# Indice



# 1 Il metodo della massima verosimiglianza

## 1.1 Definizione del problema

Nel primo quaderno abbiamo definito in modo parallelo le variabili casuali (v.c.) e le variabili statistiche (v.s.) spiegando come queste si comportino in modo formalmente identico una volta che si sia istitui-

ta la corrispondenza frequenze $\leftrightarrow$ probabilità e si sia conseguentemente definito l'operatore di media $E\{\bullet\}$.

Elemento centrale di quella corrispondenza è stato il considerare una v.s. come il riordino di una "popolazione" di valori $\{x_1, \ldots, x_N\}$, ottenuti ripetendo $N$ volte l'esperimento stocastico $\mathcal{E}$ di cui una certa v.c. $X$ descrive il comportamento aleatorio.

Diremo in questo caso che $\{x_1, \ldots, x_N\}$ costituisce un campione di tipo bernoulliano tratto dalla v.c. $X$. L'aggettivo bernoulliano si riferisce all'ipotesi, che qui supporremo sempre soddisfatta, che le ripetizioni dell'esperimento $\mathcal{E}$ siano tali da non influenzarsi stocasticamente tra loro.

**Osservazione 1.1.1:** per comprendere che vi sono casi significativi in cui invece le ripetizioni si influenzano reciprocamente, basta pensare all'esempio dell'estrazione di due numeri da un'urna che ne contenga 3, senza che il primo estratto venga riposto nell'urna stessa prima della seconda estrazione. In questo caso la prima estrazione è descritta dalla v.c.

$$X_1 \left\{ \begin{array}{ccc} 1 & 2 & 3 \\ 1/3 & 1/3 & 1/3 \end{array} \right.$$

ammesso che il primo estratto sia ad esempio $x_1 = 1$, la seconda estrazione ha la distribuzione

$$X_2 \left\{ \begin{array}{cc} 2 & 3 \\ 1/2 & 1/2 \end{array} \right.$$

con le ovvie modifiche nel caso il primo estratto fosse uno degli altri due numeri. Dunque in questo caso la distribuzione di $X_2$ dipende chiaramente dal valore assunto da $X_1$.

Spesso nell'esperimento $\mathcal{E}$ si ha una conoscenza qualitativa del meccanismo stocastico che genera il comportamento aleatorio e poi appunto se ne conosce un campione (bernoulliano) $\{x_1, \ldots, x_N\}$.

La conoscenza qualitativa di $\mathcal{E}$ può allora suggerire che la v.c. $X$ che lo descrive, abbia una distribuzione appartenente ad una certa famiglia $f_X(x, \theta)$, dove $\theta$ rappresenta il parametro (mono- o pluri-dimensionale) che specifica la particolare densità di probabilità che ci interessa: supporremo che $\theta$ possa assumere valori in un insieme $\Theta$ specificato.

Qualora questo elemento qualitativo non ci sia, talvolta, se il campione è numeroso, si può prendere in esame l'istogramma e dalla sua forma arguire l'appartenenza di $X$ ad una certa famiglia.

**Esempio 1.1.1:** sia $\mathcal{E}$ il lancio di una moneta: poiché i valori argomentali di $X$ sono solo due, questa v.c. non può che essere rappresentata da una famiglia ad un parametro

$$X \left\{ \begin{array}{cc} 0 & 1 \\ p & 1-p \end{array} \right. \tag{1.1.1}$$

dove chiaramente

$$\theta = p \ . \tag{1.1.2}$$

analogo ragionamento vale per ogni $\mathcal{E}$ che ammetta un numero finito di valori argomentali.

**Esempio 1.1.2:** sia $\mathcal{E}$ una misura di precisione di una grandezza continua; in tal caso si potrà supporre

$$X = N[\mu, \sigma^2] \tag{1.1.3}$$

e quindi si potrà porre

$$\theta = \left| \begin{array}{c} \mu \\ \sigma^2 \end{array} \right| \ . \tag{1.1.4}$$

Analogo ragionamento vale per ogni v.c. per cui si possa invocare il teorema centrale della statistica.

Il problema che ci poniamo ora è il seguente:

> supponendo di conoscere la famiglia $f_X(x, \theta)$, cui la distribuzione della v.c. $X$ appartiene per un certo $\theta = \overline{\theta}$, e conoscendo un campione (bernoulliano) $\{x_1, \ldots, x_N\}$ estratto da $X$, dare una stima $t$ di $\overline{\theta}$, intendendo con ciò per il momento almeno che $|t - \overline{\theta}|$ è piccolo, con elevata probabilità, rispetto ad un criterio prefissato.

Poiché $t$ deve essere un numero (od un vettore di numeri) calcolabile sulla base delle conoscenze che abbiamo, risulta ovvio che $t$, oltre che dalla forma della famiglia $f(x,\theta)$, dovrà dipendere dai valori del campione, cioè

$$t = t(x_1, \ldots, x_N) \ . \tag{1.1.5}$$

Il nostro problema quindi sarà di discutere i criteri di scelta della funzione (1.1.5) in base alle proprietà che ne derivano.

## 1.2 Variabile campionaria - Statistiche - Stimatori

La variabile campionaria nel caso bernoulliano è la variabile che descrive $N$ repliche identiche e indipendenti dell'esperimento $\mathcal{E}$, ovvero descrive l'esperimento complesso $\mathcal{E}^N \equiv \mathcal{E} \otimes \ldots \otimes \mathcal{E}$.

Se la v.c. $X$ è ad 1 dimensione, allora la v.c. campionaria sarà $N$-dimensionale $\underline{X}^{(N)}$, e fatta in modo tale che ogni sua componente abbia la stessa distribuzione di $X$

$$\underline{X}^{(N)} = \left| \begin{array}{c} X_1 \\ X_2 \\ \vdots \\ X_N \end{array} \right| \qquad f_{X_i}(x) = f_X(x,\theta) \ . \tag{1.2.1}$$

La distribuzione della v. campionaria $\underline{X}^{(N)}$ è poi interamente identificata tramite l'ipotesi che le sue componenti siano tutte indipendenti tra loro, così che

$$f_{\underline{X}}(\underline{x};\theta) = f_{\underline{X}}(x_1, \ldots, x_N;\theta) = \Pi_{i=1}^{N} \quad f_X(x_i;\theta) \ . \tag{1.2.2}$$

La funzione di densità di $\underline{X}^{(N)}$ è detta "likelihood" (verosimiglianza) e noi la indicheremo come

$$L(\underline{x};\theta) = f_{\underline{X}}(\underline{x};\theta) \ . \tag{1.2.3}$$

L'importanza della v. campionaria sta nel fatto che un campione ber-

noulliano $\{x_1, \ldots, x_N\}$ di $N$ estrazioni da $X$ può essere considerato come una singola estrazione della variabile $\underline{X}^{(N)}$.

Un esempio chiarirà meglio il concetto.

**Esempio 1.2.1:** si consideri il classico gioco a testa e croce ($\mathcal{E}$); un giocatore lancia due volte una moneta ottenendo il campione $(t, c)$. Ma questo stesso può essere anche considerato come un campione di una sola replica di $\mathcal{E}^2$, cioè di un esperimento costituito da due lanciatori indipendenti che lanciano contemporaneamente due monete identiche.

**Osservazione 1.2.1:** il concetto di variabile campionaria può essere applicato anche al caso (non bernoulliano) in cui la ripetizione dell'esperimento $\mathcal{E}$ non è fatta nelle identiche condizioni, così che ogni ripetizione sarà descritta da una diversa v.c. $X_i$.

Un esempio importante di questo tipo è costituito dal caso in cui una stessa quantità $\mu$ è misurata in condizioni diverse o con strumenti diversi, così che ogni misura avrà una diversa varianza $\sigma_i^2$. In questo caso si userà semplicemente la definizione (1.2.3), tenendo però conto che ora

$$f_{\underline{X}}(\underline{x}; \theta) = \Pi_{i=1}^{N} \quad f_{X_i}(x_i; \theta) \ , \tag{1.2.4}$$

senza che le $f_{X_i}$ siano uguali tra loro.

Anzi, per questa strada si potrebbe anche lasciar cadere l'ipotesi di indipendenza stocastica delle $X_i$, definendo la v. campionaria semplicemente come quella variabile che descrive il complesso degli $N$ esperimenti considerati, indipendenti o no che siano le variabili che li descrivono, e la funzione di likelihood come la densità di probabilità di tale variabile $N$-pla.

**Osservazione 1.2.2:** nel caso di variabili discrete (ad es. binomiale, poissoniana, ecc.) definiamo come likelihood in un punto $x_1 = \xi_{k(1)}, x_2 = = \xi_{k(2)}, \ldots$ direttamente il prodotto delle probabilità che ogni componente $x_i$ assuma il valore assegnato $\xi_{k(i)}$. Perciò se

$$P_X\{x = \xi_k\} = p_k \qquad k = 1, 2, \ldots$$

si ha

$$L_{\underline{X}}(\underline{x}) = \Pi_{i=1}^{N} P_X\{x_i = \xi_{k(i)}\} \ . \tag{1.2.5}$$

Definiamo *statistica* $S(\underline{X}^{(N)})$ una qualunque funzione, a una o più dimensioni, della v. campionaria $\underline{X}^{(N)}$.

Così ad esempio la v.c.

$$\mathcal{M} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1.2.6}$$

è una statistica che prende ovviamente il nome di media campionaria. È bene osservare che (1.2.6), come statistica è una v.c. e non già una costante: è chiaro però che per un certo campione dato $\{x_1, \ldots, x_N\}$, cioè per una estrazione da $\underline{X}^{(N)}$, si potrà calcolare la corrispondente media campionaria, cioè il numero

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.2.7}$$

e che tale numero può essere considerato come un'estrazione da $\mathcal{M}$.

Alla luce di queste definizioni possiamo riprendere il problema della stima definito nel §1.1 nel seguente modo: data una certa funzione di likelihood $L(\underline{x};\theta)$, cerchiamo un'opportuna statistica $T(\underline{X}^{(N)})$, con una distribuzione sufficientemente concentrata attorno al valore $\theta$, in modo tale che un'estrazione $t$ da $T$, calcolata in corrispondenza a un campione estratto $\underline{x},\ t = T(\underline{x})$, abbia un'elevata probabilità di essere vicina a $\theta$.

Con questa impostazione $T$ si chiama uno *stimatore* di $\theta$ mentre la sua estrazione $t = T(\underline{x})$ si chiama *stima*.

Vedremo nei prossimi paragrafi le proprietà di maggior interesse degli stimatori, in particolare studieremo la media di $T(\underline{X}^{(N)})$ e la sua varianza per $N \to \infty$.

## 1.3 Correttezza

Per semplicità rappresentiamo in questo paragrafo la v. campionaria col vettore $(\underline{X})$.

Avendo stabilito che uno stimatore $T(\underline{X})$ debba essere distribuito attorno

a $\theta$, la prima questione interessante sarà di vedere quale relazione vi sia tra la media di $T(\underline{X})$, parametro di posizione per eccellenza, e $\theta$.

Definiamo il bias (distorsione) di $T(\underline{X})$ come

$$b(\theta) = E\{T(\underline{X})|\theta\} - \theta \ . \tag{1.3.1}$$

Questa formula richiede un commento: intanto intendiamo con $E\{T(\underline{X})|\theta\}$ la media di $T(\underline{X})$, facendo l'ipotesi che $\theta$ sia il valore corretto del parametro, ovvero

$$E\{T(\underline{X})|\theta\} = \int_{R^N} T(\underline{x}) L(\underline{x};\theta) d_N \underline{x} \ . \tag{1.3.2}$$

Dunque, supposto che $\theta$ sia il valore giusto del parametro, $b(\theta)$ misura la deviazione della media di $T(\underline{X})$ da $\theta$.

Uno stimatore che abbia bias nullo

$$E\{T(\underline{X})|\theta\} = \theta \tag{1.3.3}$$

si dice *unbiased* o *corretto*.

**Esempio 1.3.1:** la media campionaria $\mathcal{M} = (1/N)\sum_i X_i$ è uno stimatore corretto di $\mu$, qualunque sia la distribuzione sottostante. Infatti notiamo che per ogni componente $X_i$ si ha [1]

$$E\{X_i|\mu\} = \mu \tag{1.3.4}$$

così che

$$E\{\mathcal{M}|\mu\} = \frac{1}{N}\sum_i E\{X_i \mid \mu\} = \frac{1}{N}\sum_i \mu = \mu \ . \tag{1.3.5}$$

**Esempio 1.3.2:** i momenti (non centrali) campionari di ordine $n$ qualsiasi sono stime unbiased dei corrispondenti momenti della distribuzione sottostante. Sia

[1] Ciò esprime l'ovvio fatto che ogni componente è di per sé uno stimatore unbiased di $\mu$.

$$\mathcal{M}_{(n)} = \frac{1}{N}\sum_i X_i^n \tag{1.3.6}$$

la variabile campionaria "momento campionario di ordine $n$".

Poiché chiaramente

$$E\{X_i^n|\mu_n\} = \mu_n \tag{1.3.7}$$

ragionando come per la media si ha

$$E\{\mathcal{M}_{(n)}|\mu_n\} = \mu_n \ . \tag{1.3.8}$$

**Esempio 1.3.3:** la varianza campionaria non è uno stimatore corretto della varianza teorica della popolazione sottostante. Definiamo come varianza campionaria la variabile

$$\begin{aligned} \mathcal{S}^2 &= \frac{1}{N}\sum(X_i - \mathcal{M})^2 = \\ &= \frac{1}{N}\sum X_i^2 - (\mathcal{M})^2 = \mathcal{M}_{(2)} - (\mathcal{M})^2 \ . \end{aligned} \tag{1.3.9}$$

Studiamo la media di $\mathcal{M}^2$.

Si ha tenendo conto che le diverse componenti di $\underline{X}$ sono tra loro indipendenti,

$$\begin{aligned} E\{\mathcal{M}^2\} &= E\left\{\frac{1}{N^2}\sum_{i,k} X_i X_k\right\} = \\ &= E\left\{\frac{1}{N^2}\sum_{i\neq k} X_i X_k\right\} + E\left\{\frac{1}{N^2}\sum_i X_i^2\right\} = \\ &= \frac{N^2 - N}{N^2}\mu^2 + \frac{N\mu_2}{N^2} = \\ &= \frac{N-1}{N}\mu^2 + \frac{1}{N}\mu_2 \ . \end{aligned} \tag{1.3.10}$$

Pertanto, tornando alla (1.3.9) si ha

$$
\begin{aligned}
E\{\mathcal{S}^2\} &= \mu_2 - \frac{N-1}{N}\mu^2 - \frac{1}{N}\mu_2 = \\
&= \frac{N-1}{N}\left[\mu_2 - \mu^2\right] = \frac{N-1}{N}\sigma^2 \ . \qquad (1.3.11)
\end{aligned}
$$

**Osservazione 1.3.1:** nell'esempio (1.3.3) si è dimostrato che la media di $\mathcal{S}^2$ non coincide con $\sigma^2$, ma è una funzione lineare di $\sigma^2$. In questo caso è particolarmente semplice modificare lo stimatore in modo da ottenerne un altro che sia unbiased. Basterà infatti porre

$$
\overline{\mathcal{S}}^2 = \frac{N}{N-1}\mathcal{S}^2 = \frac{1}{N-1}\sum_i (X_i - \mathcal{M})^2 \qquad (1.3.12)
$$

perché si abbia

$$
E\{\overline{\mathcal{S}}^2\} = \sigma^2 \ . \qquad (1.3.13)
$$

Lo stimatore (1.3.12) è detto varianza campionaria corretta: come si vede esso differisce poco (anche numericamente) da $\mathcal{S}^2$ quando $N$ sia grande, mentre la differenza può essere significativa quando $N$ sia pari a poche unità.

**Esempio 1.3.4:** sia $X$ una variabile doppia $X = \left| \begin{array}{c} U \\ V \end{array} \right|$ e sia $\underline{X}^{(N)}$ la corrispondente variabile campionaria ($2N$-dimensionale) le cui componenti $X_i = \left| \begin{array}{c} U_i \\ V_i \end{array} \right|$ sono tra loro indipendenti e identicamente distribuite. La covarianza campionaria

$$
\begin{aligned}
\mathcal{S}_{UV} &= \frac{1}{N}\sum (U_i - \mathcal{M}_U)(V_i - \mathcal{M}_V) \\
&= \frac{1}{N}\sum U_i V_i - \mathcal{M}_U \mathcal{M}_V \qquad (1.3.14)
\end{aligned}
$$

è uno stimatore non corretto di $\sigma_{UV}$.

Poiché tanto $\mathcal{S}_{UV}$ quanto $\sigma_{UV}$ sono indici invarianti per traslazione, possiamo supporre di metterci nella situazione più comoda, cioè $\mu_U = 0,\ \mu_V = 0$, senza alterare il risultato.

Calcoliamo prima

$$\begin{aligned} E\{\mathcal{M}_U\mathcal{M}_V\} &= E\left\{\frac{1}{N^2}\sum_{i,k} U_i V_k\right\} = \\ &= \frac{1}{N^2}\sum_i E\{U_i V_i\} = \\ &= \frac{N\sigma_{UV}}{N^2} = \frac{\sigma_{UV}}{N} \ , \end{aligned}$$

poiché $E\{U_i V_k\} = 0$ $(i \neq k)$ in base alle ipotesi fatte.

Pertanto, dalla (1.3.14)

$$E\{\mathcal{S}_{UV}\} = \frac{1}{N}\sum \sigma_{UV} - \frac{1}{N}\sigma_{UV} = \frac{N-1}{N}\sigma_{UV} \ . \tag{1.3.15}$$

Analogamente a quanto si fa per la varianza, è anche qui semplice correggere lo stimatore definendo una covarianza corretta tramite la formula

$$\overline{\mathcal{S}}_{UV} = \frac{1}{N-1}\sum (U_i - \mathcal{M}_U)(V_i - \mathcal{M}_V) \ . \tag{1.3.16}$$

**Osservazione 1.3.2:** nella definizione (1.3.3) di correttezza si è supposto che $E\{T\}$ fosse uguale al parametro $\theta$ per tutti i suoi valori: in questo caso si dirà che $T$ è uno stimatore uniformemente corretto di $\theta$. Vi sono casi invece in cui lo stimatore $T(\underline{X})$ può essere corretto solo per qualche valore di $\theta$. Si consideri ad esempio il coefficiente di correlazione campionario definito da

$$R = \frac{\mathcal{S}_{UV}}{\mathcal{S}_U \mathcal{S}_V} \ : \tag{1.3.17}$$

supponiamo che la variabile doppia sottostante sia

$$X = \left| \begin{array}{c} U \\ V \end{array} \right| \ ,$$

una normale doppia con $\rho = 0$. In tal caso, riscrivendo la (1.3.17) nella forma

$$R = \frac{1}{N} \sum \frac{U_i - \mathcal{M}_U}{\mathcal{S}_U} \cdot \frac{V_i - \mathcal{M}_V}{\mathcal{S}_V} \tag{1.3.18}$$

si vede che, per l'indipendenza stocastica di $U_i, V_i$,

$$E\{R\} = \frac{1}{N} \sum E\left\{\frac{U_i - \mathcal{M}_U}{\mathcal{S}_U}\right\} \cdot E\left\{\frac{V_i - \mathcal{M}_V}{\mathcal{S}_V}\right\} \ . \tag{1.3.19}$$

Ora, essendo le funzioni entro parentesi funzioni dispari, mentre le distribuzioni dei vettori

$$\left|\begin{array}{c} U_i \\ \vdots \\ U_N \end{array}\right| \ , \ \left|\begin{array}{c} V_i \\ \vdots \\ V_N \end{array}\right|$$

sono funzioni pari, si ha chiaramente che ciascuna delle medie di (1.3.19) è nulla, così

$$E\{R\} = 0 \ . \tag{1.3.20}$$

Pertanto $R$ è uno stimatore corretto di $\rho = 0$ (ed $X$ è normale); è possibile anche vedere che se $\rho \neq 0$, $R$ non è uno stimatore unbiased.

**Osservazione 1.3.3:** (diminuzione del bias). Se $T$ è uno stimatore corretto di $\theta$, in generale $g(T)$ non è uno stimatore corretto di $g(\theta)$, almeno in modo esatto, a meno che $g$ non sia una funzione lineare.

Tuttavia se $\sigma^2(T)$ è abbastanza piccola si può porre in modo approssimato

$$E\{g(T)\} \cong g(E\{T\}) = g(\theta) \ . \tag{1.3.21}$$

Supposto che $g$ sia una funzione liscia, ad esempio con derivata terza continua, (1.3.21) equivale a trascurare, come termine principale, la quantità $1/2 \ g''(\theta) \ \sigma_T^2$, infatti

$$E\{g(T)\} = E\{g(\theta) + (T-\theta)g'(\theta) + \frac{1}{2}g''(\theta)(T-\theta)^2 + o_2\} \ .$$

Qualora l'approssimazione (1.3.21) non bastasse, si può ridurre il bias di $g(T)$, portandolo ad un infinitesimo di grado superiore, ponendo

$$G = g(T) - \frac{1}{2}\sigma_T^2 g''(T) \ . \tag{1.3.22}$$

È infatti facile vedere che, per una distribuzione simmetrica di $T$ e per una $g$ abbastanza liscia, (1.3.22) è $O_4$, al contrario di $g(T)$ che è $O_2$.

Si osservi che (1.3.22) è utile quando si sappia calcolare $\sigma_T^2$.

## 1.4 Consistenza

Vogliamo ora caratterizzare il comportamento degli stimatori quando la numerosità del campione $N$ tenda all'infinito.

È ovvio che sia desiderabile che la distribuzione di $T(\underline{X}^{(N)})$ tenda a concentrarsi sempre più attorno a $\theta$, cioè che $T(\underline{X}^{(N)})$ tenda in probabilità a $\theta$.

Diremo che uno stimatore per cui

$$\lim_{N\to\infty} T(\underline{X}^{(N)}) = \theta \quad \text{in} \quad P \ , \tag{1.4.1}$$

è *consistente*.

In pratica spesso anziché provare la (1.4.1) si fa ricorso alla condizione sufficiente della convergenza in media quadratica, ovvero

$$\lim_{N\to\infty} E\{T(\underline{X}^{(N)}\} = \theta \tag{1.4.2}$$

$$\lim_{N\to\infty} \sigma^2\{T(\underline{X}^{(N)}\} = 0 \ . \tag{1.4.3}$$

Notiamo che la (1.4.2) equivale a dire che il bias $b(\theta)$ tende a zero per $N \to \infty$.

