Fiche TD avec le logiciel R : tdr334

Tailles des testicules et systèmes d'appariement

D. Pontier, A.B. Dufour & D. Chessel

La fiche permet de se familiariser avec l'analyse de covariance et la méthode des contrastes à partir d'un problème biologique. Les données présentées sont issues de l'article : Say L. & Pontier D. (2006) " What determines testis size in the domestic cat (*Felis catus* L.)? ". Une approche expérimentale de la manipulation des contrastes sur un facteur est proposée.

Table des matières

1 La problématique

L'objectif de cette étude est d'analyser une des prédictions de la théorie de la sélection sexuelle concernant les variations de taille des testicules. La taille des testicules est positivement liée à la production de sperme et, comme résultat, elle a évolué en réponse au risque de compétition spermatique, qui varie avec le système d'appariement. La compétition spermatique représente la compétition entre spermes de différents mâles pour fertiliser le(s) ovule(s) de la même femelle. Un mécanisme de la compétition spermatique est d'augmenter le nombre de spermatozoïdes de façon à " surpasser " ceux des mâles concurrents. Les plus gros testicules fabriquant plus de spermatozoïdes, on attend des testicules de taille plus importante dans les situations où la compétition spermatique est la plus forte.

Hypothèse de travail : la compétition spermatique et la taille des testicules sont liées.

Cette relation chez les vertébrés a été évaluée à l'échelle interspécifique et les résultats sont conformes à la théorie, à quelques exceptions près. On a montré aussi que le poids et les effets phylogénétiques ont une influence considérable sur la variabilité de la taille des testicules. Le chat domestique *Felis catus* L. constitue un excellent modèle pour réaliser à l'échelle intraspécifique un test original des prédictions de la théorie de la compétition spermatique concernant la relation entre système d'appariement et taille de testicules. La plasticité phénotypique intraspécifique devrait conduire au même patron, les plus gros testicules devant être observés dans les situations où la compétition spermatique est la plus forte. C'est ce que nous allons analyser.

2 Les populations

Cinq populations de chats domestiques sont étudiées. Elles se distinguent par la taille du groupe social et le système d'appariement.

Deux populations sont **rurales** : Saint-Just Chaleyssin (SJT) et Barisey-la-Côte (BAC). Les femelles vivent solitaires ou en petit groupe de 2 ou 3 femelles apparentées. La densité est faible (200 chats/km^2). Les mâles sont très agressifs et se battent pour monopoliser les femelles en oestrus. Le système d'appariement est la polygynie : un mâle monopolise avec succès une ou plusieurs femelles ; l'ensemble des chatons d'une même portée ont un seul père. *La compétition spermatique est attendue faible.*

Deux populations sont **urbaines** : Lyon Croix-Rousse (CxR) et Rome (ROM). Mâles et femelles vivent en grands groupes sociaux à très forte densité (1000 chats/km^2). Au moment de la reproduction, plusieurs mâles s'accouplent avec la même femelle durant la période de reproduction, sans aucune agressivité entre les mâles. Le système d'appariement est la promiscuité : une portée est, dans 80% des cas, engendrée par plusieurs mâles. *La compétition spermatique est attendue forte.*

Une population est **insulaire** : Kerguelen (KER). Les mâles et femelles sont solitaires occupant de très grands domaines vitaux. La densité est très faible (1 chat/km^2). Au moment de la reproduction, des couples peuvent se former.

Le système d'appariement est la monogamie. *La compétition spermatique est attendue faible.*

Quelle(s) prédiction(s) pouvez-vous faire ?

3 Préparation des données

Le fichier chasay.txt contient le poids (en grammes) et le volume des testicules (en cm^3) dans les cinq populations de chats : Barisey-la-côte (BAC), la Croix-Rousse (CxR), Saint-Just Chaleyssin (SJT), Kerguelen (KER) et Rome (ROM).

```
chasay <- read.table("http://pbil.univ-lyon1.fr/R/donnees/chasay.txt",
    h = TRUE)
```

Lire les données.

```
names(chasay)
```

```
[1] "poi" "pop" "vol"
```

```
chasay$pop
```

```
  [1] CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR
 [21] CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR CxR SJT SJT SJT
 [41] SJT SJT SJT SJT SJT SJT SJT SJT SJT SJT SJT SJT SJT SJT SJT BAC BAC BAC BAC BAC
 [61] BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC BAC
 [81] BAC BAC KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER
[101] KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER
[121] KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER KER
[141] KER KER KER ROM ROM ROM ROM ROM ROM ROM
Levels: BAC CxR KER ROM SJT
```

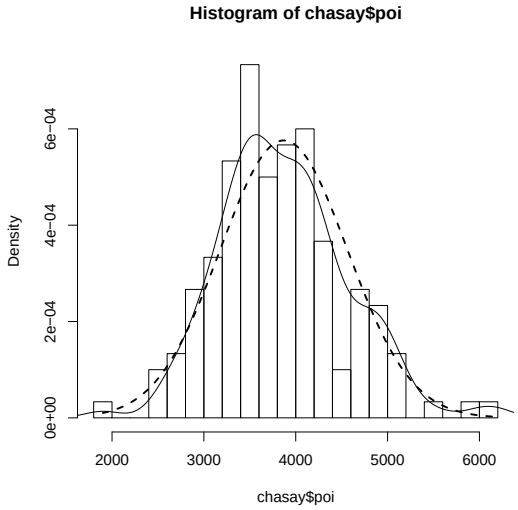
Vérifier que la lecture est correcte en retrouvant le résumé :

```
summary(chasay)
```

```
      poi         pop          vol        
 Min.   :1900   BAC:26   Min.   : 303.6  
 1st Qu.:3400   CxR:37   1st Qu.: 941.4  
 Median :3800   KER:61   Median :1232.2  
 Mean   :3870   ROM: 7   Mean   :1322.0  
 3rd Qu.:4200   SJT:19   3rd Qu.:1588.6  
 Max.   :6200            Max.   :3259.9  
```

Faire l'histogramme du poids et du volume avec 20 classes, lisser par une estimation de la densité et ajuster une loi normale. Commenter le résultat.

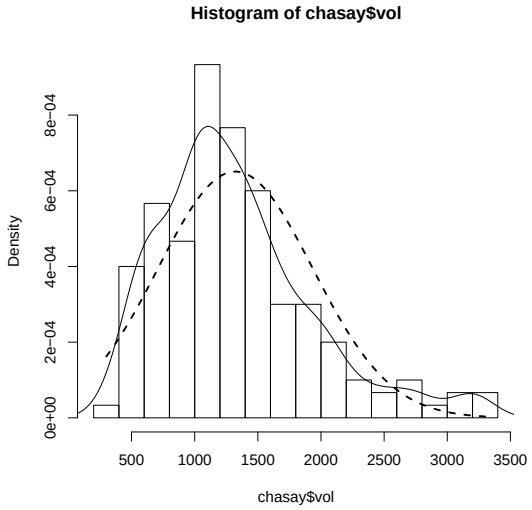
```
hist(chasay$poi, proba = T, nclass = 20)
x0 <- seq(1900, 6200, length = 100)
lines(x0, dnorm(x0, mean(chasay$poi), sd(chasay$poi)), lwd = 2,
    lty = 2)
lines(density(chasay$poi))
```



On pourrait presque accepter la normalité globale (`shapiro.test(chasay$poi)`). La distribution en tout cas est très favorable à l'analyse statistique.

```
hist(chasay$vol, proba = T, nclass = 20)
x1 <- seq(300, 3300, length = 100)
lines(x1, dnorm(x1, mean(chasay$vol), sd(chasay$vol)), lwd = 2,
    lty = 2)
lines(density(chasay$vol))
```

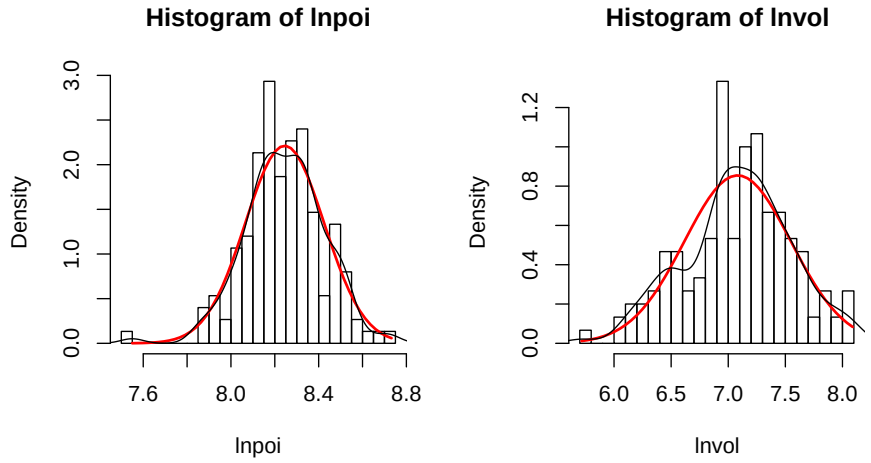
L'ajustement est moins bon pour la seconde variable, sans être vraiment mauvais. On peut faire les mêmes observations sur le logarithme des variables.



