

Bio-statistiques 2 / Première session

Bio-statistiques 2 / M1-AMIV / 2006 / Solution / D. Chessel

Éléments pour la solution.

1 Introduction

Les sciuridés sont une famille de mammifères rongeurs ...

Les données forment le tableau qui suit.

```
data <- read.table("http://pbil.univ-lyon1.fr/R/donnees/exp6.txt")
head(data)
```

```
  genre espece hiber  bw  aw
1  Amm Amm.hrr    N 3.60 136.8
2  Amm Amm.hrr    N 3.60 122.0
3  Amm Amm.lcr    N 2.90  92.0
4  Amm Amm.lcr    N 3.50 111.1
5  Amm Amm.lcr    N 2.90 100.6
6  Amm Amm.nls    N 4.88 154.5
```

```
summary(data)
```

```
  genre      espece  hiber      bw      aw
Slu   :43  Slu.bld: 5   N:24   Min.   : 2.280   Min.   : 46.63
Amm   : 6   Slu.elg: 5   0:45  1st Qu.: 4.000   1st Qu.: 136.80
Mrm   : 4   Amm.lcr: 3           Median : 6.800   Median : 233.10
Scr   : 4   Cyn.ldv: 3           Mean   : 8.278   Mean   : 455.67
Tas   : 4   Slu.clm: 3           3rd Qu.: 9.300   3rd Qu.: 436.21
Cyn   : 3   Slu.ltr: 3           Max.   :33.800   Max.   :3526.00
(Other): 5   (Other):47
```

```
options(digits = 4)
options(show.signif.stars = FALSE)
```

Il s'agit d'interactions potentielles entre traits biologiques.

```
x <- data$aw
y <- data$bw
xlog <- log(x)
ylog <- log(y)
gen <- data$genre
esp <- data$espece
hib <- data$hiber
d0 <- cbind.data.frame(xlog, ylog, esp, hib)[gen == "Slu", ]
```

2 Régressions par l'origine

2.1

Caractériser la distribution des variables x et y dans l'échantillon étudié.

```
par(mfrow = c(3, 3))
hist(x)
hist(y)
qqnorm(x)
qqline(x)
qqnorm(y)
qqline(y)
hist(xlog)
hist(ylog)
qqnorm(xlog)
qqline(xlog)
qqnorm(ylog)
qqline(ylog)
shapiro.test(x)

      Shapiro-Wilk normality test
data:  x
W = 0.5267, p-value = 1.429e-13

shapiro.test(xlog)

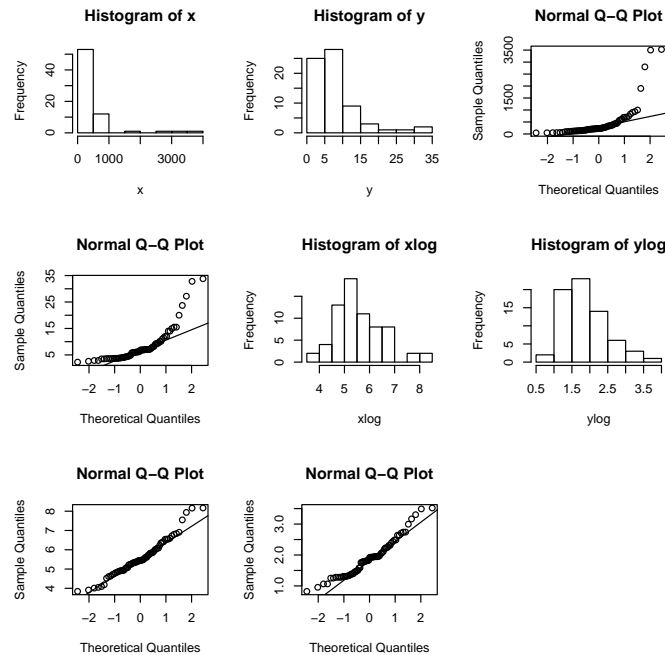
      Shapiro-Wilk normality test
data:  xlog
W = 0.9604, p-value = 0.02807

shapiro.test(y)

      Shapiro-Wilk normality test
data:  y
W = 0.7227, p-value = 4.132e-10

shapiro.test(ylog)

      Shapiro-Wilk normality test
data:  ylog
W = 0.9509, p-value = 0.008658
```



Les distributions des données brutes sont très dissymétriques, mais les distributions sont presque normales en log. On est classiquement en échelle log dans ce type de problèmes. Voir par exemple [1].

