

Cenni sulla stima di massima verosimiglianza

Maria Luisa Mancusi
Università Bocconi

Novembre 2009

In quanto segue studieremo (o ripasseremo) un metodo di stima che richiede la conoscenza dell'intera funzione di distribuzione e non semplicemente di alcuni suoi momenti. Se la distribuzione condizionale di una variabile y , $f(y|x)$, è nota a meno di un numero limitato di parametri, possiamo stimare questi parametri in modo che la distribuzione che ne risulta sia quella che assegna al campione osservato la massima probabilità di realizzazione.

1. Introduzione alla stima di massima verosimiglianza

Per capire come funziona la stima di massima verosimiglianza partiamo da un esempio molto semplice. Consideriamo una variabile casuale $x \sim N(\mu, \sigma)$ con μ e σ sconosciuti, e supponiamo di avere un campione di due osservazioni: 4 e 6. Per il momento assumiamo che $\sigma = 1$ e concentriamoci sulla stima di μ a partire dal nostro campione. Consideriamo inizialmente l'ipotesi che $\mu=3.5$. La densità di probabilità in corrispondenza di 4 è 0.3521 ed in corrispondenza di 6 è 0.0175.

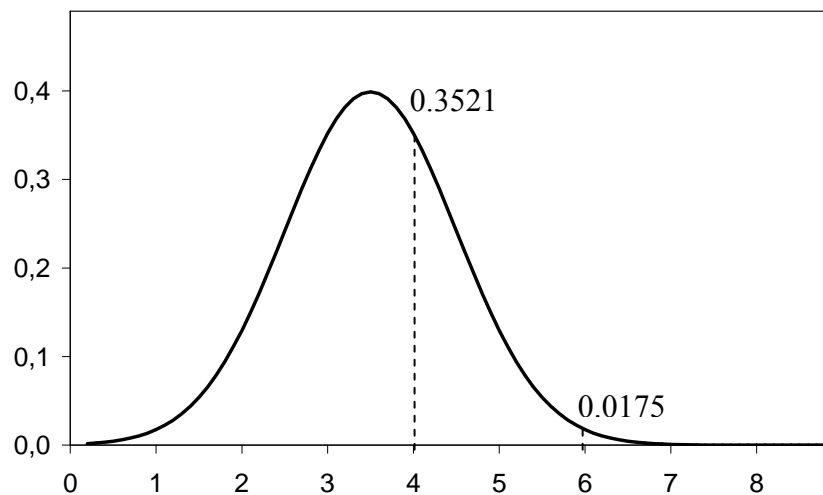


Figura 1. Funzione di densità di $x \sim N(3.5, 1)$

La densità di probabilità congiunta sarà data dal prodotto di questi due valori, ossia 0.0062. Ripetiamo il procedimento per altri possibili valori di μ . I risultati sono riportati nella Tabella 1 e rappresentati graficamente nella Figura 2. La stima di μ in base al principio della massima verosimiglianza corrisponde al valore di μ in corrispondenza del quale la densità congiunta delle osservazioni nel campione è massima. E' dunque il valore di μ che massimizza la probabilità (\leftrightarrow la verosimiglianza) di osservare il campione. Come si vede nella Figura 2, la densità congiunta raggiunge il suo valore massimo in corrispondenza di $\mu=5$, quindi $\hat{\mu}_{MLE} = 5$.

μ	$P(4 \mu)$	$P(6 \mu)$	L
3.5	0.3521	0.0175	0.0062
4.0	0.3989	0.0540	0.0215
4.5	0.3521	0.1295	0.0456
5.0	0.2420	0.2420	0.0585
5.5	0.1295	0.3521	0.0456

Tabella 1. Valori assunti dalla funzione di densità individuale e congiunta in corrispondenza di diverse ipotesi su μ

