Tutoriel : Analyse de la variance et de la covariance avec R
Master 1 EDSB

2023-2024

Plan du document

Ce document est un complément R au cours *modèle linéaire* du M1 ESDB.

Références (en plus des slides du cours) :

J.J. Faraway (2002) Practical Regression and Anova using R.
C. Jost, http://cbi-toulouse.fr/fr/page-personnelle-14#enseignement.

Outline

Quelle matrice de design est utilisée par défaut dans R ?

Exemple 1 : ANOVA un facteur

```
Donnees
   virus yield
1     cc  28.5
2     cc  21.7
3     cc  23.0
4     fc  14.9
5     fc  10.6
6     fc  13.1
```

Modèle :

$$y_{ij} = \mu + \alpha_j + e_{ij},$$

avec $1 \leq i \leq 3$, $1 \leq j \leq 2$.

Quelle matrice de design est utilisée par défaut dans R ?

Exemple 1 : ANOVA un facteur (suite)

Matrice X associée telle qu'introduite en cours :

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

qui sera multipliée par vecteur $\theta = {}^t(\mu, \alpha_1, \alpha_2)$

Dans R ?

```
> mod<-lm(yield~virus,data=Donnees)
> X <- model.matrix(mod)
```

Quelle matrice de design est utilisée par défaut dans R ?

Exemple 1 : ANOVA un facteur (suite)

```
 X       (Intercept) virusfc
            1        0
            1        0
            1        0
            1        1
            1        1
            1        1
```

qui est donc différente de celle vue en cours . . .

Le vecteur des paramètres est alors : $\theta = {}^t(\mu, \alpha_2)$

$\hookrightarrow$ dans R, la contrainte d'estimation $\alpha_1 = 0$ est donc explicitement prise en compte dans la matrice de design X du modèle.

Quelle matrice de design est utilisée par défaut dans R ?

Exemple 2 : anova 2 facteurs sans interaction

```
y <-  rnorm(4,0,1)
data <-  data.frame(y=y,A=c("A1","A2","A1","A2"),
                                    B=c("B1","B1","B2","B2"))
data
         y        A   B
1 -0.1501903    A1  B1
2 -0.2687818    A2  B1
3  1.7913320    A1  B2
4  0.6722680    A2  B2
```

Modèle considéré (additif) :

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

avec $1 \leq i \leq 2$, $1 \leq j \leq 2$.

Quelle matrice de design est utilisée par défaut dans R ?

Dans R :

```
model<-lm(data$y~data$A+data$B)
X <- model.matrix(model)
X
  (Intercept) data$AA2 data$BB2
            1        0        0
            1        1        0
            1        0        1
            1        1        1
```

qui, là encore est d'écriture différente de celle vue en cours . . .

Quelle matrice de design est utilisée par défaut dans R ?

Pour l'exemple 2, l'approche cours donne $X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix}$ (multipliée par vecteur $\theta = {}^t(\mu, \alpha_1, \alpha_2, \beta_1, \beta_2)$), et l'approche R considère $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ qui sera multipliée par vecteur $\theta = {}^t(\mu, \alpha_2, \beta_2)$, car les paramètres α_1 et β_1 sont explicitement pris comme égal à 0.

$\hookrightarrow$ plutôt que de travailler avec une matrice de design construite comme vu en cours et qui tiendrait compte que l'on a les paramètres μ, α_1, α_2 pour l'exemple 1 ou μ, α_1, α_2, β_1 et β_2 pour l'exemple 2, *on exploite ici directement les contraintes (par défaut) $\alpha_1 = 0$ (ANOVA 1 facteur), $\alpha_1 = 0$ et $\beta_1 = 0$ (ANOVA 2 facteurs additifs), ce qui fait que seuls μ, α_2 sont à considérer dans l'exemple 1 et μ, α_2 et β_2 dans l'exemple 2.*

Choix des contraintes dans R ?

C'est l'option *contrasts* qui permet de gérer les contraintes (à noter, R distingue contrainte pour les facteurs non ordonnés et contrainte pour les facteurs ordonnés). La contrainte par défaut, (dit *contraste contr.treatment*) est l'égalité à 0 de paramètres (typiquement, par défaut la première modalité par ordre alphabétique). Pour changer une contrainte, on agit soit directement dans la syntaxe de la commande *lm* pour un changement limité à l'analyse en question, soit pour toute la session en cours via la commande $options(contrasts = c(\cdot, \cdot)$. Un système de contraintes aussi couramment utilisé consiste à fixer la somme des effets à 0, c'est l'option *contr.sum*

```
#Options par défaut :
options("contrasts")
$contrasts
        unordered           ordered
"contr.treatment"      "contr.poly"
# changement
options(contrasts=c("contr.sum","contr.sum"))
options("contrasts")
$contrasts
[1] "contr.sum" "contr.sum"
```

Pour comprendre à quoi correspondent les différents contrastes considérés dans R, voir en annexe 1.

Outline

ANOVA à un facteur : exemple

Cadre général :
on dispose de p échantillons de tailles $n_1, \ldots, n_p$, correspondant par exemple aux observations d'une variable réponse obtenues selon p modalités d'un facteur. Soit $N = \sum_{i=1}^{p} n_i$ le nombre total d'observations.

Ex : on a a mesuré 24 temps de coagulation du sang en fonction d'un régime alimentaire. Les données sont disponibles sous la forme d'un dataframe, une colonne *numeric* et une autre *factor*, dans le package *faraway*.

```
library(faraway)
data(coagulation)

coagulation # jeu données

xtabs(~coagulation$diet) # effectifs modalités
```

ANOVA à un facteur

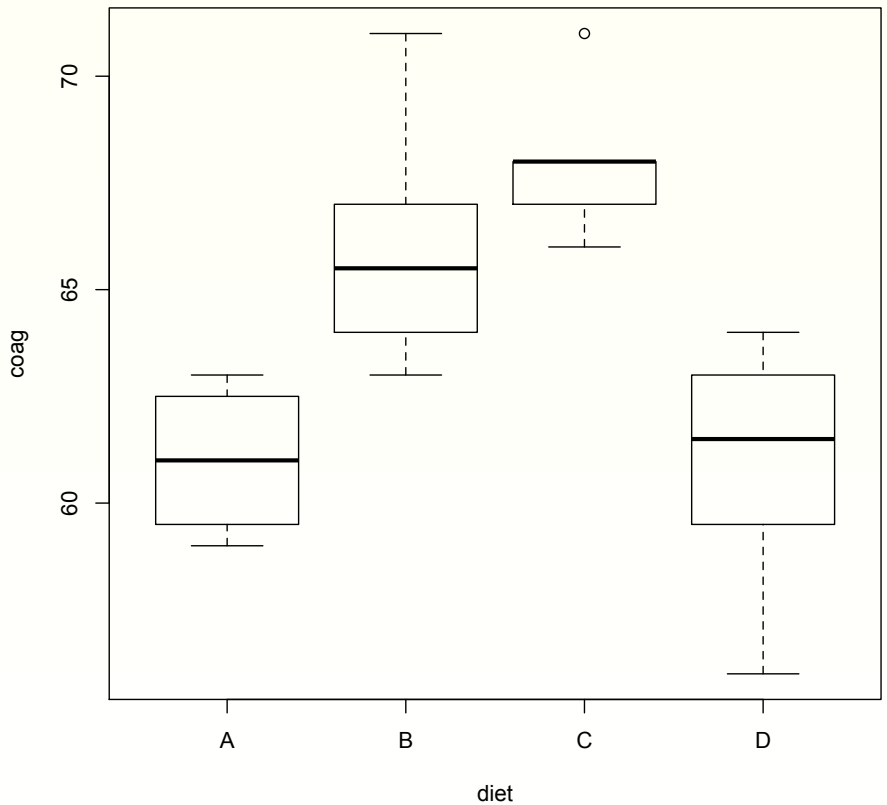
	coag	diet
1	62	A
2	60	A
3	63	A
4	59	A
5	63	B
6	67	B
7	71	B
.		
.		
.		
.		
18	62	D
19	60	D
20	61	D
21	63	D
22	64	D
23	63	D
24	59	D