2.2

En mesurant l'investissement maternel par un rapport et en supposant ce rapport constant, on utilise le modèle $\mathbf{y} = a\mathbf{x}$ à une erreur aléatoire près. Estimer a au moindres carrés.

```
a <- coefficients(lm(y ~ -1 + x))
b <- sum(y)/sum(x)
c <- mean(y/x)
plot(xlog, ylog, pch = 20)
```

Droite de régression comme modèle linéaire sans terme constant. La réponse pour a est 0.0117.

2.3

Le calcul précédent cherche a qui minimise $\sum_{i=1}^{i=n} (y_i - ax_i)^2$. Si l'imprécision (écart-type résiduel) de la réponse croit comme la racine du prédicteur, on préfère parfois estimer le coefficient du modèle $\mathbf{y} = b\mathbf{x}$ avec b qui minimise $\sum_{i=1}^{i=n} ((y_i - bx_i)^2 / x_i)$. Donner la solution générale et la valeur obtenue pour l'exemple en cours.

La solution est :

$$b = \frac{\sum_{i=1}^{i=n} ((x_i y_i) / x_i)}{\sum_{i=1}^{i=n} ((x_i x_i) / x_i)} = \frac{\sum_{i=1}^{i=n} y_i}{\sum_{i=1}^{i=n} x_i}$$

Cas particulier : pour b , on trouve 0.0182.

2.4

Si l'imprécision (écart-type résiduel) de la réponse croît comme le prédicteur, on préfère alors estimer le coefficient du modèle $\mathbf{y} = c\mathbf{x}$ avec c qui minimise $\sum_{i=1}^{i=n} ((y_i - cx_i)^2 / x_i^2)$. Donner la solution générale et la valeur obtenue pour l'exemple en cours.

La solution est :

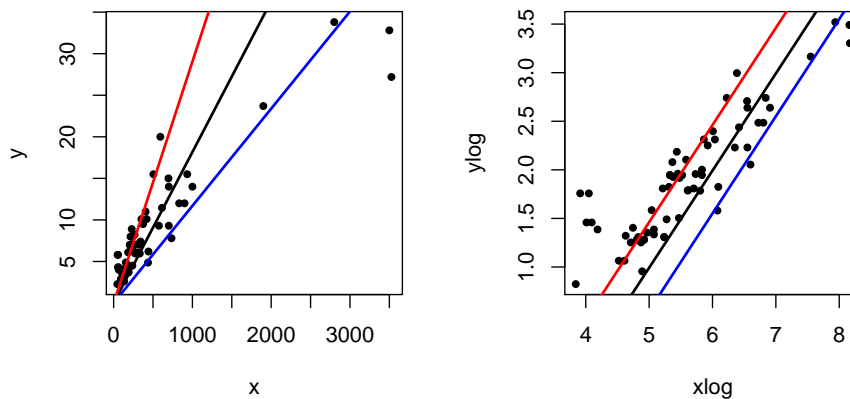
$$b = \frac{\sum_{i=1}^{i=n} ((x_i y_i) / x_i^2)}{\sum_{i=1}^{i=n} ((x_i x_i) / x_i^2)} = \frac{1}{n} \sum_{i=1}^{i=n} \frac{y_i}{x_i}$$

Cas particulier : pour c , on trouve 0.029.

