4. ANALISI DELLA VARIANZA

4.1 Introduzione

L'obiettivo di un'analisi della varianza univariata, usualmente abbreviata in ANOVA (Analysis of Variance), consiste nel confrontare i valori medi di una variabile quantitativa (detta **variabile dipendente**) in corrispondenza delle diverse determinazioni di una o più variabili quantitative e/o qualitative (dette **fattori**) sotto le condizioni standard di normalità e omoschedasticità della variabile dipendente.

In situazioni più complesse vengono rilevate due o più variabili dipendenti, sempre in corrispondenza di uno o più fattori e l'analisi della varianza multivariata viene usualmente abbreviata con l'acronimo MANOVA (Multivariate Analysis of Variance). Anche in questo caso devono essere ritenute valide le ipotesi di normalità e omoschedasticità per tutte le variabili dipendenti considerate.

Le modalità dei fattori vengono detti **livelli** e, nel caso ci sia un solo fattore, tali livelli corrispondono anche ai cosiddetti **trattamenti**. Lo scopo dell'analisi è verificare se i trattamenti influenzano i valori medi della variabile dipendente o i vettori delle medie delle variabili dipendenti.

Se i fattori sono 2 o più, la rilevazione dei valori della variabile dipendente (o delle variabili dipendenti) viene effettuata in corrispondenza di diverse combinazioni dei livelli dei fattori e sono queste combinazioni che costituiscono i trattamenti.

In questa situazione si può essere interessati a valutare l'effetto di ciascuno dei fattori considerati e l'effetto della loro eventuale **interazione**. Fra i fattori non esiste un effetto di interazione quando l'effetto complessivo sulla variabile dipendente (o sulle variabili dipendenti) sotto un determinato trattamento corrisponde alla somma degli effetti imputabili ai singoli fattori. Se invece l'effetto complessivo risulta significativamente diverso dalla somma degli effetti imputabili ai singoli fattori, si conclude dicendo che i fattori interagiscono fra di loro.

In ogni caso l'analisi viene condotta scomponendo la varianza complessiva della variabile dipendente (o delle variabili dipendenti) nella **varianza spiegata**, imputabile ai fattori (o al fattore), e nella **varianza residua**, o varianza all'interno dei trattamenti.

4.2 Analisi della varianza univariata con un fattore sperimentale

Considerato un solo fattore A che assume *u* livelli diversi, l'ipotesi da verificare è

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_u \qquad 4.2.1$$

contro l'ipotesi alternativa che almeno una media differisca dalle altre.

Se i fattori fossero 2 ed il secondo fattore, denominato B, assumesse *v* livelli diversi, il numero dei trattamenti sarebbe $t=u\times v$ e lo scopo dell'analisi diventerebbe quello di valutare l'effetto del fattore A, del fattore B e, dove possibile, l'eventuale effetto di interazione, che può essere misurato solo quando esiste un numero sufficiente di osservazioni per ciascun trattamento.

Esistono più procedure SAS che effettuano l'analisi della varianza (come la Proc ANOVA), ma è preferibile utilizzare la Proc GLM perché adeguata a gestire situazioni in cui il numero di osservazioni sotto i diversi trattamenti è variabile. La sintassi da utilizzare varia a seconda del piano degli esperimenti utilizzato.

Quando si considera un solo fattore sperimentale il listato assume la forma

```
PROC GLM <opzioni>;
CLASS <fattore>;
MODEL <variabile dipendente>=<fattore>;
```

in cui il fattore viene specificato dopo la parola chiave "CLASS", mentre l'istruzione "MODEL" assume la stessa forma utilizzata nella Proc REG.

Le opzioni più comunemente usate all'interno della **PROC GLM** sono

`data = <nome DSS>`
per specificare il dataset SAS su cui si vuole far eseguire la procedura

`outstat = <nome DSS>`
per memorizzare i principali risultati dell'analisi in un dataset SAS

L'output fornito dal SAS consiste in più tabelle, nella prima delle quali è indicato il fattore, il numero di livelli e le loro determinazioni

Tabella 4.2.1
Informazioni sui livelli di classificazione

Classe	**Livelli**	**Valori**
A	u	$a_1\, a_2 \ldots a_u$

Nella seconda tabella è indicato il numero di osservazioni lette e di quelle usate nell'analisi

Tabella 4.2.2
Osservazioni lette e usate

Numero osservazioni lette	n
Numero osservazioni usate	n

Nella terza tabella sono riportati i risultati relativi a:

- scomposizione della devianza complessiva della variabile dipendente (*SST*) in devianza spiegata (*SSA*) e devianza residua (*SSE*), con *SST*=*SSA*+*SSE*;
- gradi di libertà associati
- valori della varianza spiegata (*MSA*) e della varianza residua (*MSE*);
- test *F* sulla bontà complessiva del modello (pari al rapporto *MSA*/*MSE*) e *p*-valore associato

Tabella 4.2.3
Scomposizione della varianza per uno schema casuale semplice

Origine	**DF**	**Somma dei quadrati**	**Media quadratica**	**Valore F**	**Pr > F**
Modello	u-1	*SSA*	$MSA = SSA/(u\text{-}1)$	$F_{(u-1),(n-u)}=MSA/MSE$	*p*-valore
Errore	$n–u$	*SSE*	$MSE = SSE/(n\text{-}u)$		
Totale corretto	$n–1$	*SST*			

Scelto un livello di significatività α per la verifica dell'ipotesi 4.2.1, se la probabilità risultante nell'ultima colonna della tabella precedente è minore di α l'ipotesi va rifiutata.
Supponiamo, per esempio, di considerare i 5 diversi macchinari a_i (i=1,2,...,5) e di avere prelevato 4 fogli da ciascun lotto in modo casuale ottenendo i valori (y) della resistenza alla lacerazione indicati nel listato riportato nella figura 4.2.1.

