Biostatistiques avancées avec R

Introduction aux modèles à effets aléatoires

Christophe Lalanne

www.aliquote.org

Synopsis

Éléments de contexte

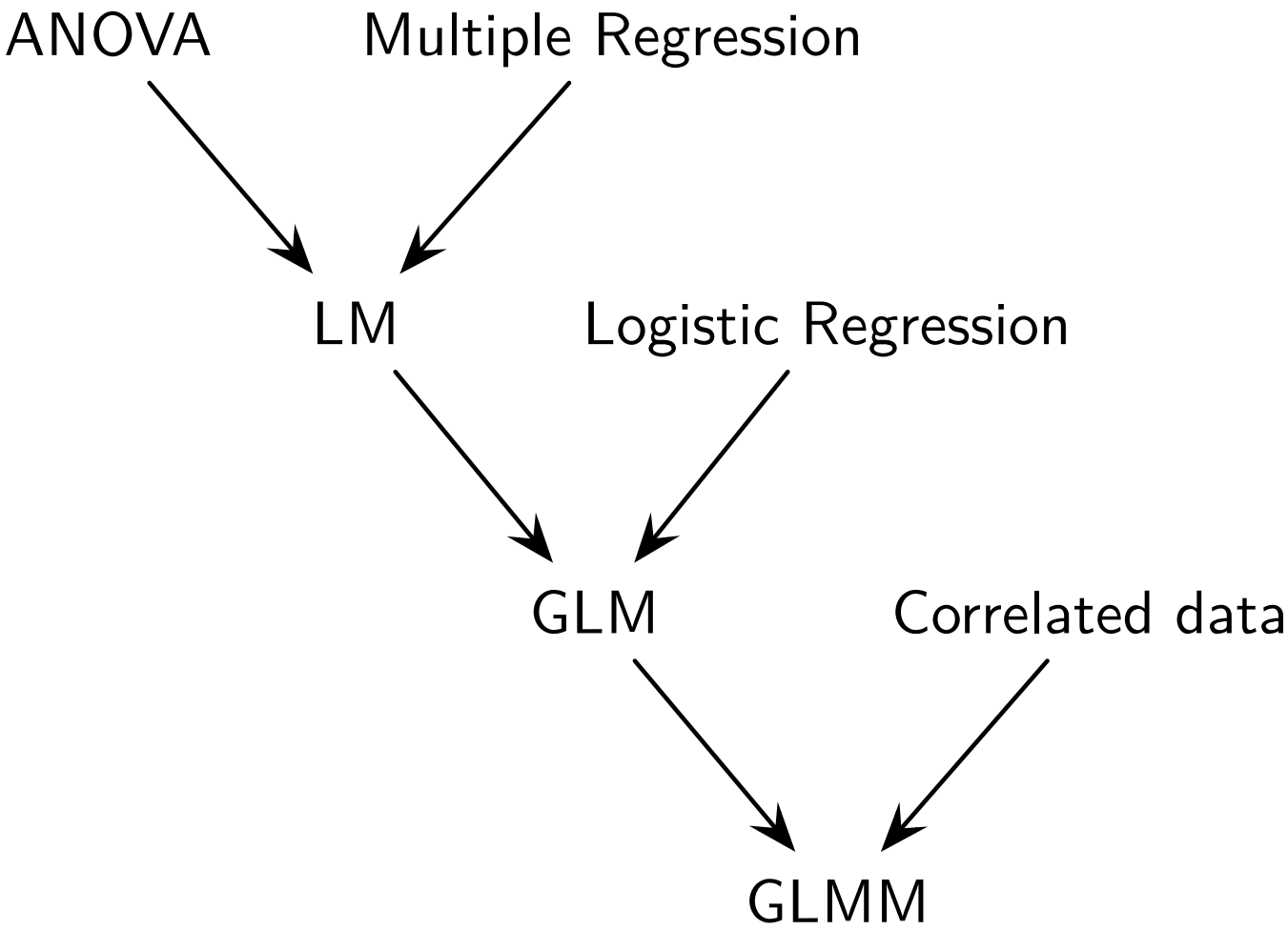
Tests de Student pour échantillons appariés

Analyse de variance et mesures répétées

Modèles à effet aléatoires

Modèles pour données longitudinales

Tour d'horizon



Types de modèles

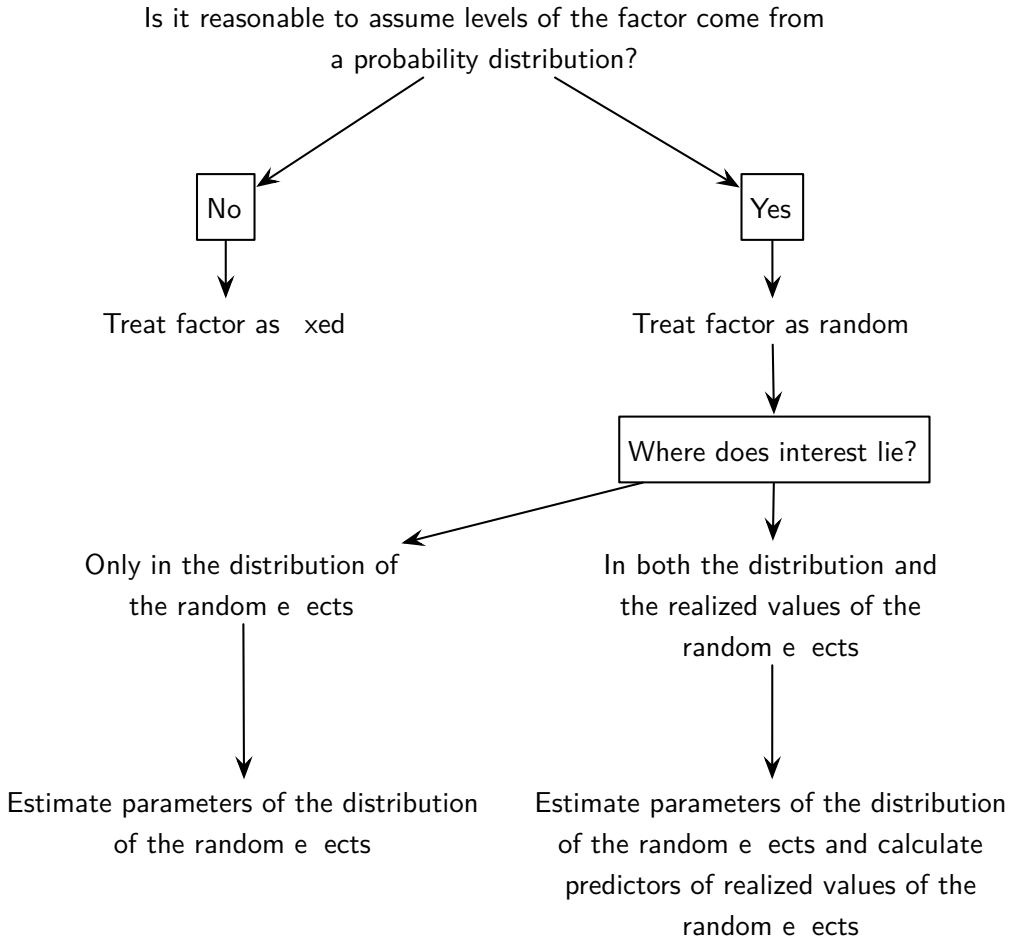
Par rapport au modèle linéaire (généralisé), les modèles à effets mixtes permettent d'introduire des termes ou effets aléatoires permettant de modéliser la corrélation entre les unités statistiques.