2.5

Représenter les trois modèles sur les graphiques en données brutes à gauche et en échelle log-log à droite.

```
par(mfrow = c(1, 2))
plot(x, y, pch = 20)
abline(c(0, a), col = "blue", lwd = 2)
abline(c(0, b), col = "black", lwd = 2)
abline(c(0, c), col = "red", lwd = 2)
plot(xlog, ylog, pch = 20)
abline(c(log(a), 1), col = "blue", lwd = 2)
abline(c(log(b), 1), col = "black", lwd = 2)
abline(c(log(c), 1), col = "red", lwd = 2)
```



On retient de cet essai qu'il est toujours délicat de mesurer un rapport, du fait de l'incertitude aléatoire présente au dénominateur. Le modèle linéaire sur les variables transformées s'impose.

3 Deux variables

3.1

$\mathbf{x} = (x_1, \dots, x_n)$ et $\mathbf{y} = (y_1, \dots, y_n)$ sont deux variables quelconques. Donner l'équation des **deux** droites de régression.

La droite de régression de y sur x est $y = ax + b$ avec :

$$a = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\text{var}(\mathbf{x})}$$

$$b = \bar{y} - a\bar{x}$$

La droite de régression de x sur y est $x = cy + d$ avec :

$$c = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\text{var}(\mathbf{y})}$$

$$d = \bar{x} - c\bar{y}$$

Donc on devra utiliser la droite $y = \frac{x}{c} - \frac{d}{c}$.

3.2

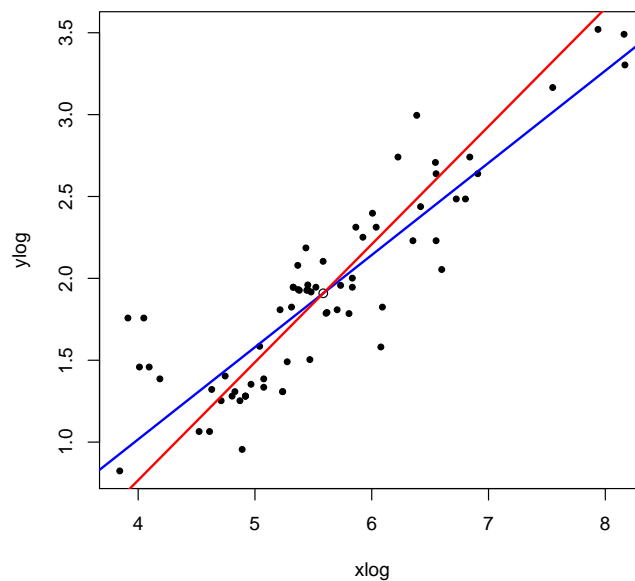
Représenter sur le graphique prévu à cet effet les deux droites de régression entre les variables `xlog` et `ylog`.

```
w = coefficients(lm(xlog ~ ylog))
w <- c(-w[1], 1)/w[2]
plot(xlog, ylog, pch = 20)
points(mean(xlog), mean(ylog))
abline(lm(ylog ~ xlog), col = "blue", lwd = 2)
abline(w, col = "red", lwd = 2)
w1 <- coefficients(lm(ylog ~ xlog))
paste("Y/X", "b", w1[1], "a", w1[2])
```

```
[1] "Y/X b -1.23450860049149 a 0.562846677319532"
```

```
paste("X/Y", "b", w[1], "a", w[2])
```

```
[1] "X/Y b -2.11441770016882 a 0.720396400001666"
```



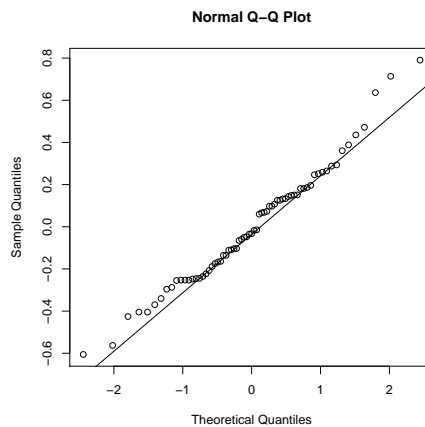
3.3

Les résidus de la régression de `ylog` sur `xlog` sont-ils normaux ?

```
w = residuals(lm(ylog ~ xlog))
shapiro.test(w)
```

```
Shapiro-Wilk normality test
data: w
W = 0.9791, p-value = 0.3002
```

```
qqnorm(w)
qqline(w)
```



Rien ne s'oppose à considérer les résidus comme normaux.