Figura 4.2.1
Listato di un programma SAS con la procedura GLM

```
data a;
input A $ y @@;
cards;
a1 112 a1 119 a1 117 a1 113
a2 108 a2 99 a2 112 a2 118
a3 120 a3 106 a3 102 a3 109
a4 110 a4 101 a4 99 a4 104
a5 100 a5 102 a5 96 a5 101
;
proc glm;
class A;
model y=A;
run;
```

Il caso considerato costituisce un esempio di **schema casuale semplice** (**o schema completamente casuale**), in cui non esistono i fattori sub-sperimentali, e ciascuno dei gruppi omogenei in A può essere costituito da un numero di unità sperimentali che è costante oppure variabile.
I principali risultati forniti dal programma SAS sono riportati nelle tabelle successive

Tabella 4.2.4
Informazioni sui livelli di classificazione

Classe	**Livelli**	**Valori**
A	5	a1 a2 a3 a4 a5

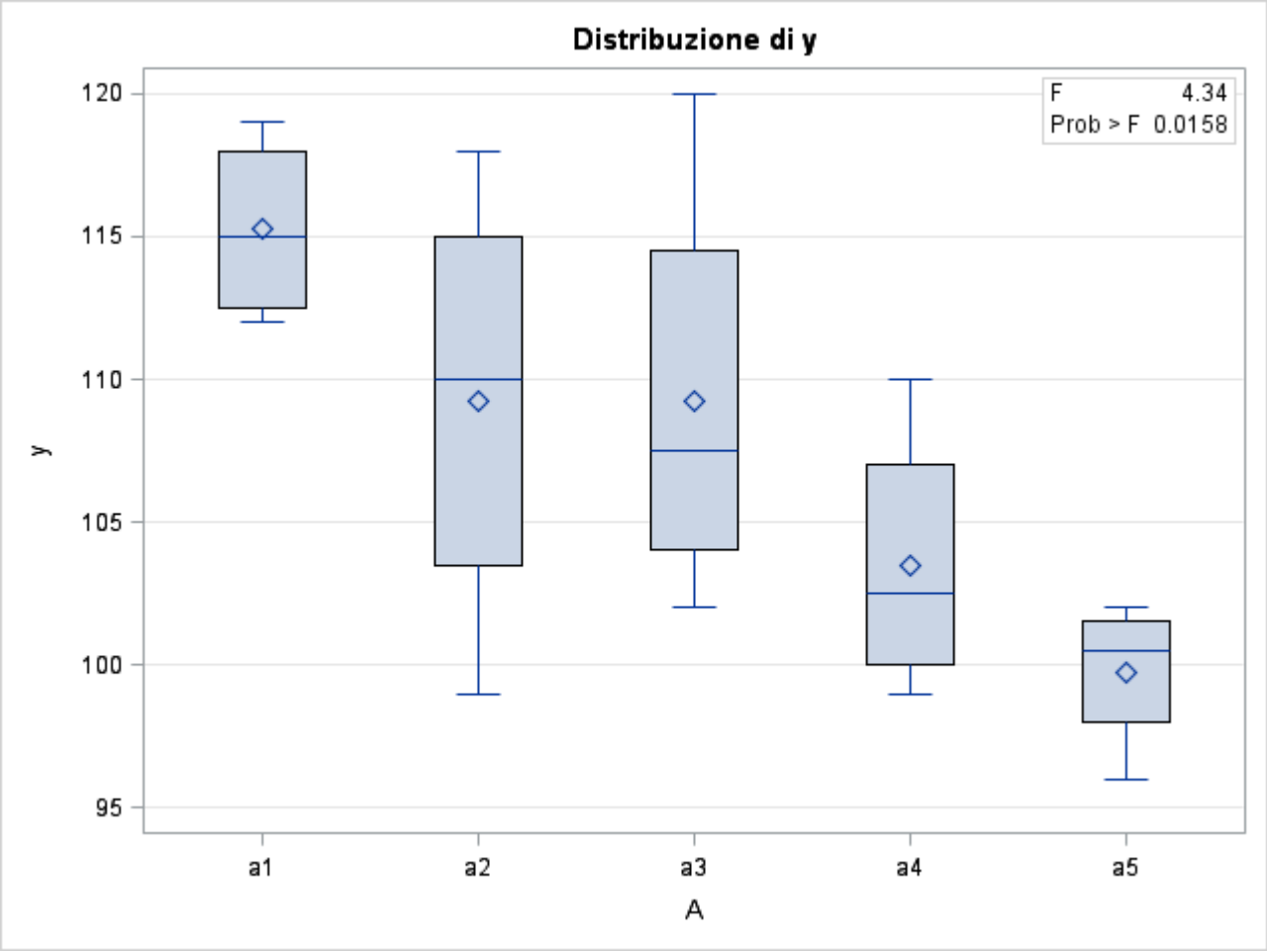
Tabella 4.2.5
Osservazioni lette e usate

Numero osservazioni lette	20
Numero osservazioni usate	20

Tabella 4.2.6
Scomposizione della varianza e test F

Origine	**DF**	**Somma dei quadrati**	**Media quadratica**	**Valore F**	**Pr > F**
Modello	4	568.800000	142.200000	4.34	0.0158
Errore	15	492.000000	32.800000		
Totale corretto	19	1060.800000			

Dai risultati riportati nella tabella precedente si vede come il test *F* porti al rifiuto dell'ipotesi nulla 4.2.1 per ogni livello di significatività α superiore all'1.58%, per cui si concluderebbe che i diversi macchinari danno luogo a fogli di carta la cui resistenza alla rottura è significativamente diversa.
Di seguito il SAS fornisce anche i box-plot della variabile dipendente in corrispondenza dei diversi trattamenti



In realtà un'analisi della varianza dovrebbe essere sempre preceduta dalla verifica delle ipotesi di normalità e omoschedasticità sottostanti il modello, che verranno illustrate qui di seguito.
La verifica dell'ipotesi di normalità della variabile dipendente all'interno di ciascun gruppo omogeneo rispetto al fattore sperimentale può essere effettuata mediante la **`proc univariate`** nella quale va specificata:

- l'opzione "`normal`",
- la variabile sulla quale effettuare il test

- l'eventuale variabile che identifica i gruppi omogenei all'interno dei quali si vuole effettuare la verifica di normalità.

Nel caso esaminato le istruzioni risultano

```
proc univariate data=a normal;
   var y;
   by A;
```

dove l'opzione "by A" richiede che la verifica di normalità sulla Y sia eseguita separatamente per ciascuna determinazione assunta dal fattore "A".
L'output fornito dal SAS comprende 4 diversi test, riportati nella prima colonna delle tabelle seguenti, di cui il preferibile è quello di Shapiro-Wilk, particolarmente adatto per numerosità contenute.