Les modèles hiérarchiques ou multi-niveaux prennent en compte le regroupement des individus en unités homogènes (e.g., étudiants dans des classes), la corrélation intraclasse (e.g., données longitudinales) ou un mélange des deux (e.g., performance au cours du temps pour différents groupes de sujets)[8,7,10,6,3].

Deux types d'approche : modèles conditionnels (typiquement, les modèles mixtes) et modèles marginaux (e.g., GLS, GEE). Dans le second cas, on cherche plutôt à modéliser un effet moyen (population) à partir d'une matrice de corrélation intra-unité pré-définie.

Effets fixes *versus* aléatoires

Peu de définitions consensuelles[5], mais on peut se demander si l'on souhaite simplement estimer les paramètres des termes aléatoires ou dériver des prédictions au niveau individuel[12].



Comparaison de moyennes

Effet de somnifères sur le temps de sommeil[11].

```
> data(sleep)
> t.test(extra ~ group, data = sleep)


        Welch Two Sample t-test

data:  extra by group
t = -2, df = 20, p-value = 0.08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.365  0.205
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

```
> t.test(extra ~ group, data = sleep, paired = TRUE)

        Paired t-test

data:  extra by group
t = -4, df = 9, p-value = 0.003
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.46 -0.70
sample estimates:
mean of the differences
                  -1.58
```

Ignorer la corrélation intra-unité résulte généralement en un test moins puissant.

Pourquoi une telle perte de puissance ?

On sait que

$$\text{Var}(X_1 - X_2) = \text{Var}(X_1) + \text{Var}(X_2) - 2\text{Cov}(X_1,X_2),$$

avec $\text{Cov}(X_1,X_2) = \rho\sqrt{\text{Var}(X_1)}\sqrt{\text{Var}(X_2)}$.

Si l'on suppose $\text{Cov}(X_1,X_2) = 0$, cela revient à sur-estimer la variance des différences (le test de Student pour échantillons appariés est essentiellement un problème à un échantillon) puisque $\text{Cov}(X_1,X_2)$ sera généralement > 0.

```
> xyplot(extra ~ group, sleep, groups = ID, type = "a")
```

Données d’illustration

Enzyme digestive et problèmes intestinaux[12].

ID	Pill type				Average
	None	Tablet	Capsule	Coated	
1	44.5	7.3	3.4	12.4	16.9
2	33.0	21.0	23.1	25.4	25.6
3	19.1	5.0	11.8	22.0	14.5
4	9.4	4.6	4.6	5.8	6.1
5	71.3	23.3	25.6	68.2	47.1
6	51.2	38.0	36.0	52.6	44.5
Average	38.1	16.5	17.4	31.1	25.8

Format large (« wide ») $\leftrightarrow$ long

```
> labs <- c("None","Tablet","Capsule","Coated")
> fat <- data.frame(fecfat = c(44.5,33.0,19.1,9.4,71.3,51.2,
+                              7.3,21.0,5.0,4.6,23.3,38.0,
+                              3.4,23.1,11.8,4.6,25.6,36.0,
+                              12.4,25.4,22.0,5.8,68.2,52.6),
+                   pilltype = gl(4, 6, labels=labs),
+                   subject = gl(6, 1))
> head(fat)
```

```
  fecfat pilltype subject
1   44.5     None       1
2   33.0     None       2
3   19.1     None       3
4    9.4     None       4
5   71.3     None       5
6   51.2     None       6
```

Composantes de variance

Il n'y a qu'un seul prédicteur (« Pill type ») qui reflète le lien entre le facteur sujet et le temps d'administration (équivalent aux différents produits administrés séquentiellement).

Différentes façons de décomposer la variance totale :

1. ANOVA à un facteur : aov(fecfat pilltype, data=fat)
2. ANOVA à deux facteurs : aov(fecfat pilltype + subject, data=fat)
3. ANOVA à mesures répétées : aov(fecfat pilltype + Error(subject), data=fat)

Tableaux d'ANOVA

Source	DF	SS	MS	M1	M2*/M3
pilltype	3	2009	669.5	669.5/359.7	669.5/107.0
				p=0.169	p=0.006
subject	5	5588	1117.7	—	1117.7/107.0
					p=0.000*
Residuals	15	1605	107.0	—	—

Le 1er modèle qui suppose des observations indépendantes n'enlève pas la variance entre sujets qui représente environ 78 % de la résiduelle.

Les deux autres modèles incorporent des effets spécifiques aux sujets :

$$y_{ij} = \mu + \text{subject}_i + \text{pilltype}_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\varepsilon).$$

Dans le 3e modèle, on suppose en plus que $\text{subject}_i \sim \mathcal{N}(0, \sigma_s^2)$, indépendant de ε_{ij}.

L'inclusion d'un effet aléatoire spécifique aux individus permet de modéliser différents types de structure de corrélation intra-unité.

Corrélation entre les mesures :

$$\text{Cor}(y_{ij}, y_{ik}) = \frac{\text{Cov}(y_{ij}, y_{ik})}{\sqrt{\text{Var}(y_{ij})}\sqrt{\text{Var}(y_{ik})}}.$$

Puisque μ et pilltype sont fixes, et que $\varepsilon_{ij} \perp \mathit{subject}_i$, on a :

$$\begin{aligned} \text{Cov}(y_{ij}, y_{ik}) &= \text{Cov}(\mathit{subject}_i, \mathit{subject}_i) \\ &= \text{Var}(\mathit{subject}_i) \\ &= \sigma_s^2, \end{aligned}$$

d'où $\text{Var}(y_{ij}) = \text{Var}(\text{subject}_i) + \text{Var}(\varepsilon_{ij}) = \sigma_s^2 + \sigma_\varepsilon^2$ (pour chaque observation).

Corrélation intraclasse

On a donc

$$\text{Cor}(y_{ij}, y_{ik}) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\varepsilon^2}$$

qui représente la proportion de variance totale expliquée par le facteur `subject`. On appelle cette quantité le coefficient de corrélation intraclasse, ρ, et il représente la proximité des observations prises sur différents sujets (similarité intra-classe).

En d'autres termes, la variabilité inter-individuelle augmente ou diminue simultanément toutes les mesures d'un même sujet.

Structure de variance-covariance

La structure de variance-covariance ($\sigma^2 = \sigma_s^2 + \sigma_\varepsilon^2$) dans le modèle précédent (modèle à « intercept » aléatoire) est appelée structure de symétrie composée.

$$\begin{bmatrix} \sigma_s^2 + \sigma_\varepsilon^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 + \sigma_\varepsilon^2 & \sigma_s^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2 + \sigma_\varepsilon^2 & \sigma_s^2 \\ \sigma_s^2 & \sigma_s^2 & \sigma_s^2 & \sigma_s^2 + \sigma_\varepsilon^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

Les observations d'un même sujet sont modélisées comme étant corrélées *via* leur effet aléatoire partagé.

Estimation de *rho*

Modèle à intercept aléatoire :

