Capitolo 5

La distribuzione normale multivariata.

5.1 Richiami sulla normale bivariata

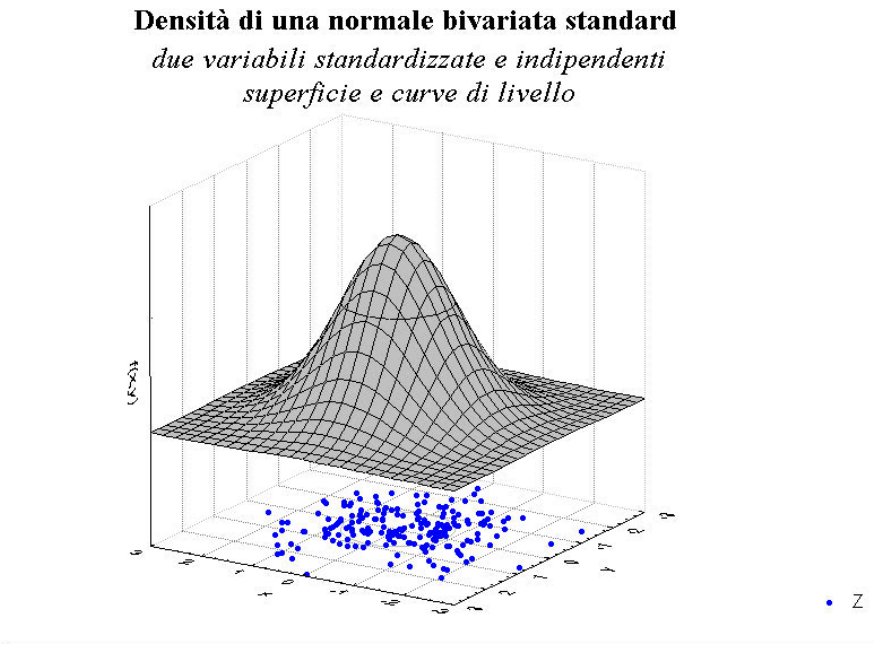

Figura 5.1: densità di normali bivariate 1

vai a indice figure

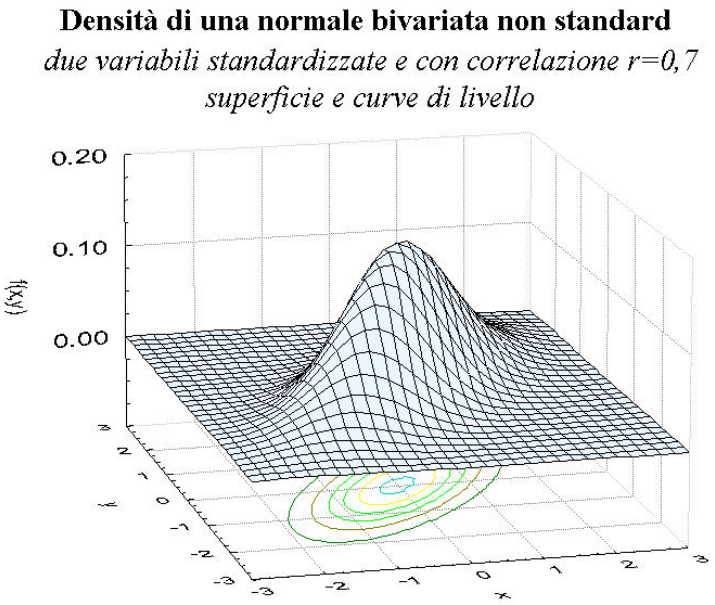

Figura 5.2: densità di normali bivariate 2

vai a indice figure

images/multinormani1.gifDensità della normale bivariata al variare di ρ images/multinormani2.gifDensità della normale bivariata al variare di ρ images/multinormani3.gifDensità della normale bivariata al variare di ρ images/multinormani4.gifDensità della normale bivariata al variare di ρ

ARGOMENTO DA COMPLETARE

La densità di una variabile aleatoria $\mathbf{X} = (X_1, X_2)$ con distribuzione normale bivariata è data da:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \tag{5.1}$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\left(-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma 2}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

Ho riportato la coppia di variabili (X_1, X_2), (e non (X,Y) perchè questo renderà più semplice poi il passaggio alla normale multivariata;

tuttavia ho mantenuto la parametrizzazione con la correlazione ρ piuttosto che con la covarianza σ_{12}.

I primi due momenti identificano completamente la distribuzione, in quanto si ha:

$$\mathrm{E}[X_1] = \mu_1 \quad \mathrm{E}[X_2] = \mu_2$$
$$\mathrm{V}[X_1] = \sigma_1^2 \quad \mathrm{V}[X_2] = \sigma_2^2 \quad CovX_1, X_2 = \rho\sigma_1\sigma_2$$

in termini matriciali:

$$\mathrm{E}[\mathbf{X}] = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \mathrm{V}[\mathbf{X}] = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

per cui la correlazione lineare è data da ρ, infatti:

$$corr(X_1, X_2) = \frac{covX_1, X_2}{\sigma_1\sigma_2} = \rho$$

Si ha l'importantissima proprietà:

Correlazione $\iff$ indipendenza nella normale bivariata

In una normale bivariata:

$$X_1 \perp X_2 \iff \rho = 0$$

ossia l'assenza di correlazione lineare implica l'indipendenza, per due variabili con distribuzione normale bivariata.

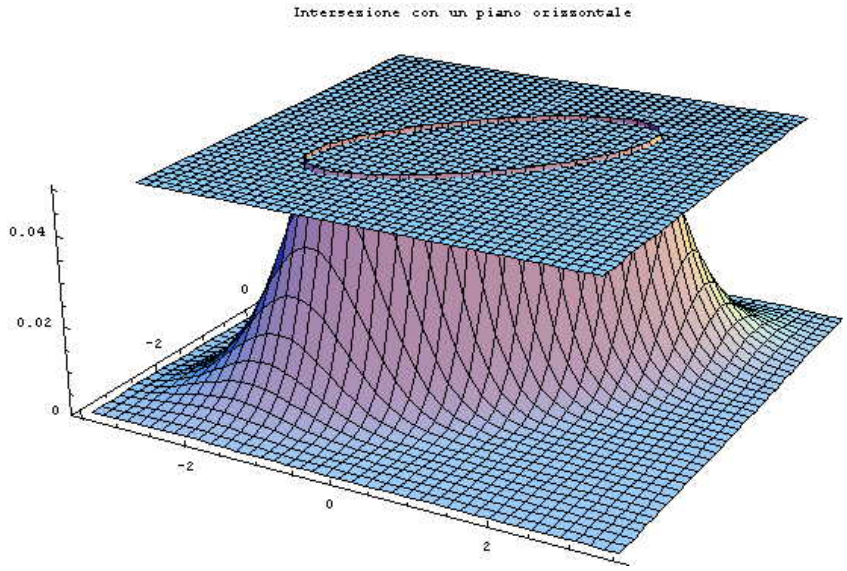

Figura 5.3: intersezioni con la normale bivariata

vai a indice figure

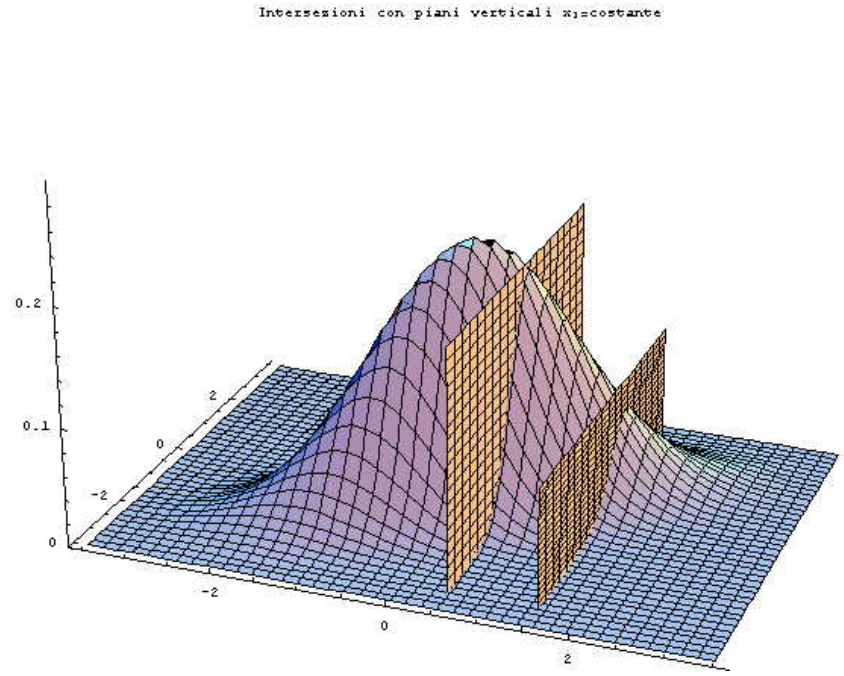

Figura 5.4: intersezioni con la normale bivariata

vai a indice figure

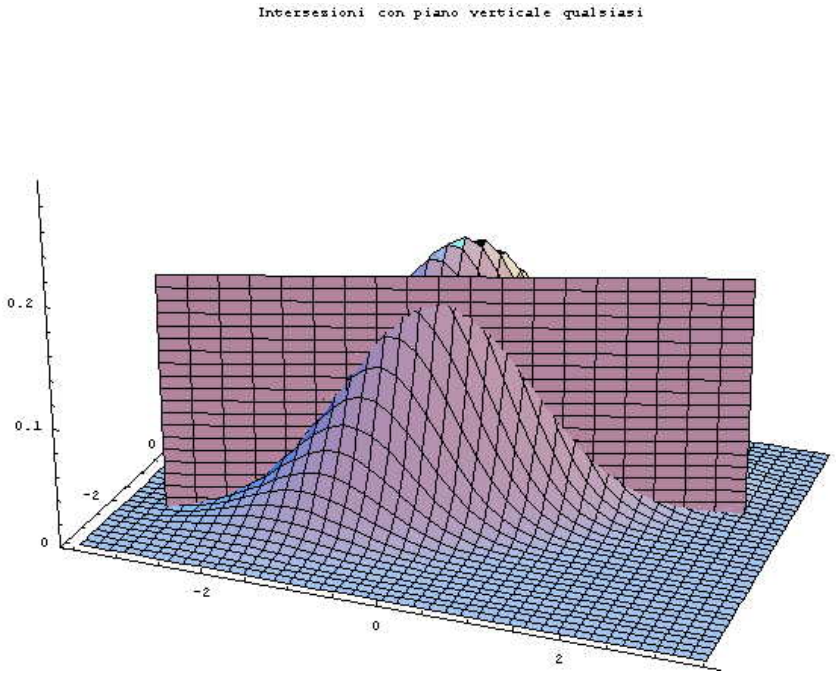

Figura 5.5: intersezioni con la normale bivariata

vai a indice figure

5.2 La normale multivariata

La distribuzione normale multipla può essere introdotta in numerosi modi, ed espressa con diverse caratterizzazioni.

Qui viene introdotta come la distribuzione congiunta di combinazioni lineari di variabili normali.

5.3 Distribuzione di variabili normali indipendenti

Sia $\mathbf{X}$ un vettore di variabili casuali a p componenti indipendenti:

$$\mathbf{X} = \{X_1, X_2, \ldots, X_i, \ldots, X_p\}^{\mathrm{T}}$$

ciascuna distribuita secondo una normale standardizzata.

La densità di tale distribuzione, data l'indipendenza, è data da:

...

Densità congiunta di p variabili normali standardizzate e indipendenti.

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{p} f(x_i) =$$

$$= (2\pi)^{-p/2} \exp[-\sum_{i=1}^{p} x_i^2/2] =$$

$$= (2\pi)^{-p/2} \exp[-\mathbf{x}^{\mathrm{T}}\mathbf{x}/2]$$

La funzione caratteristica è:

$$\phi_{\mathbf{X}}(t) = \exp\left(-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t}\right)$$

Ovviamente i primi due momenti di $\mathbf{X}$, per le ipotesi fatte, sono:

$$\mathrm{E}[\mathbf{X}] = 0_p,$$

$$V(\mathbf{X}) = \mathbf{I}_p$$

E' noto, ed è facile comunque vederlo attraverso la funzione caratteristica, che una singola combinazione lineare Z del vettore aleatorio $\mathbf{X}$ si distribuisce secondo una normale univariata, con media e varianza ricavabili dalle relazioni già viste per i momenti di combinazioni lineari di vettori aleatori qualsiasi.

Infatti se: $Z = \mathbf{b}^T\mathbf{X} + c$, allora i primi due momenti di Z sono dati da:

$$E(Z) = \mathbf{b}^T E(\mathbf{X}) + c = c$$

$$V(Z) = \mathbf{b}^T\mathbf{\Sigma}(\mathbf{X})\mathbf{b} = \mathbf{b}^T\mathbf{b} = b_1^2 + b_2^2 + \ldots + b_i^2 + \ldots + b_p^2$$

e si ha anche:

$$Z \sim N(E(Z), V(Z)).$$

`funzione caratteristica della combinazione lineare`

...

Adesso occorre però studiare la distribuzione congiunta di p combinazioni lineari di variabili normali indipendenti.

5.4 Densità della distribuzione congiunta di p combinazioni lineari di p variabili normali indipendenti

Consideriamo allora il vettore aleatorio $\mathbf{Y}$, trasformazione lineare del vettore aleatorio $\mathbf{X}$, definito dalla relazione:

$$\mathbf{Y} = \mathbf{A}^T\mathbf{X} + \boldsymbol{\mu}$$

essendo:
$\mathbf{A}$ una matrice quadrata di dimensione p e rango pieno;
$\boldsymbol{\mu}$ un vettore di p elementi;

Per ora abbiamo posto la condizione che $\mathbf{A}$ sia a rango pieno p, sarà poi possibile generalizzare a trasformazioni $\mathbf{X} \Rightarrow \mathbf{Y}$ anche singolari, ossia a rango non pieno;

(rispetto alla notazione ordinaria si è indicata la trasformazione mediante una matrice trasposta, perché di solito si dà un significato geometrico alle colonne di $\mathbf{A}$, ed ogni componente di $\mathbf{Y}$ corrisponde ad una colonna di $\mathbf{A}$; inoltre è irrilevante ai fini del risultato partire da p variabili standardizzate $\mathbf{X}_i$ oppure a varianza qualsiasi: l'importante è che siano indipendenti)

Per le proprietà sui momenti di trasformate lineari di v.a. i momenti di $\mathbf{Y}$ sono dati da:

$$E(\mathbf{Y}) = \mathbf{A}^T E(\mathbf{X}) + \boldsymbol{\mu} = \boldsymbol{\mu}$$
$$V(\mathbf{Y}) = \mathbf{A}^T V(\mathbf{X})\mathbf{A} = \mathbf{A}^T\mathbf{A}$$

Per ricavare la densità di $\mathbf{Y}$ è conveniente esplicitare la trasformazione inversa.

Dalla relazione diretta:

$$\mathbf{Y} = \mathbf{A}^T\mathbf{X} + \boldsymbol{\mu},$$

si ottiene subito la relazione inversa:

$$\mathbf{X} = \mathbf{B}^T[\mathbf{Y} - \boldsymbol{\mu}], \text{avendo posto: } \mathbf{B} = \mathbf{A}^{-1}$$

Pertanto, applicando la regola per le densità di trasformazioni di variabili aleatorie, la densità di $\mathbf{Y}$ è data da:

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{B}^T[\mathbf{y} - \boldsymbol{\mu}])J =$$
$$= J(2\pi)^{-p/2} \exp\left(-\frac{1}{2}[\mathbf{y} - \boldsymbol{\mu}]^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}[\mathbf{y} - \boldsymbol{\mu}]\right)$$

essendo J lo Jacobiano della trasformazione da $\mathbf{Y}$ a $\mathbf{X}$, ossia la matrice $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$, che ovviamente è dato da $J = \mathrm{mod}|\mathbf{B}|$, per cui si ha:

$$f_{\mathbf{y}}(\mathbf{y}) = \mathrm{mod}|B|(2\pi)^{-p/2} \exp\left(-\frac{1}{2}[\mathbf{y} - \boldsymbol{\mu}]^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}[\mathbf{y} - \boldsymbol{\mu}]\right) \quad (5.2)$$

Questa è la densità richiesta, tuttavia è meglio parametrizzare questa distribuzione in modo che sia esplicito, se possibile, il legame con i momenti di $\mathbf{Y}$.

Indichiamo con $\boldsymbol{\Sigma}$ la matrice di varianza e covarianza di $\mathbf{Y}$, ossia $V(\mathbf{Y})$, che abbiamo già visto essere uguale a $\mathbf{A}^T\mathbf{A}$.

Se vogliamo esprimere $V(\mathbf{X})$ in funzione di $V(\mathbf{Y})$ si ha:

$$V(\mathbf{X}) = \mathbf{B}^T V(\mathbf{Y})\mathbf{B} = \mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B}.$$

Per ipotesi abbiamo però che $V(\mathbf{X}) = \mathbf{I}_p$, per cui:

$$\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B} = \mathbf{I}_p$$

chiarire

e quindi la matrice $\mathbf{B}$ diagonalizza $\boldsymbol{\Sigma}$, per cui ha colonne proporzionali agli autovettori di $\boldsymbol{\Sigma}$ **citazione**

divisi per le radici dei rispettivi autovalori (si rivedano eventualmente i teoremi relativi alla diagonalizzazione di matrici, agli autovalori ed agli autovettori).

Inoltre, prendendo in esame la relazione $\mathbf{B}^T\mathbf{\Sigma}\mathbf{B} = \mathbf{I}$, premoltiplicando ambo i membri per $\mathbf{B}$ e postmoltiplicando per $\mathbf{B}^T$, si ottiene:

$$\mathbf{B}\mathbf{B}^T\mathbf{\Sigma}\mathbf{B}\mathbf{B}^T = \mathbf{B}\mathbf{B}^T$$

Postmoltiplicando (o premoltiplicando) ora ambo i membri per $(\mathbf{B}\mathbf{B}^T)^{-1}$ (che esiste sempre essendo $\mathbf{B}$, e quindi anche $\mathbf{B}\mathbf{B}^T$, a rango pieno p) si ha:

$$\mathbf{B}\mathbf{B}^T\mathbf{\Sigma}\mathbf{B}\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1} = \mathbf{B}\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\text{e quindi:}$$

$$\mathbf{B}\mathbf{B}^T\mathbf{\Sigma} = \mathbf{I}.$$

Per cui in definitiva si ha:

$$\mathbf{B}\mathbf{B}^T = \mathbf{\Sigma}^{-1}$$

e quindi nella forma quadratica ad esponente nell'espressione (5.2) di $f_{\mathbf{Y}}(\mathbf{y})$ potremo sostituire $\mathbf{B}\mathbf{B}^T$ con $\mathbf{\Sigma}^{-1}$.

Per potere ottenere il determinante di $\mathbf{B}$ che compare in $f_{\mathbf{Y}}(\mathbf{y})$, basta applicare le note regole sui determinanti delle trasposte, dei prodotti e delle inverse, per vedere che:

$$\|\mathbf{B}\| = \|\mathbf{B}^T\| = \|\mathbf{B}\mathbf{B}^T\|^{\frac{1}{2}} = \|\mathbf{\Sigma}^{-1}\|^{\frac{1}{2}} = \|\mathbf{\Sigma}\|^{-\frac{1}{2}}$$

Inoltre essendo $\mathbf{\Sigma}$ definita positiva, il suo determinante è certamente positivo.

5.5 Densità della distribuzione normale multivariata

In conclusione, sostituendo nella densità di $\mathbf{y}$:

$$f_{\mathbf{Y}}(\mathbf{y}) = \|\mathbf{B}|(2\pi)^{-p/2}\exp[-\frac{1}{2}[\mathbf{y}-\boldsymbol{\mu}]^{\mathrm{T}}\mathbf{B}\mathbf{B}^{\mathrm{T}}[\mathbf{y}-\boldsymbol{\mu}]]$$

abbiamo:

...

Densità della distribuzione normale *non singolare* multivariata di parametri $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$:

$$f_{\mathbf{Y}}(\mathbf{y}) = \|\boldsymbol{\Sigma}\|^{-\frac{1}{2}}(2\pi)^{-p/2}\exp[-\frac{1}{2}[\mathbf{y}-\boldsymbol{\mu}]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}[\mathbf{y}-\boldsymbol{\mu}]]$$

o anche:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}}\exp\left\{-\frac{1}{2}[\mathbf{y}-\boldsymbol{\mu}]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}[\mathbf{y}-\boldsymbol{\mu}]\right\}$$

oppure

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|\,(2\pi)^{p}}}e^{-\frac{1}{2}[\mathbf{y}-\boldsymbol{\mu}]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}[\mathbf{y}-\boldsymbol{\mu}]}$$

I primi due momenti multivariati sono (come già visto prima senza alcun bisogno di effettuare integrazioni p-dimensionali):

$$\mathrm{E}\,[\mathbf{Y}] = \boldsymbol{\mu}$$

$$V(\mathbf{Y}) = \boldsymbol{\Sigma}$$

e la funzione caratteristica (applicando la regola per le trasformazioni lineari di variabili aleatorie) è data da:

$$\phi_{\mathbf{Y}}(\mathbf{t}) = \exp[i\mathbf{t}^{T}\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^{T}\boldsymbol{\Sigma}t]$$

Ricordo che i momenti possono essere eventualmente ricavati dalle opportune derivate di $\phi_{\mathbf{Y}}(\mathbf{t})$, valutate in $\mathbf{t} = \mathbf{0}$.

Dalle espressioni della densità riportate sopra, è evidente l'analogia con l'espressione della densità della distribuzione normale univariata.

Si vede quindi, in analogia al caso univariato, che la distribuzione normale multivariata dipende soltanto dai primi due momenti (multivariati) di $\mathbf{Y}$.

Inoltre è possibile far vedere, rifacendo a ritroso i passaggi pre-

cedenti, che qualsiasi vettore aleatorio $\mathbf{Y}$ la cui densità è data da:

$$f_{\mathbf{Y}}(\mathbf{y}) = \|\mathbf{C}|^{\frac{1}{2}}(2\pi)^{-p/2}\exp\left(-\frac{1}{2}\quad[\mathbf{y}-\boldsymbol{\mu}]\mathbf{C}\right) \tag{5.3}$$

(con $\mathbf{C}$ definita positiva di rango p) è distribuito secondo una normale multivariata di parametri $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma} = \mathbf{C}^{-1}$.

Esiste inoltre una trasformazione lineare di $\mathbf{Y}$ che conduce ad un vettore aleatorio $\mathbf{X}$ a componenti standardizzate e indipendenti:

$$\mathbf{X} = \mathbf{B}^T[\mathbf{Y}-\boldsymbol{\mu}], \text{in cui} \mathbf{B} \text{è tale che} : \mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B} = \mathbf{I}$$

5.5.1 Distribuzioni marginali e indipendenza

Inoltre è evidente che l'indipendenza fra tutte le componenti di $\mathbf{Y}$ si può avere solo quando la $f_{\mathbf{Y}}(\mathbf{y})$ è fattorizzabile nelle rispettive densità marginali, il che può avvenire se (*e solo se*) $\boldsymbol{\Sigma}$ è diagonale, ossia con covarianze nulle, e quindi correlazioni lineari semplici nulle, il che porta un'altra fondamentale proprietà della normale multivariata:

Un vettore aleatorio $\mathbf{Y}$ con distribuzione normale multivariata, è a componenti indipendenti se *(e solo se)* le correlazioni lineari fra le sue componenti prese a due a due sono nulle, ossia se la matrice di varianza e covarianza è diagonale.
Quindi, se due variabili sono congiuntamente normali, l'assenza di correlazione implica l'indipendenza.

La distribuzione marginale di un qualsiasi sottoinsieme di componenti di un vettore aleatorio distribuito secondo una normale multivariata è ancora distribuito secondo una normale multivariata con parametri uguali ai corrispondenti sottoinsiemi di $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$: il risultato si dimostra facilmente, ad esempio prendendo la funzione caratteristica.

Infatti se il vettore $\mathbf{Y}$ è suddiviso in due sottovettori $[\mathbf{Y}_1, \mathbf{Y}_2]$, corrispondentemente suddividiamo il vettore delle medie e la matrice di varianza e covarianza:

$$\boldsymbol{\mu} = \left(\frac{\boldsymbol{\mu}_1}{\boldsymbol{\mu}_2}\right)$$

$$\mathbf{\Sigma} = \left(\begin{array}{c|c} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \hline \mathbf{\Sigma}_{12}^T & \mathbf{\Sigma}_{22} \end{array} \right)$$

Posta ora, corrispondentemente alla partizione di $\mathbf{Y}$, una partizione $\mathbf{t} = \{\mathbf{t}_1, \mathbf{t}2\}$, come si sa la funzione caratteristica di $\mathbf{Y}_1$ si ottiene da quella di $\mathbf{Y}$ ponendo $\mathbf{t}_2 = \mathbf{0}$:

$$\phi_{\mathbf{Y}_1}(\mathbf{t}_1) = \phi_{\mathbf{Y}}(\mathbf{t}_1, \mathbf{0}) = \exp[i\mathbf{t}_1^{\mathrm{T}}\boldsymbol{\mu}_1 - \frac{1}{2}\mathbf{t}_1^{\mathrm{T}}\mathbf{\Sigma}_{11}\mathbf{t}_1]$$

che è la funzione caratteristica di una normale di parametri μ_1 e $\mathbf{\Sigma}_{11}$.

In particolare tutte le distribuzioni marginali delle singole componenti sono normali univariate.

Come corollario è facile vedere che $\mathbf{Y}_1$ e $\mathbf{Y}_2$ (vettori aleatori normali) sono indipendenti se e solo se $\mathbf{\Sigma}_{12} = \mathbf{0}$.

5.5.2 Distribuzione di combinazioni lineari di variabili normali qualsiasi.

Mediante la funzione caratteristica è possibile vedere ora che qualsiasi combinazione lineare di un vettore aleatorio distribuito secondo una qualsiasi normale multivariata si distribuisce ancora secondo una distribuzione normale multivariata:

Infatti dal momento che se $\mathbf{Y} = \mathbf{A}Z$, si ha:

$$\phi_{\mathbf{Y}}(\mathbf{t}) = \phi_{\mathbf{Z}}(\mathbf{A}^{\mathrm{T}}\mathbf{t}),$$

se $\mathbf{Z}(N_p(\boldsymbol{\mu}_{\mathbf{Z}}, \mathbf{\Sigma}_{\mathbf{Z}})$ allora:

$$\phi_{\mathbf{Z}}(\mathbf{t}) = \exp\left[i\mathbf{t}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{Z}} - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{t}\right]$$

e quindi :

$$\phi_{\mathbf{Y}}(\mathbf{t}) = \phi_{\mathbf{Z}}(\mathbf{A}^T t) = \exp\left[i\mathbf{A}^{\mathrm{T}}\boldsymbol{\mu}_{\mathbf{Z}} - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{A}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{A}^{\mathrm{T}}\mathbf{t}\right]$$

per cui è immediato vedere che questa è ancora la funzione caratteristica di una normale multivariata di parametri $\mathbf{A}\boldsymbol{\mu}_{\mathbf{Z}}$ e $\mathbf{A}\mathbf{\Sigma}_{\mathbf{Z}}\mathbf{A}^{\mathrm{T}}$.

5.5.3 Caratterizzazione della distribuzione normale multivariata.

Le proprietà viste prima sulla distribuzione congiunta di combinazioni lineari di variabili normali costituiscono addirittura una caratterizzazione della distribuzione normale multivariata.

Infatti si ricorda una importante proprietà che caratterizza la distribuzione normale multivariata (di cui non si fornisce la dimostrazione) (Mardia, 1970): **citazione**

...

$\mathbf{X}$, vettore aleatorio a p componenti, è distribuito secondo una normale multivariata se e solo se $\mathbf{b}^T\mathbf{X}$ è distribuito secondo una normale (univariata) per qualsiasi vettore $\mathbf{b}$ di p componenti.

E' appena il caso di dire che il calcolo delle probabilità integrali su domini rettangolari della normale multivariata è estremamente complesso, e comunque non riconducibile a trasformazioni semplici di integrali unidimensionali, se le variabili sono correlate.

`citare software`

Ancora va chiarito, sulla genesi della normale multivariata utilizzata in queste righe, che questa è una impostazione utile per ricavare la distribuzione di combinazioni lineari di variabili normali indipendenti: nell'analisi di fenomeni reali ovviamente non è quasi mai ragionevole pensare che delle variabili osservate correlate siano state effettivamente ottenute come combinazione di fattori o variabili non correlate, anche se ovviamente è possibile, come si vede nell'analisi delle componenti principali, operare una rotazione per ricavare variabili non correlte, che non necessariamente corrispondono però a variabili osservabili o dotate di significato

5.6 Assi principali degli ellissoidi di equiprobabilità

E' immediato vedere che le curve con densità $f(\mathbf{y})$ costante per la normale multivariata di parametri $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$ sono, in uno spazio p -dimensionale, degli ellissoidi di centro in $\boldsymbol{\mu}$, e di equazione:

$$\|\boldsymbol{\Sigma}\|^{-\frac{1}{2}}(2\pi)^{-p/2}\exp(-[\mathbf{y}-\boldsymbol{\mu}]^T\boldsymbol{\Sigma}^{-1}[\mathbf{y}-\boldsymbol{\mu}]/2) = k_0$$

e quindi:

$$[\mathbf{y}-\boldsymbol{\mu}]^T\boldsymbol{\Sigma}^{-1}[\mathbf{y}-\boldsymbol{\mu}] = k_1$$

essendo k_1 e k_2 due costanti positive qualsiasi.

- E' facile vedere quindi che al variare del livello costante k , cambia solo il volume dell'ellissoide, ma le proporzioni fra gli assi restano inalterate;
- le equazioni degli assi principali di tali ellissoidi sono date dagli autovettori di $\boldsymbol{\Sigma}$;
- i quadrati delle lunghezze degli assi principali di tali ellissoidi sono proporzionali agli autovalori di $\boldsymbol{\Sigma}$.
- Se $\boldsymbol{\Sigma}$ è diagonale, gli ellissoidi hanno assi paralleli agli assi coordinati e lunghezza proporzionale agli scarti quadratici medi delle singole componenti.
- Si può fare vedere che gli autovettori danno le direzioni degli assi principali impostando ancora un problema di massimo, ossia cercando i due punti sulla superficie dell'ellisse che hanno distanza massima.

...

Fissato un qualsiasi valore di k_1, esiste una corrispondenza biunivoca fra ellissoidi in $\Re_p$ e distribuzioni normali multivariate non singolari.

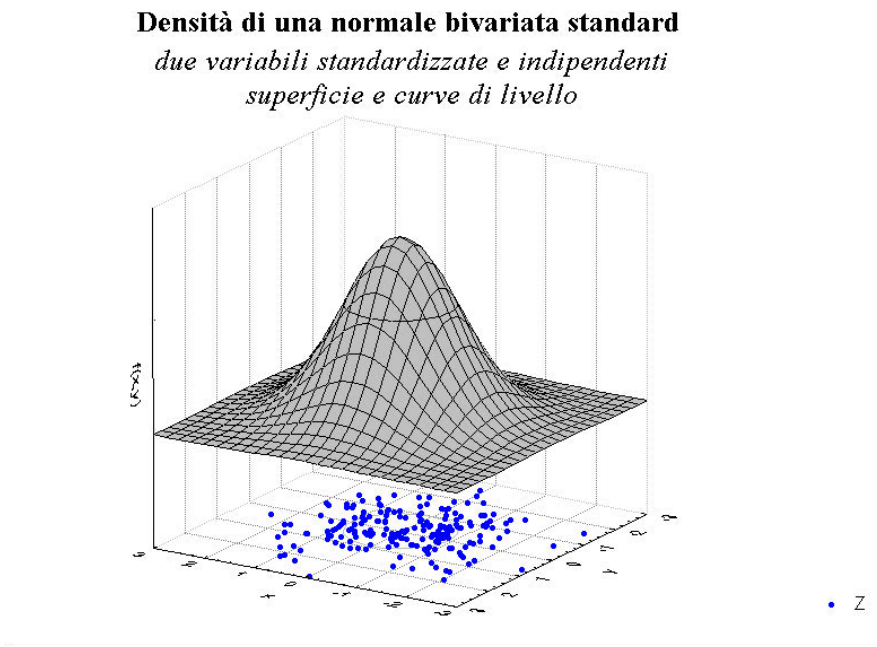

Figura 5.6: densità di normali bivariate 1

vai a indice figure

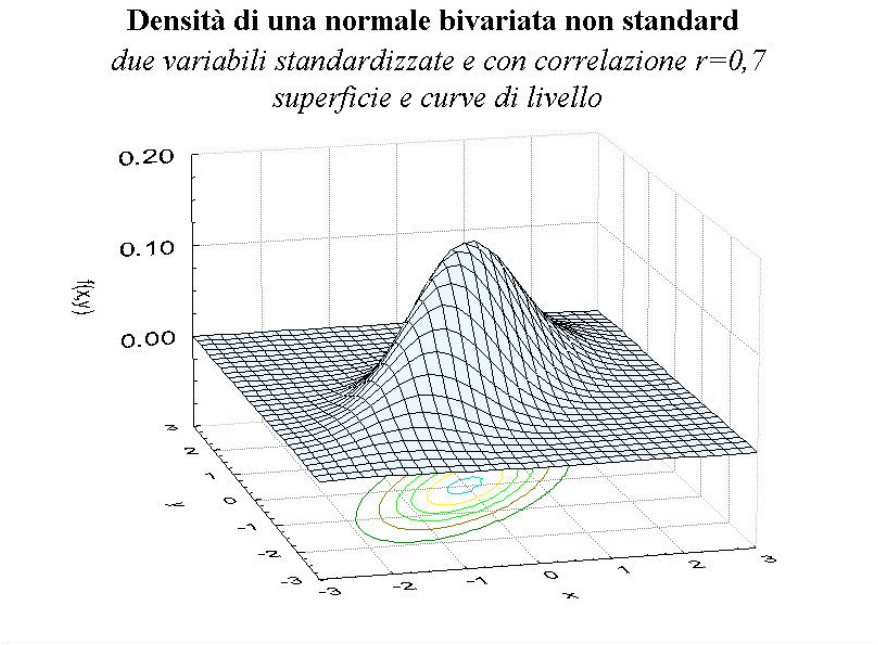

Figura 5.7: densità di normali bivariate 2

vai a indice figure

Esempi e grafici sulla normale trivariata

Normale trivariata a media nulla con Matrice di varianze e covarianze:

```
\begin{fig}

{parametric_ellissoide1_gr_3.gif}

Con autovalori:

parametric_ellissoide1_gr_5.gif

\end{fig}
```

Ellissoidi di equidensità
(sono due sezioni tridimensionali della densità (a 4D)
Scalato in modo tale che la probabilità che un punto risulti interno all'ellissoide risulti del 90%.

```
\begin{fig}
parametric_ellissoide1_gr_11.gif
\end{fig}
```

Scalato in modo tale che la probabilità che un punto risulti interno all'ellissoide è del 50%

```
\begin{fig}
parametric_ellissoide1_gr_14.gif
\end{fig}
```

Normale trivariata a media nulla con Matrice di varianze e covarianze:

```
\begin{fig}

parametric_ellissoide1_gr_17.gif

Con autovalori:

parametric_ellissoide1_gr_19.gif

\end{fig}
```

Ellissoidi di equidensità (sono due sezioni tridimensionali della densità (a 4D)

Scalato in modo tale che la probabilità che un punto risulti interno all'ellissoide sia del 90%.

```
\begin{fig}
parametric_ellissoide1_gr_25.gif
\end{fig}
```

Scalato in modo tale che la probabilità che un punto risulti interno all'ellissoide sia del 50%.

```
\begin{fig}
parametric_ellissoide1_gr_27.gif
\end{fig}

\begin{fig}
parametric_ellissoide1_gr_34.gif
\end{fig}
```

Dalla figura a fianco si vedono le caratteristiche della distribuzioni condizionate.

Normale trivariata a media nulla con Matrice di varianze e covarianze:

```
\begin{fig}

parametric_ellissoide1_gr_41.gif

Ellissoide di equidensità
(è una sezione tridimensionale della densità (a 4D)

Scalato in modo tale che la probabilità che un punto
 risulti interno all'ellissoide è del 50%
parametric_ellissoide1_gr_49.gif

\end{fig}
```

5.7 Distribuzione di forme quadratiche in variabili normali standardizzate e indipendenti.

In questa sezione affrontiamo il problema della distribuzione di particolari forme quadratiche in variabili normali, indipendenti e non: la finalità sarà chiara quando si studieranno le proprietà degli stimatori e dei test nei modelli lineari (modelli di regressione di analisi della varianza etc.); si tratta molto semplicemente di generalizzare alcuni risultati noti sulla v.c. χ^2: è ragionevole aspettarsi che forme quadratiche in variabili normali multivariate siano talora riconducibili a variabili χ^2.

Sia $\mathbf{X}$ un vettore di variabili casuali a p componenti indipendenti, ciascuna distribuita secondo una normale standardizzata, ossia

$$\mathbf{X}(N_p(0_p, \mathbf{I}_p).$$

E' noto che:

$$\sum_{i=1}^{p} \mathbf{X}_i^2 \sim \chi^2_{p,}(\text{oppure}\mathbf{X}^T\mathbf{X}(\chi^2_p).$$

In effetti questa è proprio la definizione di una variabile casuale di tipo chi-quadrato con p gradi di libertà, che risulta avere una

distribuzione gamma di parametro di forma $c = p/2$ e parametro di scala λ .

Più in generale ci si potrebbe chiedere se si può ricavare la distribuzione di una forma quadratica qualsiasi in variabili normali standardizzate, ossia

$$\mathbf{Q} = \mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X},$$

e per quali matrici $\mathbf{A}$ questa forma quadratica risulta ancora distribuita come una chi-quadrato.

E' facile vedere che la forma quadratica $\mathbf{Q} = \mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X}$
si distribuisce come $\sum_{i=1}^{p} \lambda_i \chi_1^2$,
ove i λ_i sono gli autovalori di $\mathbf{A}$;

$$\mathbf{Q} = \mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X} \sim \sum_{i=1}^{p} \lambda_i \chi_1^2$$

Questo risultato si ricava facilmente dalla decomposizione spettrale della matrice $\mathbf{A}$, in quanto si può scrivere:

$$\mathbf{A} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T, \quad \text{per cui:} \quad \mathbf{Q} = \mathbf{X}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{X},$$

e il vettore aleatorio $\mathbf{W} = \mathbf{\Gamma}^{\mathrm{T}}\mathbf{X}$ è ovviamente normale a componenti standardizzate e indipendenti, data l'ortogonalità di $\mathbf{\Gamma}$ (una rotazione ortogonale di una iper-sfera conduce sempre ad una iper-sfera!). Quindi segue facilmente in modo naturale il risultato scritto prima. Esprimendo in modo più formale si ha:

posto $\mathbf{W} = \mathbf{\Gamma}^T\mathbf{X}$, essendo le colonne di $\mathbf{\Gamma}$ gli autovettori (ortogonali: $\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{I}$) di $\mathbf{A}$, si ha per i momenti di $\mathbf{W}$:

$$E(\mathbf{W}) = \mathbf{\Gamma}^T E(\mathbf{X}) = 0$$
$$V(\mathbf{W}) = \mathbf{\Gamma}^{\mathrm{T}} V(\mathbf{X})\mathbf{\Gamma} = \mathbf{\Gamma}^{\mathrm{T}}\mathbf{I}_p\mathbf{\Gamma} = \mathbf{I}_p$$

Il vettore aleatorio $\mathbf{W}$ è dunque composto da p variabili normali, standardizzate e indipendenti.

Tornando ora alla forma quadratica $\mathbf{Q}$ si ha:

$$\mathbf{Q} = \mathbf{X}^T\mathbf{A}\mathbf{X} = \mathbf{X}^T\mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T\mathbf{X} = \mathbf{W}^T\mathbf{\Lambda}\mathbf{W} = \sum_{i=1}^{p} \lambda_i \mathbf{W}_i^2$$

Le componenti $\mathbf{W}_i^2$ sono chiaramente distribuite come delle chi-quadrato (indipendenti) con un grado di libertà.

Pertanto $\mathbf{Q}$ è distribuita come una combinazione lineare di p variabili casuali chi-quadrato indipendenti con un grado di libertà, con coefficienti dati dagli autovalori di A.

In ogni caso è possibile calcolare i momenti di $\mathbf{Q}$ in quanto combinazione lineare di v.c. χ_1^2 indipendenti:

$$E(\mathbf{Q}) = \sum_{i=1}^{p} \lambda_i E(\chi_1^2) = \sum_{i=1}^{p} \lambda_i$$

$$V(\mathbf{Q}) = \sum_{i=1}^{p} \lambda_i^2 V(\chi_1^2) = 2 \sum_{i=1}^{p} \lambda_i^2$$

Se (*e solo se*) gli autovalori di $\mathbf{A}$ sono tutti uguali a 0 o a 1, ossia se (e solo se) $\mathbf{A}$ è idempotente,
$\mathbf{Q} = \mathbf{X}^T\mathbf{A}\mathbf{X}$ si distribuisce come una variabile casuale χ_r^2
per la proprietà additiva delle v.c. χ^2 , essendo r il rango di $\mathbf{A}$,
ossia il numero degli autovalori λ_i uguali ad uno

Infatti si vede immediatamente che, se $\mathbf{A}$ è idempotente di rango r , si ha:

$$\lambda_1 = \lambda_2 = \ldots = \lambda_r = 1;$$

$$\lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_p = 0;$$

per cui:

$$\sum_{i=1}^{p} \lambda_i \chi_1^2 = \sum_{i=1}^{r} 1\,(\chi_1^2 + \sum_{i=r+1}^{p} 0\,(\chi_1^2 = \sum_{i=1}^{r} \chi_1^2 \sim \chi_r^2$$

Per dimostrare che l'idempotenza di $\mathbf{A}$ è condizione necessaria e sufficiente perhcè $\mathbf{Q}$ sia distribuita come una chi-quadrato (prima abbiano visto che l'idempotenza di $\mathbf{A}$ è condizione sufficiente), conviene ricorrere alla funzione caratteristica di $\mathbf{Q}$, che è data da:

$$\phi_{\mathbf{Q}}(t) = Eexp(it\mathbf{X}^T\mathbf{A}\mathbf{X}) = Eexp(it \sum_{i=1}^{p} \lambda_i \mathbf{X}_i^2) =$$

$$= \prod_{i=1}^{p} Eexp(it\lambda_i \mathbf{X}_i^2) = \prod_{i=1}^{p} (1 - 2it\lambda_i)^{-\frac{1}{2}}$$

(dato che $\mathbf{X}_i^2$ è distribuito come una chi-quadrato con un grado di libertà, l'ultimo passaggio deriva dalla funzione caratteristica della v.c. chi-quadrato).

Ancora si può osservare che $1-2it\lambda_i$ è un autovalore della matrice: $\mathbf{I} - 2it\mathbf{A}$ e quindi la produttoria di tali autovalori $(1 - 2it\lambda_i)$ è uguale al determinante della suddetta matrice:

$$\phi(t) = \prod_{i=1}^{p}(1 - 2it\lambda_i)^{-\frac{1}{2}} = |\mathbf{I} - 2it\mathbf{A}|^{-\frac{1}{2}}$$

Perché $\mathbf{Q}$ sia distribuita come una chi-quadrato, occorre che la sua funzione caratteristica $\phi_{\mathbf{Q}}(t)$ sia identicamente uguale a quella di una v.c. χ^2 per qualsiasi valore dell'argomento t .

È la funzione caratteristica di una v.c. χ^2 con ν gradi di libertà è data da:

$$\phi_{\chi 2}(t) = (1 - 2it)^{-\nu/2},$$

mentre per la funzione caratteristica di $\mathbf{Q}$ si è visto che:

$$\phi_{\mathbf{Q}}(t) = \prod_{i=1}^{p}(1 - 2it\lambda_i)^{-\frac{1}{2}}.$$

Per avere $\phi_{\chi 2}(t) = \phi_{\mathbf{Q}}(t)$ per qualsiasi t , occorre che i coefficienti λ_i siano o zero o uno, di modo che i corrispondenti termini della produttoria in $\phi_{\mathbf{Q}}(t)$ siano uguali ad uno (se $\lambda_i = 0$) oppure a $(1 - 2it)^{-\frac{1}{2}}$ (se $\lambda_i = 1$); se sono r (rango di $\mathbf{A}$) quelli uguali ad uno, si avrà in definitiva:

$$\phi_{\mathbf{Q}}(t) = (1 - 2it)^{-r/2},$$

che è la funzione caratteristica di una chi-quadrato con r gradi di libertà.

Esempio

Ad esempio si consideri la matrice seguente:

$$\mathbf{A} = \begin{pmatrix} 16/25 & 12/25 \\ 12/25 & 9/25 \end{pmatrix}$$

Tale matrice simmetrica risulta idempotente di rango 1, come è facile verificare effettuando il prodotto $\mathbf{AA}$, oppure verificando che $\lambda_1 = 1$ e $\lambda_2 = 0$.

Supponendo di avere un vettore aleatorio $\mathbf{X}$ costituito da due variabili casuali normali standardizzate e indipendenti, X_1 e X_2 la forma quadratica $\mathbf{Q} = \mathbf{X}^T\mathbf{A}\mathbf{X}$ risulta data da:

$$\mathbf{Q} = a_{11}X_1^2 + a_{22}X_2^2 + 2a_{12}X_1X_2 = (16X_1^2 + 9X_2^2 + 24X_1X_2)/25,$$

e infine:

$$\mathbf{Q} = [(4/5)X_1 + (3/5)X_2]^2$$

E' immediato verificare che $\mathbf{Q}$ si distribuisce secondo una chi-quadrato con un grado di libertà, senza bisogno di applicare il teorema generale sulla distribuzione delle forme quadratiche. Infatti la variabile:

$$Z = (4/5)X_1 + (3/5)X_2$$

è distribuita normalmente (in quanto combinazione lineare di variabili normali) con media zero e varianza unitaria.

Infatti:

$$E[Z] = (4/5)E[X_1] + (3/5)E[X_2] = 0$$

$$Var[Z] = (4/5)^2Var[X_1] + (3/5)^2Var[X_2] = 16/25 + 9/25 = 1$$

($Cov[X_1, X_2] = 0$ per l'indipendenza).

Quindi $\mathbf{Q}$ è uguale al quadrato di una normale standardizzata, e quindi segue una distribuzione chi-quadrato con un grado di libertà.

Forme quadratiche idempotenti: somma dei quadrati degli scarti dalla media.

Prendiamo ora in esame una forma quadratiche già nota, ossia la somma dei quadrati degli scarti dalla propria media aritmetica di n variabili casuali normali indipendenti X_i. Tipicamente le variabili saranno quelle corrispondenti ad un campione a n componenti i.i.d.(e quindi il vettore aleatorio è al solito $X = \{X_1, X_2, \ldots, X_i, \ldots, X_n\}^{\mathrm{T}}$).

Interessa dunque la distribuzione della quantità:

$$\mathbf{Q} = \sum_{i=1}^{n}(\mathbf{X}_i - M)^2$$

avendo indicato con M la variabile casuale media aritmetica delle n componenti X_i:

$$M = \sum_{i=1}^{n} X_i/n$$

che si può anche scrivere:

$$M = \frac{\mathbf{1_n}^{\mathrm{T}}\mathbf{X}}{n},$$

essendo $\mathbf{1_n}$ un vettore di n elementi uguali ad uno.

Allora la somma dei quadrati degli scarti si può scrivere in notazione vettoriale con semplici passaggi:

$$\mathbf{Q} = \sum_{i=1}^{n}(\mathbf{X}_i - M)^2 = [\mathbf{X} - \mathbf{1_n}M]^T[\mathbf{X} - \mathbf{1_n}M] =$$

$$= [\mathbf{X} - \frac{\mathbf{1_n}\mathbf{1_n}^{\mathrm{T}}\mathbf{X}}{n}]^{\mathrm{T}}[\mathbf{X} - \frac{\mathbf{1_n}\mathbf{1_n}^{\mathrm{T}}\mathbf{X}}{n}] =$$

$$\mathbf{X}^T[\mathbf{I} - \frac{\mathbf{1_n}\mathbf{1_n}^{\mathrm{T}}}{n}]^T[\mathbf{I} - \frac{\mathbf{1_n}\mathbf{1_n}^{\mathrm{T}}}{n}]\mathbf{X}$$

Posto ora $\mathbf{U} = \frac{\mathbf{1_n}\mathbf{1_n}^{\mathrm{T}}}{n}$, è facile vedere che $\mathbf{U}$ è idempotente e simmetrica di rango 1: è composta da $n \times n$ elementi tutti uguali a $\frac{1}{n}$; quindi sono idempotenti (ma di $rangon - 1$) anche $\mathbf{I} - \mathbf{U}$, e $[\mathbf{I} - \mathbf{U}]^T[\mathbf{I} - \mathbf{U}]$, per cui possiamo scrivere:

$$\mathbf{Q} = \sum_{i=1}^{n}(\mathbf{X}_i - M)^2 = \mathbf{X}^T[\mathbf{I} - \mathbf{U}]\mathbf{X}$$

e $\mathbf{Q}$ è distribuita secondo una χ^2_{n-1}.