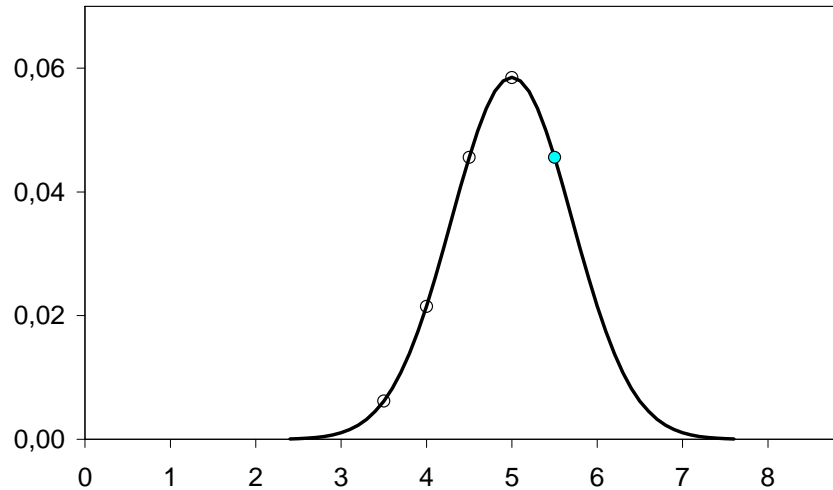


Figura 2. Rappresentazione della densità congiunta in funzione di μ

Vediamo ora come giungere alla stessa conclusione analiticamente. Partiamo dalla funzione di densità della normale con media μ e varianza σ^2 :

$$f(x | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Abbiamo assunto che $\sigma=1$, dunque l'espressione si semplifica nel modo seguente:

$$f(x | \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$$

La densità in corrispondenza delle due osservazioni del campione sarà quindi

$$f(x_1 = 4 | \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \quad f(x_2 = 6 | \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2}$$

e la densità congiunta è data dal prodotto delle due:

$$f(x_1 = 4 \text{ e } x_2 = 6 | \mu, \sigma = 1) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right)$$

L'espressione così trovata è stata ottenuta in corrispondenza di due dati valori di X (4 e 6) ed assume diversi valori al variare di μ : è quindi una funzione di μ , dati X=4 e X=6. Così interpretata, la densità congiunta diviene la *funzione di verosimiglianza*, che indicheremo con $L(\mu|4,6)$ ¹:

$$L(\mu | x_1 = 4 \text{ e } x_2 = 6) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right)$$

Per trovare la stima di massima verosimiglianza di μ occorre ora massimizzare la funzione di verosimiglianza rispetto a μ :

$$\hat{\mu}_{MLE} = \arg \max_{\mu} L(\mu | x_1 = 4 \text{ e } x_2 = 6)^2$$

Il calcolo delle condizioni del primo ordine è piuttosto laborioso e risulta quindi preferibile calcolare il logaritmo della funzione di verosimiglianza, $\ln L(\mu | x_1 = 4 \text{ e } x_2 = 6)$, e massimizzare questo rispetto a μ . Poiché il logaritmo è una funzione monotona crescente, il valore di μ che massimizza $\ln L(\mu | x_1 = 4 \text{ e } x_2 = 6)$ è lo stesso che massimizza $L(\mu | x_1 = 4 \text{ e } x_2 = 6)$:

$$\hat{\mu}_{MLE} = \arg \max_{\mu} \ln L(\mu | x_1 = 4 \text{ e } x_2 = 6)$$

Dunque:

$$\begin{aligned} \ln L &= \ln \left[\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right) \right] \\ &= \ln \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right) + \ln \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right) \\ &= \ln \left(\frac{1}{\sqrt{2\pi}} \right) + \ln \left(e^{-\frac{1}{2}(4-\mu)^2} \right) + \ln \left(\frac{1}{\sqrt{2\pi}} \right) + \ln \left(e^{-\frac{1}{2}(6-\mu)^2} \right) \\ &= 2 \ln \left(\frac{1}{\sqrt{2\pi}} \right) - \frac{1}{2}(4-\mu)^2 - \frac{1}{2}(6-\mu)^2 \end{aligned}$$

Massimizzando questa espressione rispetto a μ otteniamo:

¹ Viene indicata con L dal termine inglese *likelihood* (verosimiglianza).

