Università degli Studi di Padova

Facoltà di Scienze Statistiche

Corso di laurea in Statistica e Gestione delle Imprese

Il modello lineare a effetti misti: un caso di studio con R

Relatore: Prof.ssa Laura Ventura

Laureando: Luca Volini

Anno Accademico 2011/2012

INDICE

Prefazione

Molte ricerche sono caratterizzate da misure ripetute dello stesso tipo, compiute su ogni unità, in tempi successivi. Le misure ripetute sono molto frequenti in quasi tutti i settori scientifici in cui si ricorre a modelli statistici per condurre determinate analisi; biologia, economia e medicina sono solo alcuni esempi. I modelli adatti a descrivere dati di questo tipo sono numerosi e la letteratura scientifica in materia è molto vasta. Gran parte di questa è centrata sui modelli derivati dalla teoria normale classica, ma un grande sforzo è stato fatto per includere modelli in grado di descrivere adeguatamente dati di natura diversa, in particolare, dati categoriali e dati di durata (Armitage & Berry, 1996; Lindsey, 1993).

In questa tesi verranno presentati tre modelli che si collocano tra quelli in cui la variabile risposta, rilevata più volte sulle stesse unità, è di tipo continuo e distribuita normalmente: il primo di carattere generale, il secondo e il terzo di carattere più specifico, ma riconducibili al primo. Nell'ambito delle indagini statistiche verrà prestata particolare attenzione a quelle che si possono condurre basandosi sull'analisi della varianza (ANOVA) per misure ripetute, e si vedrà una loro applicazione in R.

Lo schema della tesi è il seguente. Nel Capitolo 1 vengono definite le misure ripetute e si focalizza l'attenzione sui principali vantaggi e svantaggi dei disegni per dati di questo tipo. Nel Capitolo 2 vengono discussi in modo approfondito il modello generale, il modello a un campione e il modello a campioni multipli, e vengono riportate le tabelle contenenti le grandezze necessarie a condurre le analisi basate sull'ANOVA per misure ripetute. Nel Capitolo 3 vengono esposti i comandi di R per fare inferenza in tali modelli e nel Capitolo 4, di carattere applicativo, viene affrontato un caso di studio. Infine, sono riportate le conclusioni riguardanti le analisi statistiche condotte.

Capitolo 1
Le misure ripetute

1.1 Introduzione

Il termine "*misure ripetute*" si riferisce a quei dati che vengono raccolti sulla stessa unità statistica in istanti temporali differenti o sotto diverse condizioni. Nell'ambito delle scienze medico-sanitarie, ad esempio, le diverse condizioni sotto le quali la variabile di interesse viene rilevata possono essere trattamenti distinti a cui lo stesso soggetto viene sottoposto. Inoltre, può essere che le diverse unità statistiche su cui vengono effettuate le misurazioni non siano soggetti singoli, ma insiemi di soggetti. Spesso ci si riferisce a misurazioni di questo tipo con il termine "*dati longitudinali*" (Davis, 2002).

I disegni per misure ripetute presentano alcuni vantaggi e svantaggi.

I principali vantaggi sono i seguenti:

- Il disegno sperimentale per misure ripetute è l'unico che permette di ottenere informazioni sull'evoluzione della variabile di interesse su un singolo soggetto.
- In questo tipo di disegno è possibile utilizzare campioni di dimensione inferiore a quella necessaria in un disegno sperimentale con una sola rilevazione. Infatti, ad esempio, se consideriamo quattro trattamenti o quattro istanti temporali per ognuno dei quali vogliamo disporre di dieci osservazioni, in un disegno sperimentale per misure ripetute sarà sufficiente considerare dieci soggetti, mentre ne servirebbero quaranta se facessimo un'unica rilevazione su ognuno di essi. Questo aspetto è rilevante quando non è facile reperire i soggetti da sottoporre a rilevazione o quando risulta particolarmente costoso reperirli.
- Poiché la variabilità tra i diversi soggetti può essere tenuta sotto controllo e quindi essere esclusa dall'errore sperimentale, rispetto a disegni *cross-section* con lo stesso numero e tipo di misurazioni, questo disegno produce spesso stimatori più efficienti dei parametri rilevanti.

Per quanto riguarda gli svantaggi, essi si riferiscono alle seguenti problematiche:

- Il primo difetto di questo tipo di analisi è legato alla dipendenza che sussiste tra le osservazioni sul medesimo soggetto. Questo problema viene chiamato "*effetto trascinamento*".
- In secondo luogo, le rilevazioni effettuate sullo stesso soggetto possono risentire di un *"effetto posizione"*, sulla base del quale, per esempio, la risposta di un soggetto a un certo trattamento messo in ultima posizione potrebbe essere diversa rispetto al caso in cui questo fosse messo in prima posizione. Una possibile soluzione a questo tipo di inconveniente è quella di randomizzare la sequenza con cui vengono somministrati i trattamenti indipendentemente dai soggetti.
- Infine, chi deve condurre l'analisi non sempre ha il controllo delle circostanze nelle quali vengono raccolti i dati; di conseguenza i dati possono non essere bilanciati o parzialmente incompleti.

Nonostante siano stati sviluppati molti approcci all'analisi delle misure ripetute, molti sono ristretti al caso in cui la variabile risposta è distribuita normalmente, i dati sono bilanciati e non ci sono valori mancanti (Davis, 2002; Wayne, 2007).

1.2 Notazione

La notazione impiegata per la rappresentazione dei dati raccolti attraverso misurazioni ripetute varia considerevolmente nella letteratura statistica (Davis, 2002). In generale, si può fare riferimento alla notazione riportata nella Tabella 1.1.

Unità	*Tempo/Trattamento*	*Indicatore di dato mancante*	*Risposta*	*Covariate*
1	*1*	δ_{11}	y_{11}	$x_{111}\ldots x_{11p}$
	.	.	.	.
	.	.	.	.
	.	.	.	.
	j	δ_{1j}	y_{1j}	$x_{1j1}\ldots x_{1jp}$
	.	.	.	.
	.	.	.	.
	.	.	.	.
	t_1	δ_{1t_1}	y_{1t_1}	$x_{1t_11}\ldots x_{1t_1p}$
i	*1*	δ_{i1}	y_{i1}	$x_{i11}\ldots x_{i1p}$
	.	.	.	.
	.	.	.	.
	.	.	.	.
	j	δ_{ij}	y_{ij}	$x_{1j1}\ldots x_{1jp}$
	.	.	.	.
	.	.	.	.
	.	.	.	.
	t_i	δ_{it_i}	y_{it_i}	$x_{1t_i1}\ldots x_{1t_ip}$
n	*1*	δ_{n1}	y_{n1}	$x_{n11}\ldots x_{n1p}$
	.	.	.	.
	.	.	.	.
	.	.	.	.
	j	δ_{nj}	y_{nj}	$x_{nj1}\ldots x_{njp}$
	.	.	.	.
	.	.	.	.
	.	.	.	.
	t_n	δ_{nt_n}	y_{nt_n}	$x_{nt_n1}\ldots x_{nt_np}$

TABELLA 1.1. Rappresentazione generale dei dati per misure ripetute.

Nella Tabella 1.1:

- $\boldsymbol{n}$ è il numero di soggetti sui quali vengono effettuate le misurazioni;
- $\boldsymbol{t_i}$ è il numero di misurazioni effettuate sul soggetto *i*-esimo*, i=1,…,n*;
- $\boldsymbol{y_{ij}}$ è il valore che la variabile risposta assume sul soggetto *i*-esimo al tempo (o occasione) *j*-esimo, *con j=1,…,*t_i e *i=1,…,n*;
- $\boldsymbol{p}$ è il numero di covariate;
- $\boldsymbol{x_{ij}}$*=(*x_{ij1}*,…,*x_{ijp}*)'* è il vettore delle covariate associato a y_{ij};
- $\boldsymbol{\delta_{ij}}$ è una variabile indicatrice tale che δ_{ij}*=1* se y_{ij} e $\boldsymbol{x_{ij}}$ sono osservate, δ_{ij}*=0* altrimenti (nel caso in cui ci siano valori mancanti per la generica riposta y_{ij} o per le componenti del vettore $\boldsymbol{x_{ij}}$).

Un caso particolare di questa rappresentazione è quello in cui t_1*=…=*t_n*=t*, ossia i dati sono bilanciati.

Una situazione che si verifica spesso è quella in cui le misurazioni ripetute sono effettuate su *s* sottopopolazioni o *gruppi* di soggetti, in corrispondenza di *t* istanti temporali o occasioni di misurazione comuni. In questo caso, può essere conveniente usare una rappresentazione alternativa a quella precedente, come quella riportata nella Tabella 1.2.

		Tempo				
Gruppo	*Soggetto*	*1*	. . .	*j*	. . .	*t*
1	*1*	y_{111}	. . .	y_{11j}	. . .	y_{11t}
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	i	y_{1i1}	. . .	y_{1ij}	. . .	y_{1it}
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	n_1	y_{1n_11}	. . .	y_{1n_1j}	. . .	y_{1n_1t}
						.
h	*1*	y_{h11}	. . .	y_{h1j}	. . .	y_{h1t}
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	i	y_{hi1}	. . .	y_{hij}	. . .	y_{hit}
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	n_h	y_{hn_h1}	. . .	y_{hn_hj}	. . .	y_{hn_ht}
s	*1*	y_{s11}	. . .	y_{s1j}	. . .	y_{s1t}
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	i	y_{si1}	. . .	y_{sij}	. . .	y_{sit}
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.
	n_s	y_{sn_s1}	. . .	y_{sn_sj}	. . .	y_{sn_st}

TABELLA 1.2. Rappresentazione dei dati per il caso di campioni multipli.

Nella Tabella 1.2:

- $\boldsymbol{n_h}$ è il numero di soggetti nel gruppo *h*, con *h=1,…,s*, con *n*=$\sum_{h=1}^{s} n_h$;
- $\boldsymbol{s}$ è il numero di gruppi (o sottopopolazioni), che possono essere identificati dagli *s* livelli di una singola covariata;
- $\boldsymbol{y_{hij}}$ rappresenta la risposta al tempo *j*-esimo per il soggetto *i*-esimo appartenente al gruppo *h*-esimo, con *j=1,…,t, i=1,…,*n_h e *h=1,…,s.*

Un secondo caso particolare è quello tipico di una situazione in cui le misure ripetute sono ottenute in *t* istanti temporali comuni su *n* soggetti distinti. In questo caso, i dati possono essere rappresentati secondo lo schema riportato nella Tabella 1.3, corrispondente a una rappresentazione matriciale di dimensioni *n x t*.

	Tempo				
Soggetto	*1*	. . .	*j*	. . .	*t*
1	y_{11}	. . .	y_{1j}	. . .	y_{1t}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
i	y_{i1}	. . .	y_{ij}	. . .	y_{it}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	y_{n1}	. . .	y_{nj}	. . .	y_{nt}

TABELLA 1.3. Rappresentazione dei dati per il caso a un campione.

Nella Tabella 1.3, $\boldsymbol{y_{ij}}$ denota la *j*-esima misurazione sull' *i-esimo* soggetto, con *j=1,…,t* e *i=1,…,n*. Inoltre, può essere definita una variabile indicatrice $\boldsymbol{\delta_{ij}}$, tale che δ_{ij}*=1* se y_{ij} è osservata e δ_{ij}*=0* altrimenti.