Esempio numerico

Con $n = 5$ si supponga di avere le 5 osservazioni $x_i : 3, 5, 8, 9, 10$, con media aritmetica $M = 7$.

La somma dei quadrati degli scarti (osservati!) è data da:

$$\mathbf{Q} = \sum_{i=1}^{n}(x_i - M)^2 = 16 + 4 + 1 + 4 + 9 = 34.$$

E' facile vedere che la matrice $\mathbf{U}$ è data da:

$$\mathbf{U} = \begin{pmatrix} 0,2 & 0,2 & 0,2 & 0,2 & 0,2 \\ 0,2 & 0,2 & 0,2 & 0,2 & 0,2 \\ 0,2 & 0,2 & 0,2 & 0,2 & 0,2 \\ 0,2 & 0,2 & 0,2 & 0,2 & 0,2 \\ 0,2 & 0,2 & 0,2 & 0,2 & 0,2 \end{pmatrix}$$

Indicato quindi con x il vettore delle 5 osservazioni, si verifichi il risultato fornito dal prodotto $\mathbf{x}^{\mathrm{T}}[\mathbf{I} - \mathbf{U}]\mathbf{x}$:

$$\mathbf{x}^{\mathrm{T}}[\mathbf{I} - \mathbf{U}]\mathbf{x} =$$

$$= \begin{pmatrix} 3 & 5 & 8 & 9 & 10 \end{pmatrix} \begin{pmatrix} 0,8 & -0,2 & -0,2 & -0,2 & -0,2 \\ -0,2 & 0,8 & -0,2 & -0,2 & -0,2 \\ -0,2 & -0,2 & 0,8 & -0,2 & -0,2 \\ -0,2 & -0,2 & -0,2 & 0,8 & -0,2 \\ -0,2 & -0,2 & -0,2 & -0,2 & 0,8 \end{pmatrix} \begin{pmatrix} 3 \\ 5 \\ 8 \\ 9 \\ 10 \end{pmatrix} =$$

$$= 3^2 \times 0,8 + 5^2 \times 0,8 + \ldots + 10^2 \times 0,8 - 2 \times 0,2 \times 3 \times 5 - \ldots = 34$$

5.7.1 La distribuzione dell'esponente della distribuzione normale multivariata.

Sappiamo già che il doppio dell'esponente della distribuzione normale univariata, $\frac{(\mathbf{X}-\mathrm{E}[\mathbf{X}])^2}{V(\mathbf{X})}$, si distribuisce secondo una variabile casuale χ^2. Vediamo come si generalizza questo risultato nel caso normale multivariato.

Sia $\mathbf{Y}$ un vettore di variabili casuali a p componenti, distribuito secondo una normale multivariata qualsiasi, ossia

$$\mathbf{Y} \sim (N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Si può dimostrare che la variabile casuale corrispondente alla forma quadratica che figura al numeratore dell'esponente della funzione di densità, ossia:

$$\mathbf{Q} = (\mathbf{Y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu}),$$

segue una distribuzione chi-quadrato con p gradi di libertà.

Infatti il risultato si mostra facilmente ricorrendo ad una opportuna trasformazione lineare (già impiegata in questo capitolo)

$$\mathbf{X} = \mathbf{B}^T[\mathbf{Y} - \boldsymbol{\mu}],$$

in cui $\mathbf{B}$ è tale che:

$$\mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B} = \mathbf{I}, \text{e} \boldsymbol{\Sigma}^{-1} = \mathbf{B}\mathbf{B}^T.$$

e quindi:

$$V(\mathbf{X}) = \mathbf{B}^{\mathrm{T}}V(\mathbf{Y} - \boldsymbol{\mu})\mathbf{B} = \mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B} = \mathbf{I}$$

Pertanto:

$$\mathbf{Q} = (\mathbf{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) = (\mathbf{Y} - \boldsymbol{\mu})^T\mathbf{B}\mathbf{B}^T(\mathbf{Y} - \boldsymbol{\mu}) =$$

$$= [(\mathbf{Y} - \boldsymbol{\mu})^T\mathbf{B}][\mathbf{B}^T(\mathbf{Y} - \boldsymbol{\mu})] = \mathbf{X}^T\mathbf{X} \sim \chi^2_p$$

Per cui $\mathbf{Q}$ si distribuisce come la somma dei quadrati di p variabili normali standardizzate e indipendenti, ossia come una chi-quadrato con p gradi di libertà.

In definitiva:

---

se $\mathbf{Y} \sim (N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, allora

$$(\mathbf{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \sim \chi^2_p$$

Esempio numerico

$$\mathbf{Y} \sim (N_2(\mathbf{0}, \boldsymbol{\Sigma}),$$

$$\text{con} \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

$$\text{e matrice di correlazione:} R = \begin{pmatrix} 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix}$$

e quindi

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix};$$

pertanto la forma quadratica:

$$\mathbf{Q} = \mathbf{Y}^T \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \mathbf{Y} = \mathbf{y}_1^2 + 2\mathbf{y}_2^2 - 2\mathbf{y}_1^2\mathbf{y}_2^2 \sim \chi^2_2$$

segue una distribuzione chi-quadro con due gradi di libertà.

```
eventualmente dimostrarlo per via
diretta nell'esempio
```

5.7.2 Indipendenza di forme quadratiche e combinazioni lineari di variabili normali.

Sia $\mathbf{X}$ un vettore di variabili casuali a p componenti indipendenti, ciascuna distribuita secondo una normale standardizzata, ossia

$$\mathbf{X} \sim N_p(0_p, \mathbf{I}_p).$$

Valgono alcuni teoremi sull'indipendenza fra forme quadratiche in $\mathbf{X}$ e combinazioni lineari in $\mathbf{X}$, che si basano sulle proprietà dei vettori dei coefficienti che determinano le forme quadratiche e le combinazioni lineari.

```
eventualmente mettere in
forma di schema
```

Si abbiano due forme quadratiche in variabili normali indipendenti $\mathbf{X}$:

$$\mathbf{Q}_1 = \mathbf{X}^T\mathbf{A}_1\mathbf{X} e \mathbf{Q}_2 = \mathbf{X}^T\mathbf{A}_2\mathbf{X}$$

Le due forme quadratiche $\mathbf{Q}_1$ e $\mathbf{Q}_2$ sono indipendenti se e solo se $\mathbf{A}_1\mathbf{A}_2 = 0_{(pxp)}$

(essendo ovviamente $\mathbf{A}_1$ e $\mathbf{A}_2$ matrici quadrate simmetriche, ed essendo $0_{(p\times p)}$ una matrice quadrata composta di zeri);

Si abbia la forma quadratica

$$\mathbf{Q} = \mathbf{X}^T\mathbf{A}^T\mathbf{X},$$

e la combinazione lineare $Z = \mathbf{b}^T\mathbf{X}$

La forma quadratica $\mathbf{Q}$ e la combinazione lineare Z sono indipendenti

se e solo se $\mathbf{Ab} = 0_p$

(essendo $\mathbf{b}$ un vettore di p elementi e 0_p il vettore nullo di p componenti)

5.7.3 Teorema di Cochran:

Supponiamo di avere una somma di quadrati di p variabili normali standardizzate e indipendenti, ossia:

$$\mathbf{Q} = \mathbf{X}^T\mathbf{X}$$

o, più in generale, una forma quadratica

$$\mathbf{Q} = \mathbf{X}^T\mathbf{A}\mathbf{X},$$

con $\mathbf{A}$ idempotente di rango p . In questo caso il numero delle componenti di $\mathbf{X}$ potrà essere in generale maggiore di p; il punto essenziale è che $\mathbf{Q}$ abbia una distribuzione chi-quadrato con p gradi di libertà.

Supponiamo di saper scomporre algebricamente $\mathbf{Q}$ nella somma di k forme quadratiche:

$$\mathbf{Q} = \mathbf{X}^T\mathbf{X} = \sum_{i=1}^{k} \mathbf{X}^T\mathbf{A}_i\mathbf{X} = \sum_{i=1}^{k} \mathbf{Q}_i,$$

avendo posto : $\mathbf{Q}_i = \mathbf{X}^T\mathbf{A}_i\mathbf{X}$, ed essendo per ipotesi:

$$\mathbf{Q}(\chi^2_p$$

Il teorema di Cochran stabilisce delle relazioni di importanza fondamentale in merito alle caratteristiche delle distribuzioni delle singole componenti $\mathbf{Q}_i$.

...

TEOREMA DI COCHRAN Una qualsiasi delle seguenti tre condizioni implica le altre due:

1. la somma dei gradi di libertà delle forme quadratiche deve eguaglire p:

$$\sum_{i=1}^{k} \rho(\mathbf{A}_i) = p = \rho(\mathbf{A})$$

(in generale la somma dei ranghi delle singole componenti deve eguagliare il rango di $\mathbf{A}$)

2. tutte le k forme quadratiche $\mathbf{Q}_i = \mathbf{X}^T\mathbf{A}_i\mathbf{X}$ hanno una distribuzione χ^2

   che corrisponde a :

   tutte le $\mathbf{A}_i$ devono essere idempotenti;

3. tutte le k forme quadratiche $\mathbf{Q}_i = \mathbf{X}^T\mathbf{A}_i\mathbf{X}$ sono a due a due indipendenti,

   che corrisponde a: $\mathbf{A}_i\mathbf{A}_j = 0$ per qualsiasi *coppiai* $\neq j$.

...

L'importanza di tale teorema nell'ambito della teoria normale sui modelli lineari è cruciale; in generale a ciascuna delle k componenti si farà corrispondere una particolare fonte di variabilità o un gruppo di parametri.

Ai fini pratici se per esempio se si vuole applicare ad una particolare scomposizione la proprietà 2, per poi dedurne la 1 e la 3, non è necessario esplicitare le singole matrici $\mathbf{A}_i$, ma è sufficiente sapere che si è scomposta $\mathbf{Q}$ in forme quadratiche nelle variabili aleatorie $\mathbf{X}_i$.

Sezione avanzata

In effetti esiste una formulazione ancora più generale del teorema, che prende in considerazione distribuzioni χ^2 non centrali, ossia forme quadratiche in variabili normali con speranza matematica diversa da zero, utile per la generalizzazione alla distribuzione di determinate quantità test non solo sotto H_0 ma anche sotto H_1. Per non appesantire questi appunti non riporto questa generalizzazione: ne farò cenno più avanti soltanto quando sarà necessario.

Esempio. Come esempio si rifletta sulla nota scomposizione per la somma dei quadrati di n variabili normali standardizzate indipendenti:

$$\sum_{i=1}^{n} \mathbf{X}_i^2 = \sum_{i=1}^{n} (\mathbf{X}_i - M)^2 + nM^2$$

Per applicare il teorema di Cochran è sufficiente far vedere che i due addendi sulla destra sono forme quadratiche in variabili normali di rango $n-1$ e 1: è immediato verificarlo senza bisogno di esplicitare le matrici, perché $\sum_{i=1}^{n}(\mathbf{X}_i - M)^2$ è palesemente una forma quadratica con un vincolo lineare ($\sum_{i=1}^{n}(\mathbf{X}_i - M) = 0$), mentre M^2

ha ovviamente un solo grado di libertà, quindi i due termini sono indipendenti e distribuiti come delle v.c. χ^2 con i rispettivi gradi di libertà.

5.8 Distribuzioni condizionate nella normale multivariata

Una proprietà fondamentale della normale, che oltretutto la caratterizza, riguarda le distribuzioni di un gruppo di componenti condizionatamente ai valori di un altro gruppo di componenti.

Questo argomento viene trattato adesso, senza limitarci ad esporre i risultati fondamentali, ma anzi entrando con un certo dettaglio, per tre ordini di ragioni:

1. La peculiarità delle caratteristiche delle distribuzioni condizionate nella normale multivariata, che ne rappresenta un aspetto fondamentale;
2. La possibilità di dare un significato statistico autonomo agli elementi dell'inversa della matrice di correlazione di una variabile multipla normale;
3. Come premessa indispensabile ai modelli lineari che tratteremo ampiamente in questo corso;

Come si vedrà nelle pagine successive, la distribuzione di un gruppo di variabili $\mathbf{Y_A}$ condizionata ad un particolare valore $\mathbf{y_B}$ assunto da un altro gruppo di $\mathbf{Y_B}$ è:

1. ancora normale ed inoltre:
2. La funzione di regressione di una componente $\mathbf{y}_A$ rispetto alle altre componenti è lineare
3. La distribuzione ha una matrice di varianze e covarianze che non dipende dai valori della componente condizionante (omoscedasticità).

I risultati esposti in queste pagine generalizzano le proprietà note per distribuzioni normali bivariate, in cui le due funzioni di regressione di ciascuna delle due variabili rispetto all'altra sono lineari, ed inoltre le distribuzioni condizionate sono normali e omoschedastiche.

In effetti ci porremo il problema nella forma più generale della distribuzione di un gruppo di variabili normali condizionatamente

ad un altro gruppo di variabili normali, nota la loro distribuzione congiunta.

Significato degli elementi dell'inversa della matrice di varianza e covarianza .

Sarà anche possibile dare un significato agli elementi dell'inversa di $\mathbf{\Sigma}$ in termini di distribuzioni condizionate.

Infatti si dimostrerà che se $\mathbf{C} = \mathbf{\Sigma}^{-1}$, allora:

teorema 5.8.1 *In una normale multivariata, $c_{ij} = 0$ è condizione necessaria e sufficiente perché le variabili $\mathbf{Y}_i$ e $\mathbf{Y}_j$ siano indipendenti condizionatamente alle altre $p-2$ variabili.*

5.8.1 Distribuzione condizionata nel caso generale di un gruppo di componenti rispetto ad un altro gruppo di componenti.

Nella versione breve del corso studiare solo i risultati finali

---

Supponiamo di avere un vettore $\mathbf{Y}$ di p componenti, con distribuzione normale multivariata, suddiviso nel caso più generale in due sottovettori $[\mathbf{Y_A}, \mathbf{Y_B}]$, con corrispondente suddivisione del vettore delle medie e della matrice di varianze e covarianze:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_A \\ \mathbf{Y}_B \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}$$

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{AA} & \mathbf{\Sigma}_{AB} \\ \mathbf{\Sigma}_{AB}^{\mathrm{T}} & \mathbf{\Sigma}_{BB} \end{pmatrix}$$

I due insiemi di indici A e B costuituiscono una partizione dell'insieme di indici $I = 1, 2, \ldots, p$ così che:

$$A \cup B = I \qquad A \cap B = \emptyset \qquad A \neq \emptyset B \neq 0\emptyset$$

per il resto A e B sono costituiti da sottoinsiemi di indici qualsiasi (con la restrizione che esistano le inverse delle matrici di varianze e covarianze che si richiederanno nel seguito).

In effetti i casi più rilevanti, che tratteremo specificatamente, sono quelli in cui $A = i$, per lo studio della distribuzione di una variabile condizionatamente alle altre e $A = i, j$, per lo studio della distribuzione condizionata di due variabili, in particolare per lo studio della dipendenza condizionata.

Ci chiediamo qual è la funzione di regressione di $\mathbf{Y}_A$ su $\mathbf{Y}_B$, ossia la speranza matematica di $\mathbf{Y}_A$ condizionata ad un particolare valore $\mathbf{y}_B$ di $\mathbf{Y}_B$:

$$E[\mathbf{Y}_A \| \mathbf{Y}_B = \mathbf{y}_B] = ??$$

In generale ci chiediamo direttamente qual è la distribuzione di $\mathbf{Y}_A$ condizionata ad un particolare valore $\mathbf{y}_B$ di $\mathbf{Y}_B$.

Per trovare la funzione di regressione nel caso generale, ricaviamo prima la densità della distribuzione di $\mathbf{Y}_A$ condizionata ad un particolare valore $\mathbf{y}_B$ assunto da $\mathbf{Y_B}$.

Per comodità lavoriamo con variabili $\mathbf{X}_A$, $\mathbf{X}_B$ con speranze matematiche nulle, ponendo:

$$\mathbf{X}_A = \mathbf{Y}_A - \boldsymbol{\mu}_A$$
$$\mathbf{X}_B = \mathbf{Y}_B - \boldsymbol{\mu}_B$$

Ovviamente la matrice di varianze e covarianze di $\mathbf{X}$ è uguale a quella di $\mathbf{Y}$:

$$V(\mathbf{X}) = V(\mathbf{Y})$$

E' opportuno richiamare le formule per la semplificazione degli elementi dell'inversa della matrice partizionata delle varianze e covarianze di $\mathbf{y}$:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{AA.B}^{-1} & -\boldsymbol{\Sigma}_{AA.B}^{-1}\boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1} \\ -\boldsymbol{\Sigma}_{AA.B}^{-1}\boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1} & \boldsymbol{\Sigma}_{BB}^{-1}[\boldsymbol{\Sigma}_{AB}^{T}\boldsymbol{\Sigma}_{AA.B}^{-1}\boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1} + \mathbf{I}] \end{pmatrix}$$

avendo posto:

$$\boldsymbol{\Sigma}_{AA.B} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{AB}^{T}.$$

Indichiamo con $\boldsymbol{\Sigma}^{IJ}$ il blocco corrispondente al posto di $\boldsymbol{\Sigma}_{IJ} (I = A, B; J = A, B)$ nell'inversa $\boldsymbol{\Sigma}^{-1}$, così che l'inversa sia data da:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{AA} & \boldsymbol{\Sigma}^{AB} \\ \boldsymbol{\Sigma}^{AB\mathrm{T}} & \boldsymbol{\Sigma}^{BB} \end{pmatrix}$$

$$\mathbf{\Sigma}^{AA} = \mathbf{\Sigma}_{AA.B}^{-1};$$

$$\mathbf{\Sigma}^{AB} = -\mathbf{\Sigma}_{AA.B}^{-1}\mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1};$$

$$\mathbf{\Sigma}^{BA} = -\mathbf{\Sigma}_{BB}^{-1}\mathbf{\Sigma}_{AB}^{T}\mathbf{\Sigma}{-}1_{AA.B};$$

$$\mathbf{\Sigma}^{BB} = \mathbf{\Sigma}_{BB}^{-1}\left[\mathbf{\Sigma}_{AB}^{T}\mathbf{\Sigma}_{AA.B}^{-1}\mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1} + \mathbf{I}\right].$$

Non si confonda ad esempio $\mathbf{\Sigma}^{AA}$ (blocco dell'inversa $\mathbf{\Sigma}^{-1}$ corrispondente agli indici AA) con $\mathbf{\Sigma}_{AA}^{-1}$ (inversa del blocco di $\mathbf{\Sigma}$ corrispondente agli indici AA) (coincidono solo se $\mathbf{\Sigma}_{AB} = 0$)

Ricaviamo dai noti teoremi del calcolo delle probabilità la densità della distribuzione condizionata di $\mathbf{X}_A$:

$$f_{\mathbf{X}_A}|_{\mathbf{X}_B=\mathbf{x}_B}(\mathbf{x}_A\|\mathbf{X}_B = \mathbf{x}_B) = \frac{f_{\mathbf{X}_A\mathbf{X}_B}(\mathbf{x}_A, \mathbf{x}_B)}{f_{\mathbf{X}_B}(\mathbf{x}_B)}$$

E' più comodo lavorare sui logaritmi ed in particolare su $-2\log f$ (in modo da trasformare solo le forme quadratiche a numeratore dell'esponente nella densità normale), indicando per brevità con K la costante di normalizzazione, che si può determinare dopo:

$$-2\log[f(\mathbf{x}_A, \mathbf{x}_B)/f(\mathbf{x}_B)] =$$

$$= K + \mathbf{x}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x} - \mathbf{x}_B^{\mathrm{T}}\mathbf{\Sigma}_{BB}^{-1}\mathbf{x}_B =$$

$$= K + \mathbf{x}_A^{\mathrm{T}}\mathbf{\Sigma}^{AA}\mathbf{x}_A + 2x_A^T\mathbf{\Sigma}^{AB}\mathbf{x}_B + \mathbf{x}_B^T\mathbf{\Sigma}^{BB}\mathbf{x}_B - \mathbf{x}_B^{\mathrm{T}}\mathbf{\Sigma}_{BB}^{-1}\mathbf{x}_B =$$

(sostituendo gli opportuni blocchi di $\mathbf{\Sigma}^{-1}$)

$$= K + \mathbf{x}_A^{\mathrm{T}}\mathbf{\Sigma}_{AA.B}^{-1}\mathbf{x}_A - 2x_A^T\mathbf{\Sigma}_{AA.B}^{-1}\mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}\mathbf{x}_B +$$

$$+\mathbf{x}_B^T\mathbf{\Sigma}_{BB}^{-1}\left[\mathbf{\Sigma}_{AB}^{T}{}^{\mathrm{T}}\mathbf{\Sigma}_{AA.B}^{-1}\mathbf{\Sigma}_{AB}^{T}\mathbf{\Sigma}_{BB}^{-1} + \mathbf{I}\right]\mathbf{x}_B - \mathbf{x}_B^{\mathrm{T}}\mathbf{\Sigma}_{BB}^{-1}\mathbf{x}_B =$$

$$= K + \mathbf{x}_A^{\mathrm{T}}\mathbf{\Sigma}_{AA.B}^{-1}\mathbf{x}_A - 2\mathbf{x}_A^T\mathbf{\Sigma}_{AA.B}^{-1}[\mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}\mathbf{x}_B] +$$

$$+[\mathbf{x}_B^T \boldsymbol{\Sigma}_{BB}^{-1} \boldsymbol{\Sigma}_{AB}^T] \boldsymbol{\Sigma}_{AA.B}^{-1} [\boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_{BB}^{-1} \mathbf{x}_B] =$$

$$= K + \left(\mathbf{x}_A - \boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_{BB}^{-1} \mathbf{x}_B\right)^{\mathrm{T}} \boldsymbol{\Sigma}_{AA.B}^{-1} \left(\mathbf{x}_A - \boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_{BB}^{-1} \mathbf{x}_B\right)$$

Per cui è chiaro dall'ultima forma quadratica, che si tratta del numeratore dell'esponente di una distribuzione normale di parametri:

$$\boldsymbol{\mu}_{\mathbf{X}_A \| \mathbf{X}_B = \mathbf{x}_B} = \boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_{BB}^{-1} \mathbf{x}_B$$

$$\boldsymbol{\Sigma}_{\mathbf{X}_A \| \mathbf{X}_B = \mathbf{x}_B} = \boldsymbol{\Sigma}_{AA.B} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_{BB}^{-1} \boldsymbol{\Sigma}_{AB}^T = (\boldsymbol{\Sigma}^{AA})^{-1}$$

(La costante K è ricavabile dalla condizione di normalizzazione, ma si può comunque verificare effettuando il rapporto fra i termini costanti delle due densità, tenendo presente che per matrici partizionate si ha:

$$\|\boldsymbol{\Sigma}\| = \|\boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_{BB}^{-1} \boldsymbol{\Sigma}_{AB}^T\| \|\boldsymbol{\Sigma}_{BB}\| = \|\boldsymbol{\Sigma}_{AA.B}\| \|\boldsymbol{\Sigma}_{BB}\|$$

Per cui la distribuzione condizionata è:

$$\mathbf{X}_{A|\mathbf{x}_B} \sim \mathcal{N}\left[\boldsymbol{\Sigma}_{AB} \boldsymbol{\Sigma}_{BB}^{-1} \mathbf{x}_B; \boldsymbol{\Sigma}_{AA.B}\right]$$

e quindi si ha:

Distribuzioni condizionate nel caso generale di vettori aleatori normali:

$$\mathbf{Y}_A \| \mathbf{Y}_B = \mathbf{y}_B \sim \mathcal{N}\left[\boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{y}_B - \boldsymbol{\mu}_B); \boldsymbol{\Sigma}_{AA.B}\right]$$

La distribuzione condizionata è normale multivariata con parametri:

$$\mathrm{E}\left[\mathbf{Y}_A \| \mathbf{Y}_B = \mathbf{y}_B\right) = \boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}(\mathbf{y}_B - \boldsymbol{\mu}_B\right]$$

la funzione di regressione (speranza matematica condizionata) è lineare in $\mathbf{y}_B$

$$V(\mathbf{Y}_A \| \mathbf{Y}_B = \mathbf{y}_B) = \boldsymbol{\Sigma}_{AA.B} = \boldsymbol{\Sigma}_{AA} - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\boldsymbol{\Sigma}_{AB}^T = (\boldsymbol{\Sigma}^{AA})^{-1}$$

la matrice di varianze e covarianze condizionate non dipende da $\mathbf{y}_B$ (omoscedasticità) i vettori casuali:

$$\mathbf{Y}_A - (\boldsymbol{\mu}_A + \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}[\mathbf{Y}_B - \boldsymbol{\mu}_B]) e \mathbf{Y}_B$$

(oppure $\mathbf{Y}_A - \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_{BB}^{-1}\mathbf{Y}_B$ e $\mathbf{Y}_B$)
risultano indipendenti (si verifica subito calcolndo $E(\mathbf{Y}_A\mathbf{Y}_B^T$)

---

link o riferimento
(vedere anche →)(figure varie)

Esempio Esempio numerico: Si consideri la matrice 3×3 di varianza e covarianza relativa ad una distribuzione normale multivariata a tre componenti:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

Consideriamo la distribuzione della variabile 1 condizionatamente alla 2 e alla 3. La matrice di varianze e covarianze va quindi partizionata seguente modo:

$$\boldsymbol{\Sigma} = \left(\begin{array}{c|cc} 2 & 1 & 1 \\ \hline 1 & 2 & 1 \\ 1 & 1 & 1 \end{array}\right)$$

Mentre

$$\boldsymbol{\Sigma}_{BB} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

`controllare`
`inserire lucidi manuali ed esercizio`
`completare`

5.8.2 Significato degli elementi dell'inversa della matrice di varianza e covarianza.

E' possibile dare anche un significato agli elementi dell'inversa di $\boldsymbol{\Sigma}$, in termini di distribuzioni condizionate, nel caso in cui $\boldsymbol{\Sigma}$ sia la matrice di varianza e covarianza di variabili aleatorie normali; si vedrà come tali concetti possano essere estesi al caso di variabili aleatorie non normali o, meglio, nell'analisi di dati multivariati, al caso di variabili statistiche osservate.

Gli elementi non diagonali dell'inversa: la correlazione parziale

Intanto, con riferimento ad una distribuzione normale multivariata con matrice di varianze e covarianze $\boldsymbol{\Sigma}$, si può dimostrare che se $\mathbf{C} = \boldsymbol{\Sigma}^{-1}$, allora:

teorema 5.8.2 *$c_{ij} = 0$ è condizione necessaria e sufficiente perché le variabili $\mathbf{Y}_i$ e $\mathbf{Y}_j$ siano indipendenti condizionatamente alle altre $p-2$ variabili $\mathbf{Y}_B$.*

Si può giungere al risultato in due modi:
Dalla densità normale multivariata si vede direttamente che:
se e solo se $c_{ij} = 0$ si ha la fattorizzazione:

$$f(\mathbf{y}) = f(\mathbf{y}_i, \mathbf{y}_B) f(\mathbf{y}_j, \mathbf{y}_B)$$

che è una condizione necessaria e sufficiente per l'indipendenza condizionata di due variabili aleatorie qualsiasi dotate di densità.

Infatti, ponendo $\mathbf{Y}_A = (\mathbf{y}_i, \mathbf{y}_j)^T$ e indicando con $\mathbf{Y}_B$ tutte le altre componenti, avendo indicato con $\mathbf{C}$ l'inversa della matrice di varianza e covarianza opportunamente partizionata:

$$\mathbf{C} = \left(\begin{array}{cc|c} c_{ii} & c_{ij} & \mathbf{c}_{iB}^T \\ c_{ij} & c_{jj} & \mathbf{c}_{jB}^T \\ \hline \mathbf{c}_{iB} & \mathbf{c}_{jB} & \mathbf{C}_{BB} \end{array} \right)$$

si ha:

$$f(\mathbf{y}) = f(\mathbf{y}_i, \mathbf{y}_j, \mathbf{y}_B) = K \times exp[-(\mathbf{y}^T \mathbf{C} \mathbf{y})/2] =$$

$$K \times \exp[-(c_{ii}\mathbf{y}_i^2 + c_{jj}\mathbf{y}_j^2 + 2c_{ij}\mathbf{y}_i\mathbf{y}_j + 2y_i\mathbf{c}_{iB}^T\mathbf{y}_B + 2y_j\mathbf{c}_{jB}^T\mathbf{y}_B + \mathbf{y}_B^T\mathbf{C}_{BB}\mathbf{y}_B)/2]$$

Se ora $c_{ij} = 0$ allora si può facilmente operare su $f(\mathbf{y})$:

$$f(\mathbf{y}) = K \times \exp[-(c_{ii}\mathbf{y}_i^2 + c_{jj}\mathbf{y}_j^2 + 2y_i\mathbf{c}_{iB}^T\mathbf{y}_B + 2y_j\mathbf{c}_{jB}^T\mathbf{y}_B + \mathbf{y}_B^T\mathbf{C}_{BB}\mathbf{y}_B)/2] =$$

$$= K \times \exp[-(c_{ii}\mathbf{y}_i^2 + 2y_i\mathbf{c}_{iB}^T\mathbf{y}_B + \mathbf{y}_B^T\mathbf{C}_{BB}\mathbf{y}_B)/2] \times \exp[-(c_{jj}\mathbf{y}_j^2 + 2y_j\mathbf{c}_{jB}^T\mathbf{y}_B)/2]$$

$$\overbrace{g(\mathbf{y}_i, \mathbf{y}_B) \times g(\mathbf{y}_j, \mathbf{y}_B)}$$

in modo da ottenere la fattorizzazione desiderata in due funzioni, in cui non compaiono simultaneamente termini in $\mathbf{y}_i$ e $\mathbf{y}_j$

Per una interpretazione in generale del significato dei termini dell'inversa, e non solo per il caso estremo $c_{ij} = 0$, conviene riferirsi alle distribuzioni condizionate.

Dalla distribuzione di $\mathbf{Y}_A$ condizionata a $\mathbf{Y}_B = \mathbf{y}_B$, ponendo $\mathbf{Y}_A = (\mathbf{y}_i, \mathbf{y}_j)^T$ (e quindi nella notazione della sezione precedente A è uguale alla coppia di indici i, j e B all'insieme degli altri $p-2$ indici) si ricava che essendo la distribuzione condizionata di $\mathbf{Y}_A$ ancora normale, l'indipendenza condizionata si ha se e solo se $\mathbf{y}_i, \mathbf{y}_j$ risultano non correlati, condizionatamente a $\mathbf{Y}_B = \mathbf{y}_B$.

Si è visto che:

$$V(\mathbf{Y}_A \| \mathbf{y}_B) = \mathbf{\Sigma}_{AA.B}^{-1} = \mathbf{\Sigma}_{AA} - \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}\mathbf{\Sigma}_{AB}^T = (\mathbf{\Sigma}^{AA})^{-1}$$

cioè la varianza condizionata di $\mathbf{Y}_A$ è uguale all'inversa del blocco di elementi corrispondenti ad $\mathbf{Y}_A$ nell'inversa di $\mathbf{\Sigma}$.

Nel caso di due variabili i e j , occorre invertire la matrice 2×2 di elementi:

$$\mathbf{\Sigma}^{AA} = \begin{pmatrix} c_{ii} & c_{ij} \\ c_{ij} & c_{jj} \end{pmatrix}$$

e quindi:

$$(\mathbf{\Sigma}^{AA})^{-1} = \begin{pmatrix} c_{jj} & -c_{ij} \\ -c_{ij} & c_{ii} \end{pmatrix} / (c_{ii}c_{jj} - c_{ij}^2)$$

pertanto $\mathbf{y}_i$ *e* $\mathbf{y}_j$ *sono non correlati condizionatamente alle altre* $p-2$ *variabili, e quindi indipendenti data la normalità della distribuzione condizionata, se e solo se* $c_{ij} = 0$.

Dagli elementi di $(\mathbf{\Sigma}^{AA})^{-1}$ è possibile calcolare l'indice di correlazione lineare fra $\mathbf{y}_i$ e $\mathbf{y}_j$ condizionatamente a $\mathbf{Y}_B$:

$$corr(\mathbf{y}_i, \mathbf{y}_j \| \mathbf{Y}_B = \mathbf{y}_B) = \frac{-c_{ij}}{\sqrt{c_{ii}c_{jj}}} = \frac{-\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$$

(essendo σ^{ij} il cofattore di σ_{ij} in $\mathbf{\Sigma}$)

---

indice di correlazione lineare parziale ossia correlazione fra due variabili eliminata l'influenza delle altre $p-2$ variabili

Esempi sulla differenza fra l'indipendenza condizionata e l'indipendenza marginale

L'indipendenza condizionata e l'indipendenza marginale sono due concetti diversi, e nessuno dei due implica l'altro.

Per chiarire la differenza fra indipendenza marginale e indipendenza condizionata, ricorro qui ad un esempio relativo alla distribuzione congiunta di tre variabili dicotomiche A,B e C.

Esempio Si ha una tavola $2 \times 2 \times 2$ di tre mutabili A,B, e C. Le due tavole $A \times B$ condizionate ai valori di C sono:

$\mathbf{C} = c_1$	b_1	b_2	$tot.$		$\mathbf{C} = c_2$	b_1	b_2	$tot.$
a_1	0,24	0,06	0,30		a_1	0,12	0,28	0,4
a_2	0,56	0,14	0,70		a_2	0,18	0,42	0,6
$tot.$	0,80	0,20	1,00		$tot.$	0,30	0,70	1,00

In queste distribuzioni condizionate A e B sono indipendenti;

se P(C=c_1)= P(C=c_2)=$\frac{1}{2}$la tavola marginale $A \times B$ è:

C*tot.*	b_1	b_2	*tot.*
a_1	0,18	0,17	0,35
a_2	0,37	0,28	0,65
tot.	0,55	0,45	1,00

Nella distribuzione marginale A e B non sono indipendenti.

Si può presentare il caso opposto, di caratteri indipendenti marginalmente e associati condizionatamente (paradosso di Simpson).

citazione

Si ha un'altra tavola $2 \times 2 \times 2$ di tre mutabili A,B, e C. Le due tavole $A \times B$ condizionate ai valori di C sono ora:

$\mathbf{C} = c_1$	b_1	b_2	*tot.*		$\mathbf{C} = c_2$	b_1	b_2	*tot.*
a_1	0,5	0	0,5		a_1	0	0,5	0,5
a_2	0	0,5	0,5		a_2	0,5	0	0,5
tot.	0,5	0,5	1		*tot.*	0,5	0,5	1

In queste distribuzioni condizionate A e B sono associati (addirittura sono massimamente associati)

Infatti se P($C = c_1$)= P(C=c_2)=$\frac{1}{2}$la tavola marginale $A \times B$ è:

C*tot.*	b_1	b_2	*tot.*
a_1	0,25	0,25	0,5
a_2	0,25	0,25	0,5
tot.	0,5	0,5	1

Nella distribuzione marginale A e B sono indipendenti (addirittura equidistribuite)

Gli elementi diagonali dell'inversa: la correlazione multipla

Anche gli elementi sulla diagonale principale di $\mathbf{\Sigma}^{-1}$ sono interpretabili tenendo conto delle distribuzioni condizionate, ma in termini di variabilità di una variabile spiegata da tutte le altre, concetto che rivedremo poi nel caso di modelli lineari generali.

Infatti se ora consideriamo l'insieme $\mathbf{Y}_A$ costituito da una sola variabile $\mathbf{y}_i$ (e quindi nella notazione adottata finora A è uguale all'indice i e B all'insieme degli altri $p-1$ indici), si ha per la varianza di $\mathbf{y}_i$ condizionata ai valori delle altre $p-1$ variabili:

Tenendo conto che $\mathbf{\Sigma}^{AA} = c_{ii}$ si ha:

$$V(\mathbf{y}_i \| \mathbf{y}_B) = (\mathbf{\Sigma}^{AA})^{-1} = 1/c_{ii} = \|\mathbf{\Sigma}|/\sigma^{ii}$$

Quindi l'inverso di un elemento diagonale dell'inversa della matrice di varianze e covarianze esprime la varianza della variabile di posto corrispondente condizionatamente alle altre $p-1$ variabili.

$$\max(\frac{1}{c_{ii}}) = \sigma_i^2 \qquad \min(c_{ii}) = \frac{1}{\sigma_i^2}$$

Il massimo di questa quantità è proprio la varianza della componente i -esima, ossia σ_i^2

Se $\mathbf{\Sigma}$ è una matrice $\mathbf{Z}$ di correlazione, allora $1/c_{ii}$indica la variabilità di $\mathbf{y}_i$ non spiegata dalle altre $p-1$ variabili, per cui si può costruire il coefficiente di determinazione multipla:

$$R^2_{i.B} = 1 - \|\mathbf{Z}|/z^{ii} = 1 - 1/c_{ii} = 1 - \frac{V(\mathbf{y}_i \| \|\mathbf{Y}_B)}{V(\mathbf{y}_i)}$$

Misura quanta parte della variabilità di $\mathbf{Y}_i$ è spiegata dalle altre p-1 variabili del vettore aleatorio $\mathbf{y_B}$

In generale l'indice di correlazione lineare multipla è dato da:

$$R_{i.B} = \sqrt{1 - \frac{|\mathbf{\Sigma}|}{\sigma_i^2 \sigma^{ii}}} = \sqrt{1 - 1/(\sigma_i^2 c_{ii})}$$

```
\begin{fig}
Esempio Date le rilevazioni di  $p=7 $ misure antropometriche
 su un insieme di  $n=1432 $ bambini,
 si è calcolata la matrice di correlazione  $\mZ $ che segue:

\mZ=

\end{fig}
```

Ad esempio la correlazione lineare (marginale, ossia senza tenere conto della presenza delle altre variabili) fra le prime due variabili è di 0,719.

Figura da inserire Da questa matrice di correlazione si è calcolata l'inversa $\mathbf{C}$: $\mathbf{C} =$

e quindi si è calcolata la matrice $\mathbf{A}$ che ha come elemento generico:

$$r_{ij.B} = \frac{-c_{ij}}{\sqrt{c_{ii}c_{jj}}}$$

correlazione parziale fra due variabili, $\mathbf{X}_i$ e $\mathbf{X}_j$, tenute costanti le altre 5:

Figura da inserire P=

(ovviamente in questa matrice gli elementi diagonali non vanno considerati Si vede che la correlazione lineare (parziale, o meglio condizionata, ossia tenute costanti le altre variabili) fra le prime due variabili è di 0,245. Buona parte quindi della correlazione marginale è indotta dall'influenza delle altre 5 variabili, ossia la covariazione delle prime due variabili insieme alle altre 5.

Se invece trasformiamo gli elementi diagonali di $\mathbf{C}$, mediante la relazione:

$$R^2_{i.B} = 1 - \frac{\|\mathbf{Z}|}{z^{ii}} = 1 - \frac{1}{c_{ii}}$$

otteniamo i 7 indici di determinazione multipla, di ciascuna variabile condizionatamente alle altre 6:

$$0.827137, 0.896544, 0.848327, 0.297231, 0.722443, 0.756753, 0.82098$$

Si noti che la matrice di correlazione ha 7 autovalori dati da:

$$\boldsymbol{\lambda}^{\mathrm{T}} = 5.06451, 0.674288, 0.635871, 0.245914, 0.207684, 0.105888, 0.06584$$

La successione di tali valori indica chiaramente la presenza di correlazioni lineari fra combinazioni lineari di variabili molto forti.

Impiego delle informazioni dell'inversa C nell'analisi di dati multivariati.

Come si è visto, l'analisi degli elementi dell'inversa della matrice di correlazione può fornire degli elementi utili per indagare sulla dipendenza fra variabili sia in termini marginali che in termini condizionati.

Esempio di variabili condizionatamente non correlate

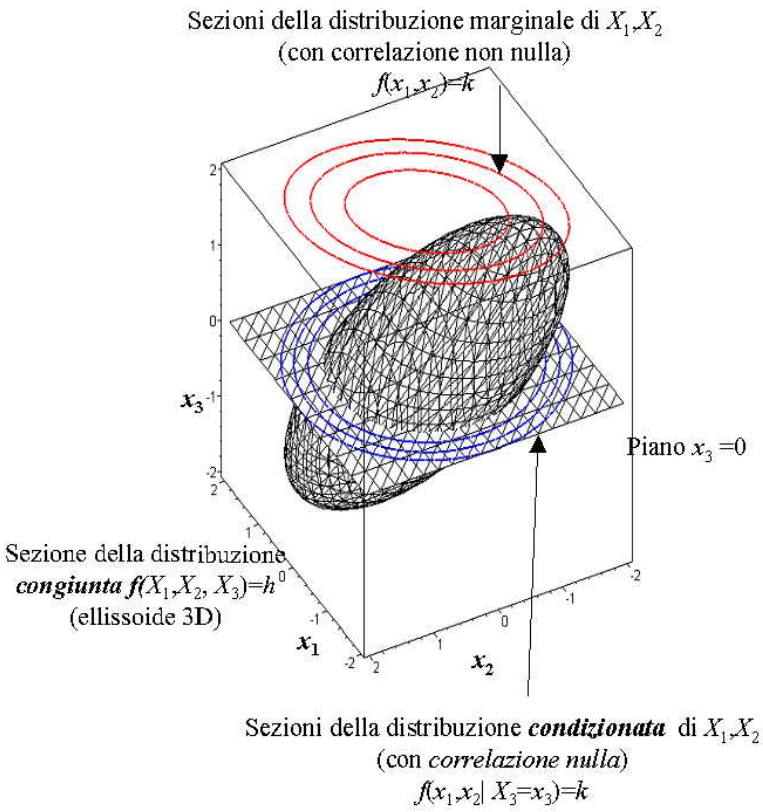

Figura 5.8: distribuzioni condizionate in una normale multivariata

vai a indice figure

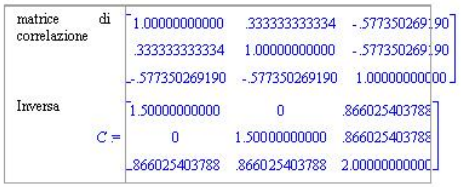

Figura 5.9: distribuzioni condizionate in una normale multivariatamatrice di varianze e covarianze e inversa

vai a indice figure

5.9 Utilità della distribuzione normale multivariata

In effetti quanto visto finora riguarda solo il modello teorico della normale multivariata, ossia le caratteristiche delle distribuzioni di vettori aleatori normali multivariati, che riassumo brevemente (e solo per le proprietà più rilevanti)

- dipende solo dai primi due momenti multivariati;
- ha contorni iper-ellissoidali;
- ha distribuzioni marginali normali multivariate;
- ha distribuzioni condizionate (o parziali) normali multivariate omoschedastiche e con funzioni di regressione lineari;
- combinazioni lineari di sue componenti sono ancora normali multivariate;

- è unimodale;

```
xxxxxxxxxxxxxxxxxxxxxxxxxxx
\item si ottiene come distribuzione
limite di processi multivariati
come teorema limite centrale multivariato
Non ci stiamo per ora ponendo
il problema di adattare una tale
distribuzione a dati osservati.
In effetti in questo corso questo
problema non verrà affrontato,
 se non marginalmente: l'importanza
del modello normale multivariato
 per questo corso sta nel fatto
che è un modello utile per la
 definizione di
relazioni di dipendenza in
 media esattamente lineari ed
omoschedastiche, che
xxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

5.10 Regressioni approssimate per vettori aleatori qualsiasi

In generale se abbiamo un vettore aleatorio $\mathbf{Z}$ a p componenti con distribuzione qualsiasi,

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_i \\ \vdots \\ Z_p \end{pmatrix}$$

possiamo essere interessati a misurare in qualche modo la dipendenza di una componente dalle altre, diciamo per semplicità per ora la dipendenza di Z_1 da tutte le altre componenti, $Z_2, \ldots, Z_p$; in altre parole vogliamo vedere se e come si modifica la distribuzione di Z_1, condizionata a particolari valori $z_2, \ldots, z_p$ assunti dalle altre $p-1$ variabili, al variare dei valori condizionanti $z_2, \ldots, z_p$.

Per semplicità supponiamo che la distribuzione condizionata di Z_1 esista sempre e sia dotata di densità per qualsiasi insieme di valori $z_2, \ldots, z_p$.

Siamo quindi interessati allo studio della distribuzione condizionata di Z_1 di densità:

$$f_{Z_1}(z_1|Z_2 = z_2, \ldots, Z_p = z_p)$$

al variare dei valori $z_2, \ldots, z_p$.

Tale densità (univariata) è ovviamente data da:

$$f_{Z_1}(z_1|Z_2 = z_2, \ldots, Z_p = z_p) = \frac{f_{\mathbf{Z}}(z_1, z_2, \ldots, z_p)}{f_{Z_2,\ldots,Z_p}(z_2, \ldots, z_p)}$$

Come esprimere adesso la dipendenza di Z_1 dai valori $z_2, \ldots, z_p$ in modo sintetico, possibilmente senza considerare l'intera distribuzione condizionata?

Una soluzione del tutto naturale è quella di considerare una funzione $g(\cdot)$ (matematica, non aleatoria), dei valori $z_2, \ldots, z_p$ che sintetizzi al meglio la distribuzione di densità $f_{Z_1}(z_1|Z_2 = z_2, \ldots, Z_p = z_p)$.

Vogliamo quindi sostituire alla variabile aleatoria $Z_1|Z_2 = z_2, \ldots, Z_p = z_p$, una funzione $g(z_2, \ldots, z_p)$ in modo tale che sia minimo la perdita di informazione . Se adottiamo un criterio di perdita quadratico, dobbiamo minimizzare quindi il valore atteso:

$$\mathrm{E}\left[((Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - g(z_2, \ldots, z_p))^2\right] \qquad (5.4)$$

in corrispondenza di ciascuna combinazione di valori $z_2, \ldots, z_p$.

Con questa impostazione, ossia una funzione di perdita quadratica, è evidente che il valore che minimizza la (5.4) è il valore atteso della distribuzione condizionata di Z_1 ossia:

$$g(z_2, \ldots, z_p) = \mathrm{E}\left[(Z_1|Z_2 = z_2, \ldots, Z_p = z_p)\right]$$

Tale funzione va sotto il nome di *funzione di regressione di* Z_1 *su* $Z_2, \ldots, Z_p$

La bontà di tale funzione di regressione nel sintetizzare la distribuzione condizionata di Z_1 è valutabile attraverso la *funzione di varianza condizionata*:

$$V\left[Z_1|Z_2 = z_2, \ldots, Z_p = z_p\right] = \mathrm{E}\left[((Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - g(z_2, \ldots, z_p))^2\right]$$

L'analisi di questa funzione mostra se le varianze sono costanti o meno e come variano in funzione dei valori $z_2, \ldots, z_p$.

La funzione di regressione ovviamente, tranne che in casi particolari, è una funzione qualsiasi: può essere lineare, polinomiale, esponenziale o altro.

Ci possiamo porre ancora un altro problema:

invece della funzione di regressione esatta, usiamo una funzione parametrica $h(z_2, \ldots, z_p; \boldsymbol{\beta})$, che dipenda da un numero ridotto di parametri $\boldsymbol{\beta}$. Anche stavolta vorremo minimizzare la perdita quadratica:

$$\mathrm{E}\left[((Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - h(z_2, \ldots, z_p; \boldsymbol{\beta}))^2\right] \tag{5.5}$$

Vediamo subito che relazione c'è fra questa perdita e quella minima realizzata con la funzione di regressione: non v'è dubbio che il minimo della quantità in (5.5) sarà superiore al valore ottimo (5.4), perchè nella (5.5) si minimizza rispetto ad una particolare funzione parametrica.

Si può poi vedere che:

$$\mathrm{E}\left[((Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - h(z_2, \ldots, z_p; \boldsymbol{\beta}))^2\right] =$$

$$= \mathrm{E}\left[(\{(Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - \mathrm{E}\left[Z_1|Z_2 = z_2, \ldots, Z_p = z_p\right]\}\right] + \left\{\mathrm{E}\left[Z_1|Z_2 = z_2, \ldots, Z\right.\right.$$

$$= \mathrm{E}\left[(\{(Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - \mathrm{E}\left[Z_1|Z_2 = z_2, \ldots, Z_p = z_p\right]\})^2 + \left\{\mathrm{E}\left[Z_1|Z_2 = z_2, \ldots\right.\right.\right.$$

-

$$2\mathrm{E}\left[\{(Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - \mathrm{E}\left[Z_1|Z_2 = z_2, \ldots, Z_p = z_p\right]\}\right]\{\mathrm{E}\left[Z_1|Z_2 = z_2, \ldots, Z_p =\right.$$

E' facile vedere che il doppio prodotto è nullo, dato che:

$$2\mathrm{E}\left[\{(Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - \mathrm{E}\left[Z_1|Z_2 = z_2, \ldots, Z_p = z_p\right]\}\right]\{\mathrm{E}\left[Z_1|Z_2 = z_2, \ldots, Z_p =\right.$$

$$= 2\left\{\mathrm{E}\left[Z_1|Z_2 = z_2, \ldots, Z_p = z_p\right] - h(z_2, \ldots, z_p; \boldsymbol{\beta})\right\}\mathrm{E}\left[\{(Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - \mathrm{E}\left[\right.\right.$$

perchè$\mathrm{E}\left[\{(Z_1|Z_2 = z_2, \ldots, Z_p = z_p) - \mathrm{E}\left[Z_1|Z_2 = z_2, \ldots, Z_p = z_p\right]\}\right] = 0.$

In definitiva abbiamo, utilizzando una notazione più sintetica ma altrettanto chiara:

$$\mathrm{E}\left[((Z_1|z_2,\ldots,z_p) - h(z_2,\ldots,z_p;\boldsymbol{\beta}))^2\right] =$$

$$= \mathrm{E}\left[(\{(Z_1|z_2,\ldots,z_p) - \mathrm{E}\left[Z_1|Z_1|z_2,\ldots,z_p\right]\})^2 + \left\{\mathrm{E}\left[Z_1|Z_1|z_2,\ldots,z_p\right] - h(z_2,\ldots,z_p;\boldsymbol{\beta})^2\right\}\right]$$

Questa relazione è molto importante per due motivi:

1. La funzione di perdita

   $$\mathrm{E}\left[((Z_1|z_2,\ldots,z_p) - h(z_2,\ldots,z_p;\boldsymbol{\beta}))^2\right]$$

   relativa alla funzione $h(z_2,\ldots,z_p;\boldsymbol{\beta})$ può essere scomposta in due componenti:

   - la funzione di varianza condizionata:

     $$\mathrm{E}\left[(\right]\{(Z_1|z_2,\ldots,z_p) - \mathrm{E}\left[Z_1|Z_1|z_2,\ldots,z_p\right]\})^2$$

   - e la cosiddetta divergenza dalla funzione $h()$:

     $$\mathrm{E}\left[(\right]\{(h(z_2,\ldots,z_p;\boldsymbol{\beta}) - \mathrm{E}\left[Z_1|Z_1|z_2,\ldots,z_p\right]\})^2$$

2. dal momento che la varianza condizionata non dipende dalla funzione $h(z_2,\ldots,z_p;\boldsymbol{\beta})$, per ottenere il valore ottimo di $\boldsymbol{\beta}$, invece di minimizzare la (5.4) possiamo minimizzare rispetto a $\boldsymbol{\beta}$ la quantità:

   $$\mathrm{E}\left[(\right]\{(h(z_2,\ldots,z_p;\boldsymbol{\beta}) - \mathrm{E}\left[Z_1|Z_1|z_2,\ldots,z_p\right]\})^2$$

