.

Alla luce dei concetti appena presentati è chiara l'esigenza di costruire un modello identificabile, altrimenti se per qualche coppia di valori del parametro si avesse $p_{\theta_1}(y) = p_{\theta_2}(y)$ per ogni possibile campione $\mathbf{y}$, non saremmo in grado di decidere quale valore vada preferito.

Indice

Verosimiglianza

Definizione

Dato il modello statistico $(\mathcal{Y}, p_\theta, \Theta)$ si dice **funzione di verosimiglianza** una qualunque funzione $L : \Theta \to \mathbb{R}_+$ tale che

$$L(\theta) = L(\theta; \mathbf{y}) = c(y)p(\mathbf{y}; \theta)$$

dove $c(\mathbf{y}) > 0$ non dipende da θ.

Si definisce **funzione di log-verosimiglianza** il logaritmo naturale della funzione di verosimiglianza.

Osservazioni sulla proporzionalità

Con verosimiglianza, in altre parole, si intende un insieme di funzioni (al variare di c) che è null'altro che una classe di equivalenza dove due funzioni sono equivalenti se sono proporzionali.

Coerentemente, la funzione di log-verosimiglianza è definita a meno di una costante additiva, infatti

$$l(\theta) = \log L(\theta) = \log(c(y)p(\mathbf{y};\theta)) = \log(c(y)) + \log(p(\mathbf{y};\theta)).$$

È usuale rappresentare la funzione di verosimiglianza con il massimo pari a 1 (e quindi la log-verosimiglianza con il massimo pari a 0).

Si noti che la FdV non è una distribuzione di probabilità (il suo valore è sì proporzionale a una distribuzione di probabilità, ma una diversa per ogni valore di θ).

Perché a meno di una costante moltiplicativa?

Si può argomentare immaginando di avere un campione $X \sim p_\theta(x)$ e di considerare una trasformazione biunivoca $Y = y(x)$, sicché

$$p_\theta(y) = p_\theta(x) \left| \frac{dx}{dy} \right| .$$

Il campione y e il campione x portano la stessa informazione su θ (sono lo stesso campione, di fatto), per cui si vuole che le conclusioni siano invarianti rispetto a qualunque trasformazione di questo tipo. Questo si realizza, appunto, se si utilizzano i rapporti, si ha infatti

$$\frac{L(\theta_2; y)}{L(\theta_1; y)} = \frac{L(\theta_2; x)}{L(\theta_1; x)}$$

quali che siano θ_1 e θ_2.

Verosimiglianza e non identificabilità

Se un modello non è identificabile, allora esistono due valori del parametro θ_1 e θ_2 tali che

$$L(\theta_1; \mathbf{y}) = L(\theta_2; \mathbf{y}) \quad \forall \mathbf{y}$$

ossia non c'è modo di discriminare tra i due sulla base di L.

Esempio: campione IID Normale

Sia $Y_1, \ldots, Y_n$ un campione *IID* da una $\mathcal{N}\left(\mu, \sigma^2\right)$, si ha allora, se $f(\cdot)$ indica la densità dell'i-esima variabile

$$
\begin{aligned}
p_\theta(\mathbf{y}) &= \prod_{i=1}^{n} f(y_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\theta)^2\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i-\theta)^2\right\}
\end{aligned}
$$

$$
L(\theta) \propto \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i-\theta)^2\right\}
$$

Esempio: processo $AR(1)$

Il modello $AR(1)$ è un modello per serie storiche in cui non si ha né indipendenza né identica distribuzione. Sia $y_1, \ldots, y_n$ il campione, e sia

$$y_t = \rho y_{t-1} + \varepsilon_t$$

dove gli ε_t sono un processo $IID(\mathcal{N}(0, \sigma^2))$, σ^2 noto, si ha allora

$$\begin{aligned}
p_\theta(\mathbf{y}) &= f(y_1) f(y_2|y_1) \ldots f(y_n|y_{n-1}) \\
&= f(y_1) \prod_{t=2}^{n} \exp\left\{ -\frac{1}{2\sigma^2} (y_t - \rho y_{t-1})^2 \right\} \\
&= f(y_1) \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=2}^{n} (y_t - \rho y_{t-1})^2 \right\} \\
L(\theta) &\propto \exp\left\{ -\frac{1}{2\sigma^2} \left(\rho^2 \sum_{t=2}^{n} y_{t-1}^2 - 2\rho \sum_{t=2}^{n} y_t y_{t-1} \right) \right\}
\end{aligned}$$

Esempio: regressione

Nella regressione si hanno osservazioni indipendenti ma non identicamente distribuite. Sia $(x_1, y_1), \ldots, (x_n, y_n)$ un campione con

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

dove gli ε_i sono un processo $IID(\mathcal{N}(0, \sigma^2))$, si ha allora $\mathcal{Y} = \mathbb{R}^2$, $\theta = (\alpha, \beta, \sigma^2) \in \Theta = \mathbb{R}^2 \times \mathbb{R}_+$

$$\begin{aligned} p_\theta(\mathbf{y}) &= \prod_{i=1}^{n} f_i(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2 \right\} \\ L(\theta) &\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right\} \end{aligned}$$

dove un ulteriore semplificazione sarebbe possibile se σ^2 fosse noto.

Indice

La stima puntuale

Definizione

Stimare un parametro significa associare al campione osservato **y** un valore all'interno dello spazio parametrico, cioè definire una statistica

$$\hat{\theta}(\cdot) : \mathcal{Y} \to \Theta$$

$$\hat{\theta} : y \mapsto \hat{\theta}(y)$$

detta **stimatore**.

Osservazione: definizione di stimatore

La definizione di stimatore è estremamente generale, l'unica relazione che essa stabilisce tra stimatore e oggetto da stimare è che i valori possibili dello stimatore siano anche valori possibili del parametro (che siano tutti non è detto).

Infatti, le procedure che si usano per proporre a partire dal campione un valore del parametro sono estremamente varie è spesso costruite ad hoc per un particolare problema, e anche i metodi più generali quali il metodo dei momenti o la massima verosimiglianza sono alternativi tra loro e privi di elementi comuni suscettibili di entrare in una definizione.

Per queste ragioni si preferisce non cercare di individuare caratteristiche comuni ma adottare una definizione vaga come quella appena riportata, essendo chiaro poi che uno stimatore è effettivamente utilizzabile se e in quanto i valori che esso assume "somigliano" in qualche senso al parametro da stimare.

Stimatore di massima verosimiglianza

Con la logica della verosimiglianza uno stimatore naturale è il valore di θ che rende massima la probabilità di osservare y se questo esiste ed è unico.

Definizione

Si dice **stimatore di massima verosimiglianza** il valore di θ che rende massima la verosimiglianza del campione $L(\theta)$ – o, equivalentemente, la log-verosimiglianza $l(\theta)$ – se questo esiste ed è unico con probabilità uno.

Stima di una proporzione (segue) I

Nell'esempio già visto sulla binomiale, in cui si osservano $y_1, \ldots, y_n$ bernoulliane di parametro θ si è già visto che il modello è

$$p_\theta(\sum_i y_i) = \binom{n}{\sum_i y_i} \theta^{\sum_i y_i} (1-\theta)^{n-\sum_i y_i}$$

dove $\theta \in [0, 1]$. Si ha dunque la verosimiglianza

$$L(\theta) = \theta^{\sum_i y_i} (1-\theta)^{n-\sum_i y_i}$$

In queste condizioni il problema di trovare lo SMV si riduce a quello di massimizzare una funzione reale di variabile reale, conviene operare con la log-verosimiglianza

$$l(\theta) = (\sum_i y_i) \log \theta + (n - \sum_i y_i) \log(1-\theta).$$

Stima di una proporzione (segue) II

Si cerca allora il massimo studiando la derivata prima, che prende anche il nome di funzione punteggio

$$l_*(\theta) = \frac{\sum_i y_i}{\theta} - \frac{n - \sum_i y_i}{1-\theta}$$

Eguagliando a zero la funzione punteggio si ottengono i punti stazionari, nel caso in ispecie si trova che $l_*(\theta) = 0$ se e solo se

$$\begin{aligned}
\frac{\sum_i y_i}{\theta} &= \frac{n - \sum_i y_i}{1-\theta} \\
\frac{1-\theta}{\theta} &= \frac{n - \sum_i y_i}{\sum_i y_i} \\
\frac{1}{\theta} &= 1 + \frac{n - \sum_i y_i}{\sum_i y_i}
\end{aligned}$$

Stima di una proporzione (segue) III

e quindi se e solo se $\theta = \sum_i y_i/n$.
Che questo sia un punto di massimo lo si mostra o osservando che

$$l_*(\theta) > 0 \Leftrightarrow \theta < \sum_i y_i/n$$

o calcolando la derivata seconda e mostrando che è negativa

$$l''(\theta) = -\frac{\sum_i y_i}{\theta^2} - \frac{n - \sum_i y_i}{(1-\theta)^2}$$

si ha quindi $l''(\theta) < 0$ per ogni θ quindi $\sum_i y_i/n$ è un punto di massimo assoluto (se fosse stato semplicemente $l''(\sum_i y_i/n) < 0$ questo avrebbe garantito soltanto che $\sum_i y_i/n$ fosse stato un massimo locale).

Normale I

Sia $Y_1, \ldots, Y_n$ un campione *IID* da una $\mathcal{N}\left(\theta, \sigma^2\right)$, si ha allora

$$
\begin{aligned}
L(\theta) = p_\theta(\mathbf{y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \theta)^2\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - \theta)^2\right\} \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_i y_i^2 - 2\theta\sum_i y_i + n\theta^2\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\left(n\theta^2 - 2\theta\sum_i y_i\right)\right\} \\
&\propto \exp\left\{-\frac{n}{2\sigma^2}(\theta - \bar{y})^2\right\}
\end{aligned}
$$

Normale II

e quindi la log verosimiglianza è la parabola

$$l(\theta) = -\frac{n}{2\sigma^2}(\theta - \bar{y})^2 .$$

con la concavità verso il basso e il massimo, dunque, in $\hat{\theta} = \bar{y}$
La funzione punteggio è

$$l_*(\theta) = -\frac{n}{\sigma^2}(\theta - \bar{y}) .$$

ed è nulla ovviamente in $\theta = \bar{y}$ dove la derivata seconda

$$l''(\theta) = -\frac{n}{\sigma^2}$$

è negativa (lo è sempre).

Lo SMV non esiste sempre

Sia $Y_1, \ldots, Y_n$ un campione *IID* dal modello

$$Y \sim \mathcal{N}(\theta, 1) \qquad \Theta \in]0, +\infty[$$

si ha allora

$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmax}} l(\theta) = \begin{cases} \bar{Y} & \text{se } \bar{Y} > 0 \\ \text{non esiste} & \text{se } \bar{Y} \leq 0 \end{cases}$$

e lo stimatore non esiste con probabilità $\Phi(-\theta\sqrt{n})$, che è positiva qualunque sia θ.

Lo SMV non esiste sempre, altro esempio I

Sia $Y_1, \ldots, Y_n$ un campione *IID* dal modello

$$p_\theta(y) = \theta^{-1} I_{]0,\theta[}(y)$$

con $\theta > 0$. La verosimiglianza è allora

$$L(\theta) = \theta^{-n} I_{]y_{(n)},+\infty[}(\theta)$$

si ha allora

$$L(\theta) < \sup_{\theta \in \Theta} L(\theta) = L(y_{(n)}).$$

SMV non unica

Sia $Y_1, \ldots, Y_n$ un campione *IID* dalla distribuzione di Laplace, si ha cioè

$$f(y_i; \theta) = \frac{1}{2} e^{-|y_i - \theta|}$$

con $y \in \mathbb{R}$ e $\theta \in \mathbb{R}$. Si ha allora

$$l(\theta) = -\sum_{i=1}^{n} |y_i - \theta|$$

SMV non unica

Sia $Y_1, \ldots, Y_n$ un campione *IID* dalla distribuzione di Laplace, si ha cioè

$$f(y_i; \theta) = \frac{1}{2} e^{-|y_i - \theta|}$$

con $y \in \mathbb{R}$ e $\theta \in \mathbb{R}$. Si ha allora

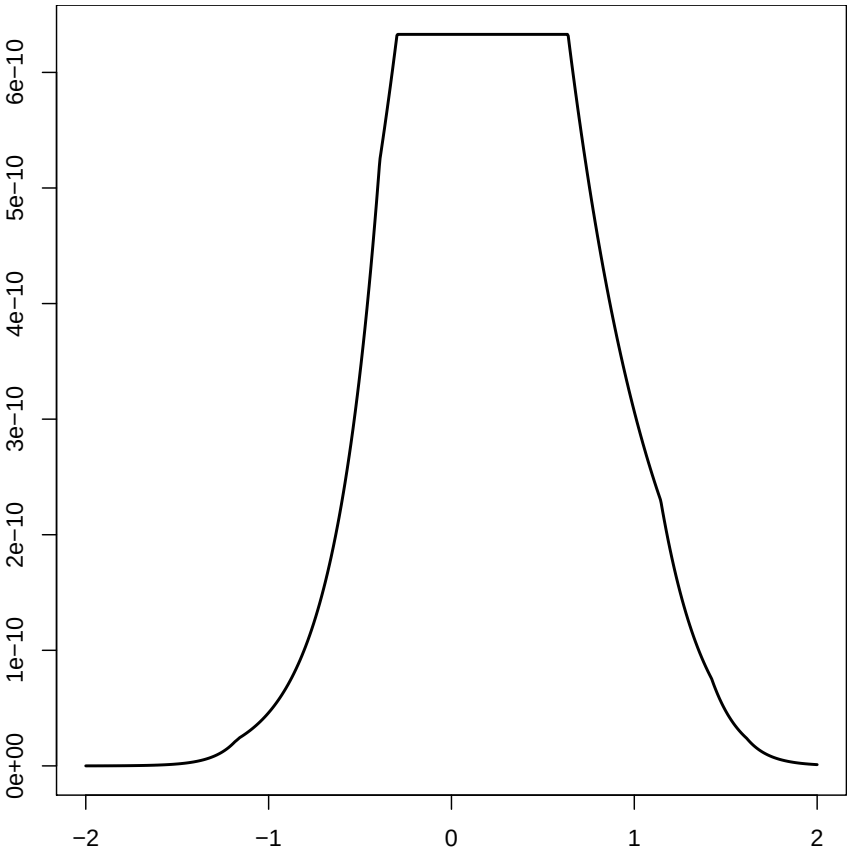
$$l(\theta) = -\sum_{i=1}^{n} |y_i - \theta|$$

