Capitolo 4

Analisi della varianza

4.1 ANOVA

La procedura per eseguire una ANOVA passa attraverso il fit del modello lineare corrispondente.

4.2 ANOVA a una via

In questo caso, che è il più semplice, si hanno r serie di dati che si classificano in base a un solo criterio. Il modello lineare si scrive come:

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1, \ldots, r \quad j = 1, \ldots, n_i$$

con $\varepsilon_{ij} \sim N(0, \sigma^2)$. α_i esprime l'effetto dovuto all'appartenenza al gruppo i-esimo. L'ipotesi nulla H_0 è che per ogni i sia $\alpha_i = 0$, mentre l'ipotesi alternativa H_1 è che almeno un α_i sia diverso da 0.

Nel seguente esempio è riportata la procedura che permette di eseguire il test in R.

Esempio

Siano A e B gli errori di battitura per minuto di due dattilografi. Ci si chiede se ci sia differenza significativa fra le loro abilità.

Si inizia inserendo i dati e unendoli in un unico vettore:

```
> A <- c(1,2,3,4,4,5,7,9)
> B <- c(3,3,5,8)
> Dati <- c(A, B)           # concateno i dati
```

È quindi necessario usare un secondo vettore in cui tenere traccia del gruppo a cui appartengono i valori in *Dati*. Si crea quindi a tale scopo il vettore *gruppo*:

```
> fA <- rep(1, length(A))       # etichetta di gruppo: 1
> fB <- rep(2, length(B))       # etichetta di gruppo: 2
> gruppo <- factor(c(fA, fB))   # concateno i gruppi in un fattore
```

Essendo l'appartenenza ai gruppi un dato categoriale, si usa la funzione *factor* che forza R a trattarli in tal modo e non come dati numerici quantitativi. A tal punto occorre fittare il modello lineare che mette in relazione gli errori di battitura (in *Dati*) con il dattilografo (in *gruppo*):

```
> modello <- aov(Dati ~ gruppo)   # fit del modello lineare
> anova(modello)                  # tabella ANOVA

Analysis of Variance Table
```

```
Response: Dati
          Df Sum Sq Mean Sq F value Pr(>F)
gruppo     1  0.375   0.375   0.058 0.8145
Residuals 10 64.625   6.462
```

da cui si conclude che non vi è differenza significativa fra i due dattilografi.

Per valutare i coefficienti del modello lineare μ, α_1 e α_2 si possono usare le chiamate seguenti:

```
> mean(Dati)
[1] 4.5

> model.tables(modello)
Tables of effects

 gruppo
        1    2
    -0.125 0.25
rep  8.000 4.00
```

da cui si ha: $\mu = 4.5$ (media generale), $\alpha_1 = -0.125$ (effetto dovuto all'appartenenza al primo gruppo) e $\alpha_2 = 0.25$ (effetto del secondo gruppo). La funzione *model.tables* è molto utile per tabelle sui fit di modelli lineari ottenuti con la funzione *aov*, soprattutto in casi complessi (si veda ad esempio la Sec. 4.10). □

Se si vogliono calcolare le medie e varianze dei due campioni un modo rapido per procedere è fare uso della funzione *tapply*:

```
> tapply(Dati, gruppo, mean)     # calcola la media dei vari gruppi
> tapply(Dati, gruppo, var)      # calcola la varianza dei vari gruppi
```

tapply suddivide i dati passati come primo argomento secondo le chiavi di classificazione passate come secondo argomento e ad ognuno di questi sottogruppi applica la funzione specificata come terzo argomento.

4.2.1 Test per l'omogeneità delle varianze

Per verificare che l'ipotesi di omogeneità delle varianze sia soddisfatta è possibile usare il test di Bartlett:

```
> bartlett.test(Dati, gruppo)
```

il cui output evidenzia che non ci sono problemi dovuti a differenza di varianza nei due gruppi:

```
        Bartlett test for homogeneity of variances

data:  Dati and gruppo
Bartlett's K-squared = 0.0373, df = 1, p-value = 0.8468
```

Il test di Bartlett ha il difetto di essere molto sensibile all'ipotesi di normalità dei dati e fornisce troppi risultati significativi se i dati provengono da distribuzioni con lunghe code. Per ovviare a questo inconveniente si può ricorrere al test di Fligner-Killeen, uno dei test per l'omogeneità delle varianze più robusto per scostamenti dalla normalità [15]. Per questo test la sintassi è:

```
> fligner.test(Dati, gruppo)

        Fligner-Killeen test for homogeneity of variances

data:  Dati and gruppo
Fligner-Killeen:med chi-squared = 0.0125, df = 1, p-value = 0.911
```

In questo caso particolare i risultati dei due test coincidono non mettendo in luce nessun problema di non omogeneità delle varianze.