```
trovare esempi semplici di
 regressioni teoriche non lineari.
 e mettere dei grafici
```

5.10.1 Regressioni lineari approssimate per vettori aleatori qualsiasi

Come visto prima, vettori aleatori con distribuzioni qualsiasi, o variabili statistiche osservate,avranno funzioni di regressione di forma qualsiasi (anche non lineare) e con varianze diverse (eteroscedasticità).

Accenniamo adesso al caso della distribuzione condizionata di un numero qualsiasi di componenti: In generale se $\mathbf{Z}$ è un vettore aleatorio con distribuzione qualsiasi, e $\mathbf{Z}_A$ e $\mathbf{Z}_B$ sono due vettori ottenuti dalle componenti di $\mathbf{Z}$, allora:

funzione di regressione la funzione di regressione di $\mathbf{Z}_A$ su $\mathbf{Z}_B$ è la speranza matematica di $\mathbf{Z}_A$ condizionatamente a particolari valori di $\mathbf{Z}_B$:

- $E(\mathbf{Z}_A \| \mathbf{Z}_B = z_B)$ (se esiste) è una funzione di z_B di forma qualsiasi (in generale non lineare).
- la distribuzione (condizionata) di $\mathbf{Z}_A$, con densità $f_{\mathbf{Z}_A}(\mathbf{z}_A \| \mathbf{Z}_B = z_B)$ è in generale non normale.
- tale distribuzione dipende in generale dai particolari valori fissati di z_B. In particolare quindi può essere con varianze $V(\mathbf{Z}_A \| \mathbf{Z}_B = z_B)$ non costanti.

Tuttavia se si considerano le regressioni parziali lineari approssimate (ossia le relazioni lineari che approssimano, secondo i minimi quadrati, le curve di regressione) si ritrovano le stesse espressioni (come funzioni della matrice di varianza e covarianza) che abbiamo trovato per la normale multivariata.

Nel caso normale però queste relazioni sono esatte.

Le relazioni di regressione lineare approssimate in generale si trovano minimizzando rispetto alla matrice $\mathbf{W}$ la quantità:

$$tr[V(\mathbf{Z}_A - \mathbf{W}\mathbf{Z}_B)]$$

(equivalente a $E(\mathbf{Z}_A - \mathbf{W}\mathbf{Z}_B)^2$ se si lavora con vettori aleatori a media nulla).

Si ottiene comunque:

$$\mathbf{W} = \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}$$

Le regressioni però saranno esattamente lineari e omoscedastiche solo nel caso normale multivariato. Figura da inserire

(figure varie)

link o riferimento

(vedere anche *rightarrow*regressione parziale e condizionata)

Analisi delle correlazioni lineari presenti in data set osservati

`forse è il caso di metterlo dopo`

Sezione avanzata

E' il caso di fare comunque delle considerazioni sui momenti del secondo ordine quando si opera con variabili che non seguono una normale multivariata, o quando si ha a disposizione un insieme di dati per il quale non si può ipotizzare che si tratti di un campione proveniente da una distribuzione normale multivariata.

Per esempio quando si utilizzano delle relazioni lineari approssimate secondo quanto visto in precedenza, si sta implicitamente ipotizzando, oltre la linearità, l'uguaglianza fra le varianze delle distribuzioni parziali e l'uguaglianza fra le correlazioni e le covarianze delle distribuzioni parziali, indipendentemente dai particolari valori fissati per le variabili indipendenti (o condizionanti). Si stanno cioè ipotizzando relazioni parziali che non cambiano forma al variare delle condizioni. Supponiamo per esempio di avere in un insieme di dati (n unità $\times p$ variabili) relativo a p variabili $\mathbf{X}_i$, in cui le n unità sono suddivise in k_Z gruppi secondo le modalità $z_h (h = 1, 2, \ldots, k_Z)$ di una ulteriore variabile $\mathbf{Z}$, supponendo quindi di avere delle osservazioni ripetute in corrispondenza di ciascuna delle k_Z modalità di $\mathbf{Z}$. Possiamo allora calcolare le varianze delle p variabili $\mathbf{X}_i$ e le loro correlazioni in coppia per ciascuno dei k_Z gruppi. Se per esempio osserviamo che le varianze di una o più variabili cambiano in modo sostanziale da un gruppo ad un altro oppure se le correlazioni fra alcune variabili cambiano in modo marcato in corrispondenza delle varie modalità di Z, questo può essere un indizio del fatto che l'approssimazione dei minimi quadrati delle vere regressioni non sarà appropriata e che quindi i dati presentano caratteristiche diverse da quelle di una normale multivariata, per cui questo può essere indizio di assenza di multinormalità. In questi casi occorrerà ricorre ad altre approssimazioni, non lineari, o eteroscedastiche, che tengano eventualmente in conto momenti diversi dai primi due. **chiarire**

5.11 Sintesi delle informazioni sui vari tipi di correlazione e dipendenza lineare ricavabile dai primi 2 momenti multivariati

I momenti multivariati primo e secondo, ossia il vettore delle speranze matematiche e la matrice di varianze e covarianze contengono tutte (e sole) le informazioni che servono per la quantificazione e l'analisi di tutti i tipi di dipendenza e correlazioni lineari relative a coppie o gruppi di variabili, sia nelle distribuzioni marginali che in quelle condizionate (si veda a proposito anche la sezione 5.8).

Riassumo nella tavola che segue le relazioni lineari e loro connessione con i momenti primi e secondi, secondo quanto fin qui studiato, per variabili multiple $\mathbf{X}$ a p componenti e con momenti primi nulli (quindi si tratta di variabili centrate):

$$\mathbf{X} = \{X_1, X_2, \ldots, X_i, \ldots, X_p\}^{\mathrm{T}} \qquad \mathrm{E}[\mathbf{X}] = \mathbf{0}$$

Ho indicato con $\mathbf{\Sigma}$ la matrice di varianze e covarianze e matrice di correlazione $\mathbf{R}$, i cui elementi sono al solito dati da:

$$r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

con σ_i^2 si è indicata la varianza della i -esima componente, ossia l'elemento i -esimo della diagonale di $\mathbf{\Sigma}$, di modo che σ_i è lo scostamento quadratico medio della i -esima variabile.

In effetti se con D indichiamo una matrice diagonale i cui elementi sono le varianze delle singole componenti, per cui $d_{ij} = 0$ se $i \neq j$ e $d_{ii} = \sigma_i^2$, si può esprime la matrice di correlazione in termini matriciali:

$$\mathbf{R} = D^{-\frac{1}{2}} \mathbf{\Sigma} D^{-\frac{1}{2}},$$

Secondo il simbolismo già adottato, con $\mathbf{C}$, di elemento generico c_{ij} , si è indicata l'inversa di $\mathbf{\Sigma}$, esprimibile al solito in termini dei cofattori σ^{ij} degli elementi di posto i, j della matrice $\mathbf{\Sigma}$:

$$c_{ij} = \sigma^{ij} / \|\mathbf{\Sigma}|$$

...

relazioni lineari e loro connessione con i momenti primi e secondi di una variabile $\mathbf{X}$

Significato statistico-probabilistico	espressione in termini di elementi di $\mathbf{\Sigma}$
varianza di una componente $\mathbf{X}_i$	σ_i^2
varianze e covarianze di una combinazione lineare $\mathbf{Y} = \mathbf{A}\ \mathbf{X}$	$\mathbf{A\Sigma A}^T$
varianza di tutte le componenti	$tr(\mathbf{\Sigma})$

Significato statistico-probabilistico	espressione in termini di elementi di $\mathbf{\Sigma}$
varianza generalizzata (di Wilks)	$\|\mathbf{\Sigma}\|$
correlazione lineare semplice fra due variabili $\mathbf{X}_i, \mathbf{X}_j$	$r_{ij} = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$
coefficiente di regressione lineare semplice di una variabile X_i rispetto ad un'altra, X_j	$b_{i,j} = \frac{\sigma_{ij}}{\sigma_j^2}$
correlazione multipla: correlazione fra una variabile X_j ed una combinazione lineare (la migliore, nel senso dei minimi quadrati!) delle altre $p-1$ variabili, che sono le componenti vettore aleatorio $\mathbf{X}_B$, con $B = 1, 2, \ldots, j-1, j+1, \ldots, p$ dipendenza lineare di una variabile dalle altre $p-1$ variabili (combinate linearmente nel miglior modo possibile). frazione della varianza di $\mathbf{X}_i$ spiegata dalle altre p-1 variabili.	$R_{i.B} = \sqrt{1 - \|\mathbf{\Sigma}\|/[\sigma_i^2 \sigma^{ii}]}$ $= \sqrt{1 - 1/(\sigma_i^2 c_{ii})} = 1 - \dfrac{V(\mathbf{X}_i \| \mathbf{X}_B)}{V(\mathbf{X}_i)}$
frazione della varianza della distribuzione di $\mathbf{X}_i$ condizionatamente a $\mathbf{X}_B$	$\dfrac{V(\mathbf{X}_i \| \mathbf{X}_B)}{V(\mathbf{X}_i)}$

Significato statistico-probabilistico	espressione in termini di elementi di $\mathbf{\Sigma}$
matrice di varianze e covarianze della regressione lineare di un gruppo di variabili $\mathbf{X}_A$ in dipendenza di una latro gruppo di variabili $\mathbf{X}_B$ (per il simbolismo sulle matrici partizionate si veda la sezione sulla normale multivariata)	$\mathbf{\Sigma}_{AA} - \mathbf{\Sigma}_{AB}\mathbf{\Sigma}_{BB}^{-1}\mathbf{\Sigma}_{AB}^{T} == (\mathbf{\Sigma}^{AA})^{-1}$
correlazione parziale fra due variabili, $\mathbf{X}_i$ e $\mathbf{X}_j$, tenute costanti le altre p-2	$r_{ij.B} = \dfrac{-c_{ij}}{\sqrt{c_{ii}c_{jj}}}$
combinazioni lineari a coefficienti normalizzati di massima varianza (e retta di minima distanza dall'insieme di dati multivariato)	Si ricavano dagli autovettori di $\mathbf{\Sigma}$
varianza massima di una combinazione lineare delle $\mathbf{X}_i$ (a coefficienti normalizzati)	λ_1
varianza minima di una combinazione lineare delle $\mathbf{X}_i$ (a coefficienti normalizzati)	λ_p
varianze delle componenti principali (combinazioni lineari delle $\mathbf{X}_i$ (a coefficienti normalizzati)	$\boldsymbol{\lambda}$: vettore degli autovalori di $\mathbf{\Sigma}$
combinazioni lineari di gruppi di variabili con correlazione massima. Analisi delle correlazioni canoniche	solo accennata

Per l'analisi di relazioni di tipo non lineare (o di regressioni lineari per esempio eteroscedastiche), occorre far ricorso ad altri momenti multivariati oltre il secondo. Si faranno degli esempi nell'ambito dell' analisi dei residui nella regressione multipla lineare.

In effetti si vedrà che anche nel modello lineare generale, l'analisi

della dipendenza lineare e delle proprietà degli stimatori, sotto certe ipotesi semplificatrici è legata *solo alla struttura delle varianze e delle covarianze fra variabili dipendenti e indipendenti.*

5.12 Stimatori di massima verosimiglianza dei parametri di una normale multivariata

Supponiamo di avere un campione(multivariato) casuale di ampiezza n estratto da una normale multivariata a p componenti, ossia una matrice $\mathbf{X} n \times p$ di dati, le cui righe sono delle determinazioni di una variabile normale multipla:

$$\mathbf{X} = \left(\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix} \right)$$

In analogia al caso univariato, i momenti primi e secondi calcolati sul campione multivariato sono le stime di massima verosimiglianza dei corrispondenti parametri della distribuzione di provenienza; in

sintesi:

Lo stimatore di massima verosimiglianza del vettore delle speranze matematiche $\boldsymbol{\mu}$ di una variabile normale multipla è dato dal vettore $M(\mathbf{X})$ delle medie aritmetiche di un campione multivariato $\mathbf{x}$ di n osservazioni i.i.d. estratto dalla corrispondente distribuzione.
Tale stimatore, come nel caso univariato, è corretto, ossia non distorto.

Lo stimatore di massima verosimiglianza della matrice di varianze e covarianze $\boldsymbol{\Sigma}$ di tale variabile è dato dalla matrice delle varianze e covarianze empiriche calcolata su un campione multivariato di n osservazioni i.i.d. estratto dalla corrispondente distribuzione.
Tale stimatore, come nel caso univariato, è invece distorto.
E' possibile costruire uno stimatore corretto moltiplicando sia le varianze che le covarianze empiriche per il fattore correttivo $\frac{n}{n-1}$, ottenendo quindi lo stimatore:

$$\hat{\boldsymbol{\Sigma}} = \mathrm{V}[\mathbf{X}] = \mathrm{V}[\mathbf{Z}] = \frac{\mathbf{Z}^{\mathrm{T}}\mathbf{Z}}{n}$$

In effetti, dal momento che gli unici parametri della distribuzione normale multivariata sono il vettore delle medie e la matrice di varianza e covarianza, per ottenere gli stimatori di massima verosimiglianza (puntuali!) di tutte le quantità necessarie per calcolare le distribuzioni congiunte, marginali, condizionate e per le componenti principali da un campione proveniente da una normale multivariata, si impiegheranno le stesse formule già viste per la distribuzione teorica, sostituendo ai momenti primi e secondi teorici quelli empirici stimati dal campione, dal momento che lo stimatore di massima verosimiglianza di una funzione dei parametri $g(\boldsymbol{\theta})$ è dato dalla stessa funzione dello stimatore di Massima verosimiglianza, $g(\hat{\theta})$

Dimostrazione

È opportuno a questo punto richiamare e rivedere le proprietà viste precedentemente sulla derivazione di forme quadratiche di determinanti e di matrici inverse.

Sezione avanzata

Per ricavare gli stimatori di massima verosimiglianza dei parametri di una normale multivariata costruiamo come sempre la verosimiglianza, o meglio il suo logaritmo, supponendo di avere n osservazioni indipendenti ciascuna con p componenti.

Per comodità e perché questo facilita i passaggi successivi, consideriamo come parametri gli elementi c_{ij} di $\mathbf{C}$, l'inversa della matrice $\mathbf{\Sigma}$ di varianze e covarianze, oltre ovviamente al vettore delle speranze matematiche $\boldsymbol{\mu}$.

Sappiamo dalle proprietà degli stimatori di massima verosimiglianza che la parametrizzazione è irrilevante ai fini della determinazione degli stimatori puntuali.

Costruiamo la quantità:

$$-2\log L(\boldsymbol{\mu};\mathbf{C})$$

(essendo $L(\boldsymbol{\mu}e\mathbf{C})$ la verosimiglianza campionaria (rispetto a $\boldsymbol{\mu}$ e $\mathbf{C}$), sulla base di un campione di n osservazioni indipendenti (si riveda la parte iniziale sulla normale multivariata, per questa parametrizzazione, in particolare l'equazione 5.3):

$$l(\boldsymbol{\mu},\mathbf{C};\mathbf{X}) = -2logL(\boldsymbol{\mu},\mathbf{C};\mathbf{X}) = k - nlog|\mathbf{C}| + \sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}(\mathbf{x}_i-\boldsymbol{\mu})$$

essendo $\mathbf{x}_i$ il vettore (p -variato) osservato relativo all'i-esima osservazione. Procedendo a derivare prima rispetto al vettore $\boldsymbol{\mu}$ si ha:

$$\frac{\partial l(\boldsymbol{\mu},\mathbf{C};\mathbf{X})}{\partial \boldsymbol{\mu}} = -2\sum_{i=1}^{n}\mathbf{C}(\mathbf{x}_i-\boldsymbol{\mu}) = -2\mathbf{C}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})$$

E' immediato vedere che $\frac{\partial l(\boldsymbol{\mu},\mathbf{C};\mathbf{X})}{\partial \boldsymbol{\mu}}$ si annulla se:

$$2\mathbf{C}\sum_{i=1}^{n}(\mathbf{x}_i-\hat{\mu}) = 0$$

ossia (dato che $\mathbf{C}$ è di rango pieno!) solo quando:

$$\sum_{i=1}^{n}(\mathbf{x}_i-\hat{\mu}) = 0,$$

ed infine:

$$\sum_{i=1}^{n}\mathbf{x}_i = n\hat{\mu}\hat{\mu} = \sum_{i=1}^{n}\mathbf{x}_i/n (= M(\mathbf{X}) = \begin{pmatrix} M_1 \\ \dots \\ M_j \\ \dots \\ M_p \end{pmatrix}$$

Per quanto riguarda invece le derivate rispetto agli elementi di $\mathbf{C}$ conviene distinguere gli elementi diagonali c_{jj} da quelli fuori dalla diagonale $c_{jk}(k \neq j)$:

$$\frac{\partial l(\boldsymbol{\mu},\mathbf{C};\mathbf{X})}{\partial c_{jj}} = -\frac{n}{|\mathbf{C}|}\frac{\partial|\mathbf{C}|}{\partial c_{jj}} + \frac{\partial\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{C}(\mathbf{x}_i-\boldsymbol{\mu})}{\partial c_{jj}} j = 1,2,\dots,p$$

Per il primo addendo a secondo membro ricordiamo il risultato generale per i determinanti di matrici simmetriche:

$$\frac{\partial|\mathbf{C}|}{\partial c_{jj}} = \mathbf{C}_{ii}$$

essendo $\mathbf{C}_{rs}$il cofattore di c_{rs} in $\mathbf{C}$,

mentre per il secondo addendo ovviamente si tratta di termini lineari in $\mathbf{C}$, per cui basterà nella sommatoria selezionare solo le componenti opportune dei vettori $(\mathbf{x}_i - \boldsymbol{\mu})$, ossia solo quelle che moltiplicano c_{jj} :

$$\frac{\partial l(\boldsymbol{\mu}, \mathbf{C}; \mathbf{X})}{\partial c_{jj}} = -n\frac{\mathbf{C}_{jj}}{|\mathbf{C}|} + \sum_{i=1}^{n}(x_{ij} - \mu_j)^2.$$

Si vede subito che:

$$\frac{\mathbf{C}_{jj}}{|\mathbf{C}|} = \sigma_j^2$$

dal momento che $\mathbf{C} = \boldsymbol{\Sigma}^{-1}$ e quindi $\boldsymbol{\Sigma} = \mathbf{C}^{-1}$ e gli elementi di un'inversa sono proprio dati dai rapporti fra cofattori e determinante.

Per trovare le espressioni degli stimatori $\hat{\sigma}_j^2$ occorre annullare le precedenti derivate, avendo sostituito alle speranza matematiche μ_j gli stimatori di massima verosimiglianza M_j. Pertanto:

$$\frac{\partial l(\boldsymbol{\mu}, \mathbf{C}; \mathbf{X})}{\partial c_{jj}} = 0 \rightarrow -n\frac{\mathbf{C}_{jj}}{|\mathbf{C}|} + \sum_{i=1}^{n}(x_{ij} - M_j)^2 = 0;$$

e quindi:

$$\hat{\sigma}_J^2 = \frac{\sum_{i=1}^{n}(x_{ij} - M_j)^2}{n}$$

Deriviamo adesso rispetto agli elementi non diagonali $c_{jk}(k \neq j$):

$$\frac{\partial l(\boldsymbol{\mu}\mathbf{C}; \mathbf{X})}{\partial c_{jk}} = -\frac{n}{|\mathbf{C}|}\frac{\partial |\mathbf{C}|}{\partial c_{jk}} + \frac{\partial \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\mathbf{C}(\mathbf{x}_i - \boldsymbol{\mu})}{\partial c_{jk}} j, k = 1, 2, \ldots, p; k \neq j$$

Procediamo come prima, e per il primo addendo a secondo membro ricordiamo il risultato generale per i determinanti di matrici simmetriche:

$$\frac{\partial |\mathbf{C}|}{\partial c_{jk}} = 2C_{jk}$$

cofattore di c_{jk} in $\mathbf{C}$, $k \neq j$.

Mentre per il secondo addendo selezioniamo le componenti dei vettori $(\mathbf{x}_i - \boldsymbol{\mu})$ che moltiplicano c_{jk}:

$$\frac{\partial l(\boldsymbol{\mu}, \mathbf{C}; \mathbf{X})}{\partial c_{jk}} = -2n\frac{\mathbf{C}_{jk}}{|\mathbf{C}|} + 2\sum_{i=1}^{n}(x_{ij} - \mu_j)(x_{ik} - \mu_k)$$

Ancora si ha:

$$\frac{\mathbf{C}_{jk}}{|\mathbf{C}|} = \sigma_{jk}$$

e per trovare le espressioni degli stimatori $\hat{\sigma}_{jk}$ occorre annullare le precedenti derivate, avendo sostituito alle speranze matematiche μ_j gli stimatori di massima verosimiglianza M_j. Pertanto:

$$\frac{\partial l(\boldsymbol{\mu}, \mathbf{C}; \mathbf{X})}{\partial c_{jk}} = 0 \Rightarrow -2n\frac{\mathbf{C}_{jk}}{|\mathbf{C}|} + 2\sum_{i=1}^{n}(x_{ij} - M_j)(x_{ik} - M_k) = 0;$$

e quindi:

$$\hat{\sigma}_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - M_j)(x_{ik} - M_k)}{n}$$

e quindi in definitiva il risultato prima anticipato:

$$\hat{\boldsymbol{\Sigma}} = V(\mathbf{X}) = V(\mathbf{Z}) = \mathbf{Z}^T\mathbf{Z}/n$$

essendo $\mathbf{X}$ il campione multivariato originario e $\mathbf{Z}$ la matrice degli scarti

Inferenza nel caso normale sugli autovalori:

Sebbene solitamente si impieghino le tecniche di analisi delle componenti principali a scopo esplorativo, è interessante accennare al caso in cui si abbia a disposizione un campione multivariato estratto da una distribuzione normale; abbiamo infatti già visto come per una distribuzione normale multivariata gli autovalori e gli autovettori assumano dei significati ben precisi.

Evidentemente gli stimatori di massima verosimiglianza degli autovalori e degli autovettori sono forniti dagli autovalori e dagli autovettori della matrice di varianza e covarianza campionaria (che è lo stimatore di massima verosimiglianza della matrice di varianza e covarianza teorica); dal momento che per gli stimatori delle varianze e delle covarianze per campioni provenienti da una normale valgono della proprietà di regolarità e dei teoremi che forniscono le distribuzioni campionarie e che garantiscono la consistenza degli stimatori insieme con loro correttezza asintotica, dobbiamo aspettarci che anche per gli autovalori e gli autovettori ricavati da tali matrici campionarie valgano delle proprietà di consistenza e di correttezza asintotica. In effetti qui mi limito a riportare un risultato asintotico che riguarda la distribuzione degli autovalori per campioni provenienti da una distribuzione normale multivariata.

Asintoticamente gli l_j , stime campionarie dei veri autovalori λ_j , ottenute da un campione di n osservazioni estratto da una normale multivariata, si distribuiscono secondo una normale multivariata a componenti indipendenti:

con valore atteso: $E[l_j] = \lambda_j$

e varianza campionaria :

$$Var[l_j] = \frac{2\lambda_j^2}{n-1}$$

(si ricordi il caso particolare di matrici di varianze e covarianze diagonali: questi risultati coincidono con quelli classici della distribuzione di una varianza campionaria!)

Casi interessanti:

$$H_0 : \lambda_j = 1, j = 1, 2, \ldots, p$$

che corrisponde al caso di indipendenza fra le variabili (standardizzate).

È da intendersi che questi risultati sono semplicemente delle approssimazioni ma danno delle indicazioni sull'ordine di grandezza dell'errore campionario.

5.12.1 Un test di Multinormalità: cenni

Quando si ha a disposizione un campione di dati multivariato, molto spesso è necessario verificare se è plausibile l'ipotesi di provenienza da un universo normale multivariato.

Un modo semplice per verificare la normalità di un campione di osservazioni multivariate, consiste ovviamente nell'effettuare dei test di normalità su ciascuna delle distribuzioni univariate.

Ricordo che la normalità delle distribuzioni marginali è una condizione necessaria ma non sufficiente per la normalità multivariata: pertanto i test sulla normalità delle distribuzioni marginali costituiscono uno sbarramento preliminare, nel senso che se danno esito negativo possiamo senz'altro scartare l'ipotesi di multinormalità, altrimenti occorrerà procedere col saggiare l'ipotesi di normalità multivariata con test basati sulla distribuzione congiunta.

Se l'insieme in esame è costituito da molte variabili non sarà possibile utilizzare i normali test di bontà dell'adattamento; tuttavia è possibile ottenere delle informazioni eventualmente anche grafiche trasformando opportunamente l'insieme di dati multivariato.

Come si è visto infatti nel capitolo sulla distribuzione normale multivariata, la forma quadratica ad esponente della densità normale ha una distribuzione proporzionale a quella di una χ^2 con p gradi di libertà.

Infatti se:

$$\mathbf{Y}(N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

si è già visto prima che la variabile casuale

$$\mathbf{Q} = (\mathbf{Y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu}) (\chi^2_p$$

Pertanto se trasformiamo ognuno degli n vettori osservati $\mathbf{x}_i$ a p componenti secondo la stessa relazione, dovremo aspettarci che questi n valori trasformati q_i seguano ciascuno una distribuzione χ^2 con p gradi di libertà:

$$q_i = (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \sim \chi^2_p$$

(le n trasformate q_i risultano indipendenti per l'indipendenza ipotizzata dei vettori osservati $\mathbf{x}_i$)

Quindi, se è valida l' ipotesi di multinormalità, il vettore delle n trasformate q_i costituisce un campione casuale semplice estratto da una distribuzione χ^2 con p gradi di libertà. In effetti le quantità che si usano effettivamente per il calcolo delle q_i sono gli stimatori di $\boldsymbol{\mu}$ e $\boldsymbol{\Sigma}$, M e $\mathbf{S}$, e non i parametri veri (usualmente incogniti); questo fa sì che le quantità:

$$\hat{q}_i = (\mathbf{x}_i - M)^T \mathbf{S}^{-1} (\mathbf{x}_i - M)$$

seguono una distribuzione χ^2_p solo approssimativamente; l'approssimazione è soddisfacente per campioni grandi.

In effetti, un'informazione utile si ricava dalla rappresentazione grafica di tali valori trasformati in corrispondenza dei percentili teorici di una variabile χ^2 ; un altro elemento di cui si potrebbe tenere conto nella costruzione di un test di normalità è dato dagli angoli che i vettori osservati formano con il centroide del campione; tuttavia adesso per semplicità non vedremo quest'ulteriore possibilità.

Esempio:

Questo esempio è tratto dall'insieme di dati antropometrici di cui si è fatto cenno in capitoli precedenti (1432 casi x 7 variabili).

```
\begin{fig}

in ese2000_correlaz1.nb

\end{fig}
```

Inserire grafici sulle distribuzioni normali condizionate

5.13 Inferenza sui parametri della normale multipla

Mi dispiace! capitolo ancora da fare

5.14 Esempi di distribuzioni multivariate non normali

Sezione avanzata

Saltare nella versione breve del corso.

L'estensione al caso multivariato di distribuzioni non-normali a componenti non indipendenti è sempre ardua, perché le possibilità di estensione di sistemi di curve univariate non normali al caso multivariato possono essere di diversa natura, mentre dalla distribuzione normale univariata si può arrivare alla sua estensione multivariata con diverse impostazioni giungendo sempre alla stessa forma multivariata; ad esempio:

- dalla densità o dalla funzione caratteristica, sostituendo ad un quadrato una forma quadratica;
- se $\mathbf{x}^{\mathrm{T}}\mathbf{a}$ è normale per qualsiasi $\mathbf{a}$, allora x è normale multivariato.
- come distribuzione di $\mathbf{x} = \boldsymbol{\mu} + \mathbf{A}\mathbf{y}$(con $\mathbf{y}$ a componenti indipendenti)
- da distribuzioni condizionate normali e omoscedastiche con funzioni di regressione lineari.

5.14.1 Una particolare distribuzione beta multivariata (distribuzione di Dirichlet)

La distribuzione di Dirichlet a k componenti, che costituisce una particolare generalizzazione multivariata della distribuzione Beta, è definita come segue:

- si considerino $k+1$ v.a. indipendenti $\mathbf{X}_i (i = 0, 1, 2, \ldots, k)$, ciascuna con distribuzione Gamma con lo stesso parametro di scala λ e di parametri di forma c_i ;
- indicata con S la loro somma, $S = \sum_{i=0}^{k} \mathbf{X}_i$, la distribuzione di Dirichlet è la distribuzione congiunta delle k nuove variabili definite dalle relazioni:

$$\mathbf{y}_i = \mathbf{X}_i / S i = 1, 2, \ldots, k.$$

La densità di tale distribuzione è data da:

$$f(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k) = \prod_{i=1}^{k} \mathbf{y}_i^{ci-1} [1 - \sum_{i=1}^{k} \mathbf{y}_i]^{c0-1} \boldsymbol{\Gamma}(\sum_{i=0}^{k} c_i) / \prod_{i=0}^{k} \boldsymbol{\Gamma}(c_i),$$

ed è definita sul simplesso:

$$\mathbf{y}_i (0, i = 1, 2, \ldots, k; \sum_{i=1}^{k} \mathbf{y}_i \leq 1.$$

Questa distribuzione è importante ad esempio per la descrizione della distribuzione simultanea di rapporti di composizione; si vedano nelle figure che seguono, per il caso bivariato, alcuni esempi di densità per diverse combinazioni dei parametri c_0, c_1 e c_2 .(indicati nel seguito con a, b, c nel caso bivariato)

Se $c_i (1 (i = 0, 1, \ldots, k)$, la densità ha sempre un massimo unico in corrispondenza di:

$$\mathbf{y}_i^* = (c_i - 1) / \sum_{i=0}^{k} (c_i - 1), (i = 1, 2, \ldots, k).$$

- Tutte le distribuzioni marginali univariate sono delle distribuzioni Beta.
- Le distribuzioni condizionate sono ancora delle Beta
- Nella distribuzione bivariata (indicando le due componenti con $\mathbf{X}$,$\mathbf{y}$, e i parametri con a,b,c)) la distribuzione di $\mathbf{y}$ condizionata a $\mathbf{X}$ =x è proporzionale ad una variabile con distribuzione Beta univariata. In particolare si dimostra che:
- $\mathbf{y}/(1 - x) \| \mathbf{X} = x$ si distribuisce come una $Beta[b, c]$
- per cui E [$\mathbf{y}$] varia linearmente con x, ma anche V [$\mathbf{y}$] varia con x

Esempi di densità di distribuzioni di Dirichlet: Figura da inserire in bivar1.nb

$$\begin{array}{ccc} c_0 = 1,2 & c_1 = 0,9 & c_2 = 0,9 \\ c_0 = 1,2 & c_1 = 1,3 & c_2 = 1,8 \\ c_0 = 3 & c_1 = 4 & c_2 = 5. \end{array}$$

\begin{fig}

FIG2000REGR_ETERO1.STG

\end{fig}

Altri esempi di distribuzioni multivariate non normali

Distribuzione Logistica Doppia di densità:

$$F(x, \mathbf{y}) = 1/(1 + Exp[-x] + Exp[-\mathbf{y}])$$

```
\begin{fig}
in bivar1.nb
\end{fig}

Distribuzione Esponenziale Bivariata  $(a=0,7)$
$$
F(\vecx,\vecy)=
(1-Exp[-\vecx])(1-Exp[-\vecy])( (1+a Exp[-\vecx-\vecy])
$$

\begin{fig}
in bivar1.nb
\end{fig}

Distribuzione Bivariata Dirichlet
 (\mBeta-bivariata)  $a=1,5;
\vecb=1,6; c=2,1$
\begin{fig}
in bivar1.nb
\end{fig}

Distribuzione Bivariata Dirichlet
 (\mBeta-bivariata)  $a=4;
\vecb=4; c=3$

\begin{fig}
```

```
in bivar1.nb

\end{fig}
Distribuzione Bivariata Dirichlet
(\mBeta-bivariata)  $a=1,1;
\vecb=1,1; c=0,9$
\begin{fig}
in bivar1.nb
\end{fig}
```

Sezione avanzata

costruzione di variabili correlate

Uno schema generale di costruzione di variabili aleatorie correlate da p+1 variabili aleatorie indipendenti $\mathbf{X}_j$ (j=0,1, . . . ,p), è quello di considerare p variabili aleatorie sommando a tutte la componente $\mathbf{X}_0$. In dettaglio otteniamo ora un nuovo vettore aleatorio $\mathbf{Y}$ a p componenti, ponendo:

$$\begin{pmatrix} \mathbf{y}_1 = \mathbf{X}_0 + \mathbf{X}_1 \\ \cdots \\ \mathbf{y}_j = \mathbf{X}_0 + \mathbf{X}_j \\ \cdots \\ \mathbf{y}_p = \mathbf{X}_0 + \mathbf{X}_p \end{pmatrix}$$

In pratica la componente $\mathbf{X}_0$ è quella che determina la covarianza fra le componenti di $\mathbf{Y}$.

E' facile calcolare i momenti di $\mathbf{Y}$ da quelli di $\mathbf{X}$, mentre può essere in generale arduo calcolare la distribuzione di $\mathbf{Y}$ (è spesso è complicato integrare rispetto a $\mathbf{X}_0$ nella densità congiunta di $\mathbf{X}_0$, $\mathbf{X}_1$, . . . , $\mathbf{X}_p$,).

Come esercizio si calcoli la correlazione e la covarianza fra due generiche componenti di $\mathbf{Y}$ o, direttamente, la matrice di varianze e covarianze e la matrice di correlazione di $\mathbf{Y}$.

$$V(\mathbf{y}_j) = V(\mathbf{X}_0) + V(\mathbf{X}_j); Cov(\mathbf{y}_j, \mathbf{y}_k) = V(\mathbf{X}_0)$$

costruzione di variabili correlate

Uno schema generale di costruzione di variabili aleatorie correlate da p+1 variabili aleatorie indipendenti $\mathbf{X}_j$ (j=0,1, . . . ,p), è quello di considerare p variabili aleatorie sommando a tutte la componente $\mathbf{X}_0$. In dettaglio otteniamo ora un nuovo vettore aleatorio $\mathbf{Y}$ a p componenti, ponendo:

In pratica la componente $\mathbf{X}_0$ è quella che determina la covarianza fra le componenti di $\mathbf{Y}$.

E' facile calcolare i momenti di $\mathbf{Y}$ da quelli di $\mathbf{X}$, mentre può essere in generale arduo calcolare la distribuzione di $\mathbf{Y}$ (è spesso è complicato integrare rispetto a $\mathbf{X}_0$ nella densità congiunta di $\mathbf{X}_0$, $\mathbf{X}_1$, . . . , $\mathbf{X}_p$,).

Come esercizio si calcoli la correlazione e la covarianza fra due generiche componenti di $\mathbf{Y}$ o, direttamente, la matrice di varianze e covarianze e la matrice di correlazione di $\mathbf{Y}$.

$$V(\mathbf{y}_j) = V(\mathbf{X}_0) + V(\mathbf{X}_j); Cov(\mathbf{y}_j, \mathbf{y}_k) = V(\mathbf{X}_0)$$

Capitolo 6

Introduzione ai Modelli Lineari

Figura da inserire
FIG2000REGR1.STG
FIG2000REGR2.STG
FIG2000REGR3.STG

6.1 Il modello lineare di dipendenza per variabili normali.

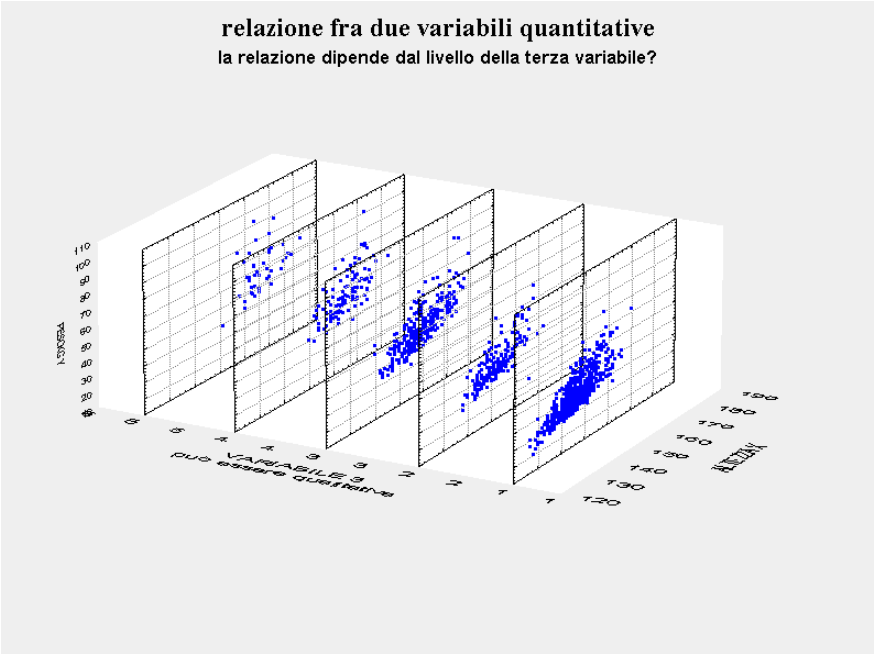

Figura 6.1: relazione fra due variabili in funzione del valore di una terza variabile

vai a indice figure

Per quanto visto nelle lezioni sulla normale multivariata, la distribuzione di un numero qualsiasi di componenti condizionatamente a valori qualsiasi $\mathbf{Z}_2$ di altre componenti del vettore aleatorio normale è normale, con valore atteso che è funzione lineare di $\mathbf{Z}_2$, e matrice di varianze e covarianze indipendente dai particolari valori condizionanti; quindi le regressioni sono tutte lineari e omoscedastiche.

Pertanto se si ha a disposizione un campione casuale semplice da una normale multivariata, non esiste alcun problema di identificazione del modello, né di scelta della funzione, perché tutte le distribuzioni condizionate sono note.

Tuttavia sono rari i casi in cui nello studio della dipendenza di uno o più fenomeni, si può ragionevolmente ipotizzare di avere un campione casuale semplice da una distribuzione normale multipla, perché spesso ci si trova in altre situazioni, fra cui essenzialmente si hanno le seguenti:

- I dati costituiscono un campione casuale semplice proveniente da una distribuzione multivariata *non normale*.
- I dati non costituiscono un campione casuale semplice ma, per esempio, i valori delle variabili indipendenti sono stati opportunamente selezionati o predisposti
- oppure si ha un campione non probabilistico o comunque un archivio di dati che non costituisce un campione.
- Il modello da cui si possono selezionare i dati è effettivamente una distribuzione normale multivariata (almeno approssimativamente normale), e si può estrarre un campione casuale semplice, tuttavia l'interesse dell'analisi è limitato allo studio della distribuzione di una delle componenti $\mathbf{y}$ condizionatamente a valori particolari o estremi delle altre componenti $\mathbf{X}$: è noto anche nell'analisi della regressione semplice, che l'inferenza è migliore (ossia le bande di confidenza della relazione di regressione sono più strette) se si selezionano unità con valori estremi delle componenti condizionanti $\mathbf{X}$ più vicine a quelle di interesse.

`spostare questo paragrafo`

Ovviamente restano rinviati (ma solo per poche pagine!) i problemi relativi alla stima dei parametri sulla base di un campione di osservazioni p-variate, che verranno affrontati estendendo opportunamente le tecniche impiegate quando si studia la dipendenza di una variabile $\mathbf{y}$ da una variabile indipendente $\mathbf{x}$.

In effetti anche nel caso di campioni casuali semplici da distribuzioni non normali multivariate, si possono cercare le migliori (nel senso dei minimi quadrati) relazioni lineari fra le speranze matematiche di $\mathbf{y}$ e particolari valori di $\mathbf{X}$. In ogni caso, come si apprestiamo a discutere diffusamente, i valori delle x possono anche non essere

delle determinazioni di variabili casuali, ma valori anche scelti in modo non casuale.

versatilità del modello lineare

Nei paragrafi che seguono verranno affrontati diversi aspetti relativi alla versatilità del modello lineare ed alle diverse possibilità interpretative del modello e dei suoi parametri: alcuni dei concetti fondamentali relativi a particolari modelli lineari vengono introdotti fra breve, prima che vengano affrontati gli aspetti inferenziali.

6.2 Funzioni di regressione

