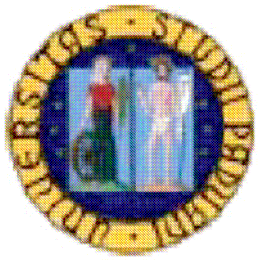
UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA, ECONOMIA E FINANZA



TESI DI LAUREA

LA REGRESSIONE MULTIVARIATA: UN'APPLICAZIONE A DATI ANTROPOLOGICI

Relatore: Ch.mo Prof. Laura Ventura

Laureanda: Marina Marzovilli

Matricola: 515937

ANNO ACCADEMICO 2006-2007

AI MIEI GENITORI

Indice

Introduzione

La regressione multivariata riguarda il problema della modellazione di più di una variabile risposta a partire dallo stesso insieme di variabili esplicative. È possibile, in questo caso, cercare di capire come si può tener conto dalle correlazioni esistenti tra le risposte per comparare l'accuratezza della previsione con quella delle procedure che generalmente si usano nel caso di regressioni univariate, nelle quali per ogni singola risposta viene utilizzato un insieme di regressori.

Nel presentare il caso in cui la risposta consiste in $p > 1$ variabili differenti si possono facilmente estendere le considerazioni fatte per la regressione multipla all'analisi multivariata. La regressione multivariata è definita dal modello

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{U}, \qquad \textbf{(1)}$$

dove $\boldsymbol{Y}$ non è più il vettore ($n \times 1$) considerato nel caso dell'analisi univariata, con n che rappresenta il numero di dati osservati, bensì una matrice ($n \times p$) contenente p variabili risposta; $\boldsymbol{X}$ è una matrice ($n \times q$) di variabili esplicative (con $q \geq 1$); $\boldsymbol{B}$ è una matrice ($q \times p$) di coefficienti di regressione e $\boldsymbol{U}$ è la matrice dei termini d'errore di dimensione ($n \times p$).

Il modello **(1)** può essere scritto come

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1q} \\ 1 & x_{22} & \cdots & x_{2q} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n2} & \cdots & x_{nq} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{1p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix} + \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ u_{n1} & u_{n2} & \cdots & u_{np} \end{bmatrix}$$

Siano $\boldsymbol{y_i}$ e $\boldsymbol{x_i}$ l'*i*-esima ($i = 1,\ldots, n$) riga, rispettivamente, delle matrici $\boldsymbol{Y}$ e $\boldsymbol{X}$; nel caso del modello con intercetta, allora la prima colonna della matrice $\boldsymbol{X}$ è il vettore unitario. Se $\boldsymbol{X}$ è una matrice di q variabili esplicative osservate su ognuno degli n individui del campione considerato, allora parliamo di *modello di regressione multivariata*. Nel modello **(1)** le colonne di $\boldsymbol{Y}$, che rappresentano le variabili dipendenti del modello osservate sugli n individui del campione, sono spiegate dalle colonne di $\boldsymbol{X}$. In particolare vale la relazione

$$\mathrm{E}(\boldsymbol{y_{ij}}) = \boldsymbol{x_i}^{\mathrm{T}}\boldsymbol{B_j}, \qquad i = 1,\ldots n,\ j = 1,\ldots,p,$$

ossia il valore atteso di $\boldsymbol{y_{ij}}$ dipende dall'*i*-esima riga della matrice $\boldsymbol{X}$ e dalla *j*-esima colonna $\boldsymbol{B_j}$ della matrice dei coefficienti di regressione $\boldsymbol{B}$. Nel caso in cui $p = 1$, abbiamo di fronte il modello di regressione multipla, con una sola variabile risposta, per il quale possiamo scrivere nel modo usuale $\boldsymbol{y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}$.

Nella maggior parte dei casi è possibile assumere che la matrice $\boldsymbol{U}$ degli errori sia distribuita normalmente: $\boldsymbol{U}$ è una matrice con una distribuzione $N_p(0, \boldsymbol{\Sigma})$, dove $\boldsymbol{\Sigma}$ è la matrice di covarianza. Si assume, inoltre, che la distribuzione di $\boldsymbol{U}$ sia incorrelata con $\boldsymbol{X}$, ossia che $\mathrm{E}(\boldsymbol{U}|\boldsymbol{X}) = \boldsymbol{0}$. Sotto l'assunzione di normalità degli errori, la log-verosimiglianza per i parametri $\boldsymbol{B}$ e $\boldsymbol{\Sigma}$ è data da

$$l(\boldsymbol{B}, \boldsymbol{\Sigma}) = -\frac{1}{2} n \log |2\pi\, \boldsymbol{\Sigma}| - \frac{1}{2} \mathrm{tr}[(\boldsymbol{Y}-\boldsymbol{XB})\, \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}-\boldsymbol{XB})^{\mathrm{T}}], \qquad \textbf{(2)}$$

dove il simbolo $\mathrm{tr}(\boldsymbol{A})$ indica la traccia della matrice $\boldsymbol{A}$, ossia la somma degli elementi che si trovano sulla diagonale di tale matrice.

Nei capitoli successivi verranno sviluppati i seguenti argomenti: il primo capitolo presenterà alcuni concetti generali relativi alla teoria della regressione multivariata e all'inferenza sui parametri stimati nel modello oggetto di studio; nel secondo capitolo si tratterà dell'analisi multivariata della varianza (MANOVA) e dei test statistici ad essa correlati; nel terzo capitolo saranno presentati i dati oggetto di studio e verranno eseguite analisi grafiche preliminari e test su di essi; infine, nell'ultimo capitolo si stimerà il modello di regressione multivariata sui dati, valutando gli effetti che hanno le variabili esplicative sulle risposte considerate; in appendice sono contenute alcune informazioni relative alle distribuzioni multivariate legate ai test statistici presentati nel seguito.

Capitolo 1

Inferenza nel modello di regressione multivariata

Nel modello **(1)**, per la stima dei parametri $\boldsymbol{B}$ e $\boldsymbol{\Sigma}$ è possibile utilizzare il metodo della massima verosimiglianza, che si basa sulla soluzione di equazioni di stima ottenute a partire dalla **(2)**. Si ottengono così gli stimatori di massima verosimiglianza per la matrice dei coefficienti $\boldsymbol{B}$ e per la matrice $\boldsymbol{\Sigma}$ di varianza e covarianza degli errori, dove $\boldsymbol{\Sigma}$ è una matrice definita positiva e simmetrica, di dimensioni ($p \times p$). A partire dai risultati ottenuti è possibile eseguire verifiche di ipotesi sui parametri del modello e costruire intervalli di confidenza. È possibile, inoltre, considerare delle generalizzazioni del modello quando alcune delle ipotesi considerate non valgono. Ad esempio, si può vedere coma cambiano le stime dei coefficienti nel caso in cui i termini d'errore siano eteroschedastici, ossia con varianze diverse.

1.1 Stima di massima verosimiglianza dei coefficienti di regressione e della matrice $\boldsymbol{\Sigma}$

Sia $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$, dove $\boldsymbol{I}$ è la matrice identità di ordine n. La matrice $\boldsymbol{P}$ gode di alcune proprietà: è simmetrica ed idempotente ($\boldsymbol{P} = \boldsymbol{P}^{\mathrm{T}}$, $\boldsymbol{P} \times \boldsymbol{P} = \boldsymbol{P}^2 = \boldsymbol{P}$) ed ha rango n-q. Essa rappresenta la proiezione ortogonale in $\boldsymbol{R}^{\boldsymbol{n}}$ delle colonne di $\boldsymbol{X}$. In particolare, $\boldsymbol{PX} = \boldsymbol{0}$.

Nel modello **(1)** considerato si ipotizza che le righe della matrice $\boldsymbol{U}$ siano indipendenti e distribuite normalmente con un vettore p-dimensionale di medie pari a zero e con una

matrice di varianza e covarianza $\boldsymbol{\Sigma}$ ($p \times p$). Ciò implica che per ogni colonna della matrice delle risposte ci sia una regressione completa col suo insieme di coefficienti di regressione. La matrice stimata dei coefficienti è ottenuta come

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y},$$

dove la matrice $\hat{\boldsymbol{B}}$ è distribuita come una *t* di Student multivariata (si veda l'Appendice) con media $\boldsymbol{B}$ e matrice di varianza $[(n - q - 1)(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})]^{-1}\boldsymbol{G}$, ossia

$$\hat{\boldsymbol{B}} \sim t\left(n - q - 1, \boldsymbol{B}, \left[(n - q - 1)(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})\right]^{-1}\boldsymbol{G}\right)$$

con $\boldsymbol{G} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^{\mathrm{T}})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})$.

In particolare $\hat{\boldsymbol{B}}_k$, che rappresenta la *k*-esima colonna di $\hat{\boldsymbol{B}}$ ($k = 1,\ldots,p$), si distribuisce come una *t* di Student multivariata con media $\boldsymbol{B}_k$ e matrice di varianza e covarianza $(n - q - p)^{-1}\boldsymbol{W}_{kk}\boldsymbol{G}$; ossia

$$\hat{\boldsymbol{B}}_k \sim t(n - q - p, \boldsymbol{B}_k, (n - q - p)^{-1}\boldsymbol{W}_{kk}\boldsymbol{G}),$$

dove $\boldsymbol{W}_{kk}$ è l'elemento di posto (k,k) della matrice $\boldsymbol{W} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$.

Inoltre, per il singolo stimatore $\hat{\boldsymbol{B}}_{jk}$ ($k = 1,\ldots,p$, $j = 1,\ldots q$), la distribuzione di riferimento è

$$\hat{\boldsymbol{B}}_{jk} \sim t(n - q - p, \boldsymbol{B}_{jk}, (n - q - p)^{-1}\boldsymbol{W}_{kk}\boldsymbol{g}_{jj}),$$

che rappresenta una t di Student univariata con media $\boldsymbol{B}_{jk}$ e varianza $(n - q - p)^{-1}\boldsymbol{W}_{kk}\boldsymbol{g}_{jj}$, dove $\boldsymbol{g}_{jj}$ è l'elemento *j*-esimo sulla diagonale di $\boldsymbol{G}$. (si veda Rowe [6])

La stima della matrice dei coefficienti di regressione $\hat{\boldsymbol{B}}$ non dipende dalla covarianza rappresentata dalla matrice $\boldsymbol{\Sigma}$ o dalla sua stima; rimane la stessa a prescindere dalla presenza o meno di correlazione.

Per il modello **(1),** $\hat{\boldsymbol{B}}$ è una stima non distorta di $\boldsymbol{B}$. Infatti sostituendo $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{U}$ nell'equazione di $\hat{\boldsymbol{B}}$ si ottiene

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{B} + \boldsymbol{U}) = \boldsymbol{B} + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{U}$$

e quindi vale la relazione $\boldsymbol{E}(\hat{\boldsymbol{B}}) = \boldsymbol{B}$ poichè $\boldsymbol{E}(\boldsymbol{U}) = \boldsymbol{0}$. Infatti $\boldsymbol{U}$ è la matrice dei termini di errore tale per cui le medie sono nulle e quindi $\boldsymbol{E}(\hat{\boldsymbol{B}}) = \boldsymbol{B}$.

La covarianza tra $\hat{\boldsymbol{B}}_{jk}$ e $\hat{\boldsymbol{B}}_{kl}$ è data da $\boldsymbol{\sigma_{jl} w_{ik}}$, dove $\boldsymbol{\sigma_{jl}}$ è l'elemento di posto (j,l) della matrice di varianza e covarianza degli errori $\boldsymbol{\Sigma}$, mentre $\boldsymbol{w}_{ik}$ è l'elemento di posto (i,k) della matrice $\boldsymbol{W} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$. Infatti dall'espressione di $\hat{\boldsymbol{B}}$, si ricava che $\hat{\boldsymbol{\beta}}_{\mathrm{ij}} - \boldsymbol{\beta}_{\mathrm{ij}} = \boldsymbol{a_i}^{\mathrm{T}}\boldsymbol{u_j}$, dove $\boldsymbol{a}_i$ è la i-esima riga della matrice $\boldsymbol{A} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$, con $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \boldsymbol{W}$,e $\boldsymbol{u_j}$ è la j-esima colonna della matrice degli errori $\boldsymbol{U}$. Allora

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{ij}}, \hat{\boldsymbol{\beta}}_{\mathrm{kl}}) = \boldsymbol{a_i}^{\mathrm{T}}\,\mathrm{E}[\boldsymbol{u_i}\,\boldsymbol{u_l}^{\mathrm{T}}]\boldsymbol{a_k} = \boldsymbol{a_i}^{\mathrm{T}}[\boldsymbol{\sigma_{jl}I}]\boldsymbol{a_k} = \boldsymbol{\sigma_{jl}}\,\boldsymbol{w_{ik}}. \qquad \textbf{(1.1)}$$

Oltre ai coefficienti di regressione, per il modello si ottiene anche una stima della matrice $\boldsymbol{\Sigma}$, come

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{n}^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Y}.$$

Si può dimostrare che per $\hat{\boldsymbol{\Sigma}}$ vale

$$n\hat{\boldsymbol{\Sigma}} \sim W_p(\boldsymbol{\Sigma}, n - q),$$

con W_p che indica una distribuzione di Wishart con $n - q$ gradi di libertà e matrice di scala $\boldsymbol{\Sigma}$. Tale distribuzione rappresenta una generalizzazione della distribuzione χ^2 nel caso multivariato (si veda l'Appendice).

Una volta calcolati i coefficienti di regressione, è possibile ottenere i valori predetti dal modello ("*fitted values*") come

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.$$

Inoltre, i residui si calcolano nel seguente modo

$$\hat{\boldsymbol{U}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{P}\boldsymbol{Y}.$$

Da ciò si ottiene che $\mathbf{E}(\hat{\boldsymbol{U}}) = \mathbf{0}$. Infatti

$$\mathbf{E}(\hat{\boldsymbol{U}}) = \mathbf{E}(\boldsymbol{Y}) - \mathbf{E}(\boldsymbol{X}\hat{\boldsymbol{B}}) = \boldsymbol{X}\boldsymbol{B} - \boldsymbol{X}\boldsymbol{B} = \mathbf{0}.$$

Nel modello di regressione multivariata può essere necessario, come nel caso univariato, effettuare una selezione delle variabili esplicative da utilizzare come regressori. Nella selezione dei regressori bisogna tener conto del fatto che la matrice $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ deve essere invertibile e che è necessario scegliere solo le variabili risposta non troppo strettamente correlate tra loro. In particolare, per la selezione dei regressori, le procedure più utilizzate sono la selezione *backward* e quella *forward* che permettono in modo quasi automatico di eliminare o aggiungere le esplicative che non influenzano, o

influenzano, significativamente le variabili risposta. Per approfondimenti si veda Mardia *et al.* (1979).

1.2 La verifica d'ipotesi

Nel modello **(1)** si possono considerare ipotesi nulla della forma H_0: $\boldsymbol{C_1BM_1} = \boldsymbol{D}$, dove $\boldsymbol{C_1}$ è una matrice ($g \times q$) di rango g, $\boldsymbol{M_1}$ è una matrice ($p \times r$) di rango r e $\boldsymbol{D}$ è una matrice ($g \times r$). In molti casi si può supporre che $\boldsymbol{D} = \boldsymbol{0}$ e $\boldsymbol{M_1} = \boldsymbol{I}$, e l'ipotesi nulla diventa semplicemente H_0: $\boldsymbol{C_1B} = \boldsymbol{0}$. Le righe di $\boldsymbol{C_1}$ permettono di capire qual è l'effetto sulla regressione di una combinazione lineare delle variabili indipendenti, mentre le colonne di $\boldsymbol{M_1}$ focalizzano l'attenzione su particolari combinazioni lineari di variabili dipendenti. Ci sono due metodi per verificare tali ipotesi: il test del rapporto di verosimiglianza (*likelihood ratio test*, LRT) ed il test unione ed intersezione (*union intersection test*, UIT).

1.2.1 LRT

Considerando l'ipotesi nulla H_0: $\boldsymbol{C_1B} = \mathbf{D}$ (con $\boldsymbol{M_1} = \boldsymbol{I}$), è opportuno definire ulteriori matrici per costruire il test. Sia $\boldsymbol{C_2}$ la matrice tale per cui $\boldsymbol{C}^{\mathrm{T}} = (\boldsymbol{C_1}^{\mathrm{T}}, \boldsymbol{C_2}^{\mathrm{T}})$ sia di ordine ($q \times q$) non singolare e sia $\boldsymbol{B_0}$ la matrice di dimensioni ($q \times p$) che soddisfa la relazione $\boldsymbol{C_1B_0} = \boldsymbol{D}$. È allora possibile riscrivere il modello **(1)** come

$$\boldsymbol{Y}_+ = \boldsymbol{Z\Delta} + \boldsymbol{U},$$

dove $\boldsymbol{Y}_+ = \boldsymbol{Y} - \boldsymbol{XB_0}$, $\boldsymbol{Z} = \boldsymbol{XC}^{-1}$ e $\boldsymbol{\Delta} = (\boldsymbol{\Delta_1}^{\mathrm{T}}, \boldsymbol{\Delta_2}^{\mathrm{T}}) = \boldsymbol{C}(\boldsymbol{B} - \boldsymbol{B_0})$. Allora l'ipotesi nulla H_0: $\boldsymbol{C_1B} = \boldsymbol{D}$ può essere riscritta come H_0: $\boldsymbol{\Delta_1} = \boldsymbol{0}$. Considerando la partizione $\boldsymbol{C}^{-1} = (\boldsymbol{C^{(1)}}, \boldsymbol{C^{(2)}})$ si può definire la matrice di proiezione $\boldsymbol{P_1}$ sul sottospazio ortogonale della colonne di $\boldsymbol{XC^{(2)}}$, come

$$\boldsymbol{P_1} = \boldsymbol{I} - \boldsymbol{XC^{(2)}}(\boldsymbol{C^{(2)}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{XC^{(2)}})^{-1}\,\boldsymbol{C^{(2)}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}.$$

Dall'espressione della funzione di verosimiglianza **(2)**, si ottiene che sotto le ipotesi nulla e alternativa, le verosimiglianze massimizzate corrispondono rispettivamente a

$$|2\pi n^{-1}\boldsymbol{Y}^{\mathrm{T}}{}_+\boldsymbol{P_1}\boldsymbol{Y}_+|^{-n/2}\exp\left\{-\frac{1}{2}np\right\} \quad \text{e} \quad |2\pi \mathrm{n}^{-1}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{PY}|^{-n/2}\exp\left\{-\frac{1}{2}np\right\}.$$

La statistica LRT si ottiene quindi come

$$\text{LRT} = |\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Y}|/|\boldsymbol{Y}_{+}\boldsymbol{P}_{\boldsymbol{1}}\boldsymbol{Y}_{+}| \, .$$

Se si definisce $\boldsymbol{P_2} = \boldsymbol{P_1} - \boldsymbol{P}$, anche $\boldsymbol{P_2}$ è una matrice di proiezione. In particolare, l'espressione di tale matrice sarà

$$\boldsymbol{P_2} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C_1}^{\mathrm{T}}[\boldsymbol{C_1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C_1}^{\mathrm{T}}]^{-1}\,\boldsymbol{C_1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}.$$

Inoltre, poiché $\boldsymbol{PX} = \boldsymbol{0}$, allora $\boldsymbol{PY} = \boldsymbol{PY}_{+}$.