**Esempio 1.4.1:** la media campionaria $\mathcal{M}$ è uno stimatore consistente di $\mu$. Infatti $\mathcal{M}$ è corretta, così che $b = 0$, ed inoltre, usando la propagazione degli errori,

$$\sigma^2(\mathcal{M}) = \frac{1}{N^2}\sum \sigma_X^2 = \frac{\sigma_X^2}{N} = 0\left(\frac{1}{N}\right) \ , \tag{1.4.4}$$

così che la (1.4.2) e (1.4.3) sono verificate.

**Esempio 1.4.2:** se la v.c. $X$ ha momento quarto limitato ($\overline{\mu}_4 < +\infty$) allora $\mathcal{S}^2$ è uno stimatore consistente di $\sigma^2$. In effetti (1.4.2) è chiaramente verificata perché $E\{\mathcal{S}^2\} = [(N-1)/N]\sigma^2$.

Quanto alla varianza di $\mathcal{S}^2$, possiamo osservare che, se $\overline{\mu}_4 < +\infty$, si ha

$$\begin{aligned} E\{\mathcal{S}^4\} &= \frac{N-1}{N}\left(1 - \frac{2}{N} + \frac{3}{N^2}\right)\sigma^4 + \frac{1}{N}\left(1 - \frac{1}{N}\right)^2 \overline{\mu}_4 = \\ &= \sigma^4 + 0\left(\frac{1}{N}\right) \ . \end{aligned} \tag{1.4.5}$$

Perciò è anche

$$\sigma^2\{\mathcal{S}^2\} = E\{\mathcal{S}^4\} - \left(1 - \frac{1}{N}\right)^2 \sigma^4 = 0\left(\frac{1}{N}\right)$$

e la consistenza è provata.

Si può notare che poiché $\mathcal{S}^2 \xrightarrow{P} \sigma^2$ si ha anche chiaramente $\overline{\mathcal{S}}^2 = N/(N-1)\mathcal{S}^2 \xrightarrow{P} \sigma^2$, cioè anche $\overline{\mathcal{S}}^2$ è uno stimatore consistente di $\sigma^2$.

Infine osserviamo che nel caso la $X$ sia una v.c. normale, dalla (1.4.5) e dalla nota relazione $\overline{\mu}_4 = 3\sigma^4$ si ricava

$$\sigma^2(\overline{\mathcal{S}}^2) = \frac{2\sigma^4}{N-1} \ . \tag{1.4.6}$$

**Esempio 1.4.3:** sia data la famiglia di distribuzioni uniformi

$$f_X(x,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & x < 0, \ x > \theta \end{cases} \quad : \tag{1.4.7}$$

si può dimostrare che la seguente statistica campionaria

$$T(\underline{X}) = \max_i X_i$$

è uno stimatore consistente di $\theta$.

Infatti, per ricavare la distribuzione di $T$, basta notare che

$$\begin{aligned} F_T(t) &= P\{T \le t\} = P\{X_1 \le t, \ldots, X_N \le t\} = \\ &= [P\{X \le t\}]^N = F_X(t)^N \ , \end{aligned}$$

così che

$$f_T(t) = N \ F_X(t)^{N-1} f_X(t) \qquad (T = \ \max \ X_i) \ .$$

Usando la (1.4.7) per la $f_X$ e per la $F_X$ si ha per $T = \ \max \ X_i$,

$$f_T(t) = \begin{cases} N \dfrac{t^{N-1}}{\theta^N} & 0 \le t \le \theta \\ \\ 0 & \theta < t \end{cases}$$

così che

$$E\{T\} = E\{\max \ X_i\} = \frac{1}{\theta^N} \int_0^\theta N \ t^N \ dt = \frac{N}{N+1} \ \theta = \theta + 0\left(\frac{1}{N}\right) \ .$$

Inoltre

$$E\{T^2\} = \frac{1}{\theta^N} \int_0^\theta N \ t^{N+1} \ dt = \frac{N}{N+2} \ \theta^2 = \theta^2 + 0\left(\frac{1}{N}\right)$$

così che $\sigma^2(T) \to 0$ per $N \to \infty$ e la consistenza di $T$ è provata.

**Osservazione 1.4.1:** contrariamente a quanto avviene per la proprietà di correttezza, la consistenza viene mantenuta se si passa da $T$ a $g(T)$ per una vasta classe di funzioni.

In effetti se $g(\bullet)$ è continua in $\theta$ si avrà che l'immagine inversa di un intorno qualsiasi di $g(\theta)$, $[g(\theta) - \varepsilon, \ g(\theta) + \varepsilon]$ contiene un intorno $I_\theta$ di $\theta$.

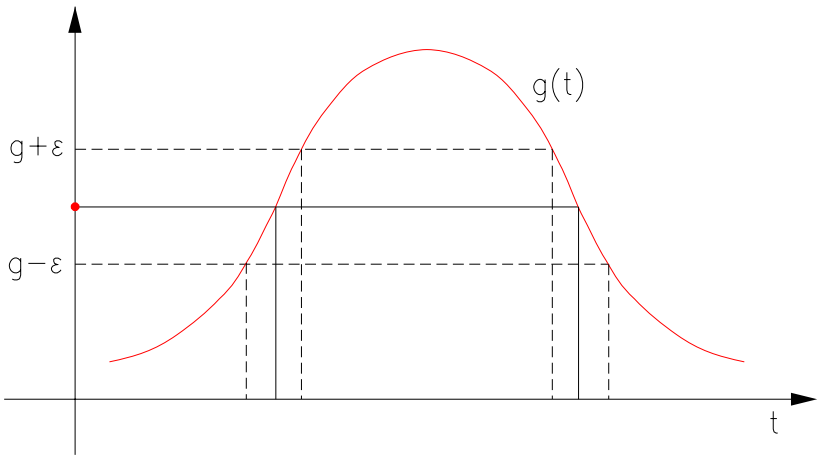



Figura 1.4.1:

Pertanto si ha

$$P(|g(T) - g(\theta)| < \varepsilon) > P(T \in I_{\theta,\varepsilon}) \ ;$$

d'altro canto per la consistenza di $T$ come stimatore di $\theta$,

$$\lim_{N\to\infty} P(T \in I_{\theta,\varepsilon}) = 1 \quad , \quad \forall \epsilon \ ,$$

così che risulta anche

$$\lim_{N\to\infty} P(|g(T) - g(\theta)| < \varepsilon) = 1 \quad , \quad \forall \varepsilon \ , \tag{1.4.8}$$

e $g(T)$ è uno stimatore consistente di $g(\theta)$.

La stessa proprietà vale a più dimensioni anche se il provare ciò richiede un più attento esame.

Premettiamo infatti un lemma.

**Lemma 1.4.1:** sia $T = (T_1, \ldots, T_p)$ uno stimatore di $\theta = = (\theta_1, \ldots, \theta_p)$ nel senso che per ogni componente $T_i \to \theta_i$

in $P$: allora $T$ è uno stimatore consistente $p$-dimensionale di $\theta$, cioè, preso un qualunque intorno $p$-dimensionale di $\theta, I_\theta$, risulta

$$\lim_{N\to\infty} P\{T_{(N)} \in I_\theta\} = 1 \ . \tag{1.4.9}$$

Infatti notiamo che data una successione $P^{(N)}$ di distribuzioni di probabilità tali che

$$\lim_{N\to\infty} P^{(N)}(A) = 1 \quad , \quad \lim_{N\to\infty} P^{(N)}(B) = 1 \ ,$$

dovrà essere anche necessariamente

$$\lim_{N\to\infty} P^{(N)}(A \cap B) = 1 \ .$$

Infatti, fissato $\varepsilon$, per $N$ abbastanza grande $P^{(N)}(A) > 1 - \varepsilon$, e quindi $P^{(N)}(B - A) < \varepsilon$. D'altra parte $P^{(N)}(B) > 1 - \varepsilon$, e quindi $P^{(N)}(A \cap B) = = P^{(N)}(B) - P^{(N)}(B - A) > 1 - 2\varepsilon$.

Tale proprietà si estende immediatamente al caso di un numero finito di intersezioni. Ne segue che la proprietà (1.4.9) è vera perché ogni intorno di $\theta$ conterrà un cubetto abbastanza piccolo
$\{\Delta\theta_1\} \otimes \{\Delta\theta_2\} \ldots \otimes \{\Delta\theta_p\} = C_\theta$ e perché tale cubo può essere visto anche come l'intersezione degli "strati" $p$-dimensionali

$$C_\theta = \{\theta_1 \in \Delta\theta_1, \ \forall\theta_2 \ldots \forall\theta_p\} \cap \{\theta_2 \in \Delta\theta_2, \ \forall\theta_1 \ldots \forall\theta_p\} \cap \ldots \ .$$

Poiché d'altronde per ipotesi

$$P\{T_1 \in \Delta\theta_1, \ \forall T_2 \ldots \forall T_p\} \to 1 \ ,$$

ed analogamente per le altre componenti, si vede che dovrà pure essere

$$P\{T \in C_\theta\} \to 1 \ ,$$

cioè vale la (1.4.9).

Una volta dimostrato il lemma, procedendo in modo identico al caso monodimensionale si vede che: se $T_i$ sono stimatori consistenti di $\theta_i$ e se $g$ è una funzione continua in $\theta = (\theta_1, \ldots, \theta_p)$, allora $g(T) = g(T_1, \ldots, T_p)$ è uno stimatore consistente di $g(\theta)$.

**Esempio 1.4.4:** il coefficiente di correlazione lineare campionario

$$R = \frac{\mathcal{M}_{XY} - \mathcal{M}_X \mathcal{M}_Y}{\mathcal{S}_X \mathcal{S}_Y} \ , \tag{1.4.10}$$

dove si è posto

$$\mathcal{M}_{XY} = \frac{1}{N} \sum X_i Y_i \ , \tag{1.4.11}$$

è uno stimatore consistente di $\rho$. Applicando l'Osservazione 1.4.1 è sufficiente supporre che $\sigma_X, \sigma_Y \neq 0$, così che (1.4.10) definisca una funzione continua di tutti i suoi argomenti, e provare che $\mathcal{M}_{XY} \to \mu_{XY}$ in $P$, cosa che viene lasciata al lettore come esercizio.

## 1.5 Sufficienza - Sufficienza minimale - Completezza

Larga parte della teoria della stima è dominata da un principio che potremmo chiamare *principio di verosimiglianza*: questo stabilisce che tutte le informazioni che si possono ottenere su un parametro $\theta$ a partire da un campione $\{x_1, \ldots, x_N\}$, debbono essere ricavate dalla forma della likelihood $L(\underline{x}; \theta)$.

Ciò introduce il concetto di sufficienza, per tener conto che spesso $\underline{x}$ entra in $L(\underline{x}; \theta)$ solo in particolari combinazioni, le quali dunque contengono tutta l'informazione proveniente dal campione, su $\theta$. Prendiamo un esempio semplice.

**Esempio 1.5.1:** consideriamo la distribuzione

$$f_X(x;\theta) = \frac{1}{\theta} e^{-x/\theta} \quad , \quad x \geq 0$$

che genera la likelihood

$$L(\underline{x};\theta) = \frac{1}{\theta^N} e^{-1/\theta \sum x_i} \quad , \quad x_i \geq 0 \ .$$

Poiché i valori campionari entrano in $L$ solo tramite $\sum x_i$, diciamo che la statistica

$$S = \sum X_i$$

contiene tutta l'informazione, per la famiglia data, concernente il parametro $\theta$. Ne segue che, accettando tale principio, si è portati a cercare i possibili stimatori di $\theta$ fra le funzioni di $S$, in modo tale che campioni diversi che danno lo stesso valore di $S$, portino anche alle stesse stime di $\theta$.

Questo concetto è posto su basi più precise con la definizione di *statistica sufficiente*:

> diciamo che $S(\underline{X}^{(N)})$ è una statistica sufficiente per il parametro $\theta$ e la famiglia $f_X(x;\theta)$, se la distribuzione della v. campionaria $\underline{X}$ condizionata a $S = s$ risulta indipendente da $\theta$.

In termini analitici, ricordando la definizione di variabile condizionata, si può verificare che

$$L(\underline{x};\theta|S=s) = \frac{L(\underline{x};\theta)|\text{grad } S(\underline{x})|^{-1}}{\displaystyle\int_{\{S=s\}} L(\underline{x};\theta)|\text{grad } S(\underline{x})|^{-1} d\sigma} = h(\underline{x}) \qquad (1.5.1)$$

dove $\underline{x} \in \{S = s\}$, e $d\sigma$ è l'elemento d'area di $\{S(\underline{x}) = s\}$.
Notando che la distribuzione marginale di $S$,

$$f_S(s;\theta) = \int_{\{S=s\}} L(\underline{x};\theta)|\text{ grad } S(\underline{x}))|^{-1} d\sigma \tag{1.5.2}$$

ovviamente non può che essere funzione di $s$ e di $\theta$

$$f_S(s,\theta) = K(s,\theta) \ ,$$

si trova che la condizione (1.5.1) può essere anche scritta nella forma equivalente

$$L(\underline{x};\theta) = K(s,\theta)\ h(\underline{x})|\text{grad } S(\underline{x})| \equiv K(s,\theta)H(\underline{x}) \qquad (s = S(\underline{x})) \ . \tag{1.5.3}$$

Si perviene così al cosiddetto *teorema di fattorizzazione* che afferma appunto che condizione necessaria e sufficiente affinché $S(\underline{X})$ sia una statistica sufficiente per $\theta$ e per la famiglia $f_{\underline{X}}(\underline{x},\theta)$, è che la likelihood si fattorizzi nel prodotto di due fattori, uno dipendente solo da $s = S(\underline{x})$ e $\theta$, e l'altro dipendente solo da $\underline{x}$.

**Esempio 1.5.2:** sia $\underline{X}^{(N)}$ un campione di variabili i.i.d. (indipendenti, identicamente distribuite) tratte da una $N[\mu,\sigma^2]$. La likelihood di $\underline{X}^{(N)}$ può essere scritta come

$$L(\underline{x};\mu,\sigma^2) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{1/(2\sigma^2)(\underline{x}-\mu\underline{e})^+(\underline{x}-\mu\underline{e})} \tag{1.5.4}$$

dove

$$\underline{e} = [1 \quad 1 \ldots 1]^+ \ .$$

Riscrivendo la (1.5.4) come

$$L = \frac{1}{(2\pi)^{N/2}} e^{-N\log\sigma - 1/(2\sigma^2)[\underline{x}^+\underline{x} - 2\mu\underline{e}^+\underline{x} + N\mu^2]} \tag{1.5.5}$$

e confrontando col teorema di fattorizzazione si riconosce che:

- se $\theta = (\mu, \sigma^2)$ allora $S = (\underline{X}^+\underline{X}, \underline{e}^+\underline{X}) = (\sum X_i^2, \sum X_i)$;
- se $\sigma^2$ è nota e $\theta = \mu$ allora $S = (\sum X_i)$;
- se $\mu$ è nota e $\theta = \sigma^2$ allora $S = \sum (X_i - \mu)^2$.

**Osservazione 1.5.1:** si noti che l'esistenza di una statistica sufficiente $S(\underline{X})$ per una famiglia $f_{\underline{X}}(x;\theta)$, non significa che $S(\underline{X})$ sia uno stimatore di $\theta$, ma in generale $S(\underline{X})$ sarà uno stimatore di una qualche funzione di $\theta$. Così nell'Esempio 1.5.2 quando $\theta = (\mu, \sigma)$ si ha $E\{S\} = \{N(\mu^2 + \sigma^2), N\mu\}$.

**Osservazione 1.5.2:** il concetto di statistica sufficiente si può illustrare notando che si tratta di una funzione (o più) le cui superfici di livello $S(\underline{x}) = s$ hanno la particolarità che, condizionando (restringendo) la distribuzione $L(\underline{x};\theta)$ ad una di esse, si trova una distribuzione indipendente da $\theta$. Ci si può chiedere allora quale sia il sistema più esteso di tali superfici. Infatti la variabile $S$ sufficiente che sia costante sul sistema più esteso di superfici, sarà anche quella che varia di meno, pur mantenendo tutta l'informazione su $\theta$: si dirà allora che $S$ è *sufficiente minimale*.

Per comprendere questa problematica si riprenda l'Esempio 1.5.1 e si osservi che chiaramente, per $\theta$ uguale alla media dell'esponenziale, la statistica bidimensionale

$$S' = (\sum_{i=1}^{N-1} X_i, X_N)$$

è sufficiente; tuttavia anche

$$S = (\sum_{i=1}^{N} X_i)$$

è sufficiente e quest'ultima compendia, per così dire, tutta l'informazione su $\theta$ che era contenuta in $S'$.

Per riconoscere se una statistica è sufficiente minimale o no, si può fare ricorso al seguente Lemma.

**Lemma 1.5.1:** se $S(\underline{X})$ è una statistica sufficiente, su ogni superficie di livello $S_s = \{S(\underline{x}) = s\}$ il rapporto di likelihood

$$\frac{L(\underline{x};\theta)}{L(\underline{y};\theta)} \qquad (\underline{x}, \underline{y} \in S_s)$$

è indipendente da $\theta$.

Infatti se $S(\underline{x}) = S(\underline{y}) = s$, per il teorema di fattorizzazione si ha

$$L(\underline{x};\theta) = K(s,\theta) \cdot H(\underline{x})$$
$$L(\underline{y};\theta) = K(s;\theta) \cdot H(\underline{y})$$

e quindi

$$\frac{L(\underline{x};\theta)}{L(\underline{y};\theta)} = \frac{H(\underline{x})}{H(\underline{y})} \ . \tag{1.5.6}$$

Ne segue che, se si vuol trovare la statistica sufficiente minimale per una certa famiglia $f_X(x;\theta)$, basta esaminare il rapporto di likelihood e vedere su quali superfici esso diventa indipendente da $\theta$. Infatti anche per la statistica minimale deve essere $S = \text{cost}$, sulle superfici di tipo (1.5.6). E viceversa, se il sistema $S = \text{cost}$ coincide con quello su cui si verifica (1.5.6), allora $S$ è minimale perché nessun'altra statistica può essere costante su superfici che non siano comprese nel sistema (1.5.6).

**Esempio 1.5.3:** per un campione normale di v.c. i.i.d. con media e varianza incognite, $S = (\sum X_i, \sum X_i^2)$ è sufficiente minimale. Infatti, scritta la likelihood come in (1.5.5) e formando il rapporto, si trova

$$\frac{L(\underline{x};\theta)}{L(\underline{y};\theta)} = e^{(1/2\sigma^2)(\underline{x}^+\underline{x} - \underline{y}^+\underline{y}) + (\mu/\sigma^2)(\underline{e}^+\underline{x} - \underline{e}^+\underline{y})} \ :$$

è chiaro che tale rapporto risulta indipendente da $\mu$ e $\sigma^2$ solo se

$$\underline{x}^+\underline{x} = \sum x_i^2 = \underline{y}^+\underline{y} = \sum y_i^2$$
$$\underline{e}^+\underline{x} = \sum x_i = \underline{e}^+\underline{y} = \sum y_i \ ,$$

cioè su quelle superfici su cui

$$\sum x_i^2 = \text{cost} \ , \sum x_i = \text{cost} \quad .$$

Pertanto $S = (\sum X_i^2 \ , \sum X_i)$ è sufficiente minimale.

**Esempio 1.5.4:** vogliamo dimostrare, mediante il rapporto di likelihood, che per la distribuzione uniforme su $(0, \theta)$ (cfr. Esempio 1.4.3) la statistica $S(\underline{X}) = \max X_i$ è sufficiente minimale. Infatti notiamo che la funzione di likelihood può essere scritta come

$$L(\underline{x};\theta) = \begin{cases} (1/\theta)^N & 0 \leq t = \max x_i \leq \theta \\ 0 & \theta < t = \max x_i \end{cases} \quad .$$

Quindi, formando il rapporto di likelihood e convenendo che questo rimanga costante se $L(\underline{x},\theta) = 0 \ , \ L(\underline{y};\theta) = 0$ entrambe, si ha

$$\frac{L(\underline{x};\theta)}{L(\underline{y};\theta)} = \begin{cases} 1 & 0 \leq t_x \ , \ t_y \leq \theta \\ 0 & \theta < t_x \ , \ 0 \leq t_y \leq \theta \\ \infty & \theta < t_y \ , \ 0 \leq t_x \leq \theta \\ 1 & \theta < t_x \ , \ t_y \ . \end{cases} \quad .$$

Questo rapporto può risultare costante solo se

$$t_x = \max \ x_i = t_y = \max \ y_i \ ,$$

il che dimostra appunto che $S(\underline{X}) = \max X_i$ è sufficiente minimale.

Chiudiamo questo paragrafo aggiungendo una definizione che ci sarà utile in seguito: diremo che la statistica $S(\underline{X})$ è completa per la famiglia $f_X(\underline{x};\theta)$ se la sola funzione $h(S)$ per cui vale

$$E\{h(S)|\theta\} = 0 \quad , \quad \forall\theta \tag{1.5.7}$$

è

$$h(S) \equiv 0 \ , \tag{1.5.8}$$

cioè se (1.5.7) implica (1.5.8).

**Esempio 1.5.5:** supposto che $\sigma^2$ sia nota, $S = (1/N)\sum X_i$ è completa per $N[\mu, \sigma^2]$. Infatti siccome $S = N[\mu, (\sigma^2/N)]$, si ha

$$\begin{aligned} E\{h(S)|\mu\} &= \int_{-\infty}^{+\infty} h(s)\frac{\sqrt{N}}{\sqrt{2\pi}\sigma}e^{-N(s-\mu)^2/(2\sigma^2)}ds = \\ &= \frac{\sqrt{N}}{\sqrt{2\pi}\sigma}e^{\mu^2 N/(2\sigma^2)}\int_{-\infty}^{+\infty}\left\{\ h(s)e^{-Ns^2/(2\sigma^2)}\ \right\}e^{\mu sN/\sigma^2}ds \end{aligned}$$

così che, per un teorema sulla trasformata di Laplace,

$$E\{h(S)|\mu\} = 0 \qquad \forall\mu$$

implica

$$h(s)e^{-s^2N/2\sigma^2} \equiv 0$$

cioè

$$h(s) \equiv 0 \ .$$

## 1.6 Stimatori di minima varianza: efficienza

Nei paragrafi 3) e 4) abbiamo studiato gli stimatori in relazione al loro valore medio e al comportamento asintotico delle loro distribuzioni, cioè quando $N \to \infty$. In questo paragrafo vogliamo studiare gli stimatori in base alla loro dispersione attorno a $\theta$, poiché più piccola sarà questa, maggiore sarà la probabilità che lo stimatore risulti vicino al valore da stimare.