Nel prossimo capitolo vedremo come modellare dati di questo tipo, a seconda che ci si trovi nel caso a un campione o in quello a campioni multipli. Inoltre, una volta stabilito il

modello, vedremo quali analisi statistiche basate sull'analisi della varianza per misure ripetute si possono condurre.

Capitolo 2
L'analisi della varianza per misure ripetute

2.1 Il modello generale

Quando la variabile di interesse, che viene rilevata più volte sugli stessi soggetti, è di tipo continuo e segue una distribuzione normale, è possibile condurre analisi statistiche basate sull'*analisi della varianza (ANOVA) per misure ripetute.*

Assumiamo che una variabile continua sia osservata ad ogni istante temporale *j*, per ognuno degli *n* soggetti oggetto di studio. Denotiamo con y_{ij} la variabile risposta relativa al soggetto *i*-esimo al tempo *j*-esimo, per *i=1,…,n* e *j=1,…,t*. Il modello generale su cui si basa l'analisi della varianza per misure ripetute è il seguente

$$y_{ij}=\mu_{ij}+\pi_{ij}+e_{ij}.$$

Questo modello ha tre componenti:

- μ_{ij} è la media al tempo *j* per gli individui estratti casualmente dalla stessa popolazione dell'individuo *i*-esimo;
- π_{ij} è lo scostamento di y_{ij} da μ_{ij} in relazione al soggetto *i*-esimo; di conseguenza ipotizzando di ripetere più volte la rilevazione sul medesimo soggetto, y_{ij} avrebbe media pari a $\mu_{ij} + \pi_{ij}$;
- e_{ij} è lo scostamento di y_{ij} dalla sua media $\mu_{ij} + \pi_{ij}$ per l'individuo *i* al tempo *j*.

Il parametro μ_{ij} è chiamato *effetto fisso*, dal momento che il suo valore non dipende dal singolo soggetto oggetto di studio. Al contrario, il valore del parametro π_{ij} è legato allo specifico soggetto che è preso in considerazione e sul quale vengono effettuate le misurazioni. Dal momento che possiamo ritenere tale individuo estratto casualmente da una certa popolazione, π_{ij} viene chiamato *effetto casuale.* Gli e_{ij} sono termini di errore casuali. Poiché questo modello di riferimento include al suo interno sia un effetto fisso che un effetto casuale, viene detto *modello misto* (Davis, 2002; Lindsey, 1993).

2.1.1 Assunzioni

Per quanto riguarda l'effetto casuale e il termine di errore del modello, le assunzioni sono le seguenti:

1. Per un dato valore *j*, il valore atteso e la varianza dell'effetto casuale π_{ij} sono rispettivamente

   $$E(\pi_{ij})=0$$

   $$Var(\pi_{ij})=\sigma^2_{\pi j}$$

   Dunque le cause che potrebbero comportare una media diversa da zero per l'effetto casuale sono assorbite nella media di popolazione e la varianza è costante rispetto ai diversi soggetti.
2. Per un dato valore *j*, il valore atteso e la varianza del termine d'errore e_{ij} sono rispettivamente

   $$E(e_{ij})=0$$

   $$Var(e_{ij})=\sigma^2_{ej}$$

   Anche per il termine di errore la varianza è costante rispetto ai diversi soggetti.
3. In merito alla struttura di correlazione degli effetti casuali si ipotizza che

   $$Cov(\pi_{ij}, \pi_{i'j})=Cov(\pi_{ij}, \pi_{i'j'})=0,$$

   per $i \neq i'$ e $j \neq j'$, cioè gli effetti casuali relativi a differenti soggetti sono incorrelati.
4. Si assume inoltre che

   $$Cov(\pi_{ij}, \pi_{ij'})=\sigma_{\pi jj'}$$

   per $j \neq j'$. Questo significa che in relazione allo stesso soggetto (*within-subject*) la covarianza degli effetti è invariante rispetto al soggetto considerato.
5. Anche i termini di errore sono assunti incorrelati, quindi

   $$Cov(e_{ij}, e_{i'j'})=0,$$

   per $i \neq i'$ e $j \neq j'$.
6. Gli effetti casuali e gli errori sono incorrelati, dunque

   $$Cov(\pi_{ij}, e_{i'j'})=0,$$

   $\forall\, i, j, i', j'$.
7. L'effetto casuale π_{ij} e il termine di errore e_{ij} sono distribuiti normalmente.

Queste assunzioni generali sono spesso semplificate. In particolare:

- la varianza dell'effetto casuale può essere assunta costante nel tempo, ossia Var(π_{ij})= σ_{π}^2;
- similmente, Cov(π_{ij}, $\pi_{ij'}$)=σ_{π}, ossia è costante $\forall\ j, j'$;
- analogamente, Var(e_{ij})=σ_e^2 (la varianza del termine di errore del modello è costante nel tempo).

Per quanto riguarda le osservazioni y_{ij} le assunzioni usuali sono:

1. E(y_{ij})=μ_{ij}, dal momento che E(π_{ij})=E(e_{ij})=0.
2. Cov(y_{ij},$y_{i'j'}$)=$\delta_{ii'}\sigma_{\pi jj'}$+$\delta_{ii'}\delta_{jj'}\sigma_{ej}^2$=$\delta_{ii'}(\sigma_{\pi jj'}$+$\delta_{jj'}\sigma_{ej}^2)$, con $\delta_{ii'}$=1 se *i*=*i'*, $\delta_{ii'}$=0 altrimenti e $\delta_{jj'}$=1 se *j*=*j'*, $\delta_{jj'}$=0 altrimenti[1].
3. Le osservazioni y_{ij} e $y_{i'j}$, con $i \neq i'$, che si riferiscono cioè a due individui distinti, sono incorrelate.
4. La correlazione tra misure effettuate sullo stesso soggetto prende il nome di *correlazione intraclasse* ed è data da

$$\text{Corr}(y_{ij}, y_{ij'})=\frac{\sigma_{\pi jj'}}{(\sigma_{\pi jj}+\sigma_{ej}^2)(\sigma_{\pi j'j'}+\sigma_{ej'})}.$$

   Ovviamente, nel caso particolare in cui σ_{ej}^2=σ_e^2 e $\sigma_{\pi jj'}$=σ_{π}^2, e dunque nel caso in cui le ipotesi generali siano semplificate, si ha

$$\rho=\text{Corr}(y_{ij}, y_{ij'})=\frac{\sigma_{\pi}^2}{(\sigma_{\pi}^2+\sigma_e^2)},$$

   con $\rho \in [0,1]$.

[1] Per il dettaglio dei passaggi algebrici necessari a ottenere questo risultato si rinvia il lettore all'Appendice A.

2.2 Il modello a un campione

Prendiamo ora in considerazione la situazione in cui le misure ripetute sono ottenute da un singolo campione di soggetti, e quindi è conveniente rappresentare i dati come in Tabella 1.3 (Armitage & Berry, 1996; Davis, 2002; Wayne, 2007). In questo caso, il modello per l'analisi della varianza per misure ripetute è definito come

$$y_{ij}=\mu+\pi_i+\tau_j+e_{ij},$$

in cui *i=1,…,n* e *j=1,…,t*. Nell'equazione del modello si ha che

- y_{ij} è la risposta misurata sul soggetto *i*-esimo al tempo *j*-esimo;
- μ è la media dell'intera popolazione;
- π_i è l'effetto casuale relativo al soggetto *i*-esimo;
- τ_j è l'effetto fisso relativo al tempo *j*-esimo;
- e_{ij} è la componente erratica casuale caratteristica del soggetto *i* al tempo *j*.

In termini del modello generale esposto in precedenza, si ha che

- $\mu_{ij}=\mu+\tau_j$ (in questo caso l'indice *i* non è necessario);
- $\pi_{ij}=\pi_i$ (in questo caso lo scostamento di y_{ij} da μ_{ij} è costante nel tempo).

Per questo modello si assume che

1. $\pi_i \sim N(0, \sigma_\pi^2)$ i.i.d.
2. $e_{ij} \sim N(0, \sigma_e^2)$ i.i.d.
3. π_i e e_{ij} sono indipendenti
4. $\sum_{j=1}^{t} \tau_j=0$
5. Var(y_{ij})=Var($\mu+\pi_i+\tau_j+e_{ij}$)= $\sigma_\pi^2+\sigma_e^2$

   Cov(y_{ij}, $y_{i'j}$)=0 per $i \neq i'$

   Cov(y_{ij}, $y_{ij'}$)=σ_π^2 per $j \neq j'$

Sulla base di queste assunzioni risulta che la matrice di varianza e covarianza del vettore $\boldsymbol{y}_i$=(y_{i1},…,y_{it})' è la matrice *t x t* data da

$$\boldsymbol{\Sigma}=\begin{bmatrix} \sigma_\pi^2+\sigma_e^2 & \cdots & \sigma_\pi^2 \\ \vdots & \ddots & \vdots \\ \sigma_\pi^2 & \cdots & \sigma_\pi^2+\sigma_e^2 \end{bmatrix},$$

che è equivalente a

$$\boldsymbol{\Sigma}=(\sigma_\pi^2+\sigma_e^2)\begin{bmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{bmatrix},$$

dove $\rho=\frac{\sigma_\pi^2}{\sigma_\pi^2+\sigma_e^2}$=Corr($y_{ij}$, $y_{ij'}$).

Anche se le variabili casuali presenti nel modello sono indipendenti, i risultati delle misurazioni effettuate sullo stesso soggetto sono correlate. La matrice di varianza e covarianza risultante, $\boldsymbol{\Sigma}$, ha elementi uguali sulla diagonale principale e elementi uguali al di fuori di questa; quindi diciamo che $\boldsymbol{\Sigma}$ ha una *simmetria composta*. Questa particolare struttura della matrice di varianza e covarianza implica che la correlazione tra una qualsiasi coppia di osservazioni sulla stessa unità statistica è uguale indipendentemente dalla distanza tra le due osservazioni. Tuttavia, questa assunzione risulta fortemente restrittiva e spesso irrealistica, specialmente quando l'elemento rispetto al quale le misure sono ripetute è il tempo.

La tabella ANOVA su cui si basano le analisi in questo contesto è riportata nella Tabella 2.1[2].

Fonte	*SS*	*Df*	*MS*	*E[MS]*
Tempo/Trattamenti	SS_T	*t-1*	$MS_T=\frac{SS_T}{t-1}$	$\sigma_e^2+n\sigma_\tau^2$
Soggetti	SS_S	*n-1*	$MS_s=\frac{SS_S}{n-1}$	$\sigma_e^2+t\sigma_\pi^2$
Residua	SS_R	*(n-1)(t-1)*	$MS_R=\frac{SS_R}{(n-1)(t-1)}$	σ_e^2

TABELLA 2.1. Somme dei quadrati, gradi di libertà, medie quadratiche e medie quadratiche attese per l'ANOVA per misure ripetute a un campione.

[2] Esiste un modello alternativo proposto da Scheffé (1959). Si veda l'Appendice B.