Se l'ipotesi di omogeneità delle varianze non è soddisfatta, è possibile utilizzare il test F con una correzione (dovuta a Satterthwaite) ai gradi di libertà della varianza d'errore. Detta n_i la numerosità di ogni gruppo e s_i^2 la relativa varianza, il numero di gradi di libertà *df* da assegnare alla varianza d'errore si calcola come:

$$df = \frac{\left(\sum \mathrm{v}_i\right)^2}{\sum \mathrm{v}_i^2/(n_i - 1)} \qquad , \qquad \mathrm{v}_i = (n_i - 1)s_i^2.$$

Nel caso dell'esempio in questione si ha:

```
> n <- tapply(Dati, gruppo, length)   # numerosita' dei gruppi
> s2 <- tapply(Dati, gruppo, var)     # varianze dei gruppi
> v <- (n - 1) * s2
> df <- sum(v)^2/sum(v^2/(n - 1))     # gdl della varianza d'errore
> df
[1] 9.921307
```

da confrontarsi con il valore 10 riportato nella tabella ANOVA. Dato che non vi sono particolari problemi dovuti alla disomogeneità delle varianze i due numeri sono quasi identici. La significatività della differenza fra i due gruppi si testa con la chiamata:

```
> 1 - pf(0.058, 1, df)
[1] 0.8145891
```

4.3 Contrasti

Si definisce contrasto fra le medie di r gruppi $\mu_1, \ldots, \mu_r$ una combinazione lineare L

$$L = \sum_i c_i \mu_i, \tag{4.1}$$

dove tutti i c_i sono noti e $\sum_i c_i = 0$. Ad esempio:

- $\mu_1 - \mu_2$ è un contrasto con $c_1 = 1$ e $c_2 = -1$, equivalente al paragone delle medie dei gruppi 1 e 2. Tutte le differenze fra coppie di gruppi sono contrasti.
- $(\mu_1 + \mu_2)/2 - \mu_3$ è un contrasto, che risulta direttamente interpretabile come il confronto fra la media dei gruppi 1 e 2 combinati contro la media del gruppo 3.

Quanto detto vale in un disegno bilanciato in cui la taglia di tutti i gruppi è uguale. Se ciò non è vero si definisce contrasto la combinazione lineare

$$L = \sum_i n_i c_i \mu_i, \tag{4.2}$$

con la condizione

$$\sum_i n_i c_i = 0. \tag{4.3}$$

È facile verificare che nel caso di gruppi di uguale taglia n, l'Eq. (4.2) si riduce all'Eq. (4.1) moltiplicata per n.

In R sono disponibili diversi tipi di contrasti fra gruppi, utili per esplorare disegni sperimentali differenti. La libreria *multcomp* mette a disposizione i test più usati. Si noti che fra di essi non vi sono i test di Duncan e di Newman-Keuls; benché molto diffusi questi test non garantiscono protezione contro errori di tipo I sull'esperimento e sono quindi sconsigliati dai più moderni testi sull'argomento [30].

4.4 Contrasti fra due gruppi: test di Tukey

Il test HSD (Honest Significant Difference) di Tukey è una tecnica tramite quale è possibile confrontare fra loro a due a due le medie dei vari gruppi.