```
> library(nlme)
> m <- lme(fecfat ~ pilltype, data = fat,
+          random = ~ 1 | subject)
> anova(m)
```

```
            numDF denDF F-value p-value
(Intercept)     1    15   14.27  0.0018
pilltype        3    15    6.26  0.0057
```

```
> intervals(m, which = "var-cov")
Approximate 95% confidence intervals

 Random Effects:
  Level: subject
                lower est. upper
sd((Intercept))     8 15.9  31.6

 Within-group standard error:
lower  est. upper
 7.23 10.34 14.79
```

Remarque : Les effets alétoires sont estimés après avoir « supprimé » les effets fixes (méthode REML). Les tests sur les effets aléatoires ($H_0 : \sigma^2 = 0$) par LRT sont conservateurs.

```
> VarCorr(m)
```

```
subject = pdLogChol(1)
            Variance StdDev
(Intercept) 253      15.9
Residual    107      10.3
```

```
> sigma.s <- as.numeric(VarCorr(m)[1,2])
> sigma.eps <- as.numeric(VarCorr(m)[2,2])
> sigma.s^2 / (sigma.s^2 + sigma.eps^2)
```

```
[1] 0.704
```

Approche par l'ANOVA

On peut retrouver les composantes de variance (sujets + résiduelle) à partir d'un tableau d'ANOVA.

```
> ms <- anova(lm(fecfat ~ pilltype + subject,
+                data=fat))[[3]]
> vs <- (ms[2] - ms[3]) / nlevels(fat$pilltype)
> vr <- ms[3]
> vs / (vs + vr)
```

```
[1] 0.703
```

Approche par GLS

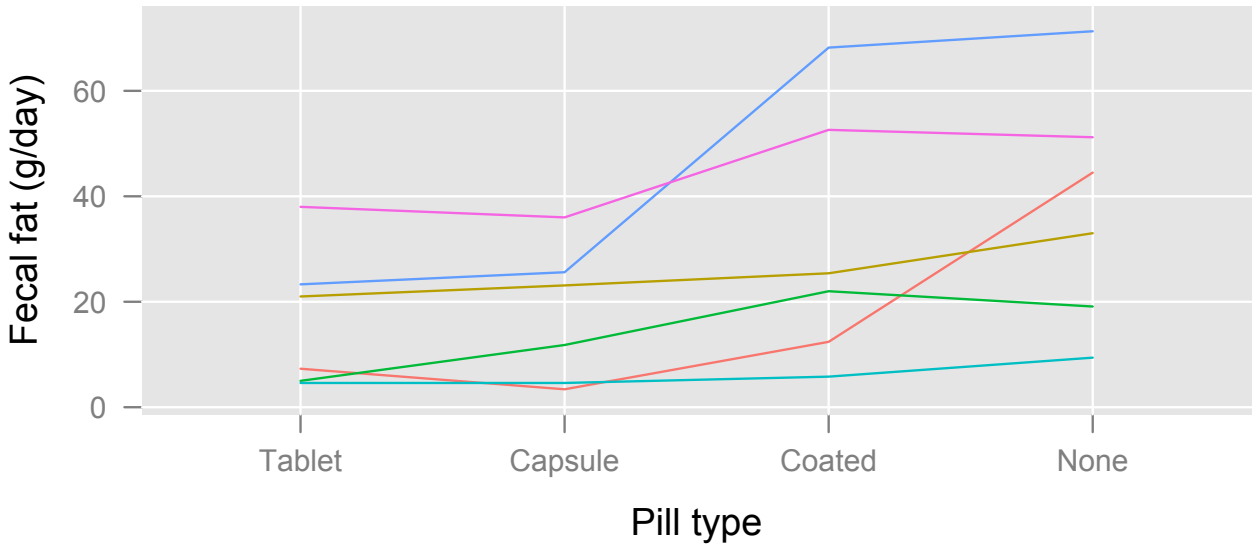
Enfin, on peut imposer une structure de symétrie composée pour la corrélation des erreurs *via* un modèle de type GLS :

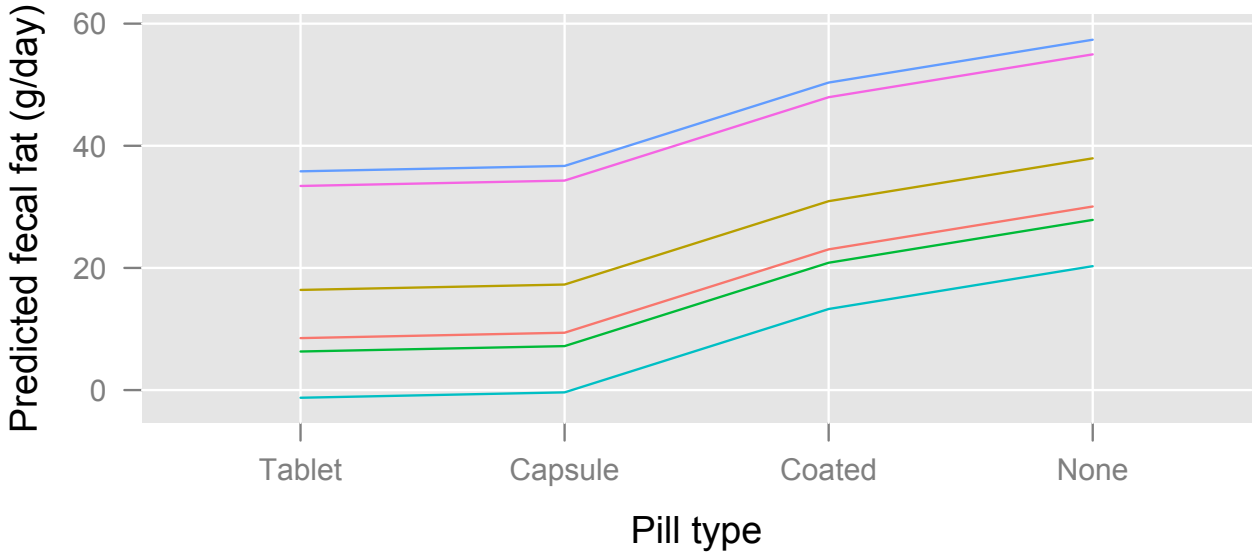
```
> gls.fit <- gls(fecfat ~ pilltype, data=fat,
+                corr=corCompSymm(form= ~ 1 | subject))
> ## anova(gls.fit)
> intervals(gls.fit, which = "var-cov")
Approximate 95% confidence intervals