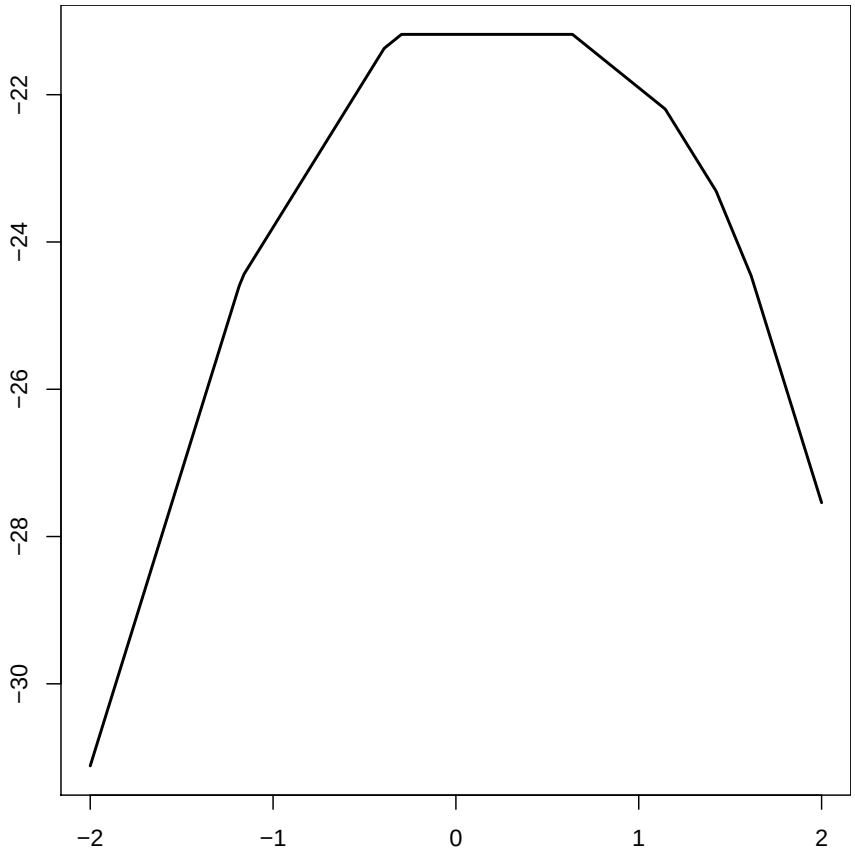
che è massima in corrispondenza della mediana campionaria, cioè si ha

$$\hat{\theta} \begin{cases} = y_{(k)} & \text{se } n = 2k - 1 \\ \in [y_{(k)}, y_{(k+1)}] & \text{se } n = 2k \end{cases}$$

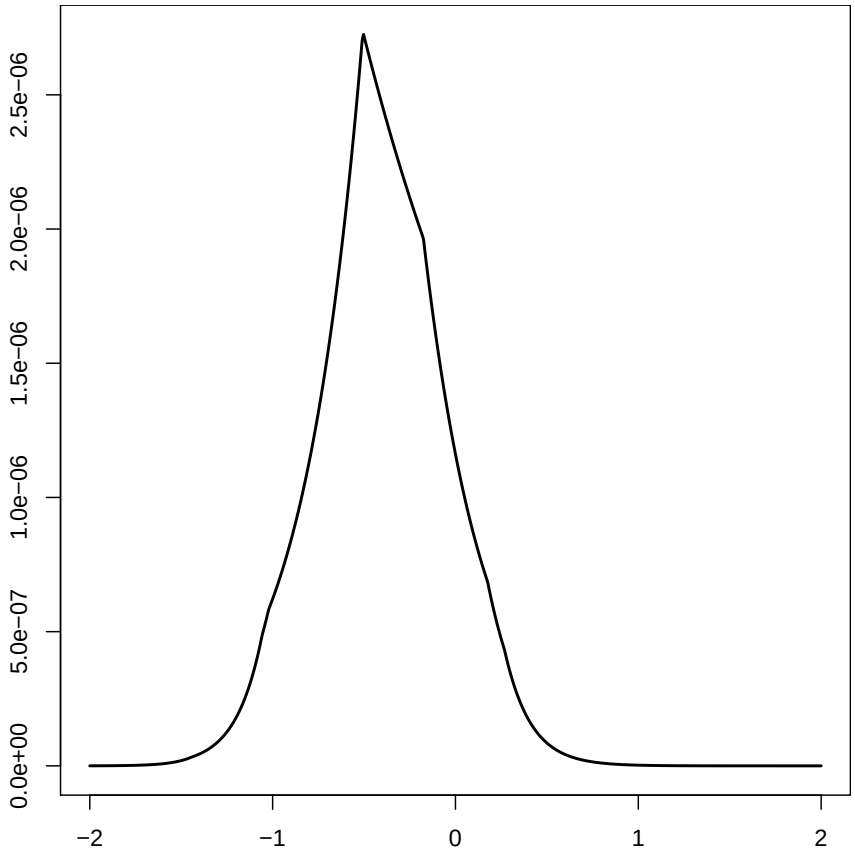
e il massimo non è unico.

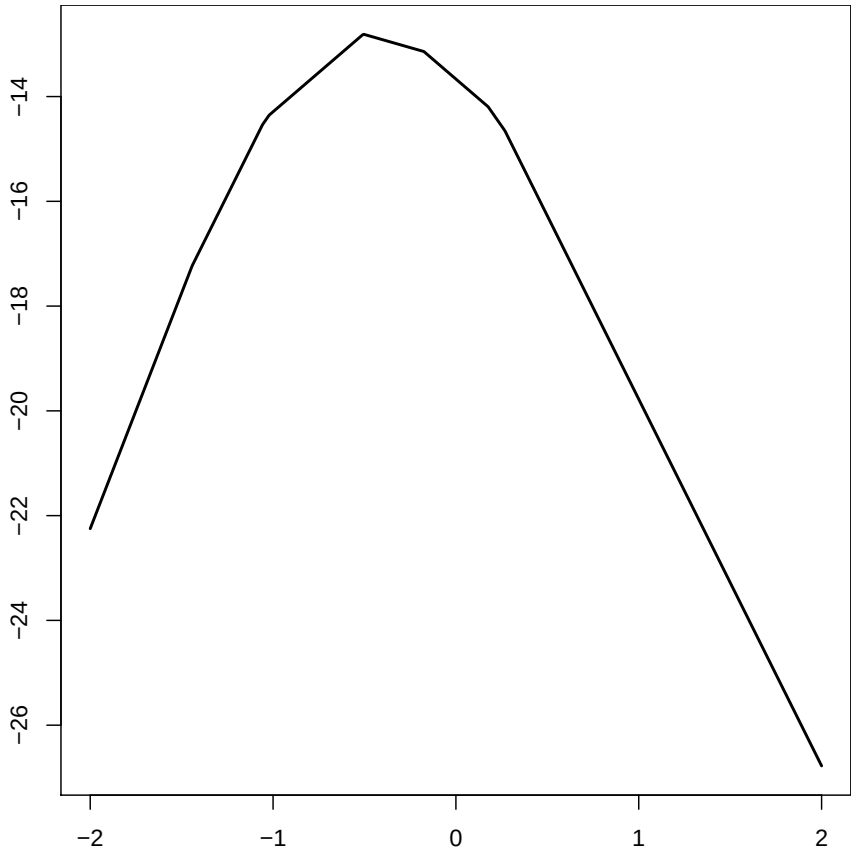
Verosimiglianza e log verosimiglianza Laplace, n pari





Verosimiglianza e log verosimiglianza Laplace, n dispari





SMV non unica II

Sia $Y_1, \ldots, Y_n$ un campione IID da

$$y_i \sim \text{Unif}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$$

dove $\theta - \frac{1}{2} \leq y_{(1)} \leq y_{(n)} \leq \theta + \frac{1}{2}$. La verosimiglianza è allora

$$L(y) = \begin{cases} 1 & \text{se } y_{(n)} - \frac{1}{2} \leq \theta \leq y_{(1)} + \frac{1}{2} \\ 0 & \text{altrimenti} \end{cases},$$

pertanto il massimo si realizza per ogni $\theta \in [y_{(n)} - \frac{1}{2}, y_{(1)} + \frac{1}{2}]$.

Indice

Uniforme

Sia

$$Y_1, \ldots, Y_n \sim \text{Unif}(0, \theta)$$

si noti che $\mathcal{Y} = [0, \theta]^n$ dipende dal parametro.
La distribuzione di probabilità di y_i è

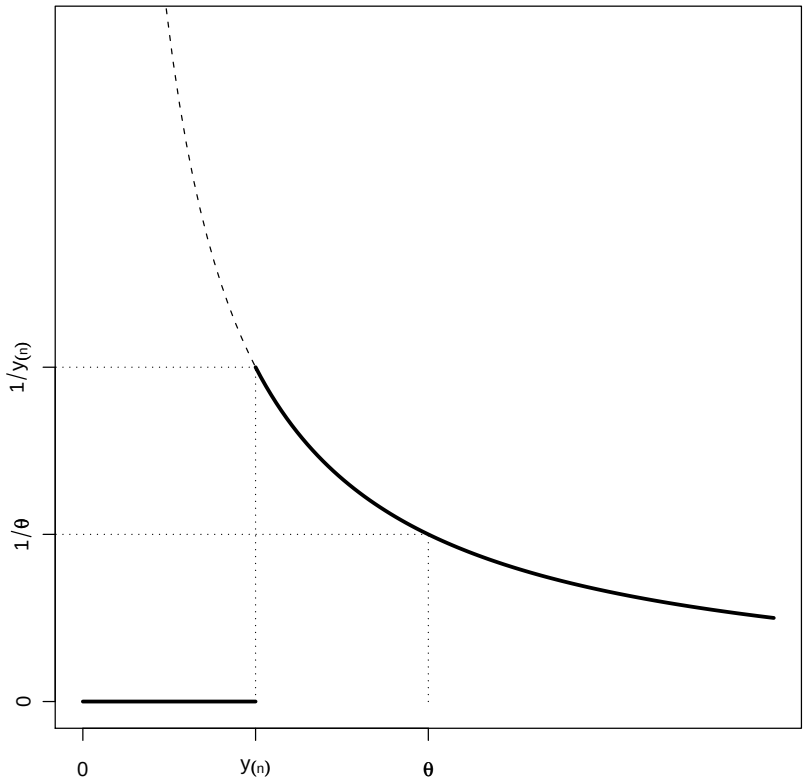
$$\begin{aligned} f_Y(y; \theta) &= \begin{cases} \theta^{-1} & \text{se } 0 \leq y \leq \theta \\ 0 & \text{altrimenti} \end{cases} \\ &= \theta^{-1} I_{[0,\theta]}(y) \end{aligned}$$

Pertanto la verosimiglianza è

$$L(\theta) = f(\mathbf{y}; \theta) = \prod_{i=1}^{n} \theta^{-1} I_{[0,\theta]}(y_i) = \theta^{-n} I_{[0,\theta]}(y_{(n)}) = \theta^{-n} I_{[y_{(n)}, +\infty]}(\theta)$$

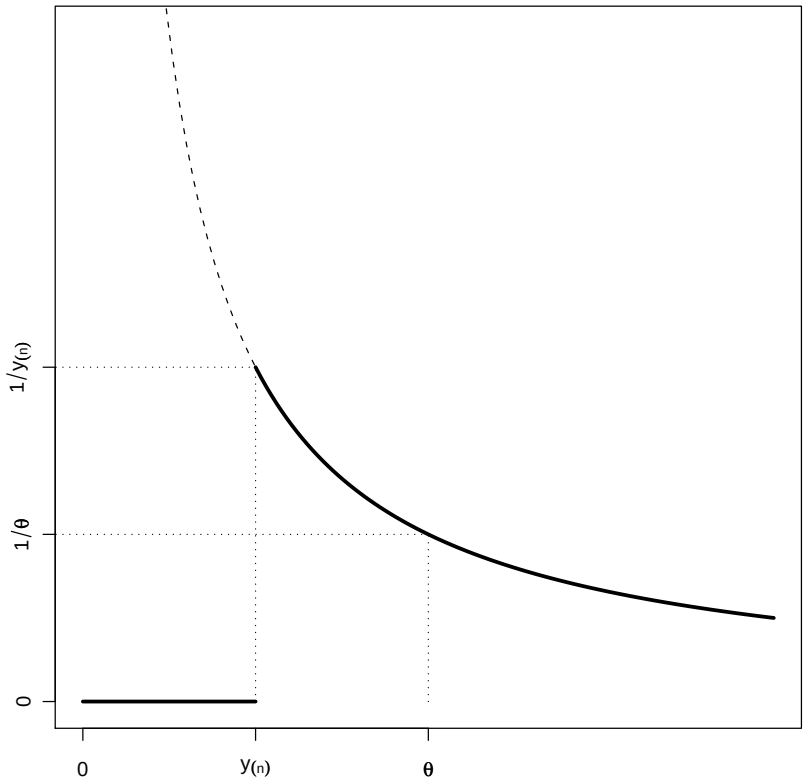
Uniforme

$$L(\theta) = f(\mathbf{y};\theta) = \prod_{i=1}^{n} \theta^{-1} I_{[0,\theta]}(y_i) = \theta^{-n} I_{[0,\theta]}(y_{(n)}) = \theta^{-n} I_{[y_{(n)},+\infty]}(\theta)$$



Uniforme

$$L(\theta) = f(\mathbf{y};\theta) = \prod_{i=1}^{n} \theta^{-1} I_{[0,\theta]}(y_i) = \theta^{-n} I_{[0,\theta]}(y_{(n)}) = \theta^{-n} I_{[y_{(n)},+\infty]}(\theta)$$



Il massimo è

$$\hat{\theta} = y_{(n)}$$

lo SMV è ben definito è unico

- è un punto di frontiera
- la derivata non si annulla

normale, caso multivariato I

Sia $Y_1, \ldots, Y_n \sim IIDN\left(\mu, \sigma^2\right)$, la verosimiglianza è (con $\theta = (\mu, \sigma^2)$)

$$L(\theta, \mathbf{y}) = (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\},$$

di conseguenza si ha la log-verosimiglianza

$$l(\theta, \mathbf{y}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Il parametro ha dimensione due, la funzione punteggio è

$$l_*(\theta) = \begin{bmatrix} \frac{\partial}{\partial\mu} l(\mu, \sigma^2) \\ \frac{\partial}{\partial\sigma^2} l(\mu, \sigma^2) \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2}(\bar{y} - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_i (y_i - \mu)^2 \end{bmatrix}.$$

Eguagliando a 0, s'ottiene l'equazione di verosimiglianza, cioè il sistema $l_*(\theta) = 0$,

normale, caso multivariato II

- da $0 = \frac{\partial}{\partial \mu} l(\mu, \sigma^2)$ si ha $\mu = \bar{y}$;
- da $0 = \frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2)$ si ha

$$0 = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_i (y_i - \mu)^2$$

sostituendo $\mu = \bar{x}$ si ottiene

$$\sigma^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2$$

normale, caso multivariato III

Si calcolano poi le derivate seconde

$$\frac{\partial^2}{\partial \mu^2} = -\frac{n}{\sigma^2}$$

$$\frac{\partial^2}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_i (y_i - \mu)^2$$

$$\frac{\partial^2}{\partial \mu \partial (\sigma^2)} = -\frac{n}{\sigma^4}(\bar{y} - \mu)$$

e si ha quindi l'hessiano

$$H(\theta) = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{n}{\sigma^4}(\bar{y} - \mu) \\ -\frac{n}{\sigma^4}(\bar{y} - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_i (y_i - \mu)^2 \end{bmatrix}$$

normale, caso multivariato IV

il minore di nord-ovest, $-\frac{n}{\sigma^2}$ è negativo, il determinante è

$$|H(\theta)| = -\frac{n^2}{\sigma^6} + \frac{n}{\sigma^8}\sum_i (y_i - \mu)^2 - \frac{n^2}{\sigma^8}(\bar{y} - \mu)^2$$