² MLE ↔ Maximum Likelihood Estimation

$$\frac{d \ln L}{d \mu} = 0 \Leftrightarrow (4 - \mu) + (6 - \mu) = 0 \Rightarrow \hat{\mu}_{MLE} = \frac{4+6}{2} = 5$$

Per essere certi che si tratti di un massimo e non di un minimo occorre controllare la condizione del secondo ordine. In questo caso, si verifica facilmente che la derivata seconda rispetto a μ è uguale a -2 e quindi si tratta effettivamente di un punto di massimo.

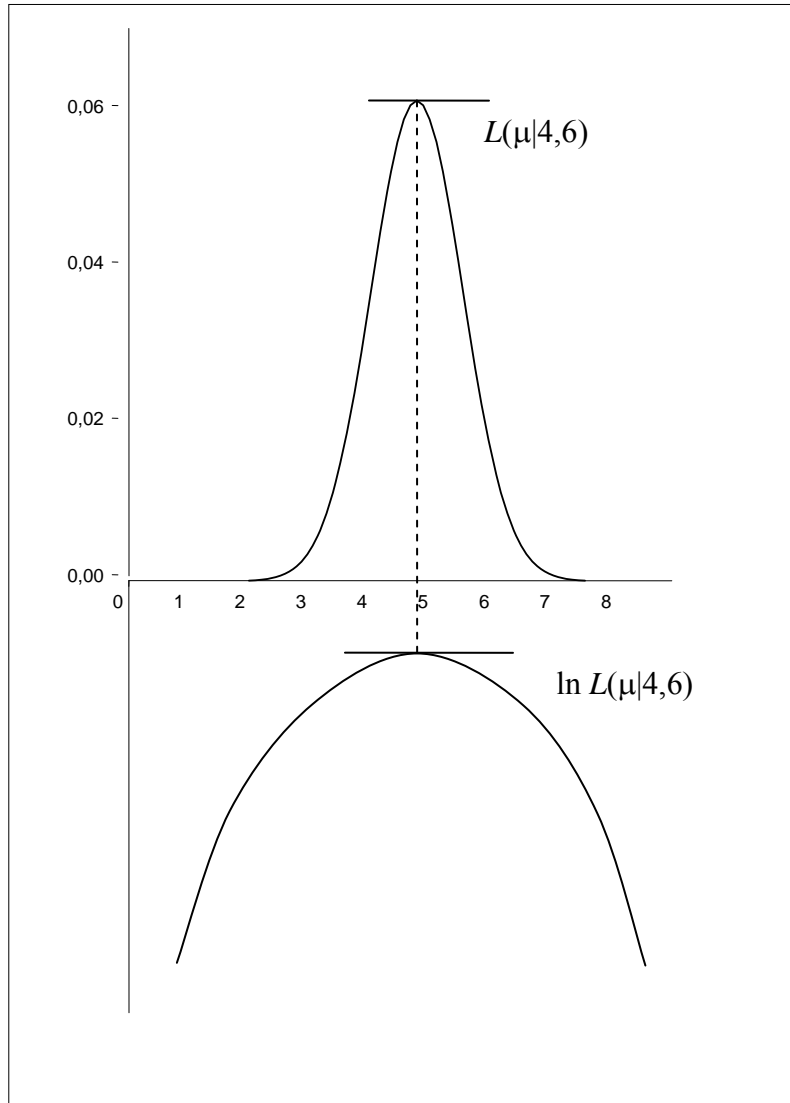


Figura 3. Relazione tra la funzione di verosimiglianza ed il suo logaritmo

Fino ad ora abbiamo assunto che $\sigma=1$, a questo punto possiamo rilassare questa ipotesi e trovare la sua stima di massima verosimiglianza. Anche in questo caso, iniziamo dall'analisi grafica, assumendo $\mu=5$ e considerando diversi valori possibili per σ . Per ciascuno di questi calcoliamo poi $P(4|\sigma)$, $P(6|\sigma)$ e la densità congiunta (vedi Tabella 2).

σ	$P(4 \sigma)$	$P(6 \sigma)$	L
0.5	0.1080	0.1080	0.0117
1.0	0.2420	0.2420	0.0586
2.0	0.1760	0.1760	0.0310

Tabella 2. Valori assunti dalla funzione di densità individuale e congiunta in corrispondenza di diverse ipotesi su σ

La funzione di verosimiglianza così ottenuta, $L(\sigma|4,6)$, risulta essere massima in corrispondenza di $\sigma=1$ (si veda la Figura 4).

