Il peso dello stimatore di massima verosimiglianza converge ad 1 quando la dimensione del campione n diverge all'infinito; intuitvamente, le nostre opinioni a priori sono schiacciate dalla forza dell'evidenza empirica.

Exercise 127 *Sia X una Gaussiana con valor medio θ e varianza σ^2, con θ essa stessa Gaussiana di parametri μ, τ^2. Abbiamo che*

$$\pi(\theta|X) \sim N(\frac{\tau^2}{\tau^2+\sigma^2}X + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}) \ .$$

16 Il Modello lineare multivariato

Nella pratica, è estremamente comune la situazione in cui si cerchi di stimare la relazione esistente tra una particolare variabile (che chiameremo genericamente Y) ed un gruppo di altre (che chiameremo genericamente $X_1, ..., X_k$), dove queste ultime sono considerate esplicative del comportamento della Y. Si tratta di un caso particolare di un ambito di ricerca al momento estremamente attivo, e che viene spesso indicato come Supervise Learning; in generale, lo studio di relazioni come $Y = f(X)$ va anche sotto il nome di Machine Learning o Statistical Learning. In questo corso, ci limiteremo a trattare il caso più comune, che è anche il più semplice; quello in cui assumiamo che $f(.)$ abbia una forma lineare in un vettore di parametri β. In particolare, supponiamo che valga la seguente relazione, per $i = 1, 2, ..., n$:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k + \varepsilon_i \ ,$$

dove le variabili ε_i sono considerate dei "residui" o degli "errori" e catturano tutto quello che non è spiegato dalla relazione lineare tra le y e le x_i. In forma matriciale, possiamo scrivere

$$Y = X\beta + \varepsilon \ , \tag{1}$$

dove Y è un vettore $n \times 1$ della forma $Y = (y_1, ..., y_n)^T$, X è una matrice $n \times k$ le cui colonne sono costituite da $(x_{j1}, ..., x_{jn})^T$, e ε è un vettore di residui; iniziamo supponendo che si tratti di residui Gaussiani, con valor medio nullo e matrice di varianza/covarianza $E[\varepsilon\varepsilon^T] = \Omega$. Comunemente la prima colonna di X viene scelta come il vettore di costanti $(1, 1, ..., 1)$; così ad esempio per $k = 2$ abbiamo

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \ .$$

Abbiamo bisogno di una condizione ulteriore sulle variabili X :

Condition 128 *Le variabili X sono deterministiche ed il rango della matrice è esattamente pari a k.*

Ambedue le condizioni possono essere rimosse ed effettivamente lo sono nella ricerca più recente; si tratta però di ipotesi semplificatrici che costituiscono un punto di partenza naturale. Si noti che questa condizione implica immediatamente $k \leq n$; questa ipotesi in particolare viene abbandonata negli studi più recenti (high-dimensional statistics/big data).

Proposition 129 *Lo stimatore di massima verosimiglianza nel modello 1 ha la seguente forma:*

$$\begin{aligned}\widehat{\beta}_{MLE} &= (X^T X)^{-1} X^T Y \\ \widehat{\sigma}^2_{MLE} &= \frac{1}{n}(\widehat{\varepsilon}^T \widehat{\varepsilon}) \ , \ \widehat{\varepsilon} = Y - \widehat{Y} = Y - X\widehat{\beta} \ .\end{aligned}$$

Si ha inoltre che

$$\begin{aligned}&\widehat{\beta}_{MLE} \stackrel{d}{=} N(\beta, \sigma^2 (X^T X)^{-1}) \ , \\ &n \times \frac{\widehat{\sigma}^2_{MLE}}{\sigma^2} \stackrel{d}{=} \chi^2_{n-k} \ .\end{aligned}$$

Proof. Notiamo innanzitutto che la funzione di verosimiglianza prende la forma

$$\begin{aligned}L(\beta, \sigma^2; Y, X) &= \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta) \right\} \ , \\ \log L(\beta, \sigma^2; Y, X) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta) \ .\end{aligned}$$