In generale definiamo l'errore quadratico medio di stima (e.q.m.s.) come

$$\begin{aligned} \mathcal{E}^2 &= E\{[T(\underline{X}) - \theta]^2\} = \sigma^2\{T(\underline{X}); \theta\} + [E\{T\} - \theta]^2 = \\ &= \sigma^2\{T(\underline{X}); \theta\} + b^2(\theta) \ . \end{aligned} \tag{1.6.1}$$

Si potrebbe pensare che minimizzando la (1.6.1) si possano direttamente trovare stimatori con proprietà ottimali: si può tuttavia vedere che in generale non esiste una $T(\underline{X})$ che minimizzi la (1.6.1) per tutti i valori di $\theta$, così che per sfruttare tale approccio si è condotti a costruire criteri di ottimalità più complessi.

Resta tuttavia vero che minimizzare $\mathcal{E}^2$, però su una classe più ristretta di possibili stimatori, permette in alcuni casi di ricavare facilmente risultati interessanti e in particolare fornisce esempi in cui si dimostra che uno stimatore biased può talvolta essere meno disperso attorno a $\theta$ di tutti gli stimatori corretti dello stesso parametro.

**Esempio 1.6.1:** supponiamo di voler stimare la varianza $\sigma^2$ di una popolazione normale, di cui sia incognita anche la media. Se si restringe la classe degli stimatori a quelli aventi le seguenti caratteristiche:

- siano invarianti per traslazione

$$T(X_i + \delta) = T(X_i) \ ,$$

- siano quadratici omogenei nella v. campionaria

$$T = \sum \lambda_{ik} X_i X_k \ ,$$

- siano simmetrici rispetto a permutazioni degli indici

$$\{\pi_1, \ldots, \pi_N\} \ ,$$

si può vedere che questi hanno necessariamente la forma

$$T(\underline{X}) = \lambda \cdot \sum (X_i - \mathcal{M})^2 = \lambda(N-1)\overline{\mathcal{S}}^2 \ . \tag{1.6.2}$$

Scegliendo $\lambda = 1/(N-1)$, si ha l'unico stimatore corretto di tale classe, mentre, lasciando libero $\lambda$, si ha in generale un bias

$$b = [\lambda(N-1) - 1]\sigma^2 \ . \tag{1.6.3}$$

Ricordando che

$$\sigma^2(\overline{\mathcal{S}}^2) = \frac{2\sigma^4}{N-1} \ , \tag{1.6.4}$$

si ha per $\mathcal{E}^2$ in tale classe l'espressione

$$\mathcal{E}^2 = 2\lambda^2(N-1)\sigma^4 + [\lambda(N-1) - 1]^2\sigma^4 \ . \tag{1.6.5}$$

Il minimo dell'e.q.m.s. è allora ottenuto per

$$\lambda = \frac{1}{N+1}$$

cioè per

$$T(\underline{X}) = \frac{1}{N+1}\sum(X_i - \mathcal{M})^2 \ \ : \tag{1.6.6}$$

in corrispondenza il valore di $\mathcal{E}^2$ è

$$\mathcal{E}^2 = \frac{2\sigma^4}{N+1} \ ,$$

che come si vede è inferiore a (1.6.4).

Volendoci limitare agli stimatori corretti, si ha $b(\theta) \equiv 0$ così che l'e.q.m.s. di $T(\underline{X})$ coincide con la sua varianza. Ci si può chiedere se esista tra gli stimatori corretti quello di *minima varianza*, poiché è chiaro che tale stimatore sia da preferirsi tra quelli della sua classe.

L'esistenza dello stimatore unbiased di minima varianza può essere provata come un teorema che, sebbene non di grande uso pratico, serve ad inquadrare bene l'argomento.

**Teorema 1.6.1:** sia data la likelihood $L(\underline{x};\theta)$; supponiamo che esista una statistica $S(\underline{x})$ sufficiente minimale completa per $\theta$ e per $L$, e sia $V$ uno stimatore unbiased qualsiasi di $\theta$. Allora,

$$T = E\{V|S\} \tag{1.6.7}$$

è uno stimatore di minima varianza tra gli stimatori corretti; inoltre esso è unico.

Notiamo, prima di dare la dimostrazione, che in conseguenza della definizione (1.6.7) come media condizionata, $T$ risulta una funzione di $S$

$$T = T(S)\ , \tag{1.6.8}$$

come è logico aspettarsi proprio perché $S$ è una statistica sufficiente minimale.

Il teorema si dimostra a passi:

**a)** *$T$ è uno stimatore corretto di $\theta$.* Infatti

$$E\{T\} = E\{T(S)\} = \underset{S}{E}\{T(S)\} = \underset{S}{E}\{E\{V|S\}\} = E\{V\} = \theta\ ;$$

**b)** *$T$ è indipendente da $V$.*

cioè se $V_1, V_2$ sono due qualsiasi stimatori corretti di $\theta$, e se $T_1 = E\{V_1|S\},\ T_2 = E\{V_2|S\}$, allora $T_1 = T_2$.

Infatti, posto $h(S) = T_1(S) - T_2(S)$, si ha che

$$E\{h(S)\} = \theta - \theta = 0 \quad , \quad \forall\theta$$

e per la completezza di $S$, segue $h(S) \equiv 0$.

**c)** *$T$ è di minima varianza.*

Infatti, sia $V$ un qualunque stimatore unbiased; si ha

$$\begin{aligned}\sigma^2(V) &= E\{(V-\theta)^2\} = E\{(V-T+T-\theta)^2\} = \qquad (1.6.9)\\ &= E\{(V-T)^2\} + \sigma^2(T) + 2E\{(V-T)(T-\theta)\}\ .\end{aligned}$$

D'altro canto, usando la relazione

$$\begin{aligned}&E\{\bullet\} = \underset{S}{E}\{E\{\bullet|S\}\}\ ,\\ &E\{(V-T)(T-\theta)\} = \underset{S}{E}\{[T(S)-\theta][E\{V|S\}-T(S)]\} \equiv 0\ ,\end{aligned}$$

così che dalla (1.6.9) deriveremo

$$\sigma^2(V) = \sigma^2(T) + E\{(V-T)^2\}\ , \qquad (1.6.10)$$

ovvero

$$\sigma^2(V) \geq \sigma^2(T)\ , \qquad (1.6.11)$$

cioè $T$ è di minima varianza.

**d)** *$T$ è unico.*

Infatti la (1.6.11) può valere col segno "=" solo se

$$E\{(V-T)^2\} = 0\ ,\ \text{cioè se } V = T \text{ con } P = 1\ .$$

**Corollario 1.6.1:** come conseguenza del Teorema 1.6.1 si vede subito anche che se uno stimatore $\overline{T}$ è funzione della statistica sufficiente minimale ed è allo stesso tempo corretto, allora esso coincide con lo stimatore corretto di minima varianza. Infatti, se $\overline{T} = \overline{T}(S)$, e inoltre $E\{\overline{T}\} = \theta$, si ha anche

$$T = E\{\overline{T}|S\} = E\{\overline{T}(S)|S\} = \overline{T}(S)\ .$$

**Esempio 1.6.2:** la varianza campionaria corretta $\overline{\mathcal{S}}^2$ è lo stimatore corretto di $\sigma^2$ di minima varianza per una famiglia normale. Infatti è

$$\overline{\mathcal{S}}^2 = \frac{1}{N-1}(\sum X_i^2) - \frac{1}{N(N-1)}(\sum X_i)^2 \ ,$$

cioè $\overline{\mathcal{S}}^2$ è funzione della statistica sufficiente, per la famiglia normale, $S = (\sum X_i, \sum X_i^2)$; inoltre $\overline{\mathcal{S}}^2$ è corretto e quindi, in base al Corollario 1.6.1, esso è lo stimatore di minima varianza.

Si noti che a questo punto resta provato anche che lo stimatore (1.6.6) ha e.q.m.s. inferiore ad ogni simatore corretto.

**Osservazione 1.6.1:** si noti che se $\theta = g(\varphi)$ dà luogo ad una nuova parametrizzazione della likelihood, ovvero se $\varphi$ è un nuovo parametro che dipende in modo invertibile da $\theta$, e se $T = h(S)$ è uno stimatore corretto di $\varphi$, allora esso è anche di minima varianza.

Infatti, se $L(\underline{x}, \theta)$ è la likelihood in funzione di $\theta$, posto

$$L'(\underline{x}; \varphi) = L(\underline{x}; g(\varphi)) \ ,$$

è facile vedere tramite il teorema di fattorizzazione che se $S$ è sufficiente per $\theta$, è anche sufficiente per $\varphi$. Pertanto l'affermazione consegue dal Corollario 1.6.1.

Anche quando esiste un minimo della varianza tra gli stimatori corretti di un parametro $\theta$, in genere non è facile calcolare tale minimo, e non è quindi semplice giudicare, per uno stimatore dato, quanto esso sia buono. Si preferisce allora ricorrere ad una maggiorazione, detta limite di Cramer-Rao, che specifica un limite inferiore per la varianza di uno stimatore, biased o unbiased che sia, limite che può essere il minimo od un numero inferiore ad esso.

**Teorema 1.6.2**: (limite di Cramer-Rao). Data una v. campionaria di varabili i.i.d., con una densità di probabilità $f_X(x; \theta)$ regolare in $\theta$, e definita la variabile, dipendente da $\theta$,

$$U = \frac{\partial}{\partial\theta} \log\ L(\underline{X};\theta) = \sum_i U_i = \sum_i \frac{\partial}{\partial\theta} \log\ f_X(X_i;\theta)\ , \tag{1.6.12}$$

per un qualsiasi stimatore $T$, con bias $b(\theta)$, vale la disuguaglianza

$$\sigma^2\{T\} \geq \frac{[1+b'(\theta)]^2}{E\{U^2\}} = -\frac{[1+b'(\theta)]^2}{E\{\partial_\theta U\}}\ ; \tag{1.6.13}$$

inoltre (1.6.13) vale col segno di uguaglianza se e solo se esiste una relazione lineare tra $T$ e $U$.

Per dimostrare la (1.6.13) partiamo dall'identità

$$\int_{R^N} L(\underline{x};\theta) d_N x = 1 \tag{1.6.14}$$

e supponiamo che $L = \Pi_i f_X(x_i;\theta)$ sia così regolare da poter derivare due volte sotto l'integrale (1.6.14). Alla prima derivazione otteniamo, ricordando la definizione (1.6.12),

$$\begin{aligned} & \int_{R^N} \partial_\theta L(\underline{x};\theta) d_N x = \\ = \quad & \int_{R^N} \sum_i \partial_\theta f_X(x_i;\theta) \prod_{k\neq i} f_X(x_k;\theta) d_N x = \\ = \quad & \int_{R^N} \sum_i \frac{\partial_\theta f_X(x_i;\theta)}{f_X(x_i;\theta)} L(\underline{x};\theta) d_N x = \\ = \quad & E\{U\} = 0\ ; \end{aligned} \tag{1.6.15}$$

cioè $U$ è una variabile a media nulla per ogni $\theta$.

Derivando una seconda volta e ragionando come in (1.6.15), si ha

$$\begin{aligned} & \int_{R^N} \{\partial_\theta U\} L(\underline{x};\theta) d_N x + \int_{R^N} U \partial_\theta (L(\underline{x};\theta)) d_N x = \\ & = E\{\partial_\theta U\} + E\{U^2\} = 0\ ; \end{aligned} \tag{1.6.16}$$

ovvero, ricordando anche (1.6.15),

$$E\{U^2\} = \sigma^2\{U\} = -E\{\partial_\theta U\} \ . \tag{1.6.17}$$

Ora si riparte dall'identità

$$\theta + b(\theta) = \int_{R^N} T(\underline{x}) L(\underline{x};\theta) d_N x \ \ : \tag{1.6.18}$$

supposto di poter derivare sotto integrale, si trova

$$1 + b'(\theta) = \int_{R^N} T(\underline{x}) \partial_\theta L(\underline{x};\theta) d_N x = E\{T(\underline{X}) U\} \ . \tag{1.6.19}$$

Ricordando che $E\{U\} = 0$, la (1.6.19) si può anche scrivere

$$1 + b'(\theta) = \sigma(T, U) \ \ :$$

infine, rammentando che il coefficiente di correlazione tra due variabili soddisfa la disuguaglianza $\rho^2 \leq 1$, e $\rho^2 = 1$ se e solo se esiste una relazione lineare tra le due variabili, si può scrivere

$$[1 + b'(\theta)]^2 \leq \sigma^2(T)\sigma^2(U) \ , \tag{1.6.20}$$

che, tramite la (1.6.17) ci dà il risultato richiesto.

Si noti che per stimatori corretti la (1.6.20) diventa semplicemente

$$\sigma^2(T) \geq -\frac{1}{E\{\partial_\theta U\}} \ .$$

La funzione

$$I(\theta) = -E\{\partial_\theta U\} = -E\{\partial_\theta^2 \log L(X;\theta)\} = \sigma^2(U) \tag{1.6.21}$$

è detta *informazione* in quanto è una misura del massimo di informazione ottenibile dai dati empirici sul parametro $\theta$.

**Osservazione 1.6.2:** il limite inferiore di Cramer-Rao è significativo, nel senso che esistono famiglie di distribuzioni per le quali esso è raggiungibile, cioè è un vero minimo. Infatti la (1.6.20) può valere col segno di uguaglianza se e solo se

$$U = \partial_\theta \log L(\underline{x};\theta) = AT(\underline{x}) + C \ : \tag{1.6.22}$$

in tale relazione $A$ e $C$ possono essere funzioni di $\theta$ senza con ciò alterare il fatto che $\sigma^2(U,T) = \sigma^2(U)\sigma^2(T)$. Integrando la (1.6.22) rispetto a $\theta$ si ha

$$\log L(\underline{x};\theta) = aT(\underline{x}) + c + d \tag{1.6.23}$$

dove

$$a = \int A \ d\theta \quad , \quad c = \int C \ d\theta$$

e la costante di integrazione $d$ può essere funzione di $\underline{x}$ in quanto $\partial_\theta d(\underline{x}) \equiv 0$.

Riscritta, la (1.6.23) ci dà

$$L(\underline{x};\theta) \equiv \exp\{a(\theta)T(\underline{x}) + c(\theta) + d(\underline{x})\} \ ; \tag{1.6.24}$$

una qualsiasi likelihood della forma (1.6.24) è detta appartenere alla famiglia esponenziale. Dunque, il limite minimo di varianza può essere raggiunto solo per funzioni di densità della famiglia esponenziale, che tuttavia come è facile provare comprende molte importanti variabili, quali ad esempio la normale, la $\Gamma$, la binomiale e la poissoniana.

**Osservazione 1.6.3:** per uno stimatore corretto che non raggiunga il limite di Cramer-Rao si potrà introdurre un indice, che misuri l'*efficienza* dello stimatore stesso, nei termini

$$\eta = [\sigma^2(T;\theta)I(\theta)]^{-1} \ . \tag{1.6.25}$$

È chiaro che

$$0 \leq \eta \leq 1 \ , \tag{1.6.26}$$

e l'efficienza può essere pari ad 1 se e solo se $f_X(x;\theta)$ è della famiglia esponenziale.

**Osservazione 1.6.4:** la trattazione del limite di Cramer-Rao (1.6.13) è essenzialmente monodimensionale. Essa vale ovviamente anche nel caso in cui $\theta$ sia la componente di un vettore e $T$ sia il corrispondente stimatore. Tuttavia nel caso multidimensionale è possibile ottenere un limite inferiore più stretto, che ci proponiamo qui di dare.

Come già nel caso monodimensionale, osserviamo anche qui che si può scrivere

$$\partial_{\theta k} L(\underline{x};\underline{\theta}) = \sum_i \frac{\partial_{\theta k} f_X(x_i;\underline{\theta})}{f_X(x_i;\underline{\theta})} L(\underline{x};\underline{\theta}) \ :$$

così, introdotta l'operazione vettoriale

$$\partial_\theta F(\underline{\theta}) = [\partial_{\theta 1} F(\underline{\theta}) \dots \partial_{\theta p} F(\underline{\theta})]^+ \tag{1.6.27}$$

e definendo

$$\underline{U}(\underline{x};\underline{\theta}) = \partial_\theta \log L(\underline{x};\underline{\theta}) \ , \tag{1.6.28}$$

si può scrivere

$$\partial_\theta L(\underline{x};\underline{\theta}) = \underline{U} \ L(\underline{x};\underline{\theta}) \ . \tag{1.6.29}$$

Supporremo $L$ abbastanza regolare in $\theta$ per poter effettuare tutte le derivazioni sotto integrale che ci serviranno. Pertanto, partendo da

$$\int L(\underline{x};\underline{\theta}) d_N x = 1$$

e differenziando, troviamo

$$E\{\underline{U}\} = \int \underline{U} \ L(\underline{x};\underline{\theta}) d_N x = 0 \ . \tag{1.6.30}$$

Se ora applichiamo l'operatore $\partial_\theta$ *a* $E\{\underline{U}\}^+$, notando anche che $\partial_\theta \partial_\theta^+ F$ è la matrice delle derivate seconde di $F$, si ha

$$\int \partial_\theta \underline{U}^+ L(\underline{x};\underline{\theta}) d_N x + \int \underline{U}\ \underline{U}^+ L(\underline{x};\underline{\theta}) d_N x = 0\ . \tag{1.6.31}$$

Poiché vale la (1.6.30), si vede che

$$E\{\underline{U}\ \underline{U}^+\} = C_{UU}\ ;$$

inoltre definiamo come *matrice di informazione* (di Fisher)

$$\begin{aligned} \mathcal{I}(\theta) &= -\int \partial_\theta \underline{U}^+ L(\underline{x};\underline{\theta}) d_N x = \\ &= -\int \{\partial_\theta \partial_\theta^+ \log L\} L d_N x = \\ &= -E\{\partial_\theta \partial_\theta^+ \log L\}\ . \end{aligned} \tag{1.6.32}$$

In questo modo la (1.6.31) diventa

$$C_{UU} = \mathcal{I}(\theta)\ . \tag{1.6.33}$$

sia ora $\underline{T}(\underline{x})$ uno stimatore corretto, $p$-dimensionale, di $\underline{\theta}$: dall'identità

$$\underline{\theta}^+ = \int \underline{T}(\underline{x})^+ L(\underline{x};\theta) d_N x\ ,$$

applicando $\partial_\theta$ si trova

$$I = \int \underline{U}\ \underline{T}^+ L(\underline{x};\underline{\theta}) d_N x\ \ :$$

dove $I$ è la matrice identità, da non confondersi con l'informazione scalare qui sempre indicata con $I(\theta)$ (vedi (1.6.21)).

Ricordando la (1.6.30), questa si può scrivere

$$C_{UT} = E\{\underline{U}(\underline{T} - \mu_T)^+\} = I\ . \tag{1.6.34}$$

Moltiplichiamo a destra e a sinistra per due vettori $\alpha^+$ e $\beta$

$$\alpha^+ C_{UT}\beta = E\{\alpha^+ \underline{U}(\underline{T}-\mu_T)^+\beta\} = \alpha^+\beta \ : \tag{1.6.35}$$

d'altro canto, come nel caso monodimensionale

$$E\{\alpha^+ \underline{U}(\underline{T}-\mu_T)^+\beta\}^2 \leq \sigma^2\{\alpha^+\underline{U}\}\sigma^2(\beta^+\underline{T}) \ ,$$

da cui, applicando la propagazione della covarianza

$$(\alpha^+\beta)^2 = E\{\alpha^+ \underline{U}(\underline{T}-\mu_T)^+\beta\}^2 \leq (\alpha^+ C_{UU}\alpha)(\beta^+ C_{TT}\beta) \ . \tag{1.6.36}$$

La (1.6.36) vale per ogni $\alpha$ e ogni $\beta$ e se vale col segno di uguaglianza significa che esiste una relazione lineare tra $\alpha^+U$ e $\beta^+T$.

Si noti che, scelto ad esempio $\alpha^+ = [1\ 0 \ldots 0]$ e $\beta = \alpha$, si ritrova lo stesso limite monodimensionale

$$\sigma^2(T_1) = \beta^+ C_{TT}\beta \geq \frac{1}{\sigma^2(U_1)} \ :$$

tuttavia dalla (1.6.36) è possibile ottenere un limite inferiore più forte. Infatti, riscritta la (1.6.36) come

$$\beta^+ C_{TT}\beta \geq \frac{(\alpha^+\beta)^2}{\alpha^+ C_{UU}\alpha} \ , \tag{1.6.37}$$

si può cercare per ogni $\beta$ dato l'$\alpha$ che rende massimo il secondo membro. Si può vedere che tale $\alpha$ è dato da

$$\alpha = C_{UU}^{-1}\beta$$

così che si giunge a

$$\sigma^2(\beta^+\underline{T}) = \beta^+ C_{TT}\beta \geq \beta^+ C_{UU}^{-1}\beta = \beta^+ \mathcal{I}(\theta)^{-1}\beta \ , \tag{1.6.38}$$

ovvero

$$C_{TT} \geq \mathcal{I}(\theta)^{-1} \ . \tag{1.6.39}$$

Questa relazione dice sostanzialmente che la varianze di $\beta^+ T$, che è uno stimatore corretto di $\beta^+\theta$, ha come limite inferiore $\beta^+\mathcal{I}^{-1}(\theta)\beta$. In particolare scegliendo

$$\begin{array}{rl} & \ldots\ldots i\ldots\ldots \\ \beta_i^+ = & [0\ldots 010\ldots 0] \end{array}$$

si trovano i limiti inferiori delle varianze delle singole componenti $T_i$.

Si può osservare che $\mathcal{I}(\theta)$ ha sulla diagonale principale $i_k(\theta) = \sigma^2(U_k)$, così che, quando $\mathcal{I}(\theta)$ sia diagonale, la (1.6.38) riproduce i singoli limiti monodimensionali.

## 1.7 Stime di massima verosimiglianza

Fino ad ora abbiamo studiato le stime dei parametri $\theta$ in base alle loro proprietà e solo in un caso abbiamo definito un criterio generale per la ricerca di tali stime: nel caso degli stimatori corretti di minima varianza. Come si è visto in quel caso, però, la soluzione generale del problema è ottenibile quando si conosca già uno stimatore corretto, cosa non sempre facile.