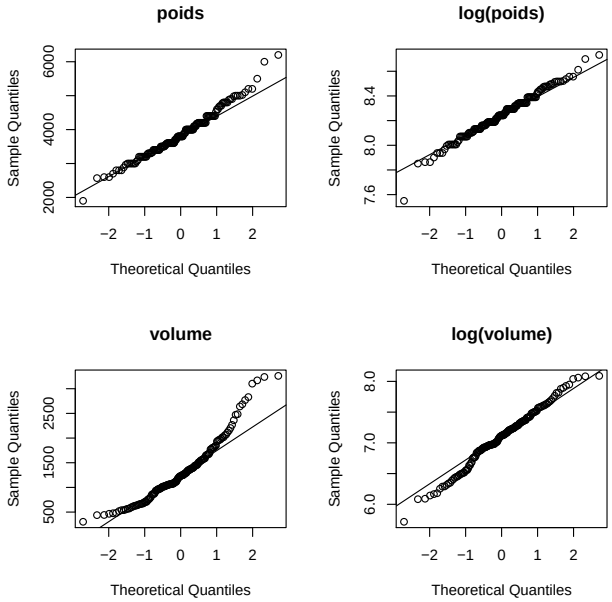

```
par(mfrow = c(1, 2))
lnpoi <- log(chasay$poi)
lnvol <- log(chasay$vol)
hist(lnpoi, proba = T, nclass = 20)
x1 <- seq(min(lnpoi), max(lnpoi), le = 50)
```

```
lines(x1, dnorm(x1, mean(lnpoi), sd(lnpoi)), lwd = 2, col = "red")
lines(density(lnpoi))
hist(lnvol, proba = T, nclass = 20)
x1 <- seq(min(lnvol), max(lnvol), le = 50)
lines(x1, dnorm(x1, mean(lnvol), sd(lnvol)), lwd = 2, col = "red")
lines(density(lnvol))
```

A gauche, le poids en logarithme, à droite le volume des testicules. La dissymétrie a disparu mais on met en évidence la présence d'un groupe à part.



Etudier la même question avec des plots quantiles-quantiles :

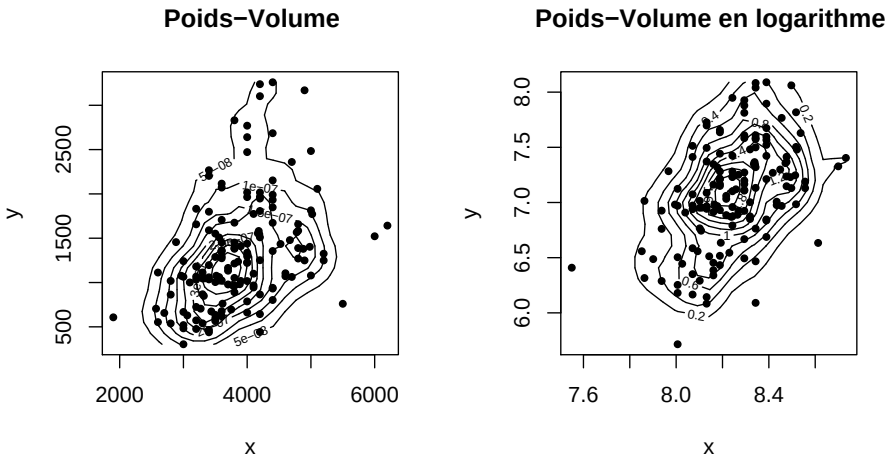



```
par(mfrow = c(2, 2))
qqnorm(chasay$poi, main = "poids")
```

```
qqline(chasay$poi)
qqnorm(lnpoi, main = "log(poids)")
qqline(lnpoi)
qqnorm(chasay$vol, main = "volume")
qqline(chasay$vol)
qqnorm(lnvol, main = "log(volume)")
qqline(lnvol)
```

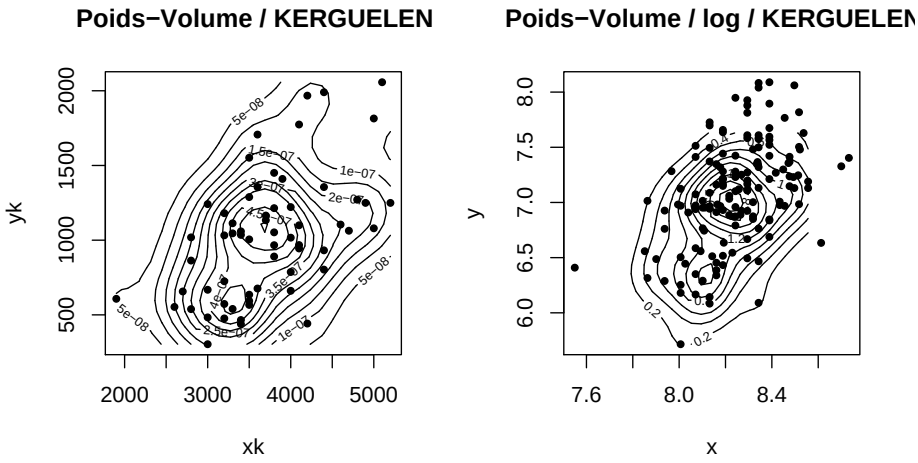
On voit mieux la dissymétrie de la seconde variable et la disparition de la dissymétrie dans le changement de variable. Remarquer qu'on ne doit pas espérer ici une normalité de l'ensemble des données. Il convient de ne pas confondre la normalité d'une variable et la normalité des résidus d'un modèle linéaire. Achever l'analyse préliminaire avec une estimation de la densité de probabilité des deux variables.

```
library(MASS)
par(mfrow = c(1, 2))
x <- chasay$poi
y <- chasay$vol
plot(x, y, pch = 20, main = "Poids-Volume")
f1 <- kde2d(x, y)
contour(f1, add = T)
x <- lnpoi
y <- lnvol
plot(x, y, pch = 20, main = "Poids-Volume en logarithme")
f1 <- kde2d(x, y)
contour(f1, add = T)
```



On peut conclure. Poids et volume des testicules ont des distributions très acceptables et la situation de l'analyse est bonne. Pour les plus curieux :

```
par(mfrow = c(1, 2))
xk <- chasay$poi[chasay$pop == "KER"]
yk <- chasay$vol[chasay$pop == "KER"]
plot(xk, yk, pch = 20, main = "Poids-Volume / KERGUELEN")
f1 <- kde2d(xk, yk)
contour(f1, add = T)
xk <- lnpoi[chasay$pop == "KER"]
yk <- lnvol[chasay$pop == "KER"]
plot(x, y, pch = 20, main = "Poids-Volume / log / KERGUELEN")
f1 = kde2d(xk, yk)
contour(f1, add = T)
```



Il n'y a pas de valeurs aberrantes, mais sans doute une hétérogénéité cachée dans le groupe de Kerguelen. Par tradition, on travaillera sur les logarithmes. En effet à la relation linéaire $y = ax + b$, se substitue sur les variables morphométriques une relation dite allométrique du type :

$$y = bx^a \Leftrightarrow \log(y) = a\log(x) + b$$

Ici la nécessité n'est pas clairement établie car volume et poids, à une constante près, sont de même dimension. On peut noter que la corrélation est un peu meilleure en log :

```
cor(chasay$poi, chasay$vol)
[1] 0.4036482
cor(lnpoi, lnvol)
[1] 0.4668886
cor(chasay$poi^(1/3), chasay$vol^(1/3))
[1] 0.4516185
```

Jusqu'à la fin, nous utiliserons les variables :

```
pop <- chasay$pop
lnvol <- log(chasay$vol)
lnpoi <- log(chasay$poi)
X <- cbind.data.frame(pop, lnpoi, lnvol)
summary(X)
 pop         lnpoi           lnvol
BAC:26   Min.   :7.550   Min.   :5.716
CxR:37   1st Qu.:8.132   1st Qu.:6.847
KER:61   Median :8.243   Median :7.117
ROM: 7   Mean   :8.245   Mean   :7.082
SJT:19   3rd Qu.:8.343   3rd Qu.:7.371
         Max.   :8.732   Max.   :8.089
```

4 L'effet global de deux facteurs

4.1 La taille des testicules en fonction du poids

On s'attend d'abord à ce que le volume des testicules soit une fonction du poids utilisé ici comme indicateur de la taille globale de l'individu. Il est clair qu'on ne compte pas ici faire une découverte. La corrélation est très positive :

```
 options(show.signif.stars = TRUE)
 cor.test(x, y)
        Pearson's product-moment correlation
data:  x and y
t = 6.423, df = 148, p-value = 1.718e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3314258 0.5834909
sample estimates:
      cor
0.4668886
 anova(lm(lnvol ~ lnpoi))
Analysis of Variance Table
Response: lnvol
           Df  Sum Sq Mean Sq F value    Pr(>F)
lnpoi       1  7.0794  7.0794  41.255 1.718e-09 ***
Residuals 148 25.3973  0.1716
---
Signif. codes:  0
```