 Correlation structure:
    lower  est. upper
Rho 0.272 0.703 0.914
attr(,"label")
[1] "Correlation structure:"

 Residual standard error:
lower  est. upper
 11.6  19.0  30.9
```

```
> fat$pred <- predict(m)
> p1 <- xyplot(fecfat ~ reorder(pilltype, fecfat), data=fat,
+              groups=subject, type="a", xlab="Pill type",
+              ylab="Fecal fat (g/day)",
+              scales=list(y=list(at=seq(0, 80, by=20))),
+              par.settings = ggplot2like(),
+              axis = axis.grid)
> p2 <- xyplot(pred ~ reorder(pilltype, fecfat), data=fat,
+              groups=subject, type="a", xlab="Pill type",
+              ylab="Predicted fecal fat (g/day)",
+              scales=list(y=list(at=seq(0, 80, by=20))),
+              par.settings = ggplot2like(),
+              axis = axis.grid)
> gridExtra::grid.arrange(p1, p2)
```


Fecal fat (g/day)
60
40
20
0
Tablet
Capsule
Coated
None
Pill type


Predicted fecal fat (g/day)
60
40
20
0
Tablet
Capsule
Coated
None
Pill type

Mesures répétées et intercept aléatoire

Pour un plan d'expérience équilibré, la variance résiduelle d'une ANOVA à mesures répétées sera identique à celle d'un modèle à intercept aléatoire (l'estimateur REML est équivalent aux CM estimés dans l'ANOVA).

Pour tester la significativité des effets fixes, on peut utiliser des tests F (ANOVA), du bootstrap[2] ou des tests du rapport de vraisemblance (LRT). Dans le dernier cas, il est nécessaire d'estimer les modèles par maximum de vraisemblance.

```
> m <- update(m, method = "ML")
> m0 <- update(m, fixed = . ~ - pilltype)
> anova(m, m0)
```