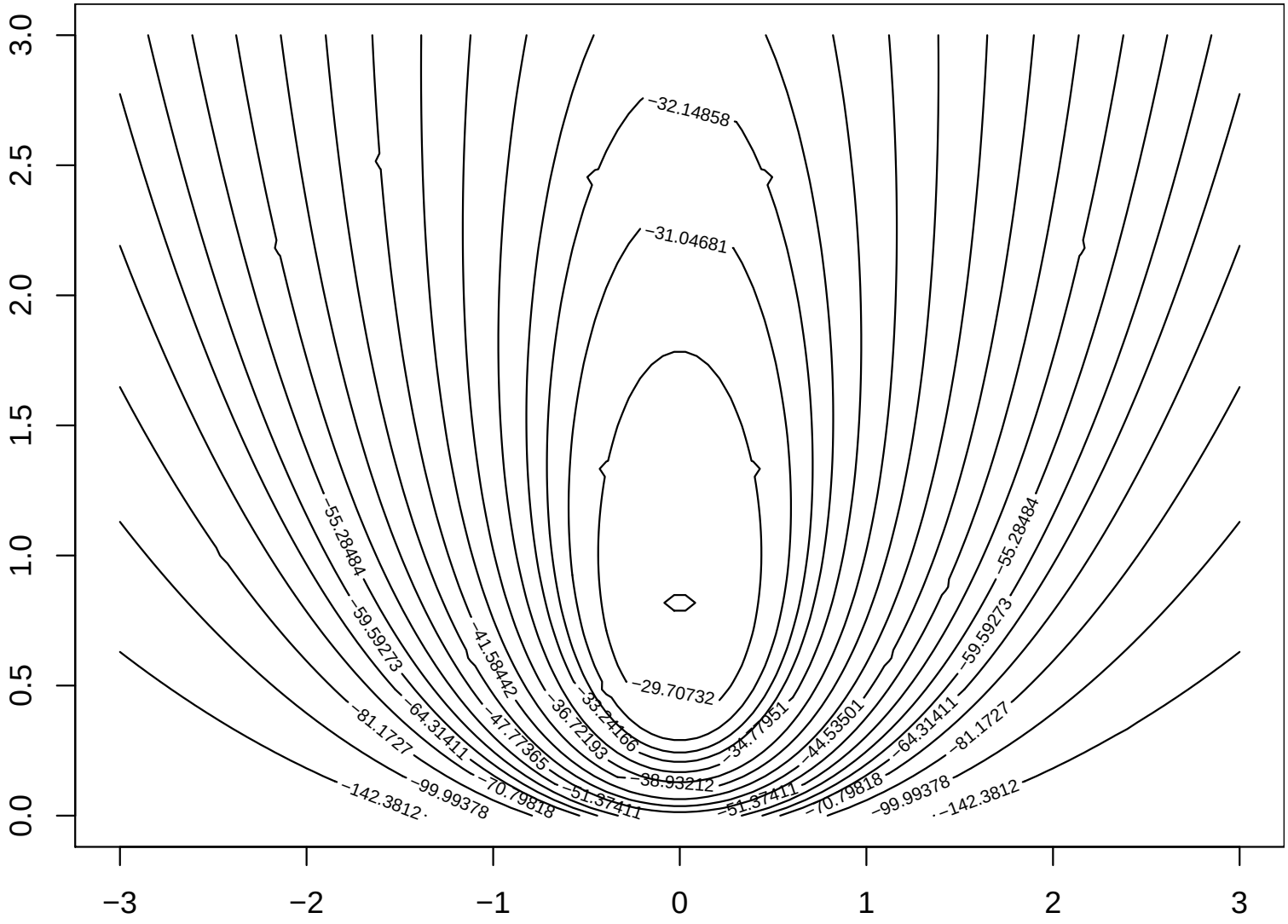
$$= \frac{n}{\sigma^6}\left(-\frac{n}{2} + \frac{1}{\sigma^2}\sum_i (y_i - \mu)^2 - \frac{n}{\sigma^2}(\bar{y} - \mu)^2\right)$$

nel punto stazionario $\mu = \bar{y}$ e $\sigma^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2$ si ha

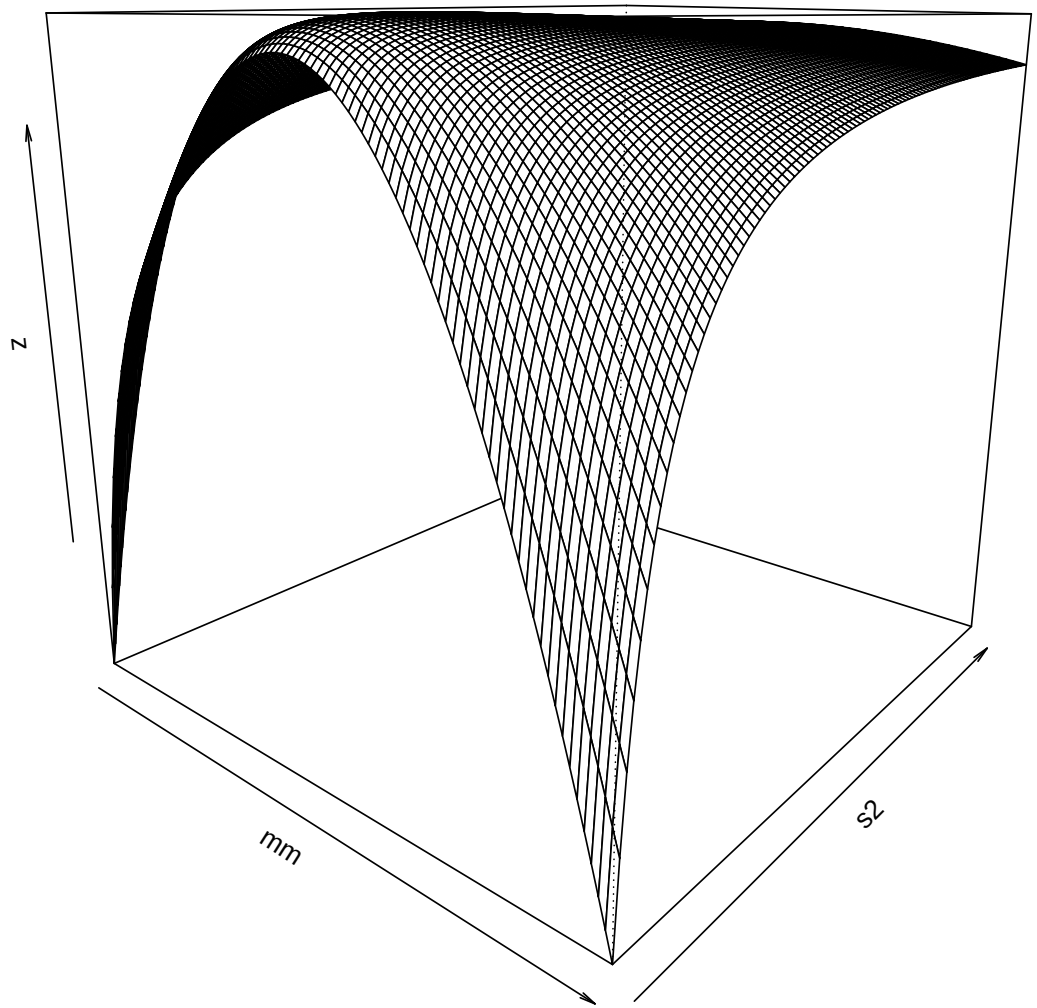
$$= \frac{n}{\sigma^6}\left(-\frac{n}{2} + n\right) > 0$$

quindi H è definita negativa in $\hat{\theta} = \left(\mu = \bar{y}, \sigma^2 = \frac{1}{n}\sum_i (y_i - \bar{y})^2\right)$ che è quindi massimo locale. La frontiera, $\sigma^2 = 0$ non può ospitare il massimo, quindi $\hat{\theta}$ è lo SMV.

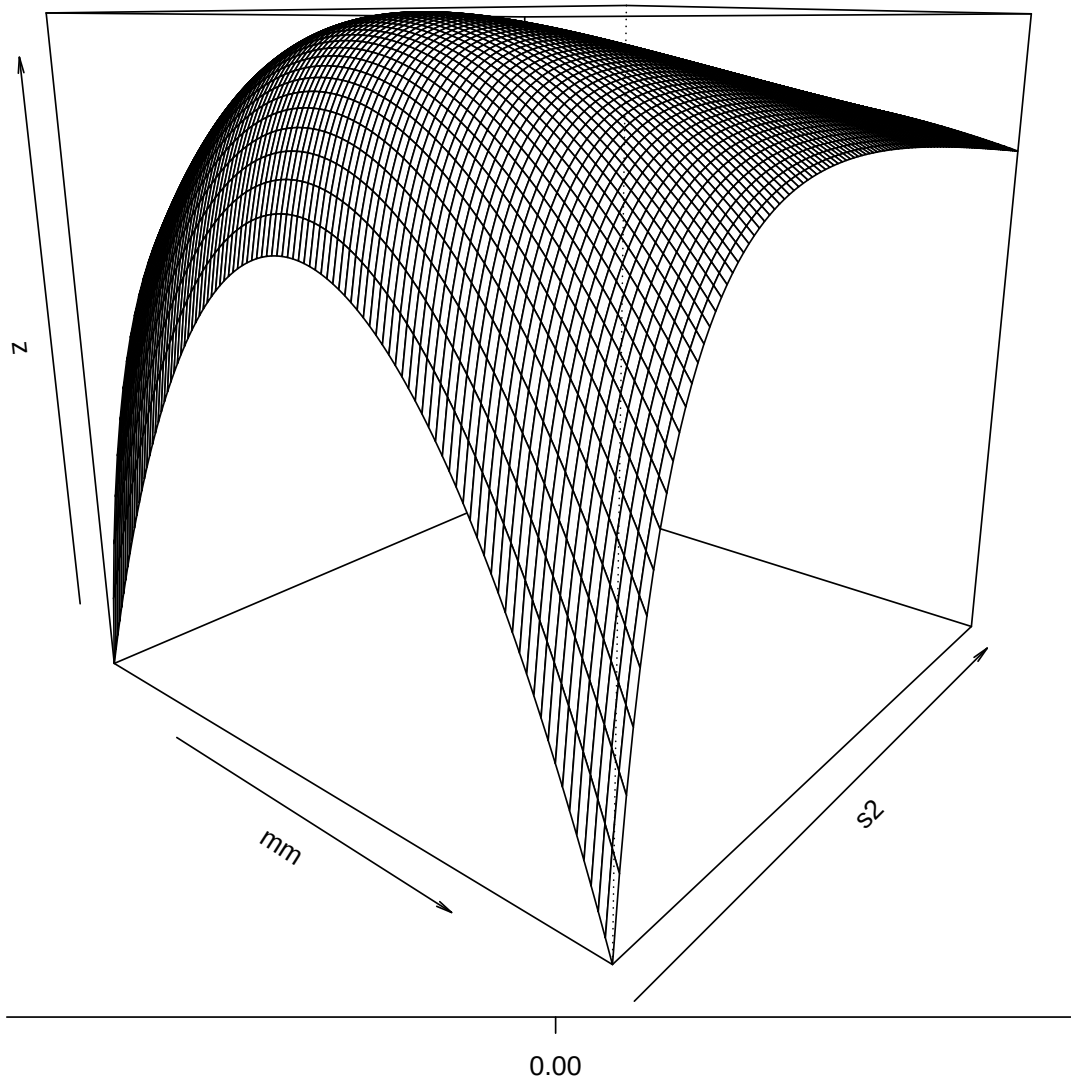
normale, caso multivariato V



normale, caso multivariato VI



normale, caso multivariato VII



Gumbel I

Sia $Y_1, \ldots, Y_n \sim IID\text{Gumbel}(\theta, 1)$

$$f(y_i; \theta) = e^{-(y-\theta)} \exp\left\{-e^{-(y-\theta)}\right\}$$

la verosimiglianza associata al campione è perciò

$$L(\theta) = \prod_{i=1}^{n} e^{-(y_i-\theta)} \exp\left\{-e^{-(y_i-\theta)}\right\} = e^{-\sum_{i=1}^{n}(y_i-\theta)} \exp\left\{-\sum_{i=1}^{n} e^{-(y_i-\theta)}\right\}$$

da cui la log-verosimiglianza

$$l(\theta) = -\sum_{i=1}^{n}(y_i - \theta) - \sum_{i=1}^{n} e^{-(y_i-\theta)}$$

Gumbel II

e la funzione punteggio

$$l_*(\theta) = n - e^{\theta} \sum_{i=1}^{n} e^{-y_i}$$

da cui

$$\hat{\theta} = \log(n / \sum_{i=1}^{n} e^{-y_i})$$

che è SMV in quanto la derivata seconda della log verosimiglianza risulta

$$l''(\theta) = -e^{\theta} \sum_{i=1}^{n} e^{-y_i} < 0 \ \ \forall \theta$$

Esponenziale I

Sia $Y_1, \ldots, Y_n \sim IIDp_\theta(y) = \theta e^{-\theta y}$, $\Theta = \mathbb{R}^+$ e $\mathcal{Y} = \mathbb{R}^n_+$.
La verosimgilianza è dunque

$$L(\theta) = \theta^n \exp\left\{-\theta \sum_i y_i\right\}$$

la log-verosimiglianza è

$$l(\theta) = \log L(\theta) = n \log \theta - \theta \sum_i y_i$$

la funzione punteggio è

$$l_*(\theta) = \frac{n}{\theta} - \sum_i y_i.$$

Esponenziale II

e l'equazione di verosimiglianza è

$$0 = l_*(\theta) = \frac{n}{\theta} - \sum_i y_i$$

che è soddisfatta per

$$\hat{\theta} = \frac{n}{\sum_i y_i} = \frac{1}{\bar{y}}$$

questo mostra che $\hat{\theta}$ è un punto stazionario, per distinguere se è un massimo locale, un minimo locale o una sella si calcola la derivata seconda, $\hat{\theta}$ è un massimo locale se $l''(\hat{\theta}) < 0$, in questo caso in particolare

$$l''(\theta) = -\frac{n}{\theta^2} < 0$$

per ogni θ, quindi l ha la concavita verso il basso su tutto Θ, questo mostra che $\hat{\theta}$ è un massimo assoluto.