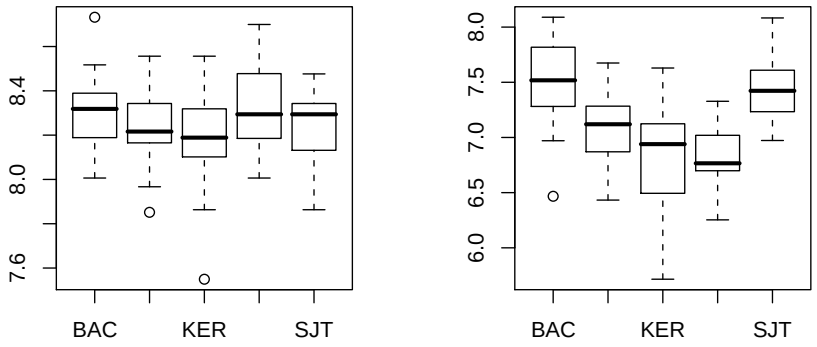
Si vous pensez que la statistique n'est pas là pour la pluie d'étoiles :

```
 options(show.signif.stars = FALSE)
 anova(lm(lnvol ~ lnpoi))
Analysis of Variance Table
Response: lnvol
           Df  Sum Sq Mean Sq F value    Pr(>F)
lnpoi       1  7.0794  7.0794  41.255 1.718e-09
Residuals 148 25.3973  0.1716
```

4.2 La taille des testicules en fonction des populations

La question plus intéressante est celle de la différence des moyennes entre les populations. Elle doit être précisée tant au niveau de la variable explicative (`lnpoi`) que de la variable à prédire (`lnvol`). La situation est très claire :

```
 par(mfrow = c(1, 2))
 tapply(lnpoi, pop, mean)
     BAC      CxR      KER      ROM      SJT
8.319972 8.243054 8.204354 8.332241 8.245662
 tapply(lnvol, pop, mean)
     BAC      CxR      KER      ROM      SJT
7.516186 7.042025 6.840336 6.826144 7.438159
 anova(lm(lnpoi ~ pop))
Analysis of Variance Table
Response: lnpoi
           Df Sum Sq Mean Sq F value  Pr(>F)
pop         4 0.3003  0.0751  2.3952 0.05312
Residuals 145 4.5454  0.0313
 anova(lm(lnvol ~ pop))
Analysis of Variance Table
Response: lnvol
           Df  Sum Sq Mean Sq F value    Pr(>F)
pop         4 11.3914  2.8479  19.584 6.632e-13
Residuals 145 21.0853  0.1454
 boxplot(lnpoi ~ pop)
 boxplot(lnvol ~ pop)
 bartlett.test(lnpoi, pop)
        Bartlett test of homogeneity of variances
data:  lnpoi and pop
Bartlett's K-squared = 3.9805, df = 4, p-value = 0.4087
 bartlett.test(lnvol, pop)
        Bartlett test of homogeneity of variances
data:  lnvol and pop
Bartlett's K-squared = 4.251, df = 4, p-value = 0.3731
```



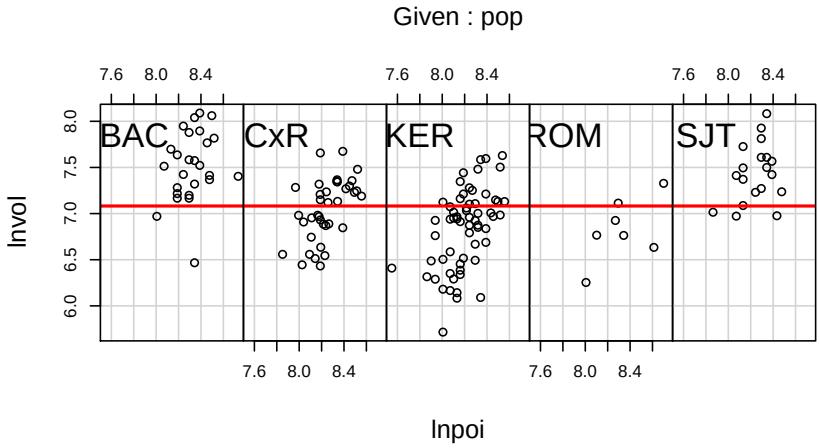
L'hypothèse nulle "les individus des différentes populations suivent une loi normale de moyenne constante (pas d'effet `pop`)" est rejetée sans risque d'erreur pour le volume (à droite) mais pas pour le poids (à gauche). Les individus des différentes populations diffèrent peu par le poids et beaucoup par le volume des testicules. On a donc deux effets très significatifs et une faible liaison entre les deux explicatives. Le volume augmente avec le poids et varie d'une population à l'autre.

4.3 Interaction ?

L'interaction, c'est la possibilité pour le poids d'avoir un effet sur le volume qui dépende de la population. La population agit sur la moyenne, agit-elle sur la pente de la droite de régression ? Posons simplement la question avant de répondre.

1. Modèle nul. Toutes les observations proviennent d'une loi unique.

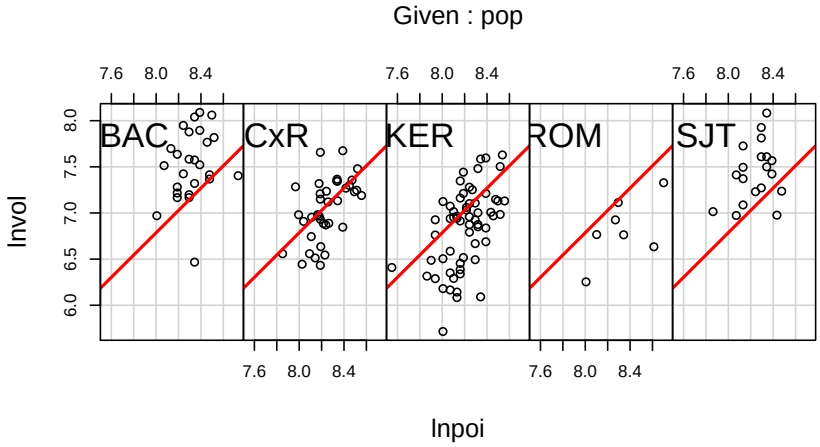
```
moytot <- mean(lnvol)
fonc <- function(x, y, subscripts, col = "black", ...) {
    points(x, y)
    abline(h = moytot, col = "red", lwd = 2)
    text(7.8, 7.85, unique(as.character(pop[subscripts])), cex = 2)
}
coplot(lnvol ~ lnpoi | pop, row = 1, subscripts = T, show = F, panel = fonc)
```



Ce modèle est évidemment insuffisant.

2. Modèle *lnpoi*. Seul le poids influence le volume.

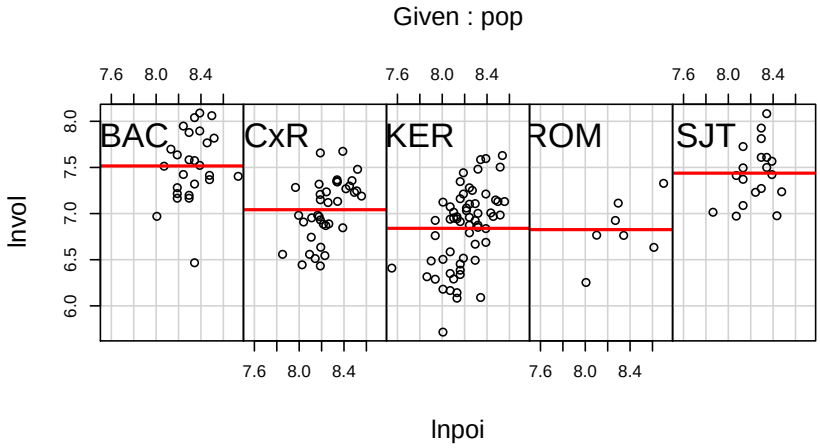
```
wt <- coefficients(lm(lnvol ~ lnpoi))
fonc <- function(x, y, subscripts, col = "black", ...) {
    points(x, y)
    abline(wt, col = "red", lwd = 2)
    text(7.8, 7.85, unique(as.character(pop[subscripts])), cex = 2)
}
coplot(lnvol ~ lnpoi | pop, row = 1, subscripts = T, show = F, panel = fonc)
```



Ce modèle est évidemment insuffisant.

3. Modèle pop. Seule la population fait varier le volume des testicules.

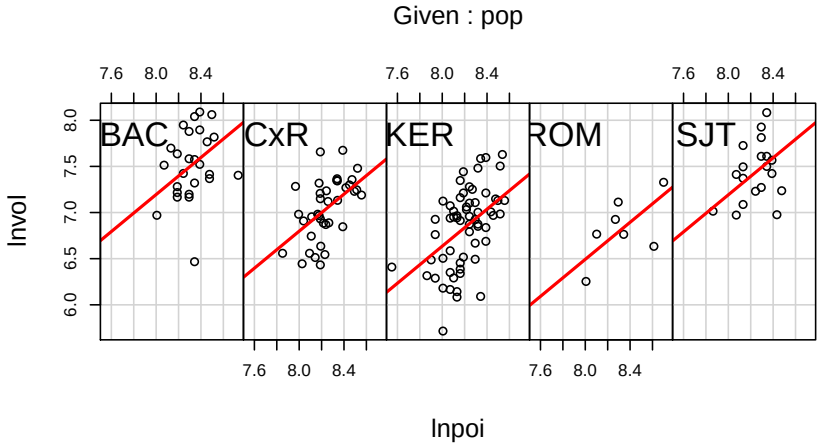
```
fonc <- function(x, y, subscripts, col = "black", ...) {
    points(x, y)
    abline(h = mean(y), col = "red", lwd = 2)
    text(7.8, 7.85, unique(as.character(pop[subscripts])), cex = 2)
}
coplot(lnvol ~ lnpoi | pop, row = 1, subscripts = T, show = F, panel = fonc)
```



Ce modèle est évidemment insuffisant.