3.4

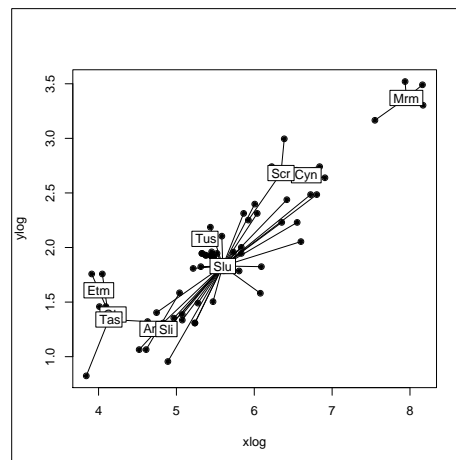
Caractériser le lien entre les deux variables `ylog` et `xlog`.

Ce lien est remarquablement linéaire. Les résidus sont normaux. On a 78% de variance expliquée. La relation linéaire est très forte. On est bien dans un modèle linéaire.

3.5

Comment est obtenue cette figure et quelle question soulève-t-elle ?

```
library(ade4)
plot(xlog, ylog)
s.class(cbind.data.frame(xlog, ylog), gen, add.p = T, cell = 0)
```



La question est claire. Le lien entre taille adulte et taille du jeune est-elle un simple effet de la taxonomie ? Il est clair en effet que la corrélation est en partie inter-générique et en partie intra-générique.

4 Modèles linéaires

4.1

Les ANOVA des modèles suivants présentent une curieuse propriété. Laquelle ? Expliquer le phénomène.

```
lm1 <- lm(ylog ~ xlog + esp)
lm2 <- lm(ylog ~ xlog + esp + hib)
lm3 <- lm(ylog ~ xlog + esp + gen)
lm4 <- lm(ylog ~ xlog + esp + hib + gen)
```

Les tableaux d'analyse de variance sont identiques. Toute classe d'un niveau est contenu dans une classe du niveau suivant. Les sous-espaces engendrés sont emboîtés :

$$\mathcal{E}_{esp} \subset \mathcal{E}_{gen} \text{ et } \mathcal{E}_{esp} \subset \mathcal{E}_{hib}.$$

Les variables ajoutées sont mathématiquement sans effets supplémentaires et éliminées. On a ici une protection assurée par la fonction contre une non-maîtrise du schéma théorique.

4.2

Les ANOVA des modèles suivants présentent une curieuse propriété. Laquelle ? Expliquer le phénomène.

```
lm1 <- lm(ylog ~ xlog + esp)
lm5 <- lm(ylog ~ esp + xlog)
```

La variable `ylog` est très liée `xlog` mais l'effet disparaît totalement après introduction de l'espèce. C'est un cas typique de redondance. Toute la variabilité de la taille adulte est inter-spécifique. Il y a *confusion de facteurs*.

4.3

Comparer les modèles :

```
lm6 <- lm(ylog ~ hib + gen + xlog)
lm7 <- lm(ylog ~ gen + hib + xlog)
lm8 <- lm(ylog ~ gen + xlog + hib)
```

Ces modèles ont même summary : c'est une propriété mathématique. Quelle que soit la place de son introduction la variable `hib` est sans aucun effet : c'est une propriété des données. Le rôle de l'hibernation est très difficile à mettre en évidence.