Tabella 4.2.7a
Test di Normalità sulla Y per A=a1

Test di normalità				
Test	**Statistica**		**P-value**	
Shapiro-Wilk	**W**	0.915686	**Pr < W**	0.5130
Kolmogorov-Smirnov	**D**	0.252059	**Pr > D**	>0.1500
Cramer-von Mises	**W-Qu**	0.044307	**Pr > W-Qu**	>0.2500
Anderson-Darling	**A-Qu**	0.27056	**Pr > A-Qu**	>0.2500

Tabella 4.2.7b
Test di Normalità sulla Y per A=a2

Test di normalità				
Test	**Statistica**		**P-value**	
Shapiro-Wilk	**W**	0.987317	**Pr < W**	0.9434
Kolmogorov-Smirnov	**D**	0.187717	**Pr > D**	>0.1500
Cramer-von Mises	**W-Qu**	0.025651	**Pr > W-Qu**	>0.2500
Anderson-Darling	**A-Qu**	0.177926	**Pr > A-Qu**	>0.2500

Tabella 4.2.7c
Test di Normalità sulla Y per A=a3

Test di normalità				
Test	**Statistica**		**P-value**	
Shapiro-Wilk	**W**	0.926606	**Pr < W**	0.5746
Kolmogorov-Smirnov	**D**	0.262918	**Pr > D**	>0.1500
Cramer-von Mises	**W-Qu**	0.045552	**Pr > W-Qu**	>0.2500
Anderson-Darling	**A-Qu**	0.276127	**Pr > A-Qu**	>0.2500

Tabella 4.2.7d
Test di Normalità sulla Y per A=a4

Test di normalità				
Test	**Statistica**		**P-value**	
Shapiro-Wilk	**W**	0.941248	**Pr < W**	0.6620
Kolmogorov-Smirnov	**D**	0.208483	**Pr > D**	>0.1500
Cramer-von Mises	**W-Qu**	0.037066	**Pr > W-Qu**	>0.2500
Anderson-Darling	**A-Qu**	0.240517	**Pr > A-Qu**	>0.2500

Tabella 4.2.7e
Test di Normalità sulla Y per A=a5

Test di normalità				
Test	**Statistica**		**P-value**	
Shapiro-Wilk	**W**	0.886912	**Pr < W**	0.3690
Kolmogorov-Smirnov	**D**	0.287866	**Pr > D**	>0.1500
Cramer-von Mises	**W-Qu**	0.058057	**Pr > W-Qu**	>0.2500
Anderson-Darling	**A-Qu**	0.338619	**Pr > A-Qu**	>0.2500

Come si vede dai *p*-valori riportati nell'ultima colonna delle precedenti tabelle, non si ha motivo di rifiutare l'ipotesi di normalità per nessuno dei trattamenti considerati.

Una volta verificata l'ipotesi di normalità si può procedere alla verifica dell'ipotesi di omoschedasticità, ossia di uguaglianza delle varianze della Y per ciascuna delle *u* popolazioni normali

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_u^2$$

può essere effettuata all'interno della stessa Proc GLM, utilizzando le istruzioni

```
MEANS <nome_del_fattore> / hovtest=bartlett;
```

che richiedono il calcolo della statistica utilizzata nel test di Bartlett, la cui distribuzione, sotto ipotesi di normalità, tende a una chi-quadrato con *u*-1 gradi di libertà.
Nell'esempio precedente le istruzioni contenute nell'ultima riga prima del "run" finale,

```
proc glm;
class A;
model y=A;
MEANS A / hovtest=bartlett;
run;
```

danno luogo ai risultati riportati nella Tabella 4.2.8, che non mostrano una evidente violazione dell'ipotesi di omoschedasticità.

Tabella 4.2.8
Test di Bartlett sui dati riportati nella Figura 4.2.1

Test di Bartlett di omogeneità della varianza y			
Origine	**DF**	**Chi-quadrato**	**Pr > ChiQuadr**
A	4	4.6395	0.3263

Quando si ipotizza che i risultati dell'esperimento siano influenzati, oltre che dal fattore sperimentale, anche da uno o più fattori sub-sperimentali, l'effetto complessivo sulla *Z* deve essere scomposto nella somma dell'effetto del fattore sperimentale e sub-sperimentali (per ipotesi, infatti, i fattori sub-sperimentali non possono interagire con il fattore sperimentale).

Per tenere sotto controllo l'effetto di questi ultimi, le unità vengono suddivise in gruppi, detti **blocchi**, che sono omogenei rispetto alle determinazioni di questi fattori. Per esempio, se si ipotizza che la variabile "sesso" sia un fattore sub-sperimentale che può influenzare i risultati di un esperimento, le unità statistiche verranno preventivamente suddivise in maschi e femmine e poi si somministreranno tutti i trattamenti all'interno di questi due blocchi distinti, in modo da poter valutare l'effetto dei trattamenti a prescindere dall'effetto del fattore sub-sperimentale "sesso".

Il SAS non distingue fra fattori sperimentali e sub-sperimentali per cui i risultati sono organizzati secondo uno schema analogo a quello riportato nella tabella successiva in cui la prima tabella scompone la devianza complessiva (SST) in devianza spiegata dal modello nel suo complesso (SSM), senza specificare a quale fattore si riferisce, e devianza residua (SSE).

La seconda tabella scinde invece la devianza spiegata nelle singole componenti (SSM=SSA+SSB) e sta al ricercatore decidere quali sono i fattori sperimentali, dei quali si deve valutare il risultato ottenuto con il test F, e quali i fattori sub-sperimentali, dei quali non si deve tener conto del risultato del test. Si ottengono perciò due tabelle analoghe a quelle successive.