La Tabella 2.1 riporta la somma dei quadrati (*Sum of Squares*, *SS*), i gradi di libertà (*Degrees of freedom*, *Df*), la media quadratica (*Mean Square*, *MS*) e la media quadratica attesa (*Expected Mean Square*, *E[MS]*) di ogni fonte di variabilità. Le somme dei quadrati riportate in tabella sono calcolate utilizzando le osservazioni y_{ij}, la media globale calcolata su tutti i valori osservati

$$\bar{y}_{..}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{t}y_{ij}}{nt},$$

le medie per ogni soggetto rispetto al tempo

$$\bar{y}_{i.}=\frac{\sum_{j=1}^{t}y_{ij}}{t}$$

e le medie per ogni istante temporale rispetto ai soggetti

$$\bar{y}_{.j}=\frac{\sum_{i=1}^{n}y_{ij}}{n}.$$

Le somme dei quadrati sono definite nel modo seguente:

$$SS_T=\sum_{i=1}^{n}\sum_{j=1}^{t}(\bar{y}_{.j}-\bar{y}_{..})^2=n\sum_{j=1}^{t}(\bar{y}_{.j}-\bar{y}_{..})^2,$$

$$SS_S=\sum_{i=1}^{n}\sum_{j=1}^{t}(\bar{y}_{i.}-\bar{y}_{..})^2=t\sum_{i=1}^{n}(\bar{y}_{i.}-\bar{y}_{..})^2,$$

$$SS_R=\sum_{i=1}^{n}\sum_{j=1}^{t}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2.$$

Infine, il simbolo σ_τ^2, presente nella colonna delle medie quadratiche attese, rappresenta una funzione degli effetti fissi τ_j.

In quest'ambito siamo interessati ad analizzare differenze in media nei valori della variabile risposta, in corrispondenza di diversi istanti temporali o diversi trattamenti. Questo può essere fatto conducendo un test statistico, il cui sistema di ipotesi è

$$H_0: \mu_h = \mu_k \ \forall\ h,k = 1,\dots,t$$
$$H_1: \overline{H}_0$$

e la statistica test è

$$F=\frac{MS_T}{MS_R}$$

che, se le assunzioni alla base del modello sono rispettate, sotto H_0 segue una distribuzione $F_{t-1,(n-1)(t-1)}$.

2.2.1 Condizione di sfericità

La matrice di varianza e covarianza definita da

$$\mathbf{\Sigma}=(\sigma_\pi^2+\sigma_e^2)\begin{bmatrix}1 & \cdots & \rho\\ \vdots & \ddots & \vdots\\ \rho & \cdots & 1\end{bmatrix},$$

ha una struttura chiamata "simmetria composta". La simmetria composta è una condizione sufficiente affinchè la statistica F, utilizzata per testare l'ipotesi nulla di uguaglianza delle medie in corrispondenza dei diversi istanti temporali, abbia una distribuzione $F_{t-1,(n-1)(t-1)}$, ma non è necessaria. La simmetria composta è infatti un caso particolare di una situazione più generale sotto la quale il test F è valido: la sfericità (Davis, 2002).

La condizione di sfericità può essere espressa in vari modi, tra cui:

1. le varianze delle differenze tra coppie di variabili sono uguali, cioè

   Var(y_{ij}-$y_{ij'}$) è costante per ogni *j,j'*;

2. ε=1, con

   $$\varepsilon=\frac{t^2(\bar{\sigma}_{ii}^2-\bar{\sigma}_{..}^2)^2}{(t-1)(S-2t\sum\bar{\sigma}_{i.}^2+t^2\bar{\sigma}_{..}^2)}, \qquad (2.1)$$

   dove $\bar{\sigma}_{ii}^2$ è la media dei valori sulla diagonale principale di **Σ,** $\bar{\sigma}_{..}^2$ è la media di tutti gli elementi di **Σ**, $\bar{\sigma}_{i.}^2$ è la media dei valori sulla riga *i*-esima di **Σ** e S è la somma dei quadrati degli elementi di **Σ**.

Se vale la condizione di simmetria composta, allora sia Var(y_{ij}) che Cov(y_{ij}, $y_{ij'}$) sono costanti per ogni *i* e *j*. Quindi, poiché

Var(y_{ij}-$y_{ij'}$)=Var(y_{ij})+ Var($y_{ij'}$)-2Cov(y_{ij}, $y_{ij'}$),

la condizione che le varianze delle differenze tra coppie di variabili siano uguali è soddisfatta. Ci sono anche altre situazioni in cui le condizioni di sfericità sopra riportate sono rispettate. Tuttavia, pur essendo la condizione di sfericità più generale rispetto a quella di simmetria composta, è difficile immaginare che si possano verificare situazioni in cui la condizione di sfericità sia soddisfatta e la matrice **Σ** abbia una struttura diversa da quella che presenta sotto la condizione di simmetria composta. In particolare, se le varianze sono uguali, allora le covarianze devono essere uguali se si vuole che la condizione di sfericità sia soddisfatta.

Wallenstein (1982) conclude che, a fini pratici, l'assunzione che dovrebbe essere fatta sulla matrice $\mathbf{\Sigma}$ è quella di simmetria composta.

Mauchly (1940) ha proposto un test per verificare se la condizione di sfericità è soddisfatta. Questo test ha una potenza limitata quando la dimensione del campione è piccola. Inoltre, quando il campione ha dimensione elevata, è probabile che ci sia evidenza a favore della condizione di sfericità anche se l'effetto sul test F è trascurabile. In aggiunta, è stato dimostrato che questo test è sensibile a scostamenti dall'ipotesi di normalità, per quanto riguarda la distribuzione della variabile risposta. In particolare, è conservativo per distribuzioni con code leggere e non conservativo per distribuzioni con code pesanti. E' anche molto sensibile agli *outlier*. A causa di queste proprietà non risulta essere di grande utilità pratica.

Quando la condizione di sfericità non sembra essere soddisfatta, una possibile soluzione consiste nel cambiare approccio all'ANOVA per misure ripetute. In questo caso, la statistica F si distribuisce approssimativamente come

$$F_{\varepsilon(t-1),\varepsilon(t-1)(n-1)},$$

dove ε è definita nella (2.1) ed è quindi una funzione della vera matrice di varianza e covarianza $\mathbf{\Sigma}$. Si può dimostrare che $\frac{1}{(t-1)} \leq \varepsilon \leq 1$.

Ci sono diversi approcci che si possono seguire per condurre test di verifica di ipotesi quando la condizione di sfericità è violata:

- Un primo approccio consiste nel correggere i gradi di libertà della distribuzione della statistica F ponendo $\varepsilon=\frac{1}{(t-1)}$. In questo modo, la distribuzione di riferimento per la statistica test sotto l'ipotesi nulla è $F_{1,n-1}$. Questo test risulta fortemente conservativo.
- Greenhouse e Geisser (1959) propongono di correggere i gradi di libertà della distribuzione della statistica F, sostituendo a ε una sua stima ottenuta a partire dallo stimatore $\hat{\varepsilon}$. Questo stimatore si ottiene a partire dall'equazione (2.1), nella quale la matrice $\mathbf{\Sigma}$ viene rimpiazzata con la sua equivalente campionaria $\mathbf{S}$. Lo stimatore $\hat{\varepsilon}$ così ottenuto è lo stimatore di massima verosimiglianza di ε.

L'utilizzo di questo stimatore comporta ugualmente che il test risultante è conservativo, e d'altro canto questo approccio presenta una distorsione elevata quando ε>0.75 e $n<2t$.

- Huynh e Feldt (1976) hanno proposto lo stimatore

$$\tilde{\varepsilon}=\min\left(1,\frac{n(t-1)\hat{\varepsilon}-2}{(t-1)(n-1-(t-1)\hat{\varepsilon})}\right)$$

che si basa su stimatori non distorti del numeratore e del denominatore di ε ed è meno distorto di $\hat{\varepsilon}$. Si può dimostrare che $\tilde{\varepsilon} \geq \hat{\varepsilon}$. Infine, anche se $\hat{\varepsilon}$ è migliore quando $\varepsilon \leq 0.5$, $\tilde{\varepsilon}$ è preferibile per $\varepsilon \geq 0.7$. Tuttavia, nella pratica il valore di ε è ignoto.

Il seguente schema, proposto da Greenhouse e Geisser (1959), mostra la procedura da seguire quando si utilizza l'ANOVA per misure ripetute nel caso in cui i soggetti facciano parte di un unico campione. I tre passi in cui si articola la procedura sono i seguenti:

1. Si conduce il test F nell'ipotesi in cui le assunzioni alla base del modello siano soddisfatte.
2. Se il test non è significativo, non si rifiuta H_0.
3. Se il test è significativo, si conduce il test usando $\varepsilon=\frac{1}{(t-1)}$, che porta alla distribuzione $F_{1,n-1}$ per la statistica F. Quindi:
    - Se questo test è significativo, si rifiuta H_0.
    - Se questo test non è significativo, si stima ε con $\hat{\varepsilon}$ ($\tilde{\varepsilon}$) e si conduce un test approssimato.

2.3 Il modello a campioni multipli

Supponiamo che le misurazioni ripetute in corrispondenza di *t* istanti temporali (o sotto *t* condizioni differenti) siano ottenute a partire dai soggetti appartenenti a *s* gruppi distinti (Davis, 2002; Wayne, 2007).

Sia n_h il numero di soggetti appartenenti al gruppo *h* e sia *n*=$\sum_{h=1}^{s} n_h$. Inoltre, sia y_{hij} il valore della variabile risposta al tempo *j* per il soggetto *i* del gruppo *h*, per *h=1,…,s*, *i=1,…,*n_h e *j=1,…,t*. In questa situazione, i dati possono essere rappresentati come in Tabella 1.2 e il modello adatto a rappresentarli può essere espresso come

$$y_{hij} = \mu + \gamma_h + \tau_j + (\gamma\tau)_{hj} + \pi_{i(h)} + e_{hij},$$

dove:

- μ è la media dell'intera popolazione;
- γ_h è l'effetto fisso del gruppo *h* con $\sum_{h=1}^{s} \gamma_h = 0$;
- τ_j è l'effetto fisso del tempo *j* con $\sum_{j=1}^{t} \tau_j = 0$;
- $(\gamma\tau)_{hj}$ è l'effetto fisso dell'interazione tra il gruppo *h* e il tempo *j* con $\sum_{h=1}^{s}(\gamma\tau)_{hj} = \sum_{j=1}^{t}(\gamma\tau)_{hj} = 0$;
- $\pi_{i(h)}$ è l'effetto casuale del soggetto *i*-esimo appartenente al gruppo *h*-esimo con $\pi_{i(h)} \sim N(0, \sigma_\pi^2)$ i.i.d.;
- e_{hij} è un termine di errore casuale con $e_{hij} \sim N(0, \sigma_e^2)$ i.i.d.

Confrontando questa formulazione del modello con quella generale, si trovano le seguenti relazioni:

- $\mu_{ij} = \mu + \gamma_h + \tau_j + (\gamma\tau)_{hj}$
- $\pi_{ij} = \pi_{i(h)}$
- $e_{ij} = e_{hij}$

La tabella ANOVA su cui si basano le analisi in questo contesto è riportata nella Tabella 2.2.

Fonte	*SS*	*Df*	*E[MS]*
Gruppo	SS_G	*s-1*	$\sigma_e^2 + t\sigma_\pi^2 + D_G$
Soggetti (Gruppo)	$SS_{S(G)}$	*n-s*	$\sigma_e^2 + t\sigma_\pi^2$
Tempo	SS_T	*t-1*	$\sigma_e^2 + D_T$
Gruppo x Tempo	SS_{GT}	*(s-1)(t-1)*	$\sigma_e^2 + D_{GT}$
Residua	SS_R	*(n-s)(t-1)*	σ_e^2

TABELLA 2.2. Somme dei quadrati, gradi di libertà e medie quadratiche attese per l'ANOVA per misure ripetute a campioni multipli.