Si abbiano ad esempio tre campioni A, B e C. Come primo passo si esegue il confronto indifferenziato:

```
> A <- c(400, 450, 420, 430, 380, 470, 300)
> B <- c(300, 350, 380, 270, 400, 320, 370, 290)
> C <- c(270, 300, 250, 200, 410)
> na <- length(A)
> nb <- length(B)
> nc <- length(C)
> dati <- c(A, B, C)
> g <- factor(c(rep("A",na), rep("B",nb), rep("C",nc)))
> anova(res <- aov(dati ~ g))

Analysis of Variance Table

Response: dati
          Df Sum Sq Mean Sq F value   Pr(>F)
g          2  45057   22529  6.5286 0.007875 **
Residuals 17  58663    3451
```

Il confronto rivela una differenza altamente significativa. Si può procedere ad analizzare da dove questa differenza tragga origine mediante un test di Tukey:

```
> TukeyHSD(res, "g", ordered=TRUE)
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = dati ~ g)

$g
         diff        lwr      upr     p adj
B-C  49.00000 -36.910577 134.9106 0.3326095
A-C 121.14286  32.903658 209.3821 0.0069972
A-B  72.14286  -5.850313 150.1360 0.0724100
```

Il test evidenzia che esiste differenza altamente significativa fra i gruppi A e C, mentre le altre differenze non sono significative. L'opzione *ordered* richiede che i gruppi siano ordinati in ordine crescente per media prima di effettuare il test. Se si vuole condurre il test a un differente livello, ad esempio $\alpha = 0.01$ si può effettuare la chiamata:

```
> TukeyHSD(res, "g", ordered=TRUE, conf.level=0.99)
```

4.5 Contrasti fra due gruppi: test di Dunnet

Questo test si impiega per confrontare un gruppo di controllo con diversi gruppi sperimentali. Per poterlo impiegare in maniera semplice e rapida è necessario installare una libreria supplementare, che non fa parte della distribuzione standard di R, ossia la libreria *multcomp* che a sua volta dipende dalla libreria *mvtnorm* (entrambe disponibili presso il sito *www.r-project.org*).

Esempio

Si stabilisca tramite test di Dunnett se qualcuno dei fattori B, C e D differisce significativamente dal controllo A.

```
> A <- c(2, 2, 3, 5, 1, 4, 6)
> B <- c(3, 5, 7)
> C <- c(6, 8, 7)
> D <- c(2, 4, 3)
> dati <- c(A, B, C, D)
> gp <- factor(c(rep("A",7), rep("B",3), rep("C",3), rep("D",3)))
> library(multcomp)
```

La libreria mette a disposizione la funzione *glht*[1], che viene chiamata nel modo seguente:

```
> mod <- aov(dati ~ gp)
> cont <- glht(mod, linfct = mcp(gp = "Dunnett"))
> confint(cont)

         Simultaneous Confidence Intervals for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = dati ~ gp)

Estimated Quantile = 2.7265

Linear Hypotheses:
           Estimate lwr     upr
B - A == 0  1.7143  -1.3305  4.7591
C - A == 0  3.7143   0.6695  6.7591
D - A == 0 -0.2857  -3.3305  2.7591

95% family-wise confidence level
```

da cui si evidenzia che solo il contrasto A vs. C è significativo. L'opzione $linfct$ serve per specificare il tipo di contrasto; in questo caso si richiede un test di Dunnett sulla variabile *gp* (nella pagina di manuale di *glht* vi sono alcuni utili esempi su come costruire i contrasti). È anche possibile utilizzare la funzione *summary* per calcolare la significatività dei confronti singoli o del test globale. Nel primo caso la chiamata è:

```
> summary(cont)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts

Fit: aov(formula = dati ~ gp)

Linear Hypotheses:
           Estimate Std. Error t value p value
B - A == 0   1.7143     1.1168   1.535  0.3548
C - A == 0   3.7143     1.1168   3.326  0.0167 *
D - A == 0  -0.2857     1.1168  -0.256  0.9903
---
(Adjusted p values reported)
```

[1] Fino alla vesione 0.4-8 le funzioni di interfaccia erano *simint* e *simtest*.

mentre se si volesse una stima di significatività globale si dovrebbe aggiungere l'opzione *test* alla chiamata precedente (si veda la pagina di manuale di *summary.glht* per le opzioni implementate).