```
questo pezzo va agganciato
con il pezzo sulla regressione
 per vettori aleatori.
```

Supponiamo di avere un vettore aleatorio di $p+1$ componenti:

$$(Y, Z_1, Z_2, \ldots, Z_p)$$

l'approccio alla misura della dipendenza di una componente Y di un vettore aleatorio dalle altre componenti, può essere affrontato in termini di funzione di regressione, ossia della funzione di dipendenza della speranza matematica di Y da particolari configurazioni di $\mathbf{Z}$:

$$\mathrm{E}[Y] = f(\mathbf{z})$$

Ovviamente questo concetto può essere esteso al caso in cui abbiamo n osservazioni relative a $p+1$ variabili statistiche, e si vuole studiare come varia una (o meglio le sue medie) in funzione delle altre.

L'approccio tecnico scelto in questo corso ci consentirà di affrontare in modo simile gli aspetti inferenziali relativi alla regressione multipla, all'analisi della varianza, della covarianza; inoltre costituirà una buona base per alcuni tipi di GLM (Generalized linear models) sia per l'interpretazione dei parametri che per l'inferenza.

6.3 I modelli statistici.

Prima di iniziare lo studio del modello lineare, che ci accompagnerà per tutto (o quasi) il corso) vale la pena di fare una citazione:

Utilità dei modelli statistici

All models are wrong, but some are useful
(G.E.P. Box)
(Tutti i modelli (statistici) sono sbagliati, ma alcuni sono utili)

6.4 Il modello lineare generale.

Per modello lineare in generale si intende un modello nel quale una variabile di risposta osservabile $\mathbf{Y}$ è spiegata da una combinazione lineare di k variabili esplicative $\mathbf{X}_j$, secondo dei parametri incogniti β_j , più una componente accidentale ε (non osservabile) , secondo la generica relazione lineare:

...

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \ldots + \mathbf{X}_k\beta_k + \boldsymbol{\varepsilon}$$

...

In generale si può avere:

$$\mathbf{y} = g(\mathbf{X}_1, \ldots, \mathbf{X}_k, \boldsymbol{\beta}, \boldsymbol{\varepsilon})$$

In particolare comunque ci occuperemo di modelli lineari di dipendenza nei quali le $\mathbf{X}_j$ non sono variabili casuali, ma costanti note, che assumono n valori in $\Re_k$ (tutti distinti oppure con ripetizioni, questo si vedrà meglio in seguito).

A differenza di quanto visto nelle lezioni precedenti, non ci stiamo occupando della distribuzione simultanea di $k+1$ variabili aleatorie, perché le $\mathbf{X}_j$ sono variabili i cui valori possono addirittura essere prefissati ed assegnati.

La generica osservazione i -esima è quindi caratterizzata da un particolare vettore di valori delle k variabili $\mathbf{X}_j$, indicato con:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \ldots \\ x_{ij} \\ \ldots \\ x_{ik} \end{pmatrix}$$

Eventualmente le x potranno essere dei valori particolari **(fissati!)** di variabili casuali, nel caso in cui studiamo le distribuzioni condizionate della variabile aleatoria $\mathbf{Y}$, condizionatamente agli n valori di k variabili aleatorie $\mathbf{X}_j$, e ipotizzeremmo in quel caso l'esistenza di k+1 variabili aleatorie *osservabili*. Anche in questa situazione però non ci occuperemmo della distribuzione congiunta delle $\mathbf{X}_j$, ma solo di $f(\mathbf{Y}|\mathbf{X}_{nxk})$, ossia la distribuzione di $\mathbf{Y}$ condizionatamente a particolari valori delle x.

E' più opportuno allora fornire l'equazione per la variabile casuale $\mathbf{y}_i$ corrispondente alla generica i-esima osservazione:

$$\ldots$$

$$\mathbf{Y}_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ik}\beta_k + \boldsymbol{\varepsilon}_i = \mathbf{x}_i^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$$

Il vettore delle n osservazioni può essere quindi così espresso formalmente:

MODELLO LINEARE GENERALE

$$\mathbf{Y}_{[n\times 1]} = \mathbf{X}_{[n\times k]}\boldsymbol{\beta}_{[k\times 1]} + \boldsymbol{\varepsilon}_{[n\times 1]}$$

L' equazione deve essere lineare *nei parametri* $\boldsymbol{\beta}$.

Rappresentando i dati in blocchi si ha:
Figura da inserire
BLOCCHI

$$\mathbf{Y}_{[n\times 1]} = \mathbf{X}_{[n\times k]}\boldsymbol{\beta}_{[k\times 1]} + \boldsymbol{\varepsilon}_{[n\times 1]}$$

$$\begin{pmatrix} \mathbf{y}_1 \\ \ldots \\ \ldots \\ \mathbf{y}_i \\ \ldots \\ \ldots \\ \mathbf{y}_n \end{pmatrix} = \begin{pmatrix} x_{11}\beta_1+ & x_{12}\beta_2+ & \ldots & +x_{1k}\beta_k \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ x_{i1}\beta_1+ & x_{i2}\beta_2+ & \ldots & +x_{ik}\beta_k \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1}\beta_1+ & x_{n2}\beta_2+ & \ldots & +x_{nk}\beta_k \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \ldots \\ \ldots \\ \boldsymbol{\varepsilon}_i \\ \ldots \\ \ldots \\ \boldsymbol{\varepsilon}_n \end{pmatrix}$$

...

L'utilità e la versatilità di tale modello per la descrizione di fenomeni reali risiede nella possibilità di dare un significato agli elementi di $\mathbf{X}$ e di $\boldsymbol{\beta}$.
Il nome lineare presuppone in generale che il modello sia lineare nei parametri β_j

6.4.1 componente sistematica e componente casuale.

Possiamo interpretare le due componenti fondamentali del modello che forniscono la risposta $\mathbf{Y}$ come:

- $\mathbf{X}\boldsymbol{\beta}$ la componente sistematica del modello;
- $\boldsymbol{\varepsilon}$ la componente accidentale, che qui sto supponendo additiva, per semplicità, e per comodità interpretativa.

Se:

$$\mathrm{E}[\boldsymbol{\varepsilon}] = 0_n$$

(come è ovvio assumere se $\boldsymbol{\varepsilon}$ è effettivamente una componente accidentale additiva) allora:

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta},$$

e quindi il modello è schematizzabile come:

$$\mathbf{Y} = \mathrm{E}[\mathbf{Y}] + \boldsymbol{\varepsilon}$$

In questo caso quindi possiamo vedere la variabile $\mathbf{Y}$ come una variabile casuale, di cui abbiamo un campione di n osservazioni, la cui speranza matematica è funzione lineare di k variabili $\mathbf{X}_j$ secondo la relazione:

$$\mathrm{E}[\mathbf{y}_i] = \sum_{j=1}^{k} x_{ij}\beta_j \qquad i = 1, 2, \ldots, n$$

questa proprietà è in stretta relazione con l'ipotesi di additività della componente accidentale.

L'assunzione $\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}_n$ presuppone la validità del modello per le speranze matematiche e quindi implicitamente si ipotizza:

- che la componente accidentale (che ha un effetto additivo) sia a media nulla: questo in effetti è quasi scontato quando parliamo di errori accidentali additivi;

- che le k variabili siano le uniche rilevanti ai fini della spiegazione della speranza matematica di $\mathbf{Y}$, o meglio della spiegazione di sue variazioni.

- Il modello per la parte sistematica non è distorto, perchè: $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$.

In ogni caso non si sta implicitamente assumendo l'esistenza di relazioni di causa effetto fra le $\mathbf{X}$ e la $\mathbf{Y}$, ma semplicemente che la conoscenza delle $\mathbf{X}$ può spiegare meglio la variabilità del fenomeno $\mathbf{Y}$ (nel senso che ne diminuisce la variabilità).
Ricordo inoltre che non è necessario ipotizzare modelli distributivi per le $\mathbf{X}_j$, perché, almeno per ora, si sta supponendo che siano dei termini fissati, secondo differenti possibili schemi che vengono adesso esaminati

---

Ad esempio nella regressione lineare semplice si ipotizza:

$$\mathbf{Y}_i = \beta_0 + x_i\beta_1 + \boldsymbol{\varepsilon}_i$$

con

$$\mathrm{E}[\mathbf{Y}_i] = \beta_0 + x_i\beta_1$$

Sezione avanzata

Adesso occorre citare e studiare opportunamente gli esempi della lezione introduttiva, che in buona parte sono tutti suscettibili di essere posti in questa forma.

6.4.2 Caratteristiche essenziali degli elementi del modello lineare

Elemento e Dimensioni	Caratteristiche
$\mathbf{Y}$ vettore n elementi	Vettore aleatorio osservabile; è la variabile di risposta di interesse, ossia quella di cui si cerca di studiare (e di spiegare) la variabilità;
$\mathbf{X}$ matrice $n \times k$ elementi	Matrice di costanti note. Le k componenti (vettori di n elementi) sono variabili non aleatorie osservate senza errori Sono le k variabili esplicative che si pensa influenzino la risposta $\mathbf{Y}$. Si vedranno dopo alcune delle numerose configurazioni che può assumere la matrice $\mathbf{X}$.
$\boldsymbol{\beta}$ vettore k elementi	Vettore di parametri incogniti; $\boldsymbol{\beta}$ andrà stimato dai dati del campione. In generale sono dei parametri fissi; in certi modelli, che tratteremo in questo corso solo marginalmente, alcuni dei coefficienti sono considerati come effetti casuali, e quindi come variabili aleatorie.
$\boldsymbol{\varepsilon}$ vettore n elementi	Vettore aleatorio non osservabile direttamente; In funzione delle diverse ipotesi fatte sulla natura della distribuzione di $\boldsymbol{\varepsilon}$ (che può dipendere in generale da un insieme di parametri $\boldsymbol{\theta}$) si possono avere differenti stime dei parametri incogniti del modello.

6.4.3 Caratteristiche più dettagliate degli elementi del modello:

Elemento	Caratteristiche
Y	Vettore aleatorio *osservabile*; vettore n elementi • è la variabile di risposta di interesse, ossia quella di cui si cerca di studiare (e di spiegare) la variabilità; • è una variabile quantitativa; • solo in casi speciali si considerano **Y** qualitative (ad esempio presenza/assenza; oppure successo/insuccesso). In questo corso non affronteremo, almeno non queste tecniche, casi di risposte **y** qualitative non dicotomiche. • Ci stiamo occupando essenzialmente di modelli nei quali la risposta $\mathbf{y}_i$ è univariata; diversamente, con risposte multiple, abbiamo modelli multivariati. • Si considera la distribuzione di **Y** come vettore aleatorio, perché si pensa che questa distribuzione possa per qualche aspetto (media, varianza, etc.) variare in funzione delle X_j . • Il modello è multiplo se si hanno diverse colonne nella matrice **X** • con **y** indichiamo il vettore dei valori osservati • Di solito è utile vedere (preliminarmente) se la variabilità osservata della **Y** è dovuta solo alla variabilità naturale o anche a fattori sistematici (ossia la dipendenza dalle **X**). • Le n unità dovrebbero essere gli elementi di un campione casuale; tuttavia questo modello viene utilizzato anche per analisi esplorative su dati osservazionali o comunque non provenienti da un campione (leggere discussione di Cox su int.stat.rev.)

Elemento	Caratteristiche e Dimensioni
$\mathbf{X}$	Matrice di *costanti note*. matrice $n \times k$ elementi • Le k componenti (vettori di n elementi) sono variabili non aleatorie osservate senza errori • o comunque con un eventuale errore di ordine di grandezza molto inferiore rispetto a quello di $\mathbf{Y}$. • I valori delle x potrebbero essere n valori particolari assunti da un vettore aleatorio p-dimensionale. In questo caso studiamo la distribuzione condizionata di $\mathbf{y}$ per quei particolari valori di $\mathbf{X}$. • Le $\mathbf{X}_j$ sono le k variabili esplicative che si pensa influenzino la risposta $\mathbf{Y}$. Le configurazioni di $\mathbf{X}$ possono essere numerose: • quantitative • variabili indicatrici (0/1 o -1/1) • variabili miste

La matrice delle $\mathbf{X}$ (o meglio l'intero insieme dei dati) può provenire da:

- studi osservazionali: in cui si scelgono le k variabili, ma gli n valori di ciascuna variabile sono quelli osservati negli n individui scelti, per cui non è possibile in generale pianificare particolari combinazioni degli $n \times k$ valori.
- esperimenti pianificati: in cui si scelgono non solo le k variabili, ma anche tutto lo schema degli $n \times k$ valori, per cui è possibile stabilire in anticipo quali valori verranno utilizzati per ciascuna delle k variabili ed inoltre quali combinazioni di valori dei fattori (o delle variabili) verranno impiegate insieme.
- dati ricavati da statistiche ufficiali o archivi e/o databases o dati prelevati da archivi remoti in rete: *possibilmente si tratta di dati raccolti non per finalità statistiche* e pertanto potrebbero essere poco affidabili, di qualità non nota e molto probabilmente non costituiscono nè un campione casuale nè una popolazione completa. [1]

---

[1] Ovviamente questa considerazione riguarda l'intero dataset osservato, compresa la $\mathbf{y}$.

Elemento	Caratteristiche
$\boldsymbol{\beta}$	Vettore di *parametri incogniti*; vettore di k elementi: $$\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_k\}^{\mathrm{T}}$$ **$\boldsymbol{\beta}$ andrà stimato dai dati del campione** • Ciascun parametro esprime la dipendenza (lineare) dalla corrispondente variabile esplicativa. • In generale gli elementi di $\boldsymbol{\beta}$ sono dei parametri fissi, se non precisato diversamente; • in certi modelli alcuni dei coefficienti sono considerati come effetti casuali, e quindi come variabili aleatorie. • Ciascun parametro esprime la dipendenza (lineare) dalla corrispondente variabile esplicativa. • Pertanto β_j misura l'incremento medio della risposta $\mathbf{Y}$ in corrispondenza di un incremento unitario della j-esima variabile $\mathbf{X}_j$. • Se $\mathbf{X}_j$ è una variabile indicatrice (0/1) della presenza di una certa caratteristica (non quantitativa), allora β_j misura l'effetto medio della presenza di tale caratteristica sulla risposta $\mathbf{Y}$. In generale: $$\beta_j = \frac{\partial \mathrm{E}[\mathbf{Y}_i]}{\partial x_{ij}}$$ se il modello è lineare però vale anche: $$x_{ij} = \frac{\partial \mathrm{E}[\mathbf{Y}_i]}{\partial \beta_j}$$

Elemento	Caratteristiche
$\boldsymbol{\varepsilon}$	Vettore aleatorio *non osservabile* direttamente; vettore di n elementi *I*n funzione delle diverse ipotesi fatte sulla natura della distribuzione di $\boldsymbol{\varepsilon}$ (che può dipendere in generale da un insieme di parametri $\boldsymbol{\theta}$) si hanno differenti stimatori dei parametri incogniti del modello. • Rappresenta la componente accidentale, che viene supposta additiva, in modo tale che se è anche con speranza matematica nulla (come spesso si può ipotizzare) si ha: $$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$ • In effetti $\boldsymbol{\varepsilon}$ a rigore dovrebbe essere una variabile aleatoria non dipendente da variabili esterne, che esprime semplicemente l'errore sperimentale, o l'errore di misurazione • nel caso in cui il modello non sia correttamente specificato, $\boldsymbol{\varepsilon}$ finirà per inglobare le variabili ed i fattori non esplicitati nella parte sistematica, e quindi perderà la sua natura di componente accidentale.

6.4.4 Versatilità del modello lineare

La formulazione di tale modello per la speranza matematica di una v.a., sebbene molto semplice, permette di trattare diversi tipi di situazioni e di risolvere differenti problemi di inferenza.

In funzione di particolari configurazioni che può assumere la matrice $\mathbf{X}$, si può adattare questa impostazione a situazioni particolari.

Ad esempio:

- per l'analisi della regressione lineare multipla, se le colonne della matrice x sono n osservazioni di k variabili quantitative,
- per l'analisi della regressione polinomiale, se le colonne della matrice $\mathbf{X}$ sono le potenze di una o più variabili quantitative,
- oppure per l'analisi della varianza se le k colonne di x so-

no delle variabili dicotomiche indicatrici (dummy variables) di appartenenza ad un gruppo;

- per l'analisi della covarianza;
- per particolari analisi di disegni sperimentali a più vie con interazioni fino ad un ordine massimo fissato.
- Analisi di superfici di risposta
- Analisi discriminante
- Analisi dei modelli di crescita

---

soltanto alcune di queste problematiche verranno trattate in questi appunti;

---

si rivedano comunque gli esempi tratti dalla sezione di problemi introduttivi

---

6.5 Problemi di inferenza

In generale in un modello lineare possiamo avere diversi problemi di inferenza, in particolare di stima e di prova delle ipotesi, in funzione della natura dei dati e del tipo di problema. Ad esempio:

- stimare il vettore dei parametri $\boldsymbol{\beta}$ nel caso generale;
- stimare il vettore dei parametri $\boldsymbol{\beta}$ nel caso in cui vengono imposti dei vincoli su alcune delle sue componenti (alcune componenti nulle o uguali, per esempio)
- Il valore del vettore dei parametri $\boldsymbol{\beta}$ è uguale ad un certo valore β_0 ?
- Costruzione di una regione di confidenza per il vettore dei parametri $\boldsymbol{\beta}$;
- Costruzione di un intervallo di confidenza per una delle componenti di $\boldsymbol{\beta}$; (o per una combinazione lineare delle componenti di $\boldsymbol{\beta}$, ad esempio $\beta_1 - (\beta_2 + \beta_3)/2$).

- Inferenza su r componenti di $\boldsymbol{\beta}$; le altre $k - r$ componenti di $\boldsymbol{\beta}$ non interessano e svolgono però il ruolo di parametri di disturbo.

- Gli effetti di alcune variabili $\mathbf{X}_j$ sono uguali? Ossia alcuni dei parametri sono uguali?

- Alcuni dei parametri sono uguali subordinatamente al valore di altre variabili $\mathbf{X}_j$?

- Qual è la combinazione di fattori che fornisce la risposta media $\mathbf{Y}$ più elevata?

- Subordinatamente al fatto che alcuni effetti siano significativamente diversi da zero, quali hanno condotto alla significatività?

- Una o più fra le variabili $\mathbf{X}_j$ può essere eliminata, senza che questo riduca in modo sostanziale la spiegazione della variabile di risposta? Eliminare una variabile esplicativa $\mathbf{X}_j$ dal modello corrisponde ad ipotizzare $\beta_j = 0$.

- Anche se β_j è significativamente diverso da zero, può comunque convenire lavorare con un modello ridotto anche se distorto?

6.5.1 Ipotesi sulle ε

Per potere dare una risposta, anche approssimativa, ad alcune di queste domande, e quindi per la costruzione di stimatori e test, e per fare in generale inferenza (almeno muovendosi in un contesto parametrico), occorrerà fare ovviamente delle ipotesi, più o meno restrittive, sulla distribuzione di $\boldsymbol{\varepsilon}$. Questa distribuzione dipenderà in generale da un vettore di parametri $\boldsymbol{\theta}$:

$$\boldsymbol{\varepsilon} \sim \phi(\boldsymbol{\theta}).$$

E' ovvio che, anche ammettendo di conoscere la forma funzionale ϕ , occorrerà stimare il vettore di parametri $\boldsymbol{\theta}$.

...

Va tenuto presente che $\boldsymbol{\varepsilon}$ non è direttamente osservabile, come accade invece, ad esempio, quando si osserva un campione proveniente da una normale univariata di parametri incogniti μ (costante) e σ^2.

$\boldsymbol{\theta}$ svolge in generale il ruolo di parametro di disturbo.

Ovviamente il numero dei parametri incogniti θ_s non dovrà essere elevato, diversamente non sarà possibile stimarli.

ESEMPIO: se si suppone $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$ non possono essere incogniti emphtutti gli elementi della matrice di varianza e covarianza $\boldsymbol{\Sigma}$ (perchè sarebbero $n(n+1)/2$ parametri)

Quanto interagiscono la stima di θ e quella di β ? E' possibile in qualche modo verificare a posteriori la validità delle ipotesi fatte sulla distribuzione delle ε ?

Le possibili scelte verranno analizzate successivamente alla discussione sul significato della parte sistematica.

6.6 La matrice delle X

La struttura ed il metodo di scelta delle X_j, insieme con la parametrizzazione scelta determina in parte il tipo di analisi.

Sostanzialmente le X_j (tutte o alcune) possono provenire da:

studi osservazionali Questo caso si presenta quando non è possibile in generale stabilire a priori la matrice $\mathbf{X}$: si sceglieranno solo le k particolari variabili da analizzare e le n unità che costituiscono il campione. Eventualmente potremo, entro certi limiti, operare alcune trasformazioni sulle x in modo da ricondurci a schemi particolari.

esperimenti pianificati con: variabili controllabili

in cui alcune variabili (h), e tutto lo schema degli $n \times h$ valori corrispondenti della matrice $\mathbf{X}$, vengono pianificati in anticipo, per cui si stabilisce in partenza il range di valori di ciascuna variabile esplicativa e le combinazioni di valori delle

variabili esplicative che si vogliono osservare, in funzione delle risposte che si vogliono ottenere dall'esperimento. *Con un esperimento mal pianificato, in cui ad esempio non sono previste alcune combinazioni di livelli di variabili, non si potranno per esempio condurre tutti i test che si possono effettuare con dati provenienti da un esperimento ben pianificato.*

variabili note ma il cui valore non è pianificabile Ad esempio vengono selezionati alcuni soggetti in base al sesso ed alla condizione lavorativa, per cui si stabilisce in anticipo quante osservazioni fare per tutte le combinazioni sesso x condizione lavorativa mentre per le altre variabili non è possibile pianificare dei valori particolari.

Figura da inserire
ESEMPI VARI

6.6.1 Osservazioni ripetute.

Alcune delle righe della matrice $\mathbf{X}$ potrebbero essere (volutamente o per caso) replicate. Nel caso di presenza di osservazioni ripetute per ciascuna combinazione di fattori, l'analisi potrà anche dire qualcosa di più:

- sulla bontà delle assunzioni fatte sulla distribuzione degli errori
- sulla forma funzionale della relazione (se lineare o meno).
- Sulla variabilità della componente accidentale per ciascuna combinazione di fattori.

Figura da inserire AMPLIARE

citazione

Figura da inserire fig2000regr5.stg INSERIRE ESEMPIO E GRAFICO a 2D e 3D

6.6.2 Disegni fattoriali

Un disegno si dice fattoriale se vengono pianificate le osservazioni di tutte le possibili combinazioni dei livelli dei k fattori.

Pertanto se ogni fattore X_j può assumere m_j livelli ($j = 1, 2, \ldots, k$), si avranno:

$C = \prod_{j=1}^{k} m_j$ distinte combinazioni,

ciascuna delle quali può essere replicata, per ottenere la matrice $\mathbf{X}$.

Esempio:

In un esperimento farmacologico si vuole stimare l'effetto di un farmaco (tre dosi: una nulla, una media, una alta) su pazienti con una particolare patologia. Si vuole verificare anche l'effetto su pazienti sani, e vedere se il sesso del paziente influenza il tipo di risposta. Complessivamente si hanno i seguenti fattori e corrispondenti livelli:

Fattore	livelli (o modalità qualitative)
dosi di un farmaco	3 livelli quantitativi di dose
Sesso	2 livelli
Condizione sperimentale	2 livelli: malati e sani
Totale:	12 combinazioni

Le 12 possibili combinazioni sono dunque:

	DOSE	SESSO	CONDIZIONE
1	Alta	F	Sano
2	Alta	F	Malato
3	Alta	M	Sano
4	Alta	M	Malato
5	Media	F	Sano
6	Media	F	Malato
7	Media	M	Sano
8	Media	M	Malato
9	Bassa	F	Sano
10	Bassa	F	Malato
11	Bassa	M	Sano
12	Bassa	M	Malato

Se si conviene di assegnare i seguenti valori numerici:

DOSE	Valore
Alta	+1
Media	0
Bassa	-1

SESSO	Valore
M	+1
F	-1

CONDIZIONE	Valore
Sano	+1
Malato	-1

Si ottiene la seguente matrice $\mathbf{X}$ dei regressori:

	DOSE	SESSO	CONDIZIO
1	+1	+1	+1
2	+1	+1	-1
3	+1	-1	+1
4	+1	-1	-1
5	0	+1	+1
6	0	+1	-1
7	0	-1	+1
8	0	-1	-1
9	-1	+1	+1
10	-1	+1	-1
11	-1	-1	+1
12	-1	-1	-1

Se i livelli sono quantitativi ed equispaziati (come in questo esempio), l'analisi risulta ortogonale

Anche nell'esempio che segue si ha un disegno bilanciato: X1: 5 LIVELLI; X2 e X3 con 3 livelli

LIVELLI ORIGINALI			SCARTI DALLE MEDIE		
X1	X2	X3	Z1	Z2	Z3
1	0	0	-2	-1	-1
2	0	0	-1	-1	-1
3	0	0	0	-1	-1
4	0	0	1	-1	-1
5	0	0	2	-1	-1
1	1	0	-2	0	-1
2	1	0	-1	0	-1
3	1	0	0	0	-1
4	1	0	1	0	-1
5	1	0	2	0	-1
1	2	0	-2	1	-1
2	2	0	-1	1	-1
3	2	0	0	1	-1
4	2	0	1	1	-1
5	2	0	2	1	-1
1	0	1	-2	-1	0
2	0	1	-1	-1	0
3	0	1	0	-1	0
4	0	1	1	-1	0
5	0	1	2	-1	0
1	1	1	-2	0	0
2	1	1	-1	0	0
3	1	1	0	0	0
4	1	1	1	0	0
5	1	1	2	0	0
1	2	1	-2	1	0
2	2	1	-1	1	0
3	2	1	0	1	0
4	2	1	1	1	0
5	2	1	2	1	0
1	0	2	-2	-1	1
2	0	2	-1	-1	1
3	0	2	0	-1	1
4	0	2	1	-1	1
5	0	2	2	-1	1
1	1	2	-2	0	1
2	1	2	-1	0	1
3	1	2	0	0	1
4	1	2	1	0	1
5	1	2	2	0	1
1	2	2	-2	1	1
2	2	2	-1	1	1
3	2	2	0	1	1
4	2	2	1	1	1
5	2	2	2	1	1

$$\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \begin{pmatrix} 90 & 0 & 0 \\ 0 & 30 & 0 \\ 0 & 0 & 30 \end{pmatrix}$$

link o riferimento
(vedere anche più avanti)

- Anche se non si riesce a costruire un disegno fattoriale, perché troppo oneroso, sarà opportuno di solito ricorrere a disegni ortogonali , ossia schemi di disegni sperimentali con variabili indipendenti non correlate.
- L'opportunità di avere l'ortogonalità dei fattori (ossia variabili non correlate) è pienamente giustificata solo nell'ambito della teoria normale completa sui minimi quadrati.
- Comunque è ragionevole fare in modo che i fattori non siano correlati (se possibile).
- In un esperimento a molti fattori sarà opportuno che siano bilanciate in corrispondenza a ciascuna coppia di fattori, le possibili combinazioni di coppie di livelli.

ESEMPIO di DISEGNO FATTORIALE completo E INCOMPLETO Figura da inserire
FATTORIALI12.bmp
FATTOR2.STG

link o riferimento
(vedere anche → esempi e grafici qualitativi e quantitativi)

Disegni 2^k

Un caso particolare di disegno fattoriale si ha nel caso di k fattori qualitativi dicotomici, per cui le variabili assumeranno il valore 1 o 0 secondo che la caratteristica è presente o assente; è conveniente anche utilizzare i valori 1 e -1, in modo che in un piano fattoriale completo le variabili risulteranno centrate (ossia con media nulla) e con varianza unitaria.

Per esaminare tutte le combinazioni (senza repliche) occorre prevedere 2^k osservazioni.

Esempio Disegno fattoriale completo 4 fattori a due livelli -1,1.
farmaco si/no;
sesso M/F;
malato si/no;
ospedalizzato si/no;

Si ottiene una matrice (centrata, ossia con medie nulle) con $16 = 2^4$ righe:

	Z1	Z2	Z3	Z4
1	1	1	1	1
2	1	1	1	-1
3	1	1	-1	1
4	1	1	-1	-1
5	1	-1	1	1
6	1	-1	1	-1
7	1	-1	-1	1
8	1	-1	-1	-1
9	-1	1	1	1
10	-1	1	1	-1
11	-1	1	-1	1
12	-1	1	-1	-1
13	-1	-1	1	1
14	-1	-1	1	-1
15	-1	-1	-1	1
16	-1	-1	-1	-1

$$\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \begin{pmatrix} 16 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 \\ 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 16 \end{pmatrix}$$

6.6.3 Regressione multipla.

L'informazione campionaria, relativa a n unità, è costituita da:

- Un vettore di n valori osservati $\mathbf{y}$ della variabile di risposta quantitativa $\mathbf{Y}$.

- La matrice $\mathbf{X}$ (n righe e k colonne) è data dai valori di k regressori quantitativi, noti, per ciascuna delle n osservazioni

Figura da inserire blocchi

$$\mathbf{y}_{[n\times 1]}, \mathbf{X}_{[n\times k]}$$

Le n unità osservate sono quindi costituite da $k+1$ variabili e sono schematizzabili nelle n righe:

$$(\mathbf{y}|\mathbf{X}) = \left(\begin{array}{c|cccc} \mathbf{y}_1 & x_{11} & x_{12} & \dots & x_{1k} \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline \mathbf{y}_i & x_{i1} & x_{i2} & \dots & x_{ik} \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline \mathbf{y}_n & x_{n1} & x_{n2} & \dots & x_{nk} \end{array}\right)$$

La dipendenza (del valore atteso) di $\mathbf{y}$ dalle X_j è espressa quindi dalla relazione:

$$\mathrm{E}\left[\mathbf{y}_i\right] = \alpha + \sum_{j=1}^{k} x_{ij}\beta_j$$

abbiamo quindi k coefficienti di regressione incogniti β_j che esprimono la dipendenza media (parziale) della risposta da ciascun regressore.

In generale nel modello si considera anche un termine noto incognito α , che esprime la risposta media corrispondente a valori nulli dei regressori;

α di solito non è oggetto di particolare interesse ed usualmente svolge il ruolo di parametro di disturbo.

La relazione è analoga, almeno formalmente, alla relazione di regressione lineare che studia la dipendenza della speranza matematica di una variabile aleatoria rispetto ai valori (fissati!) di altre k variabili aleatorie.

Non si confonda la regressione multipla (una variabile di risposta e molti regressori) con la regressione multivariata (molte variabili di risposta e uno o più regressori).

Relazione di regressione in termini di scarti

Per comodità interpretativa, e per motivi più tecnici che si vedranno al momento di affrontare i problemi di stima, convenzionalmente si può porre:

la prima colonna ($j = 0$) composta tutta da 1 (in modo da prevedere la presenza di un termine noto);

le altre colonne costituite dagli scarti semplici rispetto alla media di ciascuna variabile.

Con la posizione:

$$z_{ij} = x_{ij} - M(\mathbf{X}_j) \qquad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, k$$

la matrice $\mathbf{X}$ può essere messa nella forma più conveniente:

$$\mathbf{X} = \begin{pmatrix} 1 & z_{11} & \ldots & z_{1j} & \ldots & z_{1k} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & z_{i1} & \ldots & z_{ij} & \ldots & z_{ik} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & z_{n1} & \ldots & z_{nj} & \ldots & z_{nk} \end{pmatrix}$$

$$\begin{pmatrix} \text{Media variabile:} & 1 & 0 & \ldots & 0 & \ldots & 0 \end{pmatrix}$$

Per i parametri si ha:

$$\boldsymbol{\beta}^{\mathrm{T}} = \{\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_k\}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \text{Termine noto} \\ \beta_1 & \text{Coefficiente di regressione parziale variabile 1} \\ \ldots & \ldots\ldots \\ \beta_j & \text{variabile j} \\ \ldots & \ldots\ldots \\ \ldots & \ldots\ldots \\ \beta_k & \text{variabile k} \end{pmatrix}$$

Quindi la matrice dei regressori e il vettore dei coefficienti risultano partizionati in:

$$\mathbf{X} = [\mathbf{1_n}|\mathbf{Z}]$$

$$\boldsymbol{\beta}^{\mathrm{T}} = [\beta_0|\beta_{1,k}]$$

Il legame lineare è ora dato da:

$$E(\mathbf{y}_i) = \sum_{j=0}^{k} z_{ij}\beta_j$$

Per cui la risposta viene vista come somma di:

- un effetto generale, β_0 , corrispondente a livelli nulli degli scarti z_{ij} , e quindi a livelli medi dei regressori originari x_{ij}
- k singoli effetti proporzionali agli scarti dei singoli regressori dalla propria media.

Dal punto di vista interpretativo, la riscrittura in termini di scarti consente di dare un significato logico, ed utile per i confronti, al termine noto.

Rispetto alla parametrizzazione originaria si ha:

$$E(\mathbf{y}_i) = \sum_{j=0}^{k} z_{ij}\beta_j = \beta_0 + \sum_{j=1}^{k} z_{ij}\beta_j = \beta_0 - \sum_{j=1}^{k} M(\mathbf{X}_j)\beta_j + \sum_{j=1}^{k} x_{ij}\beta_j$$

Quindi:

i coefficienti di regressione sono sempre uguali (si sono solo effettuate delle traslazioni di assi!)

Per il termine noto:

$$\alpha = \beta_0 - \sum_{j=1}^{k} M(\mathbf{X}_j)\beta_j$$

L'utilità teorica e pratica di queste posizioni sarà chiarita nella parte relativa all'inferenza nella regressione lineare. In ogni caso continuerò ad indicare la matrice del disegno o dei regressori con $\mathbf{X}$, precisando eventualmente se si tratta di scarti o di variabili originarie.

L' ipotesi nulla che più spesso si vuole verificare (almeno preliminarmente) è:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0; con \beta_0 qualsiasi.$$

Ossia che il valore atteso della variabile dipendente sia costante ed indipendente dai regressori.

Figura da inserire ESEMPIO

6.6.4 Regressione polinomiale:

Dal momento che la linearità va intesa rispetto ai parametri, e non rispetto alle $\mathbf{X}_j$, il modello lineare comprende anche la regressione polinomiale in una o più variabili:

Regressione polinomiale di grado k in un regressore Z se

$$\mathrm{E}\left[\mathbf{y}_i\right] = \sum_{j=0}^{k} \beta_j z_i^j; \qquad i = 1, 2, \ldots, n$$

Ci si riporta al caso generale del modello lineare ponendo:

$$x_{ij} = z_i^j \beta_j; \qquad i = 1, 2, \ldots, n; \quad j = 0, 1, \ldots, k.$$

Anche in questo caso si continua a parlare di modelli lineari, pochè il termine lineare si riferisce sempre ai parametri e non ai regressori.

Si noti come però i regressori risultino in generale correlati, a meno che non si faccia ricorso a particolari trasformazioni del modello polinomiale basate sui polinomi ortogonali.

Figura da inserire esempio

Polinomi in più variabili e superfici di risposta

E' immediata la generalizzazione alle superfici polinomiali di grado k in p regressori.

Regressione polinomiale di grado k in p regressori Z_h

$$E(\mathbf{y}_i) = \sum_{j=0}^{k} \cdots \sum_{j=0}^{k} \beta_{j1,j2,\ldots,jp} \prod_{\sum j_h = j} (z_{ih})^{jh}; i = 1, 2, \ldots, n$$

In particolare se $k = 2$ e se i coefficienti dei termini di secondo grado in ciascun regressore sono nulli, si possono convenientemente quantificare ed inserire nel modello degli effetti di interazione moltiplicativi del tipo $\beta_{hr} z_{ih} z_{ir}$ (interazione del primo ordine fra il regressore r-esimo ed h-esimo; Termini moltiplicativi che coinvolgono k regressori sono relativi ad effetti di interazione di grado $k - 1$

link o riferimento
(vedere anche più avanti)

Figura da inserire

FIG2000REGPOLIN1.STG

Figura da inserire
FIG2000REGSPLINE1.STG

Regressione parametrica e non parametrica

In questo corso ci occuperemo prevalentemente di regressione parametrica, ossia modelli di dipendenza nei quali è specificata la forma di dipendenza funzionale della variabile di risposta o meglio della sua speranza matematica, dalle variabili esplicative.

In effetti, di solito supponiamo anche che sia nota la forma distribuzionale della componente accidentale, a meno di qualche parametro di disturbo (per esempio nella regressione lineare semplice supponiamo usualmente che gli errori siano distribuiti normalmente con varianza uguale ma incognita).

Nella regressione non parametrica invece, si evita il più possibile di fare delle ipotesi in merito alla forma funzionale della dipendenza; queste tecniche, che non affronteremo in modo particolare nel nostro corso, sono tipiche di una fase esplorativa dell'analisi dei dati quando non si sa, almeno con buona approssimazione, qual è la forma della relazione che lega la variabile dipendente al regressore.

Sostanzialmente si cerca direttamente di approssimare la funzione di regressione localmente, per ciascun valore di x:

$$\hat{y}(x) \approx \mathrm{E}\,[y|x]$$

evidentemente nel caso in cui si ha una sola variabile esplicativa il modo più conveniente di ottenere informazioni sul tipo di relazione è quello di effettuare una analisi grafica; chiaramente questo strumento è disponibile anche nel caso di due variabili esplicative.

Un caso molto comodo si ha per esempio quando sono disponibili n_j osservazioni ripetute in corrispondenza dello stesso valore di x_j: in questo caso, infatti, la linea spezzata che congiunge le medie aritmetiche della variabile di risposta in corrispondenza dei diversi valori della x_j , costituisce una base per la stima della vera relazione funzionale fra la speranza matematica della variabile risposta e la variabile esplicativa (o le variabili esplicative).

$$\hat{y}(x_j) = \frac{\sum_{i=1}^{n_j} y_{ij}}{n_j}$$

Nel caso in cui non si abbiano osservazioni ripetute per la stessa variabile esplicativa, sarà necessario ricorrere ad approssimazioni

analitiche: alcune delle tecniche si basano su opportune generalizzazioni di tipi di medie mobili o su adattamenti mediante particolari funzioni kernel; un metodo molto generale, senza bisogno di entrare in grande dettaglio, è dato da una media ponderata delle y_i:

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} w(x_i - x) y_i}{w(x_i - x)}$$

ove i pesi $w(x_i - x)$ sono delle funzioni decrescenti di $x_i - x$; ad esempio:

$$w(x_i - x) = e^{-\frac{(x_i - x)^2}{2h^2}}$$

essendo h un parametro di lisciamento.

Se si cercano approssimazioni sufficientemente regolari uno strumento tecnico molto utile è costituito dalle funzioni splines, particolarmente utili sia nel caso univariato sia nel caso bivariato.

Le funzioni splines sono delle particolari funzioni ottenute dalla composizione di r segmenti di polinomi $f_j(x)$, $j = 1, 2, \ldots, r$ in modo tale che la curva risulti sufficientemente liscia e regolare senza punti di discontinuità in corrispondenza dei cambi di segmento.

Uno degli approcci per trovare i parametri dei segmenti di polinomio (se $r = n$) consiste nell'imporre alcuni vincoli alle funzioni e ad alcune loro derivate in corrispondenza ai punti d'incontro dei segmenti (nodi), z_j, $\quad j = 1, 2, \ldots, r$:

$$f_j(z_j) = f_{j+1}(z_j); \quad f_j(z_j) = f_{j+1}(z_j); \quad f_j(z_j) = f_{j+1}(z_j); \quad j = 1, 2, \ldots, r$$

Sufficienti requisiti di regolarità si ottengono operando con segmenti di polinomi di 3° grado.

Un altro approccio consiste nel cercare una curva composta da segmenti polinomiali che risulti adattarsi abbastanza bene ai dati (con $r < n$) mantenendo comunque una regolarità della curvatura della curva complessiva.

In ogni caso queste tecniche di regressione non parametrica sono suscettibili di applicazione:

-nella fase esplorativa della ricerca di una relazione di dipendenza fra variabili

-oppure a scopo interpolatorio, quando un'approssimazione polinomiale localmente regolare è preferibile ad una relazione lineare o comunque ad una relazione che sia della stessa forma e con gli stessi parametri in tutto il campo di variazione della $\mathbf{X}$.

6.6.5 Regressori del tipo 0/1 (dummy variables)

Esiste un modo formale di esplicitare la matrice $\mathbf{X}$ in modo da trattare anche variabili esplicative di tipo qualitativo. Vediamo come prima con un esempio relativo ad una situazione nota.

Si supponga la situazione classica del confronto delle medie μ_1 e μ_2 di due popolazioni normali con uguale varianza sulla base delle informazioni di due campioni casuali semplici indipendenti.

Per la speranza matematica della variabile casuale associata alla generica osservazione abbiamo:

$$E(\mathbf{Y}_i) = \mu_j \text{per} j = 1, 2,$$

secondo se l'unità i-esima appartiene al primo o al secondo campione.

Possiamo indicare sinteticamente:

$$E(\mathbf{Y}_i) = x_{i1}\mu_1 + x_{i2}\mu_2$$

introducendo due regressori con la convenzione che per le unità del primo campione si ha: $x_{i1} = 1$ e $x_{i2} = 0$, per le unità del secondo campione si ha invece: $x_{i1} = 0$ e $x_{i2} = 1$.

Oppure si può parametrizzare con:

$$E(\mathbf{Y}_i) = \mu_1 + x_{i2}(\mu_2 - \mu_1)$$

e l'ipotesi da verificare sarà:

$$H_0 : \delta = (\mu_2 - \mu_1) = 0$$

con μ_1 qualsiasi.

(oppure si vorranno costruire intervalli di confidenza per δ)

L'aspetto essenziale di questo esempio è che anche questa situazione standard è riconducibile ad un modello lineare.

Esempio: Si hanno due campioni indipendenti di 14 osservazioni relative ad una variabile quantitativa, suddivise in due gruppi A e B, rispettivamente di numerosità 6 e 8.

A	2; 3; 3,1; 4; 5; 5,3.
B	3; 4,1; 4,3; 4,8; 6, 6,5; 7; 7,2.

Potremmo pensare di avere rilevato 3 variabili su 14 individui nel modo che segue:

$\mathbf{y}$	$\mathbf{x_A}$	$\mathbf{x_B}$
2	1	0
3	1	0
3,1	1	0
4	1	0
5	1	0
5,3	1	0
3	0	1
4,1	0	1
4,3	0	1
4,8	0	1
6	0	1
6,5	0	1
7	0	1
7,2	0	1

Sarà bene che da ora in poi lo studente si abitui a questa impostazione, in particolare per problemi con più variabili, perché riuslta estremamente comoda in particolare per le situazioni complesse; (per la situazione dell'esempio, ossia test t a due campioni, non v'è alcun motivo pratico di ricorrere a tale formulazione, perché l'impostazione standard è quella più uitile)

6.6.6 Analisi della varianza ad effetti fissi ed un criterio di classificazione

La versatilità del modello lineare, almeno da un punto di vista formale, si coglie per situazioni apparentemente lontane da quelle della

regressione multipla, ossia per lo studio della dipendenza in media di una variabile quantitativa da una qualitativa (o più variabili qualitative).

Si supponga di avere n osservazioni suddivise in k gruppi indipendenti secondo le k modalità di un criterio di classificazione semplice (o mutabile sconnessa).

Si suppone che i gruppi siano internamente omogenei, ma che le medie dei gruppi possano essere in generale diverse:

$$E(\mathbf{Y}_i) = \mu_j$$

La matrice $\mathbf{X}$ è ora composta da k colonne costituite dagli n indicatori dell'appartenenza delle unità ai gruppi:

(MATRICE del disegno sperimentale)

i		$Gr.1$	$Gr.2$	$\ldots$	$Gr.J$	$\ldots$	$Gr.K$
1		1	0		0		0
$\ldots$		$\ldots$	0	$\ldots$	$\ldots$	$\ldots$	0
n_1		1	0	$\ldots$	$\ldots$	$\ldots$	0
		0	1	$\ldots$	$\ldots$	$\ldots$	0
		0	$\ldots$	$\ldots$	$\ldots$	$\ldots$	0
$n_1 + n_2$		0	1	$\ldots$	$\ldots$	$\ldots$	0
	$\mathbf{X} =$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
		0	0	$\ldots$	1	$\ldots$	0
$n_1 + n_2 + \ldots + n_j$		$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
		0	0	$\ldots$	0	$\ldots$	1
		0	0	$\ldots$	0	$\ldots$	$\ldots$
$n_1 + n_2 + \ldots + n_k$		0	0	$\ldots$	0	$\ldots$	1

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & & 0 & & 0 & 1 \\ \dots & 0 & \dots & \dots & \dots & 0 & \dots \\ 1 & 0 & \dots & \dots & \dots & 0 & n_1 \\ 0 & 1 & \dots & \dots & \dots & 0 & \\ 0 & \dots & \dots & \dots & \dots & 0 & \\ 0 & 1 & \dots & \dots & \dots & 0 & n_1 + n_2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \\ 0 & 0 & \dots & 1 & \dots & 0 & \\ \dots & \dots & \dots & \dots & \dots & \dots & n_1 + n_2 + \ldots + n_j \\ 0 & 0 & \dots & 0 & \dots & 1 & \\ 0 & 0 & \dots & 0 & \dots & \dots & \\ 0 & 0 & \dots & 0 & \dots & 1 & n_1 + n_2 + \ldots + n_k \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \dots \\ \mu_j \\ \dots \\ \mu_k \end{pmatrix}$$

Si ha:

n_j osservazioni per ogni trattamento o gruppo:

$$n_j = \sum_{i=1}^{n} x_{ij}; j = 1, 2, \ldots k.$$

ogni unità $\mathbf{U}_i$ appartiene ad un solo trattamento:

$$\sum_{j=1}^{k} x_{ij} = 1; i = 1, 2, \ldots, n$$

$x_{ij} = 1$ se e solo se l'unità $\mathbf{U}_i$ appartiene al j -esimo trattamento

$$\boldsymbol{\beta}^{\mathrm{T}} = \mu_1, \ldots, \mu_j, \ldots, \mu_k$$

L' ipotesi nulla di interesse è di solito quella di omogeneità:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$$

oppure

$$H_0 : \beta_1 - \beta_k = \beta_2 - \beta_k = \ldots = \beta_{k-1} - \beta_k = 0$$

Con questa parametrizzazione $\mathbf{X}$ ha rango pieno k , ma l'ipotesi nulla di omegenità far le medie impone $k - 1$ vincoli

Altro modo di impostare l' analisi della varianza a una via:

$\beta_j = \mu_j - \mu$	effetto del trattamento (o del gruppo) $j; j = 1, 2, \ldots k.$
$\beta_{k+1} = \mu$	media generale;

e stavolta la matrice del disegno è:

i		eff. gr.1	eff. gr.2	...	eff. gr.j	...	eff. gr.k	effetto g
1		1	0		0		0	1
...		...	...	...	...	...	...	1
n_1		1	0	...	...	...	0	1
		0	1	...	...	...	0	1
		0	...	...	...	...	0	..
$n_1 + n_2$		0	1	...	...	...	0	1
	$\mathbf{X} =$	...	...	...	...	...	...	..
		0	0	...	1	...	0	1
$n_1 + n_2 + \ldots + n_j$		...	...	...	...	...	...	..
		0	0	...	0	...	1	1
		0	0	...	0	...	...	..
$n_1 + n_2 + \ldots + n_k$		0	0	...	0	...	1	1

$$\mathbf{X} = \begin{pmatrix} \text{eff. gr.1} & \text{eff. gr.2} & \ldots & \text{eff. gr.}j & \ldots & \text{eff. gr.}k & \text{effetto generale} \\ 1 & 0 & & 0 & & 0 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 1 \\ 1 & 0 & \ldots & \ldots & \ldots & 0 & 1 \\ 0 & 1 & \ldots & \ldots & \ldots & 0 & 1 \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 & \ldots \\ 0 & 1 & \ldots & \ldots & \ldots & 0 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & \ldots & 0 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & \ldots & 1 & 1 \\ 0 & 0 & \ldots & 0 & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & \ldots & 1 & 1 \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu_1 - \mu \\ \dots \\ \mu_j - \mu = \\ \dots \\ \mu_k - \mu \\ \mu \end{pmatrix}$$

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ e μ qualsiasi (k vincoli)

In questo caso però $\mathbf{x}$ ha una colonna linearmente dipendente dalle altre, per cui ha rango k invece di $k + 1$.

6.6.7 Analisi della varianza ad effetti fissi con due criteri di classificazione

E' possibile estendere il disegno precedente all'analisi della varianza a due vie, per la quale si può impostare un modello lineare con rs colonne, con:

$\mathbf{X}_{ijm} = 1$ se $\mathbf{U}_i$ appartiene al j -esimo trattamento di riga e all' m -esimo trattamento di colonna.

Oppure si può partire da una matrice del disegno sperimentale semplificata con $r + s + 1$ colonne x e Z , tali che:

$x_{i0} = 1$ effetto generale;

$x_{ij} = 1$ se $\mathbf{U}_i$ appartiene al j -esimo trattamento di riga

$z_{im} = 1$ se $\mathbf{U}_i$ appartiene all' m -esimo trattamento di colonna

e introdurre nel modello di descrizione dei dati dei termini moltiplicativi (che saranno 1 solo se $\mathbf{U}_i$ appartiene ad una riga e ad una colonna) per considerare l'effetto di interazione:

$$\mathbf{y}_{ijk} = \beta_0 + \sum_{j=1}^{r} \alpha_j x_{ij} + \sum_{m=1}^{s} \eta_m z_{imj} + \sum_{j=1}^{r} \sum_{m=1}^{s} \gamma_{jm} x_{ij} z_{im} + \varepsilon_{ijk}$$

In pratica si considerano le due matrici di appartenza ai gruppi per i due criteri di classificazione separatamente; se nel modello occorre tener conto dell'appartenza simultanea (termini di interazione) si farà riferimento ai termini moltiplicativi $x_{ij} z_{im}$, che sono uguali ad 1 solo per le unità che appartengono alla modalità j -esima del primo criterio di classificazione ed alla modalità m -esima del secondo criterio di classificazione.

Le ipotesi da verificare sono quelle usuali (si vedranno in dettaglio nella parte inferenziali relativa all'analisi della varianza a due

vie); con questa parametrizzazione però, peraltro molto comoda e naturale, il modello ha parametri ridondanti (rango $= rs$; parametri $1 + r + s + rs$).

In modo analogo si possono impostare modelli a più vie.

Figura da inserire ESEMPIO

6.6.8 Analisi della covarianza

(L'utilità dell'analisi della covarianza verrà esaminata più avanti)

Supponendo di avere n osservazioni suddivise in k gruppi secondo un criterio di classificazione semplice e relative ad una variabile di risposta $\mathbf{y}$ e ad una singola variabile concomitante x ci si può ricondurre al modello lineare generale ponendo:

$$z_{ij} = x_{ij} - M_j(x) j = 1, 2, \ldots, k$$

ove $M_j(x)$ è la media di x per le sole osservazioni del gruppo j .

La matrice $\mathbf{X}$ sarà composta da $2k$ colonne, di cui le prime k sono date da:

$$\mathbf{X}_1 = \begin{pmatrix} z_{1,1} & \ldots & 0 \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ z_{n1,1} & \ldots & 0 \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & z_{ij} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \ldots & z_{n1,k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \ldots & z_{nk,k} \end{pmatrix}$$

mentre le altre k colonne sono costituite dalla matrice di appar-

tenenza ai gruppi:

$$\mathbf{X}_2 = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \dots & 0 & \dots & \dots & \dots & 0 \\ 1 & 0 & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{pmatrix}$$

per cui la matrice $\mathbf{X}$ è costituita dalle colonne di $\mathbf{X}_1$ e $\mathbf{X}_2$ affiancate ossia:

$$\mathbf{X} = \mathbf{X}_1 | \mathbf{X}_2,$$

e i $2k$ parametri sono:

$$\boldsymbol{\beta}^{\mathrm{T}} = \beta_1, \dots, \beta_j, \dots, \beta_k, \alpha_1, \dots, \alpha_j, \dots, \alpha_k$$

Ipotesi di interesse:

$$H_0 : \beta_1 = \dots = \beta_j = \dots = \beta_k; \alpha_1 = \dots = \alpha = \dots = \alpha_k$$

con β_1 , α_1 qualsiasi ($2k - 2$ vincoli)

rette di regressione uguali nei k gruppi.

In generale si possono costruire disegni più complessi, con più variabili concomitanti e con più regressori, considerando un modello lineare del tipo:

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\varepsilon}$$

in cui:

$\mathbf{X}_1$ è una matrice a più regressori,

$\mathbf{X}_2$ è una matrice di indicatori per più criteri di classificazione,

β_1 è il vettore dei parametri che esprimono la dipendenza della variabile di risposta dalle variabili concomitanti

β_2 è il vettore dei parametri che esprimono la dipendenza della variabile di risposta dai fattori di classificazione.

6.6.9 Rette o piani di regressione con pendenze diverse: termini polinomiali moltiplicativi

Una relazione polinomiale con termini lineari e termini misti di 2° grado può esprimere la presenza di effetti di interazione in un modello lineare:

Esempio 1:

Si supponga una dipendenza in media della risposta $\mathbf{y}$ da due fattori quantitativi secondo la relazione:

$$E(\mathbf{y}_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i1}x_{i2}\beta_{12}$$

Se il parametro β_{12} fosse uguale a zero avremmo un classico piano di regressione: $E(\mathbf{y}_i) = x_{i1}\beta_1 + x_{i2}\beta_2$, in cui parametri sarebbero interpretabili nel modo già visto (modello additivo).

Se invece tale parametro è diverso da zero, è presente un effetto di interazione fra i regressori $\mathbf{X}_1$ e $\mathbf{X}_2$: infatti per esempio la dipendenza di $\mathbf{y}$ dal regressore $\mathbf{X}_1$, per ciascuno dei possibili livelli di $\mathbf{X}_2$, è sempre lineare, ma l'inclinazione, e quindi la forza della dipendenza di $\mathbf{y}$ da $\mathbf{X}_1$, dipendono dal particolare livello assunto da $\mathbf{X}_2$. Il parametro β_1 non misura più la dipendenza parziale di $\mathbf{y}$ da $\mathbf{X}_1$, per qualsiasi livello di $\mathbf{X}_2$, ma solo la dipendenza media rispetto a tutti i livelli di $\mathbf{X}_2$.

Esempio di polinomio di secondo grado per effetto interazione: Supponiamo per esempio

$$\beta_1 = 1, \qquad \beta_2 = 3, \qquad \beta_{12} = 2,$$

per cui:

$$\mathrm{E}\left[\mathbf{y}_i\right] = x_{i1}1 + x_{i2}3 + x_{i1}x_{i2}2$$

disegni ortogonali

L'effetto interazione fra $\mathbf{X}_1$ e $\mathbf{X}_2$ è tale da modificare anche il tipo di dipendenza di $\mathbf{y}$ da $\mathbf{X}_1$ (da negativa a positiva)

Si vedano nel grafico seguente le tre rette di regressione ottenute per tre diversi valori di $\mathbf{X}_2$ (-1;0;+1)

Esempio 2: (confronto fra due rette) Pendenza diversa come effetto interazione fra un fattore (o regressore) quantitativo e un fattore qualitativo:

Si supponga che la relazione di una risposta $\mathbf{y}$ da un regressore $\mathbf{X}_1$ dipenda anche da una variabile dicotomica: In questo caso la differenza di pendenza può essere inserita nel modello lineare mediante l'introduzione di un termine moltiplicativo, che non altera la linearità delle relazioni parziali, ma consente l'interpretazione dell'interazione fra i due fattori. ($\mathbf{X}_1$ può essere formato da un gruppo di regressori: l'esempio resta sostanzialmente inalterato) Per semplicità possiamo considerare la variabile dicotomica $\mathbf{X}_2$ con due livelli:-1 e +1, per cui ci riportiamo formalmente al caso precedente:

$$E(\mathbf{y}_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i1}x_{i2}\beta_{12} =$$

$$= (\beta_0 + x_{i2}\beta_2) + x_{i1}(\beta_1 + x_{i2}\beta_{12})$$

e quindi:

$$\mathrm{E}\left[\mathbf{y}_i\right] = \begin{cases} (\beta_0 - \beta_2) + x_{i1}(\beta_1 - \beta_{12}) & sex_{i2} = -1 \\ (\beta_0 + \beta_2) + x_{i1}(\beta_1 + \beta_{12}) & sex_{i2} = +1 \end{cases}$$

Da cui risulta evidente, ed utile da un punto di vista interpretativo, che β_2 rappresenta un effetto (medio) del fattore $\mathbf{X}_2$ sul livello medio di $\mathbf{y}_i$, mentre β_{12} rappresenta l'effetto (medio) del fattore $\mathbf{X}_2$ sulla relazione fra $\mathbf{y}$ e $\mathbf{X}_1$, per cui rappresenta un *effetto di interazione* (di primo ordine).

Risulta quindi irrilevante o comunque poco interessante dal punto di vista pratico, con questa interpretazione dei parametri, un test costruito per la verifica dell'ipotesi: $H_0 : \beta_1 = 0$, perché questo misurerebbe l'effetto marginale del primo regressore, senza tenere conto del livello dell'altro regressore (o meglio per un livello nullo, o medio, del secondo fattore). Se per esempio il fattore dicotomico $\mathbf{X}_2$ fosse il sesso (M=-1;F=+1), tale effetto marginale sarebbe di nessun interesse, perché ogni soggetto sarà o M o F, e quindi anche se risultasse $\beta_1 = 0$, in effetti la dipendenza della risposta dal regressore $\mathbf{X}_1$ sarebbe $-\beta_{12}$ per i maschi e $+\beta_{12}$ per le femmine. Eventualmente occorrerebbe prima saggiare l'ipotesi: $H_0 : \beta_{12} = 0$

Termini moltiplicativi con più termini possono servire per quantificare effetti di interazione di ordine superiore al primo.

Abbiamo già fatto cenno a questo argomento quando abbiamo parlato di distribuzioni condizionate nella normale multivariata; ricordo infatti che in una distribuzione normale multivariata la correlazione fra due variabili condizionata ai valori singoli di un'altra variabile o di più variabili è sempre la stessa, indipendentemente dai livelli assunti dalla III variabile. In altri termini nella distribuzione normale multivariata si è già visto che la dipendenza di $\mathbf{y}$ da x non varia in funzione dei livelli di una terza variabile z : questo è analogo al concetto di assenza di interazione, con l'avvertenza che in effetti il concetto di interazione può essere introdotto senza la necessità di riferirsi ad un modello probabilistico multivariato.

Esempio di piano fattoriale 2^k

Supponendo di avere quattro fattori dicotomici $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ e $\mathbf{X}_4$ con livelli standardizzati -1 e 1, (vedere paragrafo sui disegni fattoriali), il modello seguente:

$$\mathrm{E}[\mathbf{y}_i] = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 +$$

$$+ x_{i1}x_{i2}\beta_{12} + x_{i1}x_{i3}\beta_{13} + x_{i2}x_{i3}\beta_{23} +$$

$$+ x_{i1}x_{i2}x_{i3}\beta_{123}.$$

esprime una dipendenza della risposta dai livelli dei quattro regressori; mentre il fattore $\mathbf{X}_4$ non interagisce con nessun altro fattore, gli altri 3 fattori interagiscono sia presi a due a due (interazioni di primo ordine) che tutti e tre insieme (interazione di secondo ordine). Dal punto di vista interpretativo: *l'effetto del 4° fattore è separabile* rispetto a tutti gli altri; l'effetto degli altri 3 invece non è separabile neanche a coppie.

6.6.10 Modelli autoregressivi

Un caso speciale è costituito dall'osservazione di una serie temporale, cioè si dispone di n osservazioni eseguite ad intervalli di tempo uguali.

Si può pensare, in assenza di informazioni esterne o comunque di altre variabili, di volere studiare la dipendenza della serie dalla stessa serie spostata di uno o più unità temporali; in pratica si ipotizza che Y_t, osservazione al tempo t, o meglio, la sua speranza matematica $\mathrm{E}[Y_t]$, dipenda linearmente dall'osservazione precedente y_{t-1}.

Supponiamo quindi di volere spiegare la variabilità di una serie mediante i soli valori della serie stessa in tempi precedenti; sarà in realtà opportuno fare delle ipotesi sul processo stocastico che ha generato la serie (ossia che sia stazionario), per cui la serie non ha certamente componenti di trend.

Possiamo, prima di ipotizzare particolari processi stocastici che possono avere generato la serie, adottare un approccio analogo alla regressione lineare, cercando la relazione di regressione che fa dipendere Y_t da Y_{t-1}. In pratica impostiamo un modello di regressione (detto modello autoregressivo) nel quale la serie originaria svolge il ruolo della variabile di risposta, mentre la Y_{t-1} svolge il ruolo di regressore o variabile esplicativa.

serie originaria serie arretrata di una unità temporale

$$\left\{\begin{array}{c} y_2 \\ y_3 \\ \vdots \\ y_t \\ y_{t+1} \\ \vdots \\ y_n \end{array}\right\} \qquad \left\{\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_{t-1} \\ y_t \\ \vdots \\ y_{n-1} \end{array}\right\}$$

Evidentemente questo approccio presuppone serie equiintervallate

Ovviamente la dipendenza da valori precedenti può essere estesa anche a valori distanziati di più di un intervallo temporale:

Si può proseguire il ragionamento pensando che y_t sia influenzato non solo dalla precedente determinazione y_{t-1} ma anche da y_{t-2} e dalle precedenti osservazioni fino a y_{t-k}.

serie originaria serie y_{t-1} serie y_{t-2} $\dots$ serie y_{t-k}

$$\left\{\begin{array}{c} y_{k+1} \\ y_{k+2} \\ \vdots \\ y_t \\ y_{t+1} \\ \vdots \\ y_n \end{array}\right\} \quad \left\{\begin{array}{c} y_k \\ y_{k+1} \\ \vdots \\ y_{t-1} \\ y_t \\ \vdots \\ y_{n-1} \end{array}\right\} \quad \left\{\begin{array}{c} y_{k-1} \\ y_k \\ \vdots \\ y_{t-2} \\ y_{t-1} \\ \vdots \\ y_{n-2} \end{array}\right\} \quad \cdots \quad \left\{\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_{t-k} \\ y_{t-k+1} \\ \vdots \\ y_{n-k} \end{array}\right\}$$

6.7 Generalizzazioni e modelli non lineari (cenni)

Possiamo pensare che la speranza matematica della risposta sia una funzione qualsiasi dei parametri e delle variabili indipendenti X_j:

Modello non lineare con errori additivi.

$$\mathbf{Y} = f(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \qquad \text{con:} \qquad \mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$$

$\mathbf{f}(\cdot)$ vettore di funzioni non lineari.

Modello non lineare con legame qualsiasi fra componente accidentale e sistematica.

$$\mathbf{Y} = g(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\varepsilon})$$

Modello non lineare con errori moltiplicativi.

$$\mathbf{Y}_i = f_i(\mathbf{X}_{;}\boldsymbol{\beta}) \times \boldsymbol{\varepsilon}_i$$

GLM: Generalized Linear Models Modelli Lineari Generalizzati:

$$\mathrm{E}[\mathbf{Y}] = h(\mathbf{X}\boldsymbol{\beta}) \qquad \eta(\mathrm{E}[\mathbf{Y}]) = \mathbf{X}\boldsymbol{\beta}$$

`controllare e fare anche su dispensa2003d1.tex`

La speranza matematica della variabile di risposta è funzione ($h(.)$ non lineare) del *predittore lineare* $\mathbf{X}\,\boldsymbol{\beta}$.

Si tratta ancora di modelli non lineari, ma con la particolarità che la dipendenza dalle X_j è scomposta in due parti:

- la funzione di legame (unica)
- un predittore lineare $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$

Questa impostazione consente di attribuire alla matrice $\mathbf{X}$ e al vettore di parametri $\boldsymbol{\beta}$ significati simili a quelli assunti nei modelli lineari.

Una sottoclasse di GLM molto impiegata nelle applicazioni è quella in cui la distribuzione della componente accidentale appartiene alla famiglia di distribuzioni esponenziale.

Si avrà in sostanza:

```
$$
f(\vecy_i)=
$$
```

Regressione logistica

La probabilità del verificarsi di un evento (variabile di risposta) dipende dalle variabili $\mathbf{X}_j$.

Regressione piecewise

Una relazione di regressione può essere individuata da una spezzata, ossia da una retta che cambia inclinazione in corrispondenza dei livelli delle variabili esplicative. Nel caso in cui i punti di cambio dell'inclinazione non siano noti, il problema è configurabile nell'ambito dei modelli non lineari (*non lineari rispetto ai parametri*!)

Approssimazione di modelli non lineari

Eventualmente un modello lineare può essere visto come approssimazione del primo ordine di un modello non lineare

Regressione non parametrica

La forma funzionale $f(\mathbf{X}, \boldsymbol{\beta})$ non è precisata:viene stimata direttamente $\mathrm{E}\left[\mathbf{Y}_i|\mathbf{x}_i\right]$ (in modo *non parametrico*), ed eventualmente dopo si cerca di valutare $f(.)$. Nel caso $k = 1, 2$ questo può servire come

indizio per la scelta del tipo di funzione, o per la scelta del tipo di polinomio, etc.

modello autoregressivo

Figura 6.2: FIG2000REGR3.STG

vai a indice figure

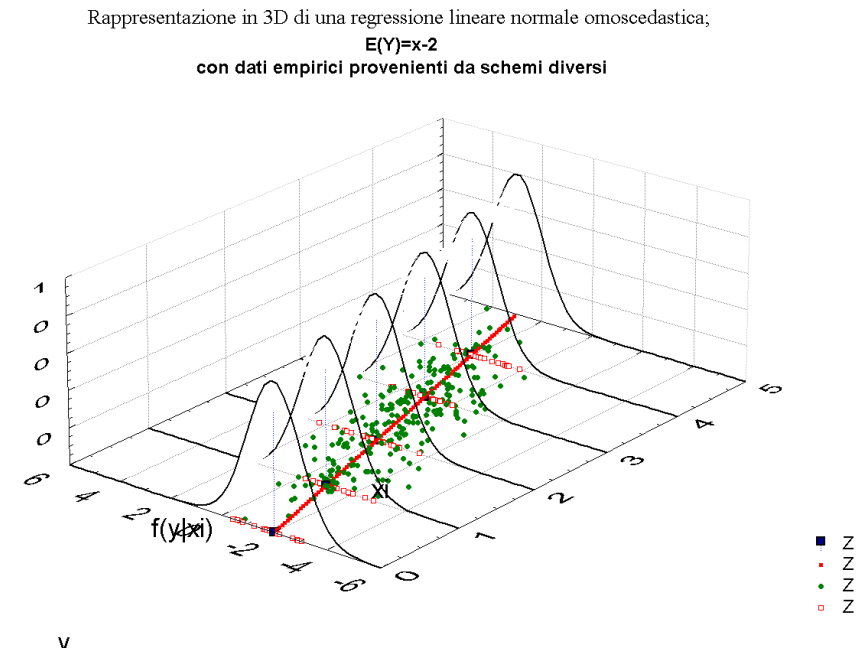

Figura 6.3: distribuzioni condizionate normali

vai a indice figure

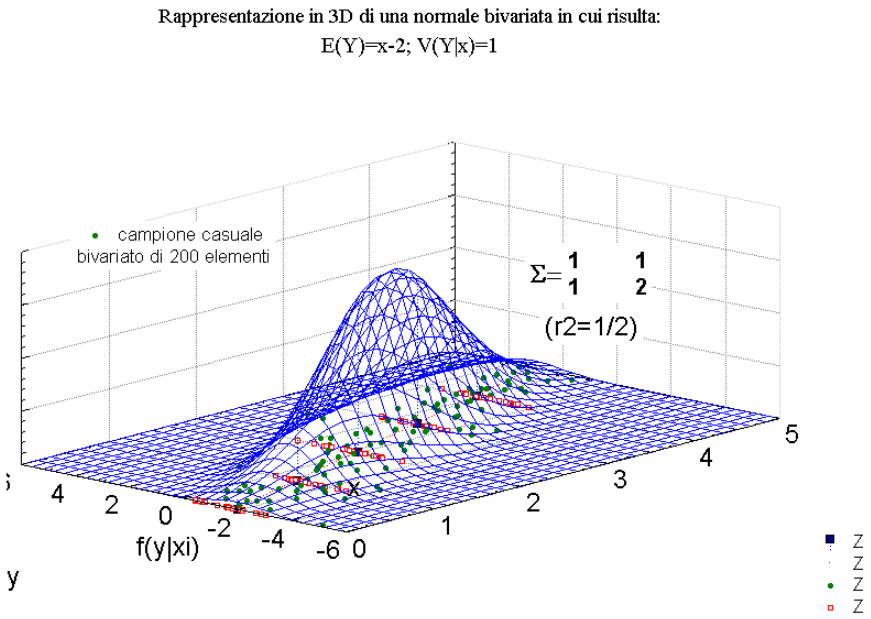

Figura 6.4: campione da una normale bivariata

vai a indice figure

Figura 6.5: distribuzioni condizionate normali in corrispondenza di valori fissati

vai a indice figure

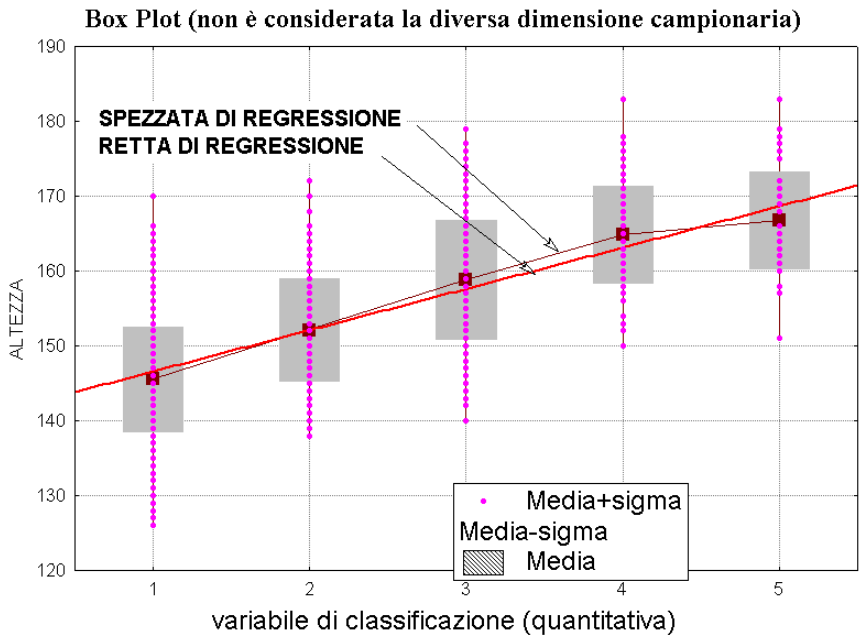

Figura 6.6: box-plot con retta di regressione e spezzata di regressione

vai a indice figure

Figura 6.7: interaz1.stg

vai a indice figure

Figura 6.8: interaz2.stg

vai a indice figure

Capitolo 7

Spunti tratti da casi reali per l'introduzione di argomenti teorici

7.1 La correlazione parziale

Si prenda in considerazione il caso relativo a dati antropometrici esposto nel grafico 3.3.

Restringiamo per semplicità per ora la nostra attenzione a tre variabili:

TORACE
ALTEZZA
PESO

e riguardiamo il grafico a matrici delle sole tre variabili che usiamo per questo esempio.

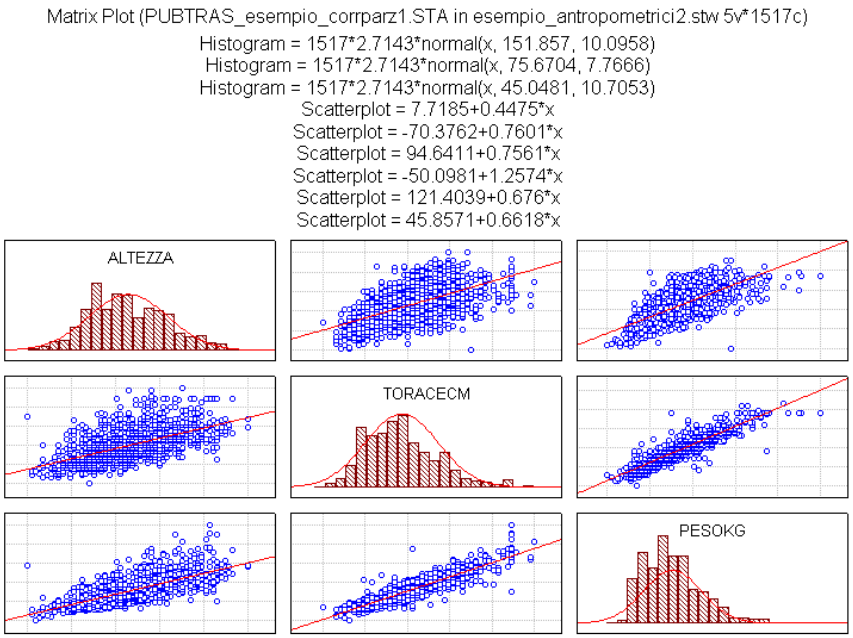

Figura 7.1: grafico a matrice delle tre variabili antropometriche

vai a indice figure

. . .

Vogliamo vedere se e come si modifica la relazione (lineare) fra due variabili, quando si vuole tenere conto dell'influenza che una terza variabile ha su di loro.
Come eliminare quest'influenza e come misurare poi la relazione?

Esaminiamo la relazione fra torace e altezza (senza considerare altre variabili).

E' una relazione crescente (prescindendo dal fatto che sia lineare o no: assumiamo per semplicità per ora di approssimare le relazioni di regressione con funzioni lineari, che nel nostro caso danno comunque una buona idea generale della relazione di regressione)

TORACECM = 7.7185+0.4475*ALTEZZA: retta di regressione lineare

r=0.58 indice di correlazione lineare semplice

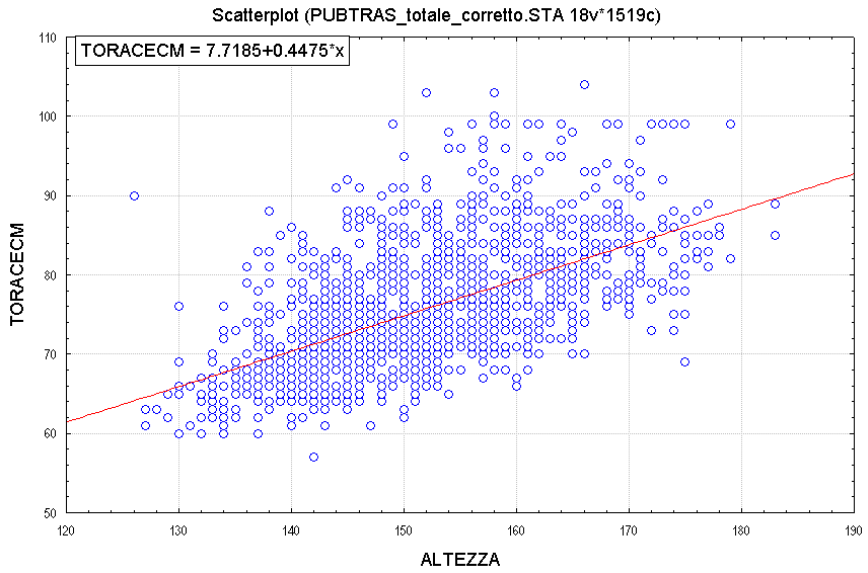

Figura 7.2: relazione fra Circonferenza toracica e altezza su 1519 ragazzi

vai a indice figure

Questa relazione non tiene conto della presenza di altre variabili.

Dal momento che si sa che esistono altre variabili che influenzano sia x che y, ci poniamo adesso una domanda un po' diversa:

che relazione esiste fra la circonferenza toracica e l'altezza a parità di altre condizioni, per ora diciamo semplicemente a parità di peso?

Oppure, che relazione esiste fra la circonferenza toracica e l'altezza dei soggetti con lo stesso peso?

Ci chiediamo: cosa succede considerando esplicitamente una terza variabile?

7.1.1 Cenno alla regressione multipla

Adesso i punti vanno rappresentati in uno spazio a tre dimensioni.

Dobbiamo adattare un piano di regressione

z=Torace

y=peso

x=altezza

Il piano di regressione

$$z = a + bx + cy$$

minimizza la somma dei quadrati degli scarti dei punti osservati

dal piano (misurati in verticale, ortogonalmente rispetto al piano xy e parallelamente a z)

(è irrilevante in questo contesto come venga calcolato)

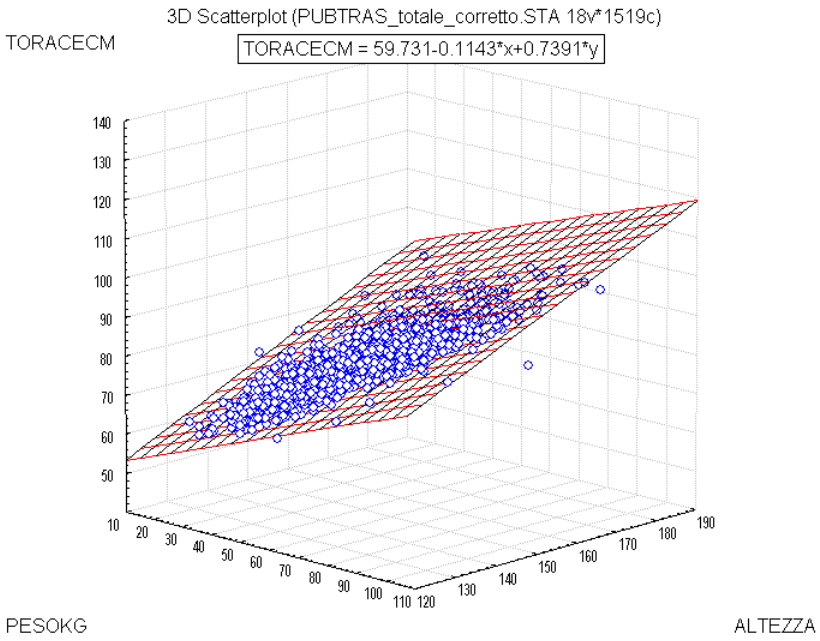

Figura 7.3: relazione fra Circonferenza toracica, altezza e peso su 1519 ragazzi

vai a indice figure

Sono riportate altre due punti di vista della nuvola di punti tridimensionale:

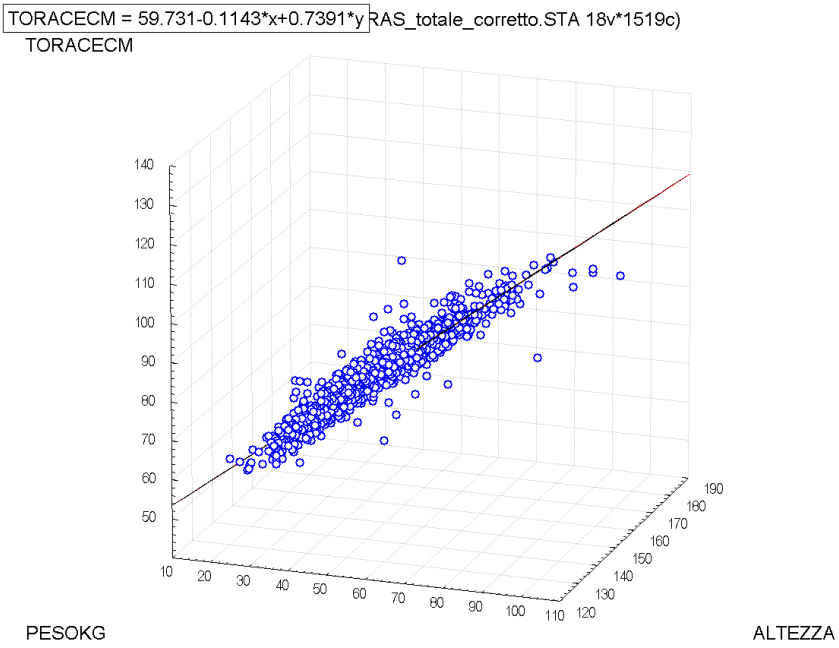

Figura 7.4: relazione fra Circonferenza toracica, altezza e peso su 1519 ragazzi

vai a indice figure

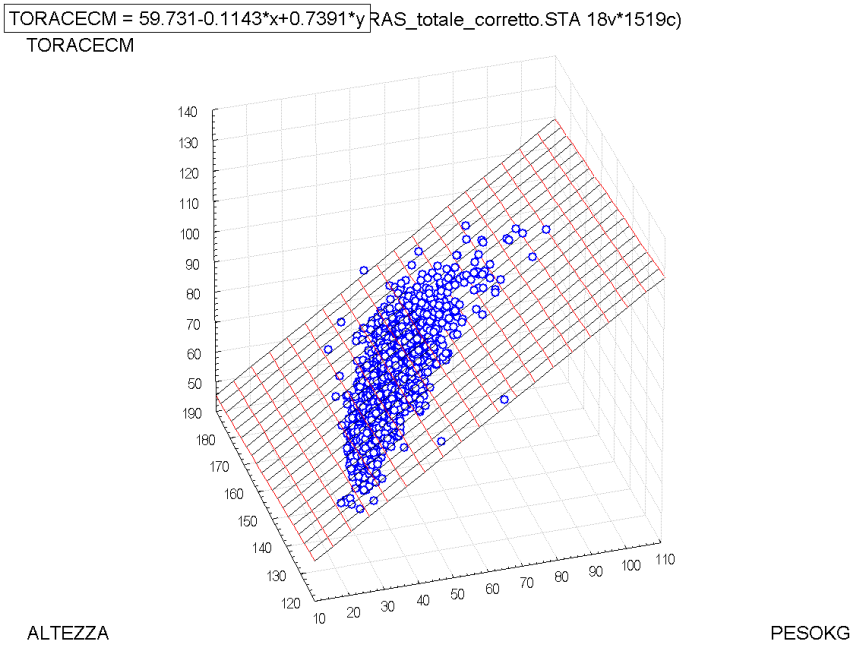

Figura 7.5: relazione fra Circonferenza toracica, altezza e peso su 1519 ragazzi

vai a indice figure

...

- Avendo utilizzato una relazione lineare (ossia l'equazione di un piano) per approssimare la relazione di regressione che fa dipendere z da x e y, piani paralleli intersecheranno il piano di regressione formando rette con la stessa pendenza
- In particolare un qualsiasi piano con y costante (ossia $y = k$ e quindi parallelo al piano X-Z) interseca il piano di regressione $z = a + bx + cy$ formando una retta di regressione di equazione:

  $$z = a + ck + bx$$

  il coefficiente b è quindi un coefficiente di regressione parziale
- L'ipotesi di linearità della regressione multipla, implica quindi regressioni parziali con la stessa pendenza: non è detto che questa sia un'ipotesi sempre realistica, ma costituisce un'approssimazione comoda.
- Si osservi ora che nel nostro caso l'intersezione del piano di regressione col piano torace- altezza (ossia a parità di peso) è una retta con pendenza negativa.

7.1.2 Correlazione parziale come correlazione fra residui

Proviamo comunque ad eliminare l'influenza della variabile peso ricorrendo solo agli strumenti tecnici della regressione lineare semplice.

...

Come eliminare l'infuenza della terza variabile sulle prime due?

Esiste un modo molto semplice per operare, che conduce agli stesi risultati che otterremo in altroi capitoli anche per altra via: calcoliamo le regressioni lineari della variabile altezza, X_1, e della variabile torace , X_2, sulla variabile peso, X_3.

Su ciascuna relazione calcoliamo i residui:

$$w_{i1} = x_{i1} - (a_{13} + b_{13} x_{i3}) \qquad \text{e} \qquad w_{i2} = x_{i2} - (a_{23} + b_{23} x_{i3}) \quad i = 1, 2, \ldots, n$$

ovviamente la nuova variabile W1 (residui Altezza) non è correlata con X3 (peso); anche la variabile W2 (residui Torace) non è correlata con X3 (peso).

coefficiente di regressione parziale

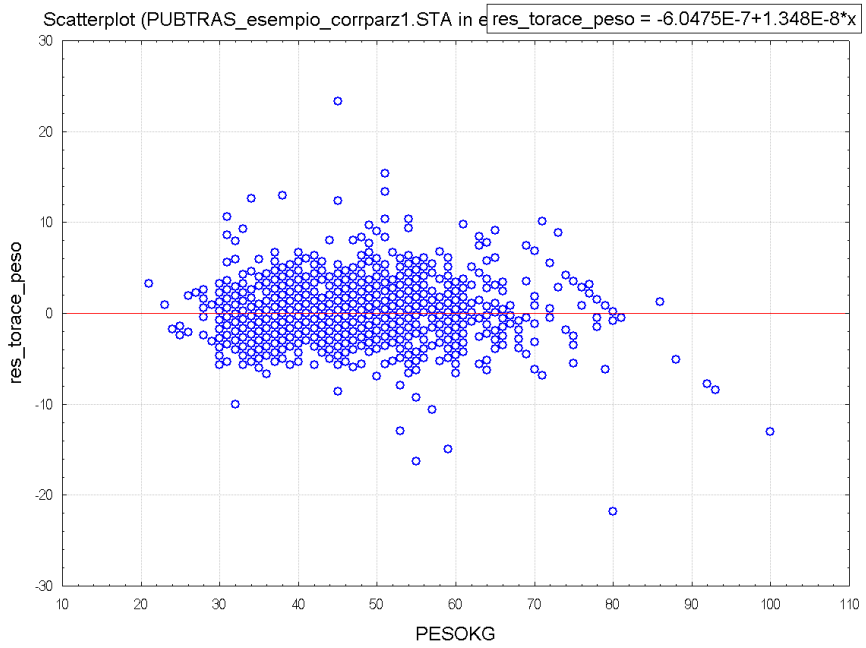

Figura 7.6: regressione dell'Altezza rispetto al peso: relazione fra i residui e la variabile esplicativa peso

vai a indice figure

...

Le due variabili W1 e W2 sono state depurate dalla dipendenza dalla variabile X3

Questa eliminazione dell'influenza di X3 è rappresentabile graficamente rappresentando nel piano le n coppie di punti $(w_{i1}, w_{i2}), i = 1, 2, \ldots, n$ insieme con la retta di regressione lineare.

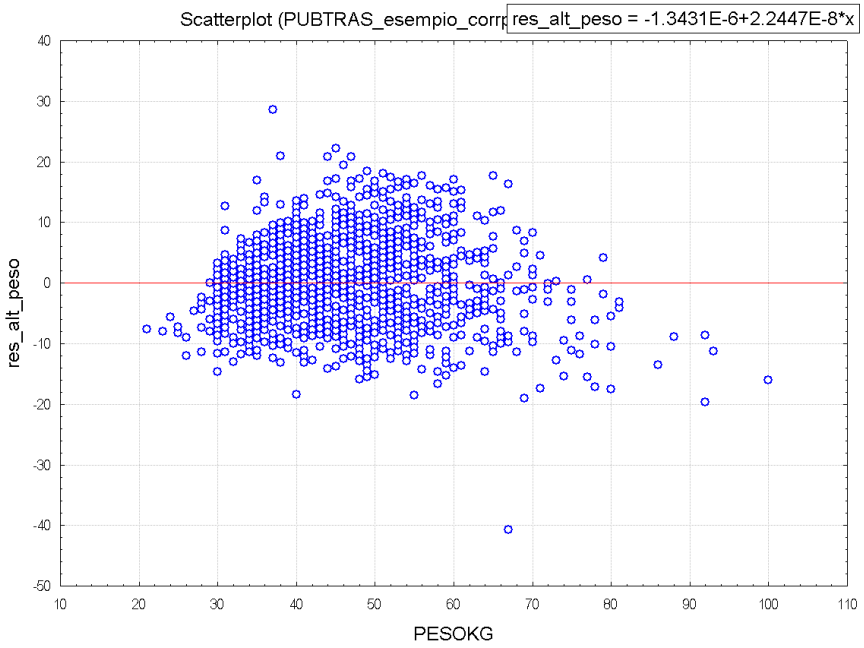

Figura 7.7: regressione della Circonferenza toracica rispetto al peso: relazione fra i residui e la variabile esplicativa peso

vai a indice figure

La retta di regressione fra il torace e l'altezza, eliminata l'influenza della variabile peso, *ha cambiato inclinazione ed è ora a pendenza negativa!*

Possiamo esprimere questo risultato dicendo che, per soggetti con lo stesso peso, la circonferenza toracica in media diminuisce all'aumentare dell'altezza.

Possiamo adesso direttamente misurare la correlazione fra le cinque variabili fin qui usate:

X1=ALTEZZA

X2=TORACE

X3=PESO

W1=res$_{a}lt_{p}eso(residuidell'altezzarispettoalpeso)$W2=res$_{t}orace_{p}eso(residuidelToracerispetto$

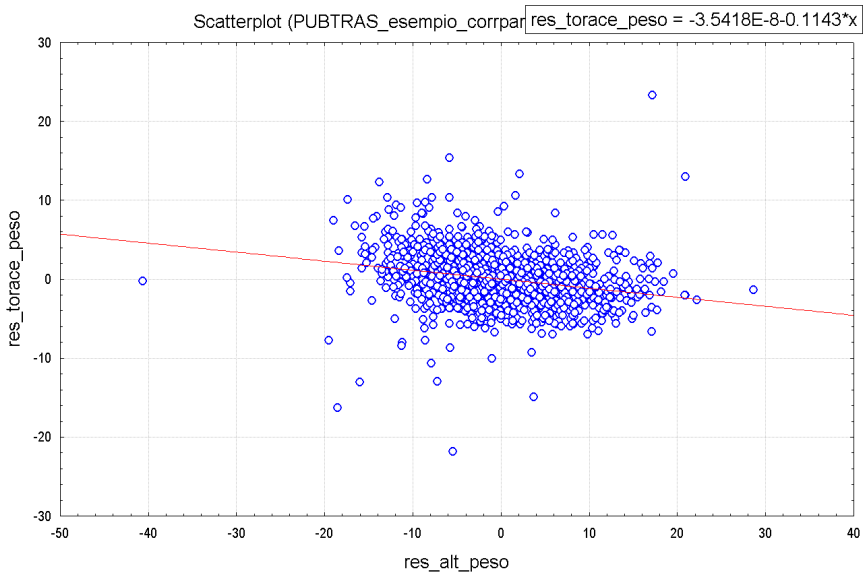

Figura 7.8: relazione fra i residui W1 della variabile torace e i residui W2 della variabile altezza

vai a indice figure

La correlazione fra le due variabili residue W1 e W2 è:

$r_{W1,W2} = -0.25$ *Correlazione fra torace e altezza a parità di peso*

Questo indice va sotto il nome di indice di correlazione lineare parziale fra le variabili X1 e X2, tenuta costante X3 e si indi con $r_{12.3}$.

Nella sezione seguente ricaviamo analiticamente $r_{12.3}$ ()se non è già noto al lettore), in funzione delle correlazioni lineari semplici.

Faccio notare soltanto che l'approccio seguito adesso per definire la correlazione parziale tenendo costante l'influenza di una variabile, è perfettamente estendibile alla correlazione parziale fra due variabili tenuta costante l'influenza di altre k variabili. Occorrerà soltanto calcolare i residui dalle regressioni multiple di X1 e X2 rispetto alle altre k variabili e poi considerarne la correlazione.

7.1.3 derivazione di $r_{12.3}$

Per derivare $r_{12.3}$ con questa impostazione, occorre richiamare soltanto alcuni risultati

della regressione lineare semplice.

Intanto ricaviamo i valori dei residui $w_{i1}, \quad w_{i2}$ in funzione dei valori originali $x_{i1}, \quad x_{i2}, \quad x_{i3}$.

Sappiamo dalla regressione lineare semplice che:

$$w_{i1} = x_{i1} - (a_{13} + b_{13}x_{i3}) = \overline{x}_{i1} - \frac{\sum_{j=1}^{n} \overline{x}_{j1}\overline{x}_{j3}}{\sum_{j=1}^{n} \overline{x}_{j3}^2}\overline{x}_{i3}$$

(con $\overline{x}$ indico lo scarto da M, media aritmetica di X)

E' più comodo adesso passare alla notazione vettoriale, per cui con $\overline{\mathbf{x}}_r\,(r = 1, 2, 3)$ indico il vettore (colonna) degli scarti relativi alla $r-$esima variabile:

$$\overline{\mathbf{x}}_r = \begin{pmatrix} x_{1r} - M_r \\ x_{2r} - M_r \\ \vdots \\ x_{jr} - M_r \\ \vdots \\ x_{nr} - M_r \end{pmatrix} \qquad , \qquad (r = 1, 2, 3)$$

Tornando all'espressione dei residui abbiamo:

$$w_{i1} = x_{i1} - (a_{13} + b_{13}x_{i3}) = \overline{x}_{i1} - \frac{\sum_{j=1}^{n} \overline{x}_{j1}\overline{x}_{j3}}{\sum_{j=1}^{n} \overline{x}_{j3}^2}\overline{x}_{i3} =$$

$$= \overline{x}_{i1} - \overline{x}_{i3}\frac{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_1}{\overline{\mathbf{x}_3}^{\mathrm{T}}\overline{\mathbf{x}_3}}$$

Adesso riesprimiamo l'intero vettore dei residui $\mathbf{w}_1$, ottenendo:

$$\mathbf{w}_1 = \mathbf{x}_1 - (a_{13} + b_{13}\mathbf{x}_3) == \overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_3\frac{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_1}{\overline{\mathbf{x}_3}^{\mathrm{T}}\overline{\mathbf{x}_3}} =$$

(mettendo in evidenza a destra il vettore $\overline{\mathbf{x}_1}$)

$$= \left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\overline{\mathbf{x}}_1$$

(si noti che $\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}$ è una matrice $(n \times n)$, mentre $\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3$ è uno scalare)

E' utile notare anche che la matrice $\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]$ è idempotente **link da creare**

A questo punto applichiamo questa formula anche alla colonna dei residui dell'altra variabile $\mathbf{w}_2$ (residui della relazione di dipendenza lineare di X_2 da X_3):

$$\mathbf{w}_2 = \left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\overline{\mathbf{x}}_2$$

Adesso finalmente costruiamo l'indice di correlazione lineare parziale:

$$r_{12.3} = \text{correlazione lineare}\,(W_1, W_2) = \frac{\mathbf{w}_2^{\mathrm{T}}\mathbf{w}_1}{\sqrt{\mathbf{w}_1^{\mathrm{T}}\mathbf{w}_1}\sqrt{\mathbf{w}_2^{\mathrm{T}}\mathbf{w}_2}} =$$

$$= \frac{\overline{\mathbf{x}}_2^{\mathrm{T}}\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\overline{\mathbf{x}}_1}{\sqrt{\overline{\mathbf{x}}_1^{\mathrm{T}}\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\overline{\mathbf{x}}_1}\sqrt{\overline{\mathbf{x}}_2^{\mathrm{T}}\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\overline{\mathbf{x}}_2}}$$

(ricordando tutte le proprietà viste in questa sezione ed applicando l'idempotenza della matrice $\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]$)

$$= \frac{\overline{\mathbf{x}}_2^{\mathrm{T}}\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\overline{\mathbf{x}}_1}{\sqrt{\overline{\mathbf{x}}_1^{\mathrm{T}}\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\overline{\mathbf{x}}_1}\sqrt{\overline{\mathbf{x}}_2^{\mathrm{T}}\left[\mathbf{I} - \frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\right]\overline{\mathbf{x}}_2}}$$

Per farla breve, si vede che le quantità a denominatore sono le radici quadrate delle devianze residue (cosa che si sapeva già dalla costruzione dell'indice di correlazione), per cui sono proporzionali a $\sqrt{1 - r_{j3}^2}\; j = 1, 2$.

A numeratore esplicitiamo il prodotto:

$$r_{12.3} = \ldots = \frac{\overline{\mathbf{x}}_2^{\mathrm{T}}\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2^{\mathrm{T}}\frac{\overline{\mathbf{x}}_3\overline{\mathbf{x}}_3^{\mathrm{T}}}{\overline{\mathbf{x}}_3^{\mathrm{T}}\overline{\mathbf{x}}_3}\overline{\mathbf{x}}_1}{\sqrt{1 - r_{13}}^2\sqrt{1 - r_{23}}^2\sqrt{DevX_1}\sqrt{DevX_2}} =$$

$$= \frac{r_{12}\sqrt{DevX_1}\sqrt{DevX_2} - \frac{r_{13}\sqrt{DevX_1}\sqrt{DevX_3}r_{23}\sqrt{DevX_2}\sqrt{DevX_3}}{DevX_3}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}\sqrt{DevX_1}\sqrt{DevX_2}} =$$

(semplificando tutte le devianze)

$$\frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

inserire poi discorso su correlazione multipla
in funzione della correlazione
parziale

Figura 7.9: correlazioni fra le 3 variabili e i due residui

vai a indice figure

Capitolo 8

Stima dei parametri del modello lineare (modelli a rango pieno)

Adesso, dopo avere visto alcuni dei più importanti impieghi del modello lineare per la descrizione di relazioni statistiche di natura varia, e le diverse interpretazioni dei parametri e delle variabili del modello, passiamo ad affrontare i problemi di stima.

L'approccio che seguiremo, di tipo parametrico, è fondato interamente sulla verosimiglianza e viene esposto prima con riferimento ad un modello generico a rango pieno; una volta esposte le caratteristiche fondamentali dell'inferenza per il caso generico, si passerà ad esaminare problemi relativi a modelli particolari, principalmente per l'analisi della regressione multipla e per l'analisi della varianza.

Si supponga che:

$$\mathbf{y}_{[n\times 1]} = \mathbf{X}_{[n\times k]}\beta_{[k\times 1]} + \boldsymbol{\varepsilon}_{[n\times 1]}$$

essendo :

$\mathbf{y}_{[n\times 1]}$	il vettore dei valori osservati
$\mathbf{X}_{[n\times k]}$	una matrice nota (i valori osservati dei regressori)
$\beta_{[k\times 1]}$	il vettore di parametri da stimare in generale completamente incognito.
$\boldsymbol{\varepsilon}_{[n\times 1]}$	un vettore di variabili casuali non osservabili, la cui distribuzione dipende in genere da un vettore $\boldsymbol{\theta}$ incognito di parametri di disturbo.

Ovviamente per potere stimare i parametri $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$ mediante il metodo della massima verosimiglianza occorre fare delle ipotesi sulla distribuzione congiunta delle componenti di $\boldsymbol{\varepsilon}$.

In ogni caso sarà necessario fare tale ipotesi se si vuole calcolare la verosimiglianza rispetto ai parametri per problemi di stima, di test e di costruzione di intervalli di confidenza di vario tipo.

In questa prima parte considereremo esclusivamente approcci di tipo parametrico.

8.0.4 Assunzioni di base nel modello lineare

Le ipotesi semplificatrici che classicamente vengono fatte nell'approccio parametrico sono:

a	$\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}_n$, momento primo	per cui $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$	$\mathbf{X}\,\boldsymbol{\beta}$ è la componente sistematica ed $\boldsymbol{\varepsilon}$ è la componente accidentale additiva.
b	$\mathrm{V}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}_n$	momento secondo b1) gli errori sono non correlati;	La matrice di varianza e covarianza della componente accidentale è diagonale con elementi uguali, ossia b2) gli errori hanno la stessa varianza (ipotesi di omoscedasticità);
c	$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}; \sigma^2\mathbf{I}_n)$	distribuzione	Nel caso di normalità degli errori, le assunzioni a) e b) che specificano i primi due momenti multivariati, identificano in modo univoco la distribuzione della comonente accidentale $\boldsymbol{\varepsilon}$.

Con queste ipotesi si vedrà che il metodo della massima verosimiglianza conduce al metodo dei minimi quadrati.

Altre implicazioni delle ipotesi di base:

- Data l'assunzione di normalità, la non correlazione fra le componenti di $\boldsymbol{\varepsilon}$ implica l'indipendenza delle componenti.

- In caso di validità della b1) e della b2) non solo si ha l'indipendenza, ma la distribuzione di ciascuna $\mathbf{y}_i$ dipende solo dalla corrispondente componente accidentale ε_i .
- Sono quindi esclusi, con questa restrizione, i modelli autoregressivi e in generale i modelli ARMA sia per l'analisi di dati temporali che di dati spaziali o territoriali.
- Le assunzioni a,b e c implicano che le ε_i abbiano la stessa distribuzione, che quindi non dipende in alcun modo né dai particolari valori x_{ij} , né dai valori dei parametri $\beta_{j.}$
- L' additività fra componente accidentale e sistematica implica che non vi sia collegamento fra l'assegnazione delle varie unità e gli errori accidentali.

8.1 La funzione di verosimiglianza nel modello lineare.

In un primo momento costruiamo la verosimiglianza del modello lineare in funzione dei parametri beta ed in funzione della varianza (o dei parametri da cui dipende la matrice di varianze e covarianze). È inutile per ora precisare se questa verosimiglianza ci servirà per costruire degli stimatori puntuali, o degli stimatori per intervallo, o per costruire dei test. In ogni caso per fare inferenza in senso lato, l'analisi della verosimiglianza è essenziale, perché ci permette di costruire un criterio per la plausibilità di determinati valori parametrici alla luce dell'evidenza campionaria.

Con le assunzioni a), b) e c) fatte prima siamo in grado di costruire la verosimiglianza campionaria, dal momento che abbiamo un campione $\mathbf{y}$ di n osservazioni estratto da una distribuzione normale di parametri (o comunque una osservazione da una normale multivariata a n componenti):

$$\mathrm{E}[\mathbf{Y}] = (\mathbf{X}\boldsymbol{\beta});$$

$$\mathrm{V}[\mathbf{Y}] = \sigma^2\mathbf{I}_n;$$

quindi in definitiva:

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$$

per cui la verosimiglianza campionaria è data da:

Verosimiglianza del modello lineare

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = (2\pi)^{-\frac{n}{2}} |\mathrm{V}[\mathbf{Y}]|^{-1/2} \times$$
$$\times \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}[\mathrm{V}[\mathbf{Y}]]^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right]$$

Funzione di verosimiglianza campionaria per il modello lineare con le ipotesi semplificatrici.

Rispetto alla notazione precedente il vettore $\boldsymbol{\theta}$ di parametri della componente accidentale è composto dal solo σ^2 , in quanto chiaramente la distribuzione di $\boldsymbol{\varepsilon}$ dipende solo da σ^2 .

Il logaritmo della verosimiglianza campionaria per i $k+1$ parametri del modello, ossia le k componenti di $\boldsymbol{\beta}$ e σ^2 è quindi dato, trascurando la costante $-(\frac{n}{2})Log(2\pi)$, da:

$$logL(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -nlog(\sigma^2)/2 - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}$$

Log Verosimiglianza per un modello lineare con l'assunzione di normalità, indipendenza e omoscedasticità

(anche uguale a:

$$-nlog(\sigma^2)/2 - \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2}{2\sigma^2})$$

Con altre ipotesi su $\mathrm{V}[\mathbf{Y}]$ si giunge a differenti funzioni di verosimiglianza e differenti stimatori.

Derivando nella rispetto a σ^2 otteniamo:

$$\frac{\partial logL[\boldsymbol{\beta}, \sigma^2|\mathbf{y}])}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2(\sigma^2)^2}$$

Uguagliando a zero e risolvendo rispetto a $\hat{\sigma}^2$ si ottiene facilmente il valore $\hat{\sigma}^2(\boldsymbol{\beta})$ che massimizza la verosimiglianza:

$$\hat{\sigma}^2(\boldsymbol{\beta}) = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{n}$$

Stima di Max. ver. di σ^2 in funzione degli altri parametri $\boldsymbol{\beta}$ per un modello con errori indipendenti e omoscedastici. (anche uguale a:

$$\hat{\sigma}^2(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n}[\mathbf{y}_i - \sum_{j=1}^{k} x_{ij}\beta_j]^2}{n}$$

Si vede dunque che con queste ipotesi la verosimiglianza campionaria dipende dalle osservazioni campionarie solo attraverso la somma dei quadrati degli scarti fra valori osservati e valori previsti.
Si vedrà più avanti il caso di osservazioni ancora distribuite normalmente ma con matrice di varianze e covarianze qualsiasi: sotto queste ipotesi più generali la verosimiglianza sarà funzione dei dati ancora attraverso una forma quadratica, ma difficilmente, o perlomeno solo in alcuni casi particolari, sarà possibile ottenere delle soluzioni esplicite per gli stimatori di massima verosimiglianza.
Tornando al nostro caso semplificato, con errori non correlati e con varianze uguali, è immediato trovare lo stimatore di massima verosimiglianza della varianza.

Si vedranno poi le caratteristiche di questo stimatore, distorsione, efficienza, etc., anche in funzione del fatto che $\boldsymbol{\beta}$ sia noto o sia da stimare.

Verosimiglianza profilo rispetto a β

Sostituendo ora nella verosimiglianza campionaria tale valore di $\hat{\sigma}^2\boldsymbol{\beta}$ al valore incognito del parametro di disturbo σ^2 , otteniamo una quantità che è funzione solo del vettore $\boldsymbol{\beta}$ dei parametri di interesse

(ossia la verosimiglianza profilo di $\boldsymbol{\beta}$)

$$L(\boldsymbol{\beta}, \hat{\sigma}^2(\boldsymbol{\beta}); \mathbf{y}) = \left(\frac{2\pi}{\hat{\sigma}^2(\boldsymbol{\beta})}\right)^{-\frac{n}{2}} \exp[-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\hat{\sigma}^2(\boldsymbol{\beta})}$$

$$= \text{cost.}\times\left(\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{n}\right)^{-\frac{n}{2}} \exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{n}}\right\}$$

In definitiva si ha:

$$L(\boldsymbol{\beta}, \hat{\sigma}^2(\boldsymbol{\beta}); \mathbf{y}) = \text{costante} \times exp(-\frac{n}{2}) \times \hat{\sigma}^2(\boldsymbol{\beta})^{-\frac{n}{2}} =$$

$$= \text{costante} \times exp(-\frac{n}{2})\left(\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{n}\right)^{-\frac{n}{2}}$$

verosimiglianza profilo rispetto a $\boldsymbol{\beta}$.

E' evidente che questa espressione è massima quando:

$$\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\text{è un minimo.}$$

Analogamente per il logaritmo di tale verosimiglianza profilo si ha:

$$\log L(\boldsymbol{\beta}, \hat{\sigma}^2(\boldsymbol{\beta}); \mathbf{y}) = \log(costante) - \frac{n}{2} - (\frac{n}{2}) Log\hat{\sigma}^2(\boldsymbol{\beta}) =$$

$$= kost - (\frac{n}{2}) \log\left(\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{n}\right)$$

(avendo posto $kost = \log(costante) - \frac{n}{2}$)

Log-verosimiglianza profilo rispetto a $\boldsymbol{\beta}$

verosimiglianza profilo normalizzata=rapporto delle verosimiglianze

La verosimiglianza profilo è uno strumento tecnico utile per fare inferenza nel caso generale di presenza di parametri di disturbo; nel nostro caso l'interesse preminente dell'inferenza è per i parametri $\boldsymbol{\beta}$: il parametro σ^2 è soltanto un parametro di disturbo, nel senso che non è necessariamente oggetto dell'inferenza ma comunque è necessario stimarlo dai dati per fare inferenza sul parametro di interesse (multiplo) $\boldsymbol{\beta}$.

Ancora vediamo che la verosimiglianza profilo è funzione dei dati solo attraverso la forma quadratica già vista: è evidente che la possibilità di ricavare la verosimiglianza profilo in modo così semplice rispetto a $\boldsymbol{\beta}$, è stata determinata dal fatto che esiste lo stimatore

di massima verosimiglianza della varianza in forma esplicita, con le assunzioni semplificatrici fatte in questo caso.

È evidente il collegamento fra verosimiglianza profilo e test basati sul rapporto delle verosimiglianze, come si vedrà fra poco; se si ricorda il metodo di costruzione del rapporto verosimiglianza si noterà come sia a numeratore sia a denominatore i parametri di disturbo vengono sostituiti dai valori massimizzano la verosimiglianza ossia dai valori più plausibili alla luce dei dati osservati.

La figura 8.1 riportata qui sotto chiarisce il significato e l'utilità dei vari tipi di verosimiglianza:

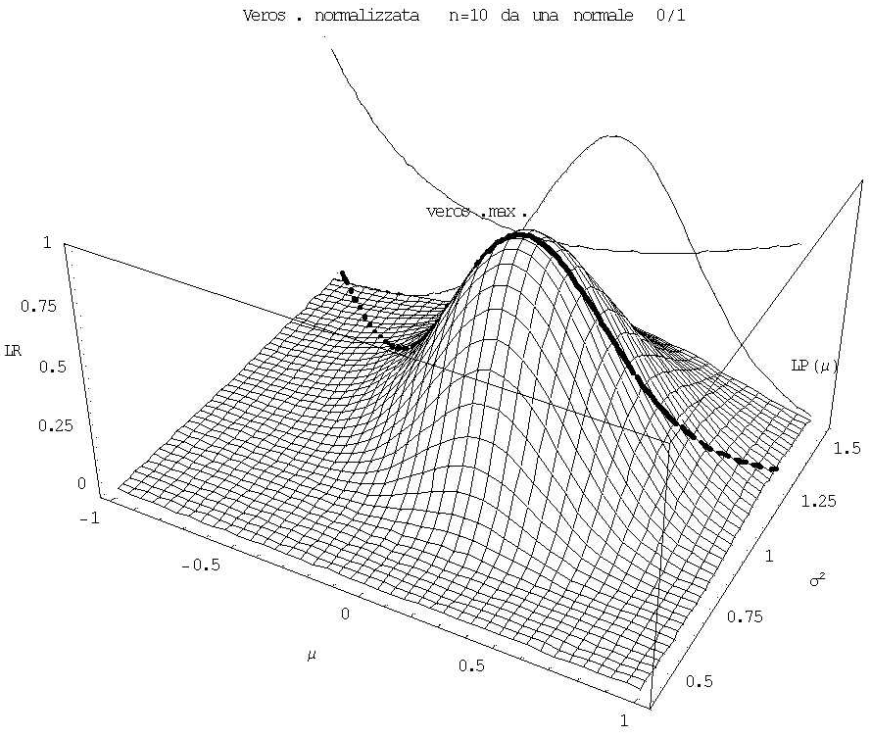

Figura 8.1: verosimiglianza rispetto a μ e σ^2 per un campione proveniente da una normale e verosimiglianza profilo su μ

vai a indice figure

La superficie rappresenta la verosimiglianza normalizzata per un campione estratto da una distribuzione normale standardizzate; tale verosimiglianza è rappresentata sull' asse z mentre sugli assi x e $\mathbf{y}$ sono rappresentati i due parametri di posizione di scala o meglio di posizione e di varianza di una distribuzione normale. Il punto di massimo è raggiunto ovviamente in corrispondenza della media campionaria e della varianza campionaria.

La curva rappresentata nel piano $x\mathbf{y}$, per comodità rappresentata sopra la superficie, rappresenta la relazione fra lo stimatore di massima verosimiglianza di sigma quadro e il parametro di posizione.

La curva in grassetto rappresentata sulla superficie è data dai valori della verosimiglianza standardizzata in corrispondenza dello stimatore ottimale della varianza. Questa è la verosimiglianza profilo rispetto al parametro medio; la curva rappresentata sul piano xz vera proiezione della verosimiglianza profilo che è funzione soltanto del parametro medio.

E' da considerare che nel caso di un modello lineare generale non

sarà possibile una tale rappresentazione grafica poiché abbiamo k parametri da stimare, ossia le componenti di; tuttavia la relazione che lega la varianza stimata ai parametri della parte sistematica è sempre la stessa, ossia di tipo quadratico.

Verosimiglianza di un campione da una normale, insieme con la verosimiglianza profilo

Costruzione del test LR

E' facile già da queste espressioni della verosimiglianza e in particolare della verosimiglianza profilo, costruire i rapporti di verosimiglianza per la verifica di particolari ipotesi sugli elementi di $\boldsymbol{\beta}$, in quanto la verosimiglianza profilo è funzione soltanto di $\hat{\sigma}^2(\boldsymbol{\beta})$ e quindi solo di $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Infatti vogliamo verificare ad esempio l'ipotesi

$$H_0 : \boldsymbol{\beta} = \beta_0$$

contro l'alternativa generica:

$$H_1 : \boldsymbol{\beta} \neq \beta_0$$

Indichiamo con $\hat{\boldsymbol{\beta}}$ la stima di massima verosimiglianza di $\boldsymbol{\beta}$ sotto H_1, costruiremo il test LR (Likelihood Ratio) rapportando la verosimiglianza massima sotto H_0 e quella massima sotto H_1. Sotto H_0 non vi sono parametri di disturbo da stimare (tranne σ^2 la cui influenza è stata eliminata in quanto stiamo lavorando con la verosimiglianza profilo su $\boldsymbol{\beta}$);

sotto H_1 a parte σ^2 va stimato il vettore $\boldsymbol{\beta}$

Per cui otteniamo la relazione:

$$LR = \frac{\max L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}|H_0)}{\max L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}|H_1)}$$

$$= \frac{L(\beta_0, \hat{\sigma}^2(\beta_0); \mathbf{y})}{L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\hat{\boldsymbol{\beta}}); \mathbf{y})} = \left(\frac{\hat{\sigma}^2(\beta_0)}{\hat{\sigma}^2(\hat{\boldsymbol{\beta}})} \right)^{\frac{n}{2}} =$$

$$\left(\frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{(\mathbf{y} - \mathbf{X}\beta_0)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta_0)} \right)^{\frac{n}{2}}$$

Come è noto valori alti di LR (vicini ad uno) indicheranno la plausibilità dell'ipotesi nulla; ci preoccuperemo dopo della costruzione effettiva dei test e della loro distribuzione campionaria.

In generale comunque se vogliamo saggiare una generica ipotesi nulla H_0 contro una più generale H_1, essendo H_0 un caso particolare di H_1, possiamo pensare ciascuna ipotesi $H_i (i = 1, 2)$ come un sistema di vincoli $g_i(\beta)$ imposti sugli elementi di $\boldsymbol{\beta}$.

Ad esempio $g_0(\boldsymbol{\beta})$ potrebbe consistere del fatto che una superficie sia di primo grado, mentre $g_1(\boldsymbol{\beta})$ potrebbe essere l'alternativa che la superficie sia di secondo grado (ma non un polinomio di grado superiore).

Indicando ora con $\hat{\boldsymbol{\beta}}_i$ la stima di massima verosimiglianza di $\boldsymbol{\beta}$ sotto H_i, possiamo nel caso generale costruire il test:

$$LR = \frac{\max L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}|g_0(\boldsymbol{\beta}))}{\max L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}|g_1(\boldsymbol{\beta}))}$$

$$\frac{L(\hat{\boldsymbol{\beta}}_0, \hat{\sigma}^2(\hat{\boldsymbol{\beta}}_0); \mathbf{y})}{L(\hat{\boldsymbol{\beta}}_1, \hat{\sigma}^2(\hat{\boldsymbol{\beta}}_1); \mathbf{y})}$$

$$= \frac{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}_1)}^{-\frac{n}{2}} =$$

$$\left\{ \frac{\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_1}{\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0} \right\}^{\frac{n}{2}}$$

Il criterio del rapporto della verosimiglianza conduce ad un test sensibile e ad uno strumento generalmente molto utile per l'inferenza statistica sebbene non possegga almeno per piccoli campioni le proprietà ottimali che un test dovrebbe avere secondo la teoria di Neyman-Pearson. Il problema della verifica di ipotesi, ossia della costruzione di un test di significatività, si può riassumere come segue: sulla base dei dati osservati la famiglia di distribuzioni dell'ipotesi alternativa H_1 si adatta significativamente meglio ai dati della famiglia parametrica rappresentata dall'ipotesi nulla H_0 ? Rifiutiamo H_0 a favore di H_1 se questo miglioramento è significativo.

Sebbene questo test non possegga tutte le proprietà ottimali richieste, risponde comunque ai requisiti fissati da Fisher per la verifica di ipotesi nell'indagine scientifica: lo scopo dei test è di attestare l'evidenza che i dati forniscono in merito a certe ipotesi più o meno definite; criteri di ottimalità quali potenza, ampiezza, non distorsione, sono importanti ma non sono necessariamente la cosa più importante nelle applicazioni.

Dalla costruzione del test del rapporto delle verosimiglianze per i parametri di un modello lineare con l'ipotesi di normalità, eteroscedasticità, non correlazione, si vede che tale rapporto dipende esclusivamente dai rapporti fra le varianze stimate sotto le diverse ipotesi;

- una varianza è quella relativa all'ipotesi più generale cioè quella che impone meno vincoli sui parametri che sarà più piccola nell'ambito della famiglia parametrica considerate;
- la varianza relativa alla verosimiglianza del numeratore è quella calcolata sotto l'ipotesi di esistenza di qualche vincolo sui parametri.

È quindi evidente che l'ipotesi di normalità implica che le quantità sufficienti per fare inferenza sono le varianze stimate.

8.1.1 MINIMI QUADRATI ORDINARI

Per trovare dunque il massimo incondizionato della verosimiglianza occorre trovare $\hat{\boldsymbol{\beta}}$ che da ora in poi indico per comodità di notazione con $\mathbf{b}$.

Minimi quadrati

Va trovato il minimo di $(\mathbf{y}-\mathbf{Xb})^{\mathrm{T}}(\mathbf{y}-\mathbf{Xb})$ ossia il minimo della somma dei quadrati degli scarti fra: valori osservati $\mathbf{y}$ e valori calcolati $\mathbf{Xb}$. (indicati con $\mathbf{y}_i^*$)

Minimi Quadrati Ordinari. (Ordinary Least Squares: OLS) In forma matriciale:

$$\min_{\mathbf{b}} R(\mathbf{b}),$$

con:

$$R(\mathbf{b}) = \sum_{i=1}^{n}(\mathbf{y}_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2 = \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i^*)^2$$

$$= (\mathbf{y}_{[n\times 1]} - \mathbf{X}_{[n\times k]}\mathbf{b}^{\mathrm{T}}_{[k\times 1]})(\mathbf{y}_{[n\times 1]} - \mathbf{X}_{[n\times k]}\mathbf{b}_{[k\times 1]}) =$$

$$= \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b}$$

essendo $\mathbf{y}_i^*$ l' i -esimo valore stimato.

Derivando $R(\mathbf{b})$ ($= \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b}$) rispetto al vettore $\mathbf{b}$ si ottiene:

$$\frac{\partial R(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b}$$

Uguagliandole a 0 (vettore nullo):

$$-2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b} = 0;$$

Occorre risolvere, in $\mathbf{b}$, il sistema:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Sistema di equazioni normali

Temporaneamente imponiamo la restrizione che $\mathbf{X}$ sia di rango k , e quindi esiste, ed è unica, l'inversa di $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

Diversamente potremmo ricorrere ad una riparametrizzazione oppure all'uso dell'inversa generalizzata

---

SOLUZIONE GENERALE DEI MINIMI QUADRATI NEI MODELLI LINEARI A RANGO PIENO

(Sono stimatori di massima verosimiglianza con le ipotesi semplificatrici)

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

la soluzione esiste unica avendo supposto $\mathbf{X}$ di rango k
e fornisce certamente il minimo di $(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b})$

Si tratta certamente di un minimo, in quanto le condizioni del secondo ordine, riguardanti l'Hessiano, sono sempre soddisfatte, è:

$$\frac{\partial R(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b}e$$

$$\frac{\partial^2 R(\mathbf{b})}{\partial \mathbf{b}\partial \mathbf{b}^{\mathrm{T}}} = 2(\mathbf{X}^{\mathrm{T}}\mathbf{X})$$

che è sempre definita positiva e quindi il punto di stazionarietà fornisce il minimo assoluto della funzione.

`inserire dimostrazione senza derivate`
`ispirata a Rao`

$$R(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) =$$

(anche uguale a: $(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])^{\mathrm{T}}(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])$)=

(Aggiungendo e sottraendo **Xb**)

$$= q[(\mathbf{y} - \mathbf{Xb}) + (\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta})] =$$

sviluppiamo il prodotto in cui compare il binomio formato dai due termini: $(\mathbf{y} - Xb)e(Xb - \mathbf{X})$

Il metodo dei minimi quadrati ordinari (OLS: Ordinary Least Squares) COINCIDE con il metodo della massima verosimiglianza se e solo se: la distribuzione di ε è una normale a n variabili a componenti indipendenti e con uguale varianza σ^2 (altrimenti occorre impiegare metodi diversi da quello dei minimi quadrati)

Pertanto gli stimatori dei minimi quadrati godranno delle proprietà asintotiche ottimali degli stimatori M.V. soltanto nel caso gaussiano, diversamente saranno soltanto i migliori stimatori lineari non distorti.

Teorema di Gauss-Markov

Date le assunzioni a) e b), ossia errori a media nulla, non correlati ed a varianze uguali, gli stimatori dei minimi quadrati hanno comunque una proprietà ottimale:

In un modello lineare, con le assunzioni ricordate sopra, gli stimatori dei minimi quadrati di un qualsiasi insieme di funzioni lineari dei parametri β_j sono a varianza minima nella classe degli stimatori non distorti e lineari nelle $\mathbf{y}_i$ In effetti si può anche dimostrare che sono gli stimatori con la minima varianza generalizzata.

In effetti questo teorema non dimostra affatto la superiorità assoluta degli stimatori dei minimi quadrati, è può considerarsi una proprietà sufficiente per rendere inutile l'assunzione di normalità: infatti il teorema asserisce solo che sono i migliori fra gli stimatori lineari nelle osservazioni non distorti.

Intanto non è detto che la non distorsione sia una proprietà in assoluto necessaria, ma fondamentalmente nulla obbliga a restringersi agli stimatori lineari.

Assumere la linearità nelle osservazioni equivale ad assumere la normalità.

Ad esempio nella derivazione della normale:

imponendo la condizione che dato un campione di n osservazioni indipendenti il miglior stimatore di E($\mathbf{X}$) sia la media aritmetica delle osservazioni, Gauss dimostrò che la distribuzione degli errori è normale.

MINIMA VARIANZA E MINIMA VARIANZA GENERALIZZATA.

Variabili a media zero (regressione in termini di scarti)

Se $\mathbf{X}$ è posta nella forma conveniente vista prima, ossia prima colonna tutta uguale ad 1, e k colonne di scarti dei regressori dalle rispettive medie, $\mathbf{X}$ avrà un totale di $k+1$ colonne, supposte linearmente indipendenti (dal momento che il rango di $\mathbf{X}$ è in questo caso $k+1$).

Questa forma della matrice dei regressori viene utilizzata quando si vuole esplicitamente inserire un'ordinata all'origine β_0 fra i parametri del modello e per semplificare alcune scomposizioni successive:

Si vede facilmente che in questo caso:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} n & \mathbf{0}_k^{\mathrm{T}} \\ 0_k & nS_{\mathbf{X}} \end{pmatrix}$$

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{pmatrix} 1/n & \mathbf{0}_k^{\mathrm{T}} \\ 0_k & (S_{\mathbf{X}})^{-1}/n \end{pmatrix}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{pmatrix} nM_{\mathbf{y}} \\ \mathrm{ncov}(\mathbf{X}, \mathbf{y}) \end{pmatrix}$$

avendo indicato:

con $\mathbf{S_X}$ matrice delle varianze e covarianze dei k regressori e

$\mathrm{cov}(\mathbf{X}, \mathbf{y})$ vettore delle covarianze fra la $\mathbf{y}$ e le x .

In questo modo è possibile separare la stima del termine noto da quella dei coefficienti di regressione:

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_0 = M_{\mathbf{y}} \\ \mathbf{b}_k = \mathbf{S}_{\mathbf{X}}^1 cov(\mathbf{X}, \mathbf{y}) \end{pmatrix}$$

8.1.2 Distribuzione campionaria di b (minimi quadrati ordinari)

In ogni caso, qualunque sia la scelta della $\mathbf{X}$, comunque di rango k (e k colonne), lo stimatore $\mathbf{b}$ è dato in generale da:

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y};$$

Per ipotesi $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$;

e quindi $\mathbf{b}$ è una combinazione lineare delle $\mathbf{y}$ per cui potremmo direttamente applicare le regole per il calcolo dei momenti di combinazioni lineari di variabili casuali.

Per la speranza matematica di $\mathbf{b}$ si ha:

$$E(\mathbf{b}) = E[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}] =$$

$$= E[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\beta + \boldsymbol{\varepsilon})] =$$

$$= E[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}] + E[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon}] =$$

$$= E(\boldsymbol{\beta}) + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}E(\boldsymbol{\varepsilon}) =$$

$$= \boldsymbol{\beta}$$

Momento primo di b

$$\mathrm{E}[\mathbf{b}] = \boldsymbol{\beta}$$

($\mathbf{b}$ è uno stimatore corretto di $\boldsymbol{\beta}$) Per ottenere il risultato è stato sufficiente assumere soltanto: $E(\varepsilon) = \mathbf{0}_n$.
Quindi perché $\mathbf{b}$ sia corretto per $\boldsymbol{\beta}$ è sufficiente che il modello lineare sia non distorto.

Per la matrice di varianze e covarianze campionarie di $\mathbf{b}$ si ha

$$\mathrm{V}[\mathbf{b}] = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathrm{V}[\mathbf{Y}]\,\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} =$$

$$= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} =$$

$$= \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} =$$

$$= \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$$

Momento secondo di b

$$V(\mathbf{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$$

avendo assunto oltre a $E(\boldsymbol{\varepsilon}) = 0_n$:

$$\mathbf{V}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$$

(omoscedasticità e non correlazione)
qualunque sia la forma della distribuzione delle ε_i

Quindi la struttura della matrice di varianze e covarianze di $\mathbf{b}$ dipende dalla struttura della matrice $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ e quindi dalla struttura delle matrici $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ e $\mathbf{X}$. Se la matrice $\mathbf{X}$ è una matrice di scarti dalle medie aritmetiche (e le variabili indipendenti sono numeriche in senso stretto), allora $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ è la matrice di devianze e codevianze dei k regressori; pertanto la struttura dei primi due momenti multivariati della distribuzione di $\mathbf{b}$ non dipende solo dalle assunzioni su $\boldsymbol{\varepsilon}$ ma anche dalla struttura della matrice $\mathbf{X}$.
Questo è uno degli aspetti di cui occorre tenere maggiormente conto tutte le volte che è possibile scegliere, in tutto o in parte, come costruire la matrice delle $\mathbf{x}$.
Se (e solo se) le $\mathbf{X}_j$ sono tutte non correlate i b_j saranno tutti non correlati; Se la matrice $(\mathbf{X}^{\mathrm{T}}\mathbf{X}$) risulta a blocchi (ossia gruppi di variabili internamente correlate ma non fra gruppi diversi), allora è a blocchi anche $V(\mathbf{b})$, ossia i corrispondenti gruppi di stimatori dei coefficienti saranno internamente correlati ma fra gruppi diversi vi sarà assenza di correlazione.
Si rivedranno in contesti particolari alcuni di questi aspetti

---

Distribuzione di b

Se, inoltre, vale l'assunzione di normalità, allora:
b segue una distribuzione normale multivariata (in quanto combinazione lineare delle **y**)
b è lo stimatore di massima verosimiglianza (come peraltro abbiamo già ottenuto)

$$\mathbf{b}(N(\boldsymbol{\beta}; \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$$

Si possono quindi costruire eventualmente delle regioni di confidenza per i parametri (se σ^2 è noto) che risulteranno in questo caso ellissoidali. Occorrerà distinguere il caso in cui σ^2 sia noto (poco plausibile) dal caso in cui venga stimato. In effetti anche senza assumere la normalità della componente accidentale, sotto condizioni non troppo restrittive sulla matrice delle x la distribuzione dello stimatore dei minimi quadrati tende alla normale al divergere di n . Si rivedrà questa proprietà quando si parlerà dell'allontanamento dalle ipotesi di normalità.

Introdurre qui discussione sull'assunzione di normalità
(verrà poi ripresa nella parte relativa all'analisi dei residui
ed agli allontanamenti dalle assunzioni di base)

8.2 Distribuzione della devianza residua nei modelli lineari

8.2.1 Devianza residua in funzione dei valori osservati

Indichiamo ancora con $\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ lo stimatore di massima verosimiglianza di $\boldsymbol{\beta}$ in un modello lineare (di rango pieno), supponendo la validità delle ipotesi semplificatrici sulla componente accidentale:

$$\boldsymbol{\varepsilon} \sim N_n(0; \sigma^2\mathbf{I})$$

Trasformiamo la devianza residua $R(\mathbf{b})$, ossia la somma dei quadrati degli scarti fra valori della variabile di risposta osservati e stimati (che è la quantità minimizzata mediante il metodo dei minimi

quadrati); l'importanza di tale quantità (e della sua distribuzione campionaria!) è evidente alla luce di quanto abbiamo visto sui test basati sui rapporti di verosimiglianze.

Il vettore $\mathbf{y} - \mathbf{Xb}$ è detto vettore dei residui (empirici). $R(\mathbf{b})$ è quindi la devianza dei residui empirici

Esprimiamo la devianza residua in funzione delle osservazioni:

$R(\mathbf{b}) = \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i^*)^2 \sum_{i=1}^{n}(\mathbf{y}_i - \sum_{j=1}^{k} x_{ij}b_j)^2 =$	
$= (\mathbf{y} - \mathbf{Xb})^{\mathrm{T}}(\mathbf{y} - \mathbf{Xb}) =$	(sostituendo a $\mathbf{b}$ il valore trovato $\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$)
$= (\mathbf{y} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}) =$	
$= [(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y}]^{\mathrm{T}}[(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y}]$	mettendo in evidenza $\mathbf{y}$
$= \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y}$	

ed infine:

$$R(\mathbf{b}) = \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y}$$

devianza residua

$(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})$ è simmetrica ed idempotente di rango $n-k$ (infatti una qualsiasi matrice $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ è idempotente di rango k, come si è visto nella parte iniziale)

$R(\mathbf{b})$ è una forma quadratica nelle $\mathbf{y}$

Inoltre:

$$(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{X} = 0_{n\times k}$$

e quindi i residui empirici risultano non correlati con le $\mathbf{X}$ è:

$$Cov(\mathbf{y} - \mathbf{Xb}, \mathbf{X}) = [\mathbf{y} - \mathbf{Xb}]^T\mathbf{X} = 0$$

si ricava direttamente dalle equazioni normali.

($\mathbf{y} - \mathbf{Xb}$ ha media nulla).

Devianza residua in funzione della componente accidentale ε :

Esprimiamo ora $R(\mathbf{b})$ in funzione della componente accidentale $\boldsymbol{\varepsilon}$:

Dall'espressione precedente:

$$R(\mathbf{b}) = \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y} =$$

(operando sul terzo fattore, esprimendo $\mathbf{y}$ come $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, secondo quanto ipotizzato)

$$= \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) =$$

aprendo la parentesi a destra

$$= \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\boldsymbol{\varepsilon} =$$

e dato che $(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{X} = 0$, ed effettuando le stesse operazioni sul termine $\mathbf{y}^{\mathrm{T}}$, si ha:

$$= \mathbf{y}^{\mathrm{T}}0_{n\times k}\boldsymbol{\beta} + (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})^{T}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\boldsymbol{\varepsilon} =$$

aprendo la parentesi a sinistra

$$= 0 + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\boldsymbol{\varepsilon} =$$

$$= 0 + 0 + \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\boldsymbol{\varepsilon}.$$

In definitiva si ha l'ulteriore espressione per la devianza residua:

$$R(\mathbf{b}) = \varepsilon^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\varepsilon$$

La devianza residua $R(\mathbf{b})$ è quindi una forma quadratica nelle ε

Quindi si può vedere facilmente che, essendo $E(\varepsilon_i\varepsilon_j) = 0(i \neq j)$, sviluppando i termini della forma quadratica si ha:

$$E(R(\mathbf{b})) = tr(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\sigma^2 edinfine :$$

$$E[R(\mathbf{b})] = (n-k)\sigma^2$$

---

avendo ipotizzato soltanto:

$$E(\boldsymbol{\varepsilon}) = 0eV(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}_n$$

(anche senza l'assunzione di normalità); quindi:

$$s^2 = R(\mathbf{b})/(n-k) = \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i^*)^2/(n-k)$$

è sempre una stima corretta della varianza.

Distribiuzione della devianza residua

Se vale l' assunzione di normalità,

$$R(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}) = \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\boldsymbol{\varepsilon}$$

si distribuisce come $\sigma^2\chi^2_{n-k}$,

perché è una forma quadratica in variabili normali indipendenti a media zero e varianze uguali ($\boldsymbol{\varepsilon}$) , con matrice dei coefficienti
($\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$) idempotente di rango $n-k$.

8.3 Scomposizione della devianza nel modello lineare e verifica di ipotesi.

Effettuiamo alcune scomposizioni delle diverse somme di quadrati (e forme quadratiche in generale) che abbiamo incontrato (fra cui ad esempio: $R(\mathbf{b})$, $R(\boldsymbol{\beta})$, $\mathbf{y}^{\mathrm{T}}\mathbf{y}$).

La scomposizione della somma dei quadrati $\mathbf{y}^{\mathrm{T}}\mathbf{y}$

Operiamo sulla devianza di $\mathbf{y}$, (o più precisamente sulla somma dei quadrati $\mathbf{y}^{\mathrm{T}}\mathbf{y}$) partendo ancora da una delle relazioni trovate per $R(\mathbf{b})$:

$$R(\mathbf{b}) = \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i^*)^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}) =$$

$$= \mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y} =$$

aprendo la parentesi

$$= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} =$$

sostituendo $\mathbf{b}$ alla espressione $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

$$= \mathbf{y}^{\mathrm{T}}\mathbf{y} - (\mathbf{y}^{\mathrm{T}}\mathbf{X})\mathbf{b} =$$

Ricordiamo che, trasponendo il sistema di equazioni normali si ha:

$$\mathbf{y}^{\mathrm{T}}\mathbf{X} = \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

$$= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b};$$

ed infine (risolvendo rispetto a $\mathbf{y}^{\mathrm{T}}\mathbf{y}$):

T avola Di Scomposizione Della Devianza Empirica (Somme Dei Quadrati)	
FORMA QUADRATICA	FONTE DI VARIABILITA'
$\mathbf{y}^{\mathrm{T}}\mathbf{y} =$	Somma dei quadrati di $\mathbf{y}$ (devianze se $\mathbf{y}$ è a media nulla)
$(\mathbf{y} - \mathbf{Xb})^{\mathrm{T}}(\mathbf{y} - \mathbf{Xb})+$	devianza residua
$\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Xb}$	Somma dei quadrati spiegata dalla regressione

8.3.1 Scomposizione di $R(\beta)$

Per potere costruire dei test, trasformiamo ora la devianza teorica $R(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}$

$$R(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) =$$

(anche uguale a: $(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])^{T}(\mathbf{Y} - \mathrm{E}[\mathbf{Y}]\ ) =$

(Aggiungendo e sottraendo **Xb**)

$$= q[(\mathbf{y} - \mathbf{Xb}) + (\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta})] =$$

sviluppiamo il prodotto in cui compare il binomio formato dai due termini: $(\mathbf{y} - Xb)e(Xb - \mathbf{X})$

$= (\mathbf{y} - \mathbf{Xb})^T(\mathbf{y} - \mathbf{Xb})$	$=R(\mathbf{b})$
$+$	
$(\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta})$	si mette in evidenza $\mathbf{X}$ sia a sinistra che a destra e si ottiene $(\mathbf{b} - \boldsymbol{\beta})^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})$
$+$	
$(\mathbf{y} - \mathbf{Xb})^T(\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta})$	$= 0$ perché $(\mathbf{y} - \mathbf{Xb})^T\mathbf{X} = 0$ dalle equazioni dei minimi quadrati
$=$	
$R(\mathbf{b}) + (\mathbf{b} - \boldsymbol{\beta})^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})$	

Si può interpretare tale scomposizione in modo leggermente diverso, ponendo l'enfasi non su $\mathbf{b}$, stimatore di $\boldsymbol{\beta}$, bensì su $\mathbf{Xb}$, stimatore lineare ottimale del valore atteso $\mathrm{E}[\mathbf{Y}]$. Pertanto

In definitiva quindi si ha:

$$R(\boldsymbol{\beta}) = R(\mathbf{b}) + (\mathbf{b} - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})$$

Oppure :

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\mathbf{b})+(\mathbf{b}-\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})$$

Possiamo rivedere questa relazione in termini di contributi alla devianza teorica di $\boldsymbol{\varepsilon}$:

$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) =$	$(\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})+$	$(\mathbf{b} - \boldsymbol{\beta})^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})$
devianza teorica complessiva di $\boldsymbol{\varepsilon}$ (rispetto al modello vero)	devianza residua	devianza delle stime

Questa scomposizione è basilare anche perché possiamo vedere che il rapporto delle verosimiglianze costruito in precedenza per saggiare l'ipotesi nulla $H_0 : \boldsymbol{\beta} = \beta_0$, contro l'alternativa generica $H_1 : \boldsymbol{\beta} \neq \beta_0$, è funzione di queste quantità. Infatti:

$$LR = \frac{\max[L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})|H_0]}{\max[L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})|H_1]} == \left\{ \frac{[\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})]}{[\mathbf{y} - \mathbf{X}\beta_0)^T(\mathbf{y} - \mathbf{X}\beta_0)]} \right\}^{\frac{n}{2}} =$$

$$= \left\{ \frac{R(\mathbf{b})}{R(\beta_0)} \right\}^{\frac{n}{2}}$$

avendo ora indicato con $\mathbf{b}$ lo stimatore di massima verosimiglianza prima indicato con $\hat{\boldsymbol{\beta}}$.

8.3.2 Test F per la verifica di ipotesi nel modello lineare: distribuzione nulla

Per esaminare la distribuzione nulla del rapporto delle verosimiglianze, o di una sua trasformazione monotona, riprendiamo in esame la

scomposizione di base di $R(\boldsymbol{\beta})$, e dividiamo tutti i termini per σ^2 :

$$\frac{R(\boldsymbol{\beta})}{\sigma^2} = \frac{R(\mathbf{b})}{\sigma^2} + \frac{(\mathbf{b}-\boldsymbol{\beta})^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}}{\sigma^2}$$

con le ipotesi che abbiamo fatto, compresa ovviamente quella di normalità, possiamo vedere che i tre termini si distribuiscono come delle χ^2 , per cui si può applicare il teorema di Cochran; infatti:

$$\frac{R(\boldsymbol{\beta})}{\sigma^2} :$$

(A) si distribuisce come una χ^2 con n gradi di libertà in quanto somma dei quadrati di n v.c. normali standardizzate:

$$\frac{R(\boldsymbol{\beta})}{\sigma^2} = \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\sigma^2} = \frac{\sum_{i=1}^{n}\varepsilon_i^2}{\sigma^2}$$

$$\frac{R(\mathbf{b})}{\sigma^2} :$$

(B) si distribuisce come una χ^2 con n-k gradi di libertà (come si è visto) in quanto:

$$R(\mathbf{b}) = (\mathbf{y}-\mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\mathbf{b}) =$$
$$= \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\boldsymbol{\varepsilon}$$

si distribuisce come $\sigma^2\chi^2_{n-k}$
essendo $(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})$ idempotente di rango $n-k$

$$\frac{[\mathbf{b}-\boldsymbol{\beta}]^T\mathbf{X}^{\mathrm{T}}\mathbf{X}[\mathbf{b}-\boldsymbol{\beta}]}{\sigma^2}$$

(C)
si distribuisce come una χ^2 con k gradi di libertà in quanto è il numeratore dell'esponente della densità di una normale multivariata:

$$\mathbf{b} \sim N(\boldsymbol{\beta}; \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})$$

Quindi si può applicare il teorema di Cochran ed i termini (B) e (C) risultano indipendenti.

In definitiva la quantità:

$$F = \frac{\frac{[\mathbf{b}-\boldsymbol{\beta}]^T\mathbf{X}^{\mathrm{T}}\mathbf{X}[\mathbf{b}-\boldsymbol{\beta}]}{k}}{\frac{[\mathbf{y}-\mathbf{X}\mathbf{b}]^T[\mathbf{y}-\mathbf{X}\mathbf{b}]}{n-k}} = \frac{[\mathbf{b}-\boldsymbol{\beta}]^T\mathbf{X}^{\mathrm{T}}\mathbf{X}[\mathbf{b}-\boldsymbol{\beta}]}{ks^2}$$

essendo il rapporto fra due variabili casuali χ^2 indipendenti divise per i rispettivi gradi di libertà, si distribuisce secondo una F di Snedecor con k ed $n-k$ gradi di libertà, essendo $\boldsymbol{\beta}$ il vero valore del vettore dei parametri.

Pertanto, per saggiare l'ipotesi nulla:

$H_0 : \boldsymbol{\beta} = \beta_0$, contro l'alternativa generica

$$H_1 : \boldsymbol{\beta} \neq \beta_0,$$

possiamo impiegare la quantità:

$$F = \frac{\frac{[\mathbf{b}-\beta_0]^T \mathbf{X}^{\mathrm{T}} \mathbf{X} [\mathbf{b}-\beta_0]}{k}}{\frac{[\mathbf{y}-\mathbf{X}\mathbf{b}]^T [\mathbf{y}-\mathbf{X}\mathbf{b}]}{n-k}}$$

che sotto H_0 si distribuisce secondo una variabile aleatoria F di Snedecor con k ed $n-k$ gradi di libertà.

La regione di rifiuto sarà costituita dai valori elevati di F, superiori ad $F_{\alpha,k,n-k}$. (ossia situati sulla coda destra della corrispondente variabile F di Snedecor)
Infatti valori osservati di F elevati danno evidenza contraria ad H_0.

F è funzione monotona del rapporto delle verosimiglianze LR costruito in precedenza. Infatti:

$$F = \frac{\frac{[\mathbf{b}-\beta_0]^T \mathbf{X}^{\mathrm{T}} \mathbf{X} [\mathbf{b}-\beta_0]}{k}}{\frac{[\mathbf{y}-\mathbf{X}\mathbf{b}]^T [\mathbf{y}-\mathbf{X}\mathbf{b}]}{n-k}} =$$

$$F = \frac{\frac{R(\beta_0)-R(\mathbf{b})}{k}}{\frac{R(\mathbf{b})}{n-k}} =$$

$$F = \left(\frac{R(\beta_0)}{R(\mathbf{b})} - 1 \right) \frac{n-k}{k} =$$

$$F = \left(\frac{1}{LR} - 1 \right) \frac{n-k}{k}$$

Statistiche sufficienti nel modello lineare.

$\mathbf{b}$ e s^2 costituiscono un set di stimatori congiuntamente sufficienti per $\boldsymbol{\beta}$ e σ^2 .

Infatti partendo dalla verosimiglianza del modello lineare, introdotta prima, con le ipotesi semplificatrici fatte, e con le scomposizioni ora viste si può giungere ad una fattorizzazione:

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) =$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\frac{R(\boldsymbol{\beta})}{2\sigma^2}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\frac{R(\mathbf{b})}{2\sigma^2} - \frac{[\mathbf{b} - \boldsymbol{\beta}]^T \mathbf{X}^{\mathrm{T}}\mathbf{X}[\mathbf{b} - \boldsymbol{\beta}]}{2\sigma^2}] =$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\frac{(n-k)s^2}{2\sigma^2}] \exp[-\frac{[\mathbf{b} - \boldsymbol{\beta}]^T \mathbf{X}^{\mathrm{T}}\mathbf{X}[\mathbf{b} - \boldsymbol{\beta}]}{2\sigma^2}].$$

Quindi la verosimiglianza campionaria rispetto a $\boldsymbol{\beta}$ e σ^2 dipende dalle osservazioni solo attraverso le statistiche $\mathbf{b}$ e s^2 .

Matrice di informazione

Dalla verosimiglianza è anche immediato vedere che l'informazione di Fisher su $\boldsymbol{\beta}$ è ancora funzione della matrice $\mathbf{X}$.

Infatti:

$$I(\boldsymbol{\beta}) = E\left\{\frac{\partial^2 \log L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})}{\partial\beta\partial\beta^{\mathrm{T}}}\right\} = -\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{\sigma^2}$$

- (La matrice delle derivate seconde comunque è costante)

$$V_{\mathrm{inf}}(\mathbf{b}) = -I^{-1}(\boldsymbol{\beta}).$$

Per cui il valore asintotico della matrice di varianze e covarianze di $\mathbf{b}$ coincide con il valore già trovato per via diretta per n qualsiasi.

8.3.3 Distribuzioni sotto H_0 e sotto H_1.

Va sottolineato che nella scomposizione vista prima la quantità(B) ossia:

$$\underline{R(\mathbf{b})} = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})}{\sigma^2}$$

si distribuisce sempre come una v.a. χ^2 con $n - k$ gradi di libertà, sia sotto H_0 che sotto H_1 ; (fatta ovviamente l'assunzione di normalità)

e quindi la stima della varianza:

$$s^2 = R(\mathbf{b})/(n-k) = \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i^*)^2/(n-k)$$

ha sempre una distribuzione proporzionale a quella di una χ^2 con $n-$ k gradi di libertà

Quindi:

$$s^2(n-k)/\sigma^2 \sim \chi^2_k$$

qualunque sia l'ipotesi vera

Infatti $R(\mathbf{b})$ dipende solo dai valori osservati e non dipende dai particolari valori delle componenti del vettore dei parametri $\boldsymbol{\beta}$.

Si noti inoltre che la distribuzione di s^2 non dipende dalla particolare configurazione (scelta a priori o osservata) della matrice $\mathbf{X}$, se non attraverso le sue dimensioni, n e k .

Diversamente la forma quadratica definita dalla quantità (C) ossia:

$$(\mathbf{b}-\boldsymbol{\beta})^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})/\sigma^2$$

si distribuisce come una χ^2 con k gradi di libertà solo se $\boldsymbol{\beta}$ è il vero valore del parametro; Pertanto la forma quadratica a numeratore del test F divisa per i gradi di libertà k

$$s_1^2 = (\mathbf{b}-\beta_0)^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}-\beta_0)/k$$

è uno stimatore corretto di σ^2 solo sotto H_0perché:

$$(\mathbf{b}-\beta_0)^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}-\beta_0)$$

si distribuisce come $\sigma^2\chi^2_k$ soltanto se è vera H_0

Infatti la distribuzione di s_1^2 dipende dal vero valore assunto dai parametri componenti del vettore $\boldsymbol{\beta}$.

Inoltre, come si vede nelle pagine successive e come si intuisce dalle formule di queste pagine, la distribuzione di s_1^2 nel caso generale (ossia sotto H_1 !) dipende anche dalla configurazione della matrice $\mathbf{X}$ (scelta a priori o osservata) attraverso il prodotto $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Pertanto è intuibile, sebbene non tratteremo tale argomento in dettaglio, che la scelta del particolare disegno della matrice $\mathbf{X}$, quando possibile, potrebbe influenzare la distribuzione di s_1^2 sotto H_1, e quindi il potere del test.

In altre parole se per la costruzione di test in particolari contesti sperimentali è necessario operare con certi valori del potere del test, questo obiettivo può essere raggiunto agendo anche sugli elementi della matrice $\mathbf{X}$, ossia sulla configurazione del disegno sperimentale.

In generale se β_0 è il valore specificato dall'ipotesi nulla e se $\boldsymbol{\beta}$ è il vero valore, allora possiamo calcolare il valore atteso della quantità $(\mathbf{b} - \beta_0)^T \mathbf{X}^{\mathrm{T}} \mathbf{X} (\mathbf{b} - \beta_0)$, effettuando alcune manipolazioni della forma quadratica:

$E(\mathbf{b} - \beta_0)^T \mathbf{X}^{\mathrm{T}} \mathbf{X} (\mathbf{b} - \beta_0) =$	Aggiungendo e sottraendo $\boldsymbol{\beta}$
$= E[(\mathbf{b} - \boldsymbol{\beta}) - (\beta_0 - \boldsymbol{\beta})]^T \mathbf{X}^{\mathrm{T}} \mathbf{X} [(\mathbf{b} - \boldsymbol{\beta}) - (\beta_0 - \boldsymbol{\beta})] =$	sviluppiamo il prodotto in cui compare il binomio formato dai due termini: $(\mathbf{b} - \boldsymbol{\beta})$ e $(\beta_0 - \boldsymbol{\beta})$
$= E(\mathbf{b} - \boldsymbol{\beta})^T \mathbf{X}^{\mathrm{T}} \mathbf{X} (\mathbf{b} - \boldsymbol{\beta})$	$= k\sigma^2$ perché la forma quadratica si distribuisce come $\sigma^2 \chi^2_k$ essendo $\boldsymbol{\beta}$ il vero valore
+	
$E(\beta_0 - \boldsymbol{\beta})^T \mathbf{X}^{\mathrm{T}} \mathbf{X} (\beta_0 - \boldsymbol{\beta})$	é la speranza matematica di una costante
-	
$2(\beta_0 - \boldsymbol{\beta})^T \mathbf{X}^{\mathrm{T}} \mathbf{X} (\mathbf{b} - \boldsymbol{\beta}) =$	=0 perché é una combinazione lineare del vettore aleatorio $\mathbf{b} - \boldsymbol{\beta}$, che é a media nulla perché: $E(\mathbf{b}) = \boldsymbol{\beta}$
=	
$k\sigma^2 + (\beta_0 - \boldsymbol{\beta})^T \mathbf{X}^{\mathrm{T}} \mathbf{X} (\beta_0 - \boldsymbol{\beta})$	

Il grafico qui sotto riporta un esempio di distribuzione nulla con due alternative: si tratta di tre densità di F non centrali con 3 e 10 gradi di libertà: la distribuzione nulla è quella corrispondente ad un parametro di non centralità nullo. La linea verticale corrisponde al valore critico per α =0,05

Distribuzione nulla e due alternative per il test F(3,10);

$$\boldsymbol{\alpha} = 0,05; \boldsymbol{\lambda} = 2,5$$