La Tabella 2.2 riporta le somme dei quadrati (*SS*), i gradi di libertà (*Df*), e le medie quadratiche attese (*E[MS]*) per ogni fonte di variabilità. Nella colonna denominata *E[MS]*, le quantità D_G, D_T e D_{GT} indicano, rispettivamente, le differenze tra gruppi, istanti temporali e le interazioni tra gruppo e tempo.

Le somme dei quadrati si basano sulla seguente decomposizione della quantità $y_{hij} - \bar{y}_{\cdots}$, che rappresenta lo scarto di ogni osservazione dalla media di tutte le osservazioni.

In particolare, vale

$$y_{hij} - \bar{y}_{\cdots} = (\bar{y}_{h\cdot\cdot} - \bar{y}_{\cdots}) + (\bar{y}_{hi\cdot} - \bar{y}_{h\cdot\cdot}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdots})$$
$$+(\bar{y}_{h\cdot j} - \bar{y}_{h\cdot\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdots}) + (y_{hij} - \bar{y}_{h\cdot j} - \bar{y}_{hi\cdot} + \bar{y}_{h\cdot\cdot}),$$

dove

$$\bar{y}_{\cdots} = \frac{\sum_{h=1}^{s}\sum_{i=1}^{n_h}\sum_{j=1}^{t} y_{hij}}{nt}$$

è la media di tutte le osservazioni,

$$\bar{y}_{h\cdot\cdot} = \frac{\sum_{i=1}^{n_h}\sum_{j=1}^{t} y_{hij}}{n_h t}$$

è la media delle osservazioni per il gruppo *h*,

$$\bar{y}_{\cdot j} = \frac{\sum_{h=1}^{s}\sum_{i=1}^{n_h} y_{hij}}{n}$$

è la media al tempo *j*,

$$\bar{y}_{h.j} = \frac{\sum_{i=1}^{n_h} y_{hij}}{n_h}$$

è la media per il gruppo *h* al tempo *j*, e

$$\bar{y}_{hi.} = \frac{\sum_{j=1}^{t} y_{hij}}{t}$$

è la media per l'*i*-esimo soggetto appartenente al gruppo *h*.

Le somme dei quadrati sono quindi definite nel modo seguente

$$SS_G = \sum_{h=1}^{s} \sum_{i=1}^{n_h} \sum_{j=1}^{t} (\bar{y}_{h..} - \bar{y}_{...})^2 = t \sum_{h=1}^{s} n_h (\bar{y}_{h..} - \bar{y}_{...})^2,$$

$$SS_{S(G)} = \sum_{h=1}^{s} \sum_{i=1}^{n_h} \sum_{j=1}^{t} (\bar{y}_{hi.} - \bar{y}_{h..})^2 = t \sum_{h=1}^{s} \sum_{i=1}^{n_h} (\bar{y}_{hi.} - \bar{y}_{h..})^2,$$

$$SS_T = \sum_{h=1}^{s} \sum_{i=1}^{n_h} \sum_{j=1}^{t} \left(\bar{y}_{..j} - \bar{y}_{...}\right)^2 = n \sum_{j=1}^{t} n_h \left(\bar{y}_{..j} - \bar{y}_{...}\right)^2,$$

$$SS_{GT} = \sum_{h=1}^{s} \sum_{i=1}^{n_h} \sum_{j=1}^{t} \left(\bar{y}_{h.j} - \bar{y}_{h..} - \bar{y}_{..j} + \bar{y}_{...}\right)^2,$$

$$SS_R = \sum_{h=1}^{s} \sum_{i=1}^{n_h} \sum_{j=1}^{t} \left(y_{hij} - \bar{y}_{h.j} - \bar{y}_{hi.} + \bar{y}_{h..}\right)^2.$$

La statistica F per verificare se ci sono differenze tra i gruppi è data da

$$F = \frac{MS_G}{MS_{S(G)}} = \frac{SS_G/(s-1)}{SS_{S(G)}/(n-s)},$$

che sotto l'ipotesi nulla segue una distribuzione $F_{s-1,n-s}$. Questo test richiede di assumere che le matrici di varianza e covarianza definite all'interno di ciascun gruppo siano uguali tra loro.

La statistica F per verificare se ci sono differenze tra istanti temporali distinti è data da

$$F = \frac{MS_T}{MS_R} = \frac{SS_T/(t-1)}{SS_R/[(n-s)(t-1)]},$$

che sotto l'ipotesi nulla si distribuisce come una $F_{t-1,(n-s)(t-1)}$.

Analogamente, per verificare se l'interazione tra il gruppo e il tempo è significativa si può usare la statistica

$$F = \frac{MS_{GT}}{MS_R} = \frac{SS_{GT}/[(s-1)(t-1)]}{SS_R/[(n-s)\,(t-1)]},$$

che segue una distribuzione $F_{(s-1)(t-1),(n-s)(t-1)}$ sotto l'ipotesi nulla.

Tutti questi test richiedono di assumere che le matrici di varianza e covarianza definite all'interno di ciascun gruppo siano uguali e che la condizione di sfericità sia soddisfatta.

Un modello alternativo a quello appena presentato include un effetto casuale aggiuntivo, quello associato all'interazione tra soggetto e tempo, e questo effetto si assume incorrelato con l'effetto casuale associato al singolo soggetto. Inoltre, anche se le medie quadratiche attese sono diverse rispetto al modello precedente, le somme dei quadrati e le statistiche test sono identiche.

Nel capitolo successivo verranno presentati i comandi di R che permettono di stimare i modelli sopra esposti e condurre le analisi statistiche basate sull'ANOVA per misure ripetute.

Capitolo 3
I comandi in R

3.1 La funzione *lme*

Come è stato illustrato nel capitolo precedente, il modello per l'analisi della varianza per misure ripetute è un modello misto; questo significa che include nella specificazione sia un effetto fisso che un effetto casuale. In R è possibile effettuare inferenza in tale modello utilizzando la funzione *lme*. Questa funzione si trova nella libreria *nlme* (*non-linear mixed effects*), che fornisce i comandi per adattare modelli misti, lineari e non lineari.
La sintassi della funzione *lme* è la seguente:

```
lme(fixed, data, random, correlation, weights,
      subset, method, na.action, control,
      contrasts = NULL, keep.data = TRUE).
```

Gli argomenti di questa funzione sono spiegati in dettaglio nella documentazione in R a cui è possibile accedere attraverso il comando `help(lme)`. I due argomenti principali sono `fixed` e `random`. L'argomento `fixed` si riferisce agli effetti fissi del modello e si esprime con una formula lineare composta da due elementi separati dal simbolo `~`. A sinistra del simbolo `~` deve essere indicata la variabile risposta e a destra, separate dall'operatore `+`, le variabili esplicative che determinano gli effetti fissi.

L'argomento `random` è esprimibile, come `fixed`, con una formula. La formula è del tipo

`~x1+...+xn|g1/.../gm`,

in cui `x1+...+xn` specificano il modello per gli effetti casuali e `g1/.../gm` specificano la struttura del raggruppamento (`m` può essere uguale a 1: in questo caso non è necessario introdurre nella formula il simbolo `/`). La formula per gli effetti casuali deve essere ripetuta per tutti il livelli di raggruppamento nel caso il loro numero sia più grande di uno.

L'esempio seguente (Fox, 2002) mostra come usare il comando *lme* per il *dataset* `MathAchieve`:

```
> bryk.lme.1<-lme(mathach~meanses*cses+sector*cses,
+ random=~cses|school, data=Bryk).
```

In questo caso, la variabile `mathach` è quella da modellare attraverso le variabili `meanses`, `cses`, e `sector` e gli effetti casuali legati alla variabile `cses` e all'intercetta definita dalla variabile `school`. Queste variabili sono rispettivamente il punteggio ottenuto dagli studenti di varie scuole in un test di matematica, la media dello stato socio-economico della famiglia degli studenti in ciascuna scuola, lo stato socio-economico della famiglia di ciascuno studente, il tipo di scuola, cattolica o pubblica e il numero identificativo di ciascuna di esse.

I modelli lineari misti possono essere rappresentati in forme diverse, ma tra loro equivalenti. La funzione *lme* si basa sul modello espresso nella forma proposta da Laird e Ware (1982). In forma matriciale compatta, il modello può essere rappresentato come

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\gamma}_i + \boldsymbol{\varepsilon}_i$$
$$\boldsymbol{\gamma}_i \sim N_g(\mathbf{0}_g, \boldsymbol{B}) \text{ i.i.d.}$$
$$\boldsymbol{\varepsilon}_i \sim N_{t_i}(\mathbf{0}_{t_i}, \boldsymbol{W}_i) \text{ i.i.d.}$$

per *i=1,…,n*. In particolare:

- $\boldsymbol{y}_i = (y_{i1}, \dots, y_{it_i})'$ è il vettore $t_i \times i$ delle risposte per il soggetto *i*-esimo;
- $\boldsymbol{X}_i$ è la matrice di disegno $t_i \times b$ del modello, relativa al soggetto *i*-esimo;
- $\boldsymbol{\beta}$ è il vettore $b \times 1$ dei coefficienti di regressione;
- $\boldsymbol{\gamma}_i$ è il vettore $g \times 1$ degli effetti casuali per il soggetto *i*-esimo;
- $\boldsymbol{Z}_i$ è la matrice di disegno $t_i \times g$ relativa agli effetti casuali;
- $\boldsymbol{\varepsilon}_i$ è il vettore $t_i \times 1$ degli errori entro il soggetto (*within-subject errors*) *i*-esimo;
- $\boldsymbol{\gamma}_i$, $\boldsymbol{\varepsilon}_i$ sono vettori indipendenti;
- $\boldsymbol{B}$ è la matrice $g \times g$ di varianza e covarianza per gli effetti casuali;
- $\boldsymbol{W}_i$ è la matrice $t_i \times t_i$ di varianza e covarianza per gli errori relativi al soggetto *i*-esimo.

Utilizzando il comando *lme*, si ottiene una stima del modello basata sul metodo della massima verosimiglianza ristretta (*REML, restricted maximum likelihood*)[3].

Per visualizzare le informazioni prodotte dall'applicazione di questa funzione, è sufficiente l'uso del comando `summary`. Applicandolo a oggetti di tipo *lme* si ottiene un output che si articola in diverse sezioni (Fox, 2002).

Un esempio di output ottenuto utilizzando il comando `summary`:

```
> summary(bryk.lme.1)
Linear mixed-effects model fit by REML
  Data: Bryk
    AIC   BIC logLik
 46525 46594 -23252

Random effects:
 Formula: ~cses | school
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 1.541177 (Intr)
Cses        0.018174 0.006
Residual    6.063492

Fixed effects: mathach ~ meanses * cses + sector * cses
                         Value   Std.Error     DF  t-value p-value
(Intercept)            12.1282     0.19920   7022   60.886 <.0001
Meanses                 5.3367     0.36898    157   14.463 <.0001
Cses                    2.9421     0.15122   7022   19.456 <.0001
sectorCatholic          1.2245     0.30611    157    4.000  1e-04
meanses:cses            1.0444     0.29107   7022    3.588  3e-04
cses:sectorCatholic    -1.6421     0.23312   7022   -7.044 <.0001
  Correlation:
                       (Intr)  meanss   cses  sctrCt  mnss:c
meanses                 0.256
cses                    0.000   0.000
sectorCatholic         -0.699  -0.356  0.000
meanses:cses            0.000   0.000  0.295   0.000
cses:sectorCatholic     0.000   0.000 -0.696   0.000   -0.351

Standardized Within-Group Residuals:
      Min         Q1        Med         Q3       Max
-3.170106  -0.724877   0.014892   0.754263  2.965498

Number of Observations: 7185
Number of Groups: 160.
```

[3] Il metodo della massima verosimiglianza ristretta non è argomento di questa tesi; per una discussione dettagliata su questo tema, si rinvia il lettore a Davis (2002).