Il calcolo del gradiente ci dà facilmente

$$\begin{aligned}\nabla_\beta \log L(\beta, \sigma^2; Y, X) \\ \frac{\partial \log L(\beta, \sigma^2; Y, X)}{\partial \sigma^2}\end{aligned} \quad \begin{aligned}&= \begin{array}{c} \nabla_\beta (Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^T (Y - X\beta) \end{array} \\ &= \begin{array}{c} -2X^T Y + 2X^T X\beta \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^T (Y - X\beta) \end{array}\end{aligned}$$

and setting these quantities to zero leads to the previous expressions for $\widehat{\beta}_{MLE}, \widehat{\sigma}^2_{MLE}$. It is also immediate to see that

$$\begin{aligned}\widehat{\beta}_{MLE} &= (X^T X)^{-1} X^T Y \\ &= (X^T X)^{-1} X^T (X\beta + \varepsilon) \\ &= \beta + (X^T X)^{-1} X^T \varepsilon \ ,\end{aligned}$$

whence

$$\begin{aligned}E[(\widehat{\beta}_{MLE} - \beta)(\widehat{\beta}_{MLE} - \beta)^T] &= E[(X^T X)^{-1} X^T \varepsilon ((X^T X)^{-1} X^T \varepsilon)^T] \\ &= E[(X^T X)^{-1} X^T \varepsilon\varepsilon^T X (X^T X)^{-1}] \\ &= (X^T X)^{-1} X^T E[\varepsilon\varepsilon^T] X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \ .\end{aligned}$$

La dimostrazione del risultato sullo stimatore della varianza richiede alcune ulteriori discussioni e sarà riportata più avanti. ■

Remark 130 *Lo stimatore $\widehat{\beta}_{MLE}$ è anche noto come OLS (Ordinary Least Squares).*

16.1 Il rapporto con i teoremi di proiezione

Lo stimatore di massima verosimiglianza nel modello lineare multivariato si presta ad una importante interpretazione usando gli strumenti dell'algebra lineare. In particolare, notiamo che il valor medio di Y dato X è dato da $E[Y] = X\beta$; il valore di β è ignoto, ma è naturale definire il valore $\widehat{Y} := X\widehat{\beta}$ come il valore "previsto" per il vettore Y sulla base delle stime $\widehat{\beta}$ e del valore dei regressori X. Analogamente, il vettore degli "errori" $\varepsilon = Y - X\beta$ si può stimare come $\widehat{\varepsilon} = Y - X\widehat{\beta}$. Notiamo ora che

$$\begin{aligned} \widehat{Y} &: = X\widehat{\beta} = X(X^T X)^{-1} X^T Y = P_X Y \ , \\ P_X &: = X(X^T X)^{-1} X^T \ . \end{aligned}$$

E' immediato verificare che P_X è una matrice di proiezione, cioè è simmetrica ed idempotente; infatti

$$P_X^2 = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X(X^T X)^{-1} X^T \ .$$

In particolare, l'azione della matrice P_X (che ha dimensioni $n \times n$) corrisponde a proiettare il vettore Y sullo spazio vettoriale di dimensione k generato dalle colonne di X (ricordiamo che queste colonne sono linearmente indipendenti per ipotesi). Analogamente abbiamo

$$\begin{aligned} \widehat{\varepsilon} &= Y - X\widehat{\beta} = M_X Y \ , \\ M_X &= I - P_X = I - X(X^T X)^{-1} X^T \ . \end{aligned}$$

Anche M_X è simmetrica ed idempotente:

$$M_X^2 = (I - P_X)^2 = I - 2P_X + P_X^2 = I - 2P_X + P_X = M_X \ .$$

In particolare, l'azione di M_X consiste nel proiettare Y nello spazio ortogonale a quello generato dalle colonne di X. Questo ha alcune conseguenze importanti; si ha infatti

$$M_X P_X = P_X M_X = \mathbf{0} \ ,$$

dove intendiamo con $\mathbf{0}$ la matrice $n \times n$ costituita da tutti zeri. Come ulteriore conseguenza, notiamo che

$$X^T \widehat{\varepsilon} = 0 \ ,$$

ed in particolare il vettore dei residui stimati è ortogonale a qualsiasi vettore che giaccia nello spazio generato dalle colonne di X.

Remark 131 *Supponiamo che la matrice X contenga un termine costante, cioè una colonna della forma $e = (1, ..., 1)^T$. Allora abbiamo $e^T \widehat{\varepsilon} = 0$, cioè $\sum_{i=1}^n \widehat{\varepsilon}_i = 0$; in altre parole, la somma dei residui di regressione è sempre nulla se tra i regressori è incluso un termine costante.*

Remark 132 *Il fatto che lo stimatore di massima verosimiglianza nel modello lineare multivariato coincida con la soluzione di un problema puramente geometrico di proiezione su sottospazi vettoriali è assolutamente degno di nota. Si tratta dell'ennesima sorprendente proprietà della legge Gaussiana.*

Possiamo ora enunciare le seguenti ulteriori proprietà delle matrici P_X, M_X.