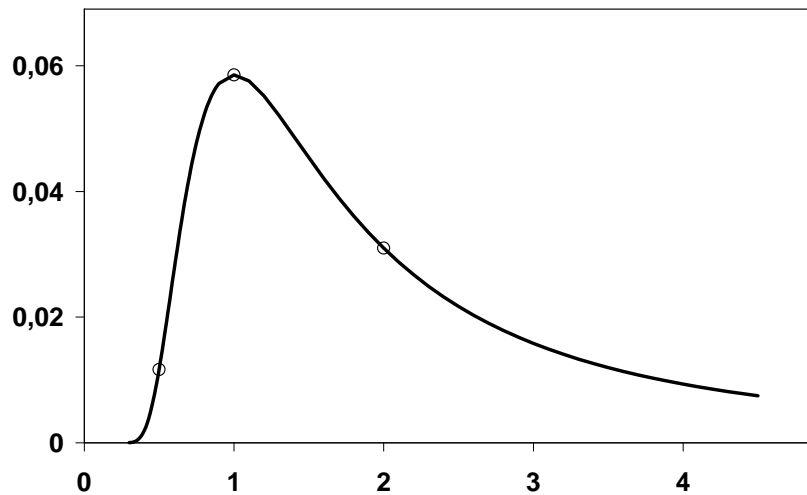


Figura 4. Rappresentazione della densità congiunta in funzione di σ

Analiticamente si procede in modo del tutto analogo a quanto fatto in precedenza, a partire dalla funzione di densità:

$$f(x | \mu = 5, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-5}{\sigma}\right)^2}$$

Una volta ottenuta $\ln L(\sigma | x_1 = 4 \text{ e } x_2 = 6)$, il valore di σ che la massimizza rappresenta la sua la stima di massima verosimiglianza:

$$\hat{\sigma}_{MLE} = \arg \max_{\sigma} \ln L(\sigma | x_1 = 4 \text{ e } x_2 = 6)$$

Risolvendo le condizioni del primo ordine si ricava che: $\hat{\sigma}_{MLE} = 1$.

E' possibile generalizzare questo esempio al caso di un campione N-dimensionale: x_1, x_2, \dots, x_N . Dati μ e σ , la densità in corrispondenza di ciascun x_i è:

$$f(x_i | \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

e la funzione di verosimiglianza:

$$L(\mu, \sigma | x_1, \dots, x_N) = \prod_{i=1}^N \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\frac{1}{2}\sum_{i=1}^N \left(\frac{x_i - \mu}{\sigma}\right)^2\right)$$

$$\ln L(\mu, \sigma | x_1, \dots, x_N) = -N \ln \sigma + N \ln\left(\frac{1}{\sqrt{2\pi}}\right) + \frac{1}{\sigma^2} \left(-\frac{1}{2}\sum_{i=1}^N (x_i - \mu)^2\right)$$

Gli stimatori di massima verosimiglianza di μ e σ si ottengono dalla massimizzazione congiunta della funzione $\ln L(\mu, \sigma | x_1, \dots, x_N)$ rispetto a μ e σ . Occorre innanzitutto calcolare le due derivate parziali:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \frac{\partial}{\partial \mu} \left(-\frac{1}{2}\sum_{i=1}^N (x_i - \mu)^2\right) = \frac{1}{\sigma^2} \left(\sum_{i=1}^N x_i - N\mu\right)$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^N (x_i - \mu)^2$$

Perché siano verificate le condizioni del primo ordine, entrambe le derivate devono essere uguali a zero. Dalla prima si ottiene:

$$\frac{\partial \ln L}{\partial \mu} = 0 \Rightarrow \hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x}$$

Sostituendo nella seconda il valore così ottenuto per μ e risolvendo per σ :

$$\frac{\partial \ln L}{\partial \sigma} = 0 \Rightarrow \hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

Dunque gli stimatori di massima verosimiglianza corrispondono alla media ed alla varianza campionarie³.