diet :	A	B	C	D
effectif :	4	6	6	8

ANOVA à un facteur : exemple

Descriptif graphique des résultats de l'expérience :

```
plot(coag~diet,data=coagulation)
```



ANOVA : tableau d'analyse de la variance - en pratique avec R

```
> mod=lm(coag~diet,coagulation)# ajuster le modele
> anova(mod) #obtenir la table d'anova associee au modele

Analysis of Variance Table

Response: coag
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       3    228    76.0  13.571 4.658e-05 ***
Residuals 20    112     5.6
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

ANOVA : résumé quantitatif ajustement - en pratique avec R

```
# obtenir un resume ajustement modele : estimations parametres, r^2,
#   p-value du modele
> summary(mod)
Residuals:
   Min     1Q Median     3Q    Max
 -5.00  -1.25   0.00   1.25   5.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
dietB       5.000e+00  1.528e+00   3.273 0.003803 **
dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
dietD       2.991e-15  1.449e+00   0.000 1.000000
---

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-squared: 0.6706,Adjusted R-squared: 0.6212
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

Retour sur l'estimation

Rappel : estimations obtenues

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
dietB       5.000e+00  1.528e+00   3.273 0.003803 **
dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
dietD       2.991e-15  1.449e+00   0.000 1.000000
```

A quoi correspondent les valeurs Estimate ?

```
> str(coagulation$diet)
 Factor w/ 4 levels "A","B","C","D": 1 1 1 1 2 2 2 2 2 2 ...
 Ici regime de ref : "A" car modalite num 1
 > coagulation$coag[coagulation$diet=='A']
 [1] 61 # intercept
> mean(coagulation$coag[coagulation$diet=="B"])-mean(
coagulation$coag[coagulation$diet=="A"])
[1] 5 # retrouve estimation pour regime B
> mean(coagulation$coag[coagulation$diet=="C"])-mean(
coagulation$coag[coagulation$diet=="A"])
[1] 7  # retrouve estimation pour regime C
```

Retour sur l'estimation

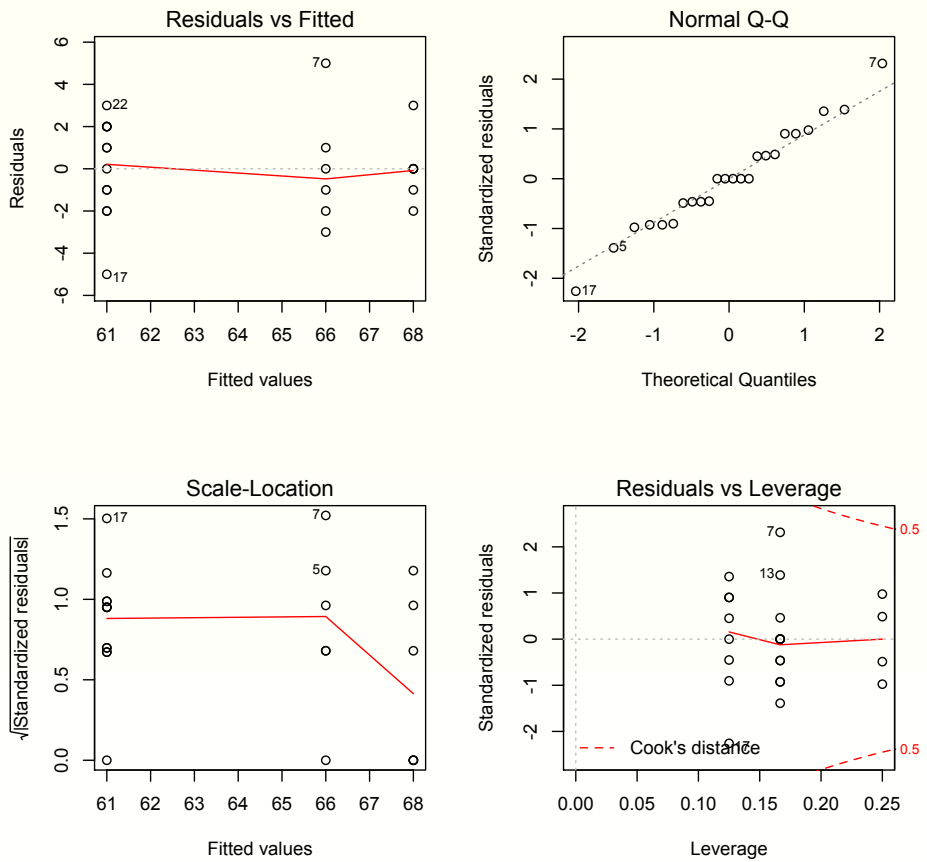
Changer la modalité de référence ?

```
> coagulation$diet<-relevel(coagulation$diet,"C")
> str(coagulation$diet)
 Factor w/ 4 levels "C","A","B","D": 2 2 2 2 3 3 3 3 3 3 ...
> mod2=lm(coag~diet,coagulation)
> summary(mod2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.0000     0.9661  70.387  < 2e-16 ***
dietA        -7.0000     1.5275  -4.583 0.000181 ***
dietB        -2.0000     1.3663  -1.464 0.158776
dietD        -7.0000     1.2780  -5.477 2.32e-05 ***

> mean(coagulation$coag[coagulation$diet=='C']  )
[1] 68
```

Validation du modèle : etudes des résidus avec R