Definita $\boldsymbol{P_2}$, è quindi possibile riscrivere la statistica LRT come

$$\text{LRT} = |\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Y}|/|\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}\,\boldsymbol{Y} + \boldsymbol{Y}_{+}\boldsymbol{P}_{\boldsymbol{2}}\boldsymbol{Y}_{+}| \qquad \textbf{(1.2)}$$

ed essa ha distribuzione $\Lambda(p, n - q, g)$, sotto H_0. Tale distribuzione si ottiene dal rapporto tra i determinanti di due matrice. Siano $\boldsymbol{A} \sim W_p(\boldsymbol{I}, m)$ e $\boldsymbol{B} \sim W_p(\boldsymbol{I}, n)$, indipendenti, dove $W_p(\boldsymbol{\Sigma}, m)$ indica una distribuzione di Wishart con matrice di scala $\boldsymbol{\Sigma}$ ed m gradi di libertà. Se $m \geq p$ allora

$$\Lambda = |\boldsymbol{A}|/|\boldsymbol{A} + \boldsymbol{B}| \;\sim \Lambda(p, m, n)$$

(si veda l'Appendice).

Per maggiori chiarimenti sulla distribuzione Λ e sulle distribuzioni ad essa collegate si rimanda al Mardia *et al.* (1979).

Nel caso in cui si voglia utilizzare questo test per dimostrare l'ipotesi nulla H_0: $\boldsymbol{C_1}\boldsymbol{B}\boldsymbol{M_1} = \boldsymbol{D}$, con $\boldsymbol{M_1} \neq \boldsymbol{I}$, la statistica test che si ottiene è una generalizzazione del LRT e prende il nome di lambda di Wilks (si veda il Cap. 2).

1.2.2. UIT

L'ipotesi nulla H_0: $\boldsymbol{C_1}\boldsymbol{B}\boldsymbol{M_1} = \boldsymbol{D}$ è vera e se e solo se vale la relazione $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{C_1}\boldsymbol{B}\boldsymbol{M_1}\boldsymbol{a} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{a}$, per ogni valore di $\boldsymbol{a}$ e $\boldsymbol{b}$. Se si sostituiscono rispettivamente $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{C_1}$ e $\boldsymbol{M_1}\boldsymbol{a}$ nelle relazioni

$$\boldsymbol{H} = \boldsymbol{M_1}^{\mathrm{T}}\boldsymbol{Y}_{+}{}^{\mathrm{T}}\boldsymbol{P_2}\boldsymbol{Y}_{+}\boldsymbol{M_1} \;\text{ ed }\; \boldsymbol{E} = \boldsymbol{M_1}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Y}\boldsymbol{M_1},$$

e scriviamo il rapporto tra $\boldsymbol{H}$ ed $\boldsymbol{E}$, si ottiene

$$\frac{\left\{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{C}_1(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}_{+}\boldsymbol{M}_1\boldsymbol{a}\right\}}{\left\{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{C}_1(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C}^{\mathrm{T}}{}_1\boldsymbol{b}\right\}\left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}{}_1\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{Y}\boldsymbol{M}_1\boldsymbol{a}\right\}} \qquad \textbf{(1.3)}$$

che, sotto l'ipotesi nulla, si distribuisce come una $F_{1,\,n-q}$ moltiplicata per $(n - q)^{-1}$, per $\boldsymbol{a}$ e $\boldsymbol{b}$ fissati.

Massimizzando rispetto a $\boldsymbol{b}$ si ottiene

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{a}/\boldsymbol{a}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{a},$$

che, sotto H_0 si distribuisce come una $[g/(n - q)]F_{g,\, n-q}$, con $\boldsymbol{a}$ fissato (perché $\boldsymbol{H}$ ed $\boldsymbol{E}$ hanno una distribuzione di Wishart e sono tra loro indipendenti).

Infine, massimizzando rispetto ad $\boldsymbol{a}$ si ottiene λ_1, che è il più grande degli autovalori di $\boldsymbol{H}\boldsymbol{E}^{-1}$.

Posto $\theta = \lambda_1 /(1+\lambda_1)$, θ rappresenta la più grande delle radici di $\boldsymbol{H}(\boldsymbol{H} + \boldsymbol{E})^{-1}$. L'ipotesi nulla viene rifiutata per valori elevati di θ, che segue, sotto H_0, la distribuzione $\theta(r, n - q, g)$. Si ricorda che per $\boldsymbol{A} \sim W_p(\boldsymbol{I}, m)$ e $\boldsymbol{B} \sim W_p(\boldsymbol{I}, n)$, indipendenti con $m \geq p$, allora il più grande degli autovalori di $(\boldsymbol{A} + \boldsymbol{B})^{-1}\boldsymbol{B}$ ha distribuzione $\theta(m, n, p)$. (Si veda l'Appendice).

Nel caso in cui il rango della matrice $\boldsymbol{M_1}$ sia $r = 1$, allora non è necessario massimizzare anche sotto $\boldsymbol{a}$, perché la statistica diventa semplicemente un rapporto tra scalari. In questo caso il test UIT diventa equivalente al test LRT.

Se invece il rango di $\boldsymbol{C_1}$ è $g = 1$, allora la massimizzazione rispetto a $\boldsymbol{b}$ non è necessaria, poiché in questo caso sia $\boldsymbol{H}$ che $\boldsymbol{H}\boldsymbol{E}^{-1}$ hanno rango 1. Ancora una volta le due statistiche UIT e LRT sono equivalenti. Infatti, l'unico autovalore non nullo di $\boldsymbol{H}\boldsymbol{E}^{-1}$ è

$$\lambda_1 = \mathrm{tr}\,(\boldsymbol{H}\boldsymbol{E}^{-1}) = \{\boldsymbol{C_1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C_1}^{\mathrm{T}}\}^{-1}\boldsymbol{C_1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}_{+}\boldsymbol{M_1}\boldsymbol{E}^{-1}\boldsymbol{M_1}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}{}_{+}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C_1}^{\mathrm{T}}.$$

Inoltre, se fissiamo $\boldsymbol{d} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{C_1}^{\mathrm{T}}$, allora sotto l'ipotesi nulla H_0 si ha che $\boldsymbol{M_1}^{\mathrm{T}}\boldsymbol{Y}_{+}{}^{\mathrm{T}}\boldsymbol{d} \sim N_p(\boldsymbol{0},\ (\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d})\boldsymbol{M_1}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{M_1})$ e sotto entrambe le ipotesi, nulla e alternativa, che $\boldsymbol{E} \sim W_r(\boldsymbol{M_1}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{M_1}, n - q)$. Allora sotto l'ipotesi nulla, λ_1 ha distribuzione

$$\lambda_1 \sim (n - q)^{-1}T^2_{r,\, n-q} = [r/(n - q - r+1)]F_{r,\, n-q-r+1},$$

dove T^2 indica la distribuzione di Hotelling (si veda l'Appendice).

Supponendo ora che $\boldsymbol{B}$ sia il vero valore della matrice dei coefficienti, dal test UIT si ottiene che la probabilità che il rapporto descritto nella **(1.2)** sia minore o uguale di $\theta_\alpha / (1 - \theta_\alpha)$, per tutti i valori di $\boldsymbol{a}$ e $\boldsymbol{b}$,è pari a $1 - \alpha$, con θ_α che indica il percentile di livello α della distribuzione $\theta(r, n - q, g)$.

1.3 Intervalli di confidenza simultanei per *B*

Utilizzando i risultati presentati nel paragrafo precedente è possibile ottenere degli intervalli di confidenza simultanei (SCIs) per le stime dei parametri $\hat{B}$, ossia tali per cui

$$P\left(b^T C_1 B M_1 a \in b^T C_1 \left(X^T X\right)^{-1} X^T Y M_1 a \pm \left\{ \frac{\theta_\alpha}{1-\theta_\alpha} \left(a^T E a\right) \left[b^T C_1 \left(X^T X\right)^{-1} C_1^T b \right] \right\}^{1/2}, \forall a, b \right) = 1-\alpha$$

Il valore $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{C_1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{M_1}\boldsymbol{a} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{C_1}\,\hat{\boldsymbol{B}}\,\boldsymbol{M_1}\boldsymbol{a}$ è uno stimatore non distorto di $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{C_1}\boldsymbol{B}\boldsymbol{M_1}\boldsymbol{a}$. In alcune applicazioni $\boldsymbol{a}$ e/o $\boldsymbol{b}$ sono forniti a priori e i limiti dell'intervallo di confidenza possono essere ristretti. Ad esempio, se $\boldsymbol{a}$ è fissato, allora gli intervalli di confidenza per $\boldsymbol{b}$ si possono ottenere sostituendo $\theta_\alpha/(1 - \theta_\alpha)$ con $[g/(n-q)]F_{r,n-q-r+1;\alpha}$, che rappresenta il percentile di livello α di una F con r e $n-q-r+1$ gradi di libertà.

Se invece è $\boldsymbol{b}$ ad essere fissato, allora $\theta_\alpha/(1 - \theta_\alpha)$ può essere sostituito da

$$(n-q)^{-1}T^2_{r,n-q;\alpha}=[r/(n-q-r+1)]F_{r,n-q-r+1;\alpha}.$$

Infine, se sono noti sia $\boldsymbol{a}$ che $\boldsymbol{b}$ si può ottenere un singolo intervallo di confidenza sostituendo $\theta_\alpha/(1 - \theta_\alpha)$ con $(n - q)^{-1}F_{1,\, n-q;\alpha}$.

1.4 Alcune considerazioni conclusive

1.4.1 Stima dei coefficienti nel caso in cui *X* non sia di rango pieno

Talvolta può succedere, per ragioni di simmetria nel disegno sperimentale, che la matrice $\boldsymbol{X}$ di dimensioni $(n \times q)$ delle esplicative non sia di rango pieno, ma abbia rango $k < q$. Ovviamente, in questa situazione, la matrice $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ non è più invertibile e quindi non è possibile calcolare le stime dei coefficienti di regressione. In questo caso, per risolvere il problema, può essere utile considerare una partizione della matrice $\boldsymbol{X}$, in modo da restringere la regressione ad un gruppo di colonne della matrice delle esplicative. In particolare, si considera $\boldsymbol{X} = (\boldsymbol{X_1},\boldsymbol{X_2})$, dove $\boldsymbol{X_1}$ è la matrice di dimensioni $(n \times k)$ di rango pieno. Allora $\boldsymbol{X}_2$ può essere scritta come $\boldsymbol{X_2} = \boldsymbol{X_1}\boldsymbol{A}$, dove $\boldsymbol{A}$ è una generica matrice di dimensione $(k \times (q-k))$. Analogamente, anche per la matrice $\boldsymbol{B}$ si considera la partizione $\boldsymbol{B}^{\mathrm{T}}= (\boldsymbol{B_1}^{\mathrm{T}},\boldsymbol{B_2}^{\mathrm{T}})$.

Il modello può essere dunque riformulato nel seguente modo

$$\boldsymbol{Y} = \boldsymbol{X_1 B^*_1} + \boldsymbol{U_1},$$

dove $\boldsymbol{B^*_1}$ indica la somma $\boldsymbol{B_1} + \boldsymbol{A B_2}$ univocamente determinata. Effettuando queste sostituzioni, è possibile ottenere comunque una stima dei coefficienti, nonostante la matrice delle esplicative $\boldsymbol{X}$ non sia di rango pieno, poiché il modello $\boldsymbol{Y} = \boldsymbol{X_1 B^*_1} + \boldsymbol{U_1}$ permette di utilizzare gli usuali metodi di stima per $\boldsymbol{B^*_1}$, che rappresenta una riparametrizzazione di $\boldsymbol{B}$.

1.4.2 Valutazione dell'effetto della media sulle risposte

Nel modello considerato talvolta può essere utile separare l'effetto della media $\boldsymbol{\mu}$ sulle risposte, dall'effetto delle altre variabili esplicative indipendenti. Il modello può essere riscritto come

$$\boldsymbol{Y} = \boldsymbol{1\mu}^{\mathrm{T}} + \boldsymbol{XB} + \boldsymbol{U}, \qquad \textbf{(1.4)}$$

dove $(\boldsymbol{1}, \boldsymbol{X})$ rappresenta una matrice $(n \times (1 + q))$. Allora

$$\left[\begin{pmatrix} 1^T \\ X^T \end{pmatrix}\begin{pmatrix} 1 & X \end{pmatrix}\right]^{-1} = \begin{pmatrix} n^{-1} & 0 \\ 0 & (X^T X)^{-1} \end{pmatrix},$$

poiché $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{1} = 0$. Allora, poiché vale la **(1.1)**, si ha $\bar{\boldsymbol{\mu}} = \bar{\boldsymbol{y}}$, con $\bar{\boldsymbol{y}}$ vettore di medie calcolate su tutti i dati, che è indipendente dalla stima di $\boldsymbol{B}$ che si ottiene come

$$\hat{\boldsymbol{B}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{1}\bar{\boldsymbol{y}}^{\mathrm{T}}).$$

In particolare, se la media di $\boldsymbol{Y}$ è stimata dal vettore $\bar{\boldsymbol{y}}$, allora la **(1.4)** si può anche scrivere come

$$\boldsymbol{Y} - \boldsymbol{1}\bar{\boldsymbol{y}}^{\mathrm{T}} = \boldsymbol{XB} + \boldsymbol{Z},$$

dove con $\boldsymbol{Z}$, si intende $\boldsymbol{Z} = \boldsymbol{U} - \boldsymbol{1}\bar{\boldsymbol{u}}^{\mathrm{T}}$, che rappresenta la matrice degli errori centrata. Ogni vettore colonna di tale matrice ha distribuzione $N_{np}(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{Q})$, dove $\boldsymbol{Q} = \boldsymbol{I} - n^{-1}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}$. Con il simbolo $\otimes$ si indica il prodotto di Kronecker. Sia $\boldsymbol{A}$ una matrice $(m \times n)$, con elementi $\boldsymbol{a_{ij}}$ e $\boldsymbol{B}$ una matrice $(p \times q)$ di elementi $\boldsymbol{b_{kl}}$, allora il prodotto di Kronecker di $\boldsymbol{A}$ e $\boldsymbol{B}$ è definito come

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}B & a_{12}B & . & . & . & a_{1n}B \\ a_{12}B & a_{22}B & . & . & . & a_{2n}B \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ a_{m1}B & a_{m2}B & . & . & . & a_{mn}B \end{bmatrix}$$

che è una matrice ($mp \times nq$).

Quanto descritto rappresenta un'estensione di quanto detto in precedenza, nel caso in cui, nello stimare i parametri $\boldsymbol{B}$, si voglia valutare separatamente l'effetto della media $\boldsymbol{\mu}$, rispetto alle altre esplicative e quindi capire qual è il peso di $\boldsymbol{\mu}$ nella procedura di determinazione di $\hat{\boldsymbol{B}}$.

1.4.3 Dalle stime OLS alle stime GLS

Nel modello univariato, il calcolo delle stime dei coefficienti di regressione si possono anche basare sul metodo dei minimi quadrati ordinari, sotto l'ipotesi che i termini d'errore abbiano media nulla e siano omoschedastici (ipotesi di Gauss-Markov). Quando la matrice di varianza e covarianza degli errori non può più essere supposta uguale a $\boldsymbol{\sigma^2 I}$, lo stimatore OLS non è più BLUE (*Best Linear Unbiased Estimator*). In queste situazioni si utilizza la stima GLS (*Generalized Least Squares*), che si basa su una trasformazione della matrice di covarianza. Partendo dalle assunzioni

$$\boldsymbol{E}(\boldsymbol{u}) = \boldsymbol{0} \quad \text{e} \quad \boldsymbol{V}(\boldsymbol{u}) = \boldsymbol{\Omega},$$

con $\boldsymbol{\Omega}$ supposta nota, si considera la trasformazione del modello

$$\boldsymbol{z} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{X}\boldsymbol{B} + \boldsymbol{v}, \qquad \textbf{(1.5)}$$

dove

$$\boldsymbol{z} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{y} \qquad \text{e} \qquad \boldsymbol{v} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{u}.$$

Il modello **(1.5)** soddisfa le ipotesi di Gauss-Markov e quindi permette di calcolare lo stimatore non distorto di $\boldsymbol{B}$.

Nel caso multivariato, ciò che si verifica è che la stima OLS, che si ottiene se $\boldsymbol{\Sigma} = \boldsymbol{\sigma^2 I}$ e la stima GLS conducono allo stesso stimatore. Riscrivendo il modello in forma vettoriale, si ottiene infatti

$$\boldsymbol{Y}^{\boldsymbol{V}} = \boldsymbol{X}^{*}\boldsymbol{B}^{\boldsymbol{V}} + \boldsymbol{U}^{\boldsymbol{V}},$$

dove $\boldsymbol{Y}^{\boldsymbol{V}} = (\boldsymbol{y}^{\mathrm{T}}_{(1)},\ldots, \boldsymbol{y}^{\mathrm{T}}_{(p)})$ e $\boldsymbol{X}^{*} = \boldsymbol{I}_p \otimes \boldsymbol{X}$. Per quanto riguarda il termine d'errore $\boldsymbol{U}^{\boldsymbol{V}}$, si assume che esso abbia media $\mathbf{0}$ e matrice di varianza e covarianza

$$\boldsymbol{V}(\boldsymbol{U}^{V}) = \boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_p.$$

Allora lo stimatore GLS di $\boldsymbol{B}$ si ottiene come

$$\begin{aligned}\hat{\boldsymbol{B}}^{V} &= (\boldsymbol{X}^{*\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{X}^{*})^{\ 1}\boldsymbol{X}^{*\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{Y}^{V} \\ &= (\boldsymbol{\Sigma} \otimes \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}^{\mathrm{T}})\boldsymbol{Y}^{V} \\ &= [\boldsymbol{I} \otimes (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}]\ \boldsymbol{Y}^{V}.\end{aligned}$$

In particolare, la stima OLS si ottiene nel caso in cui $\boldsymbol{\Sigma} = \boldsymbol{I}$; in questa situazione lo stimatore OLS e lo stimatore GLS coincidono.

È possibile valutare il comportamento dello stimatore GLS asintoticamente, ossia nell'ipotesi in cui n, che rappresenta la numerosità campionaria, tenda ad ∞.

Sia $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots$una sequenza di variabili indipendenti fissate e sia $u_1, u_2, \ldots$ una sequenza di termini d'errore casuali. Siano $\boldsymbol{X} = \boldsymbol{X}_n = (\boldsymbol{x}_1^{\mathrm{T}}, \boldsymbol{x}_2^{\mathrm{T}}, \ldots, \boldsymbol{x}_n^{\mathrm{T}})$, $\boldsymbol{\Omega} = \boldsymbol{\Omega}_n$ e $\boldsymbol{B} = \boldsymbol{B}_n$. Si supponga inoltre che

$$\lim_{n\to\infty} (\boldsymbol{X}_n^{\mathrm{T}}\boldsymbol{\Omega}_n^{-1}\boldsymbol{X}_n)^{-1} = \mathbf{0}.$$

Allora il valore atteso della stima di $\boldsymbol{B}$ sarà $\boldsymbol{B}$ per ogni valore di n considerato. Inoltre la varianza dello stimatore tenderà a 0 quando $n \to \infty$.

Ciò implica che lo stimatore di $\boldsymbol{B}$ è consistente.

Capitolo 2

MANOVA: Analisi della varianza multivariata

Nel caso della regressione multipla, una delle procedure più utilizzate per valutare l'uguaglianza delle medie in gruppi differenti e per valutare la significatività dell'effetto di un'esplicativa sulla variabile risposta è l'analisi della varianza (**ANOVA**).