4.6 Contrasti multipli

Sia L un contrasto. La sua varianza può essere stimata, nel caso di disegno bilanciato, come:

$$\text{var } L = \sum_{i=1}^{r} \text{var}(c_i \mu_i) = \sum_{i=1}^{r} c_i^2 \text{var}(\mu_i) = \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}. \tag{4.4}$$

Dove σ è la varianza d'errore ottenuta come nel test ANOVA. Per disegni non bilanciati la varianza di L si calcola come:

$$\text{var } L = \sum_{i=1}^{r} \text{var}(n_i c_i \mu_i) = \sum_{i=1}^{r} n_i^2 c_i^2 \text{var}(\mu_i) = \sigma^2 \sum_{i=1}^{r} n_i c_i^2. \tag{4.5}$$

Per il teorema di Scheffé la dimensione dell'intervallo di confidenza del contrasto L è data da:

$$\sqrt{(r-1) F^{\alpha}_{r-1,N-r}} \sqrt{\text{var } L}.$$

Il contributo di L alla devianza del fattore è:

$$d_L^2 = \sigma^2 \frac{L^2}{\text{var} L}.$$

A ciascun contrasto spetta 1 gdl.

Esempio

I vettori a, b e c registrano il numero di difetti orari di tre linee produttive. Ci si chiede se vi sia differenza fra le tre linee e da dove eventualmente tale differenza si origini.

```
> a <- c(1,1,2,3,5,3)
> b <- c(1,2,4,2,2,5)
> c <- c(5,5,6,7,4,6)
> dati <- c(a, b, c)
> gp <- factor(rep(LETTERS[1:3], each=6))
> ni <- as.vector(tapply(gp, gp, length)) # num. dati nei gruppi
```

Il punto di partenza è il confronto indifferenziato:

```
> anova(res <- lm(dati ~ gp))

Analysis of Variance Table

Response: dati
          Df Sum Sq Mean Sq F value   Pr(>F)
gp         2 34.111  17.056  9.0294 0.002667 **
Residuals 15 28.333   1.889
```

che evidenzia differenza altamente significativa nel numero di difetti orari delle tre linee. Si possono esaminare le medie dei tre gruppi:

```
> tapply(dati, gp, mean)

       A        B        C
2.500000 2.666667 5.500000
```

In questo caso l'origine della differenza sembra evidente dato che la linea C ha in media circa il doppio dei difetti delle altre due. Si procede in ogni caso a un'analisi formale provando i contrasti:

1. A $\cup$ B vs. C: verifica che la media dei gruppi A e B presi insieme differisca dalla media di C. In questo caso $c_1 = c_2 = 1/2$, $c_3 = -1$.
2. A vs. B: verifica che le medie dei gruppi A e B differiscano. Si ha $c_1 = 1$, $c_2 = -1$, $c_3 = 0$.

In questo caso particolare i vettori associati ai due contrasti sono ortogonali fra loro e si parla di *confronti ortogonali*. In generale, dato un disegno bilanciato, due contrasti $\mathbf{c} = (c_1, \ldots, c_r)$ e $\mathbf{d} = (d_1, \ldots, d_r)$ si dicono ortogonali se:

$$\sum_{i=1}^{r} c_i d_i = 0. \tag{4.6}$$

Per un disegno non bilanciato l'equazione corrispondente risulta:

$$\sum_{i=1}^{r} n_i c_i d_i = 0. \tag{4.7}$$

Per il problema in esame si definisce la matrice dei contrasti come

```
contrasti <- cbind(c(1/2,1/2,-1), c(1,-1,0))
```

Per calcolare la stima della varianza associata ai due contrasti e il loro intervallo di confidenza si calcola la varianza d'errore del modello:

```
> sigma <- summary(res)$sigma
```

e si fa uso del teorema di Scheffé:

```
> varcontrasti = sigma^2 * colSums(contrasti^2/ni)
> ngp <- length(levels(gp))  # num. gruppi
> n <- length(dati)          # num. tot. dati
> F <- qf(0.95, ngp-1, n-ngp)
> intconf <- sqrt((ngp-1)*F) * sqrt(varcontrasti)
> intconf

[1] 1.864872 2.153369
```

A partire dalla matrice dei contrasti si calcola il peso per cui moltiplicare ciascun dato:

```
> coeff <- (contrasti/ni)[gp,]
```

e infine si valutano i due contrasti:

```
> cont <- colSums(coeff*dati)
> cont

[1] -2.9166667 -0.1666667
```

Si conclude facilmente che il risultato del primo contrasto è significativo mentre non lo è il secondo.