```
> par(mfrow=c(2,2))
> plot(mod)
```

Etudes des résidus - En pratique avec R

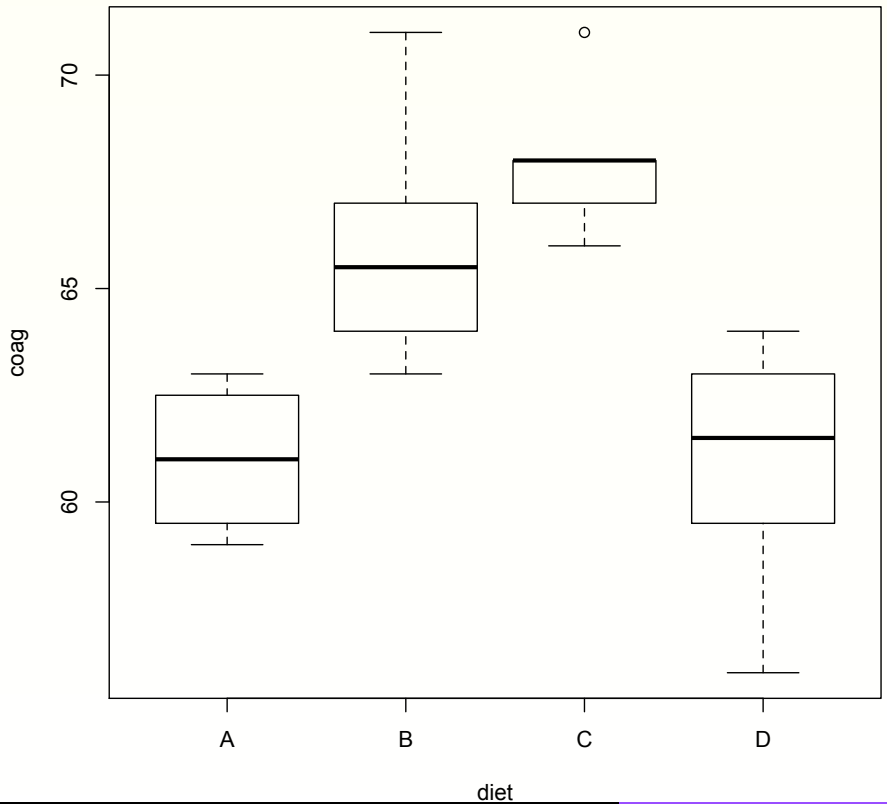


Mise en pratique R - Graphes associés - Validation graphique

- Le premier graphique trace les résidus en fonction des valeurs ajustées avec la tendance moyenne tracée en rouge et les points sortant potentiellement d'une distribution normale indiqués par un numéro. On s'attend à une ligne rouge plus ou moins horizontale (pas de tendance dans les résidus) et pas plus de 5% des points marqués. Par ailleurs, la variabilité dans les groupes doit être comparable (hypothèse de variance constante).
- Le graphique 2 (QQ-plot) permet de vérifier graphiquement l'hypothèse de normalité des résidus : si les points centraux sont à peu près alignés selon la première bissectrice des axes, on peut considérer que les résidus ne s'écartent pas trop d'une loi normale.
- Le graphique 3 répète le premier dans une autre échelle : les résidus sont standardisés et une loupe est mise sur la fluctuation sous-jacente. Cela permet de mieux apprécier les hypothèses de non-corrélation entre les résidus et de variance constante.
- Le graphique 4 (Cooks D) permet de repérer les points qui ont une (trop) forte influence (au sens de la distance de Cook) sur le modèle obtenu. Des valeurs dans la zone rouge identifient des points qui influencent beaucoup la valeur des paramètres estimés.

Comparaisons multiples - En pratique avec R

Si on rejette H_0, on sait que toutes les moyennes ne sont égales. Mais on aimerait alors des précisions... quelle(s) modalité(s) de la variable qualitative a provoqué ce rejet ? $\rightsquigarrow$ comparaisons multiples pour contrôle du risque de 1ère espèce global.



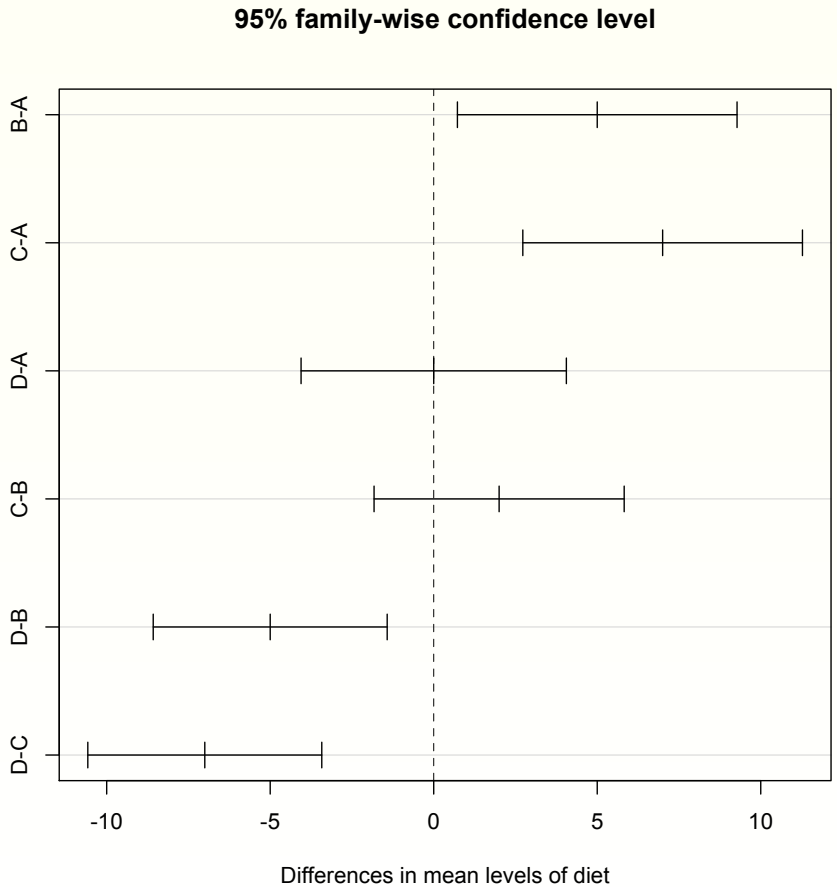
Comparaisons multiples - En pratique avec R

Une adaptation du test de Student : l'approche de Tukey.

```
> TukeyHSD(aov(coag~diet,coagulation)) #la fonction TukeyHSD() ne
#reconnait pas les objets lm() mais uniquement aov()


  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = coag ~ diet, data = coagulation)

$diet
    diff         lwr       upr     p adj
B-A    5   0.7245544  9.275446 0.0183283
C-A    7   2.7245544 11.275446 0.0009577
D-A    0  -4.0560438  4.056044 1.0000000
C-B    2  -1.8240748  5.824075 0.4766005
D-B   -5  -8.5770944 -1.422906 0.0044114
D-C   -7 -10.5770944 -3.422906 0.0001268
```

Comparaisons multiples - En pratique avec R

```
> plot(TukeyHSD(aov(coag~diet,coagulation)))
```



Comparaisons multiples - En pratique avec R

On peut aussi utiliser

```
> pairwise.t.test(coagulation$coag, coagulation$diet,p.adj="bonferroni"
```

pour une approche Bonferroni.