Vedremo in questo paragrafo invece un diverso principio di scelta degli stimatori detto di *massima verosimiglianza* (maximum likelihood). Questo criterio, inventato da Fisher, stabilisce di scegliere come stimatore di $\theta$ la statistica $T(\underline{X})$ tale che

$$L(\underline{x}; T(\underline{x})) = \max_{\theta} L(\underline{x}; \theta) \ . \tag{1.7.1}$$

Essenzialmente tale principio stabilisce di scegliere per un campione $\underline{x}$ dato, quel valore di $\theta$ per cui massima era la probabilità di estrarre proprio quell'$\underline{x}$.

Poiché il logaritmo è una funzione monotona, spesso anziché il massimo di $L(\underline{x}; \theta)$ si preferisce cercare il massimo di $\log L(\underline{x}; \theta)$. Condizione

necessaria[2] di massimo è l'annullarsi della derivata (o del gradiente) di $\log L(\underline{x};\theta)$ così che (1.7.1) viene tradotto nell'equazione

$$U(\underline{x};T(\underline{x})) = \partial_\theta \log L(\underline{x};\theta)|\theta = T(x) \equiv 0 \tag{1.7.2}$$

da risolversi rispetto a $T(\underline{x})$.

**Esempio 1.7.1:** per i campioni normali con $\theta = (\mu, \sigma^2)$ si trovano le stime $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ di m.l. delle equazioni

$$\log L(\underline{x};\theta) = -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2 + \text{cost} \tag{1.7.3}$$

$$U = \left| \begin{array}{l} \frac{1}{\sigma^2}\sum(x_i-\mu) \\ -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(x_i-\mu)^2 \end{array} \right| \tag{1.7.4}$$

$$U(\underline{X};\hat{\mu},\hat{\sigma}^2) = 0 \ . \tag{1.7.5}$$

La soluzione $\hat{\theta}$ di (1.7.5) è

$$\left\{ \begin{array}{l} \hat{\mu} = \frac{1}{N}\sum X_i \\ \hat{\sigma}^2 = \frac{1}{N}\sum(X_i-\hat{\mu})^2 \end{array} \right. \ . \tag{1.7.6}$$

Si può notare che la stima di m.l. non è in generale unbiased.

**Esempio 1.7.2:** siano $X_i$ variabili normali con la stessa varianza e medie $\mu_i$ funzioni lineari di un numero inferiore di parametri

$$\underline{X} = N[\underline{\mu};\sigma^2] \tag{1.7.7}$$

$$\underline{\mu} = A\lambda \quad , \quad \mu = \left| \begin{array}{c} \mu_1 \\ \vdots \\ \mu_N \end{array} \right| \ , \tag{1.7.8}$$

$$A = \left| \begin{array}{ccc} a_{11} & \dots & a_{1p} \\ \vdots & & \vdots \\ a_{N1} & \dots & a_{Np} \end{array} \right| , \lambda = \left| \begin{array}{c} \lambda_1 \\ \vdots \\ \lambda_p \end{array} \right| \quad p < n \ ;$$

[2]Oltre ad imporre la condizione necessaria, occorrerebbe anche verificare che la soluzione ottenuta da (1.7.2) corrisponda ad un vero massimo. Per alcune distribuzioni, come la normale, ciò è automaticamente verificato.

supponiamo inoltre che $A$ sia rango pieno[3], così che $A^+A$ sia invertibile. Posto $\theta^+ = (\lambda_1, \ldots, \lambda_p, \sigma^2)$ otteniamo la stima m.l. dalle equazioni

$$\log L = -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\underline{x} - A\lambda)^+(\underline{x} - A\lambda) + \text{cost} \tag{1.7.9}$$

$$U = \left| \begin{array}{c} \frac{\partial}{\partial\lambda}\log L \\ \frac{\partial}{\partial\sigma^2}\log L \end{array} \right| = \left| \begin{array}{c} \frac{1}{\sigma^2}A^+(\underline{x} - A\lambda) \\ -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(\underline{x} - A\lambda)^+(x - A\lambda) \end{array} \right| \tag{1.7.10}$$

$$U(\underline{x}; \hat{\lambda}, \hat{\sigma}^2) = 0 \ . \tag{1.7.11}$$

Il risultato è

$$\hat{\lambda} = (A^+A)^{-1}A^+\underline{X} \tag{1.7.12}$$

$$\hat{\sigma}^2 = \frac{1}{N}(\underline{X} - A\lambda)^+(\underline{X} - A\hat{\lambda}) \ . \tag{1.7.13}$$

Notiamo che dei due stimatori, il primo è corretto

$$E\{\hat{\lambda}\} = (A^+A)^{-1}A^+E\{\underline{X}\} = (A^+A)^{-1}A^+A\lambda = \lambda \ , \tag{1.7.14}$$

mentre il secondo non è corretto.

Infatti notiamo che $E\{X - A\hat{\lambda}\} = 0$ per la (1.7.14), così che si può scrivere, posto $U = X - A\hat{\lambda}$.

$$E\{(X - A\hat{\lambda})^+(X - A\hat{\lambda})\} = TrC_{UU} \ .$$

Poiché [$I_{(N)}$ = identità a $N$ dimensioni].

[3]In effetti $A$ di rango pieno equivale a dire che $A\lambda = 0 \to \lambda = 0$. D'altro canto se $(A^+A)\lambda = 0$ allora è anche $0 = \lambda^+A^+A\lambda = (A\lambda)^+A\lambda = |A\lambda|^2 \Rightarrow A\lambda = 0 \to \lambda = 0$: poiché $A^+A$ è quadrata e non ammette soluzioni del sistema omogeneo associato, $A^+A$ è invertibile.

$$U = [I_{(N)} - A(A^+A)^{-1}A^+]X$$

si ha

$$\begin{aligned} C_{UU} &= [I - A(A^+A)^{-1}A^+]\sigma^2 I[I - A(A^+A)^{-1}A^+] = \\ &= \sigma^2[I - A(A^+A)^{-1}A^+] \ . \end{aligned} \tag{1.7.15}$$

Ora notiamo che

$$TrI_{(N)} = N \ ,$$

mentre

$$TrA(A^+A)^{-1}A^+ = Tr(A^+A)^{-1}A^+A = TrI_{(p)} = p \ .$$

Riassumendo, dalla (1.7.15) si trova

$$TrC_{UU} = \sigma^2(N-p) \ :$$

pertanto tornando alla (1.7.13) si vede che

$$E\{\hat{\sigma}^2\} = \frac{E\{(X - A\hat{\lambda})(X - A\hat{\lambda})\}}{N} = \frac{(N-p)}{N}\sigma^2 \ .$$

Per fortuna è facile correggere tale bias ottenendo come stimatore corretto

$$\hat{\sigma}_o^2 = \frac{(X - A\hat{\lambda})^+(X - A\hat{\lambda})}{N-p} \ . \tag{1.7.16}$$

Le formule (1.7.12) e (1.7.16) forniscono la soluzione di un problema detto "di modello lineare" nell'ambito delle variabili normali. L'argomento sarà ripreso più in generale trattando del metodo dei minimi quadrati.

Gli stimatori di m.l. non hanno in generale nessuna proprietà ottimale, a parte il significato intuitivo legato alla loro definizione.

Tuttavia essi restano importanti perché godono di diverse proprietà ottimali almeno asintoticamente, fatte salve alcune condizioni di regolarità (differenziabilità) della funzione di likelihood.

In effetti le equazioni per le stime di m.l. si scrivono come

$$U(\underline{X};\hat{\theta}) = 0 \ ; \ [U = (\partial/\partial\theta)\log L] \ ; \tag{1.7.17}$$

ora notiamo che supposto che lo stimatore $\hat{\theta} = T(\underline{X})$ ottenuto dalla (1.7.17) abbia un e.q.m.s. piccolo (rispetto ad una zona di variabilità significativa di $U$ in funzione di $\theta$), si può sostituire l'equazione esatta (1.7.17), con quella linearizzata in corrispondenza al valore esatto di $\theta$,

$$U(\underline{X};\theta) + \frac{\partial U(\underline{X};\theta)}{\partial\theta}(\hat{\theta} - \theta) \cong 0 \ . \tag{1.7.18}$$

Dalla (1.7.18), che vale solo in modo approssimato, si possono ricavare interessanti proprietà, approssimate per l'appunto, dello stimatore $\theta$.

Infatti supponiamo ancora che valga

$$\sqrt{\sigma^2\{(\partial U_j)/(\partial\theta_k)\}} << E\{(\partial U_j)/(\partial\theta_k)\} = \mathcal{I}_{jk}(\theta) \qquad \forall j,k \ ,$$

nel caso monodimensionale, o analoghe condizioni per il caso multidimensionale; in tal caso la (1.7.18) può essere ulteriormente approssimata con

$$-\mathcal{I}(\theta)(\hat{\theta} - \theta) + U(\underline{X},\theta) \cong 0 \ , \tag{1.7.19}$$

che ha soluzione

$$\hat{\theta} \cong \theta + \mathcal{I}(\theta)^{-1}U(\underline{X},\theta) \ . \tag{1.7.20}$$

Ricordando che $E\{U\} = 0$ e $C_{UU} = \mathcal{I}(\theta)$, dalla (1.7.20) ricaviamo

$$E\{\hat{\theta}\} \cong \theta \tag{1.7.21}$$
$$C_{\hat{\theta}\hat{\theta}} \cong \mathcal{I}(\theta)^{-1} C_{UU} \mathcal{I}(\theta)^{-1} = \mathcal{I}(\theta)^{-1} \ . \tag{1.7.22}$$

Dunque, nell'ambito delle approssimazioni fatte, $\hat{\theta}$ è quasi unbiased e la sua matrice di covarianza è circa $\mathcal{I}(\theta)^{-1}$: si osservi ancora che più grande sarà $\mathcal{I}(\theta)$ più piccola sarà la matrice di covarianza di $\hat{\theta}$ e quindi migliore (più informativa) la stima.

In particolare notiamo che

$$U(\theta) = \partial_\theta \log L(\underline{x};\theta) = \sum \partial_\theta \log f_{X_j}(x_j;\theta) = \sum U_j(\theta) \tag{1.7.23}$$

dove le $U_j(\theta)$ sono variabili i.i.d. (possibilmente multidimensionali): supposto che per le $U_j$ si possa applicare il teorema centrale della statistica (ad esempio $U_j$ abbia momento 3° finito) si avrà che $U$ è asintoticamente normale, e anzi per le (1.7.21), (1.7.22)

$$\hat{\theta} \cong N[\theta; \mathcal{I}(\theta)^{-1}] \ . \tag{1.7.24}$$

È interessante notare che, se per ogni variabile $U_j$ si pone $E\{U_j U_j^+\} = = C_0 = \mathcal{I}_o(\theta)$, si ha

$$\mathcal{I}(\theta) = C_{UU} = N\mathcal{I}_o(\theta) \ , \tag{1.7.25}$$

così che la (1.7.24) può anche essere scritta come

$$\hat{\theta} \cong N\left[\theta; \frac{\mathcal{I}_o(\theta)^{-1}}{N}\right] \ . \tag{1.7.26}$$

Tale relazione mostra che ci si può aspettare che $\hat{\theta}$ sia anche consistente, sebbene la dimostrazione accennata sia puramente euristica: in effetti $\theta$ gode di tale proprietà sotto opportune ipotesi di regolarità per $f_X(x;\theta)$.

Il ragionamento fatto qui in modo puramente approssimato può essere reso rigoroso prendendo il lim per $N \to \infty$.

**Osservazione 1.7.1:** si vuole dimostrare che per lo stimatore di m.l. monodimensionale $\hat{\theta}$, soluzione dell'equazione $U(\hat{\theta}) = 0$, vale la proprietà asintotica, per $N \to \infty$,

$$\hat{\theta} \cong N\left[\theta, \frac{i(\theta)^{-1}}{N}\right] \ ,$$

con

$$i(\theta) = E\{[\partial_\theta \log f_X(X;\theta)]^2\} = \sigma^2(U_j) \quad , \quad \forall j$$

almeno sotto l'ipotesi restrittiva che le variabili i.i.d.
$U_j = \partial_\theta \log f_{X_j}(X_j;\theta), \partial_\theta U_j = U_j'$ soddisfino le condizioni del teorema centrale della statistica e le $\partial_\theta^2 U_j = U_j''$ siano limitate con $P = 1$. In effetti si può notare che l'equazione di $\hat{\theta}$ può essere scritta

$$0 = U(\hat{\theta}) = U(\theta) + U'(\theta)(\hat{\theta} - \theta) + \frac{1}{2}U''(\theta^*)(\hat{\theta} - \theta)^2$$

con $|\theta^* - \theta| \leq |\hat{\theta} - \theta|$. Dividendo per $\frac{1}{\sqrt{N}}$ e posto

$$\begin{aligned}
A_N &= \frac{U(\theta)}{\sqrt{N}} = \frac{\sum U_j(\theta)}{\sqrt{N}} \cong N\,[0, i(\theta)] \\
B_N &= \frac{U'(\theta)}{N} = \frac{\sum U_j'(\theta)}{N} \cong N\left[-i(\theta), \frac{\text{cost}}{N}\right] \\
C_N &= \frac{1}{2}\frac{U''(\theta)}{N^{3/2}} = \frac{1}{2}\frac{\sum U_j''(\theta)}{N^{3/2}} \cong N\left[\frac{\text{cost}}{\sqrt{N}}\ ,\ \frac{\text{cost}}{N^2}\right] \\
\xi_N &= \sqrt{N}(\hat{\theta} - \theta)
\end{aligned}$$

si ha che $\xi_N$ risolve l'equazione

$$0 = A_N + B_N\xi_N + C_N\xi_N^2 \ ,$$

così che per $N \to \infty, \xi_N$ tende in probabilità alla soluzione $\xi$ dell'equazione

$$0 = \sqrt{i(\theta)}Z - i(\theta)\xi$$

dove $Z = \lim_{N\to\infty} i(\theta)^{-1/2}A_N$ è una normale standardizzata.

# 2 Il metodo dei minimi quadrati

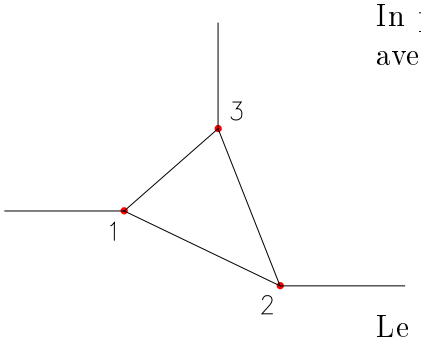
## 2.1 Introduzione

L'idea di basare la teoria della stima sul principio di massima verosimiglianza, benché attraente, presenta alcuni seri incovenienti:

- in primo luogo bisogna supporre che sia nota la famiglia di distribuzioni da cui si estrae il campione;
- inoltre il problema di trovare lo stimatore applicando le equazioni del massimo, è risolubile in forma analitica solo in pochi casi particolari;
- infine, passando a soluzioni numeriche (cioè accontentandoci del valore numerico dello stimatore in corrispondenza ai valori numerici del campione), si vede che, a parte casi particolari o con un numero piccolo di incognite, il problema è insolubile.

Si pone quindi il problema di ottenere per certe grandezze osservabili $y_i\{i = 1, \ldots, n\}$, cioè grandezze sulle quali si possono ricavare informazioni empiriche (osservazioni), stime che non implicano necessariamente la conoscenza di tutta la distribuzione ma solo certe sue caratteristiche, ad esempio medie e varianze-covarianze.

**Esempio 2.1.1:** si consideri una maglia di una rete elettrica e si supponga di aver eseguito misure di differenze di potenziale ai nodi.
In particolare (fig. 2.1.1) si supponga di aver osservato:




$$\begin{aligned} y_1 &= V_{12} = V_2 - V_1 \\ y_2 &= V_{23} = V_3 - V_2 \\ y_3 &= V_{31} = V_3 - V_1 \end{aligned} \qquad (2.1.1)$$

Le osservazioni sono effettuate commettendo errori di misura e quindi vanno

descritte da variabili casuali $Y_1, Y_2, Y_3$ di cui le variabili "osservabili" $y_1, y_2, y_3$ definite in (2.1.1) sono i valori medi:

$$E\{Y_i\} = y_i \ . \tag{2.1.2}$$

In virtù della struttura fisica del sistema sottoposto ad osservazione, le medie $y_i$ non possono assumere tre valori arbitrari, ma a causa della loro forma (2.1.1) (ovvero per la definizione stessa di potenziale) deve essere

$$y_1 + y_2 + y_3 = 0 \tag{2.1.3}$$

viceversa si potrebbe anche notare che dati tre numeri qualsiasi che soddisfano la (2.1.3) è possibile pensare ad una maglia ideale che ha quei numeri come differenza di potenziale: in altri termini la (2.1.3) è la condizione necessaria e sufficiente perché i tre numeri $y_1, y_2, y_3$ descrivano le differenze di potenziale di una maglia di vertici 1,2,3.

Naturalmente la presenza di una condizione come la (2.1.3) dà una qualche informazione sugli errori di misura connessi che non saranno conosciuti individualmente, ma in una loro combinazione lineare. Infatti, se indichiamo con $\varepsilon_i$ gli errori,

$$Y_i = y_i + \varepsilon_i \quad , \quad E\{\varepsilon_i\} = 0 \ , \tag{2.1.4}$$

deve essere

$$\sum Y_i = \sum y_i + \sum \varepsilon_i = \sum \varepsilon_i \ , \tag{2.1.5}$$

in conseguenza della (2.1.3).

Inoltre, conoscendo i procedimenti di misura, si possono fare ragionevoli ipotesi sulla matrice di covarianza di $Y$. Ad esempio, adottando particolari precauzioni da studiare caso per caso, si potrà supporre che le misure siano indipendenti e così le componenti di $Y^+ = [Y_1, Y_2, Y_3]$ saranno tra loro incorrelate: inoltre supponendo che le osservazioni abbiano tutte la stessa precisione si potrà supporre direttamente

$$C_{YY} = \sigma_o^2 I \ , \tag{2.1.6}$$

ove $\sigma_o^2$ ha il senso della varianza comune a tutte le osservazioni, e rimane una incognita del nostro problema.

Così il problema relativo a questo esempio può essere sintetizzato come:

conoscendo un vettore

$$Y_o = \left| \begin{array}{c} Y_{o1} \\ Y_{o2} \\ Y_{o3} \end{array} \right|$$

di valori osservati, estratti da una v.c. a 3 dimensioni

$$Y = \left| \begin{array}{c} Y_1 \\ Y_2 \\ Y_3 \end{array} \right| \ ,$$

di cui è definito il vettore dei valori medi, cioè delle osservabili,

$$E\{Y\} = y = \left| \begin{array}{c} y_1 \\ y_2 \\ y_3 \end{array} \right|$$

si sappia che tale media deve soddisfare la condizione (lineare)

$$\sum y_i = 0 \ , \tag{2.1.7}$$

e sapendo che la covarianza di $Y$ ha la forma

$$C_{YY} = \sigma_o^2 I \ , \tag{2.1.8}$$

si vogliono trovare stime di $y$ e di $\sigma_o^2$, nel senso che che si cercano stimatori

$$\hat{y} \ , \hat{\sigma}_o^2$$

rispettivamente di $y$ e $\sigma_o^2$, che siano funzioni della variabile campionaria $Y$, dunque calcolabili conoscendo il vettore delle osservazioni $Y_o$.

Inoltre si vorrebbe conoscere anche la dispersione delle nostre stime ovvero $C_{\hat{y}\hat{y}}$ e, dove possibile, anche $\sigma^2(\hat{\sigma}_o^2)$. Non manchiamo di osservare, chiudendo questo esempio, che per motivi di simmetria (cioè non avendo motivi di pensare che un errore sia più grande degli altri) una soluzione intuitiva del nostro probema di stima può essere data ponendo (per la (2.1.5))

$$\hat{\varepsilon}_1 = \hat{\varepsilon}_2 = \hat{\varepsilon}_3 = \frac{1}{3}\sum_{k=1}^{3}\varepsilon_k = \frac{1}{3}\sum_{k=1}^{3} Y_{ok} \quad : \tag{2.1.9}$$

inoltre si ha anche che, posto

$$\Delta = \sum Y_{ok} = \sum \varepsilon_k \quad , \quad (E\{\Delta\} = 0) \tag{2.1.10}$$

risulta per la legge di propagazione degli errori

$$E\{\Delta^2\} = \sigma^2(\Delta) = 3\sigma_o^2 \ ,$$

così che

$$\hat{\sigma}_o^2 = \frac{1}{3}\Delta^2 \tag{2.1.11}$$

è uno stimatore corretto di $\sigma_o^2$.

In base alla (2.1.9), la (2.1.11) può essere scritta come

$$\hat{\sigma}_o^2 = \frac{1}{3}\Delta^2 = 3\hat{\varepsilon}_i^2 \tag{2.1.12}$$

**Osservazione 2.1.1:** si può notare che, pur nella sua elementarità, l'Esempio 2.1.1 è stato svolto prescindendo completamente dalla distribuzione della variabile 3-D, $Y$. Una maniera per geometrizzare la scelta (2.1.9) è la seguente: si consideri lo spazio $R^3$ in cui ha sede la distribuzione di $Y$, e in esso il piano $V$ di equazione (2.1.7).

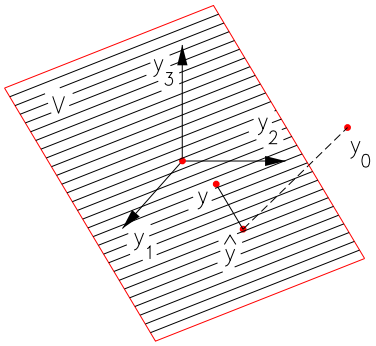



Figura 2.1.2

Geometria elementare dei minimi quadrati:

$V =$ varietà dei valori ammissibili

$Y_o =$ vettore delle osservazioni

$y =$ vettore delle osservabili

$\hat{y} =$ vettore delle stime

La distribuzione di $Y$ ha un valore medio $y \in V$ (cfr. fig. 2.1.2), detta varietà dei valori ammissibili, il vettore delle osservazioni $Y_o$ in generale sarà al di fuori di $V$; si cerca un vettore $\hat{y} \in V$ che sia "il più vicino possibile" a $y$, e sia una funzione lineare di $Y_o$.