Tabella 4.2.9
Scomposizione della devianza totale per un disegno con un fattore sperimentale e uno sub-sperimentale

Origine	**DF**	**Somma dei quadrati**	**Media quadratica**	**Valore F**	**Pr > F**
Modello	$u+v-2$	*SSM*	$MSM = SSM/(u+v-2)$	$F_{(u+v-2),(u-1)(v-1)} = MSM/MSE$	*p*-valore
Errore	$(u-1)(v-1)$	*SSE*	$MSE = SSE/[(u-1)(v-1)]$		
Totale corretto	$uv-1$	*SST*			

Tabella 4.2.10
Scomposizione della devianza spiegata per un disegno con un fattore sperimentale e uno sub-sperimentale

Origine	**DF**	**Somma dei quadrati**	**Media quadratica**	**Valore F**	**Pr > F**
Fattore sper. (A)	$u-1$	*SSA*	$MSA = SSA/(u-1)$	$F_{(u-1),(u-1)(v-1)} = MSA/MSE$	*p*-valore
Fattore sub-sper. (B)	$v-1$	*SSB*	$MSB = SSB/(v-1)$	$F_{(v-1),(u-1)(v-1)} = MSB/MSE$	*p*-valore

Supponiamo, per esempio, di voler confrontare la resa produttiva di tre diversi fertilizzanti (a_1, a_2 e a_3) e che i terreni utilizzati differiscano per 4 diverse composizioni del suolo (b_1, b_2, b_3 e b_4). Se l'interesse è rivolto solo ai fertilizzanti, al netto delle differenze indotte dalla tipologia dei terreni, si potrebbero somministrare i tre diversi tipi di fertilizzante a ciascuno dei 4 tipi di terreno, che costituisce il fattore sub-sperimentale, del quale non si desidera valutare la significatività.

Trascurando le verifiche delle ipotesi sottostanti il modello, le istruzioni che compaiono nella **proc glm** illustrata di seguito servono per specificare al SAS la presenza di due fattori (senza che sia possibile specificare quale sia il fattore sperimentale e quello sub-sperimentale).

L'opzione "ss3" che compare nelle istruzioni che specificano il modello serve per richiedere la scomposizione della varianza che si riferisce al modello ANOVA, dato che la **proc glm** è in grado di effettuare numerose altri tipi di analisi differenti, alcune delle quali dipendono anche dall'ordinamento dei dati immessi, informazione che però non è rilevante nell'analisi della varianza.

```
data a;
input A $ B $ y;
cards;
a1 b1 4.6
a1 b2 6.2
a1 b3 5
a1 b4 6.6
a2 b1 4.9
a2 b2 6.3
a2 b3 5.4
a2 b4 6.8
a3 b1 4.4
a3 b2 5.9
a3 b3 5.4
a3 b4 6.3
;

proc glm;
class A B;
model y= A B /ss3;
run;
```

Le istruzioni precedenti danno origine ai risultati riportati nelle tabelle successive dai quali consegue che, se si lavora ad un livello di significatività del 5%, il fattore A non esercita un'influenza significativa sulla variabile dipendente. Se si lavorasse invece ad un livello di significatività del 10% si dovrebbe concludere che il fertilizzante esercita un effetto significativo sulla resa produttiva.

Tabella 4.2.11
Informazioni sui livelli di classificazione

Classe	Livelli	Valori
A	3	a1 a2 a3
B	4	b1 b2 b3 b4

Tabella 4.2.12
Osservazioni lette e usate

Numero osservazioni lette	12
Numero osservazioni usate	12

Tabella 4.2.13
Scomposizione della devianza totale

Origine	DF	Somma dei quadrati	Media quadratica	Valore F	Pr > F
Modello	5	7.02333333	1.40466667	45.15	0.0001
Errore	6	0.18666667	0.03111111		
Totale corretto	11	7.21000000			

Tabella 4.2.14
Scomposizione della devianza spiegata

Origine	DF	SS Anova	Media quadratica	Valore F	Pr > F
A	2	0.26000000	0.13000000	4.18	0.0730
B	3	6.76333333	2.25444444	72.46	<.0001

Questo tipo di piano degli esperimenti può essere generalizzato per gestire un qualunque numero di fattori sub-sperimentali, eventualmente utilizzando alcune accortezze nell'esecuzione dell'esperimento stesso, in modo da risparmiare sul numero delle unità statistiche da utilizzare per valutare l'efficacia del fattore sperimentale.

4.3 Analisi della varianza univariata con due fattori sperimentali

Quando si vuole valutare l'effetto di più fattori contemporaneamente, l'effetto complessivo sulla variabile dipendente è scomponibile nell'effetto dei fattori principali più gli eventuali effetti di interazione. In questo caso si utilizza il cosiddetto **schema fattoriale**.

Dati due fattori sperimentali A e B, sia *u* il numero di livelli di A e *v* il numero di livelli di B, cosicché il numero dei trattamenti è $t=u\times v$. Se si ipotizza che i risultati dell'esperimento non siano influenzati da nessun fattore di tipo sub-sperimentale, la valutazione dell'eventuale effetto di interazione richiede che si effettuino un certo numero di osservazioni per ognuno dei t trattamenti. Se si ipotizza di effettuare uno stesso numero *r* di osservazioni (dette **replicazioni**) per ogni trattamento la numerosità campionaria necessaria sarà $n=uvr=tr$.

Un esperimento nel quale vengono considerate tutte le possibili associazioni delle modalità dei fattori sperimentali viene detto **disegno fattoriale completo**.

Anche in questo caso il SAS effettua una prima scomposizione della devianza complessiva (SST) della Z in devianza spiegata (SSM) e devianza residua (SSE) e successivamente scompone la SSM nella devianza associata al fattore A (SSA), al fattore B (SSB) e all'interazione (SSAB).

Consideriamo un esperimento che consiste nel rilevare i valori della durata (variabile Y, espressa in ore) di alcune batterie prodotte con due diverse combinazioni di materie prime ("mater" assume le determinazioni a_1 e a_2) e con due diversi livelli di temperatura ("temper" assume le determinazioni b_1 e b_2) e supponiamo che per ciascuna associazione di modalità dei due fattori vengano effettuate 4 replicazioni. In base ai valori indicati nel listato successivo si vuole verificare la significatività dei due effetti principali e dell'effetto di interazione.

```
data a;
input mater $ temper $ y @@;
cards;
a1 b1 123 a1 b1 133
a1 b1 130 a1 b1 138
a2 b1 101 a2 b1 107
a2 b1 111 a2 b1 112
a1 b2 112 a1 b2 108
a1 b2 100 a1 b2 92
a2 b2 143 a2 b2 159
a2 b2 129 a2 b2 132
;
proc glm;
class mater temper;
model y= mater temper mater*temper;
run;
```

Come si nota dalle istruzioni che compaiono nel "`model`", la valutazione dell'effetto di interazione fra due generici fattori sperimentali A e B viene richiesta al SAS scrivendo il "prodotto" fra i due fattori: "A*B".