GLM esponenziale I

Siano $Y_1, \ldots, Y_n$ osservazioni indipendenti con

$$p_\theta(y_i) = \rho_i \exp\{-\rho_i y_i\}$$

dove

$$\log \rho_i = \alpha + \beta x_i$$

con le x_i

- note
- non tutte uguali
- tali che $\sum_i x_i = 0$

GLM esponenziale II

Si ha

$$L(\alpha, \beta) = \prod_{i=1}^{n} \rho_i \exp\{-\rho_i y_i\}$$

e quindi

$$\begin{aligned} l(\alpha, \beta) &= \sum_{i=1}^{n} (\log \rho_i - \rho_i y_i) \\ &= \sum_{i=1}^{n} (\alpha + \beta x_i - y_i \exp\{\alpha + \beta x_i\}) \\ &= n\alpha + \beta \sum_i x_i - \exp\{\alpha\} \sum_i y_i \exp\{\beta x_i\} \\ &= n\alpha - \exp\{\alpha\} \sum_i y_i \exp\{\beta x_i\} \end{aligned}$$

GLM esponenziale III

la funzione punteggio è data dalle derivate

$$\frac{\partial}{\partial \alpha} = n - \exp\{\alpha\} \sum_i y_i \exp\{\beta x_i\}$$

$$\frac{\partial}{\partial \beta} = -\exp\{\alpha\} \sum_i x_i y_i \exp\{\beta x_i\}$$

eguagliando queste a 0 si ottiene il sistema

$$\alpha = \log\left(\frac{n}{\sum_i y_i \exp\{\beta x_i\}}\right)$$

$$0 = \sum_i x_i y_i \exp\{\beta x_i\}$$

la seconda equazione, in β, non ammette una soluzione esplicita, tuttavia possiamo mostrare che una soluzione esiste ed è unica.

GLM esponenziale IV

Definiamo $g(\beta) = \sum_i x_i y_i \exp\{\beta x_i\}$

- g è continua;
- g è monotona crescente, in quanto è somma di funzioni monotone crescenti,
$$\frac{d}{d\beta} x_i y_i \exp\{\beta x_i\} = x_i^2 y_i \exp\{\beta x_i\} > 0$$
- $\lim_{\beta \to +\infty} g(\beta) = +\infty$, infatti

$$\lim_{\beta \to +\infty} x_i y_i \exp\{\beta x_i\} = \begin{cases} +\infty & \text{se } x_i > 0 \\ 0 & \text{se } x_i \leq 0 \end{cases}$$

e siccome le x_i sono non tutte nulle e la somma è zero esiste almeno un i tale che $x_i > 0$, quindi il limite della somma è $+\infty$.

GLM esponenziale V

- $\lim_{\beta \to -\infty} g(\beta) = -\infty$, si ragiona come sopra a partire da

$$\lim_{\beta \to -\infty} x_i y_i \exp\{\beta x_i\} = \begin{cases} 0 & \text{se } x_i > 0 \\ -\infty & \text{se } x_i \leq 0 \end{cases}$$

Quindi

$$\text{esiste } \hat{\beta} \text{ tale che } g(\hat{\beta}) = 0$$

e questo è un punto stazionario per la verosimiglianza

GLM esponenziale VI

Per mostrare che è un massimo calcoliamo l'hessiano

$$H(\alpha,\beta)=\begin{bmatrix} -\exp\{\alpha\}\sum_i y_i \exp\{\beta x_i\} & -\exp\{\alpha\}\sum_i x_i y_i \exp\{\beta x_i\} \\ -\exp\{\alpha\}\sum_i x_i y_i \exp\{\beta x_i\} & -\exp\{\alpha\}\sum_i x_i^2 y_i \exp\{\beta x_i\} \end{bmatrix}$$

il minore di nord ovest è negativo (si ricordi che $y_i \geq 0$), il determinante è

$$|H(\alpha,\beta)| = \exp\{2\alpha\}\left(\sum_i y_i \exp\{\beta x_i\}\sum_i x_i^2 y_i \exp\{\beta x_i\} - \left(\sum_i x_i y_i \exp\{\beta x_i\}\right)^2\right)$$

se calcoliamo il determinante in corrispondenza a $(\hat{\alpha},\hat{\beta})$ si ha $\sum_i x_i y_i \exp\{\beta x_i\} = 0$ e quindi

$$|H(\hat{\alpha},\hat{\beta})| = \exp\{2\hat{\alpha}\}\exp\{-\hat{\alpha}\}\, n\sum_i x_i^2 y_i \exp\left\{\hat{\beta} x_i\right\} > 0$$

Indice

Algoritmo di Newton-Raphson e Fisher-scoring

L'algoritmo di Newton-Raphson è un metodo numerico per trovare lo zero di una funzione sapendo che questo esiste ed è unico e che la funzione è derivabile.
Si procede nel modo seguente

0 si sceglie a caso un punto x_0;
1 si calcola $x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}$

$\vdots$

m si calcola $x_m = x_{m-1} - \frac{g(x_{m-1})}{g'(x_{m-1})}$

La regola d'arresto è del tipo

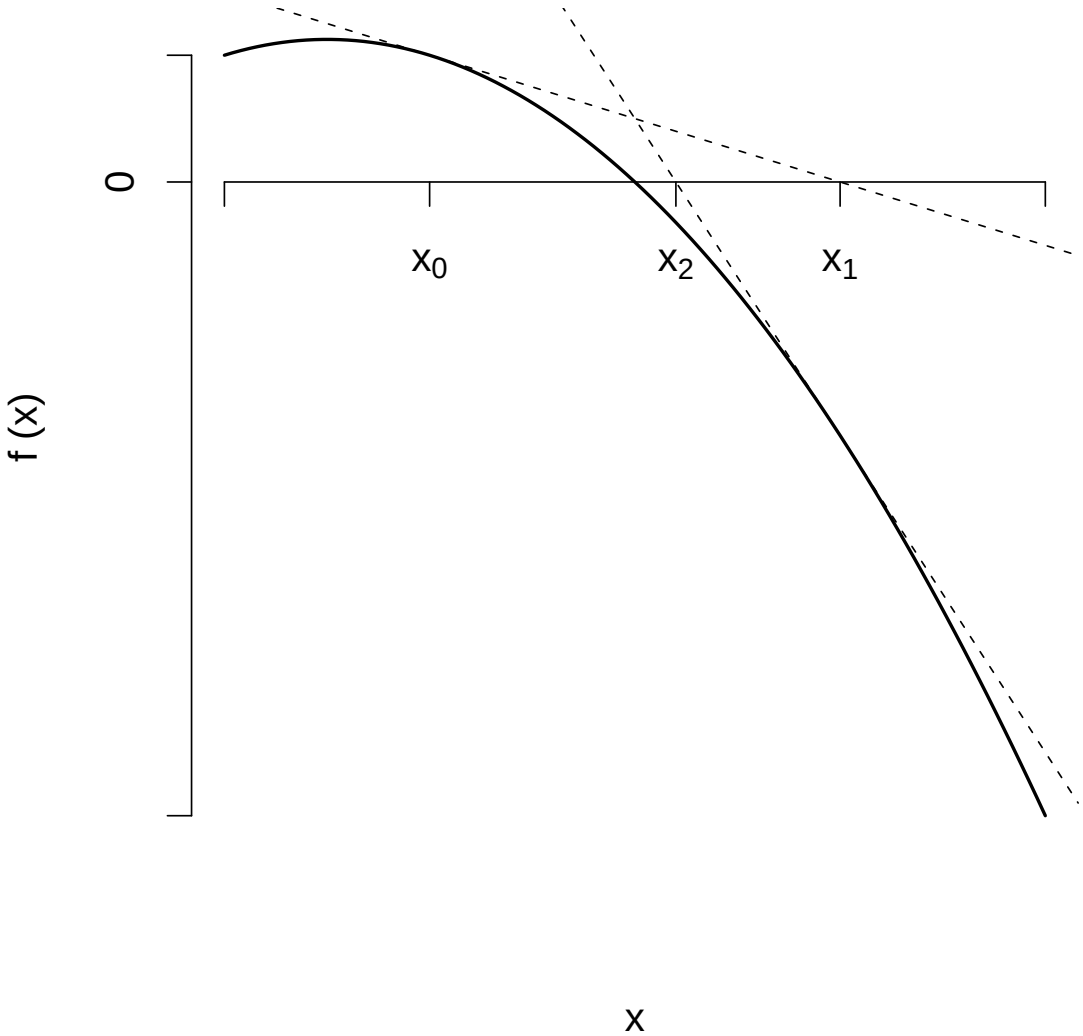
$$|x_m - x_{m-1}| < \varepsilon$$

e/o

$$|g(x_m)| < \eta$$

con ε e/o η parametri di tolleranza fissati.

Primi passi dell'algoritmo di Newton-Raphson



La logica del procedimento è che al passo m si cerca lo zero dell'approssimante lineare di g in x_{m-1}

$$g(x) \approx g(x_{m-1}) + g'(x_{m-1})(x - x_{m-1})$$

che è, appunto

$$x_m = x_{m-1} - \frac{g(x_{m-1})}{g'(x_{m-1})}.$$

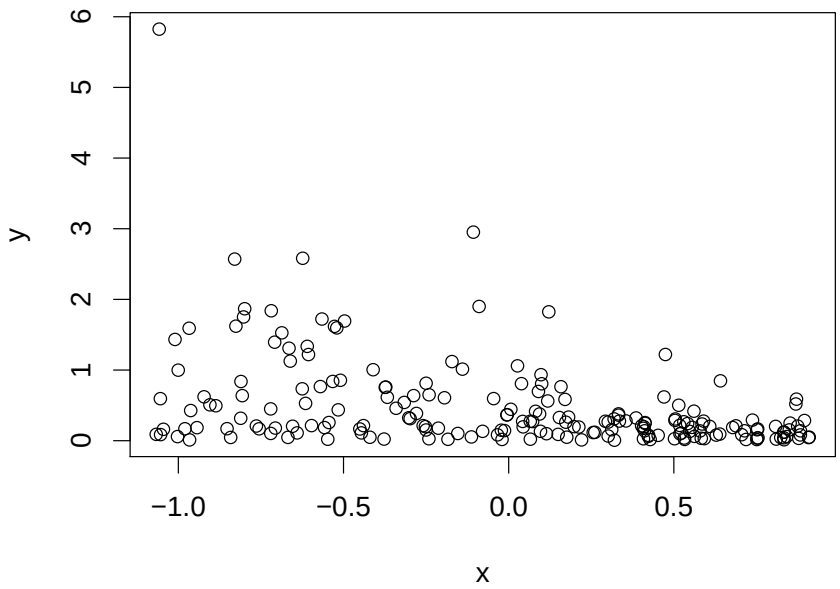
Esempio con GLM esponenziale

Nel caso del GLM esponenziale dobbiamo trovare lo 0 di

$$g(\beta) = \sum_i x_i y_i \exp\{\beta x_i\}$$

Si ha

$$g'(\beta) = \sum_i x_i^2 y_i \exp\{\beta x_i\}$$



Abbiamo un campione di 200 coppie (x_i, Y_i), rappresentato in figura

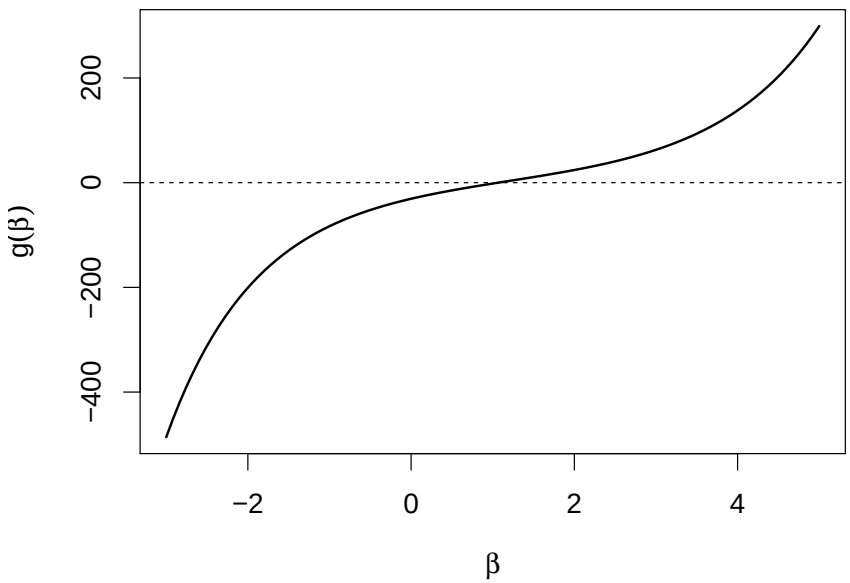
Esempio con GLM esponenziale

Nel caso del GLM esponenziale dobbiamo trovare lo 0 di

$$g(\beta) = \sum_i x_i y_i \exp\{\beta x_i\}$$

Si ha

$$g'(\beta) = \sum_i x_i^2 y_i \exp\{\beta x_i\}$$



Con queste, calcoliamo la funzione $g(\beta)$, di cui cerchiamo gli zeri.

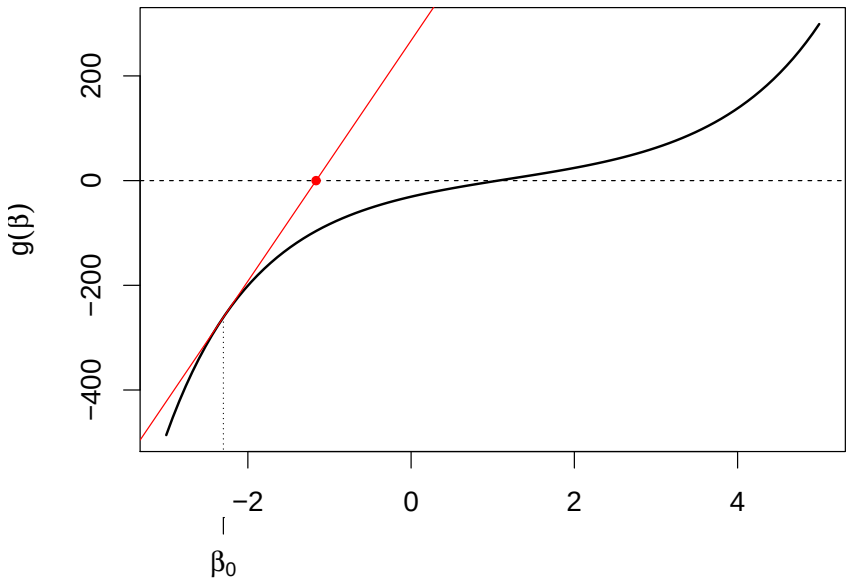
Esempio con GLM esponenziale

Nel caso del GLM esponenziale dobbiamo trovare lo 0 di

$$g(\beta) = \sum_i x_i y_i \exp\{\beta x_i\}$$

Si ha

$$g'(\beta) = \sum_i x_i^2 y_i \exp\{\beta x_i\}$$



Partiamo da un punto arbitrario, sia $\beta_0 = -2.3$, in corrispondenza ad esso approssimiamo la funzione g con la retta di pendenza $g'(\beta_0)$ (in rosso), che si annulla in