```
   Model df AIC BIC logLik   Test L.Ratio p-value
m      1  6 202 209    -95
m0     2  3 211 214   -102 1 vs 2    14.6  0.0022
```

Structures de variance-covariance alternatives

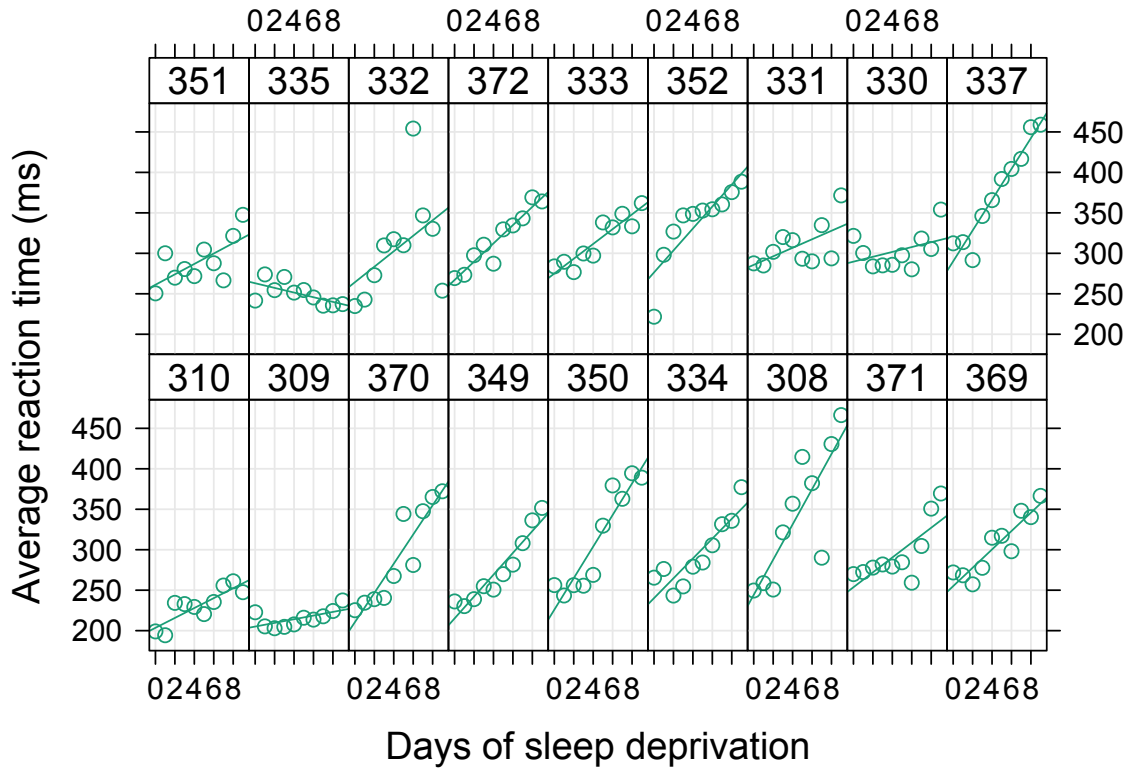
Il est possible de spécifier d'autres matrices de variance-covariance, selon le type de design[9] : non structurée, autorégressive (ordre 1), bande diagonale, AR(1) avec variance hétérogène.

Le modèle à intercept aléatoire permet de contraindre la matrice VC. Dans le cas des ANOVA à mesures répétées, la stratégie consiste à appliquer une correction de Greenhouse-Geisser ou Huynh-Feldt pour corriger les violations à l'hypothèse de symétrie composée, ou alors d'utiliser une MANOVA (moins puissante)[1,13].

Cependant, les modèles à effets alétoires ou mixtes restent plus souples d'utilisation et permettent de former des inférences sur la structure VC et de réaliser des comparaisons de modèles.

Données d'illustration

Average reaction time per day for subjects in a sleep deprivation study. On day 0 the subjects had their normal amount of sleep. Starting that night they were restricted to 3 hours of sleep per night. The observations represent the average reaction time on a series of tests given each day to each subject. D. Bates, Lausanne 2009, http://bit.ly/Kj8VVj].



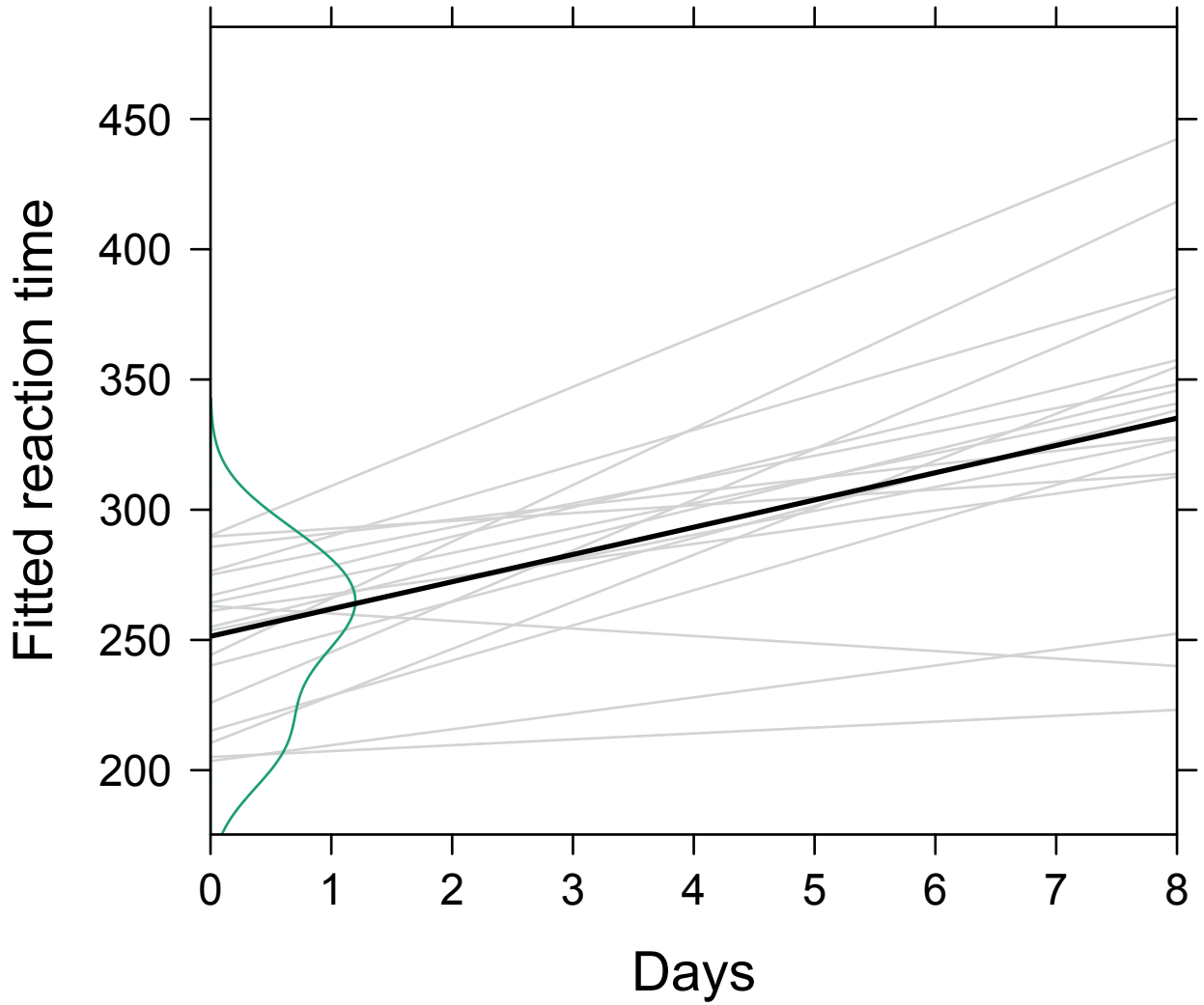
Modèle OLS classique

```
> library(lme4)
> data(sleepstudy)
> reg.subj <- lmList(Reaction ~ Days | Subject, sleepstudy)
> reg.subj.df <- data.frame(lapply(reg.subj, coef))
> apply(reg.subj.df, 1, quantile, prob = c(.25, .75))
```

```
    (Intercept)  Days
25%         229  6.19
75%         273 13.55
```

```
> coef(lm(Reaction ~ Days, data = sleepstudy))
```

```
(Intercept)        Days
      251.4        10.5
```


Fitted reaction time
450
400
350
300
250
200
0
1
2
3
4
5
6
7
8
Days

Modèles plausibles

Voici une liste de modèles incorporant un ou plusieurs effets aléatoires susceptibles de rendre compte de la corrélation des mesures intra-unité :

1. Modèle à intercept aléatoire : Reaction Days + (1 | Subject)
2. Modèle à intercept et pente aléatoires : Reaction Days + (Days | Subject)
3. Modèle à effets aléatoires non corrélés : Reaction Days + (1 | Subject) + (0 + Days | Subject)

Utilisation de la fonction `lme4::lmer` :

```
> anova(m1, m2, m3)
```

```
Data: sleepstudy
Models:
m1: Reaction ~ Days + (1 | Subject)
m3: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
m2: Reaction ~ Days + (Days | Subject)
   Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>Chisq)
m1  4 1802 1815   -897     1794
m3  5 1762 1778   -876     1752 42.08      1    8.8e-11 ***
m2  6 1764 1783   -876     1752  0.06      1         0.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prédictions

Dans un modèle à effets aléatoires, les coefficients de régression associés aux termes aléatoires (dont l'espérance mathématique vaut 0) ne sont plus des paramètres et ne peuvent être estimés comme dans le cas du modèle OLS..