La scomposizione della devianza totale porta ai risultati illustrati nelle due tabelle seguenti

Tabella 4.3.1
Scomposizione della devianza totale

Origine	DF	Somma dei quadrati	Media quadratica	Valore F	Pr > F
Modello	3	3956.250000	1318.750000	16.12	0.0002
Errore	12	981.500000	81.791667		
Totale corretto	15	4937.750000			

Tabella 4.3.2
Scomposizione della devianza spiegata

Origine	DF	SS Anova	Media quadratica	Valore F	Pr > F
mater	1	210.250000	210.250000	2.57	0.1348
temper	1	25.000000	25.000000	0.31	0.5905
mater*temper	1	3721.000000	3721.000000	45.49	<.0001

I risultati mostrano che i fattori A e B non risultano significativi neanche per un valore α pari a 0.10. Le differenze fra le medie dei gruppi omogenei sono invece dovute al solo effetto di interazione, che risulta altamente significativo.

In alcuni casi il disegno fattoriale viene applicato effettuando una sola osservazione all'interno di ciascun gruppo omogeneo nei fattori sperimentali, ma in questo caso non è possibile valutare l'effetto di interazione e la significatività degli effetti principali si verifica assumendo l'ipotesi che i fattori non interagiscano fra di loro.

In molte situazioni reali occorre considerare oltre ai fattori sperimentali anche uno o più fattori di disturbo ed in questi casi il numero di unità sperimentali può risultare eccessivamente elevato. Per questo motivo sono stati proposti numerosi schemi alternativi che consentono di valutare l'effetto dei fattori sperimentali, con o senza effetti di interazione, al netto dell'effetto dei fattori di disturbo con un notevole risparmio nel numero di osservazioni necessarie.

4.4 Le analisi successive

Quando l'analisi porta al rifiuto dell'ipotesi di uguaglianza dei valori medi della Z, i risultati ottenuti non consentono di individuare quali trattamenti producono effetti significativamente diversi da zero, ma sono proprio queste le informazioni che si vogliono ricavare dall'esperimento.

Il procedimento consiste in generale nello scomporre l'ipotesi nulla globale 4.2.1 in un certo numero ipotesi nulle **parziali** che, considerate congiuntamente, equivalgono all'ipotesi globale.

Uno di questi test si basa sul confronto dei valori medi per ogni possibile coppia di trattamenti. Se l'esperimento prevede *t* trattamenti, le ipotesi nulle parziali corrispondenti alla 4.2.1 sono

$$H_{0jl}: \mu_j = \mu_l, \qquad l > j = 1,2,...,t-1. \tag{4.4.1}$$

Queste ipotesi parziali, il cui numero è pari a $t(t\text{-}1)/2$, considerate nel loro complesso, equivalgono evidentemente all'ipotesi nulla globale. Infatti, se è vera la 4.2.1, sono vere tutte 4.4.1 e viceversa.

L'ipotesi nulla globale, quindi, può essere verificata verificando tutte le singole ipotesi parziali: la 4.2.1 viene accettata se vengono accettate tutte le ipotesi parziali, mentre viene respinta se viene respinta almeno una delle ipotesi 4.4.1.

Ma se i singoli test 4.4.1 sono effettuati al livello α risulta molto maggiore di α la probabilità di rifiutare l'ipotesi nulla globale.

Per evitare questo inconveniente sono state proposte numerose procedure per la verifica delle ipotesi parziali in modo che risulti predeterminata la probabilità di rifiutare l'ipotesi nulla globale.

Il **metodo di Tukey** può essere utilizzato sotto le condizioni standard di normalità e di omoschedasticità e sotto condizione che in ogni trattamento si abbia uno stesso numero di replicazioni.

Supponiamo, per esempio, che si voglia verificare se esiste una differenza fra i livelli medi di spesa mensili (in euro) a seconda di tre diversi tipi di clienti indicati dalle lettere a, b e c sulla base dei valori riportati nel listato seguente, nel quale si chiede al SAS l'effettuazione del test di Tukey (per default il livello di significatività è pari al 5%, ma può essere specificato un valore diverso usando l'opzione "alpha=").

```
data a;
input A $ y @@;
cards;
a 148 a 76 a 393 a 520 a 236 a 134 a 55 a 166 a 415 a 153 b 513
b 264 b 433 b 94 b 535 b 327 b 214 b 135 b 280 b 304
c 335 c 643 c 216 c 536 c 128 c 723 c 258 c 380 c 594 c 465
;
proc GLM;
   class A;
   model Y=A;
   means A / tukey;
run;
```

L'output fornito dal programma assume la forma riportata nelle tabelle seguenti

Tabella 4.4.1
Test di Tukey

Alfa	0.05
Gradi di libertà dell'errore	27
Errore quadratico medio	28543.37
Valore critico del range di Student	3.50633
Differenza minima significativa	187.33

Nella prima di queste (la cui nota ricorda che questo test può portare a commettere errori di seconda specie con una probabilità piuttosto elevata) viene calcolato il valore minimo della differenza fra medie che risulta significativo, scelto il valore del livello α.

Nella seconda tabella vengono calcolate le medie sotto ciascun trattamento, identificandole con una o più lettere (quelle che compaiono nella prima e nella seconda colonna della tabella 4.4.2) che consentono di identificare le medie significativamente diverse fra loro. Nell'esempio considerato la media dei clienti di tipo "a" differisce significativamente da quella dei clienti di tipo "c", mentre la spesa media dei clienti di tipo "b" non differisce significativamente da quella del tipo "a" o da quella del tipo "c"

Tabella 4.4.2
Medie sotto i diversi trattamenti
(quelle identificate con la stessa lettera non risultano significativamente diverse)

Raggruppamento di Tukey		**Media**	**N**	**A**
	A	427.80	10	c
	A			
B	A	309.90	10	b
B				
B		229.60	10	a

Nel caso in cui il numero di replicazioni sotto ciascun trattamento non sia costante il SAS utilizza automaticamente il **metodo di Tukey-Cramer**, che è stato proposto per casi come questi. In questo caso il SAS restituisce una tabella di output che riporta tutte le possibili differenze fra le coppie di medie campionarie, evidenziando quelle significative al livello α.