```
Pairwise comparisons using t tests with pooled SD

data:  coagulation$coag and coagulation$diet

  A       B       C
B 0.02282 -       -
C 0.00108 0.95266 -
D 1.00000 0.00518 0.00014

P value adjustment method: bonferroni
```

D'autres approches que Bonferroni sont disponibles (voir le help).

Comparaisons multiples - En pratique avec R

On peut aussi utiliser

```
> pairwise.t.test(coagulation$coag, coagulation$diet,p.adj="holm")


Pairwise comparisons using t tests with pooled SD

data:  coagulation$coag and coagulation$diet

  A       B       C
B 0.01141 -       -
C 0.00090 0.31755 -
D 1.00000 0.00345 0.00014

P value adjustment method: holm
```

Comparaisons multiples

The adjustment methods include the Bonferroni correction ('"bonferroni"') in which the p-values are multiplied by the number of comparisons. Less conservative corrections are also included by Holm (1979) ('"holm"'), Hochberg (1988)('"hochberg"'), Hommel (1988) ('"hommel"'), Benjamini & Hochberg (1995) ('"BH"'), and Benjamini & Yekutieli (2001) ('"BY"'), respectively.

The first four methods are designed to give strong control of the family wise error rate. There seems no reason to use the unmodified Bonferroni correction because it is dominated by Holm's method, which is also valid under arbitrary assumptions.

Comparaisons multiples

Hochberg's and Hommel's methods are valid when the hypothesis tests are independent or when they are non-negatively associated (Sarkar, 1998; Sarkar and Chang, 1997).
Hommel's method is more powerful than Hochberg's, but the difference is usually small and the Hochberg p-values are faster to compute.

The '"BH"' and '"BY"' method of Benjamini, Hochberg, and Yekutieli control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. The false discovery rate is a less stringent condition than the family wise error rate, so these methods are more powerful than the others.

Comparaisons multiples - approche avec R

• Approches comparaisons de groupes de moyennes : Student-Newman-Keuls (SNK) et Duncan.

SNK :

```
> library(agricolae)
> compar_SNK<- SNK.test(mod,"diet",coagulation)
> compar_SNK
# ou pour affichage direct des résultats
# SNK.test(mod,"diet",coagulation, console=TRUE)

Study: mod ~ "diet"
Student Newman Keuls Test for coag
Mean Square Error:  5.6
```

Comparaisons multiples - approche avec R

```
diet,  means
  coag       std r Min Max
A   61 1.825742 4  59  63
B   66 2.828427 6  63  71
C   68 1.673320 6  66  71
D   61 2.618615 8  56  64

Groups according to probability of means differences and
alpha level (0.05)

Means with the same letter are not significantly different.

coag groups
C   68      a
B   66      a
A   61      b
D   61      b
```

Comparaisons multiples - approche avec R

Duncan :

```
> duncan.test(mod,"diet",coagulation,console=TRUE)

Study: mod ~ "diet"
Duncan's new multiple range test
for coag

.
.
.
Means with the same letter are not significantly different.

  coag groups
C   68      a
B   66      a
A   61      b
D   61      b
```

Comparaisons multiples - approche avec R

Comparaison à un témoin : approche Dunnett.

```
> library("multcomp")# librairie "multiples comparisons"
> compar_Dun <- glht(mod2, linfct = mcp(diet = "Dunnett"))
> summary(compar_Dun)

 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts

Fit: lm(formula = coag ~ diet, data = coagulation)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
A - C == 0   -7.000      1.528  -4.583   <0.001 ***
B - C == 0   -2.000      1.366  -1.464    0.351
D - C == 0   -7.000      1.278  -5.477   <0.001 ***
---
(Adjusted p values reported -- single-step method)
```

Comparaisons multiples - approche avec R

Approche *constrates* (au sens vu en cours : combinaison linéaire des paramètres dont la somme des coefficients est nulle) :

```
library(gmodels)
attach(coagulation)

contrast2<- rbind(" : 1 versus 4"=c(-1,0,0,1),
       " 1+2 versus 3+4" = c(0.5,0.5, -0.5, -0.5))

# deux contrastes :

#tester si regime1 et regime4 sont différents ;

# tester moyenne (regime1, regime2)!= moyenne (regime3, regime4)
```

Comparaisons multiples - approche avec R

```
> fit.contrast(mod2,diet,contrast2)

                    Estimate Std. Error   t value     Pr(>|t|)
diet : 1 versus 4         -7  1.2780193 -5.477226 0.0000231827
diet 1+2 versus 3+4        1  0.9958246  1.004193 0.3272812992
attr(,"class")
[1] "fit_contrast"

# mais p-values non ajustées au nombre de tests faits ...
# on calcule les p-values ajustees selon approche à preciser

> p.adjust(c(0.0000231827,0.3272812992), "holm")
[1] 0.0000463654 0.3272812992

> p.adjust(c(0.0000231827,0.3272812992), "bonferroni")
[1] 0.0000463654 0.6545625984
```

Outline

Anova à deux facteurs - Contexte

On s'intéresse désormais aux potentiels effets du croisement de 2 facteurs A et B (à I et J niveaux respectivement) sur 1 variable réponse quantitative.

Exemples :

- Production de lait en fonction de l'exploitation (facteur A) et de la race des vaches (facteur B).
- Taux de cholestérol en fonction de la CSP (facteur A) et du sexe (facteur B).

Objectifs :

On cherche à caractériser un éventuel effet sur la réponse du facteur A, du facteur B tout en considérant si c'est possible un possible effet conjoint de A et B.

Notations

Tester le modèle avec interaction contre le modèle constant M_0 consiste à tester si au moins l'un des paramètres du modèle considé est significativement non-nul.

Sous l'hypothèse d'un test global significatif, les tests suivants seront considérés via une approche comparaisons de modèles emboités (les réécrire !).

$H_{0,1}$: $\alpha_1 = \alpha_2 = \ldots \alpha_p = 0$ (pas d'effet du facteur A).

$H_{1,1}$: Il existe un effet du facteur A

$H_{0,2}$: $\beta_1 = \beta_2 = \ldots \beta_p = 0$ (pas d'effet du facteur B)..

$H_{1,2}$: Il existe un effet du facteur B

$H_{0,3}$: $\gamma_{11} = \gamma_{12} = \ldots \gamma_{pq} = 0$ (pas d'effet de l'interaction).

$H_{1,3}$: Il existe un effet de l'interaction.

Mise en pratique sur R : l'ANOVA à deux facteurs

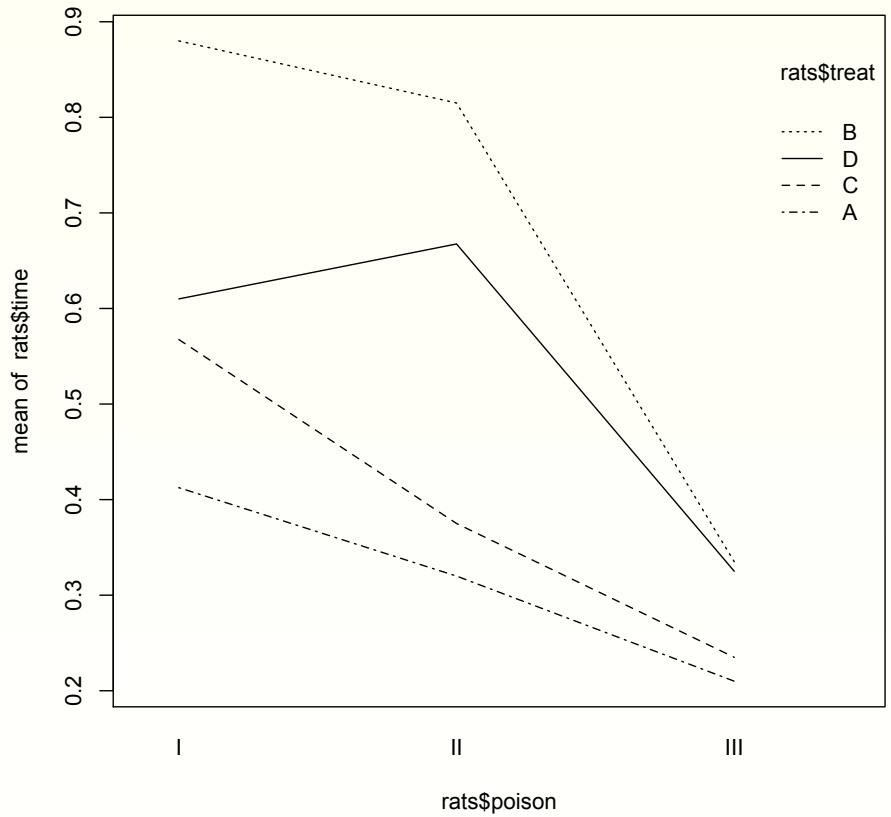
Données sur temps de survie de 48 rats soumis à 3 doses de poison (I, II, III) et 4 types de traitements (A,B,C,D).