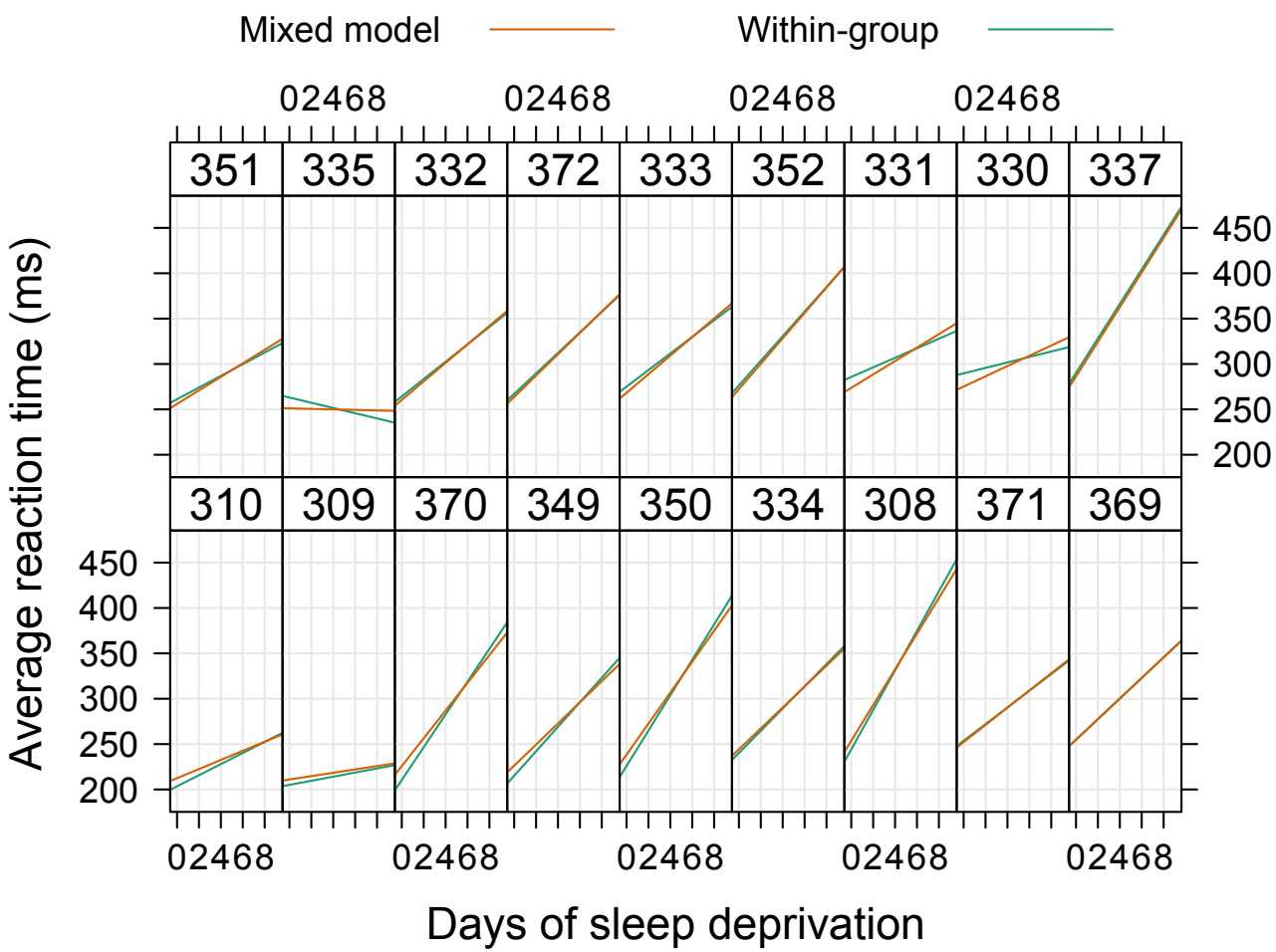
Cependant, on peut utiliser la distribution *a posteriori* (approche bayésienne). En combinant les modes conditionnels des effets aléatoires et les paramètres estimés pour les effets fixes, on obtient les modes conditionnels des coefficients intra-sujet.

On peut vérifier que les prédictions individuelles sont reliées, que le facteur sujet soit traité comme fixe ou aléatoire.

```
> m4 <- aov(Reaction ~ Days + Subject, data = sleepstudy)
> feff <- model.tables(m4, cterms="Subject")[[1]]$Subject
> as.numeric(unlist(ranef(m1)$Subject / feff)[1])
```

```
[1] 0.935
```

$$\tilde{y}_i = (\underbrace{\hat{\beta}_0}_{\text{Fixed}} + \underbrace{\hat{u}_{0i}}_{\text{Random}}) + (\hat{\beta}_1 + \hat{u}_{1i})x$$