Il contributo del contrasto alla devianza si calcola come:

```
> cont^2 /varcontrasti * sigma^2

[1] 34.02777778  0.08333333
```

Dato che i due confronti sono ortogonali, nel senso precisato sopra, si verifica che questi due numeri sommano a 34.111, ossia alla devianza del fattore esaminato. $\square$

Lo stesso risultato si può raggiungere molto più velocemente utilizzando la libreria *multcomp*. Il procedimento sarebbe stato il seguente:

```
> library(multcomp)      # carica la libreria necessaria
> a <- c(1,1,2,3,5,3)
> b <- c(1,2,4,2,2,5)
> c <- c(5,5,6,7,4,6)
> dati <- c(a, b, c)
> gp <- factor(rep(LETTERS[1:3], each=6))
> contrasti <- rbind(c(1/2,1/2,-1), c(1,-1,0))  # contrasti per riga
> mod <- aov(dati ~ gp)
> rescont <- glht(mod, linfct = mcp(gp = contrasti))
> confint(rescont)
[...]
Linear Hypotheses:
       Estimate lwr     upr
1 == 0 -2.9167  -4.6169 -1.2164
2 == 0 -0.1667  -2.1299  1.7966

95% family-wise confidence level
```

In output si osservano il valore dei contrasti (che risultano identici a quelli calcolati manualmente) e l'intervallo di confidenza al 95%. La funzione *glht* appartiene alla classe di contrasti multipli di tipo single-step, basati appunto sulla costruzione di un intervallo di confidenza simultaneo per i contrasti controllando l'errore globale di tipo I o family-wise error rate (FWER). Questi metodi garantiscono che il FWER non ecceda il livello α desiderato. Anche in questo caso si può calcolare il contributo di ciascun contrasto alla varianza dovuta al fattore; il calcolo risulta semplicemente:

```
> sigma <- summary(lm(dati ~ gp))$sigma
> coef(rescont)^2/diag(vcov(rescont)) * sigma^2
          1           2
34.02777778  0.08333333
```

Nel caso di disegni non bilanciati si procede come nel caso seguente.

Esempio

Si abbiano i tre gruppi:

```
> a <- c(1,1,2,3,5,3,2)
> b <- c(1,2,4,2,2,5)
> c <- c(5,5,6,7,4,6)
```

Si eseguano i contrasti $a \cup b$ vs. c e a vs. b.

La forma generale dei contrasti $\mathbf{c}$ e $\mathbf{d}$ è in questo caso:

$$\begin{aligned} \mathbf{c} &= (c_1, c_1, c_3) \\ \mathbf{d} &= (d_1, d_2, 0) \end{aligned} \tag{4.8}$$

Imponendo la condizione data dall'Eq. (4.3) si ottengono le equazioni:

$$\begin{aligned} c_1 &= -\frac{c_3 n_c}{n_a + n_b} \\ d_1 &= -\frac{d_2 n_b}{n_a} \end{aligned} \tag{4.9}$$

che portano ai due contrasti:

$$\begin{aligned} L_1 &= \frac{c_3 n_c}{n_a + n_b}(m_c(n_a + n_b) - m_a n_a - m_b n_b) \\ L_2 &= d_2 n_b (m_b - m_a). \end{aligned} \tag{4.10}$$

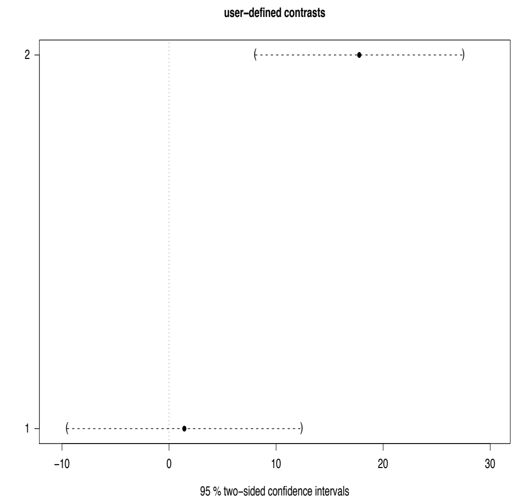

Figura 4.1: Contrasti multipli nel caso dell'esempio 4.6.