Anche nel caso multivariato si può utilizzare una procedura simile, la MANOVA (*multivariate analysis of variance*). Ad esempio, nel caso in cui si abbiano due variabili dipendenti e si voglia testare se entrambe sono influenzate da una certa esplicativa è possibile calcolare una statistica *F*, non più univariata. Ciò che si ottiene è una statistica *F* multivariata (lambda di Wilks) basata sul confronto tra la matrice di varianza e covarianza degli errori e la matrice di varianza e covarianza delle esplicative. Infatti, la formula della statistica *F* non è basata solo sulla somma dei quadrati tra ed entro i gruppi, ma anche dai prodotti incrociati (*crossproducts*).

Il calcolo della statistica *F* permette di rispondere alla domanda "Il modello è significativo per ogni variabile dipendente?". La covarianza deve essere inclusa perché le due variabili dipendenti potrebbero essere correlate ed è necessario testare se tale correlazione è statisticamente significativa. Se le due variabili risposta sono fortemente correlate si ottengono informazioni ridondanti e ciò è espresso dalla covarianza.

L'ipotesi nulla che si andrà a testare riguarda l'assenza di differenze in media in ogni variabile dipendente per gruppi differenti formati a seconda dei livelli e delle categorie delle esplicative.

Oltre a valutare se le risposte sono influenzate contemporaneamente dalla esplicative, può essere interessante valutare se le variabili dipendenti prese singolarmente vengono influenzate dalle indipendenti considerate. Infatti, oltre a calcolare la statistica *F* multivariata che considera gli effetti principali e le interazioni, è possibile esaminare i test *F* univariati per ogni singola variabile ed interpretarli singolarmente.

I test considerati nel caso univariato si possono estendere al caso multivariato. Se in precedenza si aveva il rapporto tra devianza di regressione e devianza residua, nel caso multivariato si moltiplica l'inverso della matrice $\boldsymbol{E}$ per la matrice $\boldsymbol{H}$, dove la matrice $\boldsymbol{E}$, matrice di devianza residua, si ottiene come

$$\boldsymbol{E} = \boldsymbol{u_l}^{\mathrm{T}}\,\boldsymbol{u_l} = (\boldsymbol{n_1 - 1})\,\boldsymbol{S_1}^2 + (\boldsymbol{n_2 - 1})\,\boldsymbol{S_2}^2 + \ldots + (\boldsymbol{n_p - 1})\,\boldsymbol{S_p}^2$$

con $\boldsymbol{S_1}^2$, $\boldsymbol{S_2}^2$, $\boldsymbol{S_p}^2$ matrici di varianza e covarianza all'interno dei rispettivi gruppi, $\boldsymbol{n_1}$, $\boldsymbol{n_2}$,…, $\boldsymbol{n_p}$ numerosità dei gruppi, e $\boldsymbol{H}$, matrice di devianza tra i gruppi, data da

$$\boldsymbol{H} = \Sigma_j\ \boldsymbol{n_j}\boldsymbol{\beta_{il0}}^{\mathrm{T}}\boldsymbol{B_{jl}}^{\mathrm{T}}\boldsymbol{B_{jl}}\boldsymbol{\beta_{il0}} = [\boldsymbol{\beta_l} - \boldsymbol{B_l}]^{\mathrm{T}}[\boldsymbol{\beta_l} - \boldsymbol{B_l}]\,[\boldsymbol{y_l} - \boldsymbol{m_l}]^{\mathrm{T}}[\boldsymbol{y_l} - \boldsymbol{m_l}].$$

La somma delle due matrici, in genere indicata con $\boldsymbol{T} = \boldsymbol{E} + \boldsymbol{H}$, è la matrice di devianza totale.

I test, che di seguito verranno descritti, permettono di verificare la presenza di una relazione di dipendenza lineare della matrice delle risposte $\boldsymbol{Y}$ dalla matrice delle esplicative $\boldsymbol{X}$.

2.1 Test multivariati

Esistono alcuni test multivariati che permettono di focalizzare l'attenzione sulle variabili indipendenti e sulle loro interazioni. In particolare ci sono quattro tipi di test utilizzati per questo scopo.

2.1.1 T^2 di Hotelling

Il test T^2 di Hotelling è definito come

$$T^2 = \mathrm{tr}(\boldsymbol{E}^{-1}\,\boldsymbol{H}) = \Sigma\lambda_{\mathrm{ij}},$$

dove $\boldsymbol{\lambda}$ indica l'autovalore di una matrice $\boldsymbol{A}$ ($k \times k$). È il test più utilizzato e tradizionale basato su due gruppi formati dalle variabili indipendenti. Ad esso è correlata la statistica che prende i nome di *Hotelling's Trace*. Per trasformare il coefficiente *Trace* nel coefficiente T^2 si moltiplica il *Trace* per (n - g), dove n è la numerosità complessiva e g il numero dei gruppi. Il T^2 avrà lo stesso valore della statistica F, lo stesso numero di gradi di libertà e lo stesso livello di significatività del coefficiente *Trace*.

2.1.2 Lambda di Wilks

La lambda di Wilks si può scrivere come

$$\Lambda = |\boldsymbol{E}|/|\boldsymbol{H} + \boldsymbol{E}| = \Pi(1/(1+\lambda_{ij})) \qquad \textbf{(2.1)}$$

oppure, analogamente, come

$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{H} + \boldsymbol{E}|} = \frac{\sum_{l=1}^{p}\sum_{j=1}^{q}(\boldsymbol{\beta}_{jl}\boldsymbol{B}_{jl} - \boldsymbol{\mu})(\boldsymbol{\beta}_{jl}\boldsymbol{B}_{jl} - \boldsymbol{\mu})^{\mathrm{T}}}{\sum_{l=1}^{p}\sum_{j=1}^{q}(\boldsymbol{y}_{lj} - \boldsymbol{\mu})(\boldsymbol{y}_{lj} - \boldsymbol{\mu})^{\mathrm{T}}},$$

per il quale ci sono più di due gruppi di variabili indipendenti. È una misura delle differenze tra i gruppi del vettore di medie sulle variabili indipendenti. Più piccolo è il valore della statistica Λ, più grandi saranno le differenze esistenti. Per calcolare la significatività di Λ viene spesso utilizzata una sua trasformazione, la V di Bartlett. Usate in combinazione, le due statistiche, la Λ di Wilks e la V di Bartlett, consentono il calcolo di un test per le differenze in media nella MANOVA, nel caso in cui ci siano più variabili dipendenti e più di due gruppi formati dalle esplicative. I test considerati precedentemente (t-test, Hotelling's T e test F) sono tutti casi particolari della Λ di Wilks. In particolare, in casi particolari Λ segue una distribuzione F, ossia quando

- $q = 2$ e $l \geq 2$, allora Λ si distribuirà come una F con l - 2 e 2 × (n – q - 1) gradi di libertà;
- se $q \geq 1$ e $l = 2$, allora Λ si distribuirà come una F con p e n – q - 1 gradi di libertà.

Più in generale sotto l'ipotesi nulla H_0 e con un elevata numerosità campionaria, la statistica test si distribuisce come un χ^2 con $p \times (l - 1)$ gradi di libertà.

2.1.3 Pillai-Bartlett *trace*

Il test Pillai-Bartlett *trace* è definito come

$$\mathbf{PB} = \mathrm{Tr}(\boldsymbol{E}(\boldsymbol{E} + \boldsymbol{H})^{-1}),$$

che si incontra anche nella MDA (*multiple discriminant analysis*), ossia in quella parte della MANOVA, simile all'Analisi delle Componenti Principali, che ha come obiettivo la classificazione. Tale test è il più robusto di tutti, ossia il più affidabile anche nel caso in cui certe ipotesi siano violate.

2.1.4 GCR

Il test GCR (*Roy's greatest characteristic root*) è definito come

$$\mathbf{GCR} = \max \lambda_{ij}. \qquad \mathbf{(2.2)}$$

Esso è simile al Pillai-Bartlett, ma è bastato sulla prima e più importante radice. Questo test è meno robusto dei precedenti a causa della violazione delle assunzioni di normalità multivariata.

2.1.5 Alcune conclusioni

I test presentati in questo paragrafo, permettono di riportare il modello multivariato al caso univariato in termini di verifica d'ipotesi. In particolare, essi permettono di verificare se le variabili dipendenti $\boldsymbol{Y}$ dipendono simultaneamente dalle esplicative $\boldsymbol{X}$. Infatti, l'ipotesi alternativa è quella dell'esistenza di una relazione lineare significativa tra $\boldsymbol{Y}$ e $\boldsymbol{X}$. Si dimostra che $\boldsymbol{H}$ è stima distorta della devianza fra i gruppi, mentre $\boldsymbol{E}$ è stima corretta delle devianze fra i gruppi residui.
Nel caso più semplice in cui abbiamo solo due gruppi da confrontare, estendiamo l'ipotesi nulla del modello univariato

$$H_0: \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

al caso multivariato, in cui vogliamo verificare che i vettori di medie siano uguali, ossia

$$H_0 : \begin{pmatrix} \mu_{11} \\ \mu_{21} \\ \vdots \\ \mu_{p1} \end{pmatrix} = \begin{pmatrix} \mu_{12} \\ \mu_{22} \\ \vdots \\ \mu_{p2} \end{pmatrix}.$$

L'estensione al caso multivariato del test *t* è data da

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right)^{\mathrm{T}} \boldsymbol{S}^{-1} \left(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2\right),$$

con

$$T^2 \sim F_{n_1+n_2-p-1,(n_1+n_2-2)p},$$

dove $\boldsymbol{S}$ è la matrice di covarianza "unita" (*pooled*) entro i gruppi e n_1 e n_2 sono le numerosità dei due gruppi.

Nel caso più generico in cui abbiamo *p* gruppi di osservazioni possiamo testare l'ipotesi nulla H_0: $\tau_1 = \tau_2 = \ldots = \tau_p = 0$, dove τ_i indica il vettore di medie dell'*i*-esimo gruppo. In questa situazione si utilizza il Λ di Wilks (vedi formula **(2.1)**).

Si possono utilizzare due tipi di approcci per valutare la significatività dei test effettuati e per confrontare la devianza dei trattamenti e la devianza degli errori o residua. Il primo metodo è l'approccio della verosimiglianza; il secondo è quello delle combinazioni lineari.

Il metodo della verosimiglianza si serve del Λ di Wilks e rifiuta l'ipotesi nulla H_0 se il valore della statistica test è grande.

Il secondo approccio, invece, si basa sul test GCR di Roy (vedi formula **(2.2)**). Tale statistica tenta di trovare combinazioni lineari delle osservazioni che massimizzino il rapporto tra varianza spiegata e varianza residua, massimizzando la varianza tra i gruppi e contemporaneamente minimizzando quella entro i gruppi.

Così come l'ANOVA testa le differenze in media della variabile dipendente per varie categorie delle variabili indipendenti, la MANOVA testa la differenze nei vettori di medie delle risposte per le diverse categorie delle esplicative.
Inoltre, così come nell'ANOVA la varianza spiegata è la stima distorta sotto l'ipotesi alternativa H_1 della quantità $(p - 1)\,\boldsymbol{\sigma}^2$, essendo corretta sotto l'ipotesi nulla H_0, ora nel caso della MANOVA la matrice $\boldsymbol{B}$ è la stima corretta della matrice di devianza tra i gruppi solo sotto l'ipotesi nulla H_0.

2.2 Test "post-hoc"

Il passo successivo al calcolo dei test nella procedura di analisi della varianza multivariata è rappresentato dai test a posteriori (*post-hoc*) che si eseguono nel caso in cui il test *F* mostri che il vettore di medie non sia lo stesso nei diversi gruppi formati sulla base della categorie delle esplicative. Per effettuare il test *post-hoc* si calcolano le statistiche *F* univariate che sono usate per determinare quali medie dei gruppi differiscono significativamente dalle altre. Allo stesso tempo si eseguono test (confronti multipli) su coppie di gruppi per verificarne somiglianze e differenze.
I test utilizzati per i confronti a posteriori sono:

- il test di **Bonferroni**, utilizzato se ci sono pochi gruppi, poiché produce le differenze in media delle variabili dipendenti tra coppie di gruppi (ad esempio differenze nei punteggi ottenuti nei test per ogni coppia di gruppi che differiscono per scuola di provenienza);
- il test di **Tukey**, preferibile se i gruppi sono numerosi;
- altri test utilizzati quando le assunzioni di omogeneità della varianza sono violate.

Queste procedure rappresentano le fasi conclusive della MANOVA e permettono di capire meglio le differenze rilevate tra i gruppi.

2.3 Assunzioni alla base dell'analisi multipla della varianza

La procedura di analisi multipla della varianza si basa su una serie di assunzioni che, se violate, rendono non affidabili gli indicatori calcolati.
Tali assunzioni sono:

- l'indipendenza delle osservazioni, poiché nell'ipotesi in cui questa condizione non valga, allora la MANOVA non è robusta;
- dimensioni uguali per tutti i gruppi;
- somme dei quadrati appropriate;
- dimensioni adeguate del campione;
- distribuzione casuale dei residui;
- omoschedasticità (omogeneità delle varianze e delle covarianze): la varianza di ogni variabile dipendente continua deve essere simile, come testato dal test di Levene (si veda Levene, 1960 [4]). Tale test, simile a quello di Bartlett, è utilizzato su k campioni per verificare se siano omogenei in varianza, ossia se hanno la stessa varianza. Viene utilizzato quindi per testare ipotesi del tipo

  $$H_0: \boldsymbol{\sigma}_1 = \boldsymbol{\sigma}_2 = \ldots = \boldsymbol{\sigma}_p \qquad vs \qquad H_1: \boldsymbol{\sigma}_i \neq \boldsymbol{\sigma}_j \quad \text{per almeno un } i \neq j.$$

  Se si hanno n osservazioni per una variabile $\boldsymbol{Y}$, suddivisa in k gruppi di numerosità n_i, la statistica di Levene prende la seguente forma

  $$\boldsymbol{W} = \frac{(n-k)\sum_{i=1}^{k} n_i(\overline{\boldsymbol{Z}}_i - \overline{\boldsymbol{Z}})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\boldsymbol{Z}_{ij} - \overline{\boldsymbol{Z}}_i)^2},$$

  con $\boldsymbol{Z}_{ij} = |\boldsymbol{Y}_{ij} - \overline{\boldsymbol{Y}}_i|$, dove $\overline{\boldsymbol{Y}}_i$ è la media dell'i-esimo gruppo.

  Per ognuno dei k gruppi formati dalle variabili indipendenti, la covarianza tra le coppie di variabili dipendenti deve essere la stessa. Per verificare tale assunzione si può calcolare il test M di Box (si veda Anderson, 1958 [1]; Box, 1949 [2]; Seber, 1984 [8]) che verifica l'assunzione di omoschedasticità utilizzando un test F. Tale test è basato sul rapporto di verosimiglianza e assume la forma

  $$\boldsymbol{M} = (n-g)\log|\boldsymbol{S}| - \sum_{i=1}^{g}(n_i - 1)\log|\boldsymbol{S}_i|,$$

dove $\boldsymbol{S}$ è la matrice di varianza "*pooled*", n è la numerosità complessiva, n_i è la numerosità dell'i-esimo gruppo, g è il numero di gruppi e $\boldsymbol{S_i}$ è la matrice di varianza e covarianza dell'i-esimo gruppo.

Se il *p-value* associato al test è inferiore a 0.05 allora le covarianze sono significativamente diverse;

- omogeneità della regressione: i coefficienti delle covariate sono gli stessi per ogni gruppo formato da variabili categoriali e misurato sulle variabili dipendenti;
- distribuzione normale multivariata: si assume la normalità multivariata se ogni variabile segue la distribuzione normale. La MANOVA è robusta anche in caso di violazione di questa assunzione se la dimensione del campione non è piccola. Per verificare tale assunzione si possono utilizzare procedure diverse: è possibile rappresentare graficamente i dati e verificare la presenza di simmetria, calcolare la curtosi, valutare la presenza di valori anomali; si può utilizzare il test di Shapiro-Wilks che considera come ipotesi nulla la normalità dei dati; si può rappresentare il diagramma "quantile-quantile", che confronta i quantili della distribuzione normale con quelli dei dati. Queste procedure sono quelle utilizzate anche nel caso univariato perché si assume che se tutte le variabili sono normali allora complessivamente avremo una distribuzione normale multivariata.

  Altri controlli si possono effettuare calcolando i quadrati delle distanze dalla media delle osservazioni poiché se la popolazione è normale e n e $n - p$ sono maggiori di 30, allora tali distanze dovrebbero seguire la distribuzione di un χ^2_{n-p}.
- assenza di valori anomali;
- le covariate sono linearmente correlate o in una relazione nota con le variabili dipendenti, la forma della relazione tra covariate e variabili risposta deve essere nota. Spesso le covariate sono trasformate per stabilire una relazione lineare.

Per ulteriori chiarimenti sulla MANOVA e sui test ad essa correlati si rimanda a Miles, 2003 [5]; Vittadini, 1999 [9]; Schatz, 2006 [7].

2.4 Utilizzi della MANOVA

La MANOVA può avere molteplici utilizzi. La procedura può essere utilizzata per confrontare gruppi formati da variabili indipendenti di tipo categoriale rispetto alle differenze nei gruppi su un insieme di variabili risposta continue; per usare l'assenza di differenze per un gruppo di variabili dipendenti come un criterio per rendere più piccolo e facilmente gestibile il numero di variabili; per identificare le variabili indipendenti che differenziano maggiormente le risposte.
Nell'ipotesi in cui non ci sia correlazione tra le variabili risposta, è accettabile l'utilizzo di una serie di test basati sull'ANOVA univariata piuttosto che l'utilizzo della MANOVA. Nella maggior parte dei casi, però, le variabili sono tra loro correlate ed in questo caso è preferibile l'utilizzo della MANOVA, poiché tale procedura è sensibile non solo alle differenze in media, ma anche al segno e alla forza della correlazione esistente tra le variabili dipendenti.
Nella procedura della MANOVA non c'è un limite al numero di variabili dipendenti che possono essere inserite, ma più cresce il numero di risposte e più diventa difficile l'interpretazione dei risultati ottenuti. Inoltre, si riduce la potenza dei test poiché cresce la probabilità di commettere errori del II Tipo.

2.5 Conclusioni

La procedura di analisi della varianza multivariata permette di estendere al caso di regressione multivariata le conclusioni a cui si giungeva nel caso univariato e di applicare ai dati una serie di test per eseguire verifiche d'ipotesi.
La MANOVA può essere eseguita da una serie di pacchetti statistici. Ad esempio in SAS si usa il comando `Mtest`; in SPSS è presente l'opzione MANOVA.
Anche in **R** sono stati sviluppati alcuni pacchetti che si utilizzano per la regressione multivariata e che danno in *output* i test citati con i rispettivi *p-values*.

Tali pacchetti saranno presentati nel prossimo capitolo, assieme all'analisi di un insieme di dati reali.

Capitolo 3

Un caso studio

3.1 I dati

Il modello di analisi multivariata **(1)** può essere applicato in diversi campi nel caso in cui si voglia valutare l'effetto di alcune variabili esplicative ***X*** su più variabili risposta ***Y***.

In particolare l'applicazione che si prende in considerazione in questo capitolo riguarda dati di natura antropometrica raccolti su n = 3028 unità statistiche. Tali unità sono rappresentate da adulti ed anziani sardi.

Le analisi riportate di seguito sono state eseguite con il software **R**, (www.r-project.org) che consente di eseguire analisi grafiche e calcoli statistici grazie ad alcune funzioni contenute all'interno di sue librerie. Tramite i pacchetti di base e quelli disponibili gratuitamente in rete, **R** consente di stimare modelli lineari e modelli lineari generalizzati, di eseguire regressioni non lineari, di analizzare serie storiche e di effettuare test parametrici e non-parametrici. Il *software* consente anche di costruire funzioni esterne e semplici programmi, di effettuare regressioni robuste, multivariate e non parametriche, test statistici classici e controllo della qualità.