Autour du concept de shrinkage

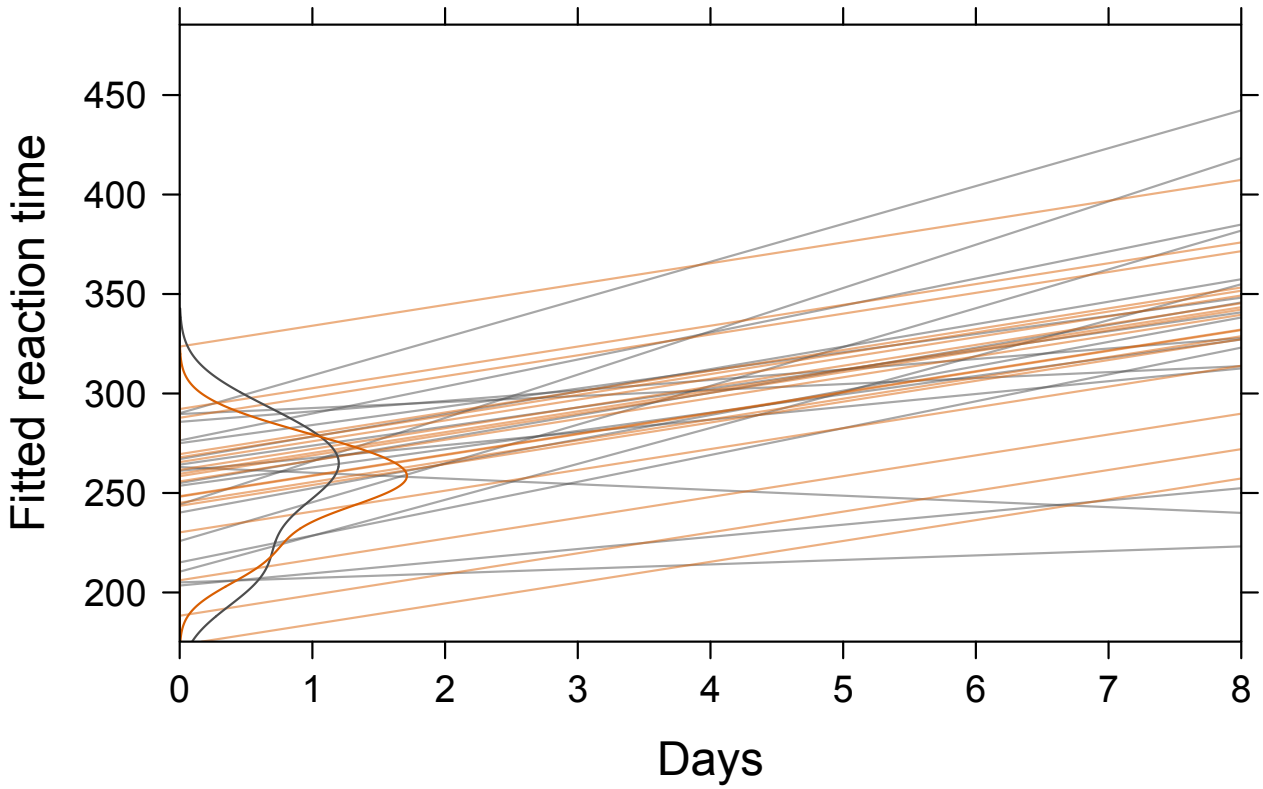
Les valeurs prédites à partir d'un modèle à effet aléatoire sont des estimations biaisées vers la tendance moyenne.

Dans les cas les plus simples, le coefficient de shrinkage revient à

$$\tau = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2 / n_i},$$

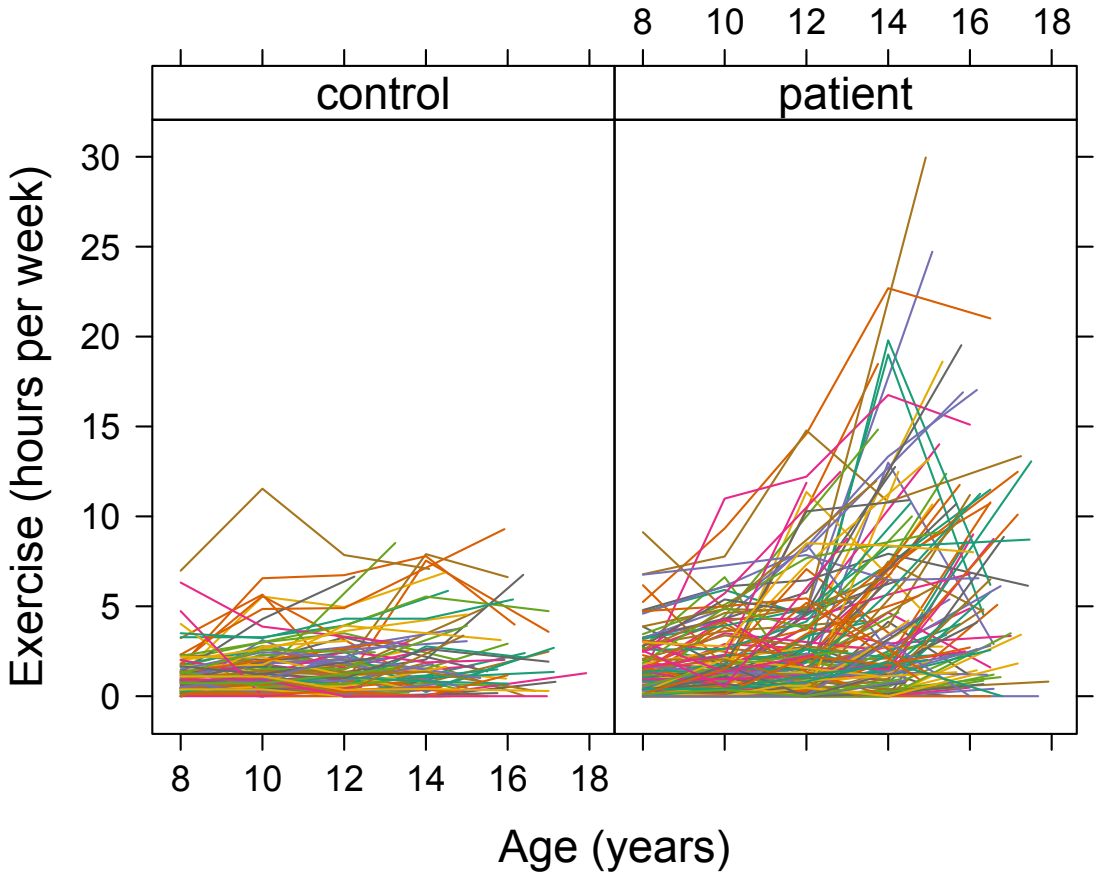
où n_i est la taille du i^{e} cluster.

Ici, $\tau = 37.1^2/(37.1^2 + 31.0^2/10) = 0.935$. Dans le cas où les mesures sont précises, lorsque les unités sont très différentes ou lorsque l'échantillon est grand il y aura généralement peu de shrinkage.