```
\begin{fig}

noncentral1_lucidi.nb

\end{fig}
```

Riassumendo in una tavola questi ultimi risultati:

Quantità		$R(\mathbf{b})$	$R(\beta_0) - R(\mathbf{b})$
Espressioni esplicite		$(\mathbf{y} - \mathbf{Xb})^{\mathrm{T}}(\mathbf{y} - \mathbf{Xb})$	$(\mathbf{b} - \beta_0)^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b} -$
Interpretazione		Devianza residua	Scostamento dall nulla
Speranza matematica	$H_0 : \boldsymbol{\beta} = \beta_0$	$(n-k)\sigma^2$	$k\sigma^2$
	$H_1 : \boldsymbol{\beta} \neq \beta_0$	$(n-k)\sigma^2$	$k\sigma^2 + (\beta_0 - \boldsymbol{\beta})^T\mathbf{X}^{\mathrm{T}}$ $\boldsymbol{\beta})$
Distribuzione	$H_0 : \boldsymbol{\beta} = \beta_0$	$\sigma^2\chi^2_{n-k}$	$\sigma^2\chi^2_k$
	$H_1 : \boldsymbol{\beta} \neq \beta_0$	$\sigma^2\chi^2_{n-k}$	$\sigma^2\chi^2(k,\lambda)$ centrale;λ:parametr di non ce $\lambda = (\beta_0 - \boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}$

Risulta evidente che $E(F(H_1) > E(F(H_0)$ e la regione di rifiuto del test va fissata sulla coda destra della distribuzione di F.

8.3.4 Scomposizione della devianza e test nel caso di gruppi di regressori ortogonali

Se r gruppi di variabili indipendenti sono ortogonali (ossia risultano non correlati linearmente se si tratta di regressori scartati dalla media) la matrice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ risulta composta da r blocchi disposti lungo la diagonale ($r \geq 2$) :

ciascun blocco è composto da un numero qualsiasi k_j di variabili, in modo tale che: $\sum_{j=1}^{r} k_j = k$;

Per esempio, termine noto e regressori, $r = 2, k_1 = 1$;

In particolare se tutti i k_j sono uguali ad uno, vuol dire che tutte le variabili risultano ortogonali

eventualmente gli indici delle variabili sono permutati in modo che le variabili di uno stesso gruppo siano consecutive

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{X}_2^T\mathbf{X}_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{X}_j^T\mathbf{X}_j & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{X}_r^T\mathbf{X}_r \end{pmatrix}$$

Ad esempio tutte le variabili del 1° blocco sono ortogonali a tutte quelle del j -esimo gruppo; all'interno di ciascun gruppo le variabili non sono ortogonali (o comunque non tutte). In corrispondenza a questi r blocchi suddividiamo il vettore dei parametri $\boldsymbol{\beta}$ e quello delle stime $\mathbf{b}$.

$$\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \dots, \boldsymbol{\beta}_j^{\mathrm{T}}, \dots, \boldsymbol{\beta}_r^{\mathrm{T}})$$

$$\mathbf{b}^{\mathrm{T}} = (\mathbf{b}_1^{\mathrm{T}}, \mathbf{b}_2^{\mathrm{T}}, \dots, \mathbf{b}_j^{\mathrm{T}}, \dots, \mathbf{b}_r^{\mathrm{T}})$$

Il vantaggio per l'inferenza è che i gruppi di stimatori dei corrispondenti parametri saranno a blocchi non correlati (indipendenti data l'assunzione di normalità):

$$Cov(\mathbf{b}_j, \mathbf{b}_l) = 0 (j \neq l)$$

Dal punto di vista numerico, ciascun gruppo di stime è ricavabile da un sottoinsieme di equazioni normali:

$$(\mathbf{X}_j^T\mathbf{X}_j)\mathbf{b}_j = \mathbf{X}_j^T\mathbf{y} equindi :$$

$$\mathbf{b}_j = (\mathbf{X}_j^T\mathbf{X}_j)^{-1}\mathbf{X}_j^T\mathbf{y}$$

è la matrice $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ risulta ora diagonale a blocchi:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{pmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1} & 0 & 0 & 0 & 0 & 0 \\ 0 & (\mathbf{X}_2^T\mathbf{X}_2)^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & (\mathbf{X}_j^T\mathbf{X}_j)^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & (\mathbf{X}_r^T\mathbf{X}_r)^{-1} \end{pmatrix}$$

La matrice di varianze e covarianze di $\mathbf{b}$ è data da:

$$V(\mathbf{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1},$$

Per cui possiamo scrivere, moltiplicando $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ per lo scalare σ^2 :

$$V(\mathbf{b}) = \begin{pmatrix} V(\mathbf{b}_1) & 0 & 0 & 0 & 0 & 0 \\ 0 & V(\mathbf{b}_2) & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & V(\mathbf{b}_j) & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & V(\mathbf{b}_r) \end{pmatrix}$$

In generale è possibile scomporre semplicemente la forma quadratica $(\mathbf{b}-\boldsymbol{\beta})^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})$ in r forme quadratiche (due o più) mutuamente indipendenti, se e solo se la matrice $\mathbf{X}$ può essere partizionata in r gruppi di regressori non correlati nel modo visto.

Possiamo in questo caso esprimere la forma quadratica:

$$\mathbf{Q}(\mathbf{b}-\boldsymbol{\beta}) = (\mathbf{b}-\boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}) =$$

$$= \sum_{j=1}^{r}(\mathbf{b}_j-\boldsymbol{\beta}_j)^{\mathrm{T}}\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j(\mathbf{b}_j-\boldsymbol{\beta}_j) = \sum_{j=1}^{r}\mathbf{Q}(\mathbf{b}_j-\beta_j);$$

Evidentemente le singole forme quadratiche si distribuiscono come delle variabili aleatorie χ^2 con k_j gradi di libertà moltiplicate per σ^2 e sono indipendenti;

Ovviamente sono anche indipendenti rispetto a $R(\mathbf{b})$

per cui le scomposizioni viste prima in questo caso si estendono ulteriormente, scomponendo ciascun termine in r termini.

Si possono quindi costruire dei test F_j con numeratori indipendenti, mettendo a denominatore sempre s^2 (stima corretta della varianza) ed a numeratore l'opportuna forma quadratica $\mathbf{Q}(\mathbf{b}_j-\beta_j)$ divisa per i rispettivi gradi di libertà k_j :

$$F_j = \frac{\frac{[\mathbf{b}_j-\beta_j]^T\mathbf{X}_j^T\mathbf{X}_j[\mathbf{b}_j-\beta_j]}{k_j}}{\frac{[\mathbf{y}-\mathbf{X}\mathbf{b}]^T[\mathbf{y}-\mathbf{X}\mathbf{b}]}{n-k}} = \frac{\frac{\mathbf{Q}(\mathbf{b}_j-\beta_j)}{k_j}}{s^2}$$

I rapporti F_j si distribuiscono secondo una F di Snedecor con k_j ed $n-k$ gradi di libertà

Pertanto, per saggiare un'ipotesi nulla:

$$H_0 : \boldsymbol{\beta} = \beta_0,$$

contro l'alternativa generica

$$H_1 : \boldsymbol{\beta} \neq \beta_0,$$

nel caso di r blocchi ortogonali, si può considerare anche per il vettore β_0 la stessa suddivisione in blocchi:

$$\beta_0^T = \beta_{10}^T \beta_{20}^T \dots \beta_{j0}^T \dots \beta_{r0}^T$$

Per cui l'ipotesi nulla può essere suddivisa in r ipotesi,

$$H_{j0} : \beta_j = \beta_{j0} j = 1, 2, \dots, r$$

per saggiare ciascuna delle quali possiamo impiegare i test:

$$F_j = \frac{\frac{[\mathbf{b}_j - \beta_{j0}]^T \mathbf{X}_j^T \mathbf{X}_j [\mathbf{b}_j - \beta_{j0}]}{k_j}}{\frac{(\mathbf{y}-\mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\mathbf{b})}{n-k}}, j = 1, 2, \dots, r;$$

ognuno dei quali sotto H_0 si distribuisce secondo una variabile aleatoria F di Snedecor con k_j ed $n - k$ gradi di libertà.

Questi test sono indipendenti. E' possibile che l'ipotesi nulla specifichi solo alcuni gruppi di parametri, e non tutti.

Es. $H_0 : \beta_s = \beta_{s0}; \beta_j$ qualsiasi per $j \neq s$

In particolare può interessare:

$$H_0 : \beta_s = 0$$

Rispetto al test che si condurrebbe in presenza di un solo gruppo di regressori, cambia solo a denominatore la stima della varianza, che ha $n-k$ gradi di libertà invece che $n-k_s$. In ogni caso è meglio procedere con la stima con $n - k$ gradi di libertà che è certamente

corretta

Se a ciascun gruppo di parametri e di regressori si può fare corrispondere una diversa fonte di variabilità, questo implica che per fare inferenza riguardo a ciascuna componente, indipendentemente dalle altre, è necessario che il gruppo di regressori corrispondente a ciascuna sorgente di variazione risulti ortogonale rispetto ai regressori corrispondenti alle altre sorgenti di variabilità.

Questi aspetti sottolineano l'importanza di operare, quando possibile, con regressori ortogonali, almeno a gruppi, perché questo implicherà essenzialmente:
L'indipendenza fra i corrispondenti gruppi di stimatori;
L'indipendenza approssimata fra i test relativi ai vari gruppi di parametri, ossia alle differenti sorgenti di variabilità