Nel dataset analizzato, l'interesse degli antropologi è particolarmente rivolto alla popolazione con età superiore ai 65 anni, per la quale si vuole valutare, separatamente per i due sessi, l'evoluzione che le variabili antropometriche hanno al crescere dell'età.

In particolare, si vuole valutare l'effetto su 23 variabili risposta di età, sesso e patologia, dove `patologia` è una variabile cha assume valore 0 se l'unità statistica è sana, 1 se è diabetica e 2 se soffre di Alzheimer.

La variabile `sesso` è codificata come un fattore a due livelli che assume valore 1 per gli uomini e 2 per le donne.

La variabile `eta` misura l'età in anni delle unità statistiche al momento della rilevazione.

Le 23 variabili risposta considerate sono:

- `peso`, che rappresenta il peso (*kg*);
- `statura`, che misura l'altezza degli individui (*cm*);
- `pm.vita`, `pm.fianchi` e `pm.polpacc`, che misurano, rispettivamente, il perimetro della vita, dei fianchi e del polpaccio (*cm*);
- `pm.Br.contr`, che misura il perimetro del braccio contratto (*cm*);
- `dm.biacrom`, `dm.omero` e `dm.bicres` che misurano, rispettivamente, diametro biacromiale, dell'omero e bicrestiliaco (*cm*);
- `pl.Bicip`, `pl.tric`, `pl.sottosc` e `pl.soprail`, che indicano plica al bicipite, al tricipite, sottoscapolare e soprailiaca e sono misurate in *cm*;
- `lung.cefal`, che misura la lunghezza cefalica (*cm*);
- `R/H`, indicatore della resistenza normalizzata per la statura;
- `Xc/H`, indicatore della reattanza normalizzata per la statura, dove la reattanza rappresenta la forza che un condensatore oppone al passaggio di una corrente elettrica;
- `FFMLohman`, che misura la massa priva di grassi;
- `FMLohman`, che indica la massa grassa;
- `BMI` (*Body Mass Index*), l'indice di massa corporea che si calcola come peso/statura^2;
- `WHR`, rapporto tra il perimetro della vita e quello dei fianchi;
- `STSR`, rapporto tra plica sottoscapolare e plica del tricipite;
- `AMA`, area muscolare del braccio, che si ottiene da una rielaborazione tra le misure del perimetro del braccio e la plica del tricipite;

- `Angolo di fase`, che si ottiene dall'arcotangente del rapporto tra `Xc` e `R`, trasformata in gradi.

Prima di effettuare qualunque tipo di analisi è necessario fattorizzare le due variabili `sesso` e `patologia`:

```
dati$patologia<-factor(dati$patologia)
dati$sesso<-factor(dati$sesso)
contrasts(sesso)
  2
1 0
2 1
contrasts(patologia)
  1 2
0 0 0
1 1 0
2 0 1
```

Per effettuare le analisi di interesse è opportuno selezionare dal dataset le unità statistiche con età superiore a 65 anni.

Il nuovo dataset contiene, quindi, 743 osservazioni.

Sui dati considerati si possono calcolare le statistiche descrittive per le variabili considerate. Delle unità statistiche del dataset, il 72,7% non presenta patologie, il 19,9% soffre di diabete e il 7,5% di Alzheimer. Inoltre, gli individui di sesso maschile sono il 47% del totale, mentre il 53% sono di sesso femminile. Nel seguito sono riportate alcune statistiche descrittive di base per le variabili risposta.

	eta	peso	statura	pm.vita	pm.fianchi	pm.polpac
Minimo	65.04	30.00	124.6	50.00	63.1	21.10
1st Qu	70.98	56.75	146.7	87.00	96.0	31.30
Mediana	76.82	67.00	153.0	95.50	102.2	34.00
Media	77.93	66.73	153.4	95.22	102.9	33.79
3rd Qu.	83.93	76.10	160.0	102.20	108.6	36.40
Std Dev	8.21	14.45	9.58	150.00	148.0	55.00
Massimo	101.38	124.00	178.4	12.24	10.30	3.71
NA's				2	5	22

	pm.br.contr	dm.biacrom	dm.omero	dm.bicres	pl.bicip
Minimo	17.80	20.40	4.60	20.10	2.00
1st Qu	26.50	31.00	6.10	28.90	8.00
Mediana	29.30	35.20	6.60	30.40	12.00
Media	29.32	34.20	6.65	30.23	13.97
3rd Qu	32.00	37.83	7.20	32.10	18.00
Std Dev	4.02	3.69	0.69	2.36	7.99
Massimo	42.30	45.00	9.00	39.00	52.00
NA's	21	275	22	275	1

	pl.tric	pl.sottosc	pl.soprail	lung.cef	rh	xch
Minimo	2.60	2.00	2.00	16.70	188.4	14.43
1st Qu	14.00	16.00	18.00	18.50	281.9	29.14
Mediana	21.00	21.50	25.00	19.00	323.8	34.42
Media	22.04	22.87	26.06	19.01	331.3	35.63
3rd Qu	28.75	28.38	32.00	19.50	372.2	40.86
Std Dev	10.16	9.52	10.52	0.47	62.06	8.42
Massimo	70.00	62.00	70.00	21.00	571.4	75.38
NA's	2	5	21	477	90	90

	ffmlohman	fmlohman	bmi	whr	stsr	ama	angolo_fase
Minimo	24.99	4.612	16.20	0.57	0.27	7.13	3.17
1st Qu	36.75	17.95	24.72	0.88	0.81	25.72	5.34
Mediana	41.77	23.251	27.94	0.93	1.05	33.42	6.13
Media	42.68	24.018	28.20	0.93	1.16	34.63	6.19
3rd Qu	48.20	29.13	31.27	0.98	1.39	42.75	7.01
Std Dev	7.77	8.06	4.84	0.08	0.52	12.22	1.19
Massimo	70.31	56.31	45.18	1.29	4.18	84.64	12.42
NA's	90	90	4	6	10	26	90

Alcune delle variabili del dataset presentano un numero elevato di valori mancanti (`NA's`). In questa situazione si può decidere di sostituire i valori mancanti con la mediana o con la media della distribuzione, a seconda che ci sia o meno simmetria, oppure di eliminare la variabile. Nel caso di variabili come `lung.cefal`, `dm.biacrom` e `dm.bicres`, sembra essere più opportuno escludere le variabili dall'analisi poiché la prima ha il 67% dei dati mancanti, mentre le ultime due il 37%.

Per quanto riguarda le altre variabili, si è deciso di sostituire i valori mancanti con la mediana della distribuzione.

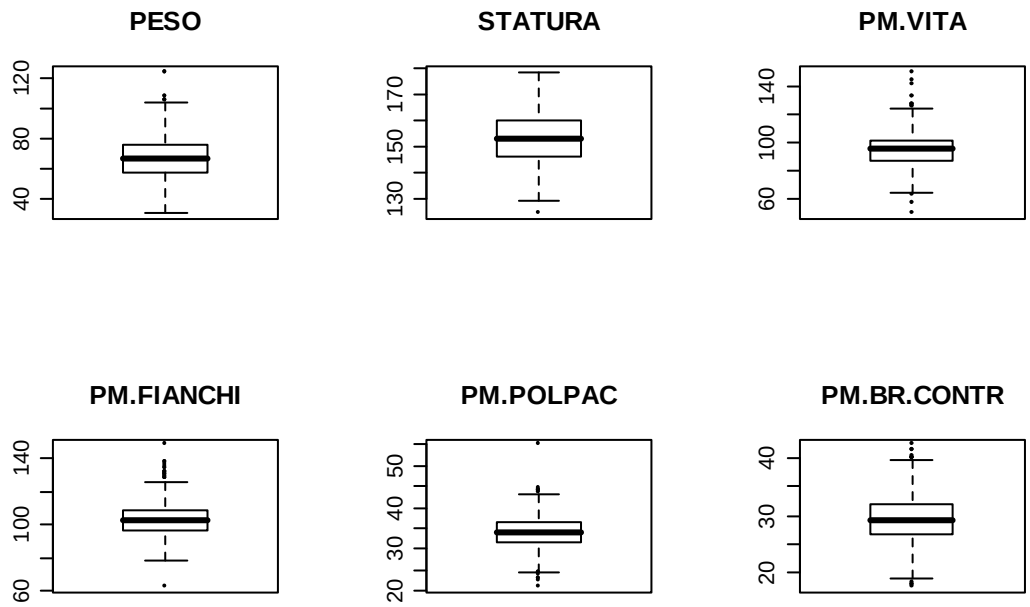
Il dataset studiato contiene pertanto 20 variabili risposta (in seguito all'eliminazione di `dm.bicres`, `dm.biacrom` e `lung.cef`) che devono essere valutate in base a sesso, età e patologia. Una prima analisi grafica permette di valutare se esiste una dipendenza effettiva delle variabili dipendenti dalle esplicative. Inoltre, l'analisi grafica consente di valutare se le variabili siano o meno simmetriche e se si possono ritenere normali. Utilizzando alcune librerie di **R** è possibile valutare, anche visivamente, la forza delle correlazioni esistenti tra le variabili risposta.

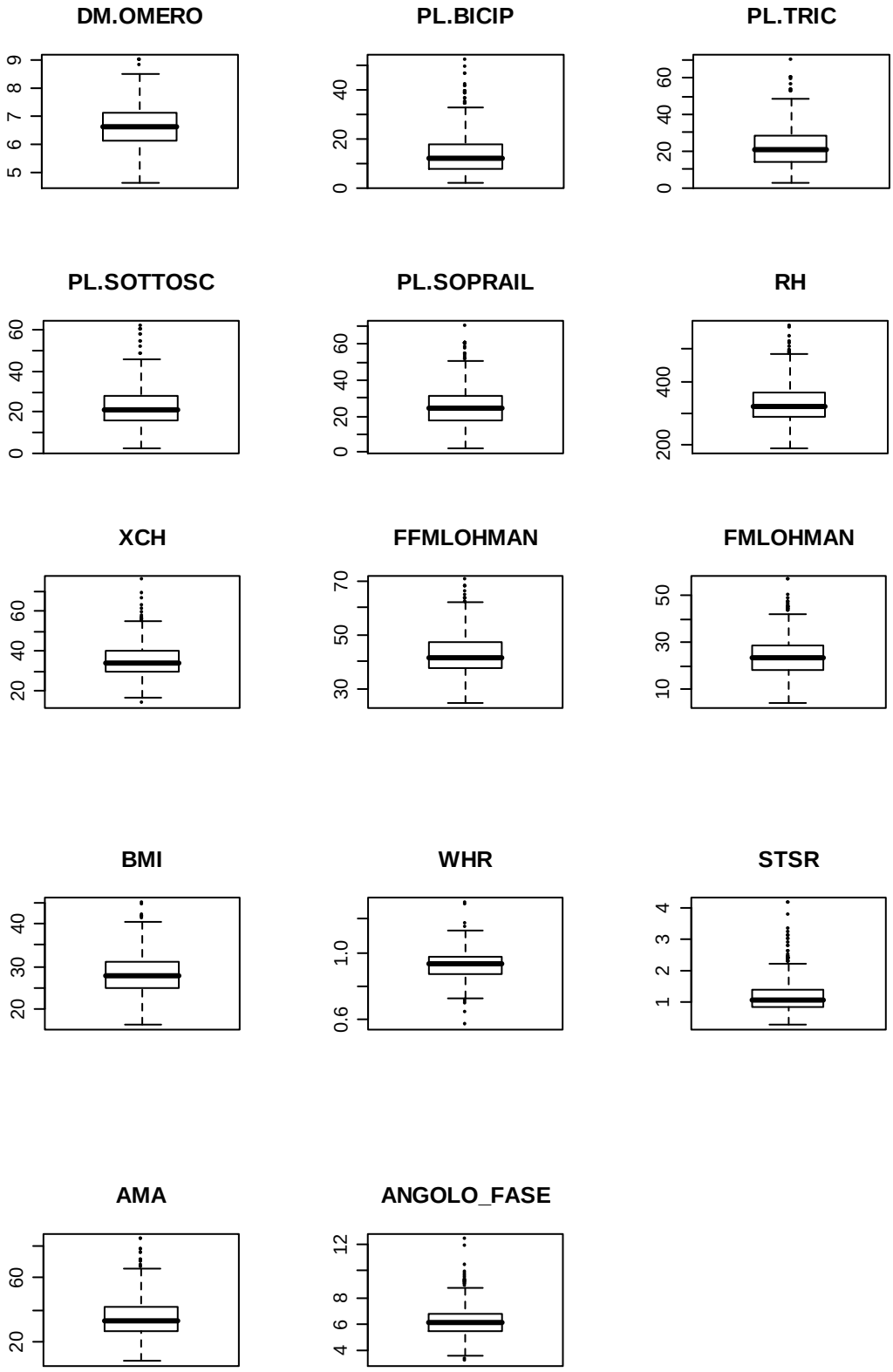
3.2 Analisi grafiche univariate

Per iniziare l'analisi sulla distribuzione delle variabili si possono considerare i boxplot, che permettono di visualizzare minimo, massimo, primo e terzo quartile e mediana della distribuzione.

Il comando di **R** per costruire boxplot è il seguente:

```
boxplot(x, data, subset,main="…")
```



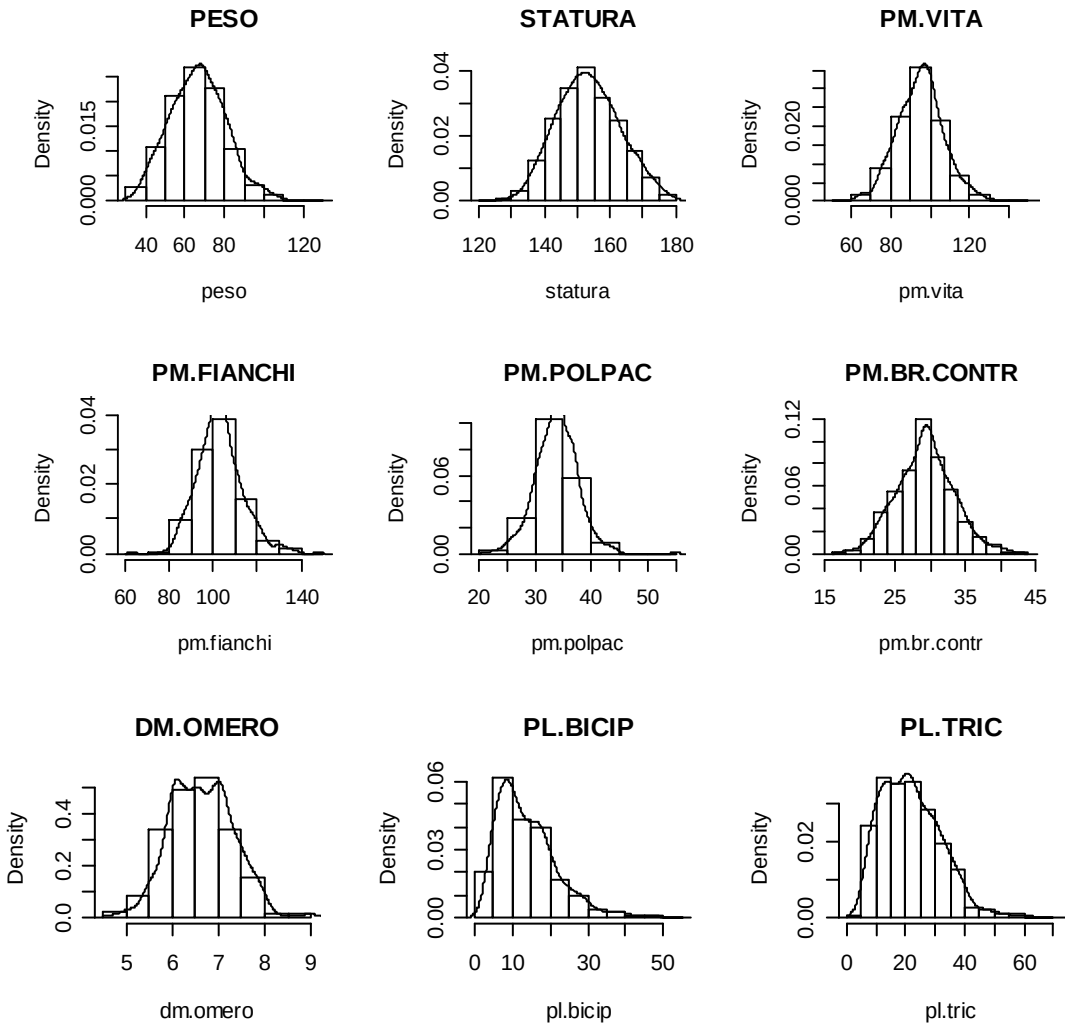

DM.OMERO
5 6 7 8 9
PL.BICIP
0 20 40
PL.TRIC
0 20 40 60
PL.SOTTOSC
0 20 40 60
PL.SOPRAIL
0 20 40 60
RH
200 400
XCH
20 40 60
FFMLOHMAN
30 50 70
FMLOHMAN
10 30 50
BMI
20 30 40
WHR
0.6 1.0
STSR
1 2 3 4
AMA
20 60
ANGOLO_FASE
4 6 8 12

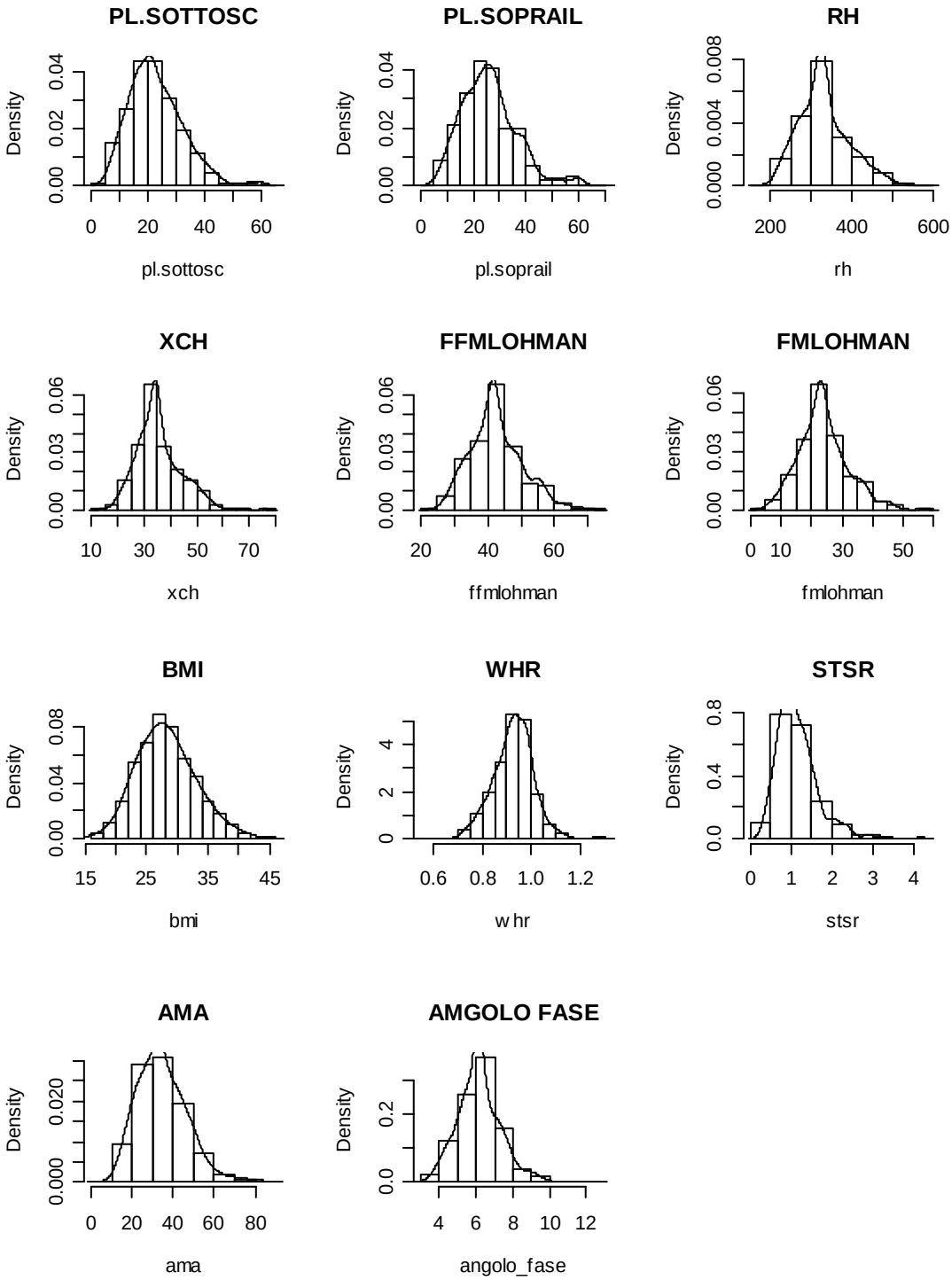
Dall'analisi dei boxplot risulta che alcune delle variabili del dataset sono asimmetriche, come `pl.bicip` e `stsr`, e questo è poco compatibile con l'ipotesi di normalità, richiesta per l'applicazione di una regressione multivariata.