- La prima sezione fornisce i valori di due criteri di informazione, l'AIC (*Akaike Information Criterion*) e il BIC (*Bayesian Information Criterion*), quantità utilizzate per confrontare modelli diversi. Inoltre, viene fornito il logaritmo della verosimiglianza ristretta massimizzata.
- La seconda sezione mostra le stime della varianza e della covarianza degli effetti casuali presenti nel modello, sotto forma di deviazioni standard e correlazioni, oltre a una stima della deviazione standard del termine di errore.
- La sezione successiva riporta una tabella collegata con gli effetti fissi del modello, ed è simile a quella che si ottiene applicando il comando `summary` a oggetti di tipo *lm*. Essa fornisce la stima dei coefficienti β che rappresentano gli effetti fissi del modello, una stima degli errori standard degli stimatori dei coefficienti β, i gradi di libertà della statistica *t*, il suo valore e il valore del *p-value* ad esso associato.
- La quarta sezione, denominata `Correlation`, fornisce le correlazioni campionarie tra le stime dei coefficienti che costituiscono gli effetti fissi. Pur non essendo di interesse diretto, quando queste quantità assumono valori elevati mettono in dubbio l'effettiva adeguatezza del modello.
- Nell'ultima sezione vengono fornite alcune informazioni sui residui standardizzati entro i soggetti/gruppi, oltre al numero delle osservazioni e il numero di soggetti/gruppi.

3.2 La funzione *lmer*

Per adattare un modello lineare misto è possibile ricorrere anche a un'altra funzione: la funzione *lmer*. Si può accedere a questa funzione attraverso la libreria *lme4*. Rispetto alla funzione *lme*, la funzione *lmer* permette di adattare un maggior numero di modelli. Inoltre, risulta essere più affidabile e veloce. Tuttavia, cambia leggermente la modalità di specificazione del modello (Bates, 2005).

La sintassi della funzione *lmer* è la seguente:

```
lmer(formula, data, family = NULL, REML = TRUE,
control = list(), start = NULL, verbose = FALSE,
doFit = TRUE, subset, weights, na.action, offset,
contrasts = NULL, model = TRUE, x = TRUE, ...).
```

Anche in questo caso è possibile accedere a una spiegazione dettagliata degli argomenti della funzione, digitando in R il comando `help(lmer)`.
Come per la maggior parte delle funzioni utilizzate per stimare modelli in ambiente R, i primi due argomenti sono costituiti da una `formula` con la quale viene specificato il modello e un oggetto, `data`, di tipo *data frame* contenente le variabili specificate in `formula`. La stima del modello avviene sulla base dei dati contenuti in questo argomento, che pur essendo opzionale, è consigliabile specificare. Se l'argomento `data` non è specificato, le variabili vengono prese dall'ambiente da cui il comando *lmer* è invocato. A differenza del comando *lme*, in cui la sintassi separa l'argomento che riguarda gli effetti fissi da quello che riguarda gli effetti casuali, nel caso della funzione *nlmer* la sintassi li include entrambi in `formula`. Questo argomento si esprime attraverso una espressione in cui a destra del simbolo ~ viene indicata la variabile risposta e a destra, separati dal segno +, gli effetti fissi del modello e quelli casuali. Questi ultimi sono riportati tra parentesi, in modo analogo a quanto avviene utilizzando il comando *lme*: `(~x1+...+xn|g1/.../gm)`.
Un esempio riguardante l'uso della funzione *lmer* (Fox, 2002):

```
> ex1<-lmer(SCIENCE ~ URBAN + (URBAN|SCHOOL), example.data).
```

Con questo comando la variabile `SCIENCE` è stimata attraverso la variabile `URBAN` e gli effetti casuali legati alla variabile `URBAN` e all'intercetta definita dalla variabile `SCHOOL`. Queste variabili sono rispettivamente la valutazione ottenuta in un test di scienze dagli studenti di varie scuole, la zona in cui si trova ciascuna scuola, urbana o rurale e il suo codice identificativo. Una volta ottenuta la stima del modello, è possibile visualizzare una serie di informazioni ricorrendo all'usuale comando `summary` o, in alternativa, al comando `print`. Usando il comando `summary`, ad esempio, le informazioni che si

ottengono possono essere racchiuse in quattro sezioni e il contenuto di ognuna di esse può essere riassunto nel modo seguente:

- la prima sezione contiene una descrizione del modello adattato;
- la seconda sezione contiene alcune statistiche caratterizzanti tale modello;
- la terza sezione riporta alcune delle caratteristiche degli effetti casuali;
- la quarta sezione contiene le stime dei parametri legati agli effetti fissi e la correlazione campionaria tra le stime dei coefficienti che costituiscono gli effetti fissi.

L'esempio seguente mostra nel dettaglio le informazioni sopra citate, ottenute con il comando `summary`:

```
> summary(ex1)
Linear mixed-effects model fit by REML
Formula: SCIENCE ~ URBAN + (URBAN | SCHOOL)
   Data: example.data
      AIC      BIC    logLik MLdeviance REMLdeviance
 424.1713 442.6223 -206.0857   413.2216     412.1713
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 SCHOOL   (Intercept) 113.60403 10.65852
          URBAN         0.25200  0.50200 -0.625
 Residual               0.27066  0.52025
# of obs: 160, groups: SCHOOL, 16

Fixed effects:
            Estimate Std. Error  DF t value  Pr(>|t|)
(Intercept)  22.39124    2.71703 158  8.2411 6.152e-14 ***
URBAN        -0.86700    0.12981 158 -6.6790 3.884e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Correlation of Fixed Effects:
      (Intr)
URBAN -0.641.
```

Come il lettore avrà certamente notato, questo output risulta molto simile a quello che si ottiene nel caso in cui si usi il comando `summary` applicato a un oggetto di tipo *lme*.

3.3 La funzione *aov*

Quando si è interessati a condurre un'analisi basta sulla ANOVA per misure ripetute è possibile ricorrere alla funzione *aov*. Questa funzione è definita nel modo seguente

```
aov(formula, data = NULL, projections = FALSE, qr = TRUE,
                    contrasts = NULL, ...).
```

Il comando `help(aov)` consente, ancora una volta, di accedere a una spiegazione dettagliata degli argomenti della funzione.

Applicando il comando `summary` alla funzione `aov`, applicata a sua volta a un oggetto di tipo *lme*, si ottiene il seguente output (Venables & Ripley, 1999)

```
> summary(aov.ex2)
             Df  Sum Sq Mean Sq F value Pr(>F)
Gender        1  76.562  76.562  2.9518 0.1115
Dosage        1   5.062   5.062  0.1952 0.6665
Gender:Dosage 1   0.063   0.063  0.0024 0.9617
Residuals    12 311.250  25.938.
```

Come si può notare, *aov* permette di ottenere la tabella ANOVA per misure ripetute con i valori delle grandezze necessarie a condurre l'analisi. In particolare, all'interno di questa tabella sono contenute le somme dei quadrati (*SS*), i gradi di libertà (*Df*), le medie quadratiche (*MS*) e i valori dei test F con i relativi *p-value*.

3.4 Il comando *mauchly.test*

Una condizione importante alla base del modello per l'analisi della varianza per misure ripetute è quella di sfericità. Questa condizione è una generalizzazione di quella di simmetria composta relativa alla matrice di varianza e covarianza $\boldsymbol{\Sigma}$ delle osservazioni y_{ij} (Cap. 2, Par. 2.2.1). Quando la condizione di sfericità è rispettata, e dunque lo è anche quella di simmetria composta, il test F per verificare l'ipotesi nulla di uguaglianza delle medie in corrispondenza dei diversi istanti temporali è valido. In R è possibile testare questa condizione utilizzando un apposito comando. Il comando *mauchly.test* permette di

condurre il test di Mauchly per verificare se la condizione di sfericità è soddisfatta o meno. La sintassi è la seguente

```
mauchly.test(object, Sigma = diag(nrow = p),
T = Thin.row(proj(M) - proj(X)), M = diag(nrow = p), X = ~0,
idata = data.frame(index = seq_len(p)), ...).
```

In generale, il test che si ottiene con questo comando verifica se una matrice di varianza e covarianza di un modello lineare multivariato, si può assumere proporzionale a una matrice data. Nell'ambito dei disegni per misure ripetute, la matrice data è quella identica. Facendo riferimento alla sintassi sopra esposta, l'argomento `object` è un oggetto che rappresenta una stima della matrice di varianza e covarianza per un modello lineare multivariato, e l'argomento `X` è la matrice a cui la matrice in `object` deve essere proporzionale.

Attraverso il comando `help(mauchly.test)` è possibile ottenere una spiegazione dei singoli argomenti di questa funzione.

Un esempio inerente all'uso del comando *mauchly.test* per un disegno per misure ripetute è il seguente

```
> mauchly.test(vcmatrix, X=~1).
```

Nel prossimo capitolo vedremo una applicazione in R del modello lineare a effetti misti e delle analisi basate sull'ANOVA per misure ripetute che si possono condurre utilizzando i comandi sopra esposti.

Capitolo 4
Un caso di studio

4.1 Il *dataset* e le variabili

In questo capitolo si presenta un'applicazione del modello misto e dell'analisi della varianza per misure ripetute a un caso di studio. Lo scopo di questa applicazione è verificare l'efficacia di due trattamenti a cui un individuo, che ha subito un infarto, può essere sottoposto, allo scopo di garantirgli il recupero delle proprie capacità motorie. I pazienti oggetto di indagine sono distinti in due sottogruppi, sulla base del trattamento a cui sono sottoposti durante la riabilitazione. Oltre a valutare l'efficacia di ognuno dei due trattamenti, e quindi valutare se c'è differenza nella capacità motoria del singolo individuo tra prima e dopo la riabilitazione, si è interessati a confrontarli. Infatti, un gruppo di pazienti è stato sottoposto a un trattamento tradizionale, un altro, a un trattamento più innovativo con realtà virtuale (Piron *et al*., 2010). Dunque, lo scopo dell'indagine è anche valutare se c'è una differenza in media tra il gruppo di pazienti sottoposti al trattamento con realtà virtuale e quelli sottoposti al trattamento tradizionale.

Il *dataset* `Retrospettiva RFVE` (Tabella 4.1) include misurazioni relative ad uno studio sull'apprendimento motorio di un gruppo di 376 pazienti, alcuni dei quali (113) sottoposti al trattamento tradizionale e altri (263) al trattamento con realtà virtuale. Le variabili presenti nel *dataset* sono:

- la variabile `trattamento`, che è una variabile che assume valore 1 se il trattamento a cui è sottoposto il paziente è di tipo sperimentale e valore 2 se invece il trattamento è di tipo tradizionale;
- la variabile `soggetto`, che identifica i pazienti attraverso una numerazione progressiva e assume valori compresi tra 1 e 376;
- la variabile `tempo`, che è una variabile binaria e assume valore 0 oppure valore 1, a seconda che la rilevazione della capacità motoria del paziente sia effettuata al momento del ricovero o al termine della riabilitazione;

- la variabile `FIM` (*Functional Independence Measure*), che è la variabile di principale interesse, che si vuole modellare. Essa rappresenta una scala sull'autonomia del paziente, con valori da 0 (non autosufficienza completa) a 126 (completa autonomia).