2. La stima di massima verosimiglianza nel modello lineare

Consideriamo il modello di regressione lineare semplice:

$$y = \beta_1 + \beta_2 x + \varepsilon$$

³ E' opportuno ricordare che lo stimatore così ottenuto della varianza è distorto: non lo sarebbe se al denominatore avessimo $(N-1)$ invece di N .

In presenza di regressori stocastici, una delle ipotesi fondamentali del modello classico è $E(\varepsilon | x) = 0$, da cui:

$$y = E(y | x) + \varepsilon$$

Ossia la retta di regressione rappresenta la media condizionale della variabile y . Se a questa aggiungiamo l'ipotesi di normalità degli errori, $\varepsilon|x \sim N(0, \sigma^2)$, otteniamo:

$$y | x \sim N(\beta_1 + \beta_2 x, \sigma^2)$$

Dunque la distribuzione condizionale di y è nota a meno di un numero limitato di parametri, $(\beta_1, \beta_2, \sigma)$ e la stima dei parametri del modello di regressione equivale alla stima della media e della varianza di tale distribuzione.

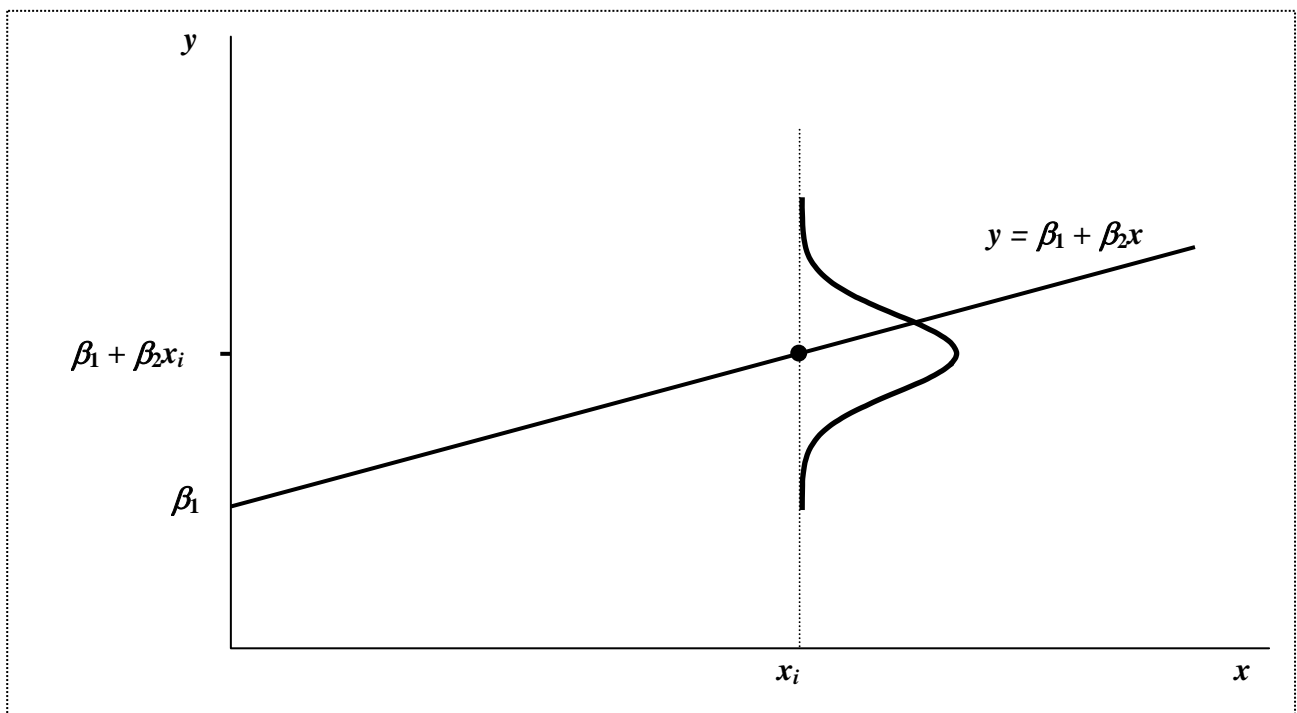


Figura 5. L'interpretazione geometrica del modello di regressione lineare

Per trovare lo stimatore di massima verosimiglianza del vettore di parametri $(\beta_1, \beta_2, \sigma)$ occorre costruire e massimizzare la funzione di verosimiglianza.