Se $Y_o$ ha una distribuzione isotropa, cioè senza una tendenza a muoversi dalla media $y$ più in una direzione che in un'altra, cosa che è indicata dalla forma di covarianza (2.1.8), è "intuitivo" scegliere per $\hat{y}$ la proiezione ortogonale di $Y_o$ su $V$, cioè quel vettore

$$\hat{y} = Y_o - \hat{\varepsilon} \tag{2.1.13}$$

che tra tutti quelli di $V$ $(\sum \hat{y}_i = 0)$ rende minima la distanza

$$\hat{\varepsilon}^+ \hat{\varepsilon} = \sum \hat{\varepsilon}_i^2 = \sum (Y_{oi} - \hat{y}_i)^2 = (Y_o - \hat{y})^+ (Y_o - \hat{y}) = \min \ . \tag{2.1.14}$$

In questo caso il quadrato della distanza minimizzata (2.1.13) serve anche a stimare $\sigma_o^2$ in base alla (2.1.12).

**Osservazione 2.1.2:** il criterio di minimizzare la somma dei quadrati (2.1.14) coincide perfettamente con quello di maximum likelihood quando si supponga che

$$Y \sim N[y, \sigma_o^2 I] \ . \tag{2.1.15}$$

Infatti in tal caso

$$L(Y;y) = \frac{1}{(\sqrt{2\pi})^n \sigma_o^n} e^{(-1/2\sigma_o^2)\sum (Y_i - y_i)^2} \ , \tag{2.1.16}$$

così che la condizione di massimo di $L$ rispetto a $y$, soggetto alla condizione $\sum y_i = 0$, si traduce nella condizione di minimo (2.1.14). In questo senso si può pensare di generalizzare la (2.1.14) al caso in cui

$$C_{YY} = \sigma_o^2 Q \tag{2.1.17}$$

con $Q \neq I$, prendendo, in analogia al caso della distribuzione normale,

$$(Y_o - \hat{y})^+ Q^{-1} (Y_o - \hat{y}) = \min \ . \tag{2.1.18}$$

Questo stesso principio sarà giustificato anche in base ad una richiesta di invarianza del procedimento di stima per trasformazioni lineari.

## 2.2 Formulazione generale del problema (caso lineare)

Analizzando l'esempio del paragrafo precedente perveniamo alla seguente formulazione generale del problema di stima.

È dato un vettore di valori osservati

$$Y_o = \left| \begin{array}{c} Y_{o1} \\ \vdots \\ Y_{on} \end{array} \right| \ , \tag{2.2.1}$$

tratto da una v.c. $n$-dimensionale, $Y = \left| \begin{array}{c} Y_1 \\ \vdots \\ Y_n \end{array} \right|$, di cui non si conosce in generale la distribuzione, ma di cui si sa che il valore medio

$$y = E\{Y\} \ , \tag{2.2.2}$$

detto anche vettore delle variabili osservabili, per motivi fisici o geometrici è ristretto a stare su una varietà lineare a $m$ dimensioni ($m < n$)

$$y \in V \ , \tag{2.2.3}$$

detta anche varietà dei valori ammissibili; questo costituisce il così detto modello deterministico.

Inoltre sulla base della conoscenza delle modalità di misura si può affermare che la covarianza di $Y$ ha la forma

$$C_{YY} = \sigma_o^2 Q \ , \tag{2.2.4}$$

con $\sigma_o^2$ incognito e $Q$ una matrice nota, strettamente definita positiva: questo costituisce il così detto modello stocastico[4].

Sulla base del vettore $Y_o$, del modello deterministico (2.2.3), del modello stocastico (2.2.4) si vogliono trovare le stime

$$\begin{array}{l} \hat{y} \sim y \\ \hat{\sigma}_o^2 \sim \sigma_o^2 \end{array} \quad \text{con matrice di covarianza} \quad C_{\hat{y}\hat{y}}$$

e, se possibile,

$$\sigma^2\{\hat{\sigma}_o^2\} \ .$$

Naturalmente per arrivare a definire le stime richieste occorre porre delle condizioni aggiuntive: queste costituiscono il *principio dei minimi quadrati*, che consiste nell'imporre che $\hat{y}$ sia tale da soddisfare

$$\begin{cases} (Y_o - \hat{y})^+ Q^{-1} (Y_o - \hat{y}) = \min \\ \hat{y} \in V \end{cases} \tag{2.2.5}$$

e che $\hat{\sigma}_o^2$ sia proporzionale alla forma quadratica minimizzata nella (2.2.5),

[4]Il modello stocastico (2.2.4) implica che siano note le varianze $\sigma^2\{Y_i\}$ a meno di un fattore proporzionale e le correlazioni tra le componenti; questa ipotesi copre molti casi metrologicamente interessanti.

$$\hat{\sigma}_o^2 = c(Y_o - \hat{y})^+ Q^{-1}(Y_o - \hat{y}) \ , \tag{2.2.6}$$

con la costante $c$ determinata in modo tale che $\hat{\sigma}_o^2$ sia uno stimatore corretto di $\sigma_o^2$.

**Osservazione 2.2.1:** nel caso $Q = I$ il principio (2.2.5) è interpretato come la ricerca del punto di minima distanza (euclidea),

$$d(Y, \hat{y}) = \sqrt{(Y - \hat{y})^+ (Y - \hat{y})} \ ,$$

tra quelli che appartengono alla varietà $V$.

Come si è visto anche nel paragrafo 2.1 la ricerca del punto più vicino equivale a riconoscere che non esiste nessuna direzione preferenziale proprio perché $Q = I$.

Vogliamo far vedere che, se $Q = I$, il principio (2.2.5) può essere derivato sulla base della seguente richiesta: sia $\hat{y}$ lo stimatore lineare a cui arriveremo in base al criterio scelto; sia ora $T$ una trasformazione lineare qualunque che trasforma $Y$ in una variabile isotropa, ad esempio

$$T = Q^{-1/2} \quad [5] \tag{2.2.7}$$

e sia

$$U = TY \ , \tag{2.2.8}$$

così che

$$C_{UU} = \sigma_o^2 I \ . \tag{2.2.9}$$

Al vettore osservato $Y_o$ corrisponderà $U_o = TY_o$, mentre alla varietà $V$ corrisponderà la varietà degli $u$ possibili del tipo $u = Ty, y \in V$. Sulla base di questo problema per $u$, sappiamo che una stima $\hat{u}$ può essere ricavata in base al principio

[5] Ricordiamo che questa è per definizione l'unica matrice simmetrica, definita positiva per cui $T^2 = Q^{-1}$.

$$\begin{cases} (U-\hat{u})^+(U-\hat{u}) = \min \\ \hat{u} = Ty \quad (y \in V) \end{cases} . \tag{2.2.10}$$

Ora richiediamo che la stima $\hat{y}$ che stiamo cercando sia legata a $\hat{u}$ proprio dalla stessa relazione, cioè che se $\hat{u}$ è la stima di $u$, allora
$\hat{y} = T^{-1}\hat{u}$ sia la stima ricercata di $y$. Ma in tal caso ispezionando la (2.2.10) si vede che $\hat{y}$ può essere ricavata direttamente dal principio

$$\begin{cases} (Y-\hat{y})^+Q^{-1}(Y-\hat{y}) = \min \\ \hat{y} \in V \end{cases} ,$$

che è appunto quanto volevamo dimostrare.

**Osservazione 2.2.2:** poiché in analisi matematica il concetto di distanza è introdotto in termini generali tramite le sue proprietà formali e in particolare negli spazi euclidei $R^n$, ogni forma quadratica omogenea, definita positiva in senso stretto

$$d(x,y) = (x-y)^+P(x-y)$$

può essere presa come nozione di distanza (subordinata alla metrica $P = \{P_{ik}\}$), si vede che il problema dei minimi quadrati (caso lineare) può essere definito in termini formalmente molto semplici dicendo che:

> si vuole trovare il vettore $\hat{y} \in V$, varietà dei valori ammissibili, più vicino ad ogni $Y$ dato, secondo una metrica definita, tramite il modello stocastico (2.2.4), come $P = Q^{-1}$.

**Osservazione 2.2.3:** dal punto di vista della simbologia notiamo che nelle formule dei paragrafi precedenti come in quelli che verranno si userà $Y$ quando si vorrà mettere in evidenza la variabile casuale che descrive il complesso delle osservazioni; si userà $Y_o$ quando si vorrà indicare lo specifico vettore di osservazioni (numeri) trovati nell'eseguire le misure. Con $\hat{y}$ si indicherà lo stimatore di $y$, il quale sarà dunque a sua volta una variabile casuale funzione della variabile campionaria $Y, \hat{y} = \hat{y}(Y)$; tuttavia indicheremo con lo stesso simbolo anche il vettore delle stime,

cioè il vettore dei numeri che si ottengono sostituendo nello stimatore $\hat{y}(Y)$ il vettore delle osservazioni, $\hat{y} = \hat{y}(Y_o)$. La funzione è ovviamente sempre la stessa, ma in un caso sottolineiamo il carattere di variabile casuale, nell'altro il carattere di vettore numerico.

## 2.3 Soluzione del problema di minimi quadrati: stimatori di osservabili e parametri

Vogliamo risolvere il problema posto col principio dei minimi quadrati (2.2.5), (2.2.6) considerando sia il caso in cui la varietà lineare $V$ sia data in forma parametrica

$$\begin{aligned} &y = Ax + a \\ &(\dim y = n,\ \dim x = m, A = [n, m], \text{rango } A = m)\ , \end{aligned} \tag{2.3.1}$$

che quello in cui essa è data in forma di equazioni di condizione

$$By = b \tag{2.3.2}$$

($B = [c, m]$ , $c = n - m = \text{N}^o$ eq. di condizione, rango $B = c$, $BA = = 0$, $Ba = b$).

Si noti che nel caso normale con $Q = I$ la soluzione è già stata trovata nell'Esempio 1.7.2.

Il metodo è perfettamente generalizzabile al caso $Q \neq I$, tuttavia si ritiene importante ricavare qui la soluzione direttamente dal principio variazionale

$$(Y_o - \hat{y})^+ Q^{-1}(Y_o - \hat{y}) = \min \tag{2.3.3}$$

sia per l'interesse del metodo in sé che per la maggiore generalità dei risultati. Per rendere i conti il più possibile sintetici conviene radunare le due forme (2.3.1), (2.3.2) in un'unica forma generale della varietà $V$, cioè

$$Dy = Ax + d\ , \tag{2.3.4}$$

ed una volta ottenuto il risultato specializzare il modello (2.3.4) ai due casi di maggior interesse, $\{D = I, d = a\}$ ovvero $\{D = B, A = 0, d = b\}$.

Prima di procedere specifichiamo le dimensioni di (2.3.4):

$$\begin{array}{lll} \dim y = n & , \quad \dim x = m & , \quad m < n \\ D = [l, n] & , \quad A = [l, m] & , \quad m < l \leq n \end{array} \tag{2.3.5}$$

inoltre supponiamo che il rango di $D$ sia pieno, cioè che tutte le equazioni di condizione siano indipendenti, e che la matrice $A$ sia o nulla (caso in cui mancano i parametri e si hanno pure equazioni di condizione (2.3.2)) oppure abbia rango pieno in modo che si possa sempre pensare di eliminare i parametri, riducendo le (2.3.4) ad un sistema di $c = l - n$ pure equazioni di condizione.

**Teorema 2.3.1:** nelle condizioni poste i vettori (stimatori) $\hat{x}, \hat{y}$ che risolvono il problema variazionale (2.3.3), (2.3.4)

$$\begin{aligned} &(Y_o - \hat{y})^+ Q^{-1} (Y_o - \hat{y}) = \min \\ &D\hat{y} = A\hat{x} + d \end{aligned}$$

sono dati da

$$\hat{x} = N^{-1} A^+ K^{-1} (DY_o - d) \tag{2.3.6}$$

$$\hat{y} = Y_o - QD^+ K^{-1} U_o \ , \tag{2.3.7}$$

dove

$$\begin{cases} K = DQD^+ & \\ N = A^+ K^{-1} A & \text{(matrice normale)} \\ U_o = DY_o - A\hat{x} - d & \text{(vettore degli scarti)} \end{cases} \tag{2.3.8}$$

Il problema è risolto cercando i punti di stazionarietà di $(Y_o - \hat{y})Q^{-1}(Y_o - \hat{y})$ sotto la condizione (2.3.4), mediante la tecnica dei moltiplicatori di Lagrange. Notiamo subito che la forma quadratica (2.3.3) è limitata inferiormente, perché definita positiva, mentre tende

all'infinito quando $Y_o - \hat{y} \to \infty$: perciò tale forma deve ammettere almeno un minimo assoluto e se le condizioni di stazionarietà hanno soluzione unica, allora tale soluzione deve necessariamente corrispondere al minimo.

Per applicare la tecnica dei moltiplicatori di Lagrange, occorre sommare a $(1/2)(Y_o - \hat{y})^+ Q^{-1}(Y_o - \hat{y})$ le componenti di $D\hat{y} - A\hat{x} - d$, ognuna moltiplicata per un suo moltiplicatore $\lambda_1, \lambda_2, \ldots, \lambda_\ell$: ovvero otteniamo la funzione obiettivo

$$\varphi(\hat{x}, \hat{y}) = (1/2)(Y_o - \hat{y})^+ Q^{-1}(Y_o - \hat{y}) + (D\hat{y} - A\hat{x} - d)^+ \lambda \qquad (2.3.9)$$

dove $\lambda^+ = [\lambda_1, \ldots, \lambda_\ell]$.

Per imporre le condizioni di stazionarietà a $\varphi$, differenziamo tale funzione[6]

$$d\varphi = -d\hat{y}^+ Q^{-1}(Y_o - \hat{y}) + d\hat{y}^+ D^+ \lambda - d\hat{x}^+ A^+ \lambda \ .$$

Annullando separatamente i coefficienti di $d\hat{y}^+$ e $d\hat{x}^+$ troviamo

$$\begin{aligned} &-Q^{-1}(Y_o - \hat{y}) + D^+ \lambda = 0 \\ &A^+ \lambda = 0 \ , \end{aligned} \qquad (2.3.10)$$

equazione cui va aggiunta la (2.3.4).

Dalla prima delle (2.3.10) si ha

$$\hat{y} = Y_o - QD^+ \lambda \qquad (2.3.11)$$

che, posta nella (2.3.4), dà

[6]Si osservi che se $S$ è simmetrica e $v$ un vettore qualunque,

$$d(v^+ Sv) = (dv^+)Sv + v^+ Sdv = (dv^+)Sv + dv^+ S^+ v = 2(dv^+)Sv \ .$$

$$DY_o - K\lambda = A\hat{x} + d \qquad (K = DQD^+) \ . \tag{2.3.12}$$

Notiamo che $D, Q$ e quindi anche $K$ sono di rango pieno. Si ricava $\lambda$

$$\lambda = K^{-1}(DY_o - d) - K^{-1}A\hat{x} \tag{2.3.13}$$

e si usa le seconda delle (2.3.10), trovando

$$N\hat{x} = A^+K^{-1}(DY_o - d) \qquad (N = A^+K^{-1}A) \ . \tag{2.3.14}$$

Ricordando che o $A$ è identicamente nulla, nel qual caso la stima di $\hat{x}$ non si pone e si può usare direttamente (2.3.13) in (2.3.11), oppure $A$ è di rango pieno, la (2.3.14) può essere risolta e dà la (2.3.6):

$$\hat{x} = N^{-1}A^+K^{-1}(DY_o - d) \ .$$

Sostituendo nella (2.3.13) si ha

$$\lambda = K^{-1}(DY_o - d - A\hat{x}) = K^{-1}U_o \ ;$$

sostituendo nella (2.3.11) si trova la (2.3.7)

$$\hat{y} = Y_o - QD^+K^{-1}U_o \ .$$

Il vettore $U_o = DY_o - d - A\hat{x}$ prende il nome di *vettore degli scarti delle equazioni*.

**Osservazione 2.3.1:** si consideri il caso in cui si abbiano pure equazioni parametriche $y = Ax + a$, dette anche *equazioni di osservazione*; rispetto al modello generale si ha

$$D = I \qquad d = a \ .$$

In questo caso la matrice $A$ prende il nome di matrice disegno. Inoltre si ha

$$K = Q \ , \ N = A^{+}Q^{-1}A$$

e risulta

$$\begin{cases} \hat{x} = N^{-1}A^{+}Q^{-1}(Y_o - a) \\ \hat{y} = Y_o - QQ^{-1}U_o = Y_o - U_o = A\hat{x} + a \end{cases} . \tag{2.3.15}$$

Si può notare che per la seconda delle (2.3.15) il vettore $U_o$ coincide in questo caso anche con il vettore delle correzioni di $Y_o, Y_o - \hat{y}$.

**Osservazione 2.3.2:** si prenda invece il caso in cui $By = b$, detto anche delle *equazioni di condizione*; rispetto al modello generale si ha

$$D = B \qquad A = 0 \qquad d = b \ :$$

dalle (2.3.14) e (2.3.14), posto $A = 0$, si ricava direttamente

$$\hat{y} = Y_o - QB^{+}K^{-1}(BY_o - b) \quad (K = BQB^{+}) \ . \tag{2.3.16}$$

In questo caso il vettore

$$U_o = BY_o - b$$

prende il nome di *vettore degli errori di chiusura* (si veda l'Esempio 2.1.1).

**Osservazione 2.3.3:** si noti che dal punto di vista della stima dei parametri $x$ serve, più che il vettore $Y_o$, il vettore $DY_o - d = V_o$: questo significa che $V_o$ contiene tutta l'informazione necessaria alla stima di $x$: in effetti il modello generale

$$Dy = Ax + d$$

può essere ritrasformato in modello puramente parametrico

$$v = Ax \ ,$$

con "osservazione" $V_o$ e con modello stocastico

$$C_{VV} = \sigma_o^2 DQD^+ = \sigma_o^2 K \ .$$

## 2.4 Covarianza degli stimatori e stima di $\sigma_o^2$

Gli stimatori $\hat{x}, \hat{y}$ sono *corretti*: infatti dalla (2.3.6), osservando che per i valori medi $x, y$ deve valere la (2.3.4), troviamo

$$E\{\hat{x}\} = N^{-1}A^+K^{-1}(DE\{Y\} - d) = N^{-1}A^+K^{-1}(Ax) = N^{-1}Nx = x \ . \tag{2.4.1}$$

Pertanto per il vettore $U$ degli scarti delle equazioni si ha

$$E\{U\} = E\{DY_o - A\hat{x} - d\} = Dy - Ax - d = 0 \tag{2.4.2}$$

e quindi

$$E\{\hat{y}\} = E\{y\} - QD^+K^{-1}E\{U\} = y \ . \tag{2.4.3}$$

Tra l'altro ciò implica che il vettore delle correzioni

$$\hat{\varepsilon} = Y_o - \hat{y}$$

è a media nulla

$$E\{\hat{\varepsilon}\} = 0 \ . \tag{2.4.4}$$

Cerchiamo ora le covarianze di $\hat{x}, \hat{y}$ e alcune matrici di covarianza e di cross-covarianza che ci serviranno in futuro.

*Covarianza di $\hat{x}$:* ricordando che per ipotesi

$$C_{YY} = \sigma_o^2 Q$$

ed applicando la propagazione della covarianza, si ha

$$C_{\hat{x}\hat{x}} = \sigma_o^2 N^{-1} A^+ K^{-1} D Q D^+ K^{-1} A N^{-1} = \sigma_o^2 N^{-1} \ . \tag{2.4.5}$$

*Covarianza di U:* dalla definizione (2.3.8) e dalla (2.3.6) si vede che

$$\begin{aligned} U &= DY - d - AN^{-1}A^+K^{-1}(DY-d) = \\ &= (I - AN^{-1}A^+K^{-1})(DY-d) \ . \end{aligned} \tag{2.4.6}$$

Poiché $V = DY - d$ ha covarianza $C_{VV} = \sigma_o^2 DQD^+ = \sigma_o^2 K$, si trova

$$\begin{aligned} C_{UU} &= \sigma_o^2 (I - AN^{-1}A^+K^{-1})K(I - K^{-1}AN^{-1}A^+) = \\ &= \sigma_o^2 (K - AN^{-1}A^+) \ . \end{aligned} \tag{2.4.7}$$

*Cross-covarianza di $U$ e $\hat{x}$:* notando che tanto $U$ quanto $\hat{x}$ sono funzioni lineari dello stesso vettore $V = DY - d$ (cfr. (2.3.6) e (2.4.6)), si ha per definizione

$$\begin{aligned} C_{U\hat{x}} &= E\{U(\hat{x}-x)^+\} = (I - AN^{-1}A^+K^{-1})C_{VV}K^{-1}AN^{-1} = \\ &= \sigma_o^2 (I - AN^{-1}A^+K^{-1})AN^{-1} = \\ &= \sigma_o^2 (A - AN^{-1}A^+K^{-1}A)N^{-1} = 0 \ . \end{aligned} \tag{2.4.8}$$

Dunque $U$ e $\hat{x}$ sono tra loro non correlati. Si osservi che di conseguenza anche il vettore delle correzioni $\hat{\varepsilon} = Y - \hat{y}$, che in base alla (2.3.7) è funzione lineare di $U$, risulterà non correlato ad $\hat{x}$

$$C_{\hat{\varepsilon}\hat{x}} = 0 \ . \tag{2.4.9}$$

*Covarianza di $\hat{y}$:* allo scopo di calcolare questa matrice, conviene prima determinare a livello di servizio $C_{\hat{\varepsilon}\hat{\varepsilon}}$ e $C_{\varepsilon y}$. Dalla definizione $\hat{\varepsilon} = Y - \hat{y} = = QD^+K^{-1}U$ e dalla (2.4.7) si trova

$$C_{\hat{\varepsilon}\hat{\varepsilon}} = \sigma_o^2 Q D^+ K^{-1}(K - AN^{-1}A^+)K^{-1}DQ \ . \qquad (2.4.10)$$

Inoltre per la (2.4.6)

$$\begin{aligned} C_{\hat{\varepsilon}Y} &= QD^+K^{-1}(I - AN^{-1}A^+K^{-1})DC_{YY} = \\ &= \sigma_o^2 QD^+K^{-1}(K - AN^{-1}A^+)K^{-1}DQ \ , \end{aligned} \qquad (2.4.11)$$

cioè $C_{\hat{\varepsilon}Y} = C_{\hat{\varepsilon}\hat{\varepsilon}}$.