```
>library(faraway)
>data(rats)
> rats
   time poison treat
1  0.31      I     A
2  0.82      I     B
3  0.43      I     C
4  0.45      I     D
5  0.45      I     A
6  1.10      I     B
7  0.45      I     C
8  0.71      I     D
9  0.46      I     A
10 0.88      I     B
11 0.63      I     C
12 0.66      I     D
13 0.43      I     A
14 0.72      I     B
15 0.76      I     C
.
.
47 0.22    III     C
48 0.33    III     D
```

Mise en pratique sur R : l'ANOVA à deux facteurs

```
> str(rats)
'data.frame': 48 obs. of  3 variables:
 $ time  : num  0.31 0.82 0.43 0.45 0.45 1.1 0.45 0.71 0.46 0.88 ...
 $ poison: Factor w/ 3 levels "I","II","III": 1 1 1 1 1 1 1 1 1 1 ...
 $ treat : Factor w/ 4 levels "A","B","C","D": 1 2 3 4 1 2 3 4 1 2 ...

 > interaction.plot(rats$poison,rats$treat,rats$time)
```

Mise en pratique sur R : l'ANOVA à deux facteurs



Mise en pratique sur R : l'ANOVA à deux facteurs

```
> mod=lm(time~poison*treat,rats)
> anova(mod)

Analysis of Variance Table

Response: time
             Df  Sum Sq Mean Sq F value    Pr(>F)
poison        2 1.03301 0.51651 23.2217 3.331e-07 ***
treat         3 0.92121 0.30707 13.8056 3.777e-06 ***
poison:treat  6 0.25014 0.04169  1.8743    0.1123
Residuals    36 0.80073 0.02224
---
```

La table permet de conclure quant aux effets de l'interaction et de la suite à donner ...

Pour obtenir $p-$value du modèle, estimations des paramètres, R^2 etc, on utilise à nouveau la fonction *summary(·)*

Mise en pratique sur R : l'ANOVA à deux facteurs

```
> summary(mod)
Call:
lm(formula = time ~ poison * treat, data = rats)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32500 -0.04875  0.00500  0.04312  0.42500
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.41250    0.07457   5.532 2.94e-06 ***
poisonII         -0.09250    0.10546  -0.877   0.3862
poisonIII        -0.20250    0.10546  -1.920   0.0628 .
treatB            0.46750    0.10546   4.433 8.37e-05 ***
treatC            0.15500    0.10546   1.470   0.1503
treatD            0.19750    0.10546   1.873   0.0692 .
poisonII:treatB   0.02750    0.14914   0.184   0.8547
poisonIII:treatB -0.34250    0.14914  -2.297   0.0276 *
poisonII:treatC  -0.10000    0.14914  -0.671   0.5068
poisonIII:treatC -0.13000    0.14914  -0.872   0.3892
poisonII:treatD   0.15000    0.14914   1.006   0.3212
poisonIII:treatD -0.08250    0.14914  -0.553   0.5836
---
Residual standard error: 0.1491 on 36 degrees of freedom
Multiple R-squared: 0.7335,Adjusted R-squared: 0.6521
F-statistic:  9.01 on 11 and 36 DF,  p-value: 1.986e-07
```

Mise en pratique sur R : l'ANOVA à deux facteurs

D'après la table d'ANOVA, l'interaction est NS $\rightsquigarrow$ continuer avec un modèle additif pour tester les effets directs de A et B.