4. Modèle additif.

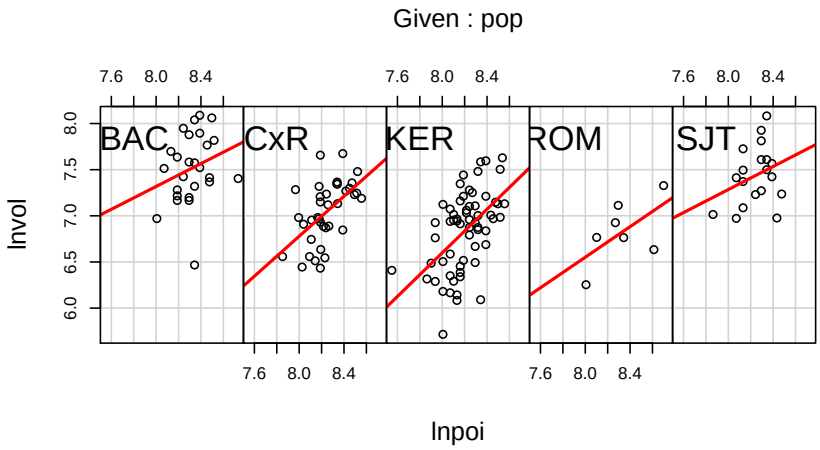
```
wp = coefficients(lm(lnvol ~ -1 + pop + lnpoi))
fonc <- function(x, y, subscripts, col = "black", ...) {
    points(x, y)
    abline(c(mean(y) - wp[6] * mean(x)), wp[6], col = "red", lwd = 2)
    text(7.8, 7.85, unique(as.character(pop[subscripts])), cex = 2)
}
coplot(lnvol ~ lnpoi | pop, row = 1, subscripts = T, show = F, panel = fonc)
```



On ne détecte plus d'insuffisance chronique. Seule se pose la possibilité d'une droite par population.

5. Modèle pop. On a une dépendance conditionnelle, dite interaction.

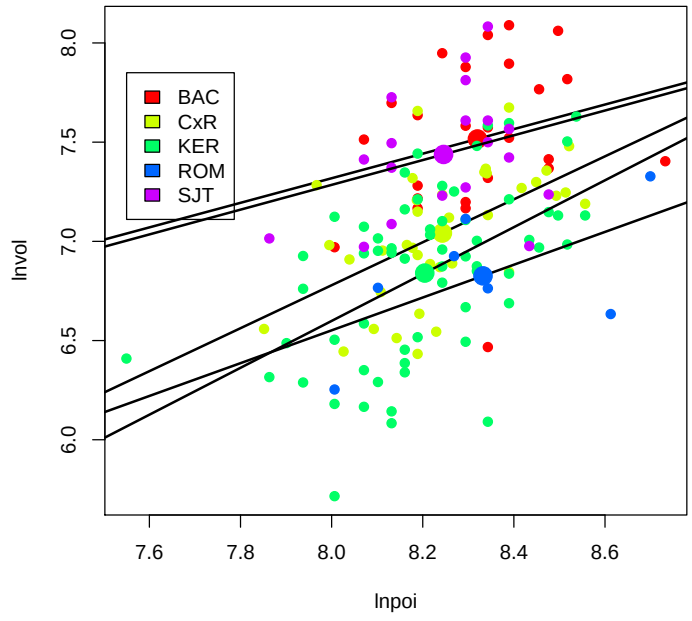
```
fonc <- function(x, y, subscripts, col = "black", ...) {
    points(x, y)
    abline(lm(y ~ x), lwd = 2, col = "red")
    text(7.8, 7.85, unique(as.character(pop[subscripts])), cex = 2)
}
coplot(lnvol ~ lnpoi | pop, row = 1, subscripts = T, show = F, panel = fonc)
```



La question est : ce modèle est-il légitime ?

Autre manière de représenter la même situation :

```
plot(lnpoi, lnvol, type = "n")
for (k in 1:5) {
    w <- levels(pop)[k]
    colo <- rainbow(5)[k]
    points(lnpoi[pop == w], lnvol[pop == w], col = colo, pch = 20,
        cex = 1.5)
    abline(lm(lnvol[pop == w] ~ lnpoi[pop == w]), lwd = 2)
    points(mean(lnpoi[pop == w]), mean(lnvol[pop == w]), pch = 20,
        cex = 3, col = colo)
}
legend(7.55, 7.85, col = rainbow(5), legend = levels(pop), fill = rainbow(5))
```



La question est de savoir si ce dessin est légitime ? Les droites n'ont évidemment pas la même ordonnée à l'origine. Ont elles les mêmes pentes ? La réponse est NON.

```
 anova(lm(lnvol ~ lnpoi * pop))
Analysis of Variance Table
Response: lnvol
           Df  Sum Sq Mean Sq F value    Pr(>F)
lnpoi       1  7.0794  7.0794 60.9040 1.260e-12
pop         4  8.8875  2.2219 19.1147 1.456e-12
lnpoi:pop   4  0.2363  0.0591  0.5081    0.7298
Residuals 140 16.2735  0.1162
```

Nous n'avons pas d'argument statistique pour rejeter l'hypothèse nulle *les pentes des droites sont égales*. Nous n'acceptons pas cette hypothèse mais actuellement la variation observée entre les pentes n'est pas incompatible avec un effet du hasard. Le modèle à deux effets additifs est pour l'instant suffisant. Nous allons donc étudier le modèle additif :

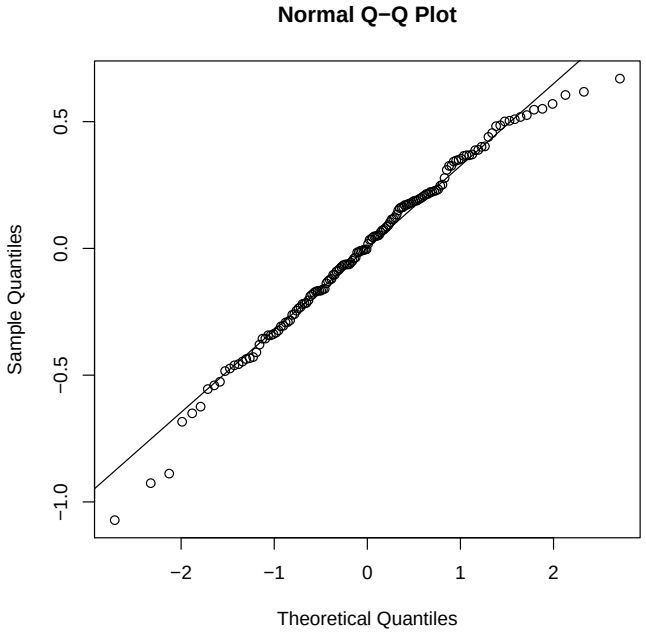
```
 lm1 <- lm(lnvol ~ lnpoi + pop)
```

C'est un bon modèle. Les deux facteurs sont très significatifs :

```
 anova(lm1)
Analysis of Variance Table
Response: lnvol
           Df  Sum Sq Mean Sq F value    Pr(>F)
lnpoi       1  7.0794  7.0794  61.748 8.292e-13
pop         4  8.8875  2.2219  19.380 8.931e-13
Residuals 144 16.5097  0.1147
```

Les résidus sont parfaitement normaux :

```
 qqnorm(lm1$residuals)
 qqline(lm1$residuals)
 shapiro.test(lm1$residuals)
        Shapiro-Wilk normality test
data:  lm1$residuals
W = 0.9854, p-value = 0.1142
```

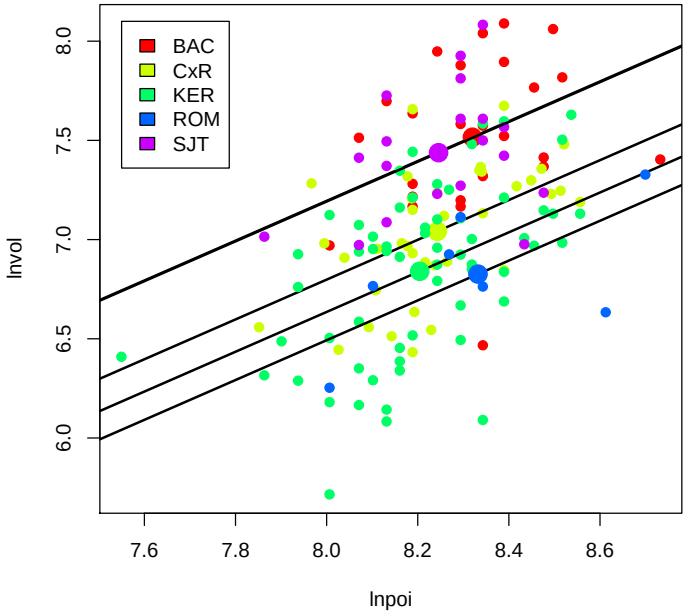


Les variances sont bien constantes par classe :

```
bartlett.test(lm1$residuals, pop)
        Bartlett test of homogeneity of variances
data:  lm1$residuals and pop
Bartlett's K-squared = 4.3626, df = 4, p-value = 0.3592
```