Il *dataset* si presenta nella forma riportata in Tabella 4.1.

trattamento	soggetto	tempo	FIM
1	1	0	122
1	2	0	122
1	3	0	89
1	4	0	116
1	5	0	120
1	6	0	102
1	7	0	125
1	8	0	65
1	9	0	90
1	10	0	27
.	.	.	.
.	.	.	.
.	.	.	.
1	376	0	89
1	1	1	122
.	.	.	.
.	.	.	.
.	.	.	.
1	367	1	100
1	368	1	106
1	369	1	103
1	370	1	99
2	371	1	73
1	372	1	86
1	373	1	86
1	374	1	96
1	375	1	109
1	376	1	89

TABELLA 4.1. *Dataset* `Retrospettiva RFVE`.

4.2 Analisi preliminari

Come esposto nel precedente paragrafo, i pazienti oggetto di indagine sono classificati in due sottogruppi, a seconda del trattamento a cui sono stati sottoposti durante la riabilitazione. Dal momento che si è interessati a studiare l'efficacia dei due trattamenti è utile, in prima battuta, osservare i *boxplot* relativi ad ognuno di essi. Questi grafici sono riportati in Figura 4.1 e in Figura 4.2 e consentono di confrontare in modo chiaro le misurazioni effettuate prima e dopo la terapia.

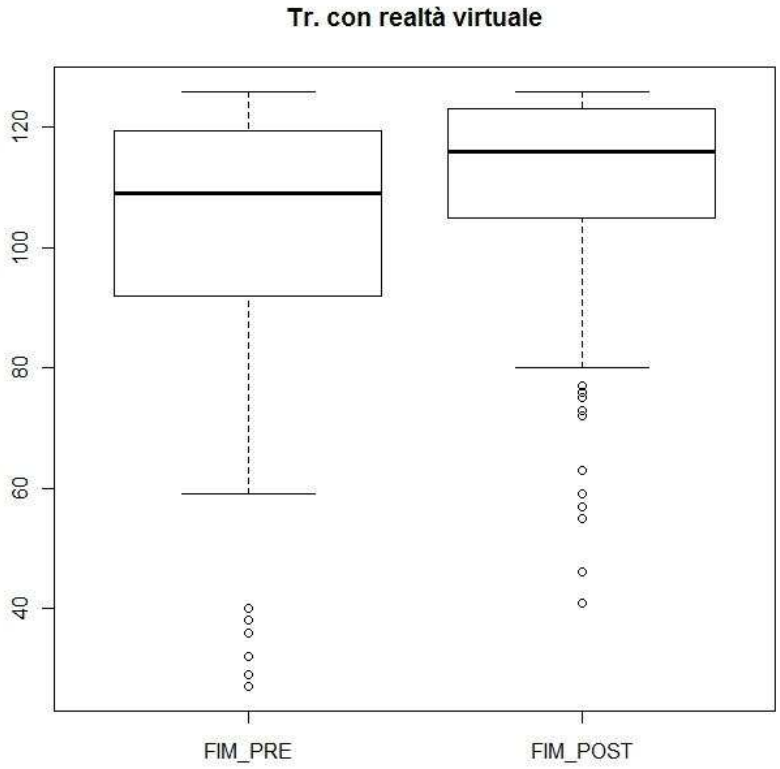

FIGURA 4.1. *Boxplot* della variabile `FIM`, prima e dopo il trattamento con realtà virtuale.

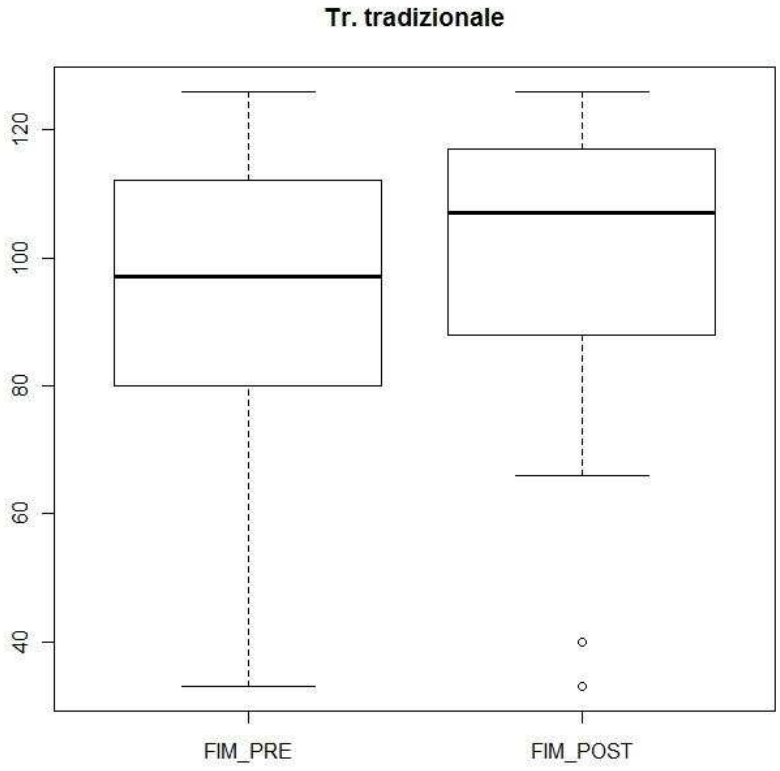

FIGURA 4.2. *Boxplot* della variabile `FIM`, prima e dopo il trattamento tradizionale.

Come si può osservare sia in Figura 4.1 che in Figura 4.2, i quantili e la mediana relativi alle misurazioni post-trattamento sono situati più in alto rispetto ai quantili e alla mediana relativi alle misurazioni pre-trattamento. Questo lascia supporre che ci sia stato un miglioramento da parte dei pazienti in termini di capacità motoria, indipendentemente dal trattamento a cui sono stati sottoposti durante la riabilitazione.

Tale aspetto può essere evidenziato anche considerando i valori di alcune statistiche descrittive. La Tabella 4.2 riporta i valori delle medie (e delle deviazioni standard), mantenendo distinte le misurazioni effettuate prima della terapia da quelle effettuate dopo averla completata e separando i due tipi di trattamento.

	Tempo	
Trattamento	Pre-trattamento	Post-trattamento
Realtà virtuale	103.21 (20.68)	110.75 (16.37)
Tradizionale	94.95 (21.39)	101.94 (19.06)

TABELLA 4.2. Medie (e deviazioni standard) per la variabile `FIM` prima e dopo ciascun trattamento.

Sulla base dei dati riportati in Tabella 4.2, si può affermare che entrambi i trattamenti abbiano sortito un effetto positivo sui pazienti, permettendogli di recuperare parte della loro capacità motoria in seguito all'infarto subito. Di conseguenza, entrambi i trattamenti risultano efficaci. È possibile averne conferma anche conducendo un test *t* di Student per dati appaiati. Prima di condurre questo test vale la pena osservare che i dati raccolti hanno una distribuzione asimmetrica (si vedano i *boxplot* riportati in Figura 4.1 e in Figura 4.2). Questa caratteristica è in contrasto con l'ipotesi alla base del test che provengano da una popolazione normale, tuttavia, in modo coerente con quanto avviene in letteratura (Piron *et al*., 2010), si assume ugualmente la normalità per questi dati. Ad ogni modo, le conclusioni a cui conduce il test *t* di Student sono le stesse a cui porta il test non parametrico di Wilcoxon, che non richiede di assumere che i dati provengano da una distribuzione normale. In questo caso il valore osservato della statistica test *t* di Student e il corrispondente *p-value*, sono riportati in Tabella 4.3.

Trattamento	**Staistica *t***	***p-value***	**Statistica *W***	***p-value***
Realtà virtuale	-11.3671	< 0.0001	897.5	< 0.0001
Tradizionale	-6.0936	< 0.0001	453	< 0.0001

TABELLA 4.3. Statistiche test *t* di Student e di Wilcoxon e *p-value* ad esse associato.

Volendo confrontare i due tipi di trattamento (si veda la Tabella 4.4), il test *t* di Student permette di concludere che la capacità motoria dei pazienti sottoposti al trattamento con realtà virtuale è in media più elevata, dopo la terapia, rispetto a quella dei pazienti di controllo. Tuttavia, il medesimo test porta ad affermare che anche prima della terapia

riabilitativa c'è una differenza in media nella capacità motoria dei pazienti appartenenti ai due sottocampioni e questo aspetto può falsare il confronto tra i due trattamenti.

Tempo	**Statistica *t***	***p-value***	**Statistica *W***	***p-value***
Pre-trattamento	3.4697	0.0003	18475	< 0.0001
Post-trattamento	4.2835	< 0.0001	19335.5	< 0.0001

TABELLA 4.4. Statistiche test *t* di Student e di Wilcoxon e *p-value* ad esse associato.

Di conseguenza, non ci si può basare esclusivamente su una analisi di questo tipo per affermare che il trattamento più innovativo è migliore rispetto a quello tradizionale. In questo caso è più utile considerare la media delle differenze tra le misurazioni effettuate prima e dopo la terapia, per avere una idea più chiara riguardo tale aspetto. La Tabella 4.5 riporta questi valori.

Trattamento	**Media (sd) delle differenze**
Realtà virtuale	7.54 (10.76)
Tradizionale	6.99 (12.20)

TABELLA 4.5. Media (e sd) delle differenze tra le misurazioni della variabile `FIM` effettuate prima e dopo ciascun trattamento.

Effettivamente, nel gruppo sottoposto al trattamento con realtà virtuale si ha in media un miglioramento più elevato rispetto al gruppo di pazienti di controllo. Tuttavia, da un punto di vista inferenziale, il test *t* di Student non evidenzia una differenza significativa tra queste medie. A tale proposito, si noti l'elevato valore del *p-value* associato alla statistica test *t* di Student riportato in Tabella 4.6.

Statistica *t*	***p-value***	**Statistica *W***	***p-value***
0.4141	0.3396	15890	0.1422

TABELLA 4.6. Statistiche test *t* di Student e di Wilcoxon e *p-value* ad esse associato.

È possibile effettuare un ulteriore confronto tra i due tipi di trattamento facendo riferimento al grafico di dispersione relativo all'analisi della covarianza, riportato in Figura 4.3.

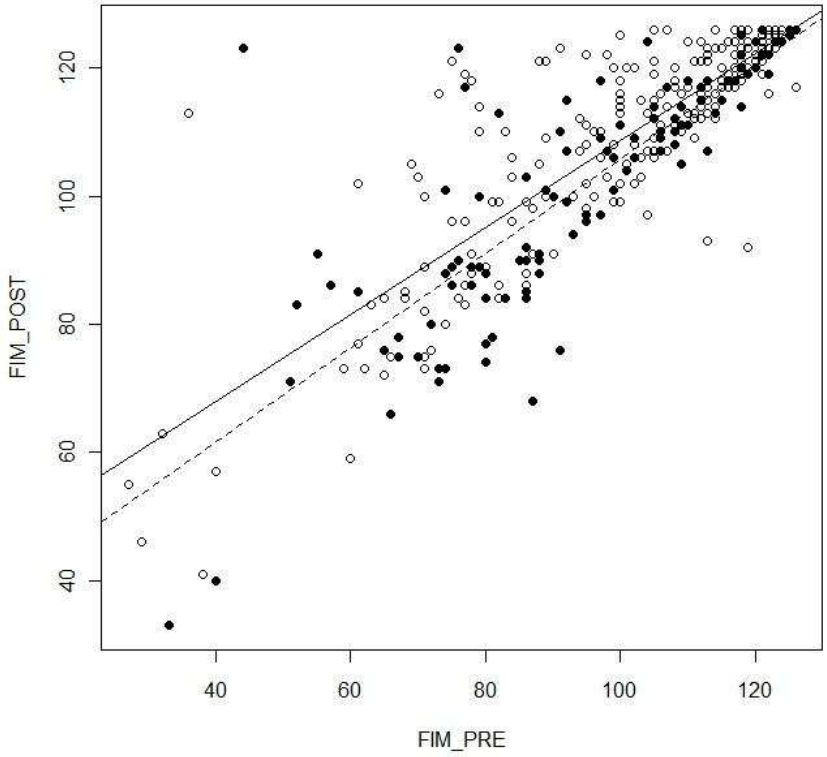

FIGURA 4.3. Analisi della covarianza.