```
> mod2=lm(time~poison+treat,rats)
```

et ensuite analyse des résidus, et si ok, pour facteur(s) significatif(s), approche comparaisons multiples ou contrastes ...

Et si intercation significative ? ?

- $\hookrightarrow$ interprétation directe des effets principaux des facteurs "impossible" !
- $\hookrightarrow$ deux possibilités : 1) création d'un nouveau facteur dont les modalités seront les traitements associés au croisement des deux facteurs sur lequel une analyse de variance à un facteur *(forcément significative ici)* sera menée et où des contrastes adaptés aux questions de comparaison que l'on souhaite aborder (comparaisons de modalités d'un facteur à modalité de l'autre fixée) seront utilisés.
  2) travailler sous un niveau de facteur fixé : anova sur un sous-ensemble du jeu de données initial (celles vérifiant la modalité que l'on s'est fixée pour l'un des facteurs) qui va donner une p-value qu'il faudra corriger en tenant compte de la variance résiduelle globale à la place de celle partielle obtenue sur le sous-ensemble de données (voir exemple dans le dossier Recap joint).

Outline

ANCOVA : qu'est-ce que c'est ?

Cette section reprend une partie du polycopié de Christian Jost disponible à l'adresse ($http://cognition.ups-tlse.fr/\_christian/poly/polycopies.html$).

ANCOVA signifie analyse de la covariance.

Dans les ANOVA vues précédemment, on cherche à comprendre l'effet d'une ou plusieurs variables qualitatives (facteurs) sur une variable quantitative.

Il arrive qu'une autre variable (quantitative) varie au long de l'expérience.

Ceci ajoute de la variabilité et il convient alors de la prendre en considération. Elle sera appelée la covariable.

Exemple : Durée de trajet d'une fourmi sur un pont entre le nid et la nourriture ($http://cognition.ups-tlse.fr/\_christian/poly/polycopies.html$)

- On se demande si cette durée augmente avec la longueur du pont.
- Sur un pont long, une fourmi va croiser en moyenne beaucoup plus de fourmis que sur un pont court et va donc avoir tendance à s'arreter plus souvent.
- Une augmentation de la durée du passage pourrait donc être due à l'augmentation du nombre de rencontres plutôt qu'à la longueur du pont.
- Il faut donc dissocier l'effet de la longeur du pont (facteur à 2 modalités, long ou court) et l'effet du nombre de rencontres (variable quantitative appelée covariable).
- Une ANCOVA va permettre de faire cette dissociation.

Exemple : Durée de trajet d'une fourmi sur une branche entre le nid et la nourriture

Réponse : Temps de trajet d'une fourmi entre nid et source de nourriture

Variables :
Type de branches : qualitative à deux niveaux (courte ; longue) ;
Nombre de contacts pendant le trajet avec autres fourmis : quantitative.

Une ANCOVA combine les techniques d'ANOVA et de régression linéaire (voir plus loin) et permet de tester les hypothèses suivantes :

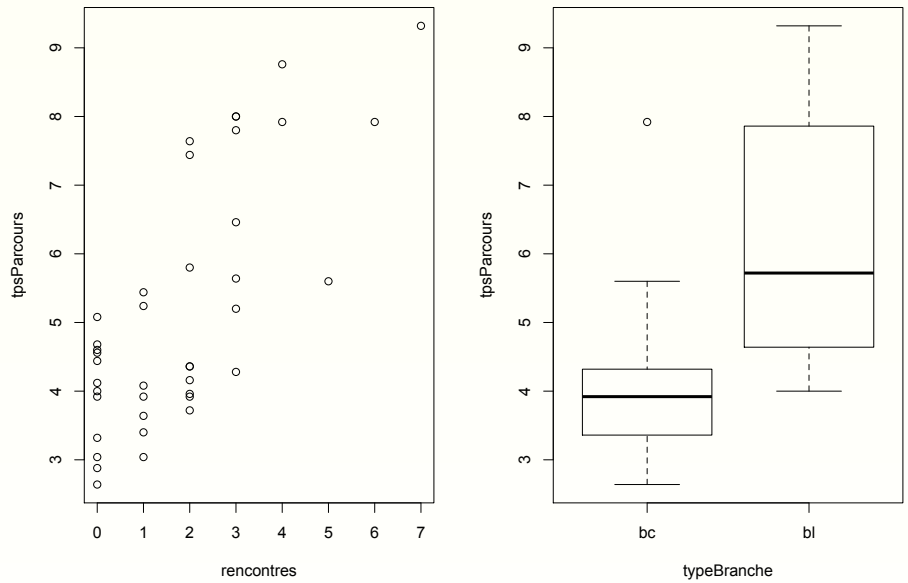
- $H_{0,1}$: Pas d'effet *type de la branche* sur la durée du trajet
- $H_{0,2}$: Pas d'effet *nombre de contacts* sur la durée du trajet
- $H_{0,3}$: Pas d'effet interaction entre *type de la branche* et *nombre de contacts* sur la durée du trajet

Exemple

```
> load("traffic.rda")
> trafficTempsParcours; attach(trafficTempsParcours)

   tpsParcours rencontres typeBranche
1         3.72          2          bc
2         4.08          1          bc
3         3.96          2          bc
.            .          .           .
.            .          .           .
.            .          .           .
38        4.44          0          bl
39        5.64          3          bl
40        4.00          0          bl
```

Exemple

```
> par(mfrow=c(1,2))
> plot(tpsParcours~rencontres)
> plot(tpsParcours~typeBranche)
```



On voit que le temps de passage augmente aussi bien avec la longueur de la branche (quantitative) qu'avec le nombre de rencontres.

Exemple

```
> mod2<-lm(tpsParcours~typeBranche*rencontres)
> anova(mod2)
Analysis of Variance Table

Response: tpsParcours
                      Df Sum Sq Mean Sq  F value    Pr(>F)
typeBranche            1 47.480  47.480 112.6355 1.238e-12 ***
rencontres             1 64.738  64.738 153.5753 1.504e-14 ***
typeBranche:rencontres 1  0.700   0.700   1.6617    0.2056
Residuals             36 15.175   0.422
---
```

Pas d'effet interaction

Exemple

```
> mod3<-lm(tpsParcours~typeBranche+rencontres)
> anova(mod3)
Analysis of Variance Table

Response: tpsParcours
             Df Sum Sq Mean Sq F value    Pr(>F)
typeBranche   1 47.480  47.480  110.66 1.134e-12 ***
rencontres    1 64.738  64.738  150.88 1.273e-14 ***
Residuals    37 15.876   0.429
---
```

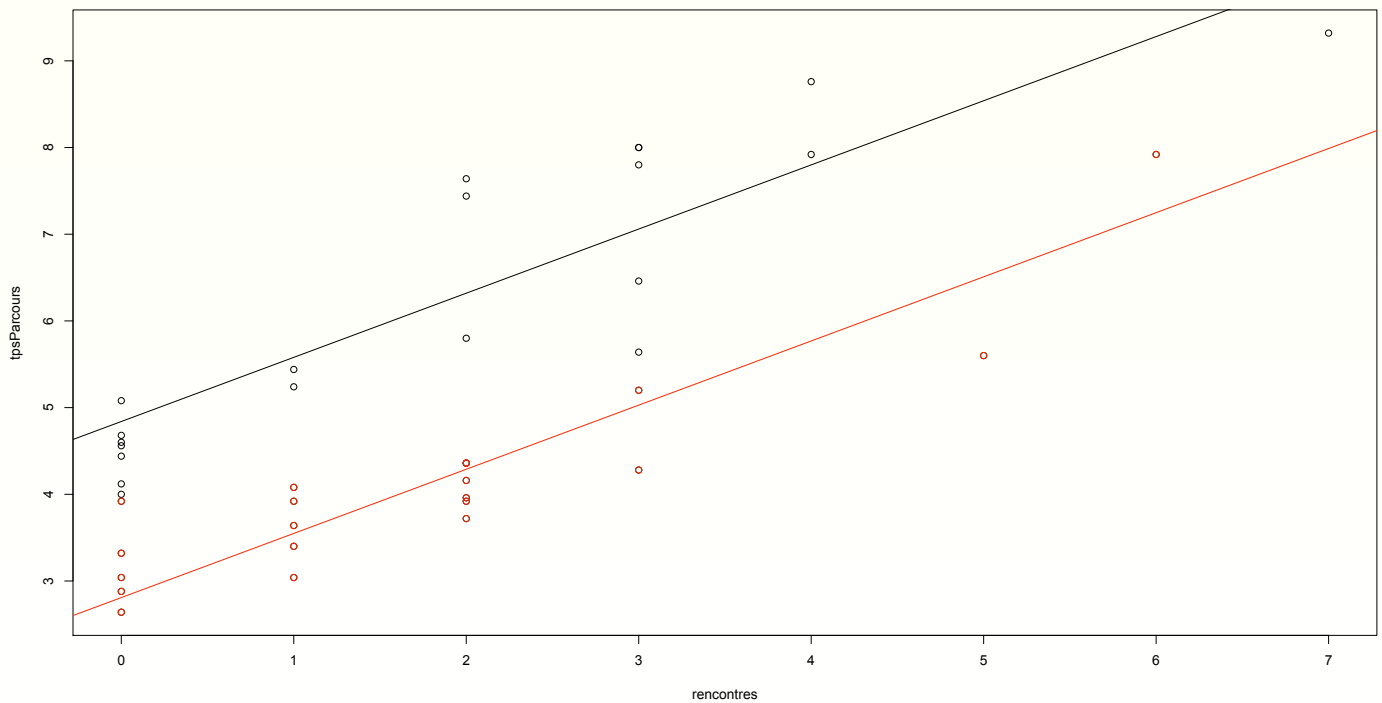
Conclusion : Les 2 facteurs ont un effet significatif sur la réponse.

Exemple