Il se résume au dessin :

```
plot(lnpoi, lnvol, type = "n")
w <- levels(pop)[1]
z <- coefficients(lm1)
colo <- rainbow(5)[1]
points(lnpoi[pop == w], lnvol[pop == w], col = colo, pch = 20, cex = 1.5)
abline(c(z[1], z[2]), lwd = 2)
points(mean(lnpoi[pop == w]), mean(lnvol[pop == w]), pch = 20, cex = 3,
    col = colo)
for (k in 2:5) {
    w <- levels(pop)[k]
    colo <- rainbow(5)[k]
    points(lnpoi[pop == w], lnvol[pop == w], col = colo, pch = 20,
        cex = 1.5)
    abline(c(z[1] + z[k + 1], z[2]), lwd = 2)
    points(mean(lnpoi[pop == w]), mean(lnvol[pop == w]), pch = 20,
        cex = 3, col = colo)
}
legend(7.55, 8.1, col = rainbow(5), legend = levels(pop), fill = rainbow(5))
```



Mais pour faire ce dessin, il faut bien comprendre où se trouvent les coefficients du modèle, ce qu'ils veulent dire, comment on s'en sert et comment on peut augmenter la **signification biologique** du modèle sans changer la **valeur statistique**. Pour comprendre le mode de fonctionnement de l'expression des paramètres d'un modèle, un détour par un exemple simple est indispensable.

5 Coefficients et contrastes d'un facteur

En 10 occasions réparties sur quatre stations, on relève le degré de pollution et la température.

```
station <- factor(rep(c("s1", "s2", "s3", "s4"), c(3, 2, 3, 2)))
station
 [1] s1 s1 s1 s2 s2 s3 s3 s3 s4 s4
Levels: s1 s2 s3 s4
```

L'objectif est de construire un modèle statistique permettant d'expliquer la pollution en fonction de la température et de la station soit, en clair :

$$pollution = a + b \times temperature + c_{station} + erreur$$

Ce modèle, en statistique, ne porte que sur les données observées. Notons $\mathbf{y}$ la variable pollution, $\mathbf{x}$ la variable température. Le modèle, pour les observations, s'écrit sous forme vectorielle :

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix} = \begin{bmatrix} a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \end{bmatrix} + b \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \end{bmatrix} + \begin{bmatrix} c_1 \\ c_1 \\ c_1 \\ c_2 \\ c_2 \\ c_3 \\ c_3 \\ c_3 \\ c_4 \\ c_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{bmatrix}$$

soit $\mathbf{y} = \mathbf{a} + b\mathbf{x} + \mathbf{c}_{station} + \varepsilon$. On peut alors observer que :

$$\begin{bmatrix} a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \end{bmatrix} = a \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = a\mathbf{1}_{10}$$

et

$$\begin{bmatrix} c_1 \\ c_1 \\ c_1 \\ c_2 \\ c_2 \\ c_3 \\ c_3 \\ c_3 \\ c_4 \\ c_4 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_1 \\ c_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ c_2 \\ c_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_3 \\ c_3 \\ c_3 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_4 \\ c_4 \end{bmatrix} = c_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + c_3 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + c_4 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

soit $\mathbf{c}_{station} = c_1\mathbf{I}_1 + c_2\mathbf{I}_2 + c_3\mathbf{I}_3 + c_4\mathbf{I}_4$, en utilisant les indicatrices des stations $\mathbf{I}_1$, $\mathbf{I}_2$, $\mathbf{I}_3$ et $\mathbf{I}_4$. Le modèle statistique exprimé devient alors :

$$\mathbf{y} = a\mathbf{1}_{10} + b\mathbf{x} + c_1\mathbf{I}_1 + c_2\mathbf{I}_2 + c_3\mathbf{I}_3 + c_4\mathbf{I}_4 + \varepsilon$$

On considère dans un premier temps, l'étude de la pollution $\mathbf{y}$ en fonction des stations $c_{station}$.

$$\mathbf{y} = a\mathbf{1}_{10} + c_1\mathbf{I}_1 + c_2\mathbf{I}_2 + c_3\mathbf{I}_3 + c_4\mathbf{I}_4 + \varepsilon$$

Simplifions encore et supposons tous les résidus (écarts au modèle) nuls.

$$\mathbf{y} = a\mathbf{1}_{10} + c_1\mathbf{I}_1 + c_2\mathbf{I}_2 + c_3\mathbf{I}_3 + c_4\mathbf{I}_4$$

```
pollution <- rep(c(12, 13, 15, 2), c(3, 2, 3, 2))
pollution
[1] 12 12 12 13 13 15 15 15  2  2
```

On ne peut faire plus simple. Et c'est déjà trop compliqué. On veut estimer les paramètres. Prenons par exemple la constante $a = 10$. Que vaut $\mathbf{c}_{station}$?

$$\begin{bmatrix} 12 \\ 12 \\ 12 \\ 13 \\ 13 \\ 15 \\ 15 \\ 15 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \\ 10 \end{bmatrix} + \begin{bmatrix} c_1 \\ c_1 \\ c_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ c_2 \\ c_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_3 \\ c_3 \\ c_3 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_4 \\ c_4 \end{bmatrix}$$

$$c_1 = 12 - 10 \Leftrightarrow c_1 = 2$$

$$c_2 = 13 - 10 \Leftrightarrow c_2 = 3$$

$$c_3 = 15 - 10 \Leftrightarrow c_3 = 5$$

$$c_4 = 2 - 10 \Leftrightarrow c_4 = -8.$$

$$\mathbf{c}_{station} = 2\mathbf{I}_1 + 3\mathbf{I}_2 + 5\mathbf{I}_3 - 8\mathbf{I}_4$$

Prenons une autre valeur de constante $a = 20$. On obtient d'autres valeurs pour les c_i.

$$c_1 = 12 - 20 \Leftrightarrow c_1 = -8$$

$$c_2 = 13 - 20 \Leftrightarrow c_2 = -7$$

$$c_3 = 15 - 20 \Leftrightarrow c_3 = -5$$

$$c_4 = 2 - 20 \Leftrightarrow c_4 = -18.$$

$$\mathbf{c}_{station} = -8\mathbf{I}_1 - 7\mathbf{I}_2 - 5\mathbf{I}_3 - 18\mathbf{I}_4$$

Le modèle (qui est ici l'observation, l'erreur étant nulle) existe, il est unique. Il s'écrit de plusieurs manières différentes :

$$\mathbf{y} = 10\times\mathbf{1}_{10}+2\times\mathbf{I}_1+3\times\mathbf{I}_2+5\times\mathbf{I}_3-8\times\mathbf{I}_4 = 20\times\mathbf{1}_{10}-8\times\mathbf{I}_1-7\times\mathbf{I}_2-5\times\mathbf{I}_3-18\times\mathbf{I}_4$$

Il s'écrit d'une infinité de façons différentes, tout en restant lui-même. Cela nuit beaucoup à son usage ! Le modèle est surparamétré. Cela est dû au fait que la somme des indicatrices des stations est égale à $\mathbf{1}_{10}$. On ne peut parler de modèles et de coefficients qu'avec des explicatives (ici $(\mathbf{1}_{10}, \mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3, \mathbf{I}_4)$ indépendantes. Que fait le programme pour résoudre cette question ?

```
 coefficients(lm(pollution ~ station))
(Intercept)    stations2    stations3    stations4
         12            1            3          -10
```

Il a décidé de l'écrire :

$$\mathbf{y} = 12 \times \mathbf{1}_{10} + 1 \times \mathbf{I}_2 + 3 \times \mathbf{I}_3 - 10 \times \mathbf{I}_4$$

Le calcul est juste, la première indicatrice a simplement disparu. On peut faire autrement :

```
 coefficients(lm(pollution ~ -1 + station))
stations1 stations2 stations3 stations4
       12        13        15         2
```

Il a décidé de l'écrire :

$$\mathbf{y} = 12 \times \mathbf{I}_1 + 13 \times \mathbf{I}_2 + 15 \times \mathbf{I}_3 + 2 \times \mathbf{I}_4$$

Vérifier qu'il s'agit toujours du même modèle. Le programme utilise effectivement et explicitement les indicatrices :

```
 model.matrix(lm(pollution ~ station))
   (Intercept) stations2 stations3 stations4
1            1         0         0         0
2            1         0         0         0
3            1         0         0         0
4            1         1         0         0
5            1         1         0         0
6            1         0         1         0
7            1         0         1         0
8            1         0         1         0
9            1         0         0         1
10           1         0         0         1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$station
[1] "contr.treatment"
 model.matrix(lm(pollution ~ -1 + station))
   stations1 stations2 stations3 stations4
1          1         0         0         0
2          1         0         0         0
3          1         0         0         0
4          0         1         0         0
5          0         1         0         0
6          0         0         1         0
7          0         0         1         0
8          0         0         1         0
9          0         0         0         1
10         0         0         0         1
attr(,"assign")
[1] 1 1 1 1
attr(,"contrasts")
attr(,"contrasts")$station
[1] "contr.treatment"
```