```
\begin{fig}
 Esempio di fattori ortogonali da
STATISTICA Esempi ripresi dai problemi introduttivi
\end{fig}
```

8.4 Configurazioni della matrice X e di $\mathbf{X}^{\mathrm{T}}\mathbf{X}$

$\mathbf{X}$	$\mathbf{X}^{\mathrm{T}}\mathbf{X}$	Significato e conseguenze per l'interpretazione del modello e per l'inferenza
	Tutte le $\mathbf{X}_j$ sono ortogonali	Diagonale È il caso migliore: si possono saggiare ipotesi e fare inferenza in generale sui singoli parametri in modo indipendente (anche i valori degli stimatori si trovano in modo indipendente)
Tutte le combinazioni di valori dei fattori	Fattoriale	Meglio ancora! Fra l'altro migliorano le proprietà delle regioni di confidenza costruite su $E(\mathbf{y}_i)$
Gruppi di $\mathbf{X}_j$ sono ortogonali	Diagonale a blocchi	È un caso importante: si possono saggiare ipotesi (e fare inferenza in generale) su gruppi di parametri separatamente
Correlazioni lineari generiche fra le $\mathbf{X}$	A rango pieno ma non diagonale	È il caso generale della regressione multipla, in particolare per studi osservazionali.
Qualcuna delle $\mathbf{X}_j$ è fortemente dipendente linearmente dalle altre $\mathbf{X}_j$	A rango pieno ma con qualche autovalore vicino a zero	MULTICOLLINEARITA'
Alcune $\mathbf{X}_j$ indicano la presenza/assenza di livelli di un fattore	A rango non pieno	Per costruzione: Alcuni casi di Analisi della varianza etc.
Alcune variabili sono esattamente proporzionali	A rango non pieno	Per errore di rilevazione (si tolgono le variabili ridondanti)

8.5 Modello lineare: Verifica di ipotesi generali

Comunque sia configurata la matrice $\mathbf{X}$ e quindi $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, non sempre l'ipotesi d'interesse riguarda tutti i parametri.

In generale siamo interessati a verificare ipotesi relativi a sottoinsiemi di valori dei parametri, come ad esempio:

$H_0 : \beta_1 = \beta_2 = 0; \beta_j$ qualsiasi per $j > 2$

comunque

$H_0 : \boldsymbol{\beta}_s = \boldsymbol{\beta}_{s\ 0}; \beta_j$ qualsiasi per $j \neq s$

relativa ad un gruppo di parametri β_s

Può però interessarci un'ipotesi che implichi un confronto fra i valori di alcuni parametri; ad esempio:

$$H_0 : \beta_1 = \beta_2 = \beta_3 (= \mu;$$

con μ non specificato) e β_j qualsiasi per $j > 3$.

quest'ultima ipotesi equivale ad imporre i due vincoli:

$$\beta_1 - \beta_3 = 0$$

$$\beta_2 - \beta_3 = 0$$

In effetti queste ipotesi nulle possono essere considerate come delle ipotesi che impongono dei vincoli lineari (anche molto generali) sui valori dei k parametri, secondo la relazione generale:

$$\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}_0$$

In dettaglio, dato il modello:

$$\mathbf{y}_{[n\times 1]} = \mathbf{X}_{[n\times k]}\beta_{[k\times 1]} + \boldsymbol{\varepsilon}_{[n\times 1]}$$

(supponiamo sempre $\mathbf{X}$ di rango k) in generale siamo interessati a verificare l'ipotesi:

$$H_0 : \mathbf{C}_{[q\times k]}\beta_{[k\times 1]} = \mathbf{C}\beta_0 = \mathbf{a}_{[q\times 1].}$$

Con $q < k$ e q rango di $\mathbf{C}$

Esempio: Analisi della varianza ad una via .

Si riveda l'impostazione della matrice $\mathbf{X}$ nella parte introduttiva sui modelli lineari; La matrice $\mathbf{X}$ è composta da k colonne indicatrici dell'appartenenza delle n unità a k gruppi disgiunti.

La parametrizzazione più naturale è quella in cui ogni parametro corrisponde al valor medio di $\mathbf{Y}$ in ciascun gruppo:

$$\boldsymbol{\beta}^{\mathrm{T}} = \mu_1, \ldots, \mu_j, \ldots, \mu_k$$

L'ipotesi che può interessare non è però che tutti i coefficienti siano nulli, ma che siano uguali fra loro:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$$

Queste $k-1$ uguaglianze corrispondono ad una scelta di $\mathbf{C}$ di $k-1$ righe e k colonne:

vincolo		$Gr.1$	$Gr.2$	$\ldots$	$Gr.J$	$\ldots$	$Gr.K$
1		1	0	$\ldots$	0	$\ldots$	-1
2		$\ldots$	1	$\ldots$	$\ldots$	$\ldots$	-1
controllare	$\mathbf{C}_{[k-1\times k]} =$	0	0	$\ldots$	0	0	-1
j		0	0	$\ldots$	1	0	-1
controllare		$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	-1
$k-1$		0	0	0	0	$\ldots$	-1

$$\mathbf{C}_{[k-1\times k]} = \begin{pmatrix} 1 & 0 & \ldots & 0 & \ldots & \ldots & -1 \\ \ldots & 1 & \ldots & \ldots & \ldots & \ldots & -1 \\ 0 & 0 & \ldots & 0 & \ldots & 0 & -1 \\ 0 & 0 & \ldots & 1 & \ldots & 0 & -1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & -1 \\ 0 & 0 & \ldots & 0 & \ldots & 1 & -1 \end{pmatrix}$$

con $\mathbf{a} = 0_{k-1}$
Scrivere ora $\mathbf{C}\boldsymbol{\beta} = \mathbf{a}$ è come scrivere:

$$\mu_1 - \mu_k = \mu_2 - \mu_k = \ldots = \mu_j - \mu_k = \ldots = \mu_{k-1} - \mu_k = 0.$$

Riprendiamo l'esempio sull'ipotesi nulla:

$$H_0 : \beta_1 = \beta_2 = \beta_3 (= \mu;$$

con μ non specificato) e β_j qualsiasi per $j > 3$.
La matrice dei vincoli è costituita da due sole righe:

vincolo		Gr. 1	Gr. 2	Gr. 3	Gr. J	...	Gr. K
1	$\mathbf{C}_{[2\times k]}=$	1	0	-1	0	...	0
2		0	1	-1	0	...	0

$$\mathbf{C}_{[2\times k]} = \begin{pmatrix} 1 & 0 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & 0 & \dots & 0 \end{pmatrix}$$

con $\mathbf{a} = (0,0)^T$ Altro esempio:
Se l'ipotesi di interesse è:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

questo corrisponde a scegliere:

$$\mathbf{C} = \mathbf{I}_k; \mathbf{a} = 0_k.$$

Esempio.

In un modello di regressione multipla si può avere un problema di scelta di variabili (vedere dopo).

L' ipotesi:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0, q < k;$$

e
$\beta_{q+1}, \beta_{q+2}, \ldots, \beta_k$ qualsiasi
corrisponde a q vincoli definiti da:

$$\mathbf{C} = \mathbf{I}_q : 0_{k-q}; \mathbf{a} = 0_k$$

ossia i vincoli non coinvolgono i $k-q$ regressori oltre β_q .

Ovviamente $q = 1$ nel caso di ipotesi concernenti un singolo parametro.

La stima dei parametri del modello lineare con vincoli lineari sui parametri

In questo caso per costruire il rapporto di verosimiglianza per la verifica dell'ipotesi generale:

$$H_0 : \mathbf{C}_{[q\times k]}\beta_{[k\times 1]} = \mathbf{a}_{[q\times 1]}.\mathbf{C}dirangoq$$

(con H_1 : ipotesi alternativa che non fissa alcun vincolo sui parametri) si ha:

$$LR = \frac{\max L[\boldsymbol{\beta}, \sigma^2, \mathbf{y}|H_0]}{\max L[\boldsymbol{\beta}, \sigma^2, \mathbf{y}|H_1]} =$$

$$\frac{\max L[\boldsymbol{\beta}, \sigma^2, \mathbf{y}|\mathbf{C}\boldsymbol{\beta} = \mathbf{a}]}{max L[\boldsymbol{\beta}, \sigma^2, \mathbf{y}|\boldsymbol{\beta} \epsilon \Re_k]} =$$

$$= \left\{ \frac{R(\mathbf{b}_0)}{R(\mathbf{b})} \right\}^{-\frac{n}{2}}$$

essendo $\mathbf{b}$ lo stimatore di massima verosimiglianza non vincolato, e $\mathbf{b}_0$ lo stimatore di massima verosimiglianza sotto i vincoli lineari imposti da H_0.

Minimi quadrati vincolati

Per trovare $\mathbf{b}_0$ occorre risolvere un problema di minimi quadrati vincolati:

$$min_{b0} R(\mathbf{b}_0) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)^T (\mathbf{y} - \mathbf{X}\mathbf{b}_0) =$$

$$\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2b_0^T \mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{b}_0^T (\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b}_0$$

soggetto a q vincoli lineari:

$$\mathbf{C}\mathbf{b}_0 = \mathbf{a}; \mathbf{C} \text{di rango} q$$

Occorre introdurre q moltiplicatori di Lagrange $2vecd_h$ ed uguagliare a 0 le derivate di $\mathbf{Q}(\mathbf{b}_0)$ rispetto al vettore $\mathbf{b}_0$ e al vettore $\mathbf{d}_{[q\times 1]}$:

$$\mathbf{Q}(\mathbf{b}_0, d) = R(\mathbf{b}_0) + 2(\mathbf{C}\mathbf{b}_0 - \mathbf{a})^T \mathbf{d} \frac{\mathbf{Q}}{\mathbf{b}_0} = -$$

$$2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b}_0 + 2\mathbf{C}^{\mathrm{T}}\mathbf{d}$$

$$\frac{\mathbf{Q}}{d} = (\mathbf{C}\mathbf{b}_0 - \mathbf{a}$$

Uguagliandole a 0 (vettore nullo):

$$-2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b}_0 + 2\mathbf{C}^{\mathrm{T}}\mathbf{d} = 0;$$

$$(\mathbf{C}\mathbf{b}_0 - \mathbf{a} = 0;$$

dal primo gruppo di equazioni:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{b}_0 = \mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{C}^{\mathrm{T}}\mathbf{d};$$

$$\mathbf{b}_0 = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{d} =$$

(sostituendo $\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, stimatore dei minimi quadrati non vincolato)

$$= \mathbf{b} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{d}$$

Dal secondo gruppo di equazioni:

$$\mathbf{C}\mathbf{b}_0 = \mathbf{a} = \mathbf{C}\mathbf{b} - \mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{d};$$

Sono q equazioni indipendenti in k incognite $\mathbf{d}$,

$$\mathbf{C}\mathbf{b} = -\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{d};$$

con soluzione data da:

$$-d = [\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}(\mathbf{a} - \mathbf{C}\mathbf{b})$$

risostituendo nel sistema che fornisce $\mathbf{b}_0$ si ha:

$$\mathbf{b}_0 = \mathbf{b} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{d} =$$

$$= \mathbf{b} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}(\mathbf{a} - \mathbf{C}\mathbf{b})$$

Si può facilmente vedere che questa soluzione fornisce il minimo e rispetta i vincoli (premoltiplicando per $\mathbf{C}$)

Tutte le inverse citate esistono, per le ipotesi fatte sui ranghi di $\mathbf{X}$ e $\mathbf{C}$.

In realtà di solito conviene risolvere il sistema dei minimi quadrati secondo la parametrizzazione fornita da H_0, se questa è esplicitabile rispetto ai parametri. La tecnica ora esposta per trovare $\mathbf{b}_0$ è utile prevalentemente a scopo teorico per vedere la relazione fra $\mathbf{b}_0$ e $\mathbf{b}$; Inoltre è utile per i casi nei quali $\mathbf{C}\boldsymbol{\beta} = \mathbf{a}$ non sia semplicemente esplicitabile.

Nell' espressione di $\mathbf{b}_0$esplicitiamo, in modo che sia evidente la relazione lineare fra $\mathbf{b}_0$ e $\mathbf{b}$:

Ponendo, per semplicità: $F = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}$

si ottiene

$$\mathbf{b}_0 = \mathbf{b}+(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}(\mathbf{a}-\mathbf{C}\mathbf{b}) == F.\mathbf{a}+(\mathbf{I}_k-F.\mathbf{C})\mathbf{b}$$

$\mathbf{b}_0$risulta corretto solo sotto H_0

Infatti

$$E(\mathbf{b}_0) = \boldsymbol{\beta} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}(\mathbf{a} - \mathbf{C}\boldsymbol{\beta}) = \boldsymbol{\beta}$$

perchè sotto $H_0 : \mathbf{a} - \mathbf{C}\boldsymbol{\beta} = 0$

Inoltre per la matrice di varianze e covarianze si ha in generale:

$$V(\mathbf{b}_0) = (\mathbf{I}_k - F.\mathbf{C})V(\mathbf{b})(\mathbf{I}_k - F.\mathbf{C})^T == \sigma^2[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} - F.\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}F^T + F.\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X}$$

Questi tre termini risultano uguali in valore assoluto.

Infine, dopo qualche semplificazione:

$$V(\mathbf{b}_0) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} - \sigma^2[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}] =$$

$$= V(\mathbf{b}) - \sigma^2[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}].$$

- Le varianze di ciascun elemento di $\mathbf{b}_0$risultano inferiori a quelle dei corrispondenti elementi di $\mathbf{b}$;
- Si ricordi però che in generale b_0 è distorto.

Modello lineare: Scomposizione della devianza per il problema soggetto a vincoli:

Anche in questo caso la devianza residua può essere scomposta in una forma conveniente

Alcune scomposizioni:

$R(\mathbf{b}_0) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)^T(\mathbf{y} - \mathbf{X}\mathbf{b}_0) =$	Sommando e sottraendo Xb e poi aprendo il quadrato del binomio
$= [(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_0)]^T[(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_0)] =$	
$= (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})$	=R(**b**)
$+$	
$(\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_0)^T(\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_0)$	si mette in evidenza $\mathbf{X}$ sia a sinistra che a destra e si ottiene $(\mathbf{b} - \mathbf{b}_0)^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})$
$+$	
$(\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_0)$	=0 perché:$(\mathbf{y} - \mathbf{X}\mathbf{b})^T\mathbf{X} = 0$ dalle equazioni dei minimi quadrati
$=$	
$R(\mathbf{b}) + (\mathbf{b} - \mathbf{b}_0)^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b} - \mathbf{b}_0)$	

In definitiva:

$$R(\mathbf{b}_0) = R(\mathbf{b}) + (\mathbf{b} - \mathbf{b}_0)^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b} - \mathbf{b}_0)$$

...

$(\mathbf{b}-\mathbf{b}_0)^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}-\mathbf{b}_0)$: Devianza residua supplementare dovuta ad H_0. Misura anche la distanza fra i due stimatori.

E inoltre, sostituendo l' espressione di $(\mathbf{b}-\mathbf{b}_0)$:

$$R(\mathbf{b}_0)-R(\mathbf{b})=(\mathbf{a}-\mathbf{Cb})^T[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}(\mathbf{a}-\mathbf{Cb})$$

Si distribuisce (sotto H_0) come una χ^2 con q gradi di libertà, indipendentemente da $R(\mathbf{b})$.

Pertanto è possibile costruire test per la verifica di una ipotesi qualsiasi semplicemente mettendo a numeratore del test F l'incremento di devianza dovuto ad H_0 (e modificando i gradi di libertà)

8.5.1 Prove di ipotesi particolari nel modello lineare

Se la matrice $\mathbf{C}$ è costituita da:

$$\mathbf{C}=\mathbf{I}_q;0_{q\times k}$$

(ossia specifica solo i valori di q parametri)

la matrice $(\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}})^{-1}$ ora risulta costituita dall'inversa del blocco $q\times q$della matrice $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ corrispondente ai q parametri specificati da H_0,ossia $[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}{}_q]^{-1}$

il vettore di q elementi $(\mathbf{a}-\mathbf{Cb})$ è semplicemente costruito dalla differenza fra valori ipotizzati e valori stimati sotto H_0.

$[\mathbf{b}_0]_q^T$ indica il vettore di q elementi coinvolto dall'ipotesi nulla particolare.$[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}{}_q]$ indica il blocco$q\times q$ nella matrice $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$

$$F=\frac{\frac{[\mathbf{b}-\beta_0]_q^T[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}{}_q]^{-1}[\mathbf{b}-\beta_0]_q}{q}}{\frac{[\mathbf{y}-\mathbf{Xb}]^T[\mathbf{y}-\mathbf{Xb}]}{n-k}}=$$

In ogni caso il rapporto:

$$F=\frac{\frac{R(\mathbf{b}_0)-R(\mathbf{b})}{q}}{\frac{R(\mathbf{b})}{n-k}}$$

si distribuisce (sotto H_0) come una F con q ed $n-k$ gradi di libertá , se è valida l' ipotesi nulla:$H_0:\boldsymbol{\beta}=\beta_0$. (con q numero di gradi di libertá del numeratore)

Ovviamente si vede facilmente che questo rapporto è funzione del rapporto delle verosimiglianze.

Va precisato che questo approccio va bene per saggiare ipotesi singole, anche concernenti q parametri, ma non gruppi di ipotesi, perché i test relativi a sottoinsiemi differenti di parametri (o di loro combinazioni lineari) non sono indipendenti, se non nel caso visto prima di matrice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ a blocchi diagonali. Condurre in parallelo test separati sugli elementi di $\boldsymbol{\beta}$ in assenza dei necessari requisiti di ortogonalità è in generale una procedura errata, nel senso che non vengono certamente rispettati i livelli di significatività nominali. Può essere utile, in analisi esplorative, a titolo comparativo, per confrontare verosimiglianze relative a modelli concorrenti, ma non per effettuare test nel vero senso del termine.

8.6 Test e regioni di confidenza nei modelli lineari

L'approccio visto prima, sui test LR per ipotesi che impongono q vincoli lineari sui parametri, a rigore va impiegato solo per saggiare un'ipotesi concernente un unico set di parametri;

oppure occorre avere set di ipotesi ortogonali In generale se $k > 1$ non esiste un test UMPU.

Regioni di confidenza simultanee per i parametri

La regione di confidenza migliore, ad un livello $1-\alpha$, è determinata dai valori $\boldsymbol{\beta}$ per i quali i valori osservati del test F non risultano superiori al valore teorico $F_{\alpha,k,n-k}$.

Pertanto, dato un campione nel quale $\mathbf{b}$ è la stima di massima verosimiglianza, tale regione è delimitata dai valori $\boldsymbol{\beta}$ per i quali:

$$[(\mathbf{b}-\boldsymbol{\beta})^T\mathbf{X}^{\mathrm{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})](ks^2F_{\alpha,k,n-k}$$

Nel caso di regressori non ortogonali, tali regioni risulteranno date da ellissoidi con assi obliqui, per cui l'interpretazione delle regioni stesse potrà essere ardua.

Anche la relazione con i singoli intervalli sarà di difficile interpretazione, infatti per ciascun valore di uno dei parametri, l'intervallo ottimo dell'altro varia, sia per posizione che per estensione.

Intervalli di confidenza e regioni di confidenza

Come si è visto la struttura di correlazione fra gli estimatori dei parametri è strettamente dipendente dalla struttura di correlazione dei regressori o comunque dalla struttura della matrice $\mathbf{X}$. Le regioni di confidenza che costruiremo per i parametri $\boldsymbol{\beta}$ saranno ovviamente quelli ellisoidali, data la normalità, ma con una inclinazione degli assi principali che dipenderà dalla correlazione fra le diverse componenti dello stimatore di $\boldsymbol{\beta}$.

È il caso adesso di riflettere sulle differenze concettuali e interpretative che esistono fra regioni di confidenza e intervalli di confidenza, l'intersezione, infatti, fra intervalli di confidenza costruiti singolarmente o per ciascun parametro anche nel caso di assenza di correlazione, conduce a risultati e ad interpretazioni diverse da quelle ottenute mediante regioni di confidenza simultanee. Si consideri, infatti, la figura seguente: dai dati dell'esempio si sono costruiti gli intervalli di confidenza per β_1 e β_2 ad un livello fiduciario di α; inoltre si è costruita la regione di confidenza simultanea per i due parametri ricavata dalla relazione vista nel paragrafo precedente, fondata sui percentili della distribuzione F.

Occorre intanto riportare le due situazioni a parità di livello di copertura ossia fare in modo che la probabilità fiduciaria complessiva dei due intervalli sia uguale alla probabilità fiduciaria della regione ellissoidale; le due situazioni o meglio i due approcci conducono a conclusioni leggermente differenti ma non contrastanti in modo stridente; il punto fondamentale consiste nell'avere in un caso un'intersezione fra segmenti che conduce ad un rettangolo e nell'altro caso una circonferenza o in generale un'ellisse con assi paralleli agli assi coordinati la differenza di area coperta è, in effetti, molto bassa. Nell'esempio si può calcolare come riportato nella figura.

Si consideri invece un esempio nel quale gli stimatori dei due parametri β_1 e sono molto correlati; in questo caso la regione di confidenza simultanea sarà costituita da un ellissoide su delle con assi non paralleli a quelli coordinati; la discrepanza fra la superficie coperta da quest'ellisse e quella coperta dall'intersezione tra i due segmenti è ora più forte;

Inoltre esiste un problema d'interpretazione molto grosso: secondo del valore assunto dal parametro β_1, l'intervallo di confidenza ottimo per il parametro β_2 è differente, non solo per ampiezza ma anche per posizione; d'altra parte il fatto che due stimatori risultino

correlati significa proprio che non è possibile fare inferenze separatamente sulle due singole componenti. La relazione con i singoli intervalli sarà di difficile interpretazione, infatti per ciascun valore di uno dei parametri, l'intervallo ottimo dell'altro varia, sia per posizione che per estensione.

VEDERE GRAFICI AGGIUNTIVI NEL FILE:

```
\begin{fig}

DISPENSA2000_FIGURE2.DOC

\end{fig}
```

regioni di confidenza per funzioni lineari dei parametri

In effetti se siamo interessati a particolari combinazioni di parametri $\mathbf{a} = \mathbf{C}\,\boldsymbol{\beta}$, possiamo direttamente costruire regioni di confidenza per tali funzioni lineari dei parametri a partire dalla quantità:

$$R(\mathbf{b}_0) - R(\mathbf{b}) = (\mathbf{a} - \mathbf{C}\mathbf{b})^T[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}(\mathbf{a} - \mathbf{C}\mathbf{b});$$

Prendendo in considerazione il corrispondente test F si può direttamente costruire la regione (q-dimensionale) costituita da tutti i valori $\mathbf{a}$ per i quali:

$$(\mathbf{a} - \mathbf{C}\mathbf{b})^T[\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}}]^{-1}(\mathbf{a} - \mathbf{C}\mathbf{b})(qs^2F_{\alpha,q,n-k}$$

regioni di confidenza relative a sottoinsiemi di parametri

Se la matrice $\mathbf{C}$ è definita da:

$$\mathbf{C} = \mathbf{I}_q; 0_{q\times k}$$

(ossia specifica solo i valori di q parametri), allora:

la matrice$(\mathbf{C}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{C}^{\mathrm{T}})^{-1}$ risulta costituita dall'inversa del blocco $q \times q$ della matrice $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ corrispondente ai q parametri specificati da H_0,ossia$[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}{}_q]^{-1}$

il vettore di q elementi $(\mathbf{a} - \mathbf{C}\mathbf{b})$ è semplicemente costruito dalla differenza fra valori dei parametri e valori degli stimatori per soli q dei k parametri.

La regione (q-dimensionale) è quindi costituita dai valori di β_q per i quali:

$$[\mathbf{b} - \boldsymbol{\beta}]_q^T[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}{}_q]^{-1}[\mathbf{b} - \boldsymbol{\beta}]_q(qs^2F_{\alpha,q,n-k}$$

$[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}{}_q]$ indica il blocco $q \times q$nella matrice $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

$[\mathbf{b} - \beta]_q$ indica l'opportuno sottovettore di q elementi

Intervalli di confidenza per $E(\mathbf{y}_i)$

Per quanto visto prima, è evidente che lo stimatore migliore di $E(\mathbf{y}_i)$ è $\mathbf{y}_i* = \mathbf{x}_i^{\mathrm{T}}\mathbf{b}$, essendo $\mathbf{x}_{(i)}$ il vettore di osservazioni dei regressori corrispondente all'unità i-esima, e quindi rientriamo nel caso i combinazioni lineari degli stimatori $\mathbf{b}$.

Pertanto, e comunque se il modello è completo e corretto:

$$E(\mathbf{y}_i*) = E(\mathbf{x}_i^{\mathrm{T}}\mathbf{b}) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} = E(\mathbf{y}_i)$$

$$V(\mathbf{y}_i*) = V(\mathbf{x}_i^{\mathrm{T}}\mathbf{b}) = \mathbf{x}_i^{\mathrm{T}}V(\mathbf{b})\mathbf{x}_{(i)} = \sigma^2\mathbf{x}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{(i)}$$

essendo al solito $\mathbf{x}_i^{\mathrm{T}}$ l'i-esima riga della matrice $\mathbf{X}$.

Applicando quindi le formule dei paragrafi precedenti, otteniamo l'intervallo di confidenza per $E(\mathbf{y}_i)$ ad un livello di probabilità fiduciaria $1 - \alpha$, dato da:

$$\mathbf{x}_i^{\mathrm{T}}\mathbf{b}(st_{\alpha,n-k}\sqrt{\mathbf{x}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{(i)}}.$$

Risulta dunque evidente che il luogo dei punti $\mathbf{x}_{(i)}$ per i quali tali intervalli risultano di uguale ampiezza, a parità di altre condizioni, è costituito dai punti per i quali

$$\mathbf{x}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{(i)} = \text{Costante},$$

ossia dai punti che hanno uguale distanze di MahalaNobis dal centroide dei regressori.

```
\begin{fig}

esempi nel notebook mathematica

\end{fig}
```

Nelle figure allegate sono mostrati gli effetti dovuti a configurazioni diverse delle $\mathbf{X}$.

```
\begin{fig}

DISPENSA2000_FIGURE3.DOC

\end{fig}
```

errori di previsione

Varianza degli errori di previsione e distorsione degli stimatori variano in senso opposto

```
\begin{fig}

INSERIRE LUCIDO FATTO A MANO

(che si trova nel blocco dopo la regressione)

esempio da rivedere e ripetere in aula

\end{fig}

\begin{fig}

regr1.ppt

\end{fig}
```

Capitolo 9

Regressione Multipla

9.1 Introduzione

Nei capitoli precedenti si è vista la teoria generale sull'inferenza nei modelli lineari, sia nel caso di modelli con vincoli che senza vincoli. Precedentemente avevamo visto come in realtà i modelli lineari siano utilizzabili per diversi problemi statistici, in funzione della particolare costruzione e configurazione della matrice $\mathbf{X}$; in questo capitolo affrontiamo il caso specifico dei modelli di regressione, e le peculiarità dell'inferenza in questo caso, insieme con una selezione dei problemi inferenziali più comunemente affrontati nelle applicazioni reali. Ricordo che nella pratica dello statistico le tecniche di regressione lineare multipla costituiscono una costante che capita di affrontare in numerosi problemi, almeno come tecnica preliminare di esplorazione dei dati.

Scomposizione della devianza empirica col termine noto e k regressori a media nulla:

Se la matrice $\mathbf{X}$ prevede una colonna di costanti uguali ad uno e altre k colonne a media nulla, abbiamo un modello con termine noto e con matrice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$partizionata a due blocchi diagonali:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} == \begin{pmatrix} n & \mathbf{0}_k^{\mathrm{T}} \\ 0_k & nS_{\mathbf{X}}(= \mathbf{Z}^{\mathrm{T}}\mathbf{Z}) \end{pmatrix}$$

Quindi tutte le forme quadratiche che hanno come matrice dei coefficienti questa matrice con $(k+1) \times (k+1)$ elementi, saranno scomponibili in una forma quadratica con matrice di $k \times k$ elementi, ed un termine singolo.

(Indichiamo ora il termine noto con α, ed il corrispondente stimatore con a, invece che con β_0 per evitare confusione con i valori β_0 dell'ipotesi nulla; con $\boldsymbol{\beta}$ indico il vettore dei parametri relativo alle k variabili e con $\mathbf{b}$ il corrispondente stimatore dei minimi quadrati);

Chiaramente risulta: $a = M_{\mathbf{y}}$

Per quanto riguarda la scomposizione della devianza empirica di $\mathbf{y}$ nel modello di regressione multipla, possiamo partire dalla relazione trovata fra $R(\mathbf{b})$ e la somma dei quadrati $\mathbf{y}^{\mathrm{T}}\mathbf{y}$. (in effetti adesso dovremmo indicarlo con $R(a, \mathbf{b})$)

$$R(\mathbf{b}) = \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i^*)^2 = (\mathbf{y} - a.1 - \mathbf{Zb})^T(\mathbf{y} - a.1 - \mathbf{Zb}) =$$

$$= \mathbf{y}^{\mathrm{T}}\mathbf{y} - nM_{\mathbf{y}}^2 - \mathbf{b}^{\mathrm{T}}\mathbf{Z}^T Zb =;$$

dato che $a = My$.