```
> summary(mod3)

Call:
lm(formula = tpsParcours ~ typeBranche + rencontres)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4218 -0.4271 -0.1353  0.4057  1.3189

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.8088     0.1788  15.711  < 2e-16 ***
typeBranchebl  2.0309     0.2075   9.788 8.21e-12 ***
rencontres     0.7407     0.0603  12.283 1.27e-14 ***
---


Residual standard error: 0.655 on 37 degrees of freedom
Multiple R-squared: 0.8761,Adjusted R-squared: 0.8694
F-statistic: 130.8 on 2 and 37 DF,  p-value: < 2.2e-16
```

Exemple

```
> plot(tpsParcours~rencontres)
> points(tpsParcours[typeBranche=="bc"]~
   rencontres[typeBranche=="bc"],col=2)
> abline(2.8088,0.74,col=2)
> abline(2.8088+2.0309,0.74,col=1)
```



ANCOVA : une approche pour éviter les confusions d'effets ?

Supposons que la question initiale était simplement de chercher à voir s'il y avait un effet *type de branche* **:** on aurait pu penser alors à une approche ANOVA à un facteur

```
> mod4<-lm(tpsParcours~typeBranche)
> anova(mod4)
Analysis of Variance Table
Response: tpsParcours
             Df Sum Sq Mean Sq F value    Pr(>F)
typeBranche   1 47.480  47.480  22.381 3.066e-05 ***
Residuals    38 80.614   2.121
```

Comme on peut le voir, la Somme des Carrés due à la prise en compte des modalités de la variable *type de branche* correspond bien à celle du modèle précédent mais ici la p-value du test de l'effet est multipliée par plus de 27×10^6 ! ...

ANCOVA : une approche pour éviter les confusions d'effets ?

le fait de ne pas tenir de la source de variabilité associée à la dépendance qu'il existe entre la réponse et la variable *nombre de rencontres* **pourrait conduire à une interprétation erronée quant à l'éventuel effet de la variable** *type de branche* . . .

Ici l'effet étant "suffisamment" marqué, il ressort même lorsque *nombre de rencontres* n'est pas pris en compte *(mais quid des résidus du modèle ! ?*.

Si l'interaction entre le facteur et le régresseur était significative, les deux pentes du graphe précédent seraient différentes . . . et l'interprétation d'un effet du facteur *type de branche* serait alors impossible car non discernable de l'effet de l'interaction qui pourrait impliquer de facto une différence des ordonnées à l'origine . . .

Outline

A quoi correspondent les contrastes considérés dans R ?

Dans *R*, les notions de contrastes et de contraintes sont confondues : l'option *contrast(·)* détermine directement la façon dont les paramètres de l'ANOVA sont traités et influencent tout aussi directement les interprétations des estimations obtenues (et des tests qui s'y associent), au titre des contraintes fixées pour les estimations des paramètres.

Contraintes et contrastes sous R

Rappel : **R "confond" contrastes et contraintes et traite les contraintes sur les paramètres comme des contrastes particuliers !** Contrastes par défaut :

```
> contrasts(coagulation$diet)=NULL #re-initialiser
> options(contrasts=c('contr.treatment','cont.poly')) #specifier les co
> contrasts(coagulation$diet)
  B C D
A 0 0 0
B 1 0 0
C 0 1 0
D 0 0 1
```

A est pris comme niveau référence (par défaut : ordre alphabétique). La matrice considérée est $(4, 3)$: 3 colonnes signifient que le modèle aura 3 paramètres (en lien avec les 3 modalités B, C et D). La lecture en colonnes permet de caractériser les différences d'espérances qui vont être testées : la présence du 1 à coté de B indique que sur la ligne "estimate de dietB" on trouvera ainsi le test d'hypothèse nulle $\mu_B - \mu_A = 0$; sur "estimate de dietC", test d'hypothèse nulle $\mu_C - \mu_A = 0$. . . **donc écart du groupe par rapport à la référence repéré par la valeur 1 de la colonne.**

Contraintes et contrastes sous R

On peut vérifier (comme déjà vu précédemment) :

```
Coefficients du modèle :
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
dietB       5.000e+00  1.528e+00   3.273 0.003803 **
dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
dietD       2.991e-15  1.449e+00   0.000 1.000000
> mean(coagulation$coag[coagulation$diet=='A'])
 [1] 61
 > coef(mod)[1]
(Intercept)
       61
> mean(coagulation$coag[coagulation$diet=="B"])-mean(
coagulation$coag[coagulation$diet=="A"])
[1] 5
> coef(mod)[2]
dietB
    5
```

Contraintes et contrastes sous R

Si l'on s'intéresse aux lignes de la matrice de design : en considérant la combinaison linéaire des coefficients du modèle associée aux valeurs de la ligne on retrouve les moyennes des groupes :

```
# moyenne groupe A --> linge 1 : 0 0 0
> coef(mod)[1] + 0*coef(mod)[2]+0*coef(mod)[3]+0*coef(mod)[4]
(Intercept)
         61
> mean(coagulation$coag[coagulation$diet=='A'])
 [1] 61

#  moyenne groupe B--> ligne 2 : 1 0 0
> coef(mod)[1] + 1*coef(mod)[2]+0*coef(mod)[3]+0*coef(mod)[4]
(Intercept)
         66
> mean(coagulation$coag[coagulation$diet=="B"])
[1] 66
```

Contraintes et contrastes sous R

Cela montre comment fonctionne l'option *contrasts(contr.treatment) :*

- *en colonne*, on retrouve la notion de contraste vue en cours mais sans que le coefficient -1 s'appliquant à la modalité prise comme témoin n'apparaisse ;
- *en ligne*, estimation des espérances de modalités.

Contraintes et contrastes sous R

Une autre possibilité :*contrasts(contr.sum)* : $\sum_i \alpha_i = 0$