$$\beta_1 = \beta_0 - \frac{g(\beta_0)}{g'(\beta_0)} = -1.16$$

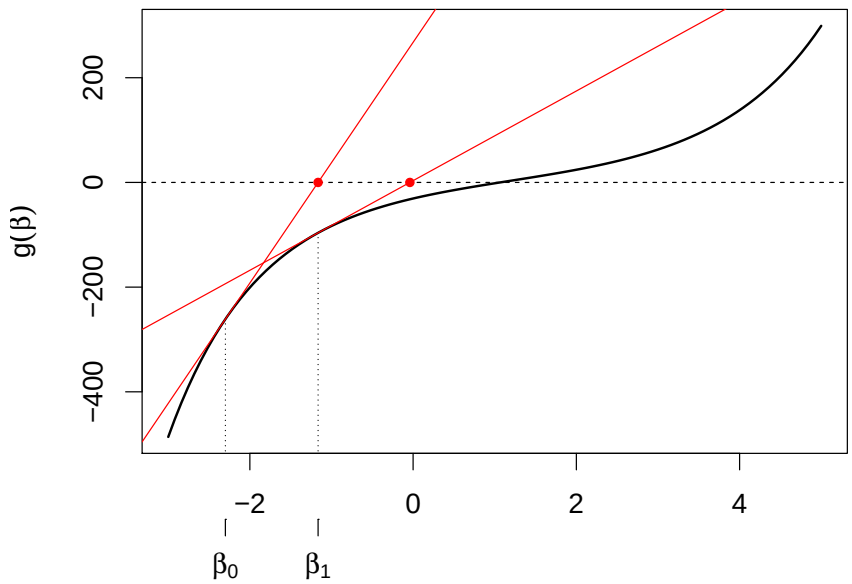
Esempio con GLM esponenziale

Nel caso del GLM esponenziale dobbiamo trovare lo 0 di

$$g(\beta) = \sum_i x_i y_i \exp\{\beta x_i\}$$

Si ha

$$g'(\beta) = \sum_i x_i^2 y_i \exp\{\beta x_i\}$$



Partiamo da $\beta_1 = -1.16$ e approssimiamo la funzione g con la retta di pendenza $g'(\beta_1)$ (in rosso), che si annulla in

$$\beta_2 = \beta_1 - \frac{g(\beta_1)}{g'(\beta_1)} = -0.039$$

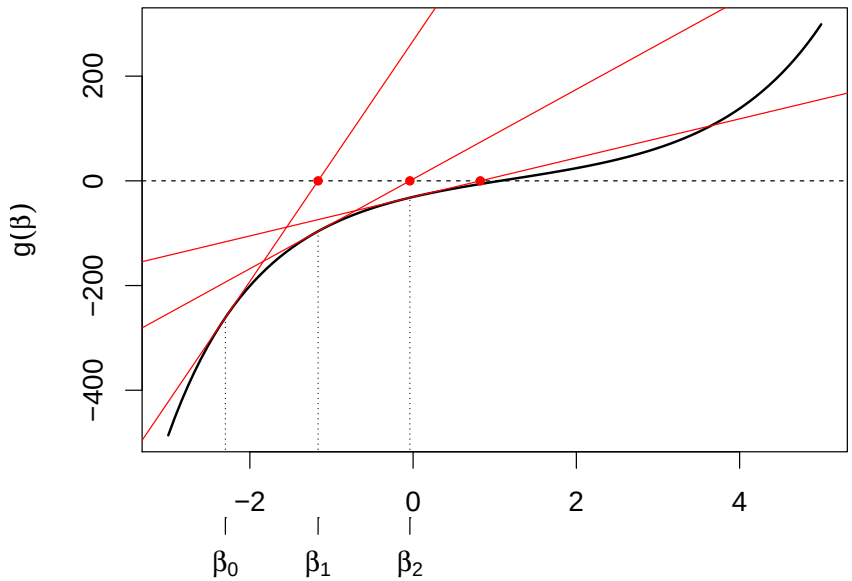
Esempio con GLM esponenziale

Nel caso del GLM esponenziale dobbiamo trovare lo 0 di

$$g(\beta) = \sum_i x_i y_i \exp\{\beta x_i\}$$

Si ha

$$g'(\beta) = \sum_i x_i^2 y_i \exp\{\beta x_i\}$$



Itero la procedura

m	β_m	$g(\beta_m)$
0	-2.30000000	-261.12575464
1	-1.16409572	-96.29197400
2	-0.03954459	-32.16339125
3	0.82351485	-6.08112329
4	1.06090494	-0.12447315
5	1.06594208	-0.00002535
6	1.06594311	-1.02×10^{-12}

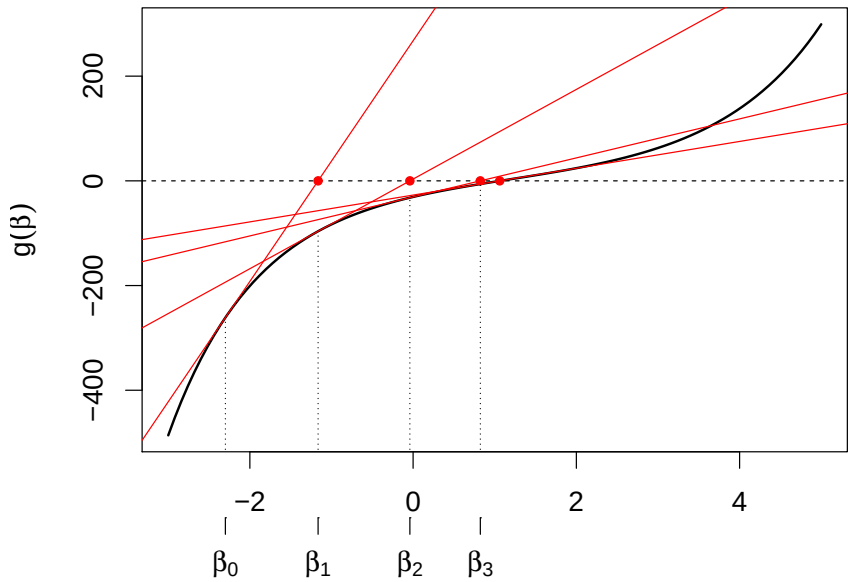
Esempio con GLM esponenziale

Nel caso del GLM esponenziale dobbiamo trovare lo 0 di

$$g(\beta) = \sum_i x_i y_i \exp\{\beta x_i\}$$

Si ha

$$g'(\beta) = \sum_i x_i^2 y_i \exp\{\beta x_i\}$$



Itero la procedura

m	β_m	$g(\beta_m)$
0	-2.30000000	-261.12575464
1	-1.16409572	-96.29197400
2	-0.03954459	-32.16339125
3	0.82351485	-6.08112329
4	1.06090494	-0.12447315
5	1.06594208	-0.00002535
6	1.06594311	-1.02×10^{-12}

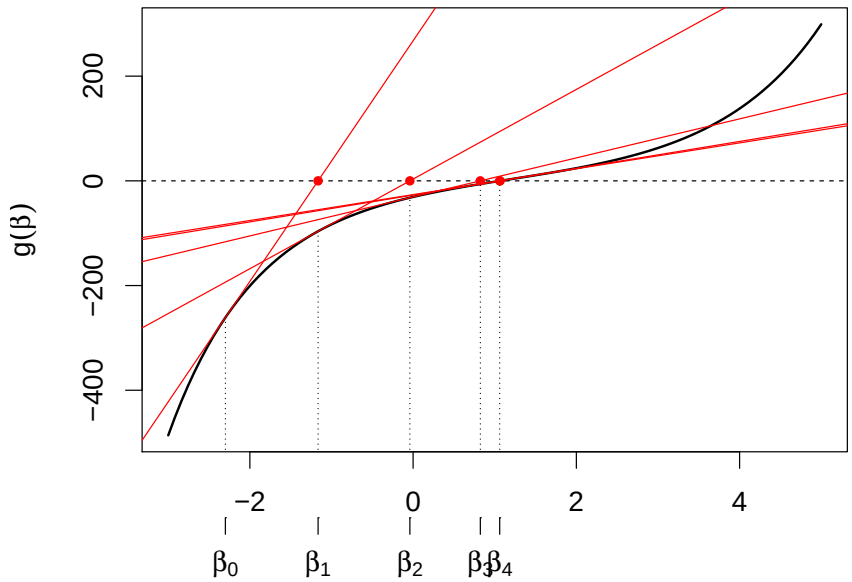
Esempio con GLM esponenziale

Nel caso del GLM esponenziale dobbiamo trovare lo 0 di

$$g(\beta) = \sum_i x_i y_i \exp\{\beta x_i\}$$

Si ha

$$g'(\beta) = \sum_i x_i^2 y_i \exp\{\beta x_i\}$$



Itero la procedura

m	β_m	$g(\beta_m)$
0	-2.30000000	-261.12575464
1	-1.16409572	-96.29197400
2	-0.03954459	-32.16339125
3	0.82351485	-6.08112329
4	1.06090494	-0.12447315
5	1.06594208	-0.00002535
6	1.06594311	-1.02×10^{-12}

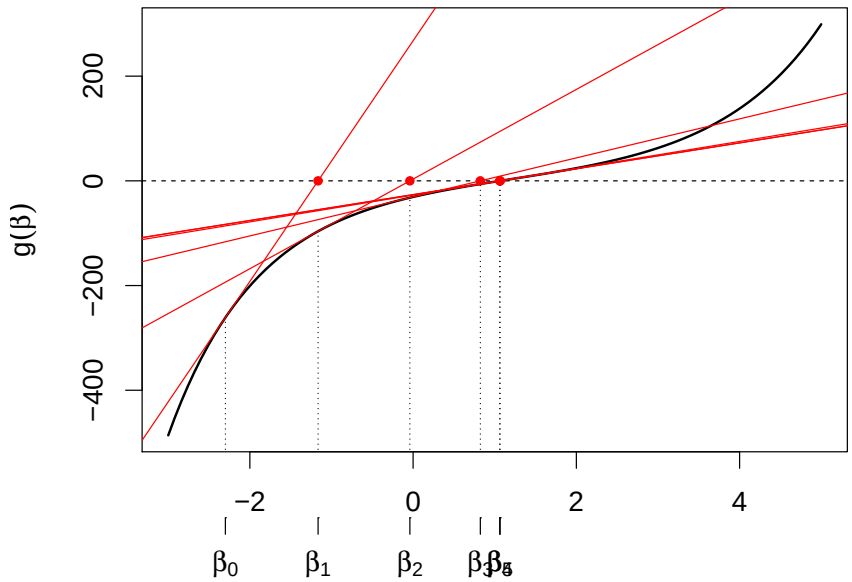
Esempio con GLM esponenziale

Nel caso del GLM esponenziale dobbiamo trovare lo 0 di

$$g(\beta) = \sum_i x_i y_i \exp\{\beta x_i\}$$

Si ha

$$g'(\beta) = \sum_i x_i^2 y_i \exp\{\beta x_i\}$$



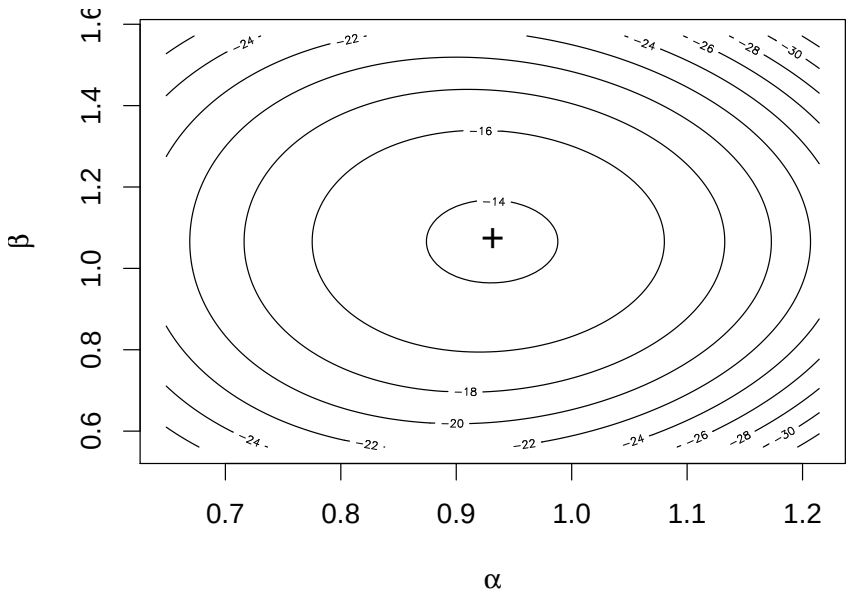
Itero la procedura

m	β_m	$g(\beta_m)$
0	-2.30000000	-261.12575464
1	-1.16409572	-96.29197400
2	-0.03954459	-32.16339125
3	0.82351485	-6.08112329
4	1.06090494	-0.12447315
5	1.06594208	-0.00002535
6	1.06594311	-1.02×10^{-12}

Esempio con GLM esponenziale

Si ha quindi $\hat{\beta} = 1.06594311$ e

$$\hat{\alpha} = \log\left(\frac{n}{\sum_i y_i \exp\left\{\hat{\beta} x_i\right\}}\right) = \log\left(\frac{200}{78.78251}\right) = 0.9316264$$



Linee di livello della verosimiglianza per α e β, la stima è rappresentata dal $+$.

N-R per la ricerca dello SMV

L'obiettivo è trovare lo zero della funzione punteggio $l_*(\cdot)$, si avrà

$$\hat{\theta}_m = \hat{\theta}_{m-1} - \frac{l_*(\hat{\theta}_{m-1})}{l''(\hat{\theta}_{m-1})} = \hat{\theta}_{m-1} + J(\hat{\theta}_{m-1})^{-1} l_*(\hat{\theta}_{m-1}) =$$

dove J rappresenta l'opposto della derivata seconda di l.

Algoritmo di Fisher scoring

Una variante dell'algoritmo di Newton-Raphson, l'algoritmo di Fisher scoring, prevede di sostituire l'informazione osservata con l'informazione attesa, sicché la formula di aggiornamento diviene

$$\hat{\theta}_m = \hat{\theta}_{m-1} - \frac{l_*(\hat{\theta}_{m-1})}{l''(\hat{\theta}_{m-1})} = \hat{\theta}_{m-1} + \mathcal{I}(\hat{\theta}_{m-1})^{-1} l_*(\hat{\theta}_{m-1}) = \tag{2}$$

dove

$$\mathcal{I}(\theta) = E(-l''(\theta))$$

è l'informazione attesa o informazione di Fisher (di cui parleremo più avanti).

Generalizzazione

L'algoritmo di Newton-Raphson (e analogamente quello di Fisher scoring) può essere generalizzato al caso di $g : \mathbb{R}^d \to \mathbb{R}$, in tal caso la formula di aggiornamento è

$$\mathbf{x}_m = \mathbf{x}_{m-1} - \left(\frac{d}{d\mathbf{x}^T} g(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_{m-1}} \right)^{-1} g(\mathbf{x}_{m-1})$$

Ancora, si può generalizzare al caso di sistemi di equazioni, cioè $g : \mathbb{R}^d \to \mathbb{R}^d$

$$\mathbf{x}_m = \mathbf{x}_{m-1} - \left[\frac{d}{dx_j} g_i(\mathbf{x}) \Big|_{\mathbf{x}=\mathbf{x}_{m-1}} \right]^{-1} g(\mathbf{x}_{m-1})$$

Convergenza dell'algoritmo, risultati teorici

Indichiamo con $\tilde{x}$ la radice di $g(\cdot)$ e con $x_0, x_1, \ldots, x_n, \ldots$ la successione generata dall'algoritmo di Newton-Raphson a partire da x_0.