Di seguito sono riportati gli istogrammi delle variabili considerate.

In R il comando per ottenere l'istogramma di una variabile è `hist(x, freq=F, main=…)`. L'opzione `freq=F` permette di ottenere, sull'asse delle ordinate, le frequenze relative anziché quelle assolute. `Main="…"` permette di indicare ad R il nome da assegnare al grafico considerato.

Con l'opzione `lines(density(x))`, R aggiunge all'istogramma delle frequenze, una stima della densità non parametrica ottenuta col metodo del nucleo (*Kernel*).

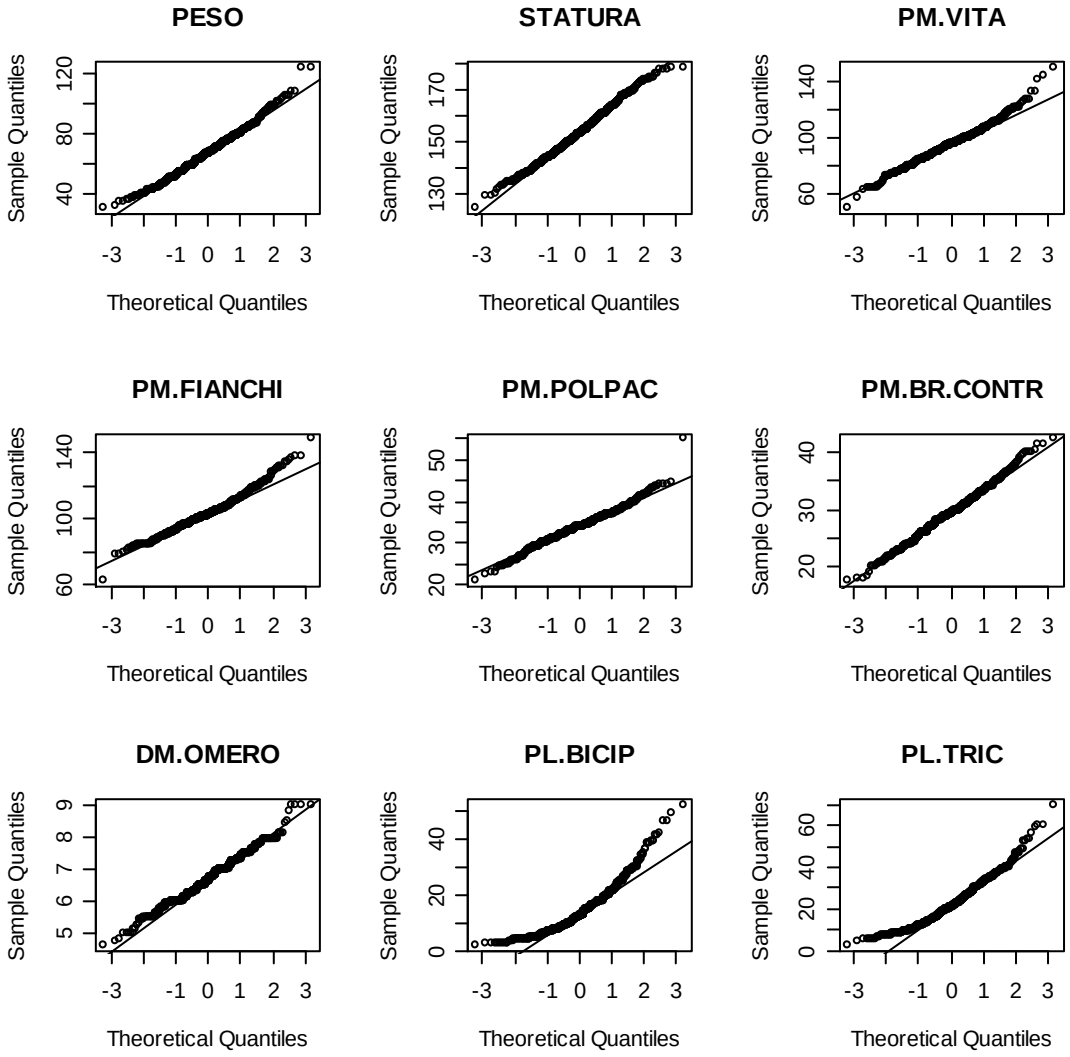


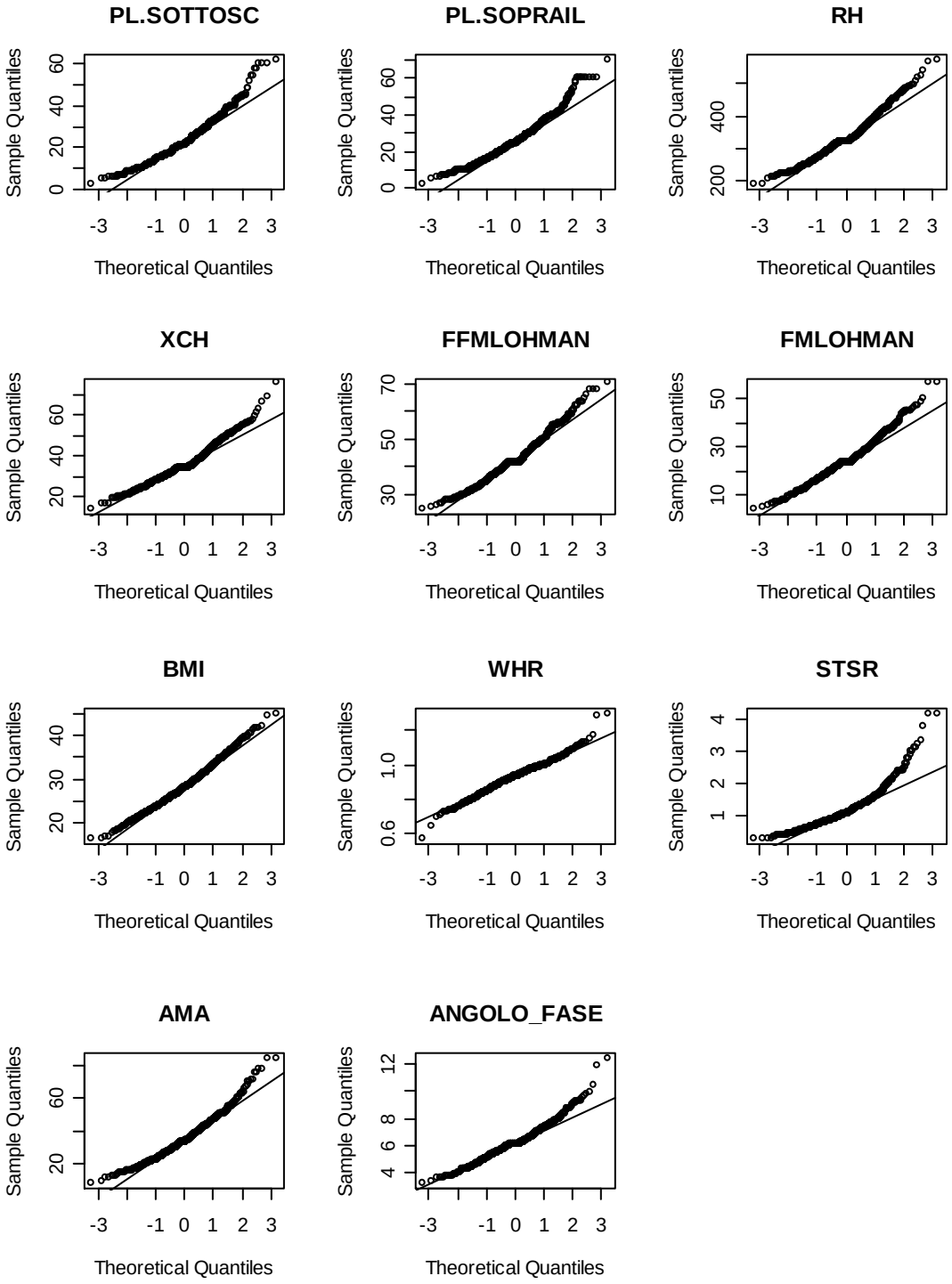


Dall'analisi degli istogrammi è possibile notare che alcune delle variabili hanno una distribuzione asimmetrica e quindi non compatibile con l'ipotesi di normalità delle

variabili risposta. In particolare le variabili `pl.bicip`, `pl.tric` e `stsr` mostrano un'asimmetria destra.

Un'ulteriore conferma del fatto che alcune delle variabili del dataset non sono normali si può ottenere utilizzando il comando `qqnorm(x)` che permette di confrontare i quantili della distribuzione della variabile di interesse con quelli teorici di una normale standard. È inoltre possibile aggiungere la retta che passa per il primo e terzo quartile della normale, per render più chiaro il confronto. Il comando per tale opzione è `qqline(x)`.



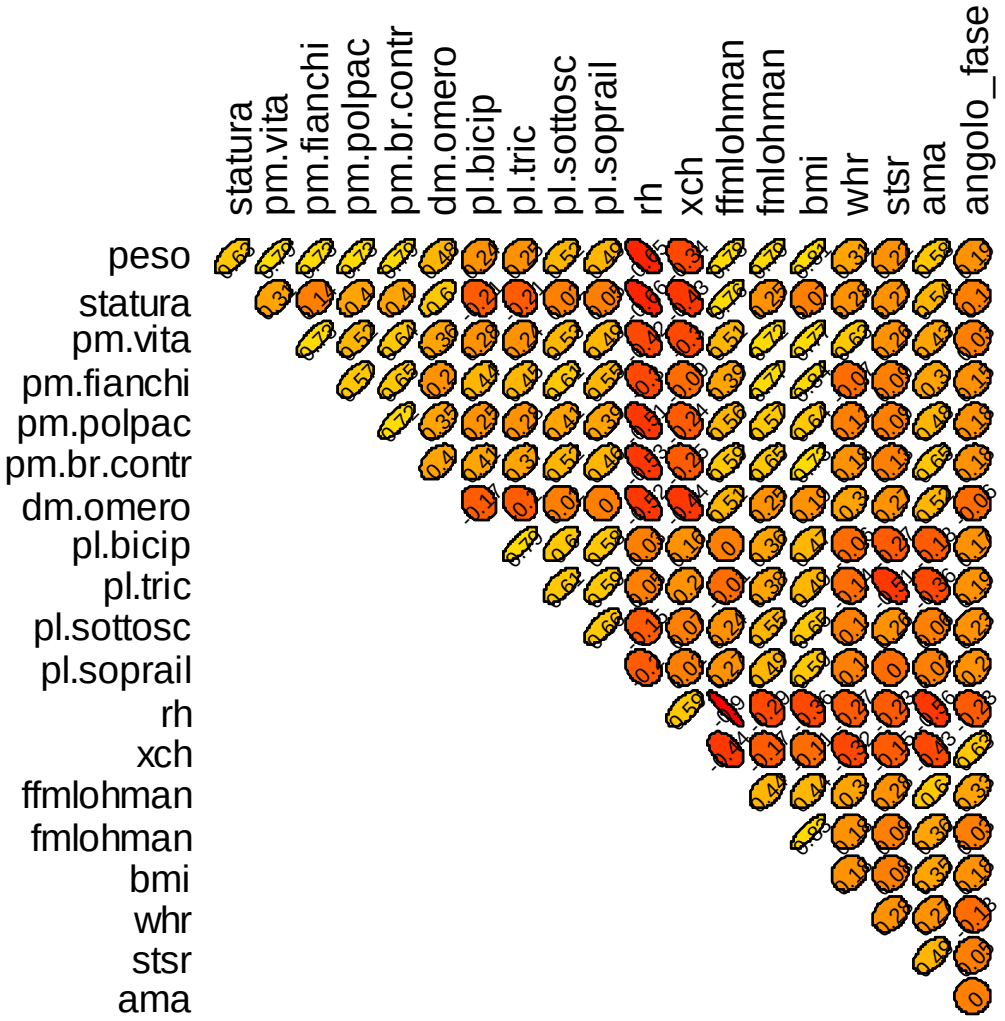


Il discostamento dei dati dalla retta fornisce la conferma definitiva che le variabili prese in esame non seguono una distribuzione normale.

In particolare risultano poco compatibili con l'ipotesi di normalità le variabili `pm.vita`, `pl.bicip`, `pl.tric`, `pl.sottosc`, `pl.soprail`, `rh`, `stsr` ed `ama`.

3.3 Analisi bivariate

R permette, col comando `cor(data.frame)`, di calcolare le correlazioni tra tutte le variabili contenute nel dataset. Esiste inoltre la libreria "*ellipse*" che consente, tramite il comando `plotcorr(cor(…))`, di rappresentare graficamente la matrice di correlazione: dove la correlazione tra le variabili è più forte **R** rappresenta forme ellittiche; più debole è la correlazione, più la rappresentazione si avvicina a quella di una circonferenza.



Alcune delle variabili risposta sono fortemente correlate, anche perché alcune sono state calcolate come trasformazioni lineari di altre. Ad esempio il *Body Mass Index* è ottenuto

come rapporto tra peso e altezza al quadrato. Infatti la correlazione tra `bmi` e `peso` è piuttosto forte. Altre variabili del dataset sembrano incorrelate. Ad esempio l'angolo di fase mostra correlazione solo con `xch`, che rappresenta la reattanza normalizzata per la statura.

Si possono considerare anche i boxplot per le variabili risposta condizionati al sesso e alla patologia.

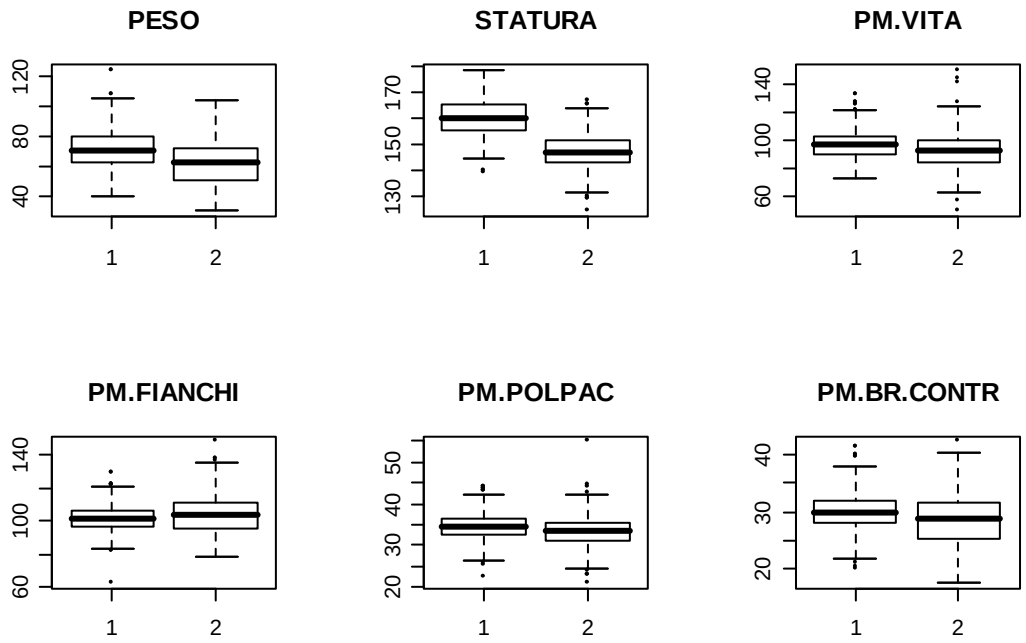
I diagrammi a scatola con baffi permettono di valutare la distribuzione di una variabile e, se condizionati ad una qualche variabile, di osservare se ci sono differenze tra le distribuzioni per i diversi valori della variabile condizionante.