In tale grafico i punti chiari si riferiscono ai pazienti sottoposti al trattamento con realtà virtuale e quelli scuri ai pazienti sottoposti al trattamento tradizionale. Inoltre, la retta continua è quella di regressione nel caso in cui il trattamento adottato sia quello con realtà virtuale, mentre la retta tratteggiata è quella di regressione nel caso in cui il trattamento adottato sia quello tradizionale. Sulla base della posizione di queste due rette sembra che il trattamento con realtà virtuale dia risultati migliori rispetto a quello tradizionale, poiché la retta di regressione relativa ad esso si colloca più in alto rispetto alla sua controparte. In effetti, si può affermare che il trattamento ha un effetto significativo ($\hat{\beta}_2 = -8.55$, *p-value*$= 0.043$).

4.3 Il modello misto

I dati da analizzare sono stati rilevati su un gruppo di pazienti che possono essere distinti in due sottocampioni sulla base del trattamento a cui sono stati sottoposti; pertanto, si è nel caso di campioni multipli. Il modello adatto per dati di questo tipo è quello presentato nel Paragrafo 2.3. In questo caso, la variabile `trattamento` è quella che permette di suddividere il campione composto da 376 pazienti in due sottogruppi, a seconda che il soggetto sia stato sottoposto al trattamento con realtà virtuale o a quello tradizionale. Inoltre, la variabile `FIM` rappresenta la variabile risposta del modello e la variabile `tempo` quella che identifica l'istante temporale a cui si riferisce ciascuna osservazione.
Il modello che si è interessati a stimare è il seguente

$$y_{hij} = \beta_0 + \beta_1 t_j + \beta_2 Tr_h + \beta_3 t_j Tr_h + \gamma_{i(h)} + \varepsilon_{hij} ,$$
$$\gamma_{i(h)} \sim N\left(0, \sigma_\gamma^2\right) \text{ i.i.d. ,}$$
$$\varepsilon_{hij} \sim N(0, \sigma_\varepsilon^2) \text{ i.i.d. ,}$$

con *h*=1,2, *i*=1,…,376 e *j*=1,2. Le quantità t_j, Tr_h e $t_j Tr_h$ sono rispettivamente le variabili `tempo`, `trattamento` e `tempo` × `trattamento`; inoltre, β_0, β_1, β_2 e β_3 rappresentano i coefficienti di regressione legati agli effetti fissi delle variabili sopra elencate e $\gamma_{i(h)}$ è l'effetto casuale relativo al soggetto *i*-esimo appartenente al gruppo *h*-esimo. Infine, ε_{hij} è il termine di errore. Quindi, oltre a un effetto fisso del tempo, del trattamento e dell'interazione tra queste due variabili, il modello prevede un'intercetta casuale generata dall'effetto casuale introdotto dall'*i*-esimo soggetto.
Utilizzando la funzione *lme* di R è possibile ottenere una stima di tale modello. La sequenza di comandi e *l'output* che si ottiene sono quelli riportati di seguito:

```
> fitted.model=lme(FIM~tempo*trattamento,random=~1|soggetto)
> summary(fitted.model)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  6137.634 6165.338 -3062.817

Random effects:
 Formula: ~1 | soggetto
        (Intercept) Residual
StdDev:     17.4298 7.924809

Fixed effects: FIM ~ tempo * trattamento
                      Value Std.Error  DF  t-value p-value
(Intercept)       111.47895  2.969836 374 37.53707  0.0000
tempo               8.08870  1.738360 374  4.65306  0.0000
trattamento        -8.26603  2.153641 374 -3.83816  0.0001
tempo:trattamento  -0.54877  1.260610 374 -0.43532  0.6636
 Correlation:
                  (Intr) tempo  trttmn
tempo             -0.293
trattamento       -0.943  0.276
tempo:trattamento  0.276 -0.943 -0.293

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.76642668 -0.31306819 -0.02218446  0.42553697  4.36661917

Number of Observations: 752
Number of Groups: 376
```

Come si può notare, il coefficiente di regressione associato all'interazione tra la variabile tempo e la variabile trattamento non è significativo; di conseguenza, è possibile escludere questa interazione dalla specificazione del modello (i.e. $\beta_3 = 0$) e il risultato che si ottiene, applicando nuovamente la funzione *lme* di R, è il seguente:

```
> fitted.model=lme(FIM~tempo+trattamento,random=~1|soggetto)
> summary(fitted.model)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  6138.124 6161.217 -3064.062

Random effects:
 Formula: ~1 | soggetto
        (Intercept) Residual
StdDev:    17.43175  7.91624

Fixed effects: FIM ~ tempo + trattamento
                Value Std.Error  DF  t-value p-value
(Intercept) 111.83580 2.8544326 375 39.17970       0
tempo         7.37500 0.5773511 375 12.77386       0
trattamento  -8.54041 2.0593413 374 -4.14716       0
 Correlation:
            (Intr) tempo
tempo       -0.101
trattamento -0.938  0.000

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-4.781171943 -0.307087456 -0.008383855  0.424135994  4.347483413

Number of Observations: 752
Number of Groups: 376
```

Questa volta tutti i coefficienti legati agli effetti fissi sono significativi e il modello stimato è

$$\hat{y} = 111.84 + 7.38t - 8.54Tr$$
$$\sigma_\gamma^2 = (17.43)^2$$
$$\sigma_\varepsilon^2 = (7.92)^2.$$

Tale modello mette in evidenza un effetto fisso della variabile `tempo` pari a $\hat{\beta}_1 = 7.38$ e un effetto fisso della variabile `trattamento` pari a $\hat{\beta}_2 = -8.54$; il valore di questi coefficienti rappresenta rispettivamente la differenza (in valore assoluto) tra il livello medio della variabile `FIM` prima e dopo ciascun trattamento, e la differenza tra il livello medio della variabile `FIM` nei due sottogruppi di pazienti in corrispondenza di un certo istante temporale. Il livello medio della variabile `FIM` per l'intera popolazione è stimato dall'intercetta ed è pari a $\hat{\beta}_0 = 111.84$.