Si hanno condizioni sufficienti per la convergenza

- Se g ha derivata continua e $g'(\tilde{x}) \neq 0$ allora esiste $\nu > 0$ tale che se $|x_0 - \tilde{x}| < \nu$ allora $x_n \to \tilde{x}$
- Se g è differenziabile due volte in $[\tilde{x}, \nu]$ $(\nu > \tilde{x})$ e $g'(x) \neq 0$, $g(x)g''(x) > 0$ per ogni $x \in [\tilde{x}, \nu]$, allora se $x_0 \in [\tilde{x}, \nu]$, $x_n \to \tilde{x}$.

La prima è poco restrittiva ma locale, la seconda è globale ma troppo restrittiva.

Convergenza dell'algoritmo, in pratica

In pratica, si impiega l'algoritmo di N-R senza avere la certezza che converga né che, qualora converga, quella sia effettivamente una soluzione.

Accorgimenti

- ripetere l'algoritmo partendo da punti diversi
- studiare la funzione (per verificare la plausibilità della soluzione trovata)
- diagnosi del modello
- riparametrizzare il modello può essere utile

Si noti anche che l'algoritmo viene impiegato anche senza avere la garanzia che la soluzione sia unica, in quel caso può convergere a qualunque delle soluzioni (massimi locali).

Esempio: SMV per via numerica I

Sia $y_1, \ldots, y_n \sim IID(Gumbel(0, \theta)$, cioè

$$f(y_i; \theta) = \theta e^{-y\theta} \exp\left\{-e^{-y\theta}\right\}$$

quindi

$$L(\theta) = \prod_{i=1}^{n} \theta e^{-y_i\theta} \exp\left\{-e^{-y_i\theta}\right\} = \theta^n e^{-\theta \sum_{i=1}^{n} y_i} \exp\left\{-\sum_{i=1}^{n} e^{-y_i\theta}\right\}$$

da cui la log-verosimiglianza

$$l(\theta) = n \log \theta - \theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{-y_i\theta}$$

Esempio: SMV per via numerica II

e la funzione punteggio

$$l_*(\theta) = n\theta^{-1} - \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i e^{-\theta y_i}$$

le radici non sono ricavabili per via analitica, si calcola allora

$$l''(\theta) = -n\theta^{-2} - \sum_{i=1}^{n} y_i^2 e^{-\theta y_i}$$

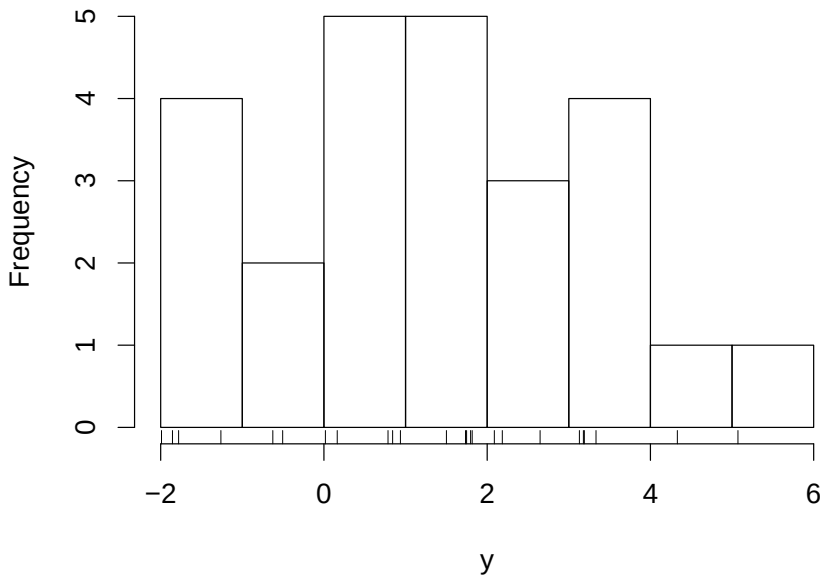
Notiamo che

- $l''(\theta) < 0$ quindi gli eventuali zeri della funzione punteggio sono massimi,
- $\lim_{\theta \to 0} l(\theta) = \lim_{\theta \to +\infty} l(\theta) = -\infty$, l è continua con derivata continua quindi esiste almeno uno zero.

Si può allora operare via algoritmo di N-R.

Esempio Gumbel parametro di scale

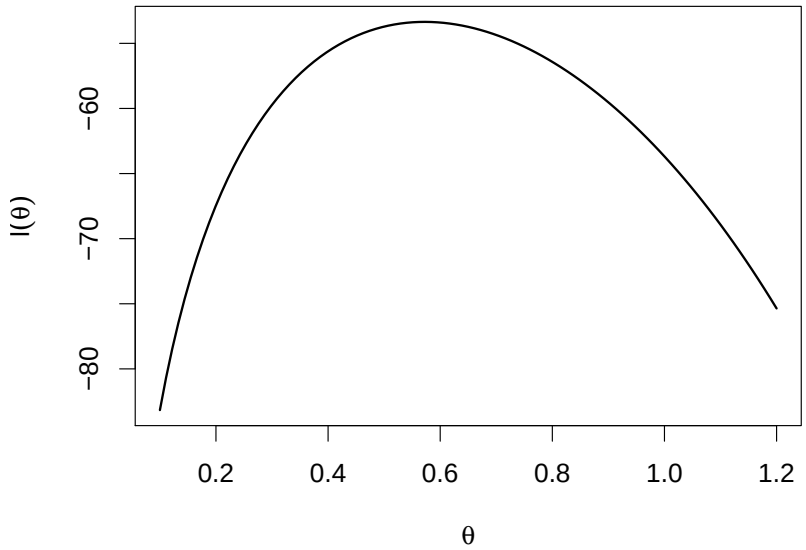
Consideriamo un campione di numerosità 25 per la Gumbel$(0, \theta)$



Esempio Gumbel parametro di scale

Consideriamo un campione di numerosità 25 per la Gumbel$(0, \theta)$

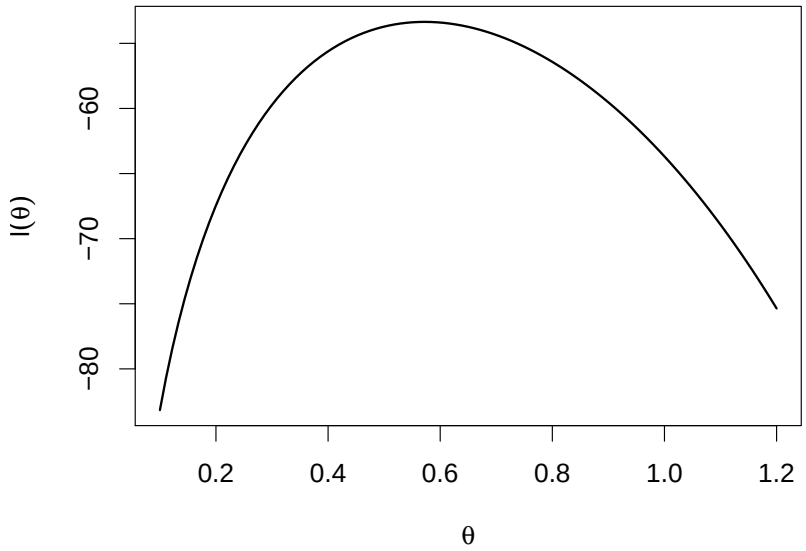
$$l(\theta) = n \log \theta - \theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{-y_i \theta}$$



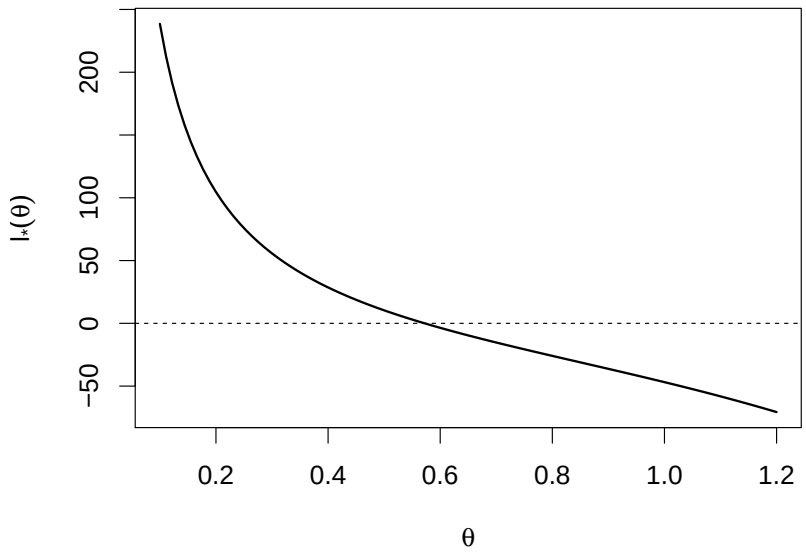
Esempio Gumbel parametro di scale

Consideriamo un campione di numerosità 25 per la Gumbel$(0, \theta)$

$$l(\theta) = n \log \theta - \theta \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} e^{-y_i \theta}$$



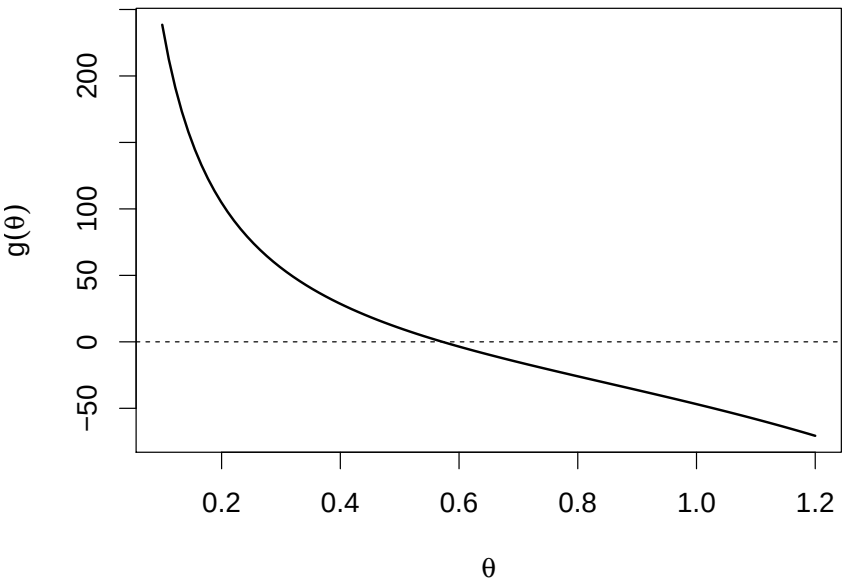
$$l_*(\theta) = n\theta^{-1} - \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i e^{-\theta y_i}$$



Soluzione via Newton-Raphson

L'aggiornamento è dato da

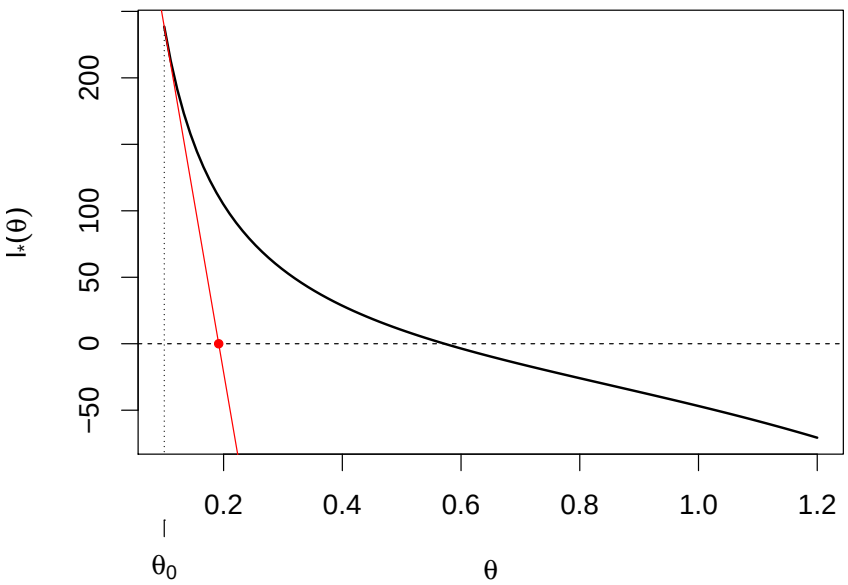
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-n\theta_{m-1}^{-2} - \sum_{i=1}^n y_i^2 e^{-\theta_{m-1} y_i}}$$



Soluzione via Newton-Raphson

L'aggiornamento è dato da

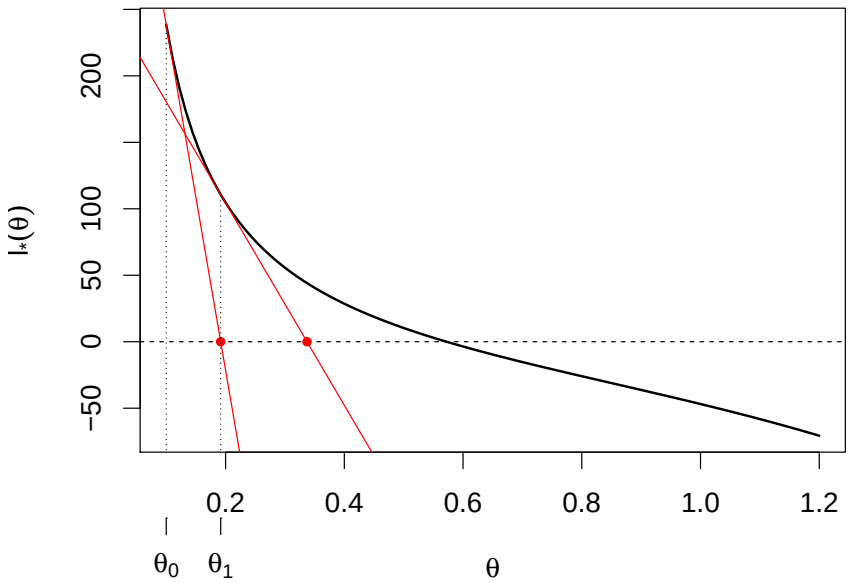
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-n\theta_{m-1}^{-2} - \sum_{i=1}^n y_i^2 e^{-\theta_{m-1} y_i}}$$



Soluzione via Newton-Raphson

L'aggiornamento è dato da

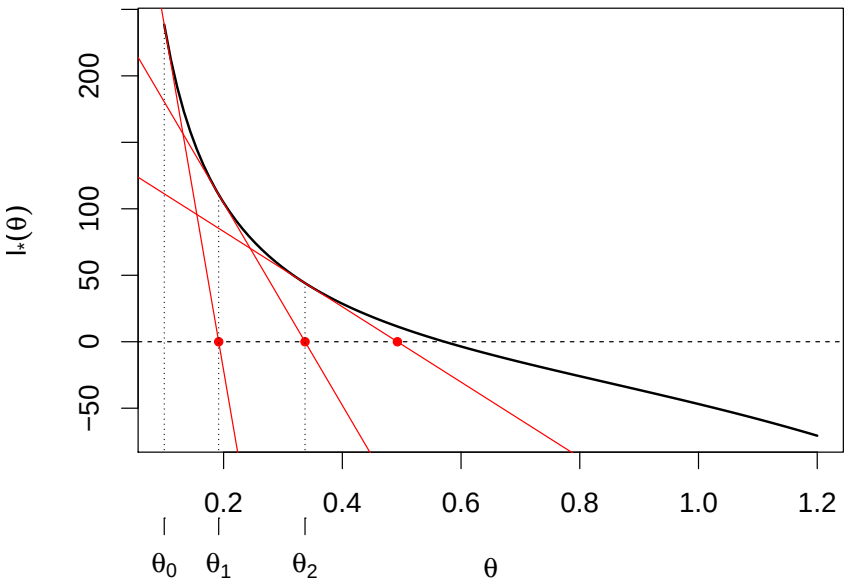
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-n\theta_{m-1}^{-2} - \sum_{i=1}^n y_i^2 e^{-\theta_{m-1} y_i}}$$