Considerato il listato seguente, in cui si è scelto un livello di significatività α=0.01 per l'effettuazione del test (`alpha=0.01`)

```
data a;
input A $ y @@;
cards;
a 14 a 76 a 39    a 52 a 23 b 13 b 64 b 33 b 94
c 135 c 143 c 116 c 86 c 128 c 93
;
proc glm;
   class A;
   model Y=A;
   means A / tukey alpha=0.01;
 run;
```

si ottiene l'output riportato nella tabella successiva, che riporta anche gli intervalli di confidenza calcolati a livello 1-α per ciascuna differenza.

Tabella 4.4.3
Differenze fra medie sotto i diversi trattamenti
(quelle significativamente diverse a livello α sono evidenziate da ***)

Confronto A	**Differenza tra medie**	**Limiti di confidenza al 99%Simultaneo**		
c - b	65.83	3.30	128.37	***
c - a	76.03	17.37	134.70	***
b - c	-65.83	-128.37	-3.30	***
b - a	10.20	-54.79	75.19	
a - c	-76.03	-134.70	-17.37	***
a - b	-10.20	-75.19	54.79	

Come si vede, risultano significative quelle differenze il cui intervallo di confidenza non contiene il valore zero.

Gli effetti dei diversi trattamenti possono essere valutati anche effettuando un confronto fra gruppi di trattamenti. Esempi di questo tipo sono

$$H_0: \alpha_1 = \alpha_t \qquad 4.4.2$$

oppure

$$H_0: \frac{\alpha_1 + \ldots + \alpha_q}{q} = \frac{\alpha_{q+1} + \ldots + \alpha_t}{t-q} \qquad 4.4.3$$

dove la 4.4.2 mette a confronto solo gli effetti del primo e dell'ultimo trattamento, mentre la 4.4.3 confronta gli effetti medi dei primi *q* e dei rimanenti *t-q* trattamenti.

Tutte le ipotesi del tipo 4.4.2 e 4.4.3 possono essere espresse nella forma generale seguente

$$H_0 : \sum c_i \alpha_i = 0, \qquad 4.4.4$$

dove i valori c_i sono degli opportuni coefficienti che servono a specificare l'ipotesi. Così per l'ipotesi 4.4.2 si ha

$$c_1 = -c_t = 1$$

mentre gli altri coefficienti sono uguali a zero; per l'ipotesi 4.4.3 i primi q coefficienti uguali a $1/q$ ed i rimanenti $t-q$ uguali a $-1/(t-q)$.

Se i valori c_i sono tali che $\sum c_i = 0$, come negli esempi precedenti, le funzioni $\sum c_i \alpha_i$ sono dette **contrasti lineari**.

Il cosiddetto **metodo di Scheffè** consente di verificare un'ipotesi su una qualsiasi combinazione lineare. Per effettuare questa verifica si utilizza la **`proc glm`** che ha una struttura (e una sintassi) molto simile alla **`proc anova`**.

Nel listato successivo, per esempio, si considerano i confronti fra i primi due trattamenti contro gli altri due e fra il primo trattamento contro gli altri 3, ed infine fra il terzo e il quarto.

```
data a;
input A $ y @@;
cards;
a 14 a 76 a 39 a 52 a 23 b 13 b 64 b 33 b 94
c 135 c 143 c 116 c 86 c 128 c 93 d 145 d 154 d 99
;
proc glm;
  class A;
  model y = A;
  means A;
  contrast 'Compara gruppi 1 e 2 contro 3 e 4' A 0.5 0.5 -0.5 -0.5;
  contrast 'Compara gruppo 1 contro i restanti' A 0.9 -0.3 -0.3 -0.3;
  contrast 'Compara gruppi 3 e 4' A 0 0 1 -1;
run;
```

Nell'opzione `contrast` va inserita una stringa di testo (compresa fra i due apici) che serve per identificare di quale contrasto si tratti, poi va elencato il fattore (in questo caso `A`) e poi i valori numerici c_i che il SAS evidenzia automaticamente in colore celeste.

In questo caso l'output assume la forma riportata nella tabella successiva, in cui viene riportata nell'ultima colonna il p-valore associato ai diversi contrasti

Tabella 4.4.4
Confronti fra medie con il metodo di Scheffè

Contrasto	**DF**	**SS contrasto**	**Media quadratica**	**Valore F**	**Pr > F**
Compara gruppi 1 e 2 contro 3 e 4	1	26178.20000	26178.20000	34.61	<.0001
Compara gruppo 1 contro i restanti	1	12439.06275	12439.06275	16.45	0.0012
Compara gruppi 3 e 4	1	501.38889	501.38889	0.66	0.4292

4.5 Analisi della varianza multivariata ad un fattore

L'analisi della varianza multivariata è un'estensione dell'analisi della varianza univariata che viene utilizzata quando sulle *n* unità statistiche si rilevano contemporaneamente i valori di *h* variabili diverse Z_k (per $k = 1, 2, .., h$) sotto ogni trattamento. Lo scopo dell'analisi è analogo a quello del caso univariato e consiste nel verificare l'*uguaglianza dei vettori delle medie* delle variabili Z_k sotto i diversi trattamenti. L'analisi della varianza multivariata viene indicata in questo caso con il termine MANOVA, dove la "M" sta per "multivariate".

Sotto le ipotesi usuali i dati osservati sono delle realizzazioni indipendenti di una variabile multinormale che ha una stessa matrice di covarianza sotto i diversi trattamenti.

L'ipotesi rilevante è che i valori medi delle Z_k nei g gruppi siano tutti uguali fra di loro

$$H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_g = \boldsymbol{\mu} \qquad 4.5.2$$

contro l'ipotesi alternativa che almeno due vettori risultino diversi.

Per verificare l'ipotesi nulla la statistica più comunemente utilizzata è il cosiddetto lambda di Wilks che si basa sul rapporto tra la matrice di covarianza entro gruppi e la matrice di covarianza totale, per cui questa statistica ha lo stesso significato della statistica F utilizzata nel caso univariato. Anche in questo caso la regione di rifiuto dell'ipotesi nulla è sulla coda destra della distribuzione, dato che valori elevati indicherebbero che la matrice di covarianza fra gruppi assume valori così elevati da non poter essere imputati a fattori puramente casuali, ma piuttosto all'effetto del fattore considerato.