$$f(y_i | x_i, \beta_1, \beta_2, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \beta_1 - \beta_2 x_i}{\sigma}\right)^2}$$

$$f(y_1, \dots, y_N | x_1, \dots, x_N; \beta_1, \beta_2, \sigma) = \prod_{i=1}^N f(y_i | x_i, \beta_1, \beta_2, \sigma) = \prod_{i=1}^N \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \beta_1 - \beta_2 x_i}{\sigma}\right)^2}$$

$$L(\beta_1, \beta_2, \sigma | y_1, \dots, y_N; x_1, \dots, x_N) = \prod_{i=1}^N \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \beta_1 - \beta_2 x_i}{\sigma}\right)^2}$$

$$\ln L(\beta_1, \beta_2, \sigma | y_1, \dots, y_N; x_1, \dots, x_N) = N \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^N (y_i - \beta_1 - \beta_2 x_i)^2\right)$$

La massimizzazione di $\ln L$ rispetto a β_1 e β_2 richiede la minimizzazione dell'ultimo termine nella formula precedente, ossia della sommatoria tra parentesi. Ma ciò equivale esattamente alla minimizzazione della somma dei residui al quadrato, da cui si ottiene lo stimatore dei minimi quadrati ordinari. Quindi:

$$\hat{\beta}_{MLE} = \hat{\beta}_{OLS}$$

Sostituendo i valori così ottenuti nella condizione del primo ordine relativa a σ , si ottiene:

$$\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 = \frac{1}{N} \sum_{i=1}^N \hat{\varepsilon}_i^2$$

che è invece diverso da $\hat{\sigma}_{OLS}^2$.

3. Proprietà dello stimatore di massima verosimiglianza

Nella sezione precedente abbiamo visto che lo stimatore di massima verosimiglianza di β coincide con lo stimatore dei minimi quadrati; ciò significa che $E(\hat{\beta}_{MLE}) = \beta$. Questa non è però una proprietà valida in generale. Lo stimatore di massima verosimiglianza non ha, infatti, proprietà in piccoli campioni, quali appunto la non distorsione (basta confrontare $\hat{\sigma}_{MLE}^2$ con $\hat{\sigma}_{OLS}^2$ per capire che il primo è distorto $\leftrightarrow E(\hat{\sigma}_{MLE}^2) \neq \sigma^2$).

Lo stimatore di massima verosimiglianza gode però di alcune fondamentali proprietà asintotiche. Dato un vettore di parametri θ^4 , è possibile dimostrare che $\hat{\theta}_{MLE}$ è:

1. consistente:

$$\text{plim} \hat{\theta}_{MLE} = \theta$$

2. asintoticamente efficiente:

tra tutti gli stimatori consistenti, lo stimatore di massima verosimiglianza è quello con varianza più "piccola"

⁴ Es. $\theta = (\beta_1, \beta_2, \sigma)$

3. asintoticamente distribuito secondo una normale:

$$\sqrt{N}(\hat{\theta}-\theta) \rightarrow N(0, V) \text{ dove } V = \{I(\theta)\}^{-1} \text{ e } I(\theta) = -E \left\{ \frac{\partial^2 \ln L}{\partial \theta \partial \theta'} \right\} \text{ (matrice di informazione)}$$

Affinché lo stimatore di massima verosimiglianza goda delle suddette proprietà è necessario che la funzione di verosimiglianza sia correttamente specificata e, quindi, che l'ipotesi sulla distribuzione degli errori sia corretta⁵. Tale ipotesi risulta dunque essere cruciale.

⁵ Ciò non è evidente nel caso del modello lineare con ipotesi di normalità qui trattato perché lo stimatore di massima verosimiglianza coincide con quello dei minimi quadrati (almeno per quanto riguarda i coefficienti delle variabili esplicative) e quindi ne assume le proprietà asintotiche di consistenza e normalità, anche se la distribuzione degli errori non è normale.