Essendo tale matrice simmetrica è anche

$$C_{\hat{\varepsilon}Y} = C_{\hat{\varepsilon}\hat{\varepsilon}} = C_{Y\hat{\varepsilon}} \ .$$

Per concludere, dalla relazione

$$\hat{y} = Y - \hat{\varepsilon} \ ,$$

si ha

$$\begin{aligned} C_{\hat{y}\hat{y}} &= C_{YY} - C_{Y\hat{\varepsilon}} - C_{\hat{\varepsilon}Y} + C_{\hat{\varepsilon}\hat{\varepsilon}} = C_{YY} - C_{\hat{\varepsilon}\hat{\varepsilon}} = \\ &= \sigma_o^2\{Q - QD^+K^{-1}[K - AN^{-1}A^+]K^{-1}DQ\} \ . \end{aligned} \qquad (2.4.12)$$

**Osservazione 2.4.1:** nel caso di un modello di equazioni parametriche si ha

$$D = I, \quad K = Q, \quad U = \varepsilon, \quad N = A^+Q^{-1}A \ ;$$

così

$$\left\{ \begin{array}{l} C_{\hat{x}\hat{x}} = \sigma_o^2 N^{-1} \\ C_{UU} = C_{\hat{\varepsilon}\hat{\varepsilon}} = \sigma_o^2[Q - AN^{-1}A^+] \\ C_{\hat{y}\hat{y}} = \sigma_o^2 AN^{-1}A^+ \ . \end{array} \right. \qquad (2.4.13)$$

Si noti come alla decomposizione

$$Y - y = Y - \hat{y} + \hat{y} - y = \hat{\varepsilon} + A(\hat{x} - x)$$

corrisponde in questo caso una perfettamente analoga decomposizione

$$C_{YY} = \sigma_o^2 Q = C_{\hat{\varepsilon}\hat{\varepsilon}} + C_{\hat{y}\hat{y}} = C_{\hat{\varepsilon}\hat{\varepsilon}} + AC_{\hat{x}\hat{x}}A^+$$

a causa della non correlazione tra i due vettori ($Q^{-1}$-ortogonali) $\hat{\varepsilon}$ e $\hat{y}-y = A(\hat{x} - x)$.

**Osservazione 2.4.2:** nel caso di un modello di pure equazioni di condizione risulta

$$\begin{aligned} &D = B\ , \quad A = 0\ , \quad K = BQB^+\ , \\ &U = BY - b \qquad \text{vettore degli errori di chiusura;} \end{aligned}$$

in tal caso si ha

$$\begin{cases} C_{UU} = \sigma_o^2 K \\ C_{\hat{\varepsilon}\hat{\varepsilon}} = \sigma_o^2 QB^+K^{-1}BQ \\ C_{\hat{y}\hat{y}} = \sigma_o^2\{Q - QB^+K^{-1}BQ\}\ . \end{cases} \tag{2.4.14}$$

Anche in questo caso alla decomposizione

$$Y - y = \hat{\varepsilon} + \hat{y} - y$$

corrisponde l'analoga relazione per le covarianze

$$C_{YY} = C_{\hat{\varepsilon}\hat{\varepsilon}} + C_{\hat{y}\hat{y}}\ .$$

*Stima di $\sigma_o^2$:* come già annunciato nella formulazione generale del problema (cfr. §2.2), cercheremo uno stimatore di $\sigma_o^2$ nella forma

$$\hat{\sigma}_o^2 = c \cdot (Y - \hat{y})^+ Q^{-1}(Y - \hat{y}) = c \cdot \hat{\varepsilon}^+ Q^{-1}\hat{\varepsilon}\ , \tag{2.4.15}$$

imponendo la condizione di correttezza

$$E\{\hat{\sigma}_o^2\} = \sigma_o^2 \ , \tag{2.4.16}$$

allo scopo di determinare la costante $c$.

Notando che[7]

$$\hat{\varepsilon}^+ Q^{-1} \hat{\varepsilon} = TrQ^{-1} \hat{\varepsilon}\hat{\varepsilon}^+$$

si vede che, (cfr. (2.4.10)):

$$\begin{aligned} E\{\hat{\varepsilon}^+ Q^{-1} \hat{\varepsilon}\} &= TrQ^{-1} E\{\hat{\varepsilon}\hat{\varepsilon}^+\} = TrQ^{-1} C_{\hat{\varepsilon}\hat{\varepsilon}} = \\ &= \sigma_o^2 Tr\ Q^{-1} Q D^+ K^{-1} [K - AN^{-1}A^+] K^{-1} DQ\} = \\ &= \sigma_o^2 \{TrD^+ K^{-1} DQ - TrD^+ K^{-1} AN^{-1} A^+ K^{-1} DQ\} \ . \end{aligned} \tag{2.4.17}$$

Per eseguire tale conto è sufficiente ricordare che

$$\begin{aligned} &K = DQD^+ = [l \cdot l] \quad (l = \text{n}^o\text{equazioni (cfr.2.3.4)}) \\ &N = A^+ K^{-1} A = [m \cdot m] \quad (m = \text{n}^\circ\text{parametri}) \ . \end{aligned}$$

In effetti

$$\begin{aligned} &TrD^+ K^{-1} DQ = TrK^{-1} DQD^+ = TrK^{-1} K = TrI(l) = l \ ; \\ &TrD^+ K^{-1} AN^{-1} A^+ K^{-1} DQ = TrA^+ K^{-1} DQD^+ K^{-1} AN^{-1} = \\ &= TrA^+ K^{-1} AN^{-1} = TrNN^{-1} = TrI(m) = m \ . \end{aligned}$$

Ciò posto dalla (2.4.17) si deriva semplicemente

[7]Ricordiamo che, data una matrice $Q = [n, n]$, la traccia è definita come $TrQ = \sum q_{ii}$: per tale funzionale valgono le seguenti due proprietà

$$\begin{aligned} &TrQ = Tr[q_{ik}] = Tr[q_{ki}] = TrQ^+ \\ &Tr(A \cdot B) = \sum_i (\sum_k a_{ik} b_{ki}) = \sum_k (\sum_i b_{ki} a_{ik}) = Tr(B \cdot A) \ . \end{aligned}$$

$$E\{\hat{\varepsilon}^+ Q^{-1}\hat{\varepsilon}\} = \sigma_o^2\{l-m\} \ . \tag{2.4.18}$$

Confrontando con (2.4.15), (2.4.16), si vede che uno stimatore corretto di $\sigma_o^2$ è dato da

$$\hat{\sigma}_o^2 = \frac{\hat{\varepsilon}^+ Q^{-1}\hat{\varepsilon}}{l-m} \ . \tag{2.4.19}$$

Questo stimatore può anche essere espresso esplicitamente in funzione degli scarti delle equazioni $U_o$, anziché delle correzioni $\hat{\varepsilon}$: infatti, ricordando che

$$\hat{\varepsilon} = QD^+K^{-1}U \quad , \quad (U_o = DY_o - A\hat{x} - d)$$

si ha

$$\sigma_o^2 = \frac{U^+K^{-1}U}{l-m} \tag{2.4.20}$$

Si può osservare che quando si sia interessati solo alla stima dei parametri $x$, la (2.4.20) può essere applicata senza cercare la stima $\hat{y}$, poiché gli scarti delle equazioni $U$ dipendono da $DY_o - d = V_o$ ($U = V_o - A\hat{x}$) e non da $Y_o$ direttamente.

**Osservazione 2.4.3:** nel caso di un modello puramente parametrico

$$U = \hat{\varepsilon} \quad , \quad K = Q$$

così che (2.4.19) e (2.4.20) coincidono in modo evidente; in più si può notare che in questo caso

$$l = \text{n°equazioni} = n = \text{n°osservabili.}$$

**Osservazione 2.4.4:** per il modello con pure equazioni di condizione, la forma (2.4.20) è certamente più comoda della (2.4.19) in quanto il vettore $U_o$ (vettore degli errori di chiusura) ha dimensioni inferiori ad $\hat{\varepsilon}$.

## 2.5 Ottimalità degli stimatori m.q.: *Teorema di Markov*

Poiché non abbiamo sin qui fatto alcuna ipotesi sulla distribuzione di $Y$, non è possibile porci in generale il problema dell'efficienza degli stimatori m.q. Tuttavia è possibile dimostrare una proprietà assai forte di tali stimatori, almeno se li si paragona ad altri stimatori lineari e corretti. In effetti supponiamo di definire la varietà dei valori ammissibili tramite le equazioni parametriche

$$y = Ax + a \ : \tag{2.5.1}$$

sappiamo che lo stimatore è poi lo stesso qualunque sia la forma con cui si descrive $V$. In tal caso gli stimatori m.q. sono dati da

$$\hat{x} = N^{-1}A^{+}Q^{-1}(Y_o - a) = \hat{M}(Y_o - a) \tag{2.5.2}$$
$$\hat{y} = AN^{-1}A^{+}Q^{-1}(Y_o - a) + a = \hat{L}(Y_o - a) + a \ . \tag{2.5.3}$$

Più in generale definiamo la classe degli stimatori lineari corretti tramite le formule

$$\tilde{x} = \tilde{M}(Y_o - a) + \tilde{m} \tag{2.5.4}$$
$$\tilde{y} = \tilde{L}(Y_o - a) + \tilde{l} \tag{2.5.5}$$

e imponiamo la condizione che quando $Y_o \in V$, cioè $Y_o - a = A\xi$ per un qualche $\xi$, allora risulti anche

$$\begin{cases} \tilde{x} = \xi \equiv \tilde{M}A\xi + \tilde{m} \\ \tilde{y} = Y_o = A\xi + a \equiv \tilde{L}A\xi + \tilde{l} \end{cases} \ . \tag{2.5.6}$$

Le (2.5.6) vanno intese come identità rispetto a $\xi$, il che comporta le condizioni

$$I = \tilde{M}A \ , \ \tilde{m} = 0 \ ; \ A = L\tilde{A} \ , \ \tilde{l} = a \ . \tag{2.5.7}$$

Dunque la classe degli stimatori lineari e corretti è definita dalle (2.5.4), (2.5.5), con le condizioni aggiuntive (2.5.6), (2.5.7): è facile vedere che gli stimatori m.q. (2.5.2), (2.5.3) appartengono appunto a tale classe.

L'ottimalità di $\hat{M}, \hat{L}$ è valida all'interno di tale classe: vale infatti il seguente teorema:

> **Teorema 2.5.1**: (di Markov). Siano $\tilde{x}, \tilde{y}$ stimatori lineari corretti qualunque di $x, y$ e siano invece $\hat{x}, \hat{y}$ gli stimatori m.q.; allora si ha
>
> $$\begin{cases} C_{\tilde{x}\tilde{x}} \geq C_{\hat{x}\hat{x}} \\ C_{\tilde{y}\tilde{y}} \geq C_{\hat{y}\hat{y}} \ . \end{cases} \tag{2.5.8}$$

**Osservazione 2.5.1:** le (2.5.8) significano essenzialmente che $C_{\tilde{x}\tilde{x}} - C_{\hat{x}\hat{x}}$ e $C_{\tilde{y}\tilde{y}} - C_{\hat{y}\hat{y}}$ sono matrici definite positive, ovvero che

$$\lambda^+ C_{\tilde{x}\tilde{x}} \lambda \geq \lambda^+ C_{\hat{x}\hat{x}} \lambda \quad \forall \lambda \tag{2.5.9}$$

$$\mu^+ C_{\tilde{y}\tilde{y}} \mu \geq \mu^+ C_{\hat{y}\hat{y}} \mu \quad \forall \mu \ . \tag{2.5.10}$$

A loro volta le (2.5.9), (2.5.10) possono essere così interpretate: si voglia stimare una funzione lineare di $x$ e una di $y$

$$\begin{aligned} u &= \lambda^+ x + u_o \\ v &= \mu^+ u + v_o \ , \end{aligned} \tag{2.5.11}$$

se nelle (2.5.11) si usano gli stimatori m.q. per ottenere $\hat{u} = \lambda^+ \hat{x} + u_o$, $\hat{v} = \mu^+ \hat{y} + v_o$ si hanno stimatori lineari corretti di varianza minima

$$\sigma^2(\hat{u}) = \lambda^+ C_{\hat{x}\hat{x}} \lambda \quad , \quad \sigma^2(\hat{v}) = \mu^+ C_{\hat{y}\hat{y}} \mu \ ,$$

tra tutti gli altri stimatori lineari corretti

$$\begin{aligned} &\tilde{u} = \lambda^+ \tilde{x} + u_o \quad , \quad \tilde{v} = \mu^+ \tilde{y} + v_o \ , \\ &\sigma^2(\tilde{u}) \geq \sigma^2(\hat{u}) \quad ; \quad \sigma^2(\tilde{v}) \geq \sigma^2(\hat{v}) \ . \end{aligned}$$

Per dimostrare il Teorema 2.5.1 notiamo che

$$\tilde{x} = \tilde{M}(Y - a)$$
$$\hat{x} = \hat{M}(Y - a)$$

con

$$\tilde{M}A = I \quad , \quad \hat{M}A = I \ . \tag{2.5.12}$$

Si può allora scrivere

$$\tilde{x} = \tilde{x} - \hat{x} + \hat{x} = \delta + \hat{x} \tag{2.5.13}$$
$$C_{\tilde{x}\tilde{x}} = C_{\delta\delta} + C_{\delta\hat{x}} + C_{\hat{x}\delta} + C_{\hat{x}\hat{x}} \ : \tag{2.5.14}$$

è chiaro che se

$$C_{\delta\hat{x}} = C^{+}_{\hat{x}\delta} = 0 \ ,$$

si ha

$$C_{\tilde{x}\tilde{x}} - C_{\hat{x}\hat{x}} = C_{\delta\delta} \geq 0 \ , \tag{2.5.15}$$

che coincide con la prima delle (2.5.8).

D'altronde

$$\begin{aligned} C_{\delta\hat{x}} &= (\tilde{M} - \hat{M})C_{YY}\hat{M}^{+} = \sigma_o^2(\tilde{M} - \hat{M})Q \cdot Q^{-1}AN^{-1} = \\ &= \sigma_o^2(\tilde{M} - \hat{M})AN^{-1} = \\ &= \sigma_o^2(\tilde{M}A - \hat{M}A)N^{-1} = 0 \ , \end{aligned}$$

in conseguenza delle (2.5.12): la (2.5.15) è così provata.

Quanto alla $2^a$ delle (2.5.8) essa discende immediatamente dalla osservazione che

$$\tilde{y} = A\tilde{x} + a \quad , \quad \hat{y} = A\hat{x} + a \ ,$$

così che

$$C_{\tilde{y}\tilde{y}} - C_{\hat{y}\hat{y}} = A[C_{\tilde{x}\tilde{x}} - C_{\hat{x}\hat{x}}]A^{+} \geq 0 \ :$$

l'ultimo passaggio è facilmente verificato calcolando la forma quadratica associata

$$\lambda^{+}(C_{\tilde{y}\tilde{y}} - C_{\hat{y}\hat{y}})\lambda \ .$$

## 2.6 Problemi di minimi quadrati con vincoli

Volendo concentrare l'attenzione sulla stima dei parametri, assumiamo in questo paragrafo il modello parametrico

$$y = Ax + a \ , \tag{2.6.1}$$

cui ci si può sempre ridurre.

Ci sono almeno due casi importanti in cui può interessare un modello deterministico come in (2.6.1) con l'aggiunta di $(k)$ vincoli sui parametri $x$, cioè di equazioni (lineari) contenenti solo $x$:

$$Hx = h \quad (H = [k, m], \quad k < m) \tag{2.6.2}$$

**a)** quando non si è sicuri del modello (2.6.1), cioè non si sa se si sono rappresentate tutte le variabili che descrivono un certo fenomeno, può essere utile provare modelli diversi in cui si parte da uno molto generale che include molti parametri, ad altri più semplici in cui una parte dei parametri è bloccata (ad esempio vincolata a zero);

**b)** quando la matrice "disegno" $A$ non sia di rango pieno: in questo caso infatti viene meno la corrispondenza biunivoca tra i punti della varietà $V$ e quelli dello spazio dei parametri. Per eliminare questa indeterminazione è necessario introdurre vincoli sulla $x$.

Illustriamo questi casi con due esempi.

**Esempio 2.6.1:** un certo prodotto ottiene delle vendite settimanali crescenti che, a parte fluttuazioni casuali, sembrano seguire un andamento lineare nel tempo

$$Y_i = x_1 + x_2 t_i + \varepsilon_i \ ; \tag{2.6.3}$$

a un certo tempo si introduce l'azione di una pubblicità, misurata ad esempio in termini di numero $p_i$ di spot televisivi per settimana: si ritiene che tale fattore possa influenzare le vendite con una legge del tipo

$$Y_i = x_1 + x_2 t_i + x_3 p_i + \varepsilon_i \ . \tag{2.6.4}$$

Per valutare l'effettiva efficacia della pubblicità, si possono compensare le "osservazioni" $Y_i$ (cioè trovare gli stimatori m.q. $\hat{y}_i$) prima col modello (2.6.4) e poi col modello (2.6.3) per confrontare tra loro i risultati. Il modello (2.6.3) risulta essere contenuto nel modello (2.6.4), con l'imposizione del vincolo

$$x_3 = 0 \ .$$

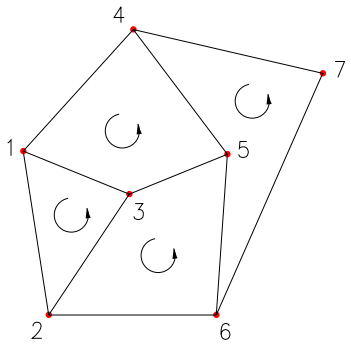



Figura 2.6.1

**Esempio 2.6.2:** in una rete di 7 punti si misurano dislivelli tra i punti stessi secondo uno schema come quello in figura: la congruenza geometrica della figura può essere espressa per mezzo delle quote dei punti stessi, considerate come parametri. Le osservabili, corrispondenti ai lati della rete, sono 10: le equazioni di osservazione si scrivono

$$Q_{ki} = Q_k - Q_i \tag{2.6.5}$$

per $i, k$ opportuni.

Le variabili $Q_{ki}$ fanno parte del vettore delle osservabili, le quote $Q_1, \ldots, Q_7$ del vettore dei parametri.

La matrice "disegno" $A$ implicata dalle (2.6.5) è deficiente di rango; infatti esiste un vettore

$$x = \begin{vmatrix} Q_1 \\ \vdots \\ Q_7 \end{vmatrix} = \begin{vmatrix} 1 \\ \vdots \\ 1 \end{vmatrix} \neq 0$$

tale per cui

$$Ax = 0 \ .$$

Questa deficienza è ovviamente legata alla mancata fissazione di una quota di riferimento; essa può essere rimossa imponendo un vincolo ai parametri.

Ad esempio si potrebbe porre

$$x_1 = Q_1 = 0 \ ,$$

cioè prendere convenzionalmente il punto 1 come riferimento, oppure

$$\sum_{i=1}^{7} x_i = \sum_{i=1}^{7} Q_i = 0 \ ,$$

ovvero fissare convenzionalmente a zero la media delle quote.

Il principio dei m.q. si scrive in questo caso come

$$\begin{cases} (1/2)(Y_o - \hat{y})^+ Q^{-1}(Y_o - \hat{y}) = \min \\ \hat{y} = A\hat{x} + a \\ 0 = H\hat{x} - h \end{cases} \tag{2.6.6}$$

Sostituendo la II nella I e usando dei moltiplicatori di Lagrange per la III, si ha la funzione obiettivo

$$\varphi = (1/2)(Y_o - A\hat{x} - a)^+ Q^{-1}(Y_o - A\hat{x} - a) + (H\hat{x} - h)^+ \lambda \ .$$

Imponendo la condizione di stazionarietà rispetto ad $\hat{x}$ si trova

$$-A^+ Q^{-1}(Y_o - A\hat{x} - a) + H^+ \lambda = 0 \ ,$$

cui vanno aggiunti i vincoli, ovvero

$$\begin{cases} N\hat{x} + H^+\lambda &= A^+ Q^{-1}(Y_o - a) \\ H\hat{x} &= h \end{cases} \tag{2.6.7}$$

La (2.6.7) costituisce un sistema normale esteso che può essere elaborato ulteriormente in modo analitico, solo se $N$ è di rango pieno.

In questo caso si può porre

$$\hat{x} = N^{-1}A^{+}Q^{-1}(Y_o - a) - N^{-1}H^{+}\lambda \tag{2.6.8}$$

che, con l'uso del vincolo, dà

$$HN^{-1}A^{+}Q^{-1}(Y_o - a) - HN^{-1}H^{+}\lambda = h \ . \tag{2.6.9}$$

Risolvendo la (2.6.9) rispetto a $\lambda$ e sostituendo in (2.6.8) si ottiene la soluzione cercata.

Questo procedimento però non è applicabile quando

$$\det N = 0$$

ed in generale non è conveniente.

Più comunemente (2.6.7) viene risolto con un metodo perturbativo che ha un'interessante interpretazione statistica.

Il vincolo

$$0 = Hx - h \ ,$$

può essere considerato come una osservazione nulla, $Z_o = 0$, di un vettore casuale $Z$ con media e covarianza nulla

$$\begin{cases} E\{Z\} = 0 \\ C_{ZZ} = 0 \end{cases} \quad : \tag{2.6.10}$$

con le condizioni (2.6.10) infatti $Z$ assume il valore zero con probabilità 1.