Ci sono molte altre statistiche che possono essere calcolate invece del lambda di Wilks, come la Traccia di Pillai, la traccia di Hotelling-Lawl e l'autovalore massimo di Roy.

In alcuni casi i risultati ottenuti con le quattro statistiche sono analoghi, ma in altri casi si può giungere a una conclusione diversa a seconda della statistica considerata. In questi casi può essere opportuno

basarsi sulla traccia di Pillai che risulta essere la statistica più robusta rispetto alle eventuali violazioni delle ipotesi di omoschedasticità. Questo vale soprattutto nei casi in cui la numerosità campionaria è piuttosto limitata.

Tutti i precedenti indici vengono di solito trasformati matematicamente in modo da ottenere delle statistiche che asintoticamente hanno delle distribuzioni F, in modo da semplificare il calcolo del *p*-valore corrispondente. Queste trasformazioni delle quattro statistiche vengono tutte calcolate dal SAS, che riporta anche i *p*-valori associati.

Si osservi infine che quando il test risulta significativo si ha generalmente l'interesse a determinare quali sono le variabili dipendenti per cui si rilevano differenze significative e quali siano i trattamenti che forniscono risultati significativamente diversi dagli altri. Per rispondere al primo tipo di quesito si utilizzano generalmente i singoli test F per l'ANOVA su ciascuna delle variabili dipendenti singolarmente considerate, mentre per rispondere al secondo quesito si possono eseguire dei confronti multivariati dei vettori di medie sotto i vari trattamenti ed eseguire quindi, per esempio, il test di Tukey o i contrasti lineari.

Si consideri, per esempio, un campione di 45 ettari coltivati con tre diverse varietà di frumento (Centauro, Eridano e Pandas). I dati rilevati sono i valori delle variabili X_1: "resa produttiva in tonnellate per ettaro", X_2: "percentuale di umidità" e X_3: "peso di un ettolitro di grano" ottenendo i valori indicati nel file grano.txt.

Prima di effettuare la verifica di ipotesi sui vettori delle medie è sempre opportuno eseguire un test di omoschedasticità sulle matrici di covarianza della variabile dipendente sotto i diversi trattamenti. Il test usualmente utilizzato è il test di Box, che è una generalizzazione del test di omoschedasticità di Bartlett. Analogamente al test di Bartlett, il test di Box si basa sul confronto assunto dagli stimatori corretti della matrice di covarianza comune sotto l'ipotesi di base

$$H_0: \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$$

di uguaglianza di queste matrici, e gli stimatori corretti delle g matrici di covarianza sotto l'ipotesi alternativa che almeno due matrici siano diverse fra di loro.

Nell'esempio considerato il test di omoschedasticità è stato effettuato singolarmente per ciascuna variabile utilizzando l'opzione

```
  means varieta /hovtest=bartlett;
```

all'interno della **proc glm**.

```
data a3;
input varieta $ resa umidita peso @@;
cards;
Centauro 6.900 13.600 78.800 Centauro 6.900 13.400 77.900
Centauro 6.600 13.800 75.800 Centauro 6.650 11.300 78.800
Centauro 7.750 11.200 79.600 Centauro 7.720 12.100 80.700
Centauro 7.530 11.900 80.700 Centauro 6.160 11.500 81.400
Centauro 3.880 11.100 77.900 Centauro 4.320 11.600 76.100
Centauro 4.390 11.100 77.000 Centauro 5.450 12.600 82.000
Centauro 4.800 12.300 81.600 Centauro 4.700 12.600 81.400
Centauro 7.500 11.400 78.900 Eridano  9.000 14.300 83.500
Eridano  7.900 13.400 79.100 Eridano  7.900 14.600 80.100
Eridano  5.200 11.500 79.800 Eridano  6.100 11.600 80.100
Eridano  7.100 11.800 81.300 Eridano  7.100 12.300 82.400
Eridano  7.330 12.500 83.100 Eridano  7.270 11.600 83.800
Eridano  4.470 11.700 80.600 Eridano  4.860 12.000 77.900
Eridano  4.440 12.200 78.800 Eridano  4.150 12.800 84.200
Eridano  4.450 12.700 83.700 Eridano  5.600 13.000 83.800
Pandas   7.750 13.900 81.100 Pandas   6.650 13.500 79.500
Pandas   7.000 13.800 79.100 Pandas   5.400 12.100 76.400
Pandas   4.700 12.400 76.000 Pandas   5.800 11.800 77.300
Pandas   7.210 12.200 81.200 Pandas   7.110 12.200 82.600
Pandas   7.140 11.800 83.000 Pandas   5.100 12.200 77.000
Pandas   5.050 12.100 78.300 Pandas   5.290 12.400 77.800
Pandas   5.600 12.900 82.200 Pandas   5.300 12.800 82.400
Pandas   5.100 12.300 82.400
;
proc print;

proc glm;
  class varieta;
  model resa umidita peso = varieta/ ss3;
  means varieta /hovtest=bartlett;
  manova h=_all_;
run;
```

Criteri di test MANOVA e approssimazioni F per l'ipotesi di nessun effetto varieta globale H = Tipo III - Matrice SSCP per varieta E = Matrice SSCP di errore S=2 M=0 N=19					
Statistica	**Valore**	**Valore F**	**DF num**	**DF den**	**Pr > F**
Lambda di Wilks	0.77749782	1.79	6	80	0.1120
Traccia di Pillai	0.23035329	1.78	6	82	0.1136
Traccia Hotelling-Lawley	0.27607931	1.82	6	51.593	0.1136
Radice massima di Roy	0.23268118	3.18	3	41	0.0339
NOTE: la statistica F per la radice massima di Roy è un limite superiore.					
NOTE: la statistica F per Lambda di Wilks è esatta.					

Il valore della statistica Λ di Wilks risulta pari a 0,78. Dalla 3.7.14 risulta che il corrispondente test F è uguale a 1,79 a cui è associata una significatività pari a 0,11 nella distribuzione $F_{6,80}$. In base a questo risultato non si ha motivo di rifiutare l'ipotesi di uguaglianza dei vettori di medie agli usuali livelli di significatività.

