Bio-statistiques 2 / Première session

Bio-statistiques 2 / M1-AMIV / 2006 / BI6003M1 P1 / D. Chessel

2 juin 2006 - 14h-16 h

Tous les documents sont autorisés. Les échanges entre étudiants sont strictement interdits. Vous devez utiliser votre micro-ordinateur personnel. R *et ade4 sont recommandés. L'accès à internet est inutile. Répondre directement sur la feuille en laissant les agrafes.*
Les questions permettent une approche plus mathématique, plus procédurale ou plus biologique. Eviter simplement de remplir sans justification.

FIG. 1 – Quelques portraits de sciuridés rencontrés sur Internet.

1) *Ammospermophilus harrisii* 2) *Glaucomys volans* 3) *Spermophilus franklini*
4) *Tamiasciurus hudsonicus* 5) *Ammospermophilus leucurus* 6) *Marmota marmota*
7) *Tamias minimus* 8) *Sciurus carolinensis* 9) *Tamias palmeri*
10) *Cynomys ludovicianus* 11) *Spermophilus elegans* 12) *Tamias striatus*

1 Introduction

Les sciuridés sont une famille de mammifères rongeurs qui groupe d'une part les écureuils volants (sous-famille des Pteromyinés) et d'autre part les écureuils arboricoles et les écureuils terrestres (chiens de prairies, marmottes, spermophiles et tamias rayés, sous- famille des sciurinés). On trouvera quelques photos de la famille dans la figure 1.

Pour 37 espèces de sciuridés, Delphine Vacheron (stage M1) a compilé dans la littérature, pour une ou plusieurs populations, une estimation du poids moyen des jeunes à la naissance et du poids moyen des adultes. On utilisera le code :

n	*Espèce*	Code	n	*Espèce*	Code
1	*Ammospermophilus harrisii*	Amm.hrr	2	*Ammospermophilus leucurus*	Amm.lcr
3	*Ammospermophilus nelsoni*	Amm.nls	4	*Cynomys ludovicianus*	Cyn.ldv
5	*Eutamias palmeri*	Etm.plm	6	*Eutamias panamintinus*	Etm.pnm
7	*Glaucomys volans*	Glc.vln	8	*Marmota flaviventris*	Mrm.flv
9	*Marmota marmota*	Mrm.mrm	10	*Marmota monax*	Mrm.mnx
11	*Sciurus carolinensis*	Scr.crl	12	*Sciurus granatensis*	Scr.grn
13	*Sciurus niger*	Scr.ngr	14	*Spermophilus armatus*	Slu.arm
15	*Spermophilus beecheyi*	Slu.bch	16	*Spermophilus beldingi*	Slu.bld
17	*Spermophilus columbianus*	Slu.clm	18	*Spermophilus elegans*	Slu.elg
19	*Spermophilus franklini*	Slu.frn	20	*Spermophilus harrisii*	Slu.hrr
21	*Spermophilus lateralis*	Slu.ltr	22	*Spermophilus mohavensis*	Slu.mhv
23	*Spermophilius mexicanus*	Sli.mxc	24	*Spermophilus parryii*	Slu.prr
25	*Spermophilus richardsonii*	Slu.rch	26	*Spermophilus saturatus*	Slu.sat
27	*Spermophilus spilosoma*	Slu.spl	28	*Spermophilus tereticaudus*	Slu.trt
29	*Spermophilus townsendii*	Slu.twn	30	*Spermophilus tridecemlineatus*	Slu.trd
31	*Spermophilus variegatus*	Slu.vrg	32	*Tamias minimus*	Tas.mnm
33	*Tamias palmeri*	Tas.plm	34	*Tamias panamintinus*	Tas.pnm
35	*Tamias striatus*	Tas.str	36	*Tamiasciurus hudsonicus*	Tus.hds

Pour chaque population, on dispose de 5 variables observées, respectivement :

1. `genre` : nom du genre des individus de la population étudiée ;
2. `especе` : nom de l'espèce (dans le code ci-dessus) ;
3. `hiber` : statut d'hibernation : facteur prenant la modalité `O` pour une espèce qui hiberne et `N` pour une espèce qui n'hiberne pas ;
4. `bw` (birth weight) moyenne des poids à la naissance des jeunes observés, en grammes ;
5. `aw` (adult weight) moyenne des poids des adultes observés, plus particulièrement des femelles lorsqu'on en dispose.

Les données forment le tableau qui suit.