Essendo interessati a valutare se ci sono differenze in media nella capacità motoria dei pazienti tra prima e dopo la terapia riabilitativa e se la variabile trattamento influenza i risultati ottenuti, possiamo ricorrere all'ANOVA per misure ripetute. La sequenza di comandi per ottenere la tabella ANOVA per misure ripetute e l'*output* risultante, sono riportati di seguito:

```
> fitted.model=lme(FIM~tempo+trattamento,random=~1|soggetto)
> aov.fitted=aov(fitted.model)
> summary(aov.fitted)
             Df Sum Sq Mean Sq F value    Pr(>F)
tempo         1  10225 10225.4  27.929 1.652e-07 ***
trattamento   1  11530 11530.1  31.492 2.820e-08 ***
Residuals   749 274229   366.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Osservando i valori riportati nell'output, si può affermare che c'è una differenza in media nella capacità motoria dei pazienti tra prima e dopo il periodo di riabilitazione. Infatti, il valore molto piccolo del *p-value* nella prima riga della tabella ANOVA, porta a rifiutare l'ipotesi nulla che la media della variabile risposta `FIM` sia uguale in corrispondenza dei due istanti temporali. Al tempo stesso, si può affermare che la variabile risposta presenta una differenza in media nei due sottogruppi, a seconda del trattamento a cui i pazienti sono stati sottoposti. Come in precedenza, anche in questo caso il valore del *p-value* è molto piccolo e porta a rifiutare l'ipotesi nulla di uguaglianza delle medie.

Per verificare se il modello stimato è adeguato, è possibile condurre l'analisi dei residui. Due grafici diagnostici basati sui residui standardizzati sono riportati in Figura 4.4 e in Figura 4.5.

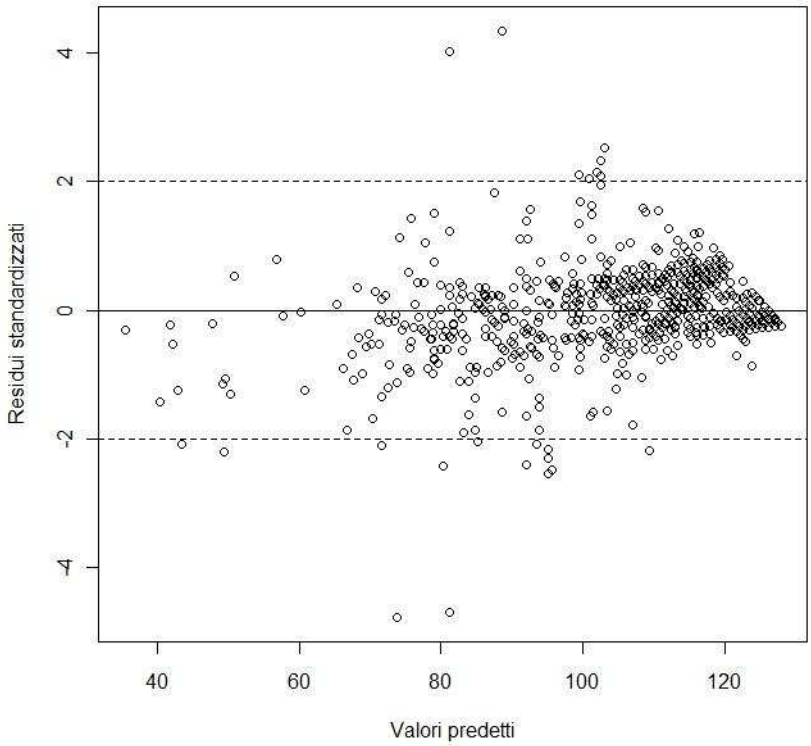

FIGURA 4.4. Grafico dei residui standardizzati rispetto ai valori predetti.

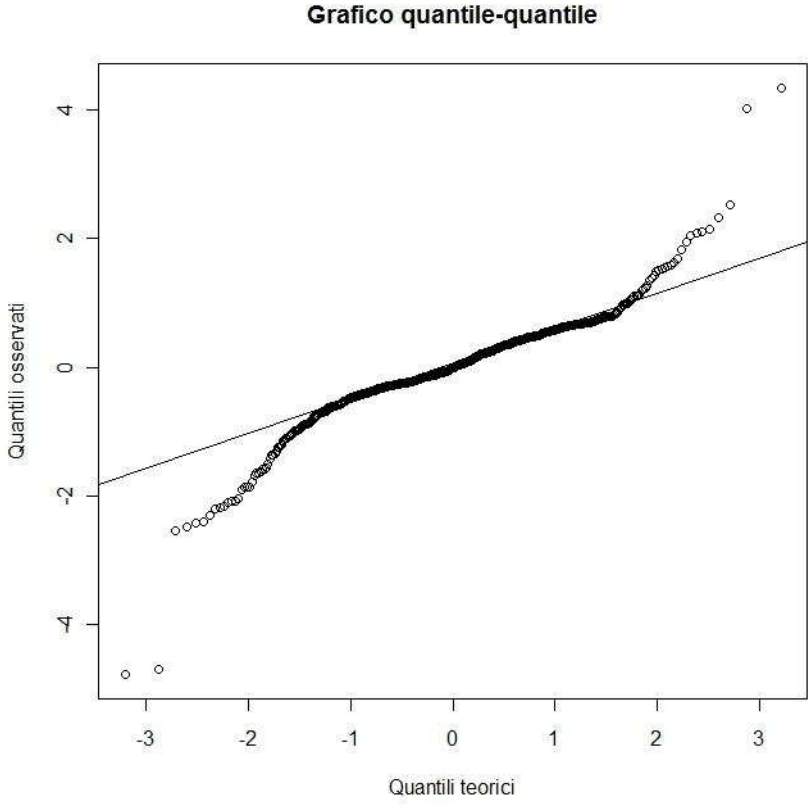

FIGURA 4.5. Grafico quantile-quantile per i residui standardizzati.

Dal grafico in Figura 4.4 si evince che i residui hanno media nulla; tuttavia, l'ipotesi di omoschedasticità non sembra soddisfatta, poiché la variabilità non appare costante. Questo aspetto, però, è coerente con la natura della variabile `FIM`, che costituisce una scala di autosufficienza e pertanto è auspicabile che tutti i pazienti, ultimata la terapia, presentino valori elevati di questa variabile e conseguentemente ci sia una "concentrazione" di valori sulla coda. Inoltre, i pazienti con un valore elevato di `FIM` all'inizio della terapia, è naturale che presentino un valore elevato di `FIM` anche al termine di quest'ultima. Osservando il grafico quantile-quantile riportato in Figura 4.5, si notano allontanamenti sistematici nelle code dalla retta di riferimento. Anche il test di Shapiro-Wilk rifiuta l'ipotesi nulla ($SW=0.9033$, $p\text{-}value<0.0001$) che i residui siano distribuiti normalmente. Tale scostamento dalle ipotesi del modello potrebbe essere dovuto alla presenza di alcuni valori anomali; tuttavia, poichè lo scopo di questa tesi è esclusivamente quello di illustrare una applicazione di tale modello, non ci si soffermerà su questo aspetto.

4.4 Conclusioni

Scopo dell'analisi appena condotta è innanzitutto illustrare una applicazione del modello lineare misto e dell'analisi della varianza per misure ripetute. In particolare, nell'ambito di questo caso di studio, si era interessati a determinare se i trattamenti a cui i pazienti erano sottoposti, risultassero efficaci al fine di garantire un recupero della capacità motoria. L'analisi esplorativa, basata prevalentemente su statistiche descrittive, congiuntamente a quella inferenziale, basata sulla stima di un modello misto e sull'ANOVA per misure ripetute, hanno permesso di concludere che entrambi i trattamenti risultano efficaci, consentendo ai pazienti di recuperare buona parte della loro capacità motoria durante la riabilitazione. Le analisi basate sull'ANOVA per misure ripetute confermano che c'è una differenza in media tra istanti temporali distinti, e quelle esplorative evidenziano che la variabile `FIM` assume mediamente valori più elevati dopo che il paziente è stato sottoposto al trattamento. Volendo confrontare i due tipi di trattamento, l'analisi della varianza consente di stabilire che il valor medio della variabile `FIM` è diverso nei due sottogruppi, e pertanto l'effetto dei due trattamenti è diverso. In particolare, le analisi

esplorative basate sulle differenze dei valori osservati nei due istanti temporali considerati, suggeriscono che il trattamento più innovativo che ricorre all'impiego della realtà virtuale produce risultati migliori, e quindi è preferibile a quello tradizionale.

APPENDICE A

La covarianza tra le osservazioni y_{ij} e $y_{i'j'}$ può essere espressa nel modo seguente

$$\mathrm{Cov}(y_{ij}, y_{i'j'})=\mathrm{E}\left[\left(y_{ij}-\mu_{ij}\right)\left(y_{i'j'}-\mu_{i'j'}\right)\right]$$

$$=\mathrm{E}[(\pi_{ij}+e_{ij})(\pi_{i'j'}+e_{i'j'})]$$

$$=\mathrm{E}[\pi_{ij}\pi_{i'j'}+\pi_{ij}e_{i'j'}+\pi_{i'j'}e_{ij}+e_{ij}e_{i'j'}].$$

A questo punto, per prima cosa si osservi che

$$\mathrm{E}\left[\pi_{ij}\pi_{i'j'}\right]=\begin{cases}\mathrm{Var}\left(\pi_{ij}\right)\text{ se } i=i', j=j'\\ \mathrm{Cov}\left(\pi_{ij},\pi_{i'j'}\right)\text{ altrimenti}\end{cases}.$$

Poiché

$$\mathrm{E}\left[\pi_{ij}\pi_{i'j'}\right]=\begin{cases}\sigma_{\pi j}^2\text{ se } i=i', j=j'\\ \sigma_{\pi jj'}\text{ se } i=i', j\neq j',\\ 0\text{ se } i\neq i'\end{cases}$$

allora $\mathrm{E}\left[\pi_{ij}\pi_{i'j'}\right]=\delta_{ii'}\sigma_{\pi jj'}$, dove

$$\delta_{ii'}=\begin{cases}1\text{ se } i=i'\\ 0\text{ altrimenti}\end{cases}$$

e $\sigma_{\pi jj}=\sigma_{\pi j}^2$.

Inoltre, poiché

$$\mathrm{Cov}\left(\pi_{ij}, e_{i'j'}\right)=0\ \ \forall\, i, j, i', j',$$

si ha che

$$\mathrm{E}\left[\pi_{ij}e_{i'j'}\right]=0,\ E\left[\pi_{i'j'}e_{ij}\right]=0.$$

Infine

$$\mathrm{E}\left[e_{ij}e_{i'j'}\right]=\begin{cases}\mathrm{Var}\left(e_{ij}\right)\text{ se } i=i', j=j'\\ \mathrm{Cov}\left(e_{ij},e_{i'j'}\right)\text{ altrimenti}'\end{cases}$$

così che

$$\mathrm{E}\left[e_{ij}e_{i'j'}\right] = \begin{cases}\sigma_{ej}^2 \text{ se } i = i', j = j' \\ 0 \text{ altrimenti}\end{cases}.$$

Quest'ultima espressione può essere riscritta come segue:

$$\mathrm{E}\left[e_{ij}e_{i'j'}\right] = \delta_{ii'}\delta_{jj'}\sigma_{ej}^2.$$

Dunque

$$\mathrm{Cov}(y_{ij}, y_{i'j'}) = \delta_{ii'}\sigma_{\pi jj'} + \delta_{ii'}\delta_{jj'}\sigma_{ej}^2$$
$$= \delta_{ii'}(\sigma_{\pi jj'} + \delta_{jj'}\sigma_{ej}^2).$$

APPENDICE B

Scheffé (1959) fornisce un modello alternativo a quello usuale, nell'ambito dell'analisi della varianza per misure ripetute a un campione. Questo modello può essere specificato come

$$y_{ij} = \mu + \pi_i + \tau_j + e_{ij},$$

per $i = 1, \dots, n$ e $j = 1, \dots, t$, dove y_{ij}, μ, τ_j sono definiti come nel Capitolo 2. La differenza tra i due modelli sta nel modo in cui è definito il termine di errore e_{ij}. Ora, la componente erratica casuale include sia l'errore di misurazione che l'interazione tra soggetto e tempo. Scheffé assume che le componenti π_i e π_j seguano una distribuzione normale multivariata. Una differenza importante nel modello di Scheffé è che Cov($e_{ij}, e_{ij,}$)≠0 e Cov(π_i, e_{ij})≠0. In ogni caso, a condizione che certe assunzioni siano soddisfatte, l'analisi conduce ai medesimi risultati indipendentemente dal modello adottato.

Adottando come modello di riferimento quello proposto da Scheffé, la tabella ANOVA per misure ripetute è quella riportata nella Tabella B.1.

Fonte	*SS*	*df*	*E[MS]*
Tempo	SS_T	*t-1*	$\sigma_e^2 + \sigma_{T\times S}^2 + n\sigma_\tau^2$
Soggetto	SS_S	*n-1*	$\sigma_e^2 + t\sigma_\pi^2$
Tempo x Soggetto	SS_{TS}	*(n-1)(t-1)*	$\sigma_e^2 + \sigma_{T\times S}^2$

TABELLA B.1. Somme dei quadrati, gradi di libertà e medie quadratiche attese per l'ANOVA per misure ripetute a un campione, basata sul modello proposto da Scheffé.

La Tabella B.1 mostra le somme dei quadrati, i gradi di libertà e le medie quadratiche attese per ogni fonte di variabilità presente nel modello di Scheffé. Le somme dei quadrati SS_T e SS_S sono calcolate nello stesso modo in cui vengono computate nel modello usuale, sulla base delle equazioni riportate nel Paragrafo 2.2. La somma dei quadrati tempo/trattamento × soggetto SS_{TS} è calcolata come la somma dei quadrati dei residui, secondo la formula riportata sempre nel Paragrafo 2.2.

La statistica

$$F = \frac{MS_T}{MS_{TS}},$$

dove

$$MS_{TS} = \frac{SS_{TS}}{(n-1)(t-1)},$$

testa l'ipotesi nulla che le medie siano uguali in corrispondenza di ogni istante temporale.

BIBLIOGRAFIA

Armitage P., Berry G. (1996). *Statistica Medica: Metodi Statistici per la Ricerca in Medicina*, McGraw-Hill, Milano.

Bates D. (2005). *R News*, Vol. 5/1.

Davis C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*, Springer, New York.

Fox J. (2002). *Linear Mixed Models: Appendix to An R and S-PLUS Companion to Applied Regression*, http://cran.r-project.org/doc/contrib/Fox-Companion/ appendix -mixed-models.pdf.

Greenhouse S. W., Geisser S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24:95-112.

Huynh H., Feldt L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, 1:69-82.

Iacus S. M., Masarotto G. (2007). *Laboratorio di Statistica con R*, McGraw-Hill, Milano.

Laird N. M., Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38:963-974.

Lindsey J. K. (1993). *Models for Repeated Measurements*, Clarendon Press·Oxford.

Mauchly J. W. (1940). Significance test for sphericity of a normal *n*-variate distribution. *Annals of Mathematical Statistics*, 29:204-209.

Piccolo D. (2004). *Statistica per le Decisioni*, Il Mulino, Bologna.

Piron L., Turolla A., Zucconi C., Agostini M., Ventura L., Tonin P., Dam. M. (2010). Motor learning principles for rehabilitation: A pilot randomized controlled study in post-stroke patients. *Neurorehabilitation and Neural Repair*, 24:501-508.

Scheffé H. (1959). *The Analysis of Variance*, John Wiley and Sons, New York.

Venables W. N., Ripley B. D. (1999). *Modern Applied Statistics with S-PLUS*, Springer, New York.

Wallenstein S. (1982). Regression models for repeated measurements. *Biometrics*, 38:849-53.

Wayne W. D. (2007). *Biostatistica: Concetti di Base per l'Analisi Statistica delle Scienze dell'Area Medico-Sanitaria*, EdiSES, Napoli.