I valori delle tracce di Pillai e di Hotelling-Lawley sono rispettivamente pari a 0,23 e a 0,28 a cui corrisponde una significatività all'incirca uguale a 0,11. Anche in questo caso non si ha motivo di rifiutare l'ipotesi nulla mentre, se si utilizza l'autovalore massimo di Roy si ottiene un risultato uguale a 0,23 a cui è associata una significatività dello 0,03 circa.

1.4 Analisi multivariata a due fattori

Se i fattori considerati sono due o più, è possibile valutare anche l'eventuale effetto di interazione.

Su un campione casuale di 60 famiglie inglesi sono stati rilevati i valori di tre diversi indici sintetici di povertà Z_1, Z_2 e Z_3, che assumono valori compresi fra zero ed uno, e i due fattori A e B che rilevano rispettivamente il possesso di un'abitazione e il possesso di un'automobile. Entrambe queste variabili vengono poste uguali ad uno in caso di possesso del bene e pari a zero in caso contrario.

I dati rilevati sono riportati nel file indici_pov.txt.

Volendo valutare l'effetto di entrambi i fattori e della loro interazione sulle variabili Z1, Z2 e Z3 (che sono tutte in formato alfanumerico nel file considerato) il programma SAS assume la forma

```
data a1;
input A  B  x1-x3 @@;
cards;
1  1  0.461  0.875  0.083 1  1  0.813  0.676  0.952 1  1  0.028  0.676  0.424
0  0  0.000  0.108  0.015 0  0  0.004  0.409  0.266 1  0  0.003  0.579  0.247
0  1  0.002  0.296  0.236 1  1  0.291  0.619  0.744 1  1  0.102  0.802  0.578
1  1  0.000  0.676  0.102 1  1  0.000  0.460  0.032 1  1  0.009  0.676  0.320
1  1  0.794  0.779  0.940 1  1  0.646  0.755  0.900 1  1  0.381  0.659  0.755
1  0  0.073  0.675  0.533 0  0  0.949  0.214  0.988 1  1  0.532  0.676  0.859
1  1  0.289  0.755  0.742 1  0  0.333  0.488  0.768 1  1  0.001  0.802  0.166
1  1  0.000  0.581  0.102 0  0  0.000  0.108  0.102 1  0  0.000  0.392  0.054
1  0  0.000  0.305  0.047 0  0  0.000  0.382  0.003 0  0  0.765  0.829  0.938
0  1  0.100  0.568  0.576 1  0  0.002  0.413  0.235 1  0  0.001  0.392  0.180
0  0  0.000  0.295  0.058 0  1  0.000  0.581  0.046 0  1  0.513  0.705  0.852
0  0  0.012  0.214  0.347 0  1  0.024  0.568  0.410 0  0  0.593  0.409  0.882
0  0  0.000  0.322  0.090 0  1  0.000  0.490  0.063 0  0  0.000  0.000  0.151
0  0  0.000  0.409  0.005 1  0  0.000  0.686  0.104 0  1  0.257  0.385  0.722
0  1  0.349  0.664  0.777 0  1  0.000  0.188  0.020 0  0  0.029  0.214  0.428
0  1  0.672  0.667  0.909 1  0  0.399  0.503  0.802 0  1  0.448  0.402  0.825
0  0  0.271  0.488  0.731 0  0  0.231  0.427  0.703 1  0  0.046  0.501  0.478
0  1  0.024  0.581  0.410 1  0  0.704  0.824  0.919 1  0  0.300  0.430  0.748
0  1  0.000  0.559  0.077 0  1  0.001  0.473  0.190 1  0  0.534  0.536  0.860
1  0  0.026  0.471  0.417 1  0  0.002  0.613  0.224 0  1  0.008  0.664  0.310
;
proc print;
proc glm;
  class A B;
  model x1 x2 x3 = A B A*B/ ss3;
  manova h=_all_;
run;
```

Per il fattore A le statistiche test risultano

Criteri di test MANOVA e statistiche F esatte per l'ipotesi di nessun effetto A globale
H = Tipo III - Matrice SSCP per A
E = Matrice SSCP di errore

S=1 M=0.5 N=26

Statistica	Valore	Valore F	DF num	DF den	Pr > F
Lambda di Wilks	0.68927448	8.11	3	54	0.0001
Traccia di Pillai	0.31072552	8.11	3	54	0.0001
Traccia Hotelling-Lawley	0.45080085	8.11	3	54	0.0001
Radice massima di Roy	0.45080085	8.11	3	54	0.0001

con una significatività pari a 0.0001, per cui i valori degli indici di povertà dipendono dal possesso o

meno di un'abitazione.

Per il fattore B le stesse statistiche assumono i valori

Criteri di test MANOVA e statistiche F esatte per l'ipotesi di nessun effetto B globale H = Tipo III - Matrice SSCP per B E = Matrice SSCP di errore S=1 M=0.5 N=26					
Statistica	**Valore**	**Valore F**	**DF num**	**DF den**	**Pr > F**
Lambda di Wilks	0.69055605	8.07	3	54	0.0002
Traccia di Pillai	0.30944395	8.07	3	54	0.0002
Traccia Hotelling-Lawley	0.44810838	8.07	3	54	0.0002
Radice massima di Roy	0.44810838	8.07	3	54	0.0002

con una significatività pari a 0.0002, cosicché anche il possesso o meno di un'automobile influisce sui valori degli indici.

Per quanto riguarda infine l'effetto di interazione le statistiche risultano

Criteri di test MANOVA e statistiche F esatte per l'ipotesi di nessun effetto A*B globale H = Tipo III - Matrice SSCP per A*B E = Matrice SSCP di errore S=1 M=0.5 N=26					
Statistica	**Valore**	**Valore F**	**DF num**	**DF den**	**Pr > F**
Lambda di Wilks	0.92985120	1.36	3	54	0.2655
Traccia di Pillai	0.07014880	1.36	3	54	0.2655
Traccia Hotelling-Lawley	0.07544088	1.36	3	54	0.2655
Radice massima di Roy	0.07544088	1.36	3	54	0.2655

per cui si conclude che i fattori non interagiscono fra di loro.