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

```
  genre  especе 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        especе   hiber        bw              aw
Slu    :43   Slu.bld: 5   N:24   Min.   : 2.280   Min.   :  46.63
Amm    : 6   Slu.elg: 5   O: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
```

Delphine Vacheron pose le problème en ces termes :

> On s'intéresse à la manière dont les femelles allouent des ressources à leurs jeunes en fonction de contraintes environnementales telles que l'hibernation. L'hibernation a pour principale conséquence de diminuer très sensiblement le temps annuel d'activité des individus : la reproduction est alors très contrainte puisque les individus doivent se reproduire puis élever leurs jeunes en un minimum de temps : leur progéniture et eux doivent être suffisamment en forme pour pouvoir passer l'hiver en état d'hibernation sans mourir. Sachant cela on se pose notamment la question de savoir si les individus non hibernants investissent plus que les autres ? On mesure cet investissement par le poids du jeune à la naissance sur le poids de l'adulte.

Il s'agit d'interactions potentielles entre traits biologiques. Dans toutes la suite on utilisera les vecteurs :

```
x <- data$aw
y <- data$bw
xlog <- log(x)
ylog <- log(y)
gen <- data$genre
esp <- data$esp
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 $\mathbf{x}$ et $\mathbf{y}$ dans l'échantillon étudié.

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.

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 $\sum_{i=1}^{i=n} \left((y_i - bx_i)^2 / x_i \right)$. Donner la solution générale et la valeur obtenue pour l'exemple en cours.

2.4

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

2.5

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

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

3.2

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

3.3

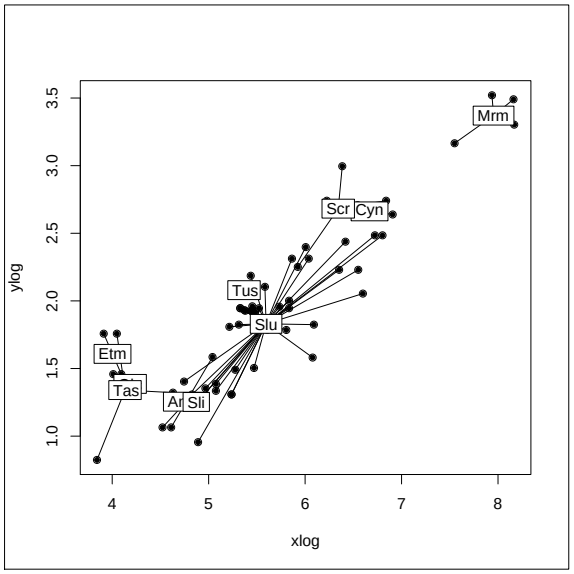
Les résidus de la régression de `ylog` sur `xlog` sont-ils normaux ?

3.4

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

3.5

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



4 Modèles linéaires

4.1

On se demande alors si ce lien est entièrement déterminé par la taxonomie (comme première approximation de la phylogénie). 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)
```

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)
```

4.3

Comparer les modèles :

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

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"))
```

4.5

Que conclure sur la question posée ?

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 $\mathbf{x} = (x_1, \ldots, x_n)$ quelle est la matrice de $\mathbf{P}_\mathcal{A}$ dans la base canonique ?

5.2

$\mathcal{B}$ est un autre sous-espace vectoriel de $\mathcal{E}$. $\mathbf{P}_\mathcal{B}$ 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 $\mathbf{P}_\mathcal{A}$, $\mathbf{P}_\mathcal{B}$ et $\mathbf{P}_{\mathcal{A}+\mathcal{B}}$? Le résultat a-t-il un intérêt en statistique ?

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`). Les résultats sont :

	dos	rep
1	fort	15.00
2	faible	11.50
3	fort	11.80
4	temoin	2.40
5	faible	9.90
6	faible	12.60
7	faible	9.90
8	fort	12.70
9	fort	10.80
10	fort	9.60
11	temoin	9.90
12	temoin	6.30

Saisir les deux variables pour obtenir :

```
summary(dos)
```

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

```
summary(rep)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.400   9.825  10.350  10.200  12.000  15.000
```

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

5.4

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

5.5

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

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.