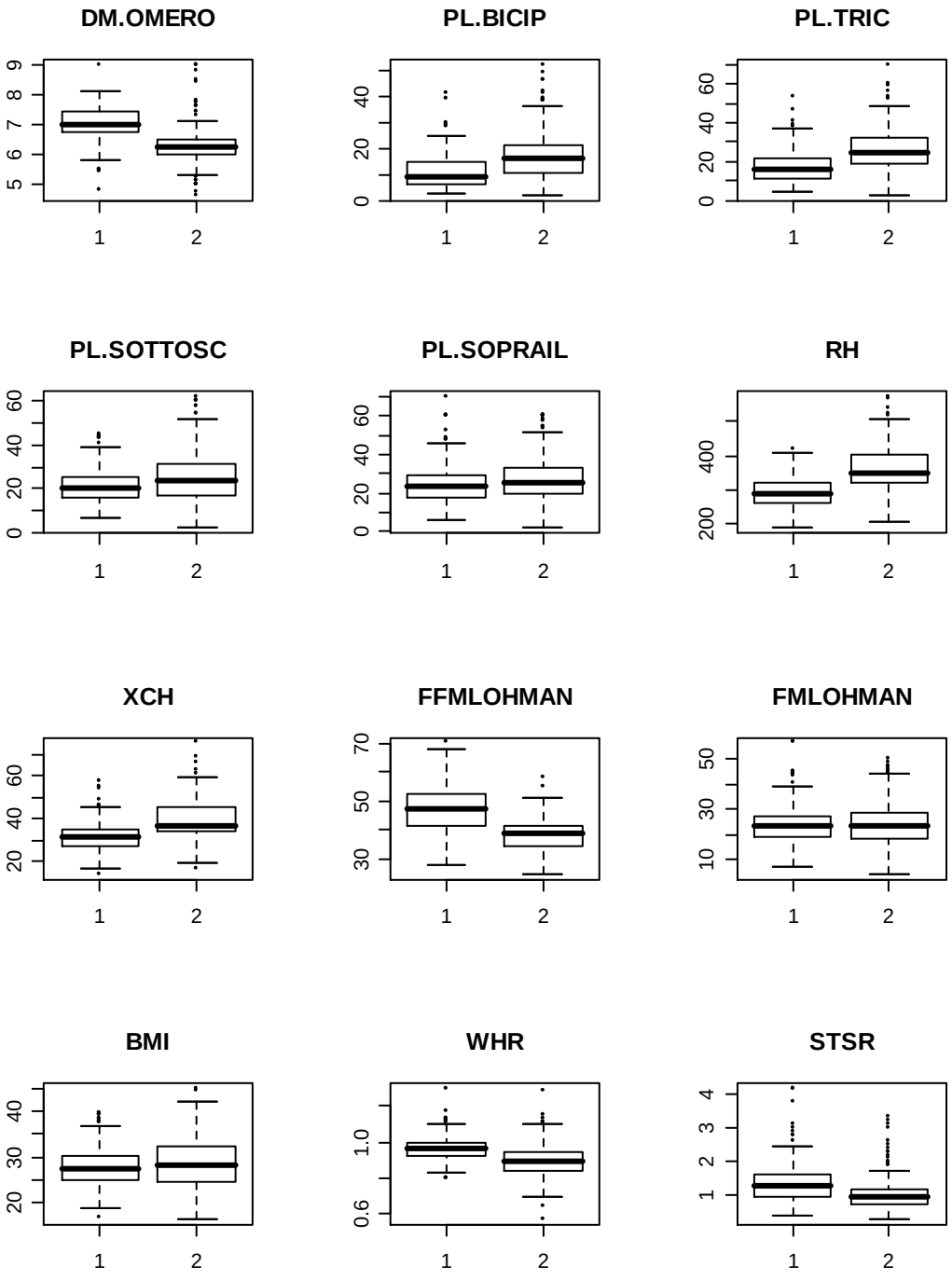
Il comando in **R** che permette di ottenere boxplot condizionati è

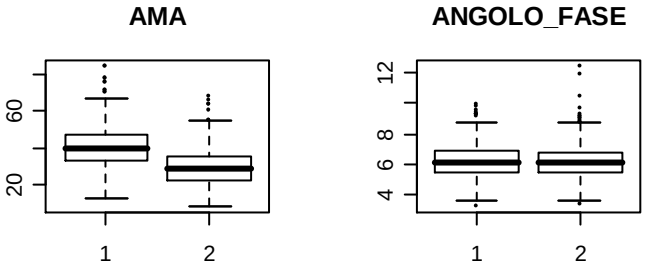
`boxplot( y ~ x, main="…")`,

dove `y` è la variabile risposta e `x` è la variabile alla quale si condiziona il grafico.

Condizionatamente al sesso si ottengono i grafici riportati nel seguito.

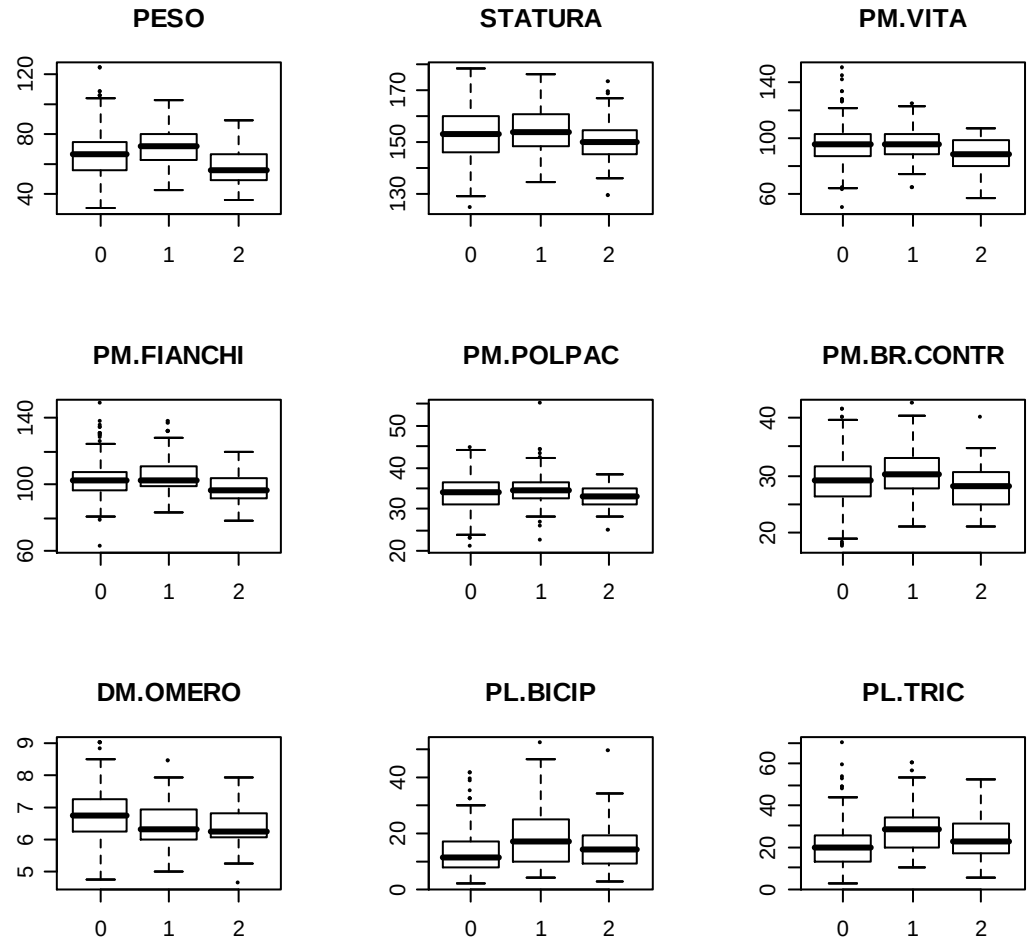


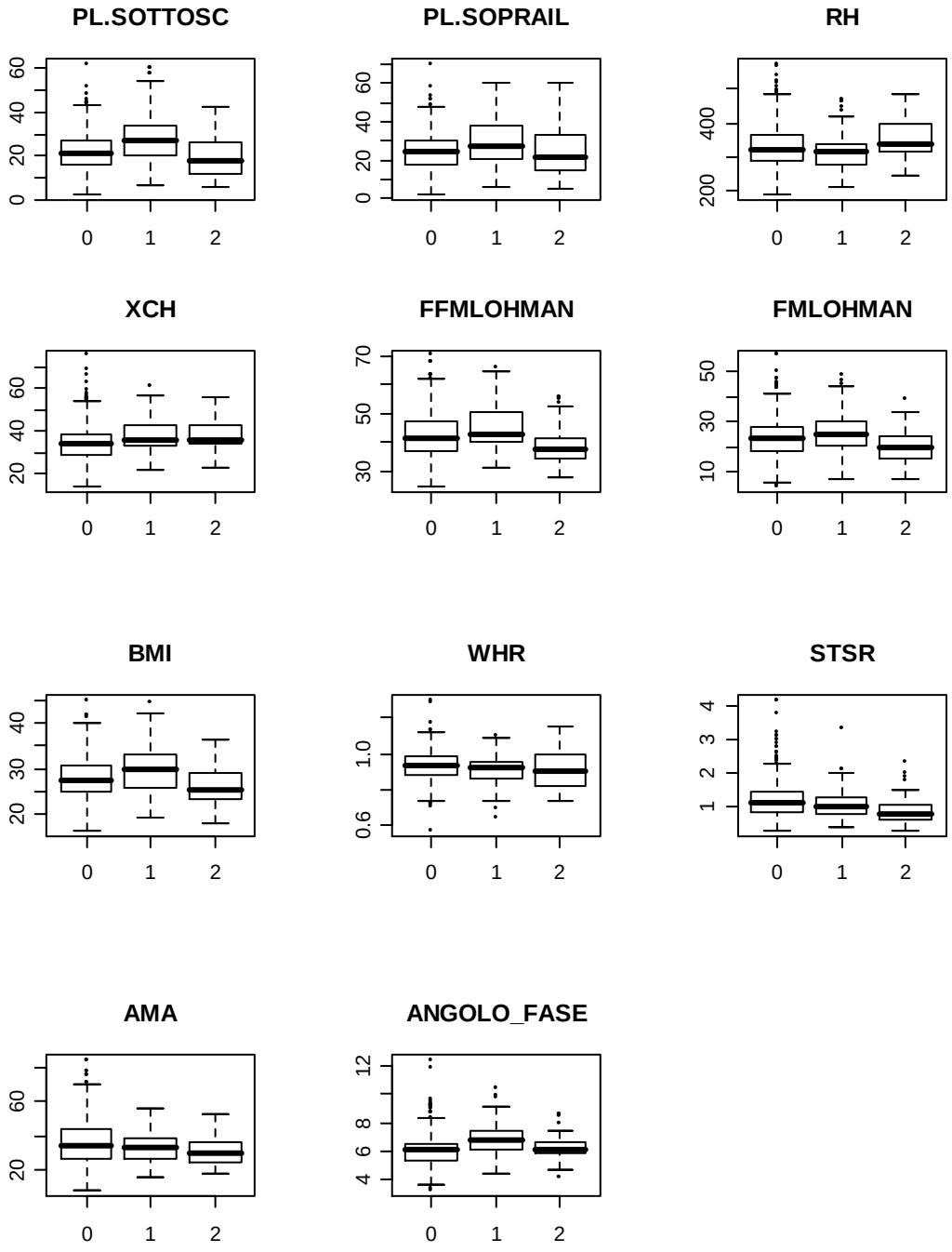

DM.OMERO
PL.BICIP
PL.TRIC
PL.SOTTOSC
PL.SOPRAIL
RH
XCH
FFMLOHMAN
FMLOHMAN
BMI
WHR
STSR



Dall'analisi dei boxplot condizionati si evince che alcune variabili assumono valori diversi a seconda del sesso, quali `dm.omero`, `stsr`, `ama` o `statura`. Altre hanno distribuzioni praticamente identiche per le due modalità, come ad esempio `angolo_fase` o `fmlohman`. Ciò significa che per alcune variabili risposta l'esplicativa sesso risulterà fortemente discriminante.

Condizionatamente alla patologia, che assume tre modalità si ottengono i seguenti grafici





Anche in questo caso si evidenzia la presenza di valori anomali nella distribuzione di alcune delle variabili che sembrano più numerosi nel caso in cui la patologia assuma

modalità 0. Le variabili `fmlohaman`, `pl.sottosc` e `pl.bicip` sembrano quelle maggiormente discriminate dalla patologia.

3.4 Conclusioni

Poiché il modello di regressione multivariata si fonda sull'ipotesi di normalità delle risposte si possono provare delle trasformazioni delle variabili che le rendano, approssimativamente, normali. Trasformazioni possibili sono le funzioni logaritmo e radice quadrata, o la trasformata di Box-Cox .
La trasformata di Box-Cox assume la seguente forma

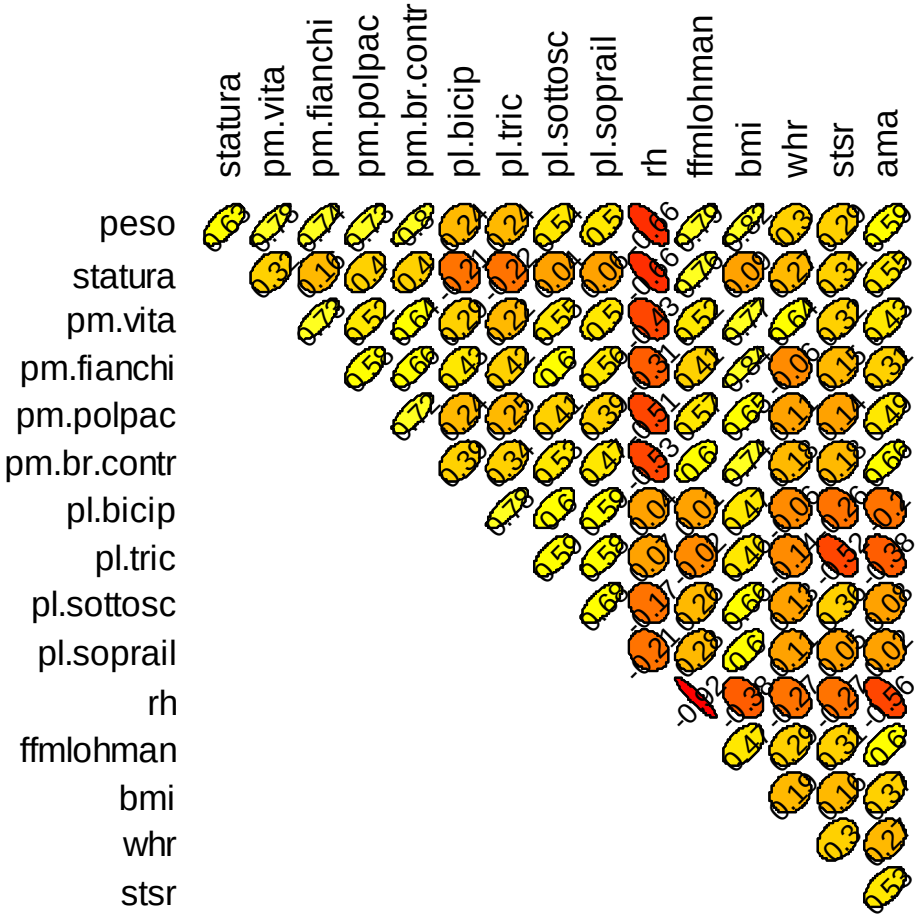
$$\begin{cases} \dfrac{Y^{\lambda}-1}{\lambda} \\ \log(Y) \end{cases}$$

ed in particolare è pari a *log(Y)* quando $\lambda = 0$. Inoltre, anche la radice quadrata si può pensare come un caso particolare della trasformata di Box-Cox quando $\lambda = 1/2$. Tale trasformata è particolarmente utile a stabilizzare la varianza (si veda Di Fonzo-Lisi, 2005 [3]).
Dalle suddette analisi si conclude che il dataset definitivo, che verrà usato per stimare il modello è costituito dalle seguenti variabili:

- `sqrt(peso);`
- `log(statura);`
- `sqrt(pm.vita);`
- `log(pm.fianchi);`
- `pm.polpac;`
- `pm.br.contr;`
- `BoxCox(pl.bicip, 0.4);`
- `BoxCox(pl.tric,0.4);`
- `BoxCox(pl.sottosc,0.4);`
- `sqrt(pl.soprail);`
- `BoxCox(rh,0.7);`
- `sqrt(ffmlohman);`
- `log(bmi);`

- `whr;`
- `log(stsr);`
- `sqrt(ama).`

Le variabili `dm.omero`, `xch`, `fmlohman` e `angolo_fase` non si possono assumere normali, neanche se trasformate. Per questo verranno escluse nelle analisi di seguito riportate. Può essere utile valutare le correlazioni tra le nuove variabili create



Sul nuovo dataset sarà stimato un modello di regressione multivariata, avente come risposte le variabili trasformate. Prima di stimare il modello sui dati è possibile cercare di valutare l'effetto marginale delle variabili esplicative sulle risposte. Per quanto riguarda il sesso e la patologia, sono già state effettuate alcune analisi in proposito tramite i boxplot condizionati. È, quindi possibile valutare l'effetto del sesso sulle variabili riposta tramite un test *t* a due campioni; l'ipotesi nulla del test è che le medie

delle due popolazioni, quella maschile e quella femminile, per la variabile considerata, siano uguali. Il comando usato in **R** per ottenere un test *t* a due campioni è `t.test(x, …)`, dove `x` è il vettore dei dati.

Di seguito è riportata la tabella contenente il valore del *t*-test e il relativo *p-value*.

Tutte le analisi successive sono state effettuate sulle variabili trasformate in modo da essere compatibili con la normalità.

***t*-test (sesso)**

Variabile	statistica test	p-value
peso	10.3521	< 2.2e-16
statura	26.9829	< 2.2e-16
pm.vita	5.719	1.561e-08
pm.fianchi	-2.7481	0.006147
pm.polpac	4.683	3.363e-06
pm.br.contr	4.7543	2.400e-06
pl.bicip	-11.8423	< 2.2e-16
pl.tric	-13.6355	< 2.2e-16
pl.sottosc	-4.3024	1.924e-05
pl.soprail	-2.7935	0.005349
rh	-20.6872	< 2.2e-16
ffmlohman	19.5933	< 2.2e-16
bmi	-1.3679	0.1718
whr	12.6781	< 2.2e-16
stsr	11.1546	< 2.2e-16
ama	13.9487	< 2.2e-16

Il sesso non risulta discriminante solo per la variabile `bmi`. Per tutte le altre variabili risposta si rifiuta l'ipotesi che ci sia uguaglianza in media nei due gruppi.

Per quanto riguarda la variabile `patologia`, è possibile effettuare il `oneway.test` che permette di verificare l'ipotesi di uguaglianza in media di più di due gruppi.

Tale test consiste in un'analisi della varianza ed è una generalizzazione del *t*-test.

Di seguito si riportano i valori della statistica test e del *p-value*.

Variabile	Valore del test F	p-value
peso	22.224	4.42e-09
statura	5.4096	0.005493
pm.vita	8.5718	0.0003142
pm.fianchi	11.6249	2.248e-05
pm.polpac	7.3813	0.0008974
pm.br.contr	9.1997	0.0001799
pl.bicip	24.5241	1.069e-09
pl.tric	45.2213	1.059e-15
pl.sottosc	27.829	9.655e-11
pl.soprail	10.835	4.676e-05
rh	9.2292	0.0001753
ffmlohman	13.588	4.231e-06
bmi	16.9252	2.825e-07
whr	4.4189	0.01394
stsr	15.7685	7.42e-07
ama	6.6231	0.001752

Al 5% di significatività si rifiuta l'ipotesi nulla di uguaglianza della media nei tre gruppi per tutte le variabili. All'1% si accetterebbe solo per la variabile `whr`.

Per valutare l'effetto marginale dell'età, che è una variabile quantitativa, si può analizzare la correlazione di `eta` con le variabili risposta.

	peso	statura	pm.vita	pm.fianchi	pm.polpac	pm.br.contr
cor	-0.471	-0.405	-0.169	-0.222	-0.458	-0.498

	pl.bicip	pl.tric	pl.sottosc	pl.soprail	rh	ffmlohman
cor	-0.078	-0.115	-0.298	-0.263	0.371	-0.473

	bmi	whr	stsr	ama
cor	-0.313	0.011	-0.178	-0.389

Si può notare che la correlazione assume valore negativo per la maggior parte delle variabili, eccetto che per `rh` e `whr`. Inoltre la correlazione tra `eta` e `whr` assume un valore piuttosto piccolo. La maggior parte delle correlazioni assume un valore maggiore di 0.2 in valore assoluto. Questo potrebbe essere un segnale significativo della presenza di un effetto marginale della variabile relativa all'età sulle risposte. In seguito a queste analisi si può effettuare la stima del modello, tenendo anche conto degli effetti congiunti delle variabili, al fine di valutarne la significatività.

Capitolo 4

Adattamento ai dati

4.1 Il pacchetto *lm*

Il comando *lm* in R è utilizzato per stimare modelli lineari, sui quali effettuare regressioni, analisi della varianza e della covarianza (in particolare l'opzione aov viene utilizzata per questi scopi).
L'espressione del comando è:

```
lm(formula, data, subset,...)
```

Argomenti:

- `formula` indica la specificazione del modello che si vuole stimare;
- `data` indica il dataset sul quale effettuare la regressione espressa nella formula e contiene le variabili del modello; se l'opzione viene omessa, **R** effettua la regressione sul dataset sul quale è stato effettuato l'*attach*;
- `subset` si utilizza per selezionare un sottoinsieme di osservazioni sul quale effettuare la regressione.