Lemma 133 *Le matrici P_X,M_X hanno tutti autovalori pari a zero o uno e rango $k, n-k$, rispettivamente.*

Proof. Poichè le matrici sono reali e simmetriche, possiamo diagonalizzarle come

$$\begin{aligned} P_X &= Q\Lambda Q^T \text{ , con } QQ^T = Q^TQ = I_n \text{ ,} \\ \Lambda &= \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & 0 & ... \\ ... & 0 & ... & 0 \\ 0 & ... & 0 & \lambda_n \end{pmatrix} . \end{aligned}$$

Si ha allora

$$P_X^2 = Q\Lambda Q^TQ\Lambda Q^T = Q\Lambda^2 Q^T = Q\Lambda Q^T \text{ ,}$$

da cui segue che necessariamente $\lambda_i = 0$ o 1, per $i = 1, 2, ..., n$; ragionamento identico si applica a M_X. Per quello che riguarda il rango, notiamo che esso eguaglia il numero di autovalori diversi da zero (con molteplicità), quindi nel caso di matrici di proiezioni la traccia (cioè la sommadi autovalori, uguale al numero di quelli che valgono 1). Ricordando che $Tr(AB) = Tr(BA)$, possiamo scrivere

$$Tr(P_X) = Tr(X(X^TX)^{-1}X^T) = Tr((X^TX)^{-1}X^TX) = Tr(I_k) = k = Rg(P_X) \text{ .}$$

La dimostrazione per M_X è identica. ■

Siamo ancora in grado di concludere la dimostrazionne sul comportamento dello stimatore di massima verosimiglianza della varianza. Ricordiamo innanzitutto la densità delle variabili aleatorie $\Gamma(\alpha, \beta)$:

$$f_\Gamma(t) = \frac{1}{\Gamma(\alpha)\beta^{\alpha-1}} t^{\alpha-1} \exp(-\frac{t}{\beta}) \mathbb{I}_{[0,\infty)}(t) \text{ ,}$$

a cui corrisponde un valor medio pari a $\alpha\beta$ ed una varianza pari a $\alpha\beta^2$. Ricordiamo che la variabile chi-quadro con p gradi di libertà corrisponde ad una Gamma di parametri $\alpha = \frac{p}{2}$, $\beta = 2$, e quindi con valor medio p e varianza $2p$. La proprietà fondamentale dellle chi-quadro e che esse sono uguali in distribuzioni alla somma di p Gaussiane standard indipendneti elevate al quadrato:

$$\chi_p^2 \stackrel{d}{=} Z_1^2 + ... + Z_p^2 \text{ , } Z_i \stackrel{d}{=} N(0,1) \text{ .}$$

Poniamo ora senza perdita di generalita $\sigma^2 = 1$, e notiamo che

$$\widehat{\varepsilon}^T\widehat{\varepsilon} = (M_X\varepsilon)^TM_X\varepsilon = \varepsilon^TM_X^2\varepsilon = \varepsilon^TM_X\varepsilon \text{ .}$$

Ricordiamo ora la diagonalizzazione $M_X = Q\widetilde{\Lambda}Q^T$, dove $\widetilde{\Lambda}$ ha $n-k$ elementi pari ad uno sulla diagonale principale ed i restanti tutti nulli. Otteniamo quindi

$$\begin{aligned}
\varepsilon^T M_X \varepsilon &= \varepsilon^T Q\widetilde{\Lambda}Q^T\varepsilon \\
&= u^T\widetilde{\Lambda}u \\
&= \sum_{i=1}^{n}\widetilde{\lambda}_i u_i^2 \\
&= \sum_{i=1}^{n-k}\widetilde{\lambda}_i u_i^2 \text{ , con } u := Q^T\varepsilon \stackrel{d}{=} N(0, I_n) \ ,
\end{aligned}$$

perchè la legge Gaussiana standard è invariante per rotazioni. L'ultimo termine ha una legge chi-quadro per quanto scritto sopra, il che completa la dimostrazione.