Ora, rilassando un poco tale condizione, possiamo prendere il vincolo come una osservazione

$$Z_o = 0 \ , \tag{2.6.11}$$

tratta da una variabile $Z$ con media nulla

$$E\{Z\} = 0 \tag{2.6.12}$$

e con varianza infinitesima

$$C_{ZZ} = \sigma_o^2 \varepsilon I \ . \tag{2.6.13}$$

Specifichiamo anche che supponiamo che $Z$ sia incorrelato con $Y$

$$C_{ZY} = C_{YZ}^{+} = 0 \ . \tag{2.6.14}$$

In questo modo abbiamo un ampliamento del modello che include i vincoli delle osservabili:

$$\begin{aligned}
& w = \left|\begin{smallmatrix} y \\ z \end{smallmatrix}\right| \text{ osservabili}, && W_o = \left|\begin{smallmatrix} Y_o \\ 0 \end{smallmatrix}\right| \text{ osservazioni} \\
& C_{WW} = \sigma_o^2 \left|\begin{smallmatrix} Q & 0 \\ 0 & \varepsilon I \end{smallmatrix}\right| \ , && \text{modello stocastico} \\
& \hat{w} = \left|\begin{smallmatrix} \hat{y} \\ \hat{z} \end{smallmatrix}\right| = \left|\begin{smallmatrix} A \\ H \end{smallmatrix}\right| |\hat{x}| + \left|\begin{smallmatrix} a \\ -h \end{smallmatrix}\right| \ , && \text{modello deterministico.}
\end{aligned}$$

La soluzione del relativo problema di m.q., fornita dalla (2.3.15) è allora

$$\begin{aligned}
\hat{x} \ &= \ \left\{ [A^{+}H^{+}] \left|\begin{smallmatrix} Q^{-1} & 0 \\ 0 & (1/\varepsilon)I \end{smallmatrix}\right| \left|\begin{smallmatrix} A \\ H \end{smallmatrix}\right| \right\}^{-1} \cdot \\
&\cdot \ [A^{+}H^{+}] \left|\begin{smallmatrix} Q^{-1} & 0 \\ 0 & (1/\varepsilon)I \end{smallmatrix}\right| \left|\begin{smallmatrix} Y_o - a \\ h \end{smallmatrix}\right| = \\
&= \ \{A^{+}Q^{-1}A + (1/\varepsilon)H^{+}H\}^{-1}\{A^{+}Q^{-1}(Y_o - a) + (1/\varepsilon)H^{+}h\} \ .
\end{aligned} \tag{2.6.15}$$

La stima di $\sigma_o^2$ è data poi da

$$\hat{\sigma}_o^2 = \frac{(Y_o - a - A\hat{x})^{+}Q^{-1}(Y_o - a - A\hat{x}) + 1/\varepsilon(h - H\hat{x})^{+}(h - H\hat{x})}{n + k - m} \tag{2.6.16}$$

in accordo con la (2.4.20), applicata nel presente caso.

**Osservazione 2.6.1:** nel caso in cui il vincolo consista semplicemente nel porre a zero l'$i$-esimo parametro, si ha

$$\begin{array}{rcl} & & \begin{array}{ccccccccc} 1 & & & 1-i & i & 1+1 & & & m \end{array} \\ H & = & [0 & \ldots\ldots & 0 & 1 & 0 & \ldots\ldots & 0] \\ h & = & 0 & : & & & & & \end{array}$$

pertanto

$$H^+H = \begin{array}{c} \begin{array}{ccccccc} 1 & \ldots & i-1 & i & i+1 & \ldots & m \end{array} \\ \left| \begin{array}{ccccccc} 0 & & & & & & \\ & 0 & & & 0 & & \\ & & & 1 & & & \\ & 0 & & & 0 & & \\ & & & & & 0 & \end{array} \right| \end{array}$$

e tutta l'operazione (2.6.15) consiste essenzialmente nel modificare la matrice normale aggiungendo il numero "grande" $1/\varepsilon$ sull'$i$-esimo elemento in diagonale.

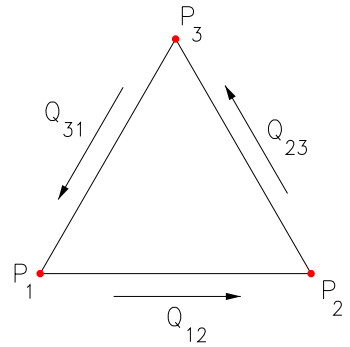



Figura 2.6.2

**Esempio 2.6.3:** si consideri una rete di livellazione (misura di dislivelli) sui tre punti $P_1, P_2, P_3$; le quantità misurabili sono dunque

$$y = \left| \begin{array}{c} Q_{12} \\ Q_{23} \\ Q_{31} \end{array} \right|$$

e il modello deterministico relativo a questo esperimento può essere espresso parametricamente in funzione del vettore $x$ costituito dalle quote dei tre punti

$$x = \left| \begin{array}{c} Q_1 \\ Q_2 \\ Q_3 \end{array} \right| .$$

Le equazioni d'osservazione sono

$$Y = \left| \begin{array}{c} Q_{12} \\ Q_{23} \\ Q_{31} \end{array} \right| = Ax = \left| \begin{array}{c} Q_2 - Q_1 \\ Q_3 - Q_2 \\ Q_1 - Q_3 \end{array} \right| = \left| \begin{array}{ccc} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{array} \right| \left| \begin{array}{c} Q_1 \\ Q_2 \\ Q_3 \end{array} \right|$$

Supposto di avere un vettore di osservazioni $Y_o$ tratto da una $Y$ che abbia come media $y$, il problema della stima di $x$ non può essere trattato in modo ordinario in quanto la matrice $A$ non è di rango pieno; in effetti, posto $e = \left| \begin{array}{c} 1 \\ 1 \\ 1 \end{array} \right|$, risulta

$$Ae = 0$$

così che $\det A = 0$, come è ovvio anche da un calcolo diretto.

Ciò significa che se $\hat{x}$ è una qualsiasi soluzione di minimi quadrati, altrettanto risulta $\hat{x} + \lambda e$ qualsiasi sia $\lambda$, in quanto gli scarti delle equazioni risultano invarianti per tale trasformazione

$$V = Y_o - A(\hat{x} + \lambda e) = Y_o - A\hat{x} .$$

Fisicamente ciò deriva dal fatto che, non avendo fissato un'origine delle quote, le quote possono tutte essere variate di una stessa costante $\lambda$, senza mutare i dislivelli $Q_{ik}$.

Proviamo ora a risolvere il corrispondente problema imponendo un vincolo che elimini l'ambiguità dell'origine; ad esempio imponiamo

$$Q_1 = [1\ 0\ 0] \left[ \begin{array}{c} Q_1 \\ Q_2 \\ Q_3 \end{array} \right] = Hx = 0$$

che ha l'ovvio significato di fissare a 0 la quota di $P_1$.

Riscriviamo allora il sistema normale nella forma (2.6.7) tenendo conto che per ipotesi le misure sono indipendenti e con uguale precisione, ovvero

$$\begin{vmatrix} 2 & -1 & -1 & 1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ 1 & 0 & 0 & 0 \end{vmatrix} \begin{vmatrix} Q_1 \\ Q_2 \\ Q_3 \\ \lambda \end{vmatrix} = \begin{vmatrix} -Q_{012} & + & Q_{031} \\ Q_{012} & - & Q_{023} \\ Q_{023} & - & Q_{031} \\ & 0 & \end{vmatrix} .$$

Dalla quarta equazione si trova ovviamente $\hat{Q}_1 = 0$, che sostituito nella seconda e terza, dà

$$\begin{array}{rrl} \hat{Q}_2 = & Q_{012} & -\frac{\Delta}{3} \\ \hat{Q}_3 = & -Q_{031} & +\frac{\Delta}{3} \end{array} \quad ; (\Delta = Q_{012} + Q_{023} + Q_{031})$$

la prima equazione definisce il valore di $\lambda$ ovvero

$$\lambda = 0 \ ,$$

nella fattispecie.

Vediamo ora come si sarebbe potuto risolvere, in modo approssimato, il problema applicando la formula (2.6.15)

In questo caso la matrice dei pesi è l'unità così che

$$\hat{x}_\varepsilon = (A^+ A + \frac{1}{\varepsilon} H^+ H)^{-1} A^+ Y_o \ , \tag{2.6.17}$$

tenuto conto che $a = 0, \ h = 0$.

Se si usa il vincolo

$$H = [1 \ 0 \ 0] \ ,$$

si può riscrivere il sistema normale di cui (2.6.17) è soluzione, nella forma

$$(D - e\ e^+)\hat{x}_\varepsilon = A^+ Y_o$$
$$D = \begin{vmatrix} \frac{3\varepsilon+1}{\varepsilon} & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{vmatrix} .$$

$(e = \gamma(1\ 1\ 1)^+)$.

Non è difficile verificare che la forma esplicita dell'inversa della normale è data da

$$(D - e\ e^+)^{-1} = D^{-1} + 3(3\varepsilon + 1)D^{-1} e\ e^+ D^{-1} .$$

Posto per brevità $\gamma = 1 + 3\varepsilon$, la forma esplicita della soluzione è allora

$$\hat{x}_\varepsilon = \begin{vmatrix} 0 \\ \varepsilon(-Q_{012} + Q_{031}) + \frac{1+\gamma}{3}(Q_{012} - Q_{023}) + \frac{\gamma}{3}(Q_{023} - Q_{031}) \\ \varepsilon(-Q_{012} + Q_{031}) + \frac{\gamma}{3}(Q_{012} - Q_{023}) + \frac{1+\gamma}{3}(Q_{023} - Q_{031}) \end{vmatrix}$$

Da qui, preso $\varepsilon \to 0$, ovvero $\gamma \to 1$, si vede che

$$\lim_{\varepsilon \to 0} \hat{x}_\varepsilon = \begin{vmatrix} 0 & & \\ Q_{012} & - & \frac{\Delta}{3} \\ -Q_{031} & + & \frac{\Delta}{3} \end{vmatrix} = \hat{x}$$

come volevasi dimostrare.

## 2.7 Problemi di stima non lineari

Vogliamo vedere come si modifica il metodo dei minimi quadrati quando il modello deterministico, che esprime il legame tra i valori medi delle osservabili, sia non lineare.

Da un punto di vista astratto il principio non differisce, come enunciazione, da quello lineare:

dato un vettore di osservazioni $Y_o \in R^n$, estratto da una v.c. $Y$ di covarianza

$$C_{YY} = \sigma_o^2 Q \tag{2.7.1}$$

e data una varietà $V, m$-dimensionale, in $R^n$, cui si sa che deve appartenere la media di $Y$

$$y = E\{Y\} \in V \ , \tag{2.7.2}$$

si cerca lo stimatore $\hat{y} \in V$, di minima distanza da $Y_o$ secondo la matrice $Q^{-1}$,

$$(Y_o - \hat{y})^+ Q^{-1} (Y_o - \hat{y}) = \min \tag{2.7.3}$$

e una stima di $\sigma_o^2$ proporzionale a tale distanza

$$\hat{\sigma}_o^2 = c \cdot (Y_o - \hat{y})^+ Q^{-1} (Y_o - \hat{y}) \ , \tag{2.7.4}$$

essendo $\hat{\sigma}_o^2$ corretto, almeno in modo approssimato.

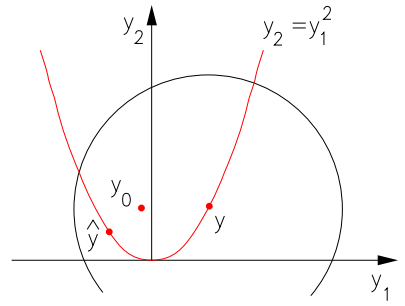
**Esempio 2.7.1:** prima di iniziare la trattazione analitica può essere utile considerare due esempi semplici, particolarmente evidenti dal punto di vista grafico.

Sia $Y_o = \left| \begin{smallmatrix} -0,2 \\ 1 \end{smallmatrix} \right|$ il vettore delle osservazioni e supponiamo in un primo caso che esso sia estratto da $Y = N\left[ \left| \begin{smallmatrix} 1 \\ 1 \end{smallmatrix} \right|, (0,8)^2 I \right]$, e in un secondo che $Y = N\left[ \left| \begin{smallmatrix} 1 \\ 0,01 \end{smallmatrix} \right|, (0,8)^2 I \right]$: i modelli deterministici sono rispettivamente $y_2 = y_1^2$ ed $y_2 = 0.01 y_1^2$. La situazione è rappresentata in fig. 2.7.1 a) e b), dove sono anche mostrati gli stimatori m.q., nonché l'area circolare contenente circa il 99% delle probabilità.

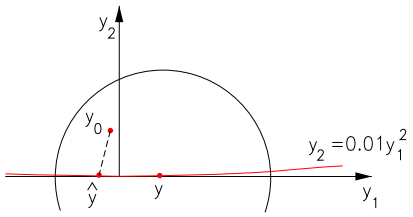
La prima osservazione è che nell'esempio a) lo stimatore m.q. $\hat{y}$ dista di più da $y$ che non $Y_o$ ($|\hat{y} - y| > |Y_o - y|$), mentre nell'esempio b) le cose vanno meglio.

La seconda osservazione è che vi sono punti nel piano in cui non è definita univocamente la proiezione sulla varietà $V$: ad esempio quelli dell'asse $x = 0$ [8], e con $y > 0,5$ per l'esempio a), con $y > 50$ per l'esempio b),

[8] Per comprendere questa affermazione basta osservare che se $(t, at^2)$ è l'equazione parametrica di una parabola e se $(o, y)$ è un punto fissato dell'asse $y$, l'equazione $d/dt\{(y - at^2)^2 + t^2\} = 0$ ha solo la radice $t = 0$ quando $y < (1/2)a$.




a)




b)

Figura 2.7.1:

ammettono due punti di minima distanza, simmetrici sui due rami della parabola. è anche utile però osservare che per il caso a) questi punti sono ben dentro la zona $P = 99\%$, mentre per il caso b) essi sono ben al di fuori.

La terza osservazione è che 0,5 e 50 non sono altro che i raggi di curvatura delle due parabole nell'origine. Pensando un minuto alla definizione di raggio di curvatura, si comprende che il concetto è generale: in tutta l'area definita da un lato dalla curva, dall'altro dalla curva descritta dai suoi centri di curvatura, i punti ammettono una sola proiezione, a minima distanza, sulla curva stessa.

L'esempio si generalizza alle varietà a $m$ dimensioni in uno spazio $R^n$, prendendo il più piccolo dei raggi di curvatura di $V^{(m)}$.

Si possono così trarre alcune conclusioni:

**Conclusione 1:** il metodo dei m.q. non può, nel caso non-lineare, godere di proprietà di ottimalità come nel caso lineare.

**Conclusione 2:** la soluzione del metodo dei m.q. con modelli non-lineari

può non essere unica.

**Conclusione 3:** l'unicità della soluzione è garantita se il raggio della zona di interesse per la distribuzione di $Y$, misurato ad esempio da $\sqrt{\chi^2_{n,\alpha}}\sigma_o$ quando $Q = I$ [9], è assai più piccolo della curvatura della varietà $V$.

In ultima analisi si può affermare che il metodo dei m.q. funziona bene anche nel caso non lineare quando nella zona di interesse la varietà $V$ non si discosta molto da una varietà lineare: qualora tale condizione fosse seriamente violata, come nell'esempio di fig. 2.7.1 a), non ha senso applicare il metodo dei m.q. in quanto l'introduzione dell'informazione data dalla varietà $V$ può peggiorare quella contenuta nel vettore $Y_o$.

Dal punto di vista analitico i problemi non lineari vengono trattati per mezzo di una linearizzazione.

Supponiamo ad esempio che la varietà $V$ sia descritta nella forma più generale, per mezzo di equazioni non lineari contenenti anche parametri aggiuntivi

$$\begin{aligned} & g(y,x) = 0 \\ & g^+(y,x) = [g_1(y,x), \ldots, g_l(y,x)] \\ & y \in R^n \quad , \quad x \in R^m \ . \end{aligned} \tag{2.7.5}$$

Ciò significa quindi che si cercano $\hat{y}, \hat{x}$ tali che

$$\begin{aligned} & (Y_o - \hat{y})Q^{-1}(Y_o - \hat{y}) = \min \\ & g(\hat{y}, \hat{x}) = 0 \end{aligned} \tag{2.7.6}$$

Supponiamo di conoscere dei "valori" approssimati $\tilde{y}, \tilde{x}$ e che nella zona di interesse $g$ sia ben approssimabile con una funzione lineare: poniamo allora

$$\begin{cases} \hat{y} = \tilde{y} + \hat{\eta} \\ \hat{x} = \tilde{x} + \hat{\xi} \end{cases} \tag{2.7.7}$$

[9] Infatti quando $Y$ è normale, e $Q = I$, si ha la relazione $|Y - y|^2 = \chi^2_n$.

e linearizziamo la seconda delle (2.7.6) come

$$g(\tilde{y},\tilde{x}) + \left(\frac{\partial \tilde{g}}{\partial y}\right)\hat{\eta} + \left(\frac{\partial \tilde{g}}{\partial x}\right)\hat{\xi} = 0 \ , \tag{2.7.8}$$

dove

$$\left(\frac{\partial \tilde{g}}{\partial y}\right) = \left[\frac{\partial g_i(\tilde{y},\tilde{x})}{\partial y_k}\right]$$
$$\left(\frac{\partial \tilde{g}}{\partial y}\right) = \left[\frac{\partial g_i(\tilde{y},\tilde{x})}{\partial x_j}\right] \ .$$

La (2.7.8) è il nostro nuovo modello, ora lineare, nelle variabili $\xi$ (parametri) ed $\eta$ (osservabili), per le quali si può anche porre

$$\begin{aligned} &\eta_o = Y_o - \tilde{y} \qquad \text{valori osservati} \\ &C_{\eta\eta} = C_{YY} = \sigma_o^2 Q \ . \end{aligned} \tag{2.7.9}$$

Notiamo che basterà porre

$$\begin{aligned} -\left(\frac{\partial \tilde{g}}{\partial y}\right) &= D \\ \left(\frac{\partial \tilde{g}}{\partial x}\right) &= A \\ g(\tilde{y},\tilde{x}) &= [g_i(\tilde{y},\tilde{x})] = d \end{aligned} \tag{2.7.10}$$

per trasformare la (2.7.8) nel modello già trattato nel paragrafo 2.2,

$$D\eta = A\xi + d \ \ :$$

ciò insieme alle (2.7.9) permette di ricavare le stime cercate $\hat{\eta}, \hat{\xi}$ applicando la teoria lineare. Una volta ottenute le stime $\hat{\eta}, \hat{\xi}$ si possono calcolare le corrispondenti $\hat{y}, \hat{x}$ dalle (2.7.7).

**Osservazione 2.7.1:** (iterazioni). ci si può chiedere ora se il vettore $\hat{y}$ così determinato appartenga veramente a $V$: infatti $\eta, \xi$ sono stati

determinati in modo che valesse solo l'equazione approssimata (2.7.8). A tale scopo si possono calcolare gli scarti

$$-\hat{v} = g(\hat{y}, \hat{x}) = g(\tilde{y} + \hat{\eta}, \tilde{x} + \hat{\xi}) \tag{2.7.11}$$

e si può valutare l'entità di $|\hat{v}|$ decidendo se questa è trascurabile rispetto agli errori di misura.[10]

Qualora si pensi che $\hat{v}$ non sia abbastanza piccolo, ciò significa che i valori approssimati $\tilde{y}, \tilde{x}$ erano troppo grossolani così che l'approssimazione lineare (2.7.8) non risulta buona.

In questo caso è necessario iterare il procedimento prendendo i valori $\hat{y}_1, \hat{x}_1$ determinati al primo passo come nuovi punti di linearizzazione.

Si può notare che nella nuova iterazione il nuovo vettore $d$ risulterà proprio uguale a

$$d = g(\hat{y}_1, \hat{x}_1) \ ,$$

in quanto $\hat{y}_1, \hat{x}_1$ sono i nuovi valori approssimati.

Ripetendo più volte lo stesso ragionamento si trova l'algoritmo iterativo

[10] Ciò naturalmente riportato alle opportune unità di misura: ad esempio si potrebbe prendere come criterio

$$|\hat{v}|^2 << (l-m)\hat{\sigma}_o^2 = U^+ K^{-1} U \ .$$

Passo i+1

$$\hat{y}_{i+1} = \hat{y}_i + \hat{\eta}_{i+1}$$
$$\hat{x}_{i+1} = \hat{x}_i + \hat{\xi}_{i+1} \ ,$$
$$D_i = -\left[\frac{\partial g(\hat{y}_i, \hat{x}_i)}{\partial y}\right]$$
$$A_i = \left[\frac{\partial g(\hat{y}_i, \hat{x}_i)}{\partial x}\right]$$
$$d_i = [g(\hat{y}_i, \hat{x}_i)]$$
$$\eta_{o,i+1} = Y_o - \hat{y}_i \qquad C_{\eta\eta} = C_{YY}$$
$$D_i\hat{\eta}_{i+1} = A_i\hat{\xi}_{i+1} + d_i$$
$$\downarrow$$
Principio m.q.
$$\downarrow$$
$$\hat{\eta}_{i+1}, \hat{\xi}_{i+1} \ ,$$

Passo 1

$$\hat{y}_o = \tilde{y} \qquad \text{(valori approssimati iniziali)}$$
$$\hat{x}_o = \tilde{x}$$

Il criterio di stop è basato sulla dimensione di $|d_i|$; quando $|d_i|^2$ è abbastanza piccolo rispetto $(l-m)\sigma_o^2$, l'iterazione è fermata.

Altri criteri di stop possono essere fissati in base a $|\xi_i|$ e $|\eta_i|$.

**Osservazione 2.7.2:** per la determinazione dei valori approssimati non esistono regole generali. Quando non vi sia di meglio, come primi valori approssimati di $y$ si può sempre scegliere

$$\tilde{y} = Y_o \ : \tag{2.7.12}$$

è bene notare che la (2.7.12) va presa come *eguaglianza tra due vettori costanti* non tra due v.c., senza di che sarebbe falsa la relazione

$$C_{\eta\eta} = C_{YY} \ .$$

Per determinare i valori $\tilde{x}$ molte volte si cerca di risolvere una parte delle equazioni (2.7.5) rispetto ad $x$ ottenendo

$$x = f(y)$$

e ponendo di conseguenza

$$\tilde{x} = f(\tilde{y}) \tag{2.7.13}$$

**Osservazione 2.7.3:** nel caso che $g(y, x)$ sia già lineare in una delle due variabili, non serve linearizzare rispetto a quella variabile. Il caso più importante è quello in cui le equazioni non lineari della varietà $V$ sono in forma puramente parametrica *(equazioni d'osservazione)*

$$y = g(x) \ . \tag{2.7.14}$$

Per linearizzare le (2.7.14) basta cercare dei valori approssimati $\tilde{x}$, ad esempio invertendo una parte di tali equazioni e usando i corrispondenti valori osservati $Y_{oi}$ come termini noti. Successivamente si pone

$$\hat{y} = g(\tilde{x} + \tilde{\xi}) \cong g(\tilde{x}) + \left(\frac{\partial \tilde{g}}{\partial x}\right) \hat{\xi} = a + A\hat{\xi} \ . \tag{2.7.15}$$

Trovata la soluzione $\hat{\xi}$ si può calcolare

$$\hat{y}_1 = g(\tilde{x} + \hat{\xi}_1) \ , \tag{2.7.16}$$

in modo tale che la stima $\hat{y}_1$ corrispondente appartiene esattamente a $V$, al contrario del caso generale (cfr. (2.7.11)). Lo schema può essere iterato ponendo

$$\begin{aligned} \hat{y}_{i+1} &= g(\hat{x}_i + \hat{\xi}_{i+1}) \cong g(\hat{x}_i) + \frac{\partial g}{\partial x}(\hat{x}_i) \cdot \hat{\xi}_{i+1} \\ &= a_i + A_i \hat{\xi}_{i+1} \ : \end{aligned} \tag{2.7.17}$$

spesso, per semplificare i conti risparmiando tempo di calcolo, la matrice $A$ non viene ricalcolata ogni volta, ma piuttosto viene lasciata fissa al valore iniziale

$$A = \frac{\partial g(\tilde{x})}{\partial x} \ . \tag{2.7.18}$$

Si può dimostrare che ciò provoca una distorsione del 2° ordine nella stima.