La formula del modello, nel comando *lm*, è del tipo "risposte ~ esplicative", dove "risposte" indica il vettore o la matrice delle variabili dipendenti del modello, mentre "esplicative" indica la matrice delle variabili indipendenti, che, combinate linearmente, modellano la risposta. Se, ad esempio, nella formula abbiamo due esplicative, x_1 e x_2, e

il modello è espresso come y ~ x_1 + x_2, ciò vuol dire che sulla variabile risposta si considera l'effetto di x_1 e l'effetto di x_2 separatamente.
Infine se il modello è del tipo y ~ x_1 * x_2, si valutano, sulla risposta y, gli effetti di x_1 e x_2 presi separatamente e l'effetto della loro interazione; x_1 * x_2 corrisponde ad x_1 + x_2 + x_1:x_2.
Nel caso in cui la formula sia del tipo y ~ x - 1, ciò significa che si sta considerando il modello di regressione senza intercetta.
Un caso particolare della regressione col comando *lm* si ha quando ***Y*** è una matrice ($n \times p$) di risposte e non un vettore; questo è il caso di interesse nel caso di regressione multipla multivariata.
In questa situazione R stima separatamente le regressioni ai minimi quadrati per ogni colonna della matrice ***Y***.
Utilizzando il comando `summary( )` si ottiene un sommario dell'output, comprendente le stime e gli *standard error* dei coefficienti, con i corrispondenti valori dei test per la verifica della significatività e i relativi *p-value*. Inoltre vengono visualizzati il minimo, il massimo, il primo e il terzo quartile e la mediana della distribuzione dei residui del modello.
Infine compaiono anche i valori dell'R^2 e dell'R^2 corretto, il *residual standard error*, che rappresenta la quantità di varianza che non è spiegata dal modello, con i relativi gradi di libertà e la statistica *F* per il modello con il solo termine d'intercetta, con il *p-value* ad essa associato.
Una volta stimato il modello di regressione multivariata, si può utilizzare la funzione `anova( )` per confrontare modelli "annidati" e per valutare se sia opportuno togliere o aggiungere una determinata variabile esplicativa nel modello; ciò si effettua considerando se risulta significativo il peggioramento o il miglioramento in termini di varianza residua, che si ottiene con la modifica adottata.

4.1.1 Il comando `anova.mlm()`

Nel caso di modelli multivariati, l'estensione dell'opzione `anova()` è il comando `anova.mlm()`.

L'espressione di tale comando è:

```
anova.mlm(object,…, test = c("Pillai", "Wilks", "Hotelling-Lawley",
"Roy", "Spherical"),…)
```

dove:

- `object` è un oggetto della classe `mlm`;
- `test` indica il tipo di test che si desidera visualizzare in output: Λ di Wilks, T^2 di Hotelling, Pillai's *Trace*, etc..

4.2 La stima del modello completo

Il modello più complesso che si può stimare sui dati, presentati nel capitolo precedente, inserisce tra le esplicative sesso, patologia, età e tutte le interazioni possibili tra le tre variabili.
Il comando per stimare con R questo modello è il seguente

```
fit<-lm(cbind(peso, statura, pm.vita, pm.fianchi, pm.polpac, pm.br.contr,
pl.bicip, pl.tric, pl.sottosc, pl.soprail, rh, ffmlohman, bmi, whr, stsr,
ama) ~ sesso * patologia * eta)
```

Con la stima di questo modello si ottiene una matrice $\hat{\boldsymbol{B}}$ di coefficienti stimati di dimensioni (11 × 16), perché le variabili risposta sono 16 e le esplicative, poiché si considerano tutte le interazioni possibili, sono 11.
Più che valutare il valore dei singoli coefficienti, ha senso cercare di determinarne la significatività, per stabilire se sia possibile semplificare il modello.
Si può inoltre valutare, in termini di devianza residua, la bontà del modello e il grado di adattamento ai dati.
Inoltre per tutte le variabili si rifiuta l'ipotesi che sia buono il modello con la sola intercetta come variabile esplicativa.
L'intercetta è l'unico parametro del modello significativo, a tutti i livelli di significatività usualmente considerati, per tutte le variabili.

L'età ha un effetto rilevante e negativo per la maggior parte delle risposte, eccetto che per `pl.bicip`, `pl.tric` ed `whr` per le quali il parametro risulta non significativo e `rh`, che è l'unica variabile per la quale la stima del coefficiente di `eta` è positivo.
Poiché gli antropologi sono interessati alla costruzione di curve di crescita, è importante notare che l'età è una variabile esplicativa importante per spiegare gli indicatori antropometrici calcolati e che ha una relazione negativa con questi ultimi: l'aumentare dell'età porta, indicativamente, ad una riduzione delle quantità considerate.
La variabile `sesso` sembra avere un effetto particolarmente significativo sulle variabili `statura`, `pm.vita`, `pl.bicip`, `pl.tric`, `pl.sottosc`, `rh`, `ffmlohman`, `whr`, ed `ama`. Il *p-value* associato all'ipotesi di nullità del coefficiente della variabile `sesso`, per queste variabili, è inferiore a 0.05.
I coefficienti associati alle due variabili dicotomiche che codificano la patologia sono significativi, entrambi o almeno uno dei due, al 5% per le seguenti variabili: `pl.bicip`, `pl.tric`, `pl.sottosc` ed `ama`.
Riguardo alle interazioni tra le variabili esplicative del modello, l'interazione a tre è significativa solo per le variabili `pl.tric`, `rh` e `ffmlohman`. L'interazione tra età e sesso è rilevante per spiegare `pm.vita`, `pl.bicip`, `pl.tric`, `pl.sottosc`, `ffmlohman` e `whr`.
L'interazione tra patologia e sesso ha un coefficiente significativo al 5% per `pl.tric`, `rh`, `ffmlohman` e `stsr`.
Infine l'interazione tra patologia ed età è indicativa per le variabili `pl.bicip`, `pl.tric`, `pl.sottosc` ed `ama`.

```
anova.mlm(fit)
Analysis of Variance Table

                    Df   Pillai approx F num Df den Df    Pr(>F)
(Intercept)          1        1 24071478     16    716 < 2.2e-16 ***
sesso                1        1      116     16    716 < 2.2e-16 ***
patologia            2  0.30496        8     32   1434 < 2.2e-16 ***
eta                  1  0.42672       33     16    716 < 2.2e-16 ***
sesso:patologia      2  0.13195        3     32   1434 9.412e-09 ***
sesso:eta            1  0.20541       12     16    716 < 2.2e-16 ***
patologia:eta        2  0.09351        2     32   1434 0.0001422 ***
```

```
sesso:patologia:eta    2  0.06736        2     32   1434 0.0241187 *
Residuals            731
```

Dall'analisi dell'output del comando `anova.mlm()` sul modello `fit`, risulta che eliminando il termine di interazione a tre dalla regressione multipla, il peggioramento in termini di devianza residua che si otterrebbe non è significativo all'1% (*p-value* = 0.02). Si può provare ad eliminare l'interazione a tre e vedere come cambia il modello.

4.2.1 Semplificazione del modello

Per eliminare il termine che rappresenta l'interazione a tre tra le esplicative del modello si usa il comando

```
fit1<-update(fit,.~.-sesso:patologia:eta)
```

dove il comando `update` indica che si sta aggiornando, modificando il modello `fit`, escludendo il termine `sesso:patologia:eta`.
La nuova matrice di coefficienti stimati è ora di dimensioni 9 × 16.
I *residual standard errors* del nuovo modello sono poco differenti da quelli del modello precedente. Inoltre in questo caso vanno confrontati con 733 e non 731 gradi di libertà. Se valutato per ogni singola variabile risposta, il modello più semplice sembra essere buono tanto quanto il precedente.

```
anova(fit1)
Analysis of Variance Table

                 Df   Pillai approx F num Df den Df    Pr(>F)
(Intercept)       1        1 23875740     16    718 < 2.2e-16 ***
sesso             1        1      116     16    718 < 2.2e-16 ***
patologia         2  0.30144        8     32   1438 < 2.2e-16 ***
eta               1  0.42497       33     16    718 < 2.2e-16 ***
sesso:patologia   2  0.13153        3     32   1438 9.586e-09 ***
sesso:eta         1  0.20529       12     16    718 < 2.2e-16 ***
patologia:eta     2  0.09311        2     32   1438 0.0001470 ***
```

```
Residuals        733
```

I termini inseriti nel modello sembrano tutti ampiamente significativi. Se si confrontano i due modelli si ottiene

```
anova.mlm(fit,fit1)
Analysis of Variance Table
  Res.Df   Df Gen.var.  Pillai approx F num Df den Df  Pr(>F)
1    731       0.14217
2    733    2  0.14240 0.06736  1.56199     32   1434 0.02412 *
```

Da ciò si evidenzia che mentre per le singole variabili il termine di interazione a tre tra `sesso`, `patologia` ed `eta` non era significativo, la sua eliminazione dal modello comporta una perdita significativa nel modello complessivo, nel quale si tiene conto anche della correlazione tra le variabili ***Y***.
Se si eliminasse anche una interazione a due, il peggioramento del modello risulterebbe ancora significativo. Si possono stimare tre modelli

```
fit21<-update(fit1,.~.-patologia:eta)
fit22<-update(fit1,.~.-sesso:eta)
fit23<-update(fit1,.~.-sesso:patologia)
```

e confrontandoli con `fit1` si ottiene

```
anova.mlm(fit1,fit21)
Analysis of Variance Table
  Res.Df   Df Gen.var.  Pillai approx F num Df den Df    Pr(>F)
1    733       0.14240
2    735    2  0.14286 0.09311  2.19420     32   1438 0.0001470 ***
```

```
anova.mlm(fit1,fit22)
Analysis of Variance Table
  Res.Df  Df Gen.var.  Pillai approx F num Df den Df    Pr(>F)
1    733       0.1424
2    734   1   0.1443  0.2095  11.8914     16    718 < 2.2e-16 ***
```

```
anova.mlm(fit1,fit23)
Analysis of Variance Table
  Res.Df   Df Gen.var.  Pillai approx F num Df den Df    Pr(>F)
1    733       0.14240
2    735    2  0.14304 0.11119  2.64535     32   1438 2.007e-06 ***
```

Si conclude quindi che il modello più significativo per spiegare i dati è quello completo. Si può quindi, stimare sui dati il modello ritenuto opportuno.

Dall'analisi del modello ridotto risulta che il coefficiente dell'intercetta è significativo per tutte e sedici le variabili risposta.

Il coefficiente della variabile dicotomica `sesso` risulta significativo al 5% per le variabili `pm.vita`, `pl.tric`, `pl.sottosc`, `rh`, `ffmlohman`, `whr`, `statura` e `ama`. Per le ultime due, però, il coefficiente risulta non significativo se si fissa come livello d'accettazione l'1%.

La variabile `patologia` risulta significativa al 5% nelle regressioni delle seguenti variabili: `peso`, `statura`, `pm.fianchi`, `pm.polpac`, `pl.sottosc`, `bmi`, `ama`, `pm.vita` e `pm.br.contr`. Per le ultime due variabili, la patologia non risulta avere un effetto significativo se il livello considerato è l'1%. Inoltre al 10% la patologia è significativa anche per la variabile `ffmlohman`.

Per quanto riguarda la variabile `eta`, il suo effetto è significativo su tutte le risposte al 5% di significatività, eccetto che su `pl.bicip`, `pl.tric` e `whr`. Inoltre, se il livello di accettazione è l'1%, l'età non risulta rilevante neanche nella regressione della variabile `stsr`.

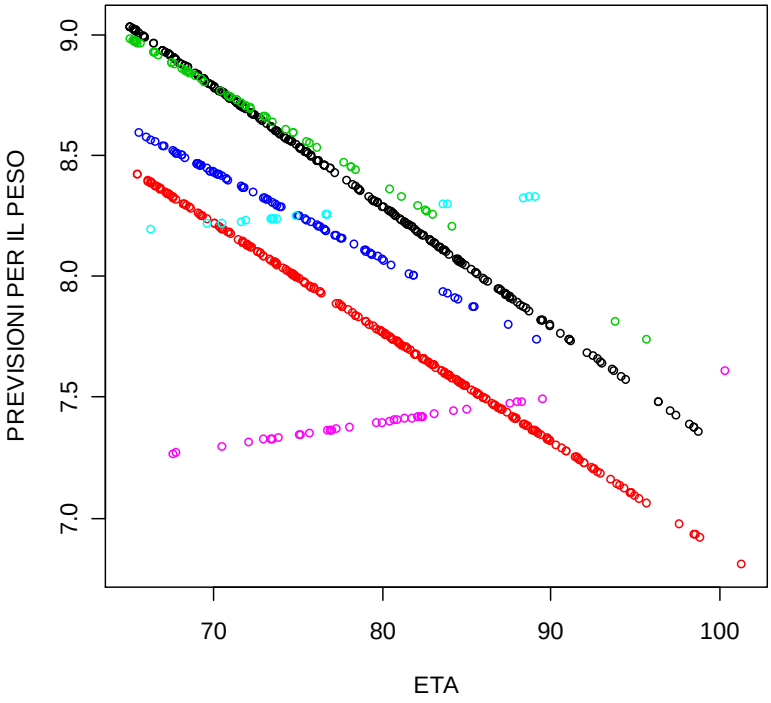
Oltre all'effetto delle singole esplicative, si può valutare se le interazioni tra esse risultano significative.

L'interazione tra sesso ed età, ha un effetto significativo al 5% solo sulle variabili `pm.vita`, `pl.bicip`, `pl.tric`, `pl.sottosc`, `ffmlohman` e `whr`; i coefficienti, per altro, risultano significativi all'1% solo per le ultime due variabili. Tenendo conto, però, delle correlazioni tra le risposte, complessivamente il coefficiente dell'interazione tra sesso ed età risulta rilevante.

I coefficienti dell'interazione tra sesso e patologia sono significativi, entrambi o uno dei due, al 5% per le variabili `pm.vita`, `pm.polpac`, `pm.br.contr`, `pl.soprail`, `stsr`, `ama`, `pm.fianchi`, `pl.sottosc` e `whr`. In particolare, per le ultime tre variabili il coefficiente dell'interazione non risulterebbe significativo se il livello di significatività considerato fosse l'1%. Al 10% l'interazione tra sesso e patologia è rilevante anche nella regressione per `bmi`.
Infine, l'interazione tra patologia e età è significativa al 5% nelle regressioni delle seguenti risposte: `ama`, `peso`, `statura`, `pm.fianchi`, `pm.polpac`, `pm.br.contr`, `pl.sottosc` e `bmi`. Nelle regressioni relative alle ultime tre variabili, il coefficiente di interazione non risulterebbe significativo all'1%. Se fa riferimento ad una probabilità del 10%, risultano significativi anche i coefficienti dell'interazione nelle regressioni delle variabili `pm.vita`, `rh` e `ffmlohman`.

4.3 Previsioni

A partire dal modello considerato è possibile costruire i valori predetti con l'opzione `fitted(fit)`del comando `lm`. Con tale comando si ottiene una matrice di previsioni di dimensioni (743×16) poiché nel caso in esame la regressione effettuata è multivariata. Si può, quindi, dalla matrice estrarre il vettore delle previsioni per una variabile di interesse. In particolare, le previsioni costruite possono essere rappresentate graficamente in un diagramma di dispersione in cui la variabile esplicativa è l'età. Inoltre, è possibile distinguere in tali grafici le unità statistiche in base al sesso e alla patologia rilevata. Discriminando in base ai due fattori, le curve di crescita che si ottengono per le previsioni della variabile risposta `peso` sono le seguenti

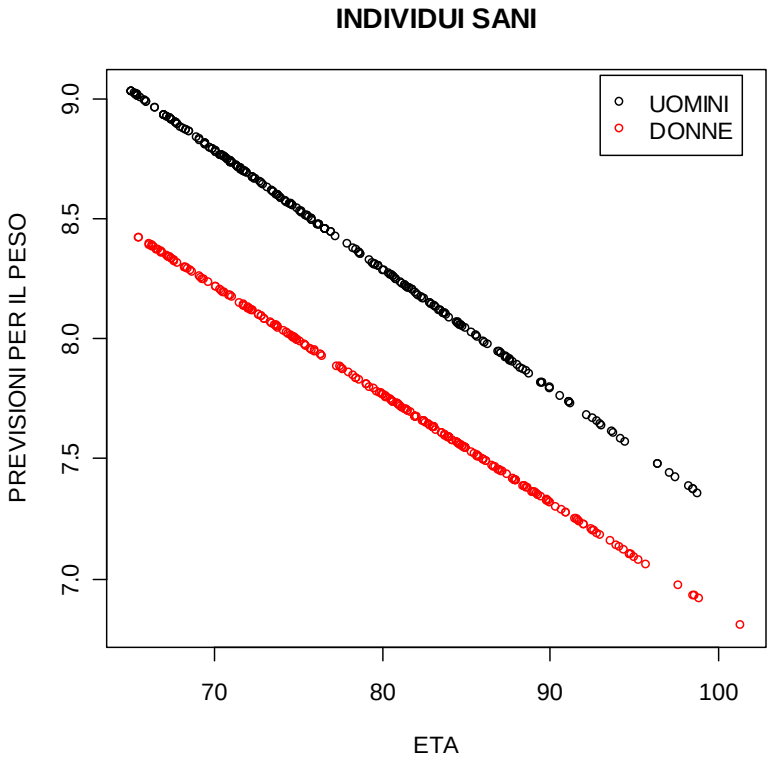


LEGENDA	
uomini sani	donne diabetiche
donne sane	uomini con Alzheimer
uomini diabetici	donne con Alzheimer

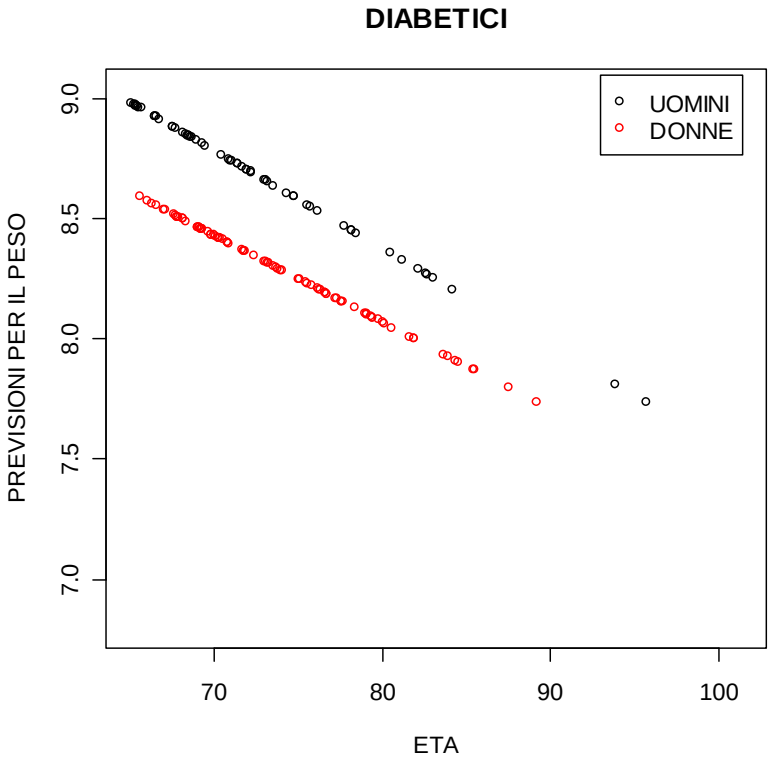
Si ricorda che le previsioni per la variabile `peso` sono ottenute a partire dalla radice quadrata della variabile contenuta nel dataset di partenza, effettuata ai fini della normalizzazione. A causa di ciò i valori in ordinata sul grafico, sono compresi tra 7.0 e 9.0, circa.