4.4

Pour éviter toute difficulté, on s'en tient au sous-ensemble des données formé par les spermophiles. Donner une légende à la figure proposée, qui a été obtenue par :

```
lmred = lm(ylog ~ xlog + hib, data = d0)
par(mfrow = c(1, 3))
plot(lmred, 1:3, c("A", "B", "C"))
```

On pourra donner la définition mathématique des figures ou leur fonction dans la lecture des données.

- * A Tracé des résidus (donnée - modèle) vs. les valeurs du modèles. Il n'y a pas d'erreur systématique dans une région du modèle qui permet de détecter un écart à la linéarité.
- * B QQ-plot ou tracé des quantiles observés et des quantiles du modèle de la loi normale associés. Le faible écart à la droite est le signe d'une excellente normalité des résidus.
- * C Résidus normalisés qui indique seulement le niveau de l'erreur indépendamment du signe de l'erreur. Permet de détecter des régions du modèle où la précision est moins grande, ce qui nuit à la valeur des tests. La régression locale est plutôt trompeuse : il est peu vraisemblable, au vu du dessin, que la précision ne soit pas constante. Simplement, aux deux extrémités, on a peu de données d'espèces particulièrement grosses ou petites, donc différentes.

4.5

Que conclure sur la question posée ?

L'équilibre des groupes est maintenant défavorable. Il n'y a aucun effet. En fait, le problème est difficile. Quand on travaille sur des espèces, le lien phylogénétique introduit un ensemble de contraintes qui cachent certains effets ou les rendent inaccessibles.

5 Pour tous les goûts

5.1

Soit l'espace vectoriel $\mathcal{E} = \mathbb{R}^n$ muni du produit scalaire canonique. \mathcal{A} est un sous-espace vectoriel de \mathcal{E} . $\mathbf{P}_{\mathcal{A}}$ désigne le projecteur orthogonal sur \mathcal{A} . Si \mathcal{A} est

engendré par un unique vecteur $\mathbf{x} = (x_1, \dots, x_n)$ quelle est la matrice de \mathbf{P}_A dans la base canonique ?

$$\text{Mat}(\mathbf{P}_A) = \frac{1}{\|\mathbf{x}\|^2} \mathbf{x}\mathbf{x}^T$$

C'est la conséquence directe de la définition.

5.2

\mathcal{B} est un autre sous-espace vectoriel de \mathcal{E} . \mathbf{P}_B désigne le projecteur orthogonal sur \mathcal{B} . Si tout vecteur de \mathcal{A} est orthogonal à tout vecteur de \mathcal{B} , que peut-on dire de \mathbf{P}_A , \mathbf{P}_B et \mathbf{P}_{A+B} ? Le résultat a-t-il un intérêt en statistique ?

Les sous-espace sont orthogonaux. Chacun d'entre eux possède des bases orthonormales. Prendre une base orthonormale de \mathcal{A} , une base orthonormale de \mathcal{B} : la réunion des deux donne une base orthonormale de $\mathcal{A} + \mathcal{B}$ et le projecteur sur la somme est la somme des projecteurs.

5.3

On fait, en conditions constantes, 12 expériences induisant une réponse (`rep`) en fonction d'un produit `dos` qui est soit absent (modalité `temoin`), soit présent en dose faible (modalité `faible`), soit présent en dose forte (modalité `fort`). Implanter les données :

```
dos = factor(c("fort", "faible", "fort", "temoin", "faible", "faible",
              "faible", "fort", "fort", "fort", "temoin", "temoin"))
rep = c(15, 11.5, 11.8, 2.4, 9.9, 12.6, 9.9, 12.7, 10.8, 9.6, 9.9,
        6.3)
summary(dos)
```

```
faible  fort temoin
      4      5      3
```

```
summary(rep)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 2.40   9.83   10.40  10.20  12.00  15.00
```

```
summary(lm(rep ~ dos))
```

```
Call:
lm(formula = rep ~ dos)
Residuals:
   Min     1Q  Median     3Q    Max
-3.800 -1.101 -0.040  0.946  3.700
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.98      1.18     9.30 6.5e-06
dosfort         1.01      1.58     0.64  0.541
dostemoin      -4.77      1.80    -2.65  0.026
```

```
Residual standard error: 2.36 on 9 degrees of freedom
Multiple R-Squared: 0.569, Adjusted R-squared: 0.474
F-statistic: 5.95 on 2 and 9 DF, p-value: 0.0226
```

Expliquer en quoi le résultat de `summary(lm(rep~dos))` n'est pas acceptable.