Par défaut, si on ne précise rien, la station 1 est considérée comme une station témoin. Nous avons dans ce cas :

$$\begin{bmatrix} 12 \\ 12 \\ 12 \\ 13 \\ 13 \\ 15 \\ 15 \\ 15 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ c_2 \\ c_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_3 \\ c_3 \\ c_3 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_4 \\ c_4 \end{bmatrix}$$

C'est un système de quatre équations et quatre inconnues.

$$a = 12$$

$$13 = a + c_2 \Leftrightarrow c_2 = 1$$

$$15 = a + c_3 \Leftrightarrow c_3 = 3$$

$$2 = a + c_4 \Leftrightarrow c_4 = -10.$$

Ce modèle implique que la constante est la valeur des témoins et que chacune des autres modalités correspond à en effet particulier. On peut changer de groupe témoin. Si on veut que ce soit le groupe 3 qui serve de témoin, on dit :

```
tem3 = station
contrasts(tem3) = contr.treatment(4, base = 3)
model.matrix(lm(pollution ~ tem3))
```

```
   (Intercept) tem31 tem32 tem34
1            1     1     0     0
2            1     1     0     0
3            1     1     0     0
4            1     0     1     0
5            1     0     1     0
6            1     0     0     0
7            1     0     0     0
8            1     0     0     0
9            1     0     0     1
10           1     0     0     1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$tem3
   1 2 4
s1 1 0 0
s2 0 1 0
s3 0 0 0
s4 0 0 1
```

```
coefficients(lm(pollution ~ tem3))
```

```
(Intercept)       tem31       tem32       tem34
         15          -3          -2         -13
```

$$\begin{bmatrix} 12 \\ 12 \\ 12 \\ 13 \\ 13 \\ 15 \\ 15 \\ 15 \\ 2 \\ 2 \end{bmatrix} = 15 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - 3 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - 2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - 13 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Il existe ainsi une infinité de manières pour décrire un même modèle à partir d'un facteur. Le facteur définit un sous-espace et la description est basée sur une base de ce sous-espace. Toutes les bases d'un sous-espace ont le même nombre de vecteurs et avec la facteur station on aura toujours 4 vecteurs, donc 3 si on laisse devant une constante (`intercept`). Cette description considère, par défaut, une constante et une première classe de témoins. Mais toute autre manière est disponible par la fonction `contrasts`. La fonction donne la valeur des contrastes d'un facteur et permet de les modifier.

```
contrasts(station)
```

```
   s2 s3 s4
s1  0  0  0
s2  1  0  0
s3  0  1  0
s4  0  0  1
```

```
contrasts(tem3)
```

```
   1 2 4
s1 1 0 0
s2 0 1 0
s3 0 0 0
s4 0 0 1
```

Cette écriture signifie que la matrice des indicatrices des classes multipliée à droite par la matrice des contrastes définit la base qui sera utilisée.

```
A = matrix(rep(c(1, 0, 1, 0, 1, 0, 1), c(3, 10, 2, 10, 3, 10, 2)),
    10)
A
```

```
      [,1] [,2] [,3] [,4]
 [1,]    1    0    0    0
 [2,]    1    0    0    0
 [3,]    1    0    0    0
 [4,]    0    1    0    0
 [5,]    0    1    0    0
 [6,]    0    0    1    0
 [7,]    0    0    1    0
 [8,]    0    0    1    0
 [9,]    0    0    0    1
[10,]    0    0    0    1
```

Cette matrice est celle des indicatrices des classes, avec 10 lignes et 4 colonnes (4 classes). Si on la multiplie à droite par une matrice de contrastes (4 lignes et 3 colonnes) on aura des matrices à 10 lignes et 3 colonnes, 3 vecteurs qui avec $\mathbf{1}_{10}$ formeront une base.

```
A %*% contrasts(station)
```

```
      s2 s3 s4
 [1,]  0  0  0
 [2,]  0  0  0
 [3,]  0  0  0
 [4,]  1  0  0
 [5,]  1  0  0
```

```
 [6,]  0  1  0
 [7,]  0  1  0
 [8,]  0  1  0
 [9,]  0  0  1
[10,]  0  0  1
```

```
model.matrix(~station)
```

```
   (Intercept) stations2 stations3 stations4
1            1         0         0         0
2            1         0         0         0
3            1         0         0         0
4            1         1         0         0
5            1         1         0         0
6            1         0         1         0
7            1         0         1         0
8            1         0         1         0
9            1         0         0         1
10           1         0         0         1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$station
[1] "contr.treatment"
```

Dans le dernier attribut est conservé, pour les analyses de variance, la valeur 0 pour `intercept` et 1 pour les autres qui définiront l'effet station globalement.

```
A %*% contrasts(tem3)
```

```
     1 2 4
 [1,] 1 0 0
 [2,] 1 0 0
 [3,] 1 0 0
 [4,] 0 1 0
 [5,] 0 1 0
 [6,] 0 0 0
 [7,] 0 0 0
 [8,] 0 0 0
 [9,] 0 0 1
[10,] 0 0 1
```

```
model.matrix(~tem3)
```

```
   (Intercept) tem31 tem32 tem34
1            1     1     0     0
2            1     1     0     0
3            1     1     0     0
4            1     0     1     0
5            1     0     1     0
6            1     0     0     0
7            1     0     0     0
8            1     0     0     0
9            1     0     0     1
10           1     0     0     1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$tem3
   1 2 4
s1 1 0 0
s2 0 1 0
s3 0 0 0
s4 0 0 1
```

Quand il y a plusieurs facteurs, le vecteur `intercept` appartient à tous les sous-espaces associés aux facteurs. Il faut le mettre directement de côté et ne conserver que $m - 1$ contrastes pour un facteur à m modalités. On peut enfin changer radicalement les contrastes pour intégrer des hypothèses sur le rôle des modalités. Aucune des stations n'est témoin et l'approche traditionnelle est mal venue. Supposons au contraire que deux stations (1 et 2) se situent au nord tandis que 3 et 4 sont au sud de la région. Une partie de l'information station mesure l'effet nord-sud. Si on concentre le premier contraste sur cet effet, les autres effets consistent à comparer les deux stations du nord entre elles et les

deux stations du sud entre elles. On construit le nouveau modèle.

$$\begin{bmatrix} 12 \\ 12 \\ 12 \\ 13 \\ 13 \\ 15 \\ 15 \\ 15 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \end{bmatrix} + \begin{bmatrix} c_{NS} \\ c_{NS} \\ c_{NS} \\ c_{NS} \\ c_{NS} \\ -c_{NS} \\ -c_{NS} \\ -c_{NS} \\ -c_{NS} \\ -c_{NS} \end{bmatrix} + \begin{bmatrix} c_N \\ c_N \\ c_N \\ -c_N \\ -c_N \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_S \\ c_S \\ c_S \\ -c_S \\ -c_S \end{bmatrix}$$

$$\begin{bmatrix} 12 \\ 12 \\ 12 \\ 13 \\ 13 \\ 15 \\ 15 \\ 15 \\ 2 \\ 2 \end{bmatrix} = a\mathbf{1}_{10} + c_{NS} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix} + c_N \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + c_S \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}$$

Nous avons un système de quatre équations et quatre inconnues.

$$12 = a + c_{NS} + c_N$$

$$13 = a + c_{NS} - c_N$$

$$15 = a - c_{NS} + c_S$$

$$2 = a - c_{NS} - c_S$$

La solution du système est : $a = \frac{21}{2}$, $c_{NS} = \frac{4}{2}$, $c_S = -\frac{1}{2}$ et $c_S = \frac{13}{2}$. Dans R, on écrit le nouveau contraste qui intègre la nouvelle information.

```
neocontr <- matrix(c(1, 1, -1, -1, 1, -1, 0, 0, 0, 0, 1, -1), nrow = 4)
dimnames(neocontr)[[2]] = c("NS", "intraN", "intraS")
dimnames(neocontr)[[1]] = c("s1", "s2", "s3", "s4")
```

On associe les nouveaux contrastes à une copie du facteur.

```
newfac <- station
contrasts(newfac) <- neocontr
model.matrix(~newfac)
   (Intercept) newfacNS newfacintraN newfacintraS
1            1        1            1            0
2            1        1            1            0
3            1        1            1            0
4            1        1           -1            0
5            1        1           -1            0
6            1       -1            0            1
7            1       -1            0            1
8            1       -1            0            1
9            1       -1            0           -1
10           1       -1            0           -1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$newfac
   NS intraN intraS
s1  1      1      0
s2  1     -1      0
s3 -1      0      1
s4 -1      0     -1
```

On peut alors s'en servir.

```
coefficients(lm(pollution ~ newfac))
```
```
(Intercept)     newfacNS newfacintraN newfacintraS
       10.5          2.0         -0.5          6.5
```