Come si può osservare, per ognuna delle due modalità del sesso si distinguono tre differenti curve, che corrispondono alla tre modalità della variabile `patologia`.

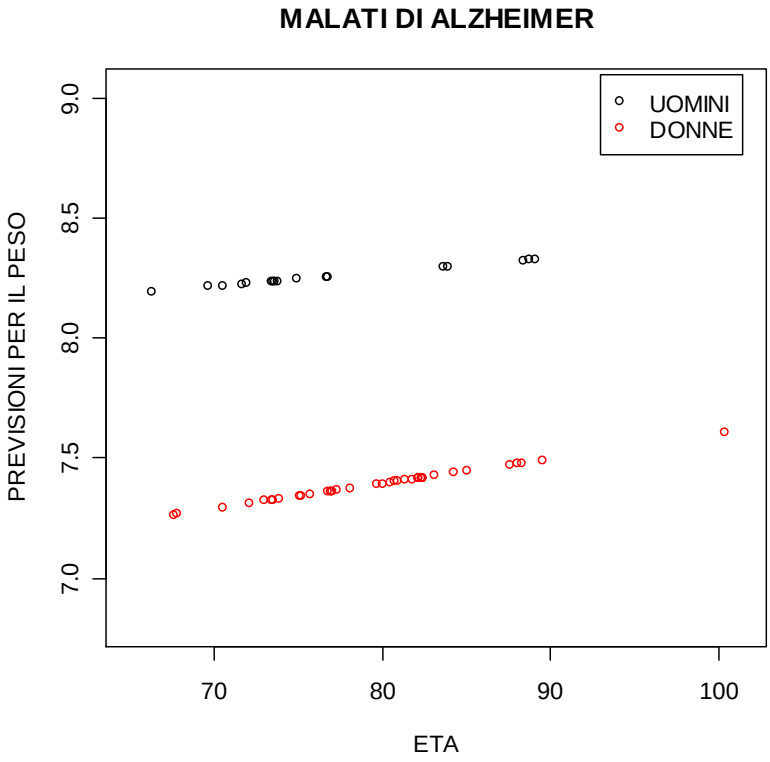
Se, invece, si rappresentano, separatamente per la patologia, le curve relative ai maschi e alle femmine, si ottiene, per gli individui sani, il seguente grafico



Per i diabetici il grafico ottenuto è il seguente



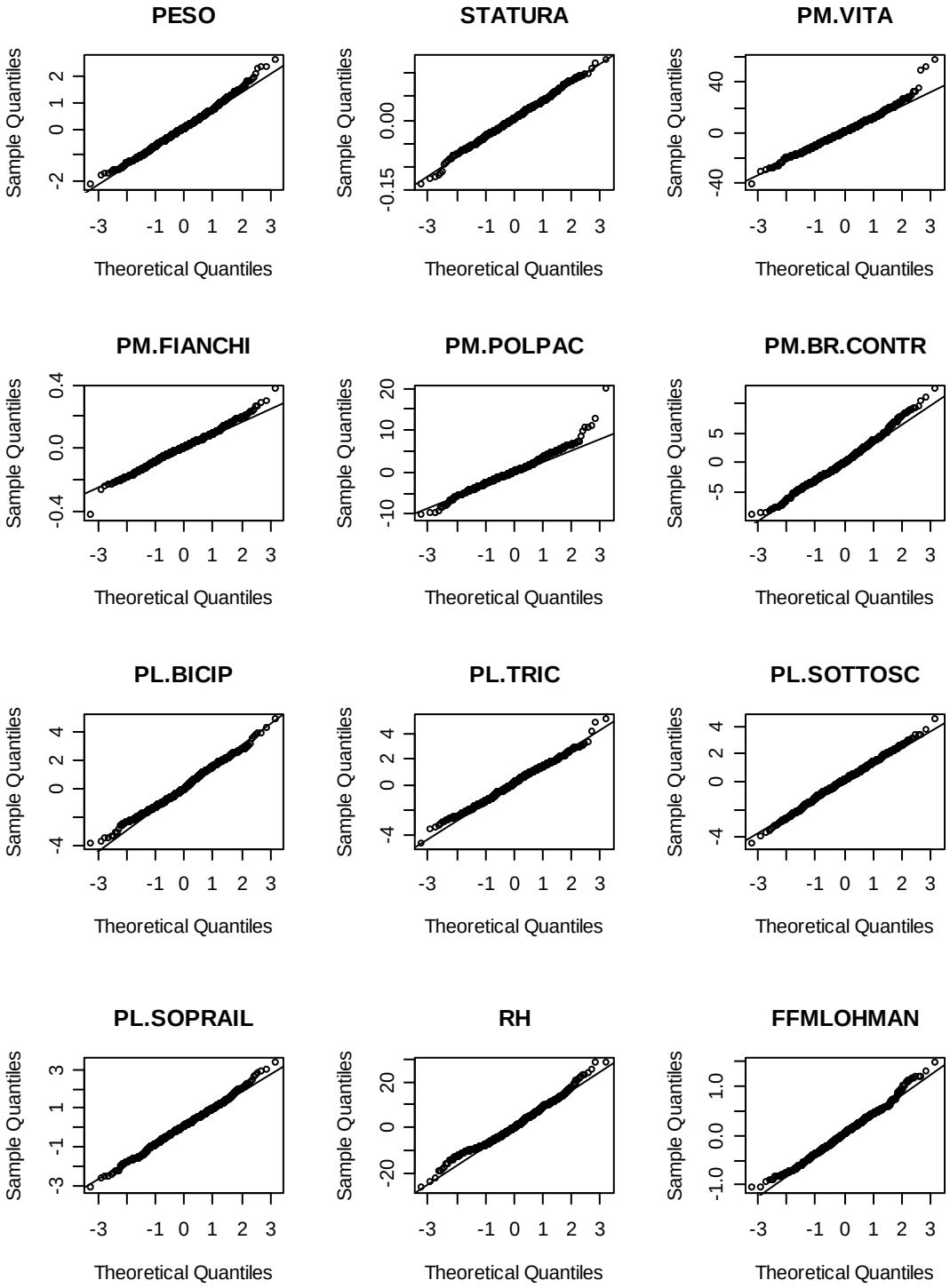
Infine, per i malati di Alzheimer si ha

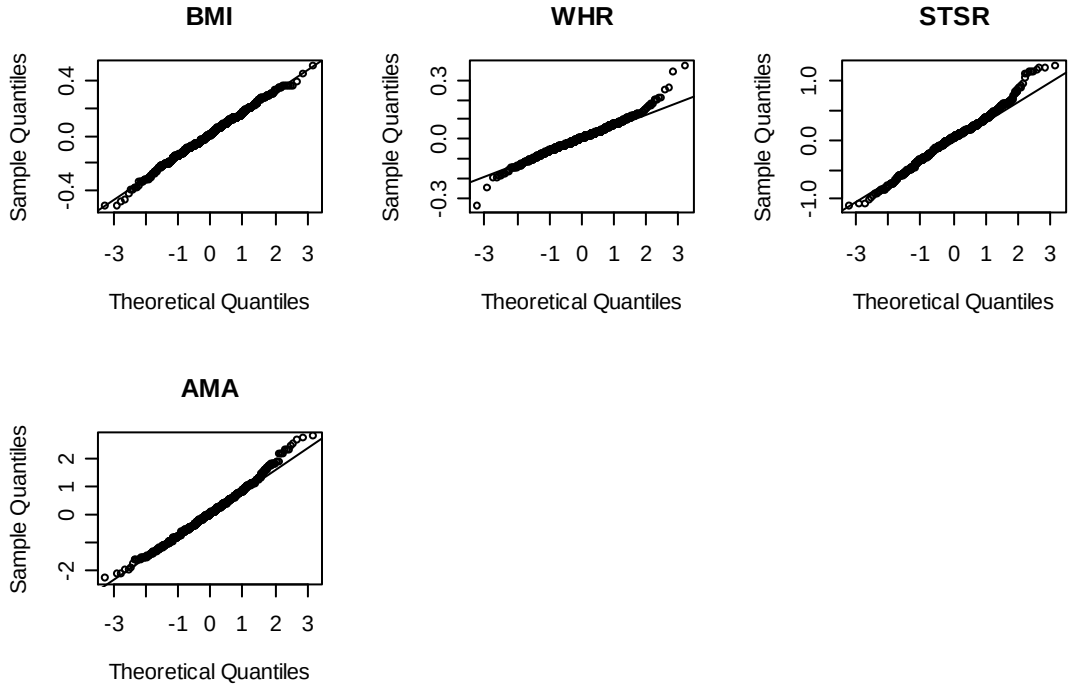


Come si può notare, le curve sono decrescenti per gli individui sani e per i diabetici, appaiono invece costanti o al più debolmente crescenti per i malati di Alzheimer. Il peso, dunque, sembra diminuire con l'età per i primi due gruppi di individui, a prescindere dal sesso. Per i malati di Alzheimer, invece, il peso sembra rimanere praticamente costante al variare dell'età. Inoltre, mentre per gli individui sani e per i malati di Alzheimer le curve distinte per i due sessi sembrano parallele, per i diabetici sembra esserci un cambiamento di pendenza.

4.4 Conclusioni

Il modello stimato può essere validato anche a partire dall'analisi dei residui, che si possono calcolare come differenza tra valori osservati e valori predetti. Secondo le ipotesi alla base del modello, i residui dovrebbero provenire da una distribuzione normale. È quindi possibile valutare graficamente se provengono da una normale, attraverso i *q-qplot*.


PESO
STATURA
PM.VITA
PM.FIANCHI
PM.POLPAC
PM.BR.CONTR
PL.BICIP
PL.TRIC
PL.SOTTOSC
PL.SOPRAIL
RH
FFMLOHMAN
Sample Quantiles
Theoretical Quantiles



I grafici quantile-quantile sembrano compatibili con l'ipotesi di normalità per tutte le variabili, eccetto che per `pm.vita`, `pm.polpac`, `whr` e `stsr` per le quali potrebbero esserci dei dubbi.

Dalle analisi effettuate in seguito alla stima del modello senza interazione a tre, si può concludere che, benché valutati per ogni singola variabile risposta, talvolta gli effetti delle variabili esplicative del modello sembrino non essere rilevanti, tuttavia, tenendo conto della correlazione delle risposte, che assume per alcune valori significativamente alti, tutti i coefficienti stimati nel modello sono significativi.

Ciò porta a concludere che le variabili esplicative considerate siano degli indicatori adeguati a spiegare le variabili risposta considerate. In particolare, le variabili di natura antropometrica sembrano avere una loro evoluzione collegata all'età, in particolare, sembra che le risposte tendano a decrescere proporzionalmente all'aumentare della variabile `eta` e ciò è coerente con la natura dei dati. Infatti, è risaputo e comprovato da studi medici, che col passare degli anni la statura tenda a ridursi, e con essa la lunghezza delle ossa e di tutte le parti del corpo umano.

Inoltre, è ragionevole pensare che ci siano differenze tra i due sessi nelle grandezze osservate. Infine, nel caso di particolari patologie, si possono riscontrare variazioni significative delle dimensioni di grandezze corporee.

Si può, per altro, ritenere sensato che ci sia un effetto anche delle interazioni: in genere, alcune patologie sono più comuni in età avanzata. L'effetto congiunto di `patologia` ed `eta` risulterà quindi rilevante; inoltre, se ci sono differenze tra i due sessi nelle misurazioni antropometriche, è logico anche pensare che tali discrepanze possano aumentare proporzionalmente all'età.

In conclusione, l'applicazione del modello di regressione multivariata ai dati porta a concludere che esiste una relazione significativa tra la misurazione delle grandezze antropometriche più rilevanti e l'età, il sesso e la presenza o meno di determinate patologie. È quindi possibile fornire stime delle grandezze che rappresentano le variabili risposta del caso-studio, applicando il modello ritenuto ottimale, a partire dai valori osservati delle esplicative. Inoltre, poiché l'interesse degli antropologi era rivolto soprattutto ai soggetti in età avanzata, è anche possibile, a partire dai dati e dal modello stimato, costruire curve di crescita, che mettano in evidenza l'evoluzione delle variabili risposta al variare dell'età.

Appendice

Distribuzioni multivariate

Distribuzione di Wishart

Sia $\boldsymbol{M}$ una matrice ($p \times p$) che può essere scritta come $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, dove $\boldsymbol{X}$ è una matrice ($m \times p$) che segue una distribuzione normale p-variata con media nulla e matrice di varianza e covarianza $\boldsymbol{\Sigma}$. Allora $\boldsymbol{M}$ ha distribuzione di Wishart con matrice di scala $\boldsymbol{\Sigma}$ e m gradi di libertà. Se $\boldsymbol{\Sigma} = \boldsymbol{I_p}$, la distribuzione si dice standard.

Nel caso in cui p sia uguale ad 1, la distribuzione $W_1(\sigma^2, m)$ è data da $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$, con $\boldsymbol{x}$ vettore ($m \times 1$) distribuito come una da $N(0,\sigma^2)$. In questo caso $W_1(\sigma^2,m)$ è la distribuzione χ^2_m moltiplicato per $\boldsymbol{\sigma^2}$.

La matrice di scala $\boldsymbol{\Sigma}$ gioca, per la distribuzione di Wishart, lo stesso ruolo che ha σ^2 nel caso particolare in cui $p = 1$.

Se $\boldsymbol{M} \sim W_p(\boldsymbol{\Sigma},m)$ e $\boldsymbol{a}$ è un vettore ($p \times 1$) tale per cui $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma a} \neq 0$, allora

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{Ma}/\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma a} \sim \chi^2_m .$$

Distribuzione Λ

Sia $\mathbf{A} \sim W_p(\boldsymbol{I},m)$ e sia $\mathbf{B} \sim W_p(\mathbf{I},n)$, indipendenti. Se $m \geq p$ allora

$$\Lambda = |\boldsymbol{A}|/|\boldsymbol{A} + \boldsymbol{B}|$$

ha distribuzione $\Lambda(p,m,n)$, dove m rappresenta il numero di gradi di libertà al numeratore ed n il numero di gradi di libertà al denominatore. La quantità $m + n$ rappresenta il numero di gradi di libertà totali. La distribuzione è invariante a cambiamenti di scala.

Un'ulteriore distribuzione legata alla Λ è la *Beta*. Infatti, se $\boldsymbol{u_i} \sim B\left(\frac{1}{2}(m+i-p), \frac{1}{2}p\right)$,

con $i = 1,\ldots, n$, e se $\boldsymbol{u_1},\ldots,\boldsymbol{u_n}$ sono indipendenti, allora

$$\prod_{i=1}^{n} u_i \sim \Lambda(p,m,n).$$

La distribuzione *Beta* si ottiene dal rapporto di una variabile *u* e la somma tra *u* e *v*, con $u \sim \chi_p^2$ e $v \sim \chi_q^2$. In questo caso la *Beta* ha parametri $\frac{1}{2}p$ e $\frac{1}{2}q$.

Infine, la distribuzione Λ è anche legata alla *F*. Infatti

$$\frac{1-\Lambda(p,m,1)}{\Lambda(p,m,1)} \sim \frac{p}{m-p+1} F_{p,m-p+1}.$$

Distribuzione θ

Se $\boldsymbol{A} \sim W_p(\boldsymbol{I},m)$ e $\boldsymbol{B} \sim W_p(\boldsymbol{I},n)$, con $\boldsymbol{A}$ e $\boldsymbol{B}$ indipendenti ed $m \geq p$, allora il più grande degli autovalori di $(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{B}$ ha distribuzione $\theta(p,m,n)$, con *m* ed *n* che rappresentano, rispettivamente, il numero di gradi di libertà del numeratore e al denominatore.
Tale distribuzione ha alcune proprietà:

- $\theta(p, m, n)$ e $\theta(p, m+n-p, p)$ hanno la stessa distribuzione;
- $\frac{\theta(p,m,1)}{1-\theta(p,m,1)} \sim \frac{1-\Lambda(p,m,1)}{\Lambda(p,m,1)}$;
- $\frac{\theta(p,m,1)}{1-\theta(p,m,1)} \sim \frac{p}{m-p+1} F_{p,m-p+1}$.

Distribuzione di T^2 Hotelling

Se una statistica $\boldsymbol{\alpha}$ può essere scritta come $m\mathbf{d}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{d}$, dove **d** e **M** sono indipendenti e distribuite rispettivamente come una $N_p(0,\boldsymbol{I})$ e una $W_p(\boldsymbol{I},m)$, allora $\boldsymbol{\alpha}$ ha distribuzione T^2 di Hotelling di parametri *p* e *m*. Si può quindi scrivere $\boldsymbol{\alpha} \sim T^2(p,m)$. Più in generale, se **d** è una normale *p*-variata di media $\boldsymbol{\mu}$ e varianza $\boldsymbol{\Sigma}$ e $\boldsymbol{M}$ proviene da una distribuzione Wishart con matrice di scala $\boldsymbol{\Sigma}$, allora si può scrivere

$$m(\boldsymbol{d}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{M}^{-1}(\boldsymbol{d}-\boldsymbol{\mu}) \sim T^2(p,m)$$

Inoltre, per quanto riguarda le relazioni esistenti con altre distribuzioni notevoli, con $\boldsymbol{\alpha} \sim T^2(p,m)$, si ha

$\boldsymbol{\alpha} = m\chi_p^2/\chi_{m-p+1}^2 = \{mp/(m-p+1)\}F_{\boldsymbol{p,\,m-p+1}}$.

Distribuzione t di Student multivariata

Sia $\boldsymbol{X}$ distribuita come una normale p-variata con media $\boldsymbol{\mu}$ e matrice di varianza e covarianza $\boldsymbol{\Sigma}$ e indipendente da $\boldsymbol{y}$, distribuita come un χ^2_υ . Posto $\boldsymbol{u}_i = \boldsymbol{x}_i/(\boldsymbol{y}/\upsilon)^{1/2}$, per i = 1,…, p, allora il vettore $\boldsymbol{u}$ segue una distribuzione t di Student multivariata con vettore delle medie $\boldsymbol{\mu}$, matrice di varianza e covarianza $\boldsymbol{\Sigma}$ e υ gradi di libertà ossia

$$\boldsymbol{u} \sim t(\upsilon, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Per maggiori chiarimenti sulle distribuzioni che sono state qui brevemente introdotte, e su altre ad esse correlate, si veda Mardia *et al*.(1979).

Bibliografia

[1] Anderson, T. W. (1958) *Introduction to multivariate statistical analysis*;

[2] Box, G. E. P. (1949) *A general distribution theory for a class of likelihood criteria*;

[3] Di Fonzo T. , Lisi F. (2005) *Serie Storiche Economiche*;

[4] Levene, H. (1960). In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., Stanford University Press;

[5] Miles, D. B. (2003) *Multivariate Analysis of Variance*;

[6] Rowe , D. B. *Multivariate Regression Generalized Likelihood Ratio Test for FMRI*;

[7] Schatz, P.(2006) *Manova*;

[8] Seber, G. A. F. (1984) *Multivariate observations*;

[9] Vittadini, G. (1999). *On the use of Multivariate Regression models in the context of Multilevel Analysis*.