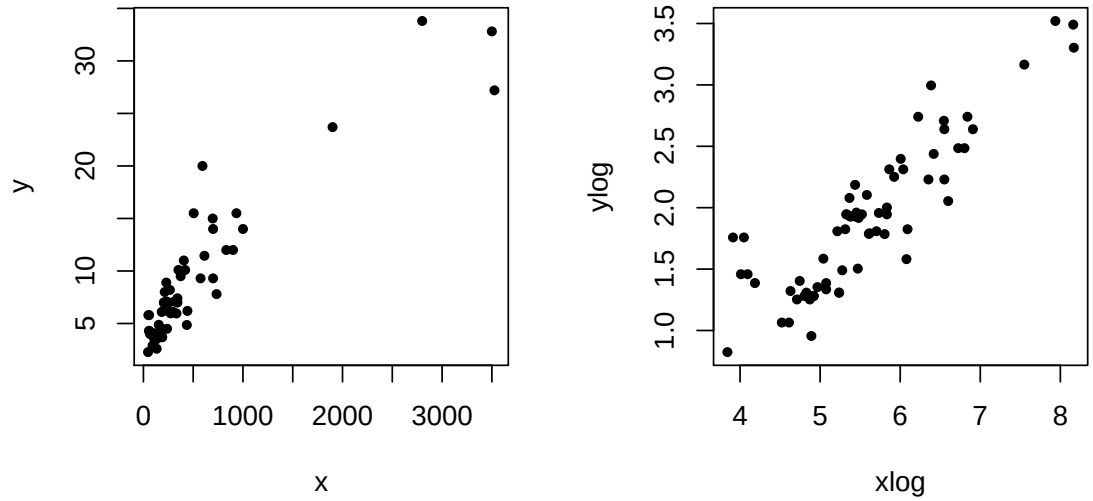
Indiquer ici clairement votre numéro d'anonymat

2.1

2.2

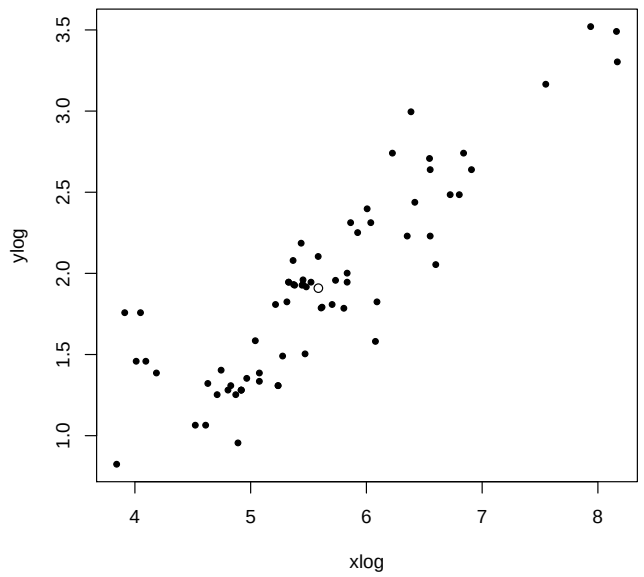
2.3

2.4



2.5

3.1



3.2

3.3

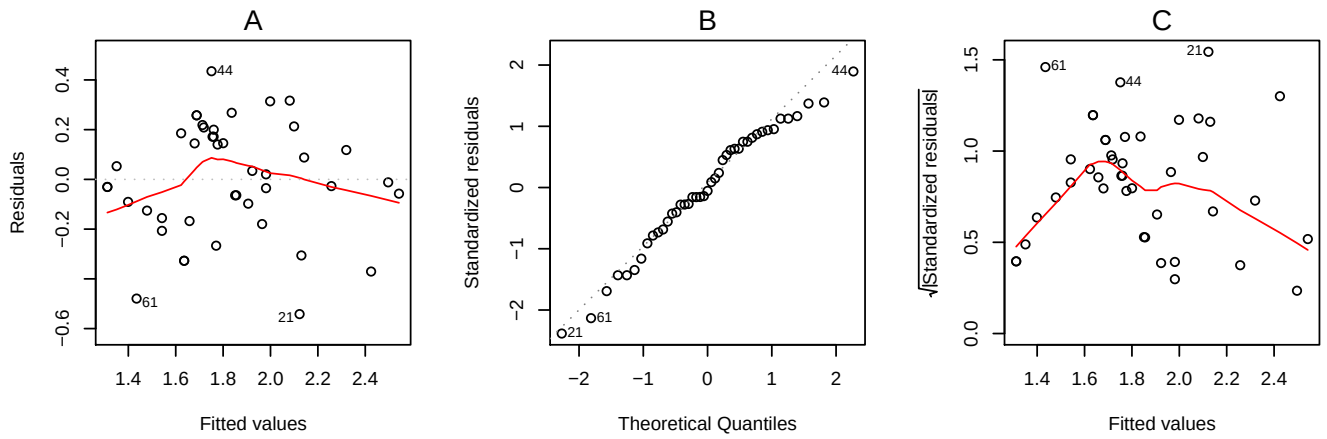
3.4

3.5

4.1

4.2

4.3


A
Residuals
Fitted values
44
61
21
B
Standardized residuals
Theoretical Quantiles
44
61
21
C
√|Standardized residuals|
Fitted values
61
44
21

4.4

4.5

5.1

5.2

5.3

5.4

5.5

5.6