Ce sont bien les résultats attendus. Noter alors l'essentiel : *les modèles sans erreur, ça n'existe pas*. Mais lorsque un facteur a un effet significatif, donné par l'analyse de variance, on dispose d'un test de signification de chaque hypothèse nulle sur chaque coefficient. Le même modèle est écrit de trois façons différentes, par exemple :

$$\mathbf{y} = a\mathbf{1}_{10} + c_2\mathbf{I}_2 + c_3\mathbf{I}_3 + c_4\mathbf{I}_4 + \varepsilon$$

$$\mathbf{y} = a'\mathbf{1}_{10} + c'_1\mathbf{I}_1 + c'_2\mathbf{I}_2 + c'_3\mathbf{I}_3 + \varepsilon$$

$$\mathbf{y} = a''\mathbf{1}_{10} + c_{NS}\mathbf{w_{NS}} + c_N\mathbf{w_N} + c_S\mathbf{w_S} + \varepsilon$$

Le modèle est indépendant de la décomposition, l'erreur aussi, le F de l'anova aussi. Mais suivant les cas, on pourra tester différentes hypothèses. La classe 1 est témoin, on teste l'effet 2, l'effet 3 et l'effet 4 par rapport au témoin :

$$H_2 \rightarrow c_2 = 0 \quad H_3 \rightarrow c_3 = 0 \quad H_4 \rightarrow c_4 = 0$$

La classe 3 est témoin, on teste l'effet 1, l'effet 2 et l'effet 4 par rapport au témoin :

$$K_1 \rightarrow c'_1 = 0 \quad K_2 \rightarrow c' = 0 \quad K_3 \rightarrow c'_3 = 0$$

Il y a un effet nord-sud, les stations nord sont différentes, les stations sud sont différentes :

$$L_{NS} \rightarrow c_{NS} = 0 \quad L_N \rightarrow c_N = 0 \quad L_S \rightarrow c_S = 0$$

Après la définition des contrastes, ceci se fait avec la fonction `summary`

6 Importance statistique des contrastes

Revenons en aux testicules des matous. Pour éviter des wagons de chiffres significatifs inutiles, faisons :

```
options(digits = 4)
lm1 <- lm(lnvol ~ lnpoi + pop)
coefficients(lm1)
```
```
(Intercept)       lnpoi      popCxR      popKER      popROM      popSJT
  -0.831269    1.003303   -0.396989   -0.559851   -0.702352   -0.003472
```

On a perdu perdu `BAC` (*Barisey-la-Côte*). On sait pourquoi. C'est le premier par ordre alphabétique (ce n'est pas une raison statistique !). Pour un chat de Barisey un lnpoi de 8 permet de prédire un lnvol de :

```
predict(lm1, new = list(lnpoi = 8, pop = as.factor("BAC")))
```
```
    1
7.195
```
```
8 * coefficients(lm1)[2] + coefficients(lm1)[1]
```
```
lnpoi
7.195
```

Le coefficient de correction pour `BAC` est nul. Pour un chat des Kerguelen de même poids, on prédit :

```
    1
6.635
lnpoi
6.635
```

Le coefficient de `popKER` est clairement un facteur de correction par rapport à BAC. Les chats de Barisey-la-Côte sont des chats témoins. A cause de l'ordre alphabétique ? Biologiquement c'est un non sens. Plus raisonnable serait alors de faire :

```
coefficients(lm2 <- lm(lnvol ~ -1 + lnpoi + pop))
 lnpoi  popBAC  popCxR  popKER  popROM  popSJT
1.0033 -0.8313 -1.2283 -1.3911 -1.5336 -0.8347
predict(lm2, new = list(lnpoi = 8, pop = as.factor("BAC")))
    1
7.195
8 * coefficients(lm2)[1] + coefficients(lm2)[2]
lnpoi
7.195
predict(lm2, new = list(lnpoi = 8, pop = as.factor("KER")))
    1
6.635
8 * coefficients(lm2)[1] + coefficients(lm2)[4]
lnpoi
6.635
```

Ceci ne change rien au modèle, mais bouleverse les définitions des coefficients. Comparez `anova(lm1)`, `anova(lm2)`, `summary(lm1)` et `summary(lm2)`. C'est la panique. Plus rien n'a de sens. On peut ne pas changer de modèle et changer bien des aides à l'interprétation.

Notons qu'il y a le test de l'analyse de variance et le test des coefficients. Ceci a rempli des générations de biologistes de perplexité. Deux tests, bonjour les dégâts. S'ils disent la même chose, ça va, sinon, il faut comprendre d'où ça sort et comment ça marche. L'anova qui a un sens, qu'on a déjà utilisé, est :

```
anova(lm1)
Analysis of Variance Table
Response: lnvol
           Df Sum Sq Mean Sq F value  Pr(>F)
lnpoi       1   7.08    7.08    61.8 8.3e-13
pop         4   8.89    2.22    19.4 8.9e-13
Residuals 144  16.51    0.11
```

On peut décider que les deux facteurs sont significatifs et qu'il est légitime d'interpréter le mode de leur action. Quand on parle de facteurs, on entend par là l'espace associé à la prise en compte mathématique des données, donc une variable pour les quantitatives, un ensemble de contrastes pour les qualitatives. Pour les premières, il n'y aura jamais de problème : le test sur l'effet du facteur et le test sur le coefficient sont de même nature.

```
summary(lm1)[[4]][2, ]
 Estimate Std. Error    t value   Pr(>|t|)
1.003e+00  1.588e-01  6.317e+00  3.122e-09
```

Editer l'objet entier. L'extrait ci-dessus signifie que la pente de la droite du modèle a été estimée dans `Estimate`, que si le facteur n'avait pas d'effet on aurait une vraie valeur de 0 avec une écart-type estimé dans `Std. Error`, ce qui donne une observation de la `t value` par

(observation-valeur)/écart-type)

et une probabilité critique `Pr(>|t|)`. La loi t vient du dénominateur qui est une estimation et non une vraie valeur.

```
 anova(lm1)[1, ]
Analysis of Variance Table
Response: lnvol
      Df Sum Sq Mean Sq F value  Pr(>F)
lnpoi  1   7.08    7.08    61.8 8.3e-13
```

Editer l'objet entier. L'extrait ci-dessus signifie que la somme des carrés des écarts au modèle `Sum Sq` pour une espace de dimension `Df` rapportée à la même quantité pour le modèle sans effet `F value` qui suit une loi F sous l'hypothèse nulle d'absence d'effet et donne une probabilité critique `Pr(>F)`. Il n'y a généralement aucune contradiction entre les deux.

Pour un facteur, c'est beaucoup plus compliqué. Pour l'ANOVA, rien n'est changé :

```
 anova(lm(lnvol ~ lnpoi + pop))
Analysis of Variance Table
Response: lnvol
           Df Sum Sq Mean Sq F value  Pr(>F)
lnpoi       1   7.08    7.08    61.8 8.3e-13
pop         4   8.89    2.22    19.4 8.9e-13
Residuals 144  16.51    0.11
 anova(lm(lnvol ~ pop + lnpoi))
Analysis of Variance Table
Response: lnvol
           Df Sum Sq Mean Sq F value  Pr(>F)
pop         4  11.39    2.85    24.8 1.2e-15
lnpoi       1   4.58    4.58    39.9 3.1e-09
Residuals 144  16.51    0.11
```

Ne pas oublier cependant que le test est emboîté, l'effet du second tenant compte de l'effet du premier. Ici, aucune contradiction entre les deux, la situation est décontractée. Oui, MAIS :

```
 summary(lm1)
Call:
lm(formula = lnvol ~ lnpoi + pop)
Residuals:
    Min      1Q  Median      3Q     Max
-1.0719 -0.2169  0.0073  0.2203  0.6698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.83127    1.32303   -0.63     0.53
lnpoi        1.00330    0.15882    6.32  3.1e-09
popCxR      -0.39699    0.08751   -4.54  1.2e-05
popKER      -0.55985    0.08140   -6.88  1.7e-10
popROM      -0.70235    0.14419   -4.87  2.9e-06
popSJT      -0.00347    0.10287   -0.03     0.97

Residual standard error: 0.339 on 144 degrees of freedom
Multiple R-squared: 0.492,      Adjusted R-squared: 0.474
F-statistic: 27.9 on 5 and 144 DF,  p-value: <2e-16
```

Le test dit que l'indicatrice `popCxR` a un coefficient de régression non nul, exactement comme pour une variable quantitative comme `lnpoi`, donc que le facteur de correction qu'on ajoute par rapport à `BAC` est non nul, donc que les chats de Barisey-la-Côte sont différents des chats de la Croix-Rousse. Le test suivant dit que les chats de Kerguelen sont différents des chats de Barisey-la-Côte, le suivant dit que les chats de Rome sont différents des chats de Barisey-la-Côte, et le dernier dit qu'on ne peut pas dire que les chats de Saint-Just Chaleyssin sont différents des chats de Barisey-la-Côte.