Soluzione via Newton-Raphson

L'aggiornamento è dato da

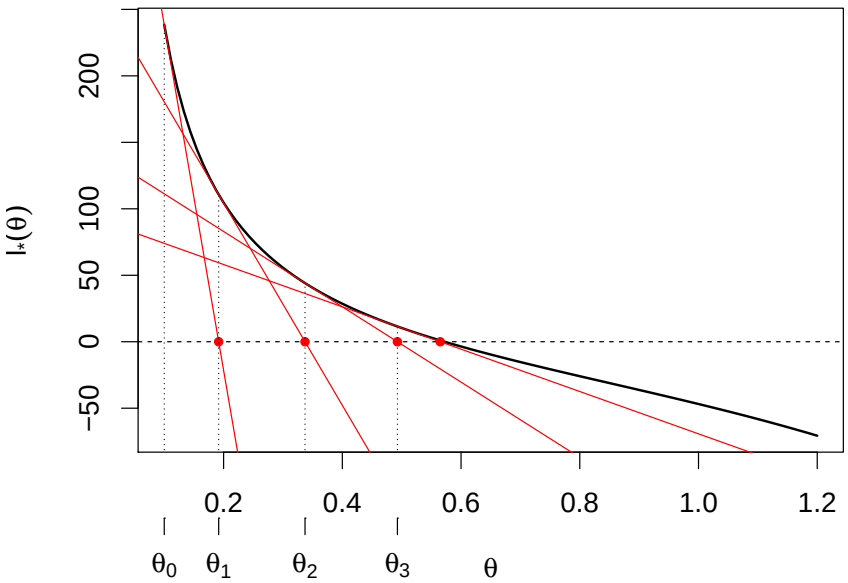
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-n\theta_{m-1}^{-2} - \sum_{i=1}^n y_i^2 e^{-\theta_{m-1} y_i}}$$



Soluzione via Newton-Raphson

L'aggiornamento è dato da

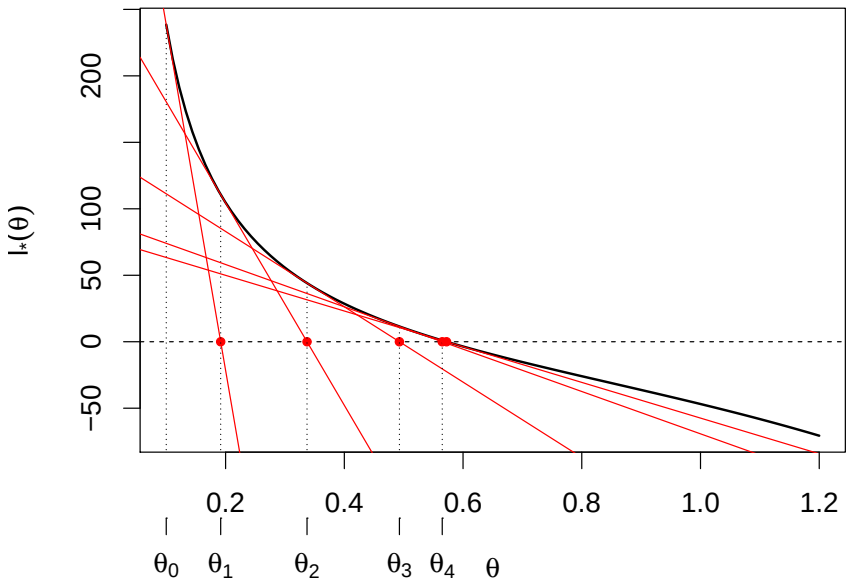
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-n\theta_{m-1}^{-2} - \sum_{i=1}^n y_i^2 e^{-\theta_{m-1} y_i}}$$



Soluzione via Newton-Raphson

L'aggiornamento è dato da

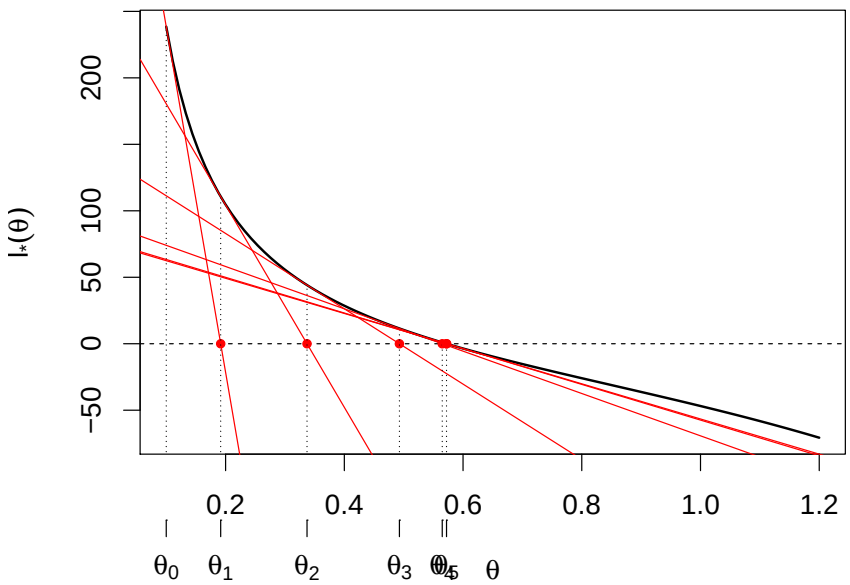
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-n\theta_{m-1}^{-2} - \sum_{i=1}^n y_i^2 e^{-\theta_{m-1} y_i}}$$



Soluzione via Newton-Raphson

L'aggiornamento è dato da

$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-n\theta_{m-1}^{-2} - \sum_{i=1}^n y_i^2 e^{-\theta_{m-1} y_i}}$$



m	θ_m	$l_*(\theta_m)$
0	0.10000000	238.52215883
1	0.19173669	110.66754155
2	0.33717220	44.08974160
3	0.49279281	11.48675921
4	0.56493467	0.97189526
5	0.57217524	0.00700918
6	0.57222822	0.00000036

Soluzione via Fisher-Scoring

Calcoliamo anzitutto

$$E(-l''(\theta)) = +n\theta^{-2} + nE\left(y^2 e^{-\theta y}\right)$$

dove il valore atteso si ottiene via integrazione per parti

$$\begin{aligned} M &= E\left(y^2 e^{-\theta y}\right) = \int_{-\infty}^{+\infty} y^2 e^{-\theta y} f(y;\theta)dy \\ &= \left[y^2 e^{-\theta y} F(y;\theta)\right]_{-\infty}^{+\infty} - \frac{2}{\theta}\int y\theta e^{-\theta y} F(y;\theta)dy + \int y^2 \theta e^{-\theta y} F(y;\theta)dy \\ &= -\frac{2}{\theta}\int y f(y;\theta)dy + \int y^2 f(y;\theta)dy \\ &= -2\frac{\gamma}{\theta^2} + \left(\frac{\pi^2}{6} + \gamma^2\right)\frac{1}{\theta^2} \end{aligned}$$

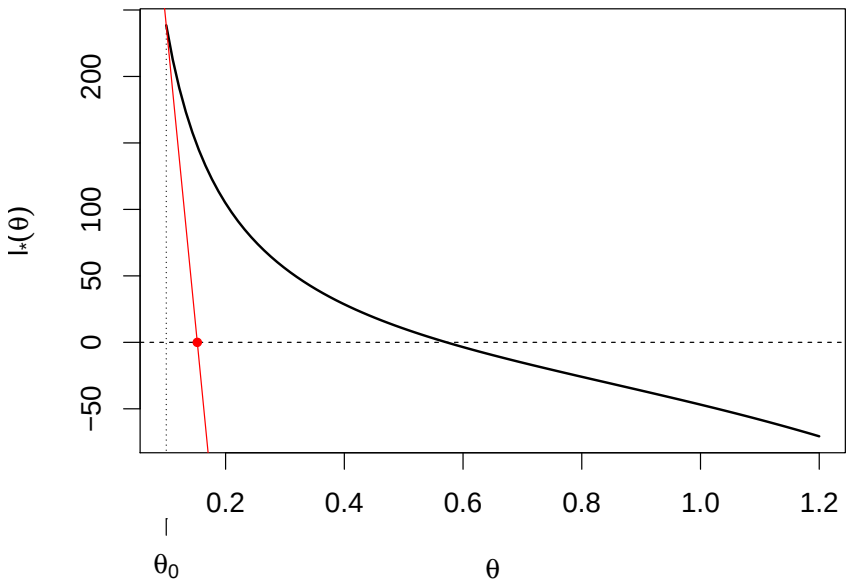
dove si è usato

- $\theta e^{-\theta y} F(y;\theta) = f(y;\theta)$
- $E(Y) = \gamma/\theta$ e $V(Y) = \pi^2/(6\theta^2)$ (γ=costante di Eulero).

Algoritmo di Fisher-scoring

L'aggiornamento è dato da

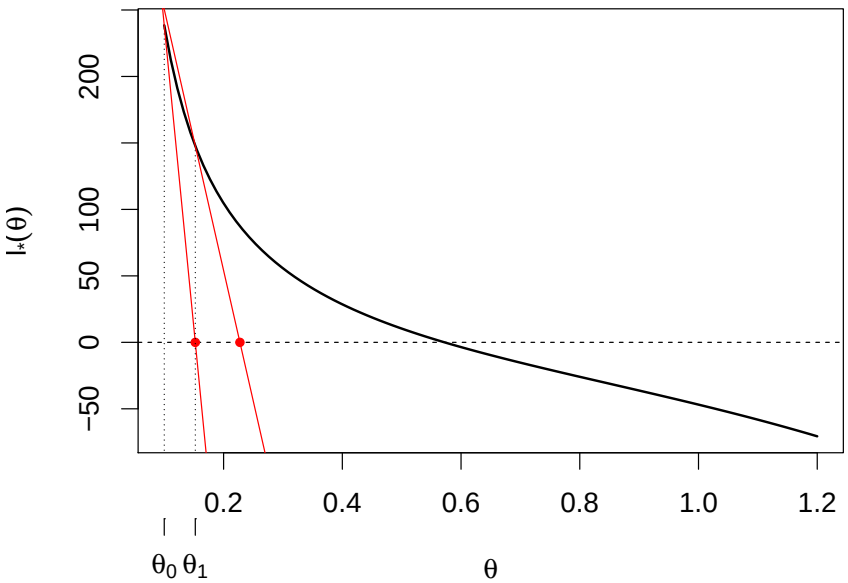
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-\frac{n}{\theta_{m-1}^2}\left(1 - 2\gamma + \gamma^2 + \frac{\pi^2}{6}\right)}$$



Algoritmo di Fisher-scoring

L'aggiornamento è dato da

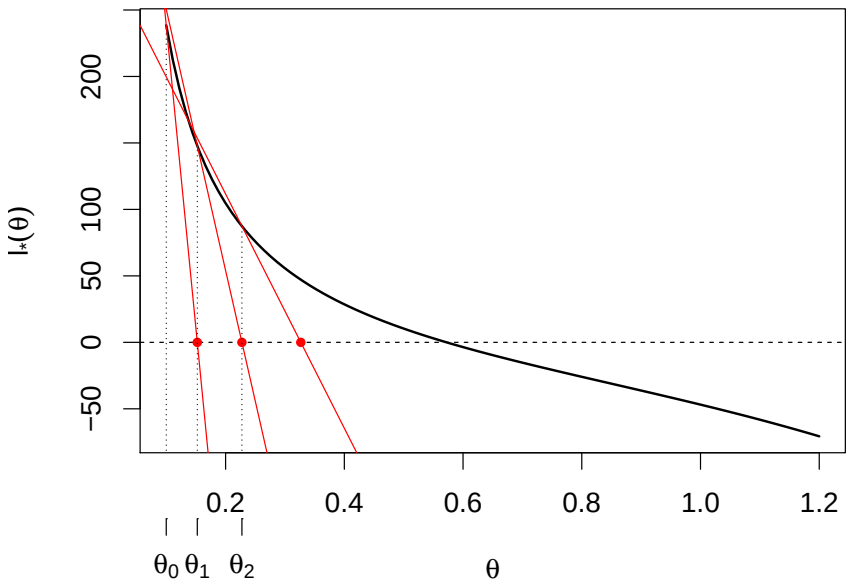
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-\frac{n}{\theta_{m-1}^2}\left(1 - 2\gamma + \gamma^2 + \frac{\pi^2}{6}\right)}$$



Algoritmo di Fisher-scoring

L'aggiornamento è dato da

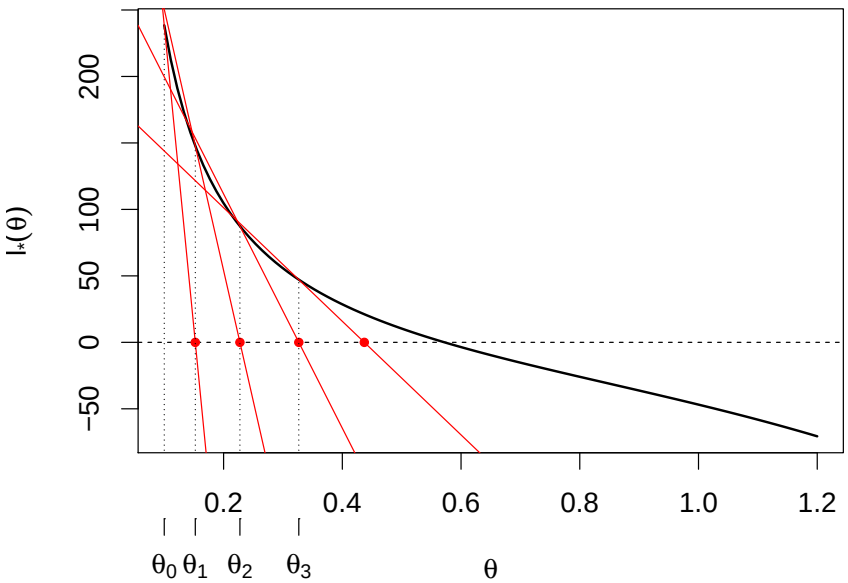
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-\frac{n}{\theta_{m-1}^2}\left(1 - 2\gamma + \gamma^2 + \frac{\pi^2}{6}\right)}$$