Lo schema iterativo è bloccato qundo il $\hat{\sigma}^2_{oi}$ non diminuisce più che di una piccola percentuale

$$\frac{\hat{\sigma}^2_{oi} - \hat{\sigma}^2_{oi+1}}{\hat{\sigma}^2_{oi}} < \varepsilon \quad \text{prefissato} \ . \tag{2.7.19}$$

## 2.8 Applicazione alla regressione lineare

Supponiamo di avere un processo descrivibile mediante una (o più) variabili $y$ su cui compiamo delle osservazioni $Y_{oi}$ che comprendono una parte stocastica e un valore medio $y_i$ che si sa essere funzione di altre $p$ variabili $t_1, t_2, \ldots, t_p$ completamente note per ognuna delle $i = 1, \ldots, n$ osservazioni

$$\begin{cases} Y_{oi} = y_i + \varepsilon_i \\ y_i = g(t_{1i}, t_{2i}, \ldots, t_{pi}) \end{cases} \ , \tag{2.8.1}$$

$$E\{\varepsilon_i\} = 0 \ , \ C_{\varepsilon\varepsilon} = \sigma^2_o I \ . \tag{2.8.2}$$

Ciò che importa sottolineare nel modello (2.8.1), (2.8.2) è che la parte stocastica va considerata aggiunta, per ogni osservazione $i$, alla variabile dipendente $y$, detta anche *variabile criterio*, mentre le variabili indipendenti $t_1, \ldots, t_p$, per la stessa osservazione, sono considerate costanti, cioè comunque note con una precisione tale da non modificare significativmente il modello stocastico di $\varepsilon$; le variabili $t_1, \ldots, t_p$ sono dette *variabili concomitanti* o di predizione.

Spesso siamo in una situazione di analisi puramente empirica dei dati in cui il legame $y = g(t_1, \ldots, t_p)$ non è noto perché troppo complesso: tuttavia da una prima indagine dei dati stessi o da una conosenza qualitativa del fenomeno sotto indagine possiamo ritenere che un'approssimazione lineare di $g$ sia ben giustificata almeno nell'ambito dei valori assunti dalle variabili concomitanti in corrispondenza alle nostre osservazioni. Si potrà così ragionevolmente sostituire alla II delle (2.8.1) l'equazione

$$\begin{cases} y_i = g(\overline{t}_1, \ldots, \overline{t}_p) + \displaystyle\sum_{k=1}^{p} \left( \frac{\partial \overline{g}}{\partial t_k} \right) \cdot (t_{ki} - \overline{t}_k) \\ \overline{t}_k = \frac{1}{n} \displaystyle\sum_{i=1}^{n} t_{ki} \end{cases} . \qquad (2.8.3)$$

**Osservazione 2.8.1:** si noti che l'uso delle medie aritmetiche $\overline{t}_k$ per linearizzare $g$, è naturale per ottimizzare la validità della approssimazione lineare stessa. Per semplicità nei conti successivi, sebbene ciò non sia strettamente necessario, supporremo di aver già scelto l'origine delle variabili concomitanti in modo tale che si abbia il baricentro dei punti d'osservazione nell'origine

$$\overline{t}_k = 0 \qquad k = 1, \ldots, p \; ; \qquad (2.8.4)$$

qualora tale condizione non fosse già soddisfatta basterà una semplice traslazione per ottenere nuove variabili indipendenti

$$\tau_{ki} = t_{ki} - \overline{t}_k \; ,$$

che invece soddisferanno la (2.8.4).

Supponendo dunque che valga la (2.8.4), il modello (2.8.3) diventa

$$y_i = c_o + c_1 t_{1i} + \ldots + c_p t_{pi} \; ; \qquad (2.8.5)$$

i parametri incogniti di questo modello sono $c_o, c_1, \ldots, c_p$ che organizzeremo nel vettore

$$\left| \begin{array}{c} x_o \\ \underline{x} \end{array} \right| , (x_o = c_o, \; \underline{x} = \left| \begin{array}{c} c_1 \\ \vdots \\ c_p \end{array} \right|) \; .$$

In forma matriciale la (2.8.5) può essere riscritta come

$$y = \underline{e} x_o + T \underline{x} \qquad (2.8.6)$$

dove

$$e = \left| \begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right| \quad T = \left| \begin{array}{ccc} t_{11} & \dots & t_{pi} \\ \vdots & & \vdots \\ t_{1n} & \dots & t_{pn} \end{array} \right| \ . \tag{2.8.7}$$

Il modello deterministico (2.8.6) e quello stocastico (2.8.2) insieme definiscono un tipico problema di minimi quadrati lineari in forma parametrica, la cui compensazione fornirà le stime ricercate $\hat{x}_o, \underline{\hat{x}}$.

**Osservazione 2.8.2:** notiamo che in virtù della condizione (2.8.4) si ha

$$T^+ \underline{e} = n \left| \begin{array}{c} \overline{t}_1 \\ \vdots \\ \overline{t}_p \end{array} \right| = 0 \ ,$$

così che

$$(T\underline{x})^+ (\underline{e} x_o) = 0 \ . \tag{2.8.8}$$

Ciò significa che nel modello (2.8.6) la varietà dei valori ammissibili (ovviamente a $p+1$ dimensioni) è decomposta in due sottovarietà ortogonali $\{\underline{e} x_o\}, \{T\underline{x}\}$ rispettivamente a 1 e a $p$ dimensioni.

Se ora formiamo il sistema normale per la (2.8.6), indicato al solito con $Y_o$ il vettore delle osservazioni e osservato che nella (2.8.6) non si ha termine noto $a$, si trova

$$\left| \begin{array}{cc} \underline{e}^+ \underline{e} & 0 \\ 0 & T^+ T \end{array} \right| \left| \begin{array}{c} \hat{x}_o \\ \underline{\hat{x}} \end{array} \right| = \left| \begin{array}{c} \underline{e}^+ Y_o \\ T^+ Y_o \end{array} \right| \ . \tag{2.8.9}$$

Si ha $\underline{e}^+ \underline{e} = n$. Poniamo inoltre

$$\begin{cases} (1/n)(T^+T)_{kl} = (1/n) \sum_{i=1}^{n} t_{ki} t_{li} = C_{tt} \\ (1/n) \underline{e}^+ Y_o = (1/n) \sum_{i=1}^{n} Y_{oi} = \overline{Y} \\ (1/n) T^+ Y_o = \left| \begin{array}{c} (1/n) \sum_i t_{1i} Y_{oi} \\ (1/n) \sum_i t_{pi} Y_{oi} \end{array} \right| = C_{ty} \end{cases} \tag{2.8.10}$$

e notiamo che queste relazioni sembrano proprio quelle di media, covarianza e cross-varianza di valori campionari, motivo questo della simbologia adottata: tuttavia è necessario sottolineare che né $C_{tt}, C_{ty}$ possono

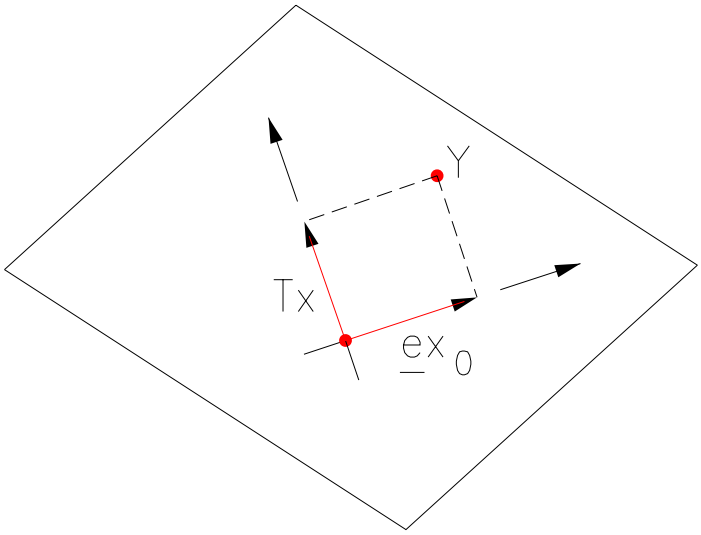



Figura 2.8.1:

essere covarianze, né $\overline{Y}$ è una media, perché da un lato $\{t_{ki}\}$ non sono campioni estratti da variabili casuali $t_k$ (le variabili concomitanti sono solo costanti note esattamente), dall'altro lato le componenti $Y_{oi}$ non sono estrazioni dalla stessa v.c. $Y$ in quanto le medie di $Y_{oi}$ hanno valori diversi. Con questa simbologia il sistema normale (2.8.9), dividendo entrambi i membri per $n$, dà la soluzione

$$\begin{cases} \hat{x}_o = \overline{Y} \\ \underline{\hat{x}} = C_{tt}^{-1} C_{ty} \end{cases} , \tag{2.8.11}$$

con matrice di covarianza

$$\sigma^2(\hat{x}_o) = \frac{\sigma_o^2}{n}$$

$$\begin{cases} C_{\underline{\hat{x}}\underline{\hat{x}}} = \sigma_o^2 (nC_{tt})^{-1} = \dfrac{\sigma_o^2}{n} C_{tt}^{-1} \\ C_{\hat{x}_o \underline{\hat{x}}} = 0 \ . \end{cases} \tag{2.8.12}$$

Per la stima di $\sigma_o^2$ formiamo dapprima il vettore degli scarti

$$U = Y_o - \overline{Y}\underline{e} - T\underline{\hat{x}} \ : \tag{2.8.13}$$

ricordando la (2.8.8) e la (2.8.9) si ha

$$\begin{aligned} |U|^2 &= U^+U = |Y_o - \overline{Y}\underline{e}|^2 - 2\hat{x}^+T^+Y_o + \hat{x}^+T^+T\hat{x} = \\ &= |Y_o - \overline{Y}\underline{e}|^2 - |T\hat{x}|^2 = |Y_o - \overline{Y}\underline{e}|^2 - n\hat{x}^+C_{ty} = \\ &= \sum_i (Y_{oi} - \overline{Y})^2 - \sum_{k,i} \hat{c}_k t_{ki} Y_{oi} \ . \end{aligned} \tag{2.8.14}$$

Da qui in poi si trova $\hat{\sigma}_o^2$ secondo la regola generale

$$\hat{\sigma}_o^2 = \frac{1}{n-p-1} U^+U \ .$$

**Osservazione 2.8.3:** l'equazione (2.8.14) esprime semplicemente la decomposizione pitagorica di fig. 2.8.2 e fornisce uno strumento comodo per la stima di $\sigma_o^2$.

Inoltre, notando che $Y_{oi} - \overline{Y}$ può essere interpretato "formalmente" come scarto generale di $Y_{oi}$ dalla media $\overline{Y}$, $(T\hat{x})_i = \sum_{k=1}^p t_{ki}\hat{c}_k$ come scarto tra la regressione $(\hat{y}_i = \hat{c}_o + \sum_k t_{ki}\hat{c}_k)$ e la media $\overline{Y}(\hat{c}_o = \overline{Y})$, cioè come lo scarto spiegato dall'equazione di regressione, $U_i = Y_{oi} - \overline{y}_i$ come scarto residuo tra il valore osservato ed il valore della regressione (cfr. fig. 2.8.3), si può riscrivere la (2.8.14) nella forma

$$S_G^2 = S_R^2 + S_S^2 \tag{2.8.15}$$

con

$$\begin{aligned} S_G^2 &= \sum (Y_{oi} - \overline{Y})^2 && = \text{scarti generali} \\ S_R^2 &= \sum (Y_{oi} - \overline{y}_i)^2 && = \text{scarti residui} \\ S_S^2 &= \sum (\overline{y}_i - \overline{Y})^2 && = \text{scarti spiegati} \end{aligned}$$

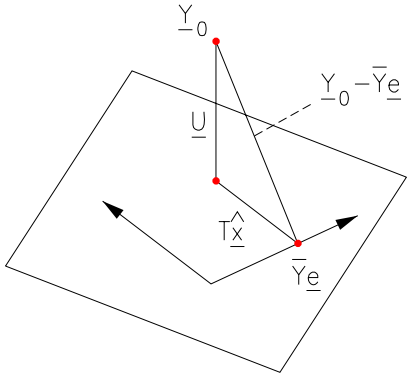



Figura 2.8.2:

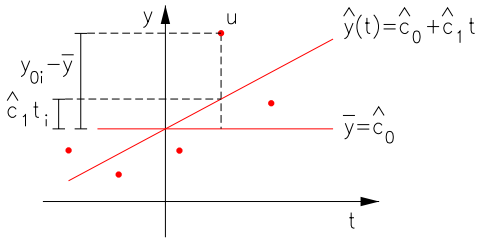



Figura 2.8.3:

**Osservazione 2.8.4:** si noti che nel caso di una sola variabile concomitante, la retta di regressione stimata

$$\hat{y} = \hat{c}_o + \hat{c}_1 t$$

diventa, secondo le (2.8.11) ($\overline{t} = 0$)

$$\hat{y} = \overline{Y} + \frac{\sigma_{ty}}{\sigma_t^2} t \ ,$$

che coincide formalmente con la retta di regressione per una variabile doppia $(Y, t)$, sebbene in questo caso l'interpretazione sia notevolmente diversa.

**Osservazione 2.8.5:** in generale il motivo per cui interessa costruire il modello (2.8.5) non è tanto quello di stimare $c_o, c_1, \ldots, c_p$ e quindi

anche i valori della regressione $\hat{y}_i$ per, diciamo così, regolarizzare i valori osservati $Y_{oi}$, ma piuttosto quello di poter prevedere pur con un margine di errore, il valore della regressione $\hat{y}$ in corrispondenza di valori delle variabili concomitanti per i quali non si hanno osservazioni.

Naturalmente il valore predetto sarà quello della regressione

$$\hat{y}(t_1, \ldots, t_p) = \hat{c}_o + \sum_{k=1}^{p} \hat{c}_k t_k \ ,$$

che, introdotto il vettore

$$\underline{t} = \left| \begin{array}{c} t_1 \\ \vdots \\ t_p \end{array} \right| \ ,$$

può anche essere scritto come

$$\hat{y}(\underline{t}) = \hat{x}_o + \underline{\hat{x}}^{+} \underline{t} \ . \tag{2.8.16}$$

Ricordando le (2.8.12), possiamo accompagnare il valore predetto con la sua varianza (ovvero l'errore di predizione)

$$\sigma^2[\hat{y}(\underline{t})] = \frac{\sigma_o^2}{n}[1 + \underline{t}^{+} C_{tt}^{-1} \underline{t}] \ : \tag{2.8.17}$$

nel caso di una sola variabile $t$, la (2.8.17) assume la forma

$$\sigma^2[\hat{y}(\underline{t})] = \frac{\sigma_o^2}{n} \left[ 1 + \frac{t^2}{(1/n) \sum_{i=1}^{n} t_i^2} \right] \ : \tag{2.8.18}$$

Per stimare queste varianze ovviamente occorrerà avvalersi del valore stimato $\hat{\sigma}_o^2$.

Dalla (2.8.18) appare in modo più chiaro che l'errore di predizione si mantiene limitato finché $t$ rimane all'interno dei valori di osservazione

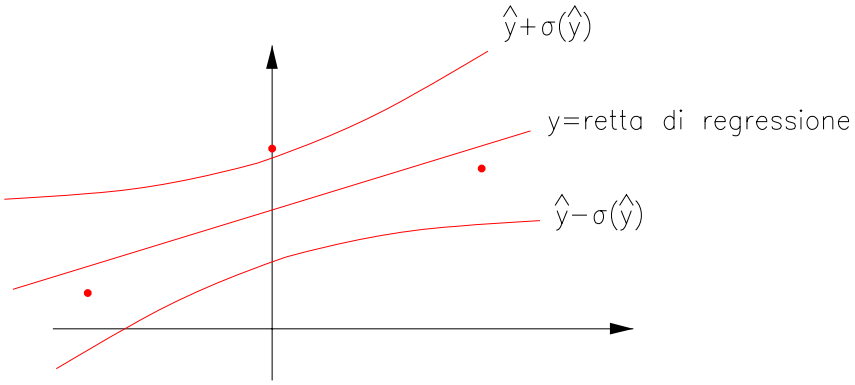



Figura 2.8.4:

($|t| \leq \max |t_i|$), mentre esso aumenta rapidamente quando si tenti di estrapolare la regressione all'esterno dei valori di osservazione.

**Esempio 2.8.1:** nelle operazioni di controllo di una struttura, si misurano le quote di un punto per 6 volte con intervalli di due mesi per verificare gli eventuali movimenti di assestamento del punto stesso. Poiché però si sa che la forma geometrica della struttura dipende anche dalla sua temperatura, contestualmente alla misura della quota si fa un lettura di temperatura.

Si ipotizza che valga un modello di regressione lineare del tipo

$$Q = c_o + c_1 t + c_2 T \quad (Q = \text{quota},\ t = \text{tempo},\ T = \text{temperatura})$$
$$Q_{oi} = Q_i + \varepsilon_i \qquad C_{\varepsilon\varepsilon} = \sigma_o^2 I \quad .$$

I dati ottenuti sono:

| | $t$ (in mesi) | $T$ (in $C^\circ$) | $Q$ (in mm) ($Y_o$) |
|---|---|---|---|
| | 2 | 4,8 | 127,02 |
| | 4 | 9,3 | 129,22 |
| | 6 | 12,5 | 130,98 |
| | 8 | 19,2 | 133,74 |
| | 10 | 10,9 | 129,98 |
| | 12 | 5,2 | 127,01 |
| Medie: | $\overline{t} = 7$ | $\overline{T} = 10,3167$ | $\overline{Y} = 129,6583$ |

Si osservi in primo luogo che né $t$ né $T$ sono centrate, perciò è utile passare a nuove variabili centrate ponendo

$$t = \overline{t} + \tau$$
$$T = \overline{T} + \theta$$

e riformando il modello come $y = Q = c_o + c_1\tau + c_2\theta$.

Si ottiene così la nuova tabella

| $\tau$ | $\theta$ | $Y_o - \overline{Y}$ | |
|---|---|---|---|
| -5 | -5,5167 | -2,6383 | $n$=6 |
| -3 | -1,0167 | -0,4383 | |
| -1 | 2,1833 | 1,3217 | $p = 2$ |
| 1 | 8,8833 | 4,0817 | |
| 3 | 0,5833 | 0,3217 | $(m = 3)$ |
| 5 | -5,1167 | -2,6483 | |
| $S^2_\tau = 70$ | $S^2_\theta = 141,6683$ | $S^2_G = 32,6770$ | grad. lib. |
| | $S_{\tau\theta} = 13,500$ | $S_{\tau Y_o} = 4,9900$ | di $\hat{\sigma}^2_o = 3$ |
| | | $S_{\theta Y_o} = 67,8832$ | |

Da cui

$$C_{tt} = \begin{vmatrix} 11,\overline{6} & 2,2500 \\ -2,2500 & 23,6114 \end{vmatrix}, \; C^{-1}_{tt} = \begin{vmatrix} 0,0873 & -0,0082 \\ -0,0082 & 0,0431 \end{vmatrix}$$
$$C_{ty} = \begin{vmatrix} 0,831\overline{6} \\ 11,3139 \end{vmatrix} .$$

Dunque le stime sono

$$\hat{c}_o = \hat{x}_o \quad = 129,6583$$
$$\begin{vmatrix} \hat{c}_1 \\ \hat{c}_2 \end{vmatrix} = \underline{\hat{x}} \quad = \begin{vmatrix} -0,021535 \\ 0,481227 \end{vmatrix} .$$

Inoltre, ricordando le (2.8.14), (2.8.15)

$$S_R^2 = S_G^2 - 6 \cdot \underline{\hat{x}}^+ C_{ty} = 0,1172 \ .$$

così che

$$\hat{\sigma}_o^2 = 0,0391 \quad , \quad \hat{\sigma}_o = 0,1977 \ .$$

In tal modo si ha anche la stima di varianza e covarianza di $x_o, \underline{x}$, cioè

$$\begin{array}{ll} \sigma^2(\hat{x}_o) & = 0,0065 \\ C_{\underline{\hat{x}}\hat{x}} & = 0,0065 \end{array} \left| \begin{array}{cc} 0,0873 & -0,0082 \\ -0,0082 & 0,0431 \end{array} \right| \ :$$

in particolare è

$$\begin{aligned} \sigma(\hat{c}_1) &= 0,0238 \\ \sigma(\hat{c}_2) &= 0,0167 \ . \end{aligned}$$

Come si vede $\hat{c}_1$ e $\sigma(\hat{c}_1)$ sono quasi uguali, il che fa pensare ad una dipendenza da $\tau$ (cioè da $t$) non molto significativa; resta invece sensibile la dipendenza da $\theta$ (cioè da $T$). Supponiamo ora di voler predire il valore di $y$ in corrispondenza a $t = 7,\ T = 15°$, cioè $\tau = 0,\ \theta = 4,6833$: si ha

$$\overline{y}(0; 4,6833) = \hat{c}_o + \hat{c}_1 \cdot 0 + \hat{c}_2 \cdot 4,6833 = 131,9120 \ .$$

Infine, volendo calcolare l'errore di stima di tale predizione, si ha

$$\begin{aligned} \sigma^2(\hat{y}) &= \frac{\sigma_o^2}{6} \left\{ 1 + [0 \quad 4,6833] C_{tt}^{-1} \left| \begin{array}{c} 0 \\ 4,6833 \end{array} \right| \right\} = \\ &= \frac{0,0391}{6} \{1 + 0,9453\} = \frac{0,0761}{6} = 0,0127 \end{aligned}$$

ovvero

$$\sigma(\hat{y}) = 0,11 \text{ mm} \ .$$