Mais je n'ai pas voulu faire ça, direz-vous ! Silence, on ne vous demande pas votre avis. Par défaut le premier niveau d'un facteur désigne la classe des

témoins. C'est une convention. Si vous ne voulez pas du défaut, il faut ouvrir la documentation, ce que nous faisons pour vous, pour vous être agréable. Les chats de Barisey-la-Côte sont bien gentils mais il n'y a aucune raison qu'ils servent de témoins dans cette expérience, tout simplement à cause de l'ordre alphabétique. Non, non et non. Bon, vous préférez dire que les témoins sont ceux des Kerguelen ? C'est facile. Pour le moment, on utilise les contrastes par défaut :

```
 contrasts(pop)
    CxR KER ROM SJT
BAC   0   0   0   0
CxR   1   0   0   0
KER   0   1   0   0
ROM   0   0   1   0
SJT   0   0   0   1
```

On en change :

```
 contrasts(pop) <- contr.treatment(5, 3)
 contrasts(pop)
    1 2 4 5
BAC 1 0 0 0
CxR 0 1 0 0
KER 0 0 0 0
ROM 0 0 1 0
SJT 0 0 0 1
 summary(lm1 <- lm(lnvol ~ lnpoi + pop))
Call:
lm(formula = lnvol ~ lnpoi + pop)
Residuals:
    Min      1Q  Median      3Q     Max
-1.0719 -0.2169  0.0073  0.2203  0.6698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.3911     1.3037   -1.07    0.288
lnpoi         1.0033     0.1588    6.32  3.1e-09
pop1          0.5599     0.0814    6.88  1.7e-10
pop2          0.1629     0.0708    2.30    0.023
pop4         -0.1425     0.1366   -1.04    0.299
pop5          0.5564     0.0892    6.24  4.7e-09

Residual standard error: 0.339 on 144 degrees of freedom
Multiple R-squared: 0.492,      Adjusted R-squared: 0.474
F-statistic: 27.9 on 5 and 144 DF,  p-value: <2e-16
 anova(lm1)
Analysis of Variance Table
Response: lnvol
           Df Sum Sq Mean Sq F value  Pr(>F)
lnpoi       1   7.08    7.08    61.8 8.3e-13
pop         4   8.89    2.22    19.4 8.9e-13
Residuals 144  16.51    0.11
```

L'effet du facteur vu par l'ANOVA n'a pas changé. Par contre les tests sur les coefficients disent que les chats de Kerguelen sont les témoins, que ceux de Barisey sont très différents, que ceux de la Croix-Rousse sont un peu différents, que ceux de Rome ne sont pas significativement différents mais que ceux de Saint-Just sont très différents de *ceux de Kerguelen*. Oui, mais, les chats de Kerguelen sont bien gentils mais il n'y a aucune raison qu'ils servent de témoins dans cette expérience, tout simplement pour faire un exercice. Ceci reste idiot, parce qu'il n'y a pas de témoins tout simplement.

7 Contrastes et hypothèses biologiques

Parmi les 5 populations deux sont de même type par rapport à la compétition spermatique : les populations " polygynes " de `BAC` et `SJT`. Deux autres sont

basées sur la promiscuité en milieu urbain. Une dernière est insulaire et illustre la monogamie. L'essentiel est de savoir si les deux premières sont différentes des secondes et positionner la dernière. Nous allons donc construire les contrastes qui correspondent au test de nos questions :

1. Les populations urbaines sont-elles différentes des populations rurales ? Un premier contraste permettra de comparer les deux population urbaines aux deux populations rurales ;
2. La population insulaire est-elle différente des autres ? Un deuxième contraste qui comparera la population insulaire aux populations non insulaires ;
3. Les populations urbaines sont-elles différentes entre elles ? Un troisième contraste permettra de comparer les deux populations urbaines.
4. Les populations rurales sont-elles différentes entre elles ? Un dernier contraste permettra de comparer les deux populations rurales.

```
w <- matrix(c(-1, 1, 0, 1, -1, 1, 1, -1, 1, 1, -1, 0, 0, 0, 1, 0,
    1, 0, -1, 0), ncol = 4)
dimnames(w) <- list(levels(pop), c("RU", "I", "WR", "WU"))
w
```

```
    RU  I WR WU
BAC -1  1 -1  0
CxR  1  1  0  1
KER  0 -1  0  0
ROM  1  1  0 -1
SJT -1  1  1  0
```

On a noté `RU` pour Rural *vs.* Urbain, `I` pour insulaire, `WR` pour intra-rural et `WU` pour intra-urbain. Le résultat est :

```
contrasts(pop) <- w
summary(lm1 <- lm(lnvol ~ lnpoi + pop))
```

```
Call:
lm(formula = lnvol ~ lnpoi + pop)
Residuals:
    Min      1Q  Median      3Q     Max
-1.0719 -0.2169  0.0073  0.2203  0.6698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.24905    1.30979   -0.95    0.342
lnpoi        1.00330    0.15882    6.32  3.1e-09
popRU       -0.27397    0.04325   -6.34  2.9e-09
popI         0.14207    0.03128    4.54  1.2e-05
popWR       -0.00174    0.05144   -0.03    0.973
popWU        0.15268    0.07014    2.18    0.031

Residual standard error: 0.339 on 144 degrees of freedom
Multiple R-squared: 0.492,        Adjusted R-squared: 0.474
F-statistic: 27.9 on 5 and 144 DF,  p-value: <2e-16
```

Biologiquement, on répondra donc oui aux deux premières questions sans hésitation. On discutera du reste avec prudence. A noter que la logique des contrastes se poursuit dans les études avec interaction. Il suffit de savoir que l'interaction est définie par les vecteurs produits, comme dans l'exemple qui précède avec les

contrastes ordinaires :

$$\begin{bmatrix} 12 \\ 12 \\ 12 \\ 13 \\ 13 \\ 15 \\ 15 \\ 15 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \\ a \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ c_2 \\ c_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_3 \\ c_3 \\ c_3 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ c_4 \\ c_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ b_2x_4 \\ b_2x_5 \\ b_3x_6 \\ b_3x_7 \\ b_3x_8 \\ b_4x_9 \\ b_4x_{10} \end{bmatrix}$$

Ceci peut se réécrire :

$$a \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + c_3 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + c_4 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} + b_2 \begin{bmatrix} 0 \\ 0 \\ 0 \\ x_4 \\ x_5 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + b_3 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ x_6 \\ x_7 \\ x_8 \\ 0 \\ 0 \end{bmatrix} + b_4 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ x_9 \\ x_{10} \end{bmatrix}$$

soit

$$a1_n + \sum_{j \geq 2} c_j \mathbf{I}_j + \sum_{j \geq 2} b_j \mathbf{x} \bullet \mathbf{I}_j$$

L'interprétation peut alors être raffinée.

Notons enfin que le choix des contrastes est un choix expérimental. Il n'appartient pas au statisticien de dire si telle hypothèse est préférable à telle autre. Par exemple, pour être cohérent avec l'introduction, on aurait pu dire :

1. La question fondamentale est la compétition spermatique. Lorsque les femelles sont fécondées par plusieurs mâles, un chat avec des testicules volumineux est avantagé. Il est bien placé pour avoir plus de descendants. Par sélection, on doit trouver un volume plus grand en moyenne dans ce système. Donc la première question est : le volume des testicules est-il plus élevé en milieu urbain ? Un premier contraste permettra de comparer les deux populations urbaines aux autres ;
2. Dans les populations à faible densité, le système insulaire est à part. On testera la polyginie des populations rurales contre la monogamie de la population insulaire ?
3. Les populations urbaines sont-elles différentes entre elles ? Un troisième contraste permettra de comparer les deux populations urbaines.
4. Les populations rurales sont-elles différentes entre elles ? Un dernier contraste permettra de comparer les deux populations rurales.

```
w <- matrix(c(-1, 1, -1, 1, -1, 1, 0, -1, 0, 1, -1, 0, 0, 0, 1,
    0, 1, 0, -1, 0), ncol = 4)
dimnames(w) <- list(levels(pop), c("U", "I", "WR", "WU"))
w
```

```
     U  I WR WU
BAC -1  1 -1  0
CxR  1  0  0  1
KER -1 -1  0  0
ROM  1  0  0 -1
SJT -1  1  1  0
```

On a noté `U` pour Urbain , `I` pour insulaire, `WR` pour intra-rural et `WU` pour intra-urbain. Le résultat est :

```
 contrasts(pop) <- w
 summary(lm1 <- lm(lnvol ~ lnpoi + pop))
Call:
lm(formula = lnvol ~ lnpoi + pop)
Residuals:
    Min      1Q  Median      3Q     Max
-1.0719 -0.2169  0.0073  0.2203  0.6698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.24650    1.31330   -0.95  0.34414
lnpoi        1.00330    0.15882    6.32  3.1e-09
popU        -0.13444    0.03886   -3.46  0.00071
popI         0.27906    0.03408    8.19  1.3e-13
popWR       -0.00174    0.05144   -0.03  0.97312
popWU        0.15268    0.07014    2.18  0.03112

Residual standard error: 0.339 on 144 degrees of freedom
Multiple R-squared: 0.492,      Adjusted R-squared: 0.474
F-statistic: 27.9 on 5 and 144 DF,  p-value: <2e-16
```

Dans tous les cas, il reste à faire *l'interprétation biologique*.