```
> contrasts(diet)<-"contr.sum"
> contrasts(diet)
  [,1] [,2] [,3]
A    1    0    0
B    0    1    0
C    0    0    1
D   -1   -1   -1
```

• Les noms des colonnes ne sont plus associées aux niveaux du facteur ! (juste colonne 1, 2 et 3) . . . la lecture en colonne ne sera pas en lien avec les modalités du facteur ;

• La matrice de design à 3 colonnes donc estimations de 3 paramètres . . .

• Si on multiplie par $\theta = {}^t(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, on retrouve $(\alpha_1, \alpha_2, \alpha_3, -\alpha_1 - \alpha_2 - \alpha_3)$, ce qui intègre directement la contrainte $\sum_{i=1}^{4} \alpha_i = 0$.

Contraintes et contrastes sous R

• du fait de la contrainte, on a forcément (en reprenant l'écriture du modèle) $\sum_{ij} y_{ij} = \sum_{ij} \mu = n\mu$ d'où $\hat{\mu} = y_{..}$ et la moyenne générale va jouer le rôle de valeur de référence (pas d'effet du facteur).

• La lecture en ligne va permettre de retrouver quelle modalité du facteur est comparée à la référence : sur la ligne *diet1* on aura l'estimation du contraste $\mu_A - \mu$; sur celle de *diet2*, estimation du contraste $\mu_B - \mu$ etc

```
> mod_sum=lm(coag~diet,coagulation)
> summary(mod_sum)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.0000     0.4979 128.537  < 2e-16 ***
diet1        -3.0000     0.9736  -3.081 0.005889 **
diet2         2.0000     0.8453   2.366 0.028195 *
diet3         4.0000     0.8453   4.732 0.000128 ***
```

Contraintes et contrastes sous R

L'intercept correspond bien à la moyenne générale

```
> mean(coagulation$coag)
[1] 64
```

pour *diet1*, estimation de $\mu_A - \mu$:

```
> mean(coagulation$coag[coagulation$diet == 'A'])
    -mean(coagulation$coag)
[1] -3
```

et même raisonnement pour ligne 2 : *diet2* estimation de $\mu_B - \mu$:

```
> mean(coagulation$coag[coagulation$diet == 'B'])
    -mean(coagulation$coag)
[1] 2
```

et idem ligne 3, *diet3* pour estimation de $\mu_C - \mu$.

Note : la ligne 4 de la matrice de design n'est pas utilisée pour estimation car du fait de la contrainte l'estimation du paramètre 4 du modèle se déduit de l'estimation des 3 premiers.

Contraintes et contrastes sous R

Une autre possibilité :*contrasts(contr.helmert)*

```
> contrasts(diet)<-"contr.helmert"
```

A nouveau, les noms des colonnes ne sont plus associées aux niveaux du facteur donc pas d'interprétation en lien avec le facteur.

```
> contrasts(diet)
  [,1] [,2] [,3]
A   -1   -1   -1
B    1   -1   -1
C    0    2   -1
D    0    0    3
> mod.helmert<-lm(coag~diet)
> summary(mod.helmert)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.0000     0.4979 128.537  < 2e-16 ***
diet1         2.5000     0.7638   3.273 0.003803 **
diet2         1.5000     0.4105   3.654 0.001577 **
diet3        -1.0000     0.2578  -3.880 0.000932 ***
```

Contraintes et contrastes sous R

• 3 colonnes donc 3 paramètres sont estimés dans le modèle considéré ;

• ligne 4 = - somme des 3 autres ; et pour obtenir les moyennes des différentes modalités de *diet*, il faudra appliquer la combinaison linéaire spécifiée par la ligne aux paramètres du modèle.

Mais ici une interprétation directe des estimations obtenues à partir de la matrice de design est plus complexe car les coefficients obtenus comparent chaque niveau avec la moyenne des précédents !
↪ L'intercept du modèle est à nouveau la moyenne générale

```
> mean(coag)
[1] 64
```

↪ *Estimation correspondant à diet1 du tableau Estimates ?*
moyenne des deux premières modalités - moyenne 1 ère modalité (i.e. "A")

```
> 0.5*mean(coagulation$coag[coagulation$diet == 'A'])
+ 0.5*mean(coagulation$coag[coagulation$diet == 'B'])
-mean(coagulation$coag[coagulation$diet == 'A'])
[1] 2.5
```

donc estimation de $\frac{(\mu_A+\mu_B)}{2} - \mu_A$

Contraintes et contrastes sous R

↪ *Colonne 2 : Estimation correspondant à diet2 du tableau Estimates ?*
moyenne des trois 1ères modalités - moyenne des deux 1ères

```
> (mean(coagulation$coag[coagulation$diet == 'A'])+
     mean(coagulation$coag[coagulation$diet == 'B'])+
     mean(coagulation$coag[coagulation$diet == 'C']))/3
     - (mean(coagulation$coag[coagulation$diet == 'A'])+
mean(coagulation$coag[coagulation$diet == 'B']))/2
[1] 1.5
```

donc estimation de $\frac{\mu_A+\mu_B+\mu_C}{3} - \frac{\mu_A+\mu_B}{2}$.

↪ *Colonne 3 : Estimation correspondant à diet3 du tableau Estimates ?*
moyenne des quatre 1ères modalités - moyenne des trois 1ères

Contraintes et contrastes sous R

estimation diet3 . . .

```
> (mean(coagulation$coag[coagulation$diet == 'A'])
+mean(coagulation$coag[coagulation$diet == 'B'])
+mean(coagulation$coag[coagulation$diet == 'C'])
+mean(coagulation$coag[coagulation$diet == 'D']))/4
[1] 64
> (mean(coagulation$coag[coagulation$diet == 'A'])
+mean(coagulation$coag[coagulation$diet == 'B'])
+mean(coagulation$coag[coagulation$diet == 'C']))/3
[1] 65
```

et on obtient bien $64 - 65 = -1$ . . .

donc estimation de $\frac{\mu_A+\mu_B+\mu_C+\mu_D}{4} - \frac{\mu_A+\mu_B+\mu_C}{3}$.

Changer les contraintes : plusieurs possibilités . . .

```
# annuler une eventuelle def de contrastes deja utilisee avant
 > contrasts(coagulation$diet)=NULL

# definir des contrastes
 > options(contrasts=c('contr.treatment','cont.poly'))

# changer la modalité de reference
  # uniquement pour le modele  d'anova considéré :
  # on utilise ~C(.,base = .) dans  la fonction lm
  > mod2ter=lm(coag~C(diet,base=2),coagulation) # ref B au lieu de A

  # ou bien rlevel dans le facteur
  > coagulation$diet<-relevel(coagulation$diet,"B")
  > mod2bis=lm(coag~diet,coagulation) # ref B au lieu de A

# choix de la contrainte \sum \alpha_j=0 ;
  # on utilise ~C(.,sum) dans la fonction lm
  > mod3=lm(coag~C(diet,sum),coagulation)
```