Le varianze dei due contrasti risultano rispettivamente:

$$\begin{aligned}\mathrm{var}L_1 &= \sigma^2 \frac{c_3^2 n_c}{n_a + n_b}(n_a + n_b + n_c) \\ \mathrm{var}L_2 &= \sigma^2 \frac{d_2^2 n_b}{n_a}(n_a + n_b) \end{aligned} \tag{4.11}$$

e quindi i contributi dei due contrasti alla devianza del fattore si calcolano come:

$$\begin{aligned} d_{F1}^2 &= \frac{n_c(m_a n_a + m_b n_b - m_c(n_a + n_b))^2}{(n_a + n_b)(n_a + n_b + n_c)} \\ d_{F2}^2 &= \frac{n_a n_b (m_a - m_b)^2}{n_a + n_b}. \end{aligned} \tag{4.12}$$

Con un po' di algebra si dimostra che le due componenti sommano alla devianza complessiva attribuibile al fattore in esame.

Tornando al problema in esempio, ponendo $c_3 = d_2 = 1$, si ha:

```
> ma <- mean(a)
> mb <- mean(b)
> mc <- mean(c)
> na <- 7
> nb <- 6
> nc <- 6
> L1 <- (-ma*na-mb*nb + mc*(na+nb))*nc/(na+nb)
> L2 <- nb*(mb-ma)
> vL1 <- nc*(na+nb+nc)/(na+nb)
> vL2 <- nb*(na+nb)/na
> L <- c(L1, L2)
> L

[1] 17.769231  1.428571
> vL <- c(vL1, vL2)
> sum(L^2/vL)

[1] 36.18922
```

E si verifica facilmente che questo numero coincide con la devianza del fattore *gp*:

```
> dati <- c(a,b,c)
> gp <- factor(c(rep("A",na), rep("B",nb), rep("C", nc)))
> anova(res <- lm(dati ~ gp))

Analysis of Variance Table

Response: dati
          Df Sum Sq Mean Sq F value  Pr(>F)
gp         2 36.189  18.095  10.141 0.00143 **
Residuals 16 28.548   1.784
```

Si noti che la funzione *glht* può essere chiamata anche in questo caso, dopo aver opportunamente impostato la matrice dei contrasti (ponendo anche in questo caso $c_3 = d_2 = 1$). In questo caso la struttura della matrice tiene conto anche della numerosità dei gruppi:

$$\begin{aligned} \mathbf{c} &= (n_a c_1, n_b c_1, n_c c_3) \\ \mathbf{d} &= (n_a d_1, n_b d_2, 0) \end{aligned} \tag{4.13}$$

con c_1 e d_1 dati in Eq. 4.9. Si ha quindi:

$$c_1 = -\frac{1 \cdot 6}{7+6} = -\frac{6}{13} \quad , \quad d_1 = -\frac{1 \cdot 6}{7} = -\frac{6}{7}$$

La matrice dei contrasti è quindi:

```
> contr <- rbind( c(-7*6/13, -6*6/13, 6), c(-6, 6, 0))
```

I test si eseguono con le chiamate:

```
> rescont <- glht(res, linfct = mcp(gp = contr))
> confint(rescont)
[...]
Linear Hypotheses:
       Estimate lwr     upr
1 == 0 17.7692   8.0460 27.4924
2 == 0  1.4286  -9.5318 12.3890

95% family-wise confidence level
```

Si nota che i contrasti risultano uguali al caso precedente (per la scelta appropriata della matrice *contr*). Si può anche verificare che il quadrato dell'errore standard ($res1\$sd^2$) diviso per la varianza d'errore σ^2 risulta identico ai valori calcolati in vL in precedenza. La libreria *multcomp* possiede anche una funzione *plot*:

```
> plot(ct)
```

che produce in output il grafico di Fig. 4.1.

4.7 ANOVA a due vie senza repliche

In questo caso i dati possono essere classificati secondo due chiavi di classificazione e disposti in matrice. La logica del test non cambia, dato che occorrerà semplicemente costruire il modello lineare appropriato e fittarlo. Il modello lineare si scrive come:

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad i = 1, \ldots, r \quad j = 1, \ldots, c$$

con $\varepsilon_{ij} \sim N(0, \sigma^2)$. α_i esprime l'effetto dovuto all'appartenenza all'i-esimo livello del fattore di riga, mentre β_j è l'effetto dovuto al livello j-esimo del fattore di colonna.