Le premier niveau `faible` sert de témoins et les contrastes sont complètement inadaptes.

5.4

Pour remédier au problème donner une solution utilisant la fonction `factor`.

```
z <- factor(dos, levels = c("temoin", "faible", "fort"))
summary(lm(rep ~ z))

Call:
lm(formula = rep ~ z)
Residuals:
    Min       1Q   Median       3Q      Max
-3.800 -1.101 -0.040  0.946  3.700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.20      1.36     4.55  0.0014
zfaible        4.78      1.80     2.65  0.0265
zfort          5.78      1.72     3.35  0.0085

Residual standard error: 2.36 on 9 degrees of freedom
Multiple R-Squared:  0.569,    Adjusted R-squared:  0.474
F-statistic: 5.95 on 2 and 9 DF,  p-value: 0.0226
```

5.5

Pour remédier au problème, donner une solution utilisant la fonction `contrasts`.

```
zz <- factor(dos)
contrasts(zz) =_matrix(c(1, 0, 0, 0, 1, 0), 3)
summary(lm(rep ~ zz))

Call:
lm(formula = rep ~ zz)
Residuals:
    Min       1Q   Median       3Q      Max
-3.800 -1.101 -0.040  0.946  3.700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.20      1.36     4.55  0.0014
zz1            4.78      1.80     2.65  0.0265
zz2            5.78      1.72     3.35  0.0085

Residual standard error: 2.36 on 9 degrees of freedom
Multiple R-Squared:  0.569,    Adjusted R-squared:  0.474
F-statistic: 5.95 on 2 and 9 DF,  p-value: 0.0226
```

Autre solution :

```
zz1 <- factor(dos)
contrasts(zz1) = contr.treatment(3, base = 3)
summary(lm(rep ~ zz1))

Call:
lm(formula = rep ~ zz1)
Residuals:
    Min       1Q   Median       3Q      Max
-3.800 -1.101 -0.040  0.946  3.700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.20      1.36     4.55  0.0014
zz11           4.78      1.80     2.65  0.0265
zz12           5.78      1.72     3.35  0.0085

Residual standard error: 2.36 on 9 degrees of freedom
Multiple R-Squared:  0.569,    Adjusted R-squared:  0.474
F-statistic: 5.95 on 2 and 9 DF,  p-value: 0.0226
```

5.6

Étendre la solution précédente pour obtenir des tests des hypothèses nulles :

1. *la présence du produit n'a pas d'effet* ;
2. *le produit n'intervient que par sa présence seulement et la variation de dose est sans effet.*

Conclure.

```
contrasts(dos) = matrix(c(1, 1, 0, -1, 1, 0), 3)
summary(lm(rep ~ dos))
```

```
Call:
lm(formula = rep ~ dos)
Residuals:
    Min       1Q   Median       3Q      Max
-3.800 -1.101 -0.040  0.946  3.700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.200     1.362    4.55  0.0014
dos1          5.278     1.575    3.35  0.0085
dos2          0.503     0.791    0.64  0.5412

Residual standard error: 2.36 on 9 degrees of freedom
Multiple R-Squared:  0.569,    Adjusted R-squared:  0.474
F-statistic: 5.95 on 2 and 9 DF,  p-value: 0.0226
```

On rejette la première hypothèse, on n'a pas d'argument pour rejeter le seconde.
Le produit a de l'effet, le rôle de la dose n'est pas significatif.

Références

- [1] L. W. Aarssen. Why don't bigger plants have proportionately bigger seeds? *Oikos*, 111(1) :199–207, 2005.