```
--------------------------
controllare Zb e
M_{\vecy}
--------------------------
```

Possiamo anche scrivere:

$$R(\mathbf{b}) = (\mathbf{y} - M_{\mathbf{y}})^T(\mathbf{y} - M_{\mathbf{y}}) - \mathbf{b}^{\mathrm{T}}\mathbf{Z}^T Zb.$$

Quindi nei modelli di regressione multipla, per eliminare l'influenza del termine noto, che svolge il ruolo di parametro di disturbo, si può direttamente lavorare in termini di scarti, sia per le x che per $\mathbf{y}$.

In ogni caso sarà possibile fare inferenza indipendente su questo termine.

T AVOLA DI SCOMPOSIZIONE DELLA DEVIANZA EMPIRICA NELLA REGRESSIONE

Z è la matrice degli scarti dalle medie		
TOTALE	RESIDUA	SPIEGATA
$(\mathbf{y} - M_{\mathbf{y}})^T(\mathbf{y} - M_{\mathbf{y}})$	$(\mathbf{y} - M_{\mathbf{y}} - Zb)^T(\mathbf{y} - M_{\mathbf{y}} - Zb)$	$\mathbf{b}^{\mathrm{T}}\mathbf{Z}^T Zb$
$\sum_{i=1}^n(\mathbf{y}_i - M_{\mathbf{y}})^2$	$\sum_{i=1}^n(\mathbf{y}_i - \mathbf{y}_i^*)^2$	$\sum_{i=1}^n(\mathbf{y}_i^* - M_{\mathbf{y}})^2$
Devianza totale osservata di **y**	devianza residua (deviazioni dal valore stimato)	Devianza spiegata dalla regressione lineare sui k regressori (presi globalmente)
gradi di libertà:		
$n-1$	$n-k-1$	k

Il coefficiente di determinazione lineare multipla R^2

E' utile almeno da un punto di vista descrittivo, generalizzare l'indice già visto per quanto riguarda le distribuzioni condizionate di vettori aleatori normali.

La bontà della regressione lineare sulle x per spiegare la variabilità della **y** può essere misurata dall'indice (compreso fra 0 e 1):

$$\mathbf{R}^2_{\mathbf{y}.12\ldots k} = \frac{\text{DEVIANZA SPIEGATA}}{\text{DEVIANZA TOTALE}}$$

Se $k = 1$ è ovvio che $R^2_{\mathbf{y}.1} = r^2$

Si può eventualmente calcolare R^2 mediante la formula vista per le distribuzioni condizionate di vettori aleatori normali.

Evidentemente possiamo anche utilizzare il complemento ad 1 per misurare l'incidenza del residuo sul totale:

$$1 - \mathbf{R}^2_{\mathbf{y}.12\ldots k} = \frac{\text{DEVIANZA RESIDUA}}{\text{DEVIANZA TOTALE}}$$

Il valore di questa quantità fornisce la porzione di variabilità di **y** che non è spiegata dalla regressione sulle k variabili.

Scomposizione della devianza teorica nella regressione multipla

Scomponiamo ora la devianza teorica:

Si riveda eventualmente la parte relativa alla stima dei parametri con questa particolare matrice $\mathbf{X}$

$$\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}/\sigma^2 =$$

$$= R(\mathbf{b})/\sigma^2 + (a-\alpha)n(a-\alpha)/\sigma^2 + (\mathbf{b}-\boldsymbol{\beta})^T\mathbf{Z}^T\mathbf{Z}(\mathbf{b}-\boldsymbol{\beta})/\sigma^2 =$$

$$= R(\mathbf{b})/\sigma^2 + (M_{\mathbf{y}}-\alpha)^2/(\sigma^2/n) + (\mathbf{b}-\boldsymbol{\beta})^T\mathbf{Z}^T\mathbf{Z}(\mathbf{b}-\boldsymbol{\beta})/\sigma^2.$$

Palesemente vale ancora il teorema di Cochran, per la scomposizione in tre parti della devianza complessiva:

il nuovo termine

$\boldsymbol{\alpha})^2/(\sigma^2/n)$ si distribuisce come una χ^2_1,

e per il teorema di Cochran risulta indipendente dalle altre due forme quadratiche.

Si ha, considerando quindi il termine noto:

$$R(\alpha,\boldsymbol{\beta}) = R(a,\mathbf{b}) + (\mathbf{b}-\boldsymbol{\beta})^T(\mathbf{Z}^T\mathbf{Z})(\mathbf{b}-\boldsymbol{\beta}) + n(M_{\mathbf{y}}-\alpha)^2$$

oppure

$$\sum_{i=1}^{n}[\mathbf{y}_i - E(\mathbf{y}_i)]^2 = (\mathbf{y}-\alpha.1-\mathbf{Z}\boldsymbol{\beta})^T(\mathbf{y}-\alpha.1-\mathbf{Z}\boldsymbol{\beta}) =$$

$$(\mathbf{y}-M_{\mathbf{y}}-Zb)^T(\mathbf{y}-M_{\mathbf{y}}-Zb)+(\mathbf{b}-\boldsymbol{\beta})^T(\mathbf{Z}^T\mathbf{Z})(\mathbf{b}-\boldsymbol{\beta})+n(M_{\mathbf{y}}-\alpha)^2+$$

(rispetto al simbolismo adottato precedentemente si consideri che adesso il valore atteso è: $\mathrm{E}[\mathbf{Y}] = \alpha.1 + \mathbf{Z}\boldsymbol{\beta}$)

?

Possiamo rivedere questa relazione in termini di contributi alla devianza teorica di $\boldsymbol{\varepsilon}$:

Forma Quadratica	fonte	gradi di libertà
$(\mathbf{y} - \alpha.1 + \mathbf{Z}\boldsymbol{\beta})^T(\mathbf{y} - \alpha.1 - \mathbf{Z}\boldsymbol{\beta})$	devianza teorica complessiva di ε. (rispetto al modello vero)	n
$(\mathbf{y} - M_{\mathbf{y}} - Zb)^T(\mathbf{y} - M_{\mathbf{y}} - Zb)$	devianza residua	$n - k - 1$
$(\mathbf{b} - \boldsymbol{\beta})^T(\mathbf{Z}^T\mathbf{Z})(\mathbf{b} - \boldsymbol{\beta})$	devianza delle stime dei coefficienti di regressione	k
$n(M_{\mathbf{y}} - \alpha)^2$	devianza dovuta alla stima del termine noto	1

9.1.1 Prova dell'ipotesi di coefficienti di regressione nulli nella regressione multipla.

Dai risultati visti in precedenza e che scaturiscono sostanzialmente dall'ortogonalità fra termine noto e regressori, risulta immediato il test per saggiare l'ipotesi nulla:

$$H_0 : \boldsymbol{\beta} = 0_k,$$

con α qualsiasi contro l'alternativa generica:

$$H_1 : \boldsymbol{\beta} \neq 0_k.$$

TEST NELLA REGRESSIONE LINEARE MULTIPLA

Si può infatti impiegare la quantità test:

$$F = \frac{\mathbf{b}'\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\mathbf{b}}{ks^2}$$

che sotto H_0 si distribuisce secondo una variabile aleatoria F di Snedecor con k ed $n - k - 1$ gradi di libertà.

Avendo indicato al solito con s^2 la stima corretta della varianza, con $n - k - 1$ gradi di libertà, data da:

$$s^2 = (\mathbf{y} - M_{\mathbf{y}} - Zb)^T(\mathbf{y} - M_{\mathbf{y}} - Zb)/(n - k - 1) =$$

$$= \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{y}_i^*)^2/(n - k - 1)$$

E' facile vedere che, dal momento che in fondo il test è dato da:

$$F = \frac{\frac{Devianzaspiegata}{k}}{\frac{Devianzaresidua}{n-k-1}}$$

si può esprimere questo test in funzione di R^2 :

$$F = \frac{\frac{\mathbf{R}^2_{\mathbf{y}.12\ldots k}}{k}}{\frac{1-\mathbf{R}^2_{\mathbf{y}.12\ldots k}}{n-k-1}}$$

Per saggiare ipotesi particolari, rare nelle applicazioni della regressione multipla, del tipo:

$H_0 : \boldsymbol{\beta} = \beta_0$,

con α qualsiasi

si impiegherà ovviamente il test:

$$F = \frac{[\mathbf{b} - \beta_0]^T \mathbf{Z}^{\mathrm{T}} \mathbf{Z} [\mathbf{b} - \beta_0]}{k s^2}$$

LA REGIONE DI RIFIUTO

La regione di rifiuto sarà costituita dai valori elevati di F, superiori ad $F_{\alpha,k,n-k-1}$ (ossia situati sulla coda destra della corrispondente variabile F di Snedecor). Valori osservati di F elevati danno evidenza contraria ad H_0. Infatti sotto H_1 il valore atteso di $\mathbf{b}^{\mathrm{T}} \mathbf{Z}^T Z b$ nel numeratore del test F per saggiare l'ipotesi $\boldsymbol{\beta} = 0_k$, è dato, dalle formule precedenti, da:

$$E(\mathbf{b}^{\mathrm{T}} \mathbf{Z}^T Z b | H_1) = k\sigma^2 + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}^T \mathbf{Z} \boldsymbol{\beta}.$$

mentre

$$E(\mathbf{b}^{\mathrm{T}} \mathbf{Z}^T Z b | H_0) = k\sigma^2$$

Risulta sempre (al solito):

$$E(\mathbf{b}^{\mathrm{T}} \mathbf{Z}^T Z b | H_1) > E(\mathbf{b}^{\mathrm{T}} \mathbf{Z}^T Z b | H_0)$$

è nella forma quadratica $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}^T \mathbf{Z} \boldsymbol{\beta}$, $\mathbf{Z}^T \mathbf{Z}$ è definita positiva; in ogni caso si vede subito che $\boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}^T \mathbf{Z} \boldsymbol{\beta} = (\mathbf{Z}\boldsymbol{\beta})^T \mathbf{Z}\boldsymbol{\beta}$ che è palesemente una somma di quadrati.

Prova di ipotesi particolari nella regressione multipla.

Si può essere interessati ad una particolare ipotesi, quale un vincolo lineare sui coefficienti di regressione, oppure il fatto che, semplicemente:

alcuni dei coefficienti di regressione siano nulli e quindi,

che i corrispondenti regressori $\mathbf{X}_j$ siano ininfluenti ai fini della spiegazione di $\mathbf{y}$.

Si può seguire la metodologia generale vista precedentemente: si badi però che quella tecnica è soddisfacente solo se applicata:

per una ipotesi soltanto oppure

per più ipotesi relative a regressori ortogonali a gruppi.

L' ipotesi:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0 \text{con} q < k;$$

e

$$\beta_{q+1}, \beta_{q+2}, \ldots, \beta_k \text{qualsiasi}$$

(che corrisponde a q vincoli definiti da $\mathbf{C} = \mathbf{I}_q : 0_{k-q}; \mathbf{a} = 0_q$)

stabilisce che q coefficienti di regressione siano nulli

e quindi stabilisce che i corrispondenti q regressori siano eliminabili dal modello generale di spiegazione della variabile di risposta.

Possiamo effettuare il test generale:

$$F = \frac{\frac{[R(\mathbf{b}_0) - R(\mathbf{b})]}{q}}{\frac{R(\mathbf{b})}{n-k-1}} =$$

$$\frac{\frac{[\mathbf{b}-\beta_0]_q^T [(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}{}_q]^{-1} [\mathbf{b}-\beta_0]_q}{q}}{\frac{[\mathbf{y}-\mathbf{X}\mathbf{b}]^{\mathrm{T}}[\mathbf{y}-\mathbf{X}\mathbf{b}]}{n-k-1}}$$

in cui $\mathbf{b}_0$ è lo stimatore di massima verosimiglianza di $\boldsymbol{\beta}$ sotto H_0 (quindi ha q elementi uguali a zero se $H_0 : [\beta_0]_q = 0$).

In effetti si vede facilmente che il test è ora dato da:

$$F = \frac{[\mathbf{b}]_q^T [(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}{}_q]^{-1} [\mathbf{b}]_q}{q s^2}$$

$[\mathbf{b}]_q^T$ indica il vettore di q elementi coinvolto dall'ipotesi nulla .$(\mathbf{Z}^T\mathbf{Z})^{-1}{}_q$ indica il blocco qxq di $(\mathbf{Z}^T\mathbf{Z})^{-1}$

in cui è esplicito il fatto che la quantità a numeratore misura la distanza da zero di un particolare sottoinsieme di stimatori di coefficienti di regressione.

Ovviamente si distribuisce come una F con q e $n-k-1$ gradi di libertà.

Test per un singolo coefficiente (uno solo!)

Nel caso particolare in cui $q = 1$, evidentemente stiamo saggiando l'ipotesi che un singolo coefficiente di regressione sia nullo:

$H_0 : \beta_j = 0$

e gli altri $\boldsymbol{\beta}$ qualsiasi

Il test in questo caso diventa:

$$F = \frac{b_j[(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}{}_{jj}]^{-1}b_j}{s^2} = \frac{b_j^2}{c_{jj}s^2}$$

essendo c_{jj} il j -esimo elemento sulla diagonale di $(\mathbf{Z}^T\mathbf{Z})^{-1}$; essendo $q = 1$ possiamo prendere la radice quadrata di questa quantità, che si distribuisce come una t di Student con $n-k-1$ gradi di libertà, per ottenere il test:

$$t = \frac{b_j}{s(c_{jj}}(t_{n-k-1}$$

Si può eventualmente considerare in questo caso un'alternativa unidirezionale che conduce a regioni di rifiuto sulla coda destra o sulla sinistra. Si noti anche che c_{jj} è la varianza campionaria di b_j

---

Con questo test possiamo saggiare una ipotesi su un coefficiente (uno e uno solo!!!);
Utilizzare questo test per più di un regressore è una procedura distorta.

Test per l'eliminazione di q regressori in termini di perdita in R^2

Riscriviamo il test per saggiare l'ipotesi che q regressori siano nulli:

$$F = \frac{\frac{[R(\mathbf{b}_0) - R(\mathbf{b})]}{q}}{\frac{R(\mathbf{b})}{n-k-1}}$$

$$F = \frac{\frac{\text{Devianza spiegata da } k \text{ regressori-Devianza spiegata da } k-q \text{ regressori}}{q}}{\frac{\text{Devianza residua [nel modello completo]}}{n-k-1}}$$

Dividendo ora ambo i termini della frazione per $Dev(\mathbf{y})$ si può esprimere questo test in funzione di due diversi indici R^2 :

$$F = \frac{\frac{\mathbf{R}^2_{\mathbf{y}.12\ldots k} - \mathbf{R}^2_{\mathbf{y}.q+1\ldots k}}{q}}{\frac{1-\mathbf{R}^2_{\mathbf{y}.12\ldots k}}{n-k-1}}$$

in cui:

- $\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}$ è la frazione di varianza di $\mathbf{y}$ spiegata dai $k-q$ regressori $\mathbf{X}_{q+1}, \mathbf{X}_{q+2}, \ldots, \mathbf{X}_k$;
- $\mathbf{R}^2_{\mathbf{y}.12\ldots k}$ è la frazione di varianza di $\mathbf{y}$ spiegata da tutti i regressori;

Quindi il test corrisponde a saggiare l'ipotesi che il decremento in $\mathbf{R}^2_{\mathbf{y}.12\ldots k}$ dovuto all'eliminazione dei q regressori $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_q$ non si discosti significativamente da 0.

Evidentemente il numeratore del test F è sempre positivo (si tratta sempre, come si era visto prima) di una frazione di varianza.

Il test è relativo ad una ipotesi relativa ad un insieme fissato di q regressori.

Successivamente si utilizzeranno queste scomposizioni per arrivare ad un criterio di scelta di k-q particolari regressori

Possiamo impostare una tavola di analisi della varianza per la riduzione di variabili:

TOTALE=	RESIDUA	SPIEGATA da $k-q$ regressori	SPIEGATA gressori (al altri $k-q$)
frazioni di varianza			
1	$1-\mathbf{R}^2_{\mathbf{y}.12\ldots k}$	$\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}$	$\mathbf{R}^2_{\mathbf{y}.12\ldots k}-\mathbf{R}$
gradi di libertà:			
$n-1$	$n-k-1$	$k-q$	q

Rappresentazione grafica della suddivisione delle frazioni di devianza.

$0\mathbf{R}^2_{\mathbf{y}.12\ldots k}1$	devianza spiegata da tutti i k regressori
$0\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}1$	devianza spiegata dagli ultimi $k-q$ regressori
$0\mathbf{R}^2_{\mathbf{y}.12\ldots k}-\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}1$	devianza in più spiegata dai primi q regressori
$01-\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}1$	devianza non spiegata dagli ultimi $k-q$ regressori

Un indice normalizzato è dato da:

$$\frac{\mathbf{R}^2_{\mathbf{y}.12\ldots k}-\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}}{1-\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}}$$

coefficiente di determinazione parziale di $\mathbf{Y}$ sui primi q regressori, al netto degli altri k-q regressori

L'indice che è ancora palesemente compreso fra 0 e 1;

misura la frazione ulteriore di varianza spiegata dai q regressori, tenuto conto della regressione sugli altri k-q.

---

incremento di R^2 in funzione dell'indice di correlazione parziale; trovare (forse sul kendall o Rao)

9.2 La multicollinearità nella regressione multipla.

In questa sezione affrontiamo un problema cruciale nell'analisi della regressione, in particolare per dati economici o comunque provenienti da indagini osservazionali che si può riassumere nella domanda:
Avere regressori linearmente correlati ha qualche influenza negativa sull'analisi della regressione?
Banalmente si potrebbe pensare che l'unica cosa importante è la correlazione (multipla) della Y con le X. Vedremo in questa sezione che è anche importantissimo analizzare la struttura di correlazione interna delle X

Caso di due soli regressori Supponiamo una caso molto semplice con due soli regressori. Consideriamo per semplificare le cose, e focalizzare l'attenzione solo sulle correlazioni, che le variabili (y e X) siano tutte standardizzate.

Sappiamo che $\mathrm{V}[\mathbf{b}] = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2$. Se quindi $k = 2$ si ha:

$$\mathrm{V}[\mathbf{b}] = \sigma^2 \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}^{-1} = \sigma^2 \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

per cui la varianza di uno dei due stimatori è data da:

$$\mathrm{V}[\mathbf{b}_1] = \mathrm{V}[\mathbf{b}_2] = \sigma^2 \frac{1}{1 - r_{12}^2}$$

Collinearità nella regressione a due regressori

La varianza degli stimatori dei coefficienti di regressione è funzione crescente della correlazione fra i regressori r_{12}
ed è funzione crescente della varianza σ^2 della componente accidentale

Studiare anche come varia la dipendenza di y dalle x passando da due a un regressore

Passando alla situazione generale, se i k regressori non sono ortogonali, possono avere una struttura di interdipendenza di vario tipo.

Si sono già viste alcune delle conseguenze della non ortogonalità dei regressori o fattori sulla distribuzione degli stimatori di massima verosimiglianza e di altre quantità collegate:

- Lo stimatore $\mathbf{b}$ è a componenti correlate (dal momento che ha varianza proporzionale a $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$);
- I contributi alla spiegazione di $\mathbf{Y}$ di ciascuna variabile non sono separabili.
- Non si possono condurre test indipendenti su tutti i singoli coefficienti.
- Le regioni di confidenza dei parametri $\boldsymbol{\beta}$ costruite sulla base del valore critico di F risultano ellissoidali e non sferiche.
- Il luogo dei punti $\mathbf{x}_i$ nello spazio dei regressori che conduce ad intervalli di confidenza di eguale ampiezza per $E(\hat{\psi}_i)$ è il contorno di un ellissoide di equazione:

$$\sigma^2 \mathbf{x}_{(i)^T} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{x}_{(i)} = Cost.$$

Il caso estremo è quello in cui il rango di $\mathbf{X}$ (e quindi di $\mathbf{X}^{\mathrm{T}}\mathbf{X}$) è inferiore a k: supponiamo di non trovarci comunque in questa situazione, perché l'eventuale variabile combinazione lineare esatta delle altre è stata individuata ed eliminata.

Supporremo di trovarci invece, nell'ambito delle situazioni con dei regressori correlati, vicino a questa situazione estrema.

Nei casi non estremi occorrerà misurare il grado di collinearità fra le variabili indipendenti ossia quanto complessivamente incidono le correlazioni fra le $\mathbf{X}_j$ sulla distribuzione di $\mathbf{b}$ ed in generale sull'inferenza nella regressione multipla

Consideriamo una matrice delle x a media nulla (quindi è una matrice di scarti) ed a varianza unitaria (quindi è una matrice di variabili standardizzate); evidentemente ciò corrisponde ad effettuare una traslazione ed un cambiamento di scala sugli assi che non

alterano in alcun modo lo studio della dipendenza lineare di $\mathbf{y}$ dalle $\mathbf{X}_j$.

(Anzi in questo modo si possono fare valutazioni comparative fra i coefficienti di regressione, in quanto non influenzati dalle diverse unità di misura).

Lo studio della multicollinearità riguarda la struttura di correlazione fra le $\mathbf{X}$ e successivamente l'influenza di questa struttura sullo studio della dipendenza di $\mathbf{Y}$ dalle $\mathbf{X}$, sulle proprietà degli stimatori, delle regioni di confidenza, etc.

In questa lezione sulla multicollinearità, sto esaminando solo le implicazioni di tipo statistico: lascio volutamente da parte le implicazioni di tipo computazionale. È noto, infatti, che dal punto di vista numerico la risoluzione di sistemi di equazioni lineari, in presenza di collinearità, comporta dei problemi di stabilità numerica delle soluzioni.
Con determinante della matrice dei coefficienti prossimo a zero gli errori di troncamento potrebbero svolgere un ruolo determinante sul calcolo delle soluzioni del sistema di equazioni normali.

Se le x sono standardizzate la matrice di varianze e covarianze $\mathbf{S}$ è anche la matrice di correlazione, ed è data da:

$$\mathbf{S} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n.$$

Quindi è lo stesso studiare la struttura di $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ o quella di $\mathbf{S}$.

Dal momento che le x sono a media nulla e a varianza unitaria, si avrà che combinazioni lineari delle x sono a media nulla, e inoltre:

dal momento che la somma degli autovalori di $\mathbf{S}$ è uguale alla sua traccia (ossia alla somma delle varianze), è quindi uguale a k se si lavora con variabili standardizzate

Occorre che le x siano standardizzate per poter valutare la grandezza di ciascun autovalore.

Infatti:

$$\lambda_i > 0 i = 1, 2, \ldots k;$$

($\mathbf{S}$ è definita positiva e di rango pieno)

Inoltre:

$$\sum_{i=1}^{k} \lambda_i = k$$

Per cui gli autovalori sono limitati fra 0 e k:

$$k > \lambda_i > 0 i = 1, 2, \ldots k;$$

e

$$M(\lambda_i) = \sum_{i=1}^{k} \lambda_i / k = 1$$

Nella situazione ideale di assenza di correlazioni fra le x si ha:

$$\lambda_1 = \lambda_2 = \ldots = \lambda_k = 1$$

perché $\mathbf{S} = \mathbf{I}$
La situazione è ideale perché le stime dei regressori risultano non correlate e le inferenze sui regressori sono indipendenti.

Si parla di multicollinearità quando, pur essendo la matrice $\mathbf{S}$ a rango pieno, alcuni dei suoi autovalori sono molto vicini a zero, avvicinandosi alla situazione estrema di collinearità esatta.

Questo si verifica quando qualcuna delle variabili x è quasi uguale

ad una combinazione lineare di alcune delle altre variabili $\mathbf{X}$.

---

la situazione limite $\lambda_k = 0$ corrisponde al caso di rango inferiore a k, ossia una variabile è esattamente combinazione lineare delle altre (oppure q variabili sono combinazioni lineari delle altre se $\lambda_{k-q+1} = \lambda_{k-q+2} = \ldots = \lambda_k = 0$)

Nella regressione multipla ci interessa che la $\mathbf{Y}$ sia molto correlata con le $\mathbf{X}$, ma è preferibile che le $\mathbf{X}$ siano poco correlate internamente

---

Si riveda per analogia la parte relativa all'analisi delle componenti principali per vettori aleatori. Si riveda anche l'interpretazione dell'analisi in componenti principali per variabili statistiche osservate.

Si riveda anche lo schema riportato in un capitolo precedente sull'influenza delle possibili configurazioni di matrice $\mathbf{x}$ sull'inferenza nei modelli lineari.

Legami lineari fra regressori

Adesso esamineremo con dettaglio l'influenza delle correlazioni fra i regressori nel caso generale: esistono infatti delle situazioni nelle quali la presenza di correlazioni potrebbe essere importante anche se non si è in una situazione di multicollinearità vera e propria; si vedrà più avanti a proposito la relazione che lega la varianza delle previsioni con la varianza degli stimatori.

Dall'equazione che definisce gli autovettori e gli autovalori della matrice delle varianze e covarianze $\mathbf{S}$ (gli autovalori son proporzionali a quelli della matrice delle devianze e codevianze $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, essendo $\mathbf{X}$ una matrice di variabili scartate dalle rispettive medie e possibilmente standardizzate) si ha:

$$\mathbf{S}\boldsymbol{\gamma}_j = (\mathbf{X}^{\mathrm{T}}\mathbf{X}/n)\boldsymbol{\gamma}_j = \lambda_j\boldsymbol{\gamma}_j \approx 0 se \lambda_j \approx 0$$

(dato che tutti gli elementi di $\boldsymbol{\gamma}_j$, i-esimo autovettore sono compresi fra 0 e 1, per la condizione di normalizzazione $\boldsymbol{\gamma}_j^T\boldsymbol{\gamma}_j = 1$)

Allora premoltiplicando per $\boldsymbol{\gamma}_j^T$ si ha:

$$(\boldsymbol{\gamma}_j^T\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\gamma}_j)/n = \boldsymbol{\gamma}_j^T\lambda_j\boldsymbol{\gamma}_j = \lambda_j \approx 0$$

Poniamo:

$$u_j = \mathbf{X}\boldsymbol{\gamma}_j/\sqrt{n}$$

così che u_j è una combinazione lineare nelle $\mathbf{X}$, e quindi:

$$(\boldsymbol{\gamma}_j^T\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\gamma}_j)/n = u_j^T u_j = \lambda_j \approx 0 (\text{per l'ipotesi fatta})$$

Allora se λ_j è piccolo si ha: il vettore u_j è una combinazione lineare delle $\mathbf{X}$, con media zero e varianza molto piccola, per cui si ha anche:

$u_j \approx 0$ ossia $\Rightarrow \mathbf{X}\boldsymbol{\gamma}_j \approx 0$

Quindi esiste una combinazione lineare delle variabili quasi nulla Le variabili maggiormente coinvolte corrispondono ai più alti coefficienti di $\boldsymbol{\gamma}_j$

ossia le variabili $\mathbf{X}_r$ corrispondenti ai più alti elementi $\boldsymbol{\gamma}_{rj}$; avendo inteso le colonne della matrice $\boldsymbol{\Gamma}$ di elemento $\boldsymbol{\gamma}_{rj}$ costituite dagli autovettori di $\mathbf{S}$

Si può giungere a questo tipo di risultato (ossia esistenza di combinazioni lineari quasi esatte fra i regressori), anche considerando che in questo caso una o più variabili risulta combinazione lineare quasi esatta delle altre, ossia avrà una dipendenza lineare elevata dalle altre variabili, in termini di regressione multipla .

In effetti, ricordando le relazioni fra R^2 e gli elementi dell'inversa di $\mathbf{S}$ (si rivedano nella parte relativa alle distribuzioni condizionate di v.a. normali), si può arrivare a:

$$R_i^2 = 1/c_{ii}$$

R_i^2 è il coefficiente di determinazione multipla di $\mathbf{X}_i$ rispetto alle altre $k-1$ variabili, ossia quanta variabilità di $\mathbf{X}_j$ è spiegata dalle altre $k-1$ variabili $\mathbf{X}_j (j \neq i)$

c_{ii} è l'elemento diagonale di $\mathbf{C}$, l'inversa di $\mathbf{S}$

Ricordando anche che:

$$\lambda_j(\mathbf{C}) = \lambda_j(S^{-1}) = 1/\lambda_j(S);$$

si ha:

$$R_i^2 = 1 - 1/c_{ii}; 1/(1-R_i^2) = c_{ii}$$

quindi sommando queste ultime relazioni per tutte le variabili si ha:

$$\sum_{i=1}^{k} 1/(1-R_i^2) = \sum_{i=1}^{k} c_{ii} = tr[\mathbf{C}] = \sum_{i=1}^{k} 1/\lambda_i$$

Quindi se qualche autovalore è molto piccolo, la traccia di $\mathbf{C}$ è molto grande e questo è direttamente collegato al fatto che qualche correlazione multipla delle x è elevata.

CITARE OUTPUT DI STATISTICA
(ridondanza, etc.)

Conseguenze sulla distribuzione campionaria di b

$$V(\mathbf{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \sigma^2(n\mathbf{S})^{-1} = \mathbf{S}^{-1}(\sigma^2/n) = \mathbf{C}(\sigma^2/n)$$

Quindi a parte il fattore (σ^2/n) la struttura delle correlazioni interne fra gli elementi di $\mathbf{b}$ è funzione della struttura delle correlazioni interne fra le $\mathbf{X}$, e non dipende in alcun modo dalla variabile di risposta $\mathbf{y}$: dipende solo dallo schema di valori assunti dai regressori (siano essi osservati o prestabiliti prima di un esperimento).

Si noti inoltre che invece le varianze dipendono al solito dai valori osservati, attraverso il fattore (σ^2/n)

$$\sum_{i=1}^{k} V(b_i) = tr(V(\mathbf{b})) = tr(\mathbf{S}^{-1})(\sigma^2/n) = (\sigma^2/n)tr[\mathbf{C}] == (\sigma^2/n)\sum_{i=1}^{k} 1/\lambda_i$$

Quindi se vi è multicollinearità (ossia qualche λ_i molto piccolo) la traccia di $\mathbf{C}$ sarà elevata e quindi sarà elevata la somma delle varianze campionarie degli stimatori dei coefficienti di regressione.
Sarà conseguentemente elevata anche la varianza di $\mathbf{y}_i^*$

Indici di multicollinearitá:

$$I_p = \frac{\sum_{i=1}^{p} \lambda_j}{\sum_{i=1}^{k} \lambda_j} = \frac{\text{varianza delle prime } p \text{ componenti}}{\text{somma di tutte le varianze}}$$

$$I_p = \frac{\sum_{i=1}^{p} \lambda_j}{k}$$

nel caso di variabili standardizzate.

Più che regole automatiche, l'analisi grafica dell'andamento di $I_p$al variare di p può guidare nell'analisi della multicollinearità in insiemi di dati reali.

ESEMPI VARI
Collinearità: confronto fra k e k-1 regressori attraverso i λ

Costruzione di un stimatore distorto di β

Per esaminare meglio gli effetti della multicollinearità sulla varianza campionaria dello stimatore $\mathbf{b}$, si può sfruttare la decomposizione spettrale o canonica della matrice $\mathbf{S}^{-1}$, introdotta a proposito delle proprietà degli autovalori e degli autovettori di matrici simmetriche:

$$\mathbf{S}^{-1} = \mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma}^T = \sum_{i=1}^{k} \boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^T/\lambda_i$$

mentre per la matrice originaria S abbiamo la decomposizione di base:

$$\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T = \sum_{i=1}^{k} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T$$

Se invece di prendere tutti i k termini di questa decomposizione, ci limitiamo a prendere i primi q termini, otteniamo un'approssimazione della matrice $\mathbf{S}$ tanto migliore, quanto più sono piccoli gli autovalori corrispondenti aai termini scartati:

$$\mathbf{S} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T = \sum_{i=1}^{k} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T \approx \sum_{i=1}^{q} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T = S_{(q)}$$

in corrispondenza di questa approssimazione costruiamo una inversa modificata:

$$\mathbf{S}^{-1} = \sum_{i=1}^{k} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T / \lambda_i \rightarrow \sum_{i=1}^{q} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T / \lambda_i = \mathbf{S}_{(q)}^{-1},$$

in cui stavolta mancano i termini più elevati in valore assoluto.

(evidentemente le stesse scomposizioni, a meno del fattore n, si possono fare sulla matrice $\mathbf{X}^{\mathrm{T}}\mathbf{X}$)

Pertanto, se invece di $\mathbf{b}$ si definisse:

$$b^0 = \mathbf{S}_{(q)}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}/n$$

si otterrebbe uno stimatore distorto ma con minore varianza!

Infatti:

---

controllare bene il seguito

$$E(b^0) = \mathbf{S}_{(q)}^{-1}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{\beta}) =$$

$$\mathbf{S}_{(q)}^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\beta} = (\mathbf{S}_{(q)}^{-1}/n)(n\mathbf{S}_{(q)} + \mathbf{R}(q))\boldsymbol{\beta} =$$

$$\sum_{i=1}^{q} \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T / \lambda_i (\lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T) + \sum_{i=q+1}^{k} \lambda_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T$$

A parte l'eventuale impiego effettivo di questo stimatore, l'utilità della sua introduzione sta nell'esplicitazione del legame fra distorsione e varianza campionaria degli stimatori di $\boldsymbol{\beta}$.

9.2.1 Esempi (sulla collinearità e simili)

Figura da inserire ESECOLL2.RTF esecollinear2.STA esecoll2.stg dove sono???

9.3 La scelta delle variabili nella regressione lineare multipla.

Motivazioni

Si è detto prima di sottoinsiemi di variabili predittive stabiliti a priori e quindi senza riferimento ai particolari dati osservati. Spesso però, date k variabili esplicative, si vuole scegliere un sottoinsieme di q di tali variabili con diverse finalità: per effettuare stime o previsioni statistiche a costo inferiore, riducendo il numero di variabili che occorrerà rilevare in futuri studi.

Per migliorare l'accuratezza delle previsioni eliminando variabili poco informative o comunque poco rilevanti ai fini della previsione di $\mathrm{E}[\mathbf{y}]$ per descrivere un data-set multivariato, o comunque una relazione multipla in modo parsimonioso e con pochi parametri. per stimare coefficienti di regressione con errori standard piccoli, in particolare se alcuni dei regressori sono molto correlati.

Stime carenti dei coefficienti possono portare buone stime predittive (ossia al solo scopo di stimare valori di $\mathbf{y}$ o di $\mathrm{E}[\mathbf{y}]$).

Strategie di scelta

La strategia complessiva della scelta di variabili si può articolare in alcune fasi generali:

- decidere quali sono le variabili che costituiscono l'insieme più ampio dei k regressori (e quindi procedere alla rilevazione)
- trovare uno o più sottoinsiemi di variabili che spiegano bene la variabile di risposta;
- applicare una regola di arresto per decidere quante variabili esplicative (regressori) usare;
- stimare i coefficienti di regressione
- saggiare la bontà del modello ottenuto (analisi dei residui, aggiunta di nuove variabili, aggiunta di termini polinomiali, etc.).

Per quanto riguarda il punto b), possiamo esplicitarlo in questo modo:

fissato un numero di regressori ridotto, diciamo p, quale dei ${}^kC_q$ sottoinsiemi dei k regressori originari scegliere?

Sembra logico, e comunque più semplice, almeno in prima istanza, scegliere quello che fornisce la maggior quota di varianza spiegata, ossia il maggiore fra gli R^2;

In aggiunta a questo criterio di massimizzazione globale, avendo fissato q, si può comunque pensare a scopo esplorativo di prendere in esame alcuni sottoinsiemi che forniscono le soluzioni migliori.

Occorrerà possibilmente un qualche algoritmo per ridurre il numero di R^2 da calcolare.

Fonti di distorsioni

Le distorsioni nella stima dei coefficienti sono dovute a due diverse fonti:

una distorsione dovuta all'avere omesso variabili, di ui è possibile fornire una valutazione (in termini di deviazione dal modello completo) una distorsione dovuta al procedimento di selezione, che non viene in generale fatto indipendentemente dai dati; in altri termini i dati mediante i quali si stimano i coefficienti sono gli stessi che hanno portato alla selezione di un particolare sottoinsieme.

quest'ultima distorsione, dovuta alla selezione, può essere distinta in due ulteriori componenti: una dovuta alla scelta fra sottoinsiemi delle stesse dimensioni l'altra dovuta alla regola di arresto impiegata per scegliere il numero q migliore di regressori. Queste ultime fonti di distorsione in generale non sono valutabili con precisione.

Criteri di scelta

Che criterio usare per scegliere il numero p più opportuno di variabili da includere nel modello?

Si tenga presente che se A_p è l'insieme ottimo di p variabili e A_{p+1} è l'insieme ottimo con $p+1$ variabili, si ha sempre:

$$R^2_{\mathbf{y}}(A_p)(R^2_{\mathbf{y}}(A_{p+1})$$

(l'uguaglianza in effetti vale solo in caso di collinearità esatta, che a rigore abbiamo escluso se $\mathbf{S}$ è di rango pieno).

Inoltre se I_{q+1} è un insieme con p+1 variabili e se I_p^T è un suo sottoinsieme, ossia un insieme di p variabili ottenuto da I_{q+1} eliminando

una variabile, si ha ancora:

$$R^2_{\mathbf{y}}(I_p^T)(R^2_{\mathbf{y}}(I_{p+1}).$$

Eventuali test F condotti sugli R^2 saranno comunque distorti, almeno in termini di livelli di significatività.

Infatti la devianza che si mette a numeratore non è calcolata su un set dato a priori, ma in base al fatto che il residuo sia il più basso possibile.

Algoritmi di scelta delle variabili.

Si possono comunque avere diversi algoritmi di scelta di variabili, a prescindere dal problema della scelta di q.

Tutte le regressioni possibili

Selezione in avanti (forward selection)

Selezione all'indietro, o eliminazione (backward selection);

Regressione passo (stepwise regression)

(algoritmi di sostituzione).

Il metodo di tutte le regressioni possibili prevede l'esame di tutti i $2^k - 1$ possibili sottoinsiemi di variabili;

$$(2^k - 1 = \sum_{p=1}^{k} {}_kC_p)$$

Computazionalmente oneroso, sebbene esistano ora degli algoritmi di ricerca che consentono di limitare il numero dei confronti, pur trovando l'ottimo assoluto per ciascun numero di regressori q.

Un problema interpretativo si ha quando si ottengono soluzioni non nidificate: alcuni software (S-Plus, per esempio) possono fornire oltre l'ottimo assoluto per ciascun valore di p, anche un certo numero di soluzioni sub-ottimali, ossia gli r migliori sottoinsiemi.

Metodi che conducono ad ottimi locali

Il metodo della selezione in avanti prevede di partire da un modello senza regressori, e di introdurli uno alla volta secondo che producano il valore più elevato fra i test F.

Evidentemente si trovano soluzioni sub-ottimali, e si rischia di non prendere mai in esame simultaneamente determinati sottoinsiemi di regressori.

Il metodo della selezione all'indietro, consiste nel partire dal modello completo, e ad ogni passo si elimina la variabile cui corrisponde il valore di F più basso.

Anche questo fornisce soluzioni sub-ottimali; tuttavia è molto usato e abbastanza ben interpretabile, in quanto prende comunque in esame una volta tutte le variabili simultaneamente;

inoltre fornisce una graduatoria delle variabili in ordine decrescente di importanza secondo l'ordine di eliminazione;

Il metodo stepwise unisce le due tecniche prima menzionate:

si parte da un modello senza regressori e si segue la tecnica della selezione in avanti; ad ogni passo con una nuova variabile introdotta, si riesamina l'insieme delle variabili introdotte, per vedere se si può eliminarne qualcuna (con procedura backward); successivamente i continua con la selezione in avanti fino a che non si effettuano più modifiche dell'insieme di repressori:

test di ingresso: $F > F_{in}$

test di uscita: $F < F_{out}$

Questa tecnica, che risale al 1960, essenzialmente rispondeva all'esigenza pratica di non prendere in esame simultaneamente grossi insiemi di regressori; inoltre nella versione originaria considerava la possibilità di valutare le varie inverse e determinanti di ogni passo a partire da quelli trovati al passo precedente.

Distorsione degli stimatori con modelli ridotti

Come si è visto:

$$E(\mathbf{y}_i*) = E(\mathbf{x}_{(i)^T}\mathbf{b}) = \mathbf{x}_{(i)^T}\boldsymbol{\beta} = E(\mathbf{y}_i)$$

$$V(\mathbf{y}_i*) = V(\mathbf{x}_{(i)^T}\mathbf{b}) = \mathbf{x}_{(i)^T}V(\mathbf{b})\mathbf{x}_{(i)} = \sigma^2\mathbf{x}_{(i)^T}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}_{(i)}$$

Ovviamente questa relazione presuppone la correttezza del modello; se adesso prendiamo in considerazione la possibilità di lavorare con modelli distorti, vediamo cosa succede all'errore quadratico medio della singola previsione:

$$E.q.m(\mathbf{y}_i*) = E(\mathbf{x}_{(i)^T}(\mathbf{b}-\boldsymbol{\beta}))^2 = E\mathbf{x}_{(i)^T}[(\mathbf{b} - E(\mathbf{b})) + (E(\mathbf{b}) - \boldsymbol{\beta})]^2 =$$

$$Ex_{(i)^T}V(\mathbf{b})\mathbf{x}_{(i)} + \mathbf{x}_{(i)^T}\mathbf{x}_{(i)}(E(\mathbf{b}) - \boldsymbol{\beta})^2.$$

Vediamo ora cosa accade per la media di tutti gli e.q.m. di previsione, almeno per i valori effettivamente osservati:

$$\sum_{i=1}^{n} ??? x_i$$

errore quadratico medio degli stimatori

Figura da inserire LUCIDI SCRITTI A MANO
cenni al Cp di Mallows

9.3.1 Esempio di correlazioni osservate fra molte variabili

Quando si rilevano molte variabili su n soggetti, in particolare in studi osservazionali, è possibile rilevare nella fase esplorativa delle correlazioni, sia semplici che multiple, anche molto consistenti, semplicemente per effetto di fluttuazioni campionarie dovute al cercare correlazioni empiriche alte in una matrice di correlazione con molti elementi.

Infatti si supponga per semplicità che la matrice $n \times p$ delle osservazioni costituisca un campione (multivariato) di ampiezza n proveniente da una distribuzione normale multivariata a p componenti indipendenti, e quindi con correlazioni lineari teoriche $\rho_{ij} = 0$; semplicemente per il fatto che nella matrice di correlazione stimata $p \times p$ si avranno $p(p-1)/2$ indici r_{ij} empirici di correlazione lineare, stime di massima verosimiglianza delle corrispondenti correlazioni lineari ρ_{ij} della popolazione multinormale di provenienza (sebbene tali $p(p-1)/2$ non siano indipendenti perché calcolate su p variabili):

Il più grande di tali indici chiaramente ha una distribuzione campionaria che non ha come valore atteso il valore teorico $\rho_{ij} = 0$.

Per un r_{ij} qualsiasi vale l'usuale trasformazione:

$$r_{ij} \sqrt{\frac{n-2}{1-r_{ij}^2}}$$

che si distribuisce come una t di student, con $n-2$ gradi di libertà, quando $\rho_{ij} = 0$, tuttavia in questo caso stiamo *scegliendo* dalla matrice di correlazione *l'elemento* (o gli elementi) *più grande*, per cui non valgono i normali risultati sulla distribuzione di r_{ij}.

Esempio:

Da una distribuzione normale multivariata con 30 componenti indipendenti e standardizzate è stato estratto un campione di 100 osservazioni (la matrice dei dati è stata costruita per simulazione, ossia mediante generazione di numeri pseudo-casuali). Dal campione di osservazioni, con $n = 100$ e $p = 30$ è stata calcolata la matrice delle stime delle correlazioni lineari:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
X1	1.00	-.07	-.03	-.13	.08	-.10	-.04	.06	.15	-.08	-.00	-.02	-.03	.00	-.03
X2	-.07	1.00	-.00	.03	.07	.07	.02	.04	-.00	.01	-.25	.05	-.14	.01	-.02
X3	-.03	-.00	1.00	.14	-.09	-.02	-.01	.08	.16	-.04	-.08	-.02	-.08	.04	-.05
X4	-.13	.03	.14	1.00	-.13	.07	.18	-.14	.18	-.12	-.05	-.03	-.09	.10	.15
X5	.08	.07	-.09	-.13	1.00	.02	.02	-.10	.06	.03	-.08	.01	-.02	-.17	.12
X6	-.10	.07	-.02	.07	.02	1.00	-.01	-.17	-.01	.00	.03	.13	.04	.02	.03
X7	-.04	.02	-.01	.18	.02	-.01	1.00	-.05	-.05	.09	-.04	-.04	.02	-.09	.02
X8	.06	.04	.08	-.14	-.10	-.17	-.05	1.00	.05	-.02	-.04	-.01	-.18	-.03	.03
X9	.15	-.00	.16	.18	.06	-.01	-.05	.05	1.00	.08	.09	-.20	-.05	-.01	.00
X10	-.08	.01	-.04	-.12	.03	.00	.09	-.02	.08	1.00	.16	-.18	-.01	.04	.14
X11	-.00	-.25	-.08	-.05	-.08	.03	-.04	-.04	.09	.16	1.00	-.23	.04	-.08	-.20
X12	-.02	.05	-.02	-.03	.01	.13	-.04	-.01	-.20	-.18	-.23	1.00	.25	.09	.05
X13	-.03	-.14	-.08	-.09	-.02	.04	.02	-.18	-.05	-.01	.04	.25	1.00	-.10	-.17
X14	.00	.01	.04	.10	-.17	.02	-.09	-.03	-.01	.04	-.08	.09	-.10	1.00	-.02
X15	-.03	-.02	-.05	.15	.12	.03	.02	.03	.00	.14	-.20	.05	-.17	-.02	1.00
X16	.08	.00	-.19	-.13	.05	.04	-.09	.03	-.18	-.03	-.03	.09	.25	-.01	.12
X17	.20	-.04	.05	.06	-.08	-.13	.14	-.05	.01	-.01	-.03	.00	.02	.02	.00
X18	-.02	-.06	.05	.03	-.03	-.13	-.09	.26	.10	-.07	.10	.08	-.00	.10	.01
X19	.22	-.04	-.08	.02	.01	.19	-.05	.02	-.09	-.13	.04	.21	.23	.00	.04
X20	.19	-.04	.08	-.13	.01	-.06	-.03	.23	.01	-.07	-.10	-.11	-.09	.15	.03
X21	.03	.06	-.11	-.09	.10	.06	.12	-.23	-.27	-.08	.04	.20	.09	-.02	-.17
X22	-.18	-.03	.14	.01	.12	-.05	.02	.12	-.13	.02	-.13	-.06	-.15	.17	.07
X23	.04	-.13	.04	.05	.04	-.18	.14	.10	.05	.08	.17	.19	-.11	-.11	.03
X24	-.05	.10	-.06	-.03	-.05	-.11	.13	.00	-.13	.06	-.01	.07	-.11	.01	.13
X25	.02	-.01	-.08	-.05	-.00	-.14	.08	-.09	-.08	-.14	-.11	.15	.06	.17	.01
X26	-.08	.16	-.12	-.01	.12	.13	-.10	-.16	-.06	.00	.13	-.07	.01	-.07	.02
X27	.00	.07	-.08	-.09	-.12	-.10	-.01	.05	.01	-.01	.00	.08	.11	.09	-.03
X28	-.02	-.03	.03	-.03	.12	-.22	.03	-.05	-.09	-.00	-.20	-.07	-.03	.02	.02
X29	.09	-.02	-.07	-.04	-.13	.06	-.03	-.06	-.14	-.17	.00	-.01	-.07	-.08	.04
X30	.10	-.01	.13	-.17	.08	-.14	-.05	-.06	.16	.03	-.11	.00	.03	.16	-.02

	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
X1	.08	.20	-.02	.22	.19	.03	-.18	.04	-.05	.02	-.08	.00	-.02	.09	.10
X2	.00	-.04	-.06	-.04	-.04	.06	-.03	-.13	.10	-.01	.16	.07	-.03	-.02	-.01
X3	-.19	.05	.05	-.08	.08	-.11	.14	.04	-.06	-.08	-.12	-.08	.03	-.07	.13
X4	-.13	.06	.03	.02	-.13	-.09	.01	.05	-.03	-.05	-.01	-.09	-.03	-.04	-.17
X5	.05	-.08	-.03	.01	.01	.10	.12	.04	-.05	-.00	.12	-.12	.12	-.13	.08
X6	.04	-.13	-.13	.19	-.06	.06	-.05	-.18	-.11	-.14	.13	-.10	-.22	.06	-.14
X7	-.09	.14	-.09	-.05	-.03	.12	.02	.14	.13	.08	-.10	-.01	.03	-.03	-.05
X8	.03	-.05	.26	.02	.23	-.23	.12	.10	.00	-.09	-.16	.05	-.05	-.06	-.06
X9	-.18	.01	.10	-.09	.01	-.27	-.13	.05	-.13	-.08	-.06	.01	-.09	-.14	.16
X10	-.03	-.01	-.07	-.13	-.07	-.08	.02	.08	.06	-.14	.00	-.01	-.00	-.17	.03
X11	-.03	-.03	.10	.04	-.10	.04	-.13	.17	-.01	-.11	.13	.00	-.20	.00	-.11
X12	.09	.00	.08	.21	-.11	.20	-.06	.19	.07	.15	-.07	.08	-.07	-.01	.00
X13	.25	.02	-.00	.23	-.09	.09	-.15	-.11	-.11	.06	.01	.11	-.03	-.07	.03
X14	-.01	.02	.10	.00	.15	-.02	.17	-.11	.01	.17	-.07	.09	.02	-.08	.16
X15	.12	.00	.01	.04	.03	-.17	.07	.03	.13	.01	.02	-.03	.02	.04	-.02
X16	1.00	-.05	.02	.26	-.02	.20	-.12	-.01	.11	-.02	-.14	.06	-.12	.14	.08
X17	-.05	1.00	.01	.10	.02	*.20*	-.20	-.08	.10	.16	-.15	-.05	-.02	-.11	.11
X18	.02	.01	1.00	-.01	-.11	-.16	.02	.01	.01	-.06	-.10	.14	.08	-.18	-.00
X19	.26	.10	-.01	1.00	-.03	.05	-.13	-.06	.10	.13	-.26	-.11	-.02	.00	-.05
X20	-.02	.02	-.11	-.03	1.00	-.13	.07	.02	.03	-.10	.05	-.10	.10	.12	.07
X21	.20	.20	-.16	.05	-.13	1.00	.14	.01	-.00	.23	.11	.11	-.06	-.08	-.06
X22	-.12	-.20	.02	-.13	.07	.14	1.00	.04	-.01	.12	.11	-.06	.22	-.18	.02
X23	-.01	-.08	.01	-.06	.02	.01	.04	1.00	.20	.05	-.20	-.16	.19	-.06	-.08
X24	.11	.10	.01	.10	.03	-.00	-.01	.20	1.00	.08	-.12	.12	.04	-.15	.02
X25	-.02	.16	-.06	.13	-.10	.23	.12	.05	.08	1.00	-.08	.01	.13	-.24	-.04
X26	-.14	-.15	-.10	-.26	.05	.11	.11	-.20	-.12	-.08	1.00	-.05	.04	.05	-.04
X27	.06	-.05	.14	-.11	-.10	.11	-.06	-.16	.12	.01	-.05	1.00	-.21	-.01	.07
X28	-.12	-.02	.08	-.02	.10	-.06	.22	.19	.04	.13	.04	-.21	1.00	.02	-.13
X29	.14	-.11	-.18	.00	.12	-.08	-.18	-.06	-.15	-.24	.05	-.01	.02	1.00	.02
X30	.08	.11	-.00	-.05	.07	-.06	.02	-.08	.02	-.04	-.04	.07	-.13	.02	1.00

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
X1	1.00	-.07	-.03	-.13	.08	-.10	-.04	.06	.15	-.08	-.00	-.02	-.03	.00	-.03
X2	-.07	1.00	-.00	.03	.07	.07	.02	.04	-.00	.01	*-.25*	.05	-.14	.01	-.02
X3	-.03	-.00	1.00	.14	-.09	-.02	-.01	.08	.16	-.04	-.08	-.02	-.08	.04	-.05
X4	-.13	.03	.14	1.00	-.13	.07	.18	-.14	.18	-.12	-.05	-.03	-.09	.10	.15
X5	.08	.07	-.09	-.13	1.00	.02	.02	-.10	.06	.03	-.08	.01	-.02	-.17	.12
X6	-.10	.07	-.02	.07	.02	1.00	-.01	-.17	-.01	.00	.03	.13	.04	.02	.03
X7	-.04	.02	-.01	.18	.02	-.01	1.00	-.05	-.05	.09	-.04	-.04	.02	-.09	.02
X8	.06	.04	.08	-.14	-.10	-.17	-.05	1.00	.05	-.02	-.04	-.01	-.18	-.03	.03
X9	.15	-.00	.16	.18	.06	-.01	-.05	.05	1.00	.08	.09	-.20	-.05	-.01	.00
X10	-.08	.01	-.04	-.12	.03	.00	.09	-.02	.08	1.00	.16	-.18	-.01	.04	.14
X11	-.00	-.25	-.08	-.05	-.08	.03	-.04	-.04	.09	.16	1.00	-.23	.04	-.08	-.20
X12	-.02	.05	-.02	-.03	.01	.13	-.04	-.01	-.20	-.18	-.23	1.00	.25	.09	.05
X13	-.03	-.14	-.08	-.09	-.02	.04	.02	-.18	-.05	-.01	.04	.25	1.00	-.10	-.17
X14	.00	.01	.04	.10	-.17	.02	-.09	-.03	-.01	.04	-.08	.09	-.10	1.00	-.02
X15	-.03	-.02	-.05	.15	.12	.03	.02	.03	.00	.14	-.20	.05	-.17	-.02	1.00
X16	.08	.00	-.19	-.13	.05	.04	-.09	.03	-.18	-.03	-.03	.09	.25	-.01	.12
X17	.20	-.04	.05	.06	-.08	-.13	.14	-.05	.01	-.01	-.03	.00	.02	.02	.00
X18	-.02	-.06	.05	.03	-.03	-.13	-.09	.26	.10	-.07	.10	.08	-.00	.10	.01
X19	.22	-.04	-.08	.02	.01	.19	-.05	.02	-.09	-.13	.04	.21	.23	.00	.04
X20	.19	-.04	.08	-.13	.01	-.06	-.03	.23	.01	-.07	-.10	-.11	-.09	.15	.03
X21	.03	.06	-.11	-.09	.10	.06	.12	-.23	-.27	-.08	.04	.20	.09	-.02	-.17
X22	-.18	-.03	.14	.01	.12	-.05	.02	.12	-.13	.02	-.13	-.06	-.15	.17	.07
X23	.04	-.13	.04	.05	.04	-.18	.14	.10	.05	.08	.17	.19	-.11	-.11	.03
X24	-.05	.10	-.06	-.03	-.05	-.11	.13	.00	-.13	.06	-.01	.07	-.11	.01	.13
X25	.02	-.01	-.08	-.05	-.00	-.14	.08	-.09	-.08	-.14	-.11	.15	.06	.17	.01
X26	-.08	.16	-.12	-.01	.12	.13	-.10	-.16	-.06	.00	.13	-.07	.01	-.07	.02
X27	.00	.07	-.08	-.09	-.12	-.10	-.01	.05	.01	-.01	.00	.08	.11	.09	-.03
X28	-.02	-.03	.03	-.03	.12	-.22	.03	-.05	-.09	-.00	-.20	-.07	-.03	.02	.02
X29	.09	-.02	-.07	-.04	-.13	.06	-.03	-.06	-.14	-.17	.00	-.01	-.07	-.08	.04
X30	.10	-.01	.13	-.17	.08	-.14	-.05	-.06	.16	.03	-.11	.00	.03	.16	-.02

	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
X1	.08	.20	-.02	*.22*	.19	.03	-.18	.04	-.05	.02	-.08	.00	-.02	.09	.10
X2	.00	-.04	-.06	-.04	-.04	.06	-.03	-.13	.10	-.01	.16	.07	-.03	-.02	-.01
X3	-.19	.05	.05	-.08	.08	-.11	.14	.04	-.06	-.08	-.12	-.08	.03	-.07	.13
X4	-.13	.06	.03	.02	-.13	-.09	.01	.05	-.03	-.05	-.01	-.09	-.03	-.04	-.17
X5	.05	-.08	-.03	.01	.01	.10	.12	.04	-.05	-.00	.12	-.12	.12	-.13	.08
X6	.04	-.13	-.13	.19	-.06	.06	-.05	-.18	-.11	-.14	.13	-.10	-.22	.06	-.14
X7	-.09	.14	-.09	-.05	-.03	.12	.02	.14	.13	.08	-.10	-.01	.03	-.03	-.05
X8	.03	-.05	.26	.02	.23	-.23	.12	.10	.00	-.09	-.16	.05	-.05	-.06	-.06
X9	-.18	.01	.10	-.09	.01	-.27	-.13	.05	-.13	-.08	-.06	.01	-.09	-.14	.16
X10	-.03	-.01	-.07	-.13	-.07	-.08	.02	.08	.06	-.14	.00	-.01	-.00	-.17	.03
X11	-.03	-.03	.10	.04	-.10	.04	-.13	.17	-.01	-.11	.13	.00	-.20	.00	-.11
X12	.09	.00	.08	.21	-.11	.20	-.06	.19	.07	.15	-.07	.08	-.07	-.01	.00
X13	.25	.02	-.00	.23	-.09	.09	-.15	-.11	-.11	.06	.01	.11	-.03	-.07	.03
X14	-.01	.02	.10	.00	.15	-.02	.17	-.11	.01	.17	-.07	.09	.02	-.08	.16
X15	.12	.00	.01	.04	.03	-.17	.07	.03	.13	.01	.02	-.03	.02	.04	-.02
X16	1.00	-.05	.02	.26	-.02	.20	-.12	-.01	.11	-.02	-.14	.06	-.12	.14	.08
X17	-.05	1.00	.01	.10	.02	.20	-.20	-.08	.10	.16	-.15	-.05	-.02	-.11	.11
X18	.02	.01	1.00	-.01	-.11	-.16	.02	.01	.01	-.06	-.10	.14	.08	-.18	-.00
X19	.26	.10	-.01	1.00	-.03	.05	-.13	-.06	.10	.13	-.26	-.11	-.02	.00	-.05
X20	-.02	.02	-.11	-.03	1.00	-.13	.07	.02	.03	-.10	.05	-.10	.10	.12	.07
X21	.20	.20	-.16	.05	-.13	1.00	.14	.01	-.00	.23	.11	.11	-.06	-.08	-.06
X22	-.12	-.20	.02	-.13	.07	.14	1.00	.04	-.01	.12	.11	-.06	.22	-.18	.02
X23	-.01	-.08	.01	-.06	.02	.01	.04	1.00	.20	.05	-.20	-.16	.19	-.06	-.08
X24	.11	.10	.01	.10	.03	-.00	-.01	.20	1.00	.08	-.12	.12	.04	-.15	.02
X25	-.02	.16	-.06	.13	-.10	.23	.12	.05	.08	1.00	-.08	.01	.13	-.24	-.04
X26	-.14	-.15	-.10	-.26	.05	.11	.11	-.20	-.12	-.08	1.00	-.05	.04	.05	-.04
X27	.06	-.05	.14	-.11	-.10	.11	-.06	-.16	.12	.01	-.05	1.00	-.21	-.01	.07
X28	-.12	-.02	.08	-.02	.10	-.06	.22	.19	.04	.13	.04	-.21	1.00	.02	-.13
X29	.14	-.11	-.18	.00	.12	-.08	-.18	-.06	-.15	-.24	.05	-.01	.02	1.00	.02
X30	.08	.11	-.00	-.05	.07	-.06	.02	-.08	.02	-.04	-.04	.07	-.13	.02	1.00

Inoltre nella tavola che segue sono riportati nella parte sinistra, per ciascuna variabile, la minima e la massima correlazione lineare r_{ij} ; nella parte destra si riporta per ciascuna variabile, il coefficiente di determinazione multipla R^2 che esprime la porzione di variabilità spiegata dalla regressione multipla (lineare) su tutte le altre 29 variabili:

	Min. r_{ij}	Max r_{ij}			R^2 (Var. $\mathbf{X}_i$ con tutte le altre)
X1	-.18	.22		X1	.249
X2	-.25	.16		X2	.211
X3	-.19	.16		X3	.164
X4	-.17	.18		X4	.336
X5	-.17	.12		X5	.222
X6	-.22	.19		X6	.288
X7	-.10	.18		X7	.183
X8	-.23	.26		X8	.364
X9	-.27	.18		X9	.306
X10	-.18	.16		X10	.259
X11	-.25	.17		X11	.427
X12	-.23	.25		X12	.434
X13	-.18	.25		X13	.380
X14	-.17	.17		X14	.303
X15	-.20	.15		X15	.280
X16	-.19	.26		X16	.367
X17	-.20	.20		X17	.301
X18	-.18	.26		X18	.271
X19	-.26	.26		X19	.384
X20	-.13	.23		X20	.265
X 21	-.27	.23		**X** 21	.442
X22	-.20	.22		X22	.355
X23	-.20	.20		X23	.419
X24	-.15	.20		X24	.253
X25	-.24	.23		X25	.296
X26	-.26	.16		X26	.352
X27	-.21	.14		X27	.252
X28	-.22	.22		X28	.344
X29	-.24	.14		X29	.320
X30	-.17	.16		X30	.261

Ricordo che i valori critici di r ad un livello di significatività α per un test bilaterale sono:

$$r_\alpha = \sqrt{\frac{t_\alpha^2}{t_\alpha^2 + n - 2}}$$

essendo t_α il valore critico ad un livello α per una t con $n-2$ gradi di libertà.

Nel nostro caso, lavorando al 5

$$r_\alpha = \sqrt{\frac{1.9845^2}{1.9845^2 + 98}} = 0.197$$

Per quanto riguarda R^2 analogamente ricaviamo (dalla distribuzione F):

$$R_\alpha^2 = \frac{kF_\alpha}{kF_\alpha + n - k - 1}$$

essendo k il numero dei regressori e F_α il valore critico ad un livello α per una F di Snedecor con k ed $n-k-1$ gradi di libertà.

Nel nostro caso:

$$R_\alpha^2 = \frac{29x1.6294}{29x1.6294 + 70} = 0,403$$

Di seguito sono riportati anche gli autovalori ricavati dalle 30 variabili standardizzate:

	C		omponenti principali
i	Autovalore	varianza	varianza cumulata
1	2.300	7.668	7.67
2	1.999	6.662	14.33
3	1.925	6.417	20.75
4	1.690	5.634	26.38
5	1.621	5.402	31.78
6	1.560	5.200	36.98
7	1.529	5.098	42.08
8	1.429	4.764	46.85
9	1.332	4.440	51.29
10	1.206	4.021	55.31
11	1.135	3.784	59.09
12	1.105	3.682	62.77
13	1.009	3.363	66.14
14	.968	3.227	69.36
15	.899	2.996	72.36
16	.885	2.949	75.31
17	.854	2.845	78.15
18	.824	2.745	80.90
19	.760	2.532	83.43
20	.740	2.466	85.90
21	.656	2.187	88.08
22	.583	1.945	90.03
23	.542	1.808	91.84
24	.527	1.756	93.59
25	.391	1.305	94.90
26	.380	1.267	96.16
27	.360	1.201	97.36
28	.323	1.078	98.44
29	.243	.811	99.25
30	.224	.747	100.00