Fitted reaction time
450
400
350
300
250
200
0
1
2
3
4
5
6
7
8
Days

Prédicteurs continus et catégoriels

Blackmoor and Davis's data on exercise histories of 138 teenaged girls hospitalized for eating disorders and 98 control subjects. John Fox[4]



Modélisation d'un terme d'interaction

Construction de trois modèles emboîtés :

```
> data(Blackmore, package="car")
> lex <- log(Blackmore$exercise + 5/60, 2)
> m0 <- lme(lex ~ I(age-8)*group,
+           random= ~ I(age-8) | subject,
+           data=Blackmore)
> m1 <- update(m0, random= ~ 1 | subject)
> m2 <- update(m0, random= ~ I(age-8) - 1 | subject)
```

```
> anova(m0, m1)
```

```
   Model df  AIC  BIC logLik   Test L.Ratio p-value
m0     1  8 3630 3669  -1807
m1     2  6 3644 3673  -1816 1 vs 2    18.1   1e-04
```

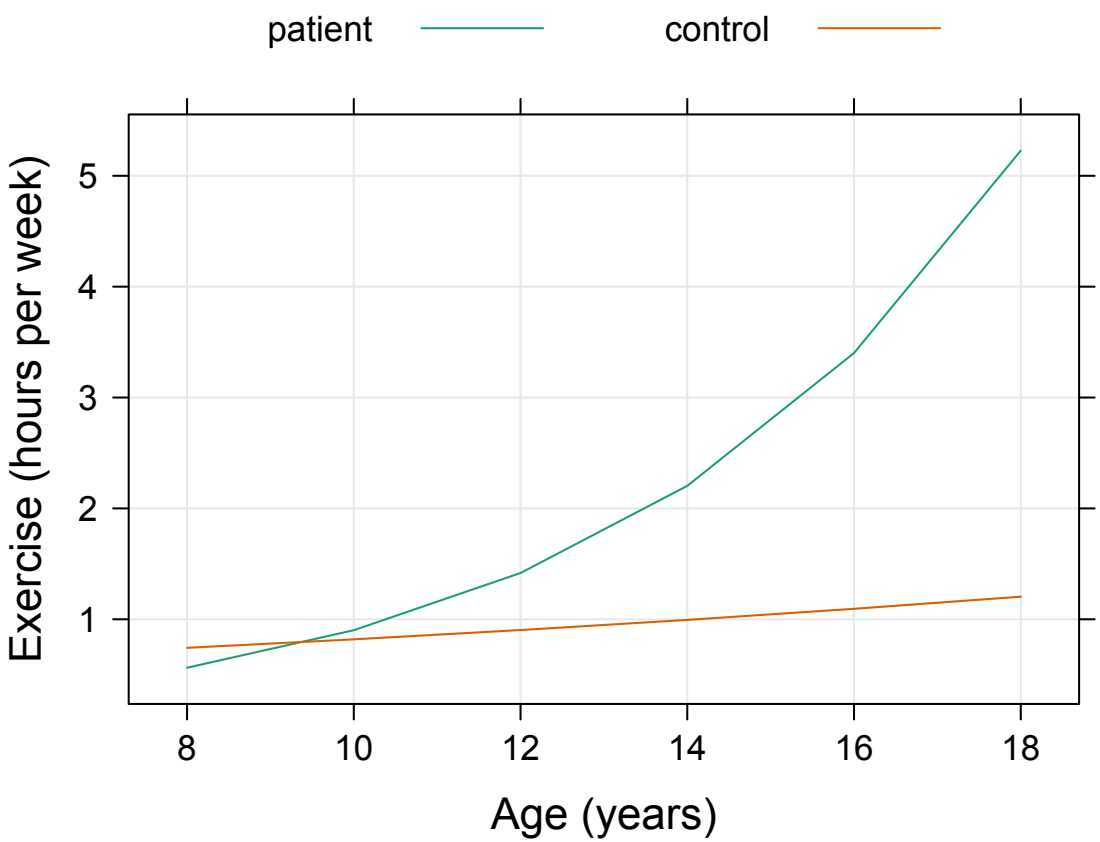
```
> anova(m0, m2)
```

```
   Model df  AIC  BIC logLik   Test L.Ratio p-value
m0     1  8 3630 3669  -1807
m2     2  6 3834 3863  -1911 1 vs 2     208  <.0001
```

```
> summary(m0)
Linear mixed-effects model fit by REML
 Data: Blackmore
   AIC  BIC logLik
  3630 3669  -1807

Random effects:
 Formula: ~I(age - 8) | subject
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev Corr
(Intercept) 1.444  (Intr)
I(age - 8)  0.165  -0.281
Residual    1.244

Fixed effects: lex ~ I(age - 8) * group
                        Value Std.Error  DF t-value p-value
(Intercept)            -0.276    0.1824 712   -1.51  0.1306
I(age - 8)              0.064    0.0314 712    2.04  0.0416
grouppatient           -0.354    0.2353 229   -1.50  0.1338
I(age - 8):grouppatient 0.240    0.0394 712    6.09  0.0000
```


patient
control
Exercise (hours per week)
1
2
3
4
5
8
10
12
14
16
18
Age (years)

Références I

1. H Abdi. The greenhouse-geisser correction. In N Salkind, editor, *Encyclopedia of Research Design*. Thousand Oaks, CA : Sage, 2010.
2. JJ Faraway. *Extending the linear model with R*. Chapman & Hall/CRC, 2006.
3. GM Fitzmaurice. *Longitudinal Data Analysis*. CRC Press, 2009.
4. J Fox. Linear mixed models. App. to An R and S-PLUS Companion to Applied Regression, May 2002.
5. A Gelman. Analysis of variance–why it is more important than ever. *Annals of Statistics*, 33(1) :1–53, 2005.
6. A Gelman and J Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.
7. JK Lindsey. *Models for Repeated Measurements*. Oxford University Press, 2nd edition, 1999.
8. CE McCulloch and SR Searle. *Generalized, Linear, and Mixed Models*. Wiley, 2001.
9. JC Pinheiro and DM Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.
10. SW Raudenbush and AS Bryk. *Hierarchical Linear Models : Applications and Data Analysis Methods*. Thousand Oaks CA : Sage, 2nd edition, 2002.
11. Student. The probable error of a mean. *Biometrika*, 6(1) :1–25, 1908.

Références II

12. E Vittinghoff, DV Glidden, SC Shiboski, and McCulloch. *Regression Methods in Biostatistics. Linear, Logistic, Survival, and Repeated Measures Models*. Springer, 2005.
13. JH Zar. *Biostatistical Analysis*. Pearson, Prentice Hall, 4th edition, 1998.

Index des commandes