Algoritmo di Fisher-scoring

L'aggiornamento è dato da

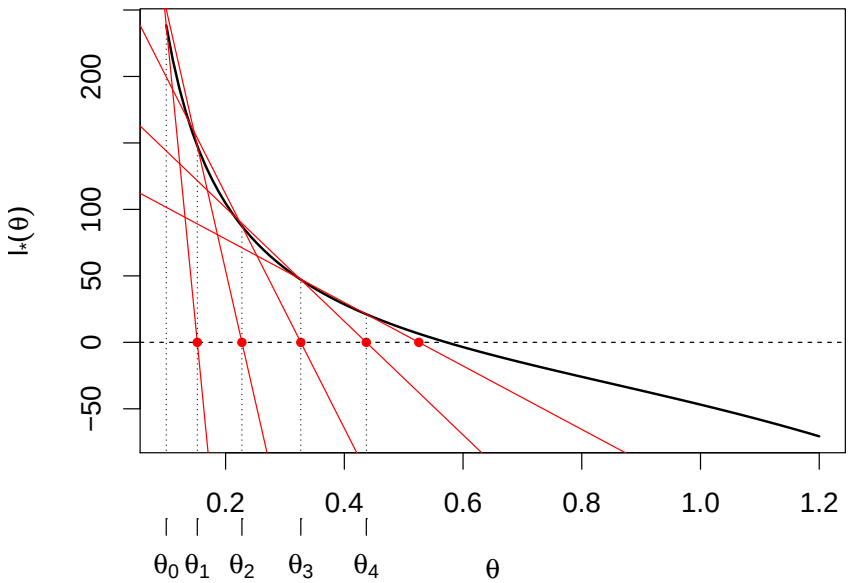
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-\frac{n}{\theta_{m-1}^2}\left(1 - 2\gamma + \gamma^2 + \frac{\pi^2}{6}\right)}$$



Algoritmo di Fisher-scoring

L'aggiornamento è dato da

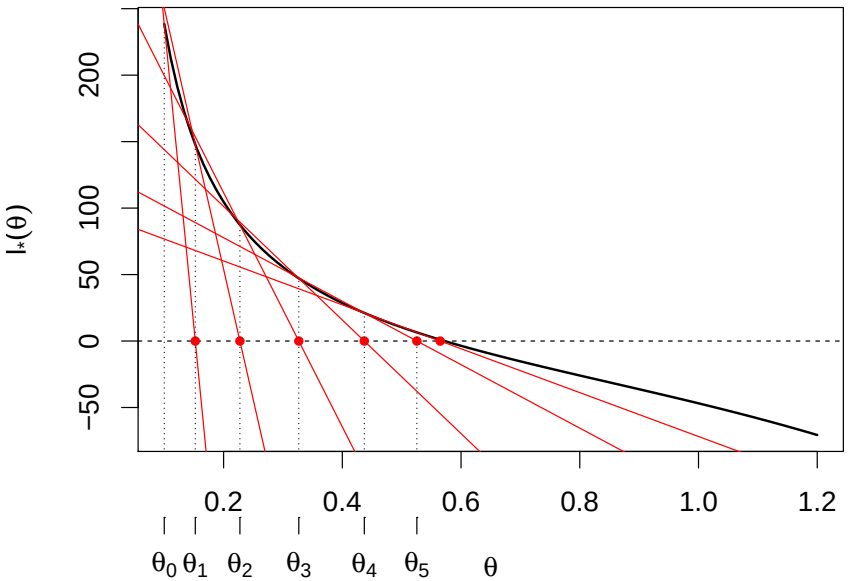
$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-\frac{n}{\theta_{m-1}^2}\left(1 - 2\gamma + \gamma^2 + \frac{\pi^2}{6}\right)}$$



Algoritmo di Fisher-scoring

L'aggiornamento è dato da

$$\theta_m = \theta_{m-1} - \frac{l_*(\theta_{m-1})}{l''(\theta_{m-1})} = \theta_{m-1} - \frac{n\theta_{m-1}^{-1} - \sum_{i=1}^n y_i + \sum_{i=1}^n y_i e^{-\theta_{m-1} y_i}}{-\frac{n}{\theta_{m-1}^2}\left(1 - 2\gamma + \gamma^2 + \frac{\pi^2}{6}\right)}$$



m	θ_m	$l_*(\theta_m)$
0	0.10000000	238.52215883
1	0.15231665	147.74091471
2	0.22749742	87.38063168
3	0.32669003	47.13789415
4	0.43703514	21.13297429
5	0.52556801	6.48563354
6	0.56486158	0.98170723
7	0.57173191	0.06569195
8	0.57220289	0.00335126
9	0.57222696	0.00016727
10	0.57222816	0.00000834
11	0.57222822	0.00000042
12	0.57222822	0.00000002
13	0.57222822	0.00000000

Esempio: distribuzione gamma I

Sia $(Y_1, \ldots, Y_n) \sim IID\text{Gamma}(a, b)$,

$$p_\theta(y) = \frac{b^a}{\Gamma(a)} y^{a-1} \exp\{-by\}$$

$\mathcal{Y} = \mathbb{R}_+^n$ e $\Theta = \mathbb{R}_+^2$.
La verosimiglianza è

$$L(\theta) = \frac{b^{na}}{\Gamma(a)^n} \left(\prod_{i=1}^n y_i \right)^{a-1} \exp\left\{ -b \sum_{i=1}^n y_i \right\},$$

la log verosimiglianza è

$$l(\theta) = na \log b - n \log \Gamma(a) + (a-1) \sum_i \log y_i - b \sum_i y_i$$

Esempio: distribuzione gamma II

e quindi si ha la funzione punteggio

$$\frac{\partial l}{\partial a} = n \log b - n \left(\frac{d \log \Gamma(a)}{da} \right) + \sum_i \log y_i$$

$$\frac{\partial l}{\partial b} = \frac{na}{b} - \sum_i y_i$$

che è poco trattabile.

Esempio: distribuzione gamma III

Conviene in questo caso riparametrizzare la densità gamma come

$$f(y;\mu,\omega) = \frac{1}{\Gamma(\omega)}\left(\frac{\omega}{\mu}\right)^{\omega} y^{\omega-1} e^{-\frac{\omega y}{\mu}}$$

dove $\mu, \omega \in \mathbb{R}^+$. La funzione di log-verosimiglianza è

$$l(\omega,\mu) = -\frac{\omega}{\mu}\sum_i y_i + (\omega-1)\sum_i \log y_i + n\omega(\log\omega - \log\mu) - n\log\Gamma(\omega)$$

Si verifica che tende a $-\infty$ sulla frontiera, cioè si verifica che sono pari a $-\infty$ i limiti

$$\lim_{\mu\to\infty} l(\omega,\mu) \qquad \lim_{\omega\to\infty} l(\omega,\mu)$$
$$\lim_{\mu\to 0} l(\omega,\mu) \qquad \lim_{\omega\to 0} l(\omega,\mu)$$
$$\lim_{\omega\to 0,\mu\to 0} l(\omega,\mu) \qquad \lim_{\omega\to\infty,\mu\to\infty} l(\omega,\mu).$$

Esempio: distribuzione gamma IV

Ne segue che esiste un massimo ed è interno, questo è pertanto soluzione del sistema

$$\begin{cases} \frac{\partial l}{\partial \mu} &= & \frac{\omega \sum_i y_i}{\mu^2} - \frac{n\omega}{\mu} & = & 0 \\ \frac{\partial l}{\partial \omega} &= & -\frac{\sum_i y_i}{\mu} + \sum_i \log y_i + n(\log \omega + 1 - \log \mu) - n\psi(\omega) & = & 0 \end{cases}$$

dove $\psi(x) = \frac{d}{dx} \log \Gamma(x)$.

Esempio: distribuzione gamma V

Si ottiene facilmente lo SMV per μ

$$\mu : \widehat{\mu} = \frac{1}{n}\sum_i y_i$$

lo SMV di ω è la soluzione di

$$g(\omega) = \sum_i \log y_i + n(\log\omega - \log\widehat{\mu}) - n\psi(\mu) = 0$$

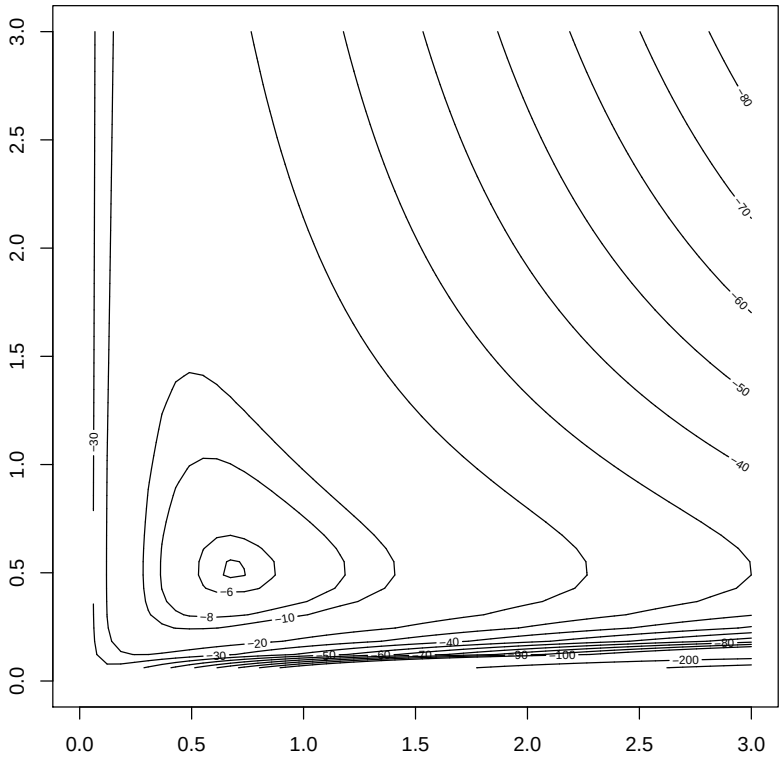
che va ottenuta per via numerica.
L'hessiano della log verosimiglianza è

$$\begin{pmatrix} -\frac{2\omega\sum_i y_i}{\mu^3} + \frac{n\omega}{\mu^2} & \frac{\sum_i y_i}{\mu^2} - \frac{n}{\mu} \\ \frac{\sum_i y_i}{\mu^2} - \frac{n}{\mu} & \frac{n}{\omega} - n\psi'(\omega) \end{pmatrix}$$

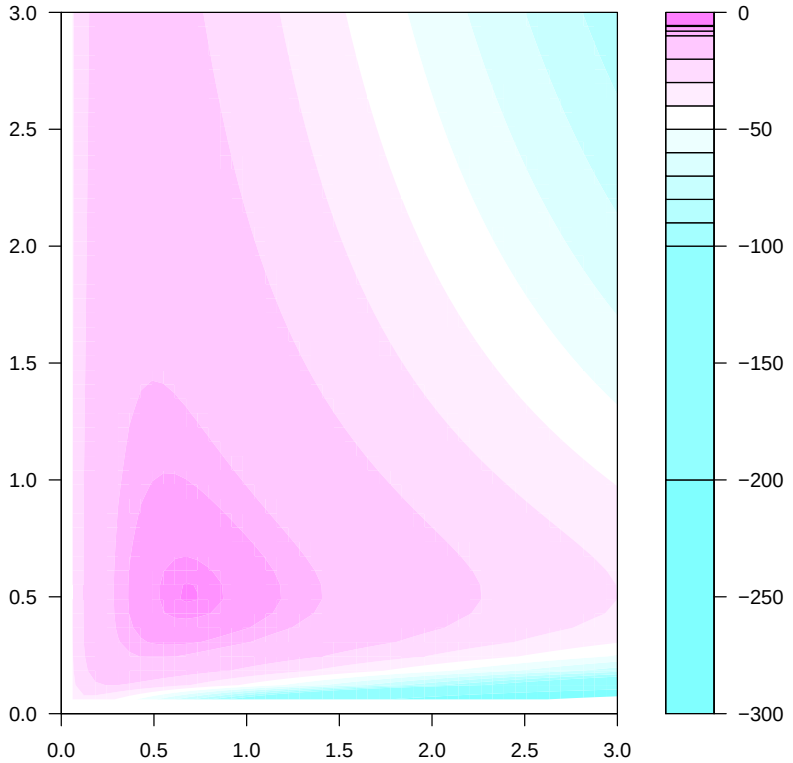
da questo si ottengano l'informazione osservata e attesa.

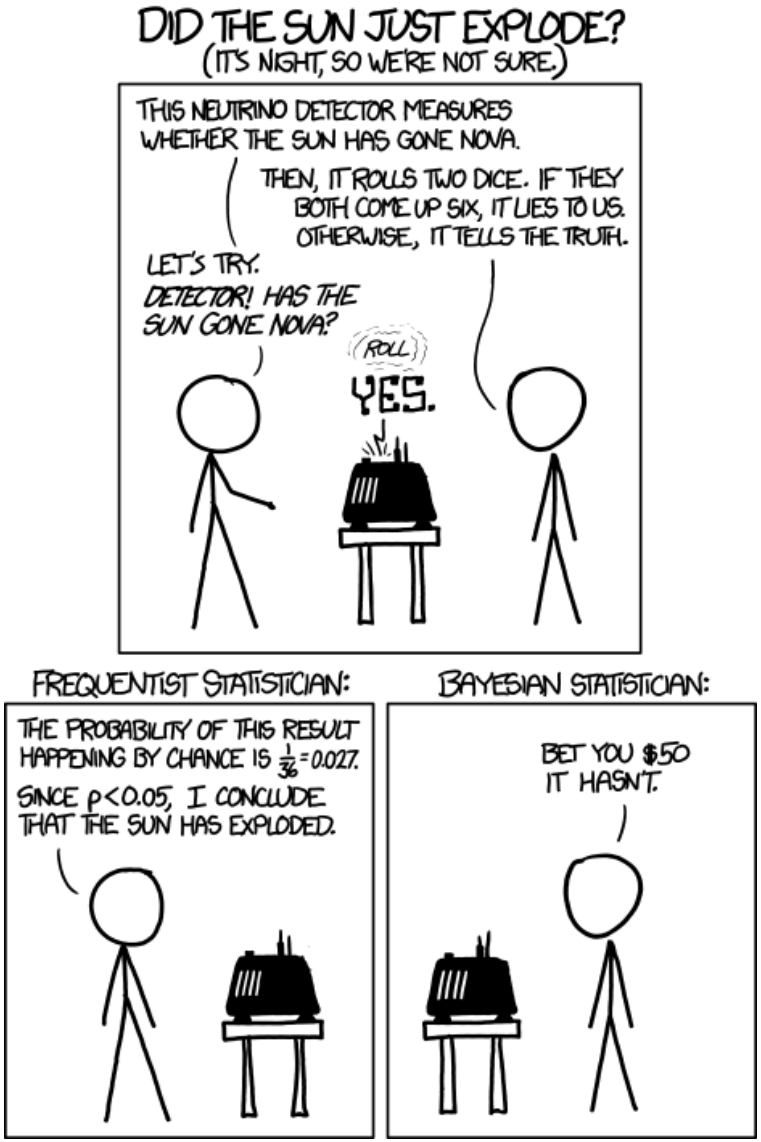
Esempio: distribuzione gamma VI

contour



filled.contour





(https://xkcd.com/1132/)

Autovalutazione

Spiegare il viz della vignetta.

Per i non triestini
viz: spiritosaggine, gioco di parole,
dal tedesco Witz, stesso significato
(fonte: wikipedia)