Example 134 *Si noti che anche il problema della stima del valor medio per variabili indipendenti $Y_i \sim N(\mu, \sigma^2)$ può essere riletta in termini di regressione. Possiamo infatti scrivere*

$$Y = \begin{pmatrix} 1 \\ 1 \\ ... \\ ... \\ 1 \end{pmatrix} \mu + \varepsilon \ , \ \varepsilon = Y - \begin{pmatrix} 1 \\ 1 \\ ... \\ ... \\ 1 \end{pmatrix} \mu \ ;$$

lo stimatore diventa

$$\widehat{\mu} = (e^T e)^{-1} e^T Y = (\sum_{i=1}^{n} 1)^{-1} \times \sum_{i=1}^{n}(1 \cdot Y_i) = \overline{Y}_n \ ,$$

con

$$\widehat{\mu} \sim N(\mu, \sigma^2(e^T e)^{-1}) = N(\mu, \frac{\sigma^2}{n}) \ .$$

Nel caso $k = 2$ (con costante) il modello diviene

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

ed un poco di algebra mostra che

$$\begin{aligned}
\widehat{\beta}_1 &= \overline{Y}_n - \widehat{\beta}_2\overline{X}_n \ , \\
\widehat{\beta}_2 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} \ ,
\end{aligned}$$

con

$$\widehat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}) \ .$$

16.2 Lo stimatore GLS (Generalized Least Squares)

Possiamo ora generalizzare il modello che abbiamo studiato sinora, immaginando che i residui ε abbiano una struttura di dipendenza molto più complessa di variabili indipendenti. In particolare, ipotizziamo che $E[\varepsilon\varepsilon^T] = \Omega$ con matrice positiva definita di rango (pieno) n. La funzione di verosimiglianza prende la forma

$$L(\beta; Y, X) = \frac{1}{(2\pi)^n} \frac{1}{\sqrt{\det(\Omega)}} \exp\left\{ -\frac{1}{2}(Y - X\beta)^T \Omega^{-1/2} (Y - X\beta)^T \right\} .$$

Potremmo procedere come nel Caso ordinario, ma c'è una strategia più semplice. La matrice Ω si può diagonalizzare come $\Omega = Q\Lambda_\Omega Q^T$, per qualche matrice ortonormale Q che non corrisponde a quelle che abbiamo introdotto prima. Possiamo definire quindi $\Omega^{-1/2} = Q\Lambda_\Omega^{-1/2} Q^T$, con l'ovvia proprietà che

$$\begin{aligned} \Omega^{-1/2}\Omega\Omega^{-1/2} &= Q\Lambda_\Omega^{-1/2} Q^T Q\Lambda_\Omega Q^T Q\Lambda_\Omega^{-1/2} Q^T \\ &= Q\Lambda_\Omega^{-1/2}\Lambda_\Omega\Lambda_\Omega^{-1/2} Q^T \\ &= QQ^T \\ &= I_n . \end{aligned}$$

Possiamo dunque definire il vettore $\widetilde{\varepsilon} := \Omega^{-1/2}\varepsilon$, che (si verifica facilmente) ha matrice di varianza/covarianza pari all'identità. Quindi

$$Y = X\beta + \varepsilon \mapsto \Omega^{-1/2} Y = \Omega^{-1/2} X\beta + \Omega^{-1/2}\varepsilon \mapsto \widetilde{Y} = \widetilde{X}\beta + \widetilde{\varepsilon} ,$$

con

$$\widetilde{Y} := \Omega^{-1/2} Y , \widetilde{X} := \Omega^{-1/2} X , \widetilde{\varepsilon} := \Omega^{-1/2}\varepsilon .$$

Lo stimatore diventa quindi

$$\widetilde{\beta}_{GLS} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}\widetilde{Y} = (X^T\Omega^{-1}X)^{-1} X\Omega^{-1} Y ,$$

che è uno stimatore non distorto con legge Gaussiana e matrice di varianza e covarianza

$$E[(\widetilde{\beta}_{GLS} - \beta)(\widetilde{\beta}_{GLS} - \beta)^T] = \Omega^{-1}.$$

Remark 135 *Lo stimatore GLS può essere visto come i coefficienti di proiezione su un sottospazio generato dalle colonen di X quando la proiezione sia effettuata con una metrica Riemanniana indotta dalla matrice positive definita Ω^{-1}.*

Exercise 136 *Calcolare la forma dello stimatore GLS quando la matrice Ω sia nulla al di fuori della diagonale, nel caso $k = 1$.*

Exercise 137 *Calcolare la matrice di varianza e covarianza dello stimatore OLS (non GLS) quando la matrice di varianza di ε sia Ω non diagonale (cioè quando lo stimatore OLS sia applicato assumendo ipotesi in realtà non soddisfatte).*