Modèle linéaire et applications

Introduction à R pour la recherche biomédicale

http://www.aliquote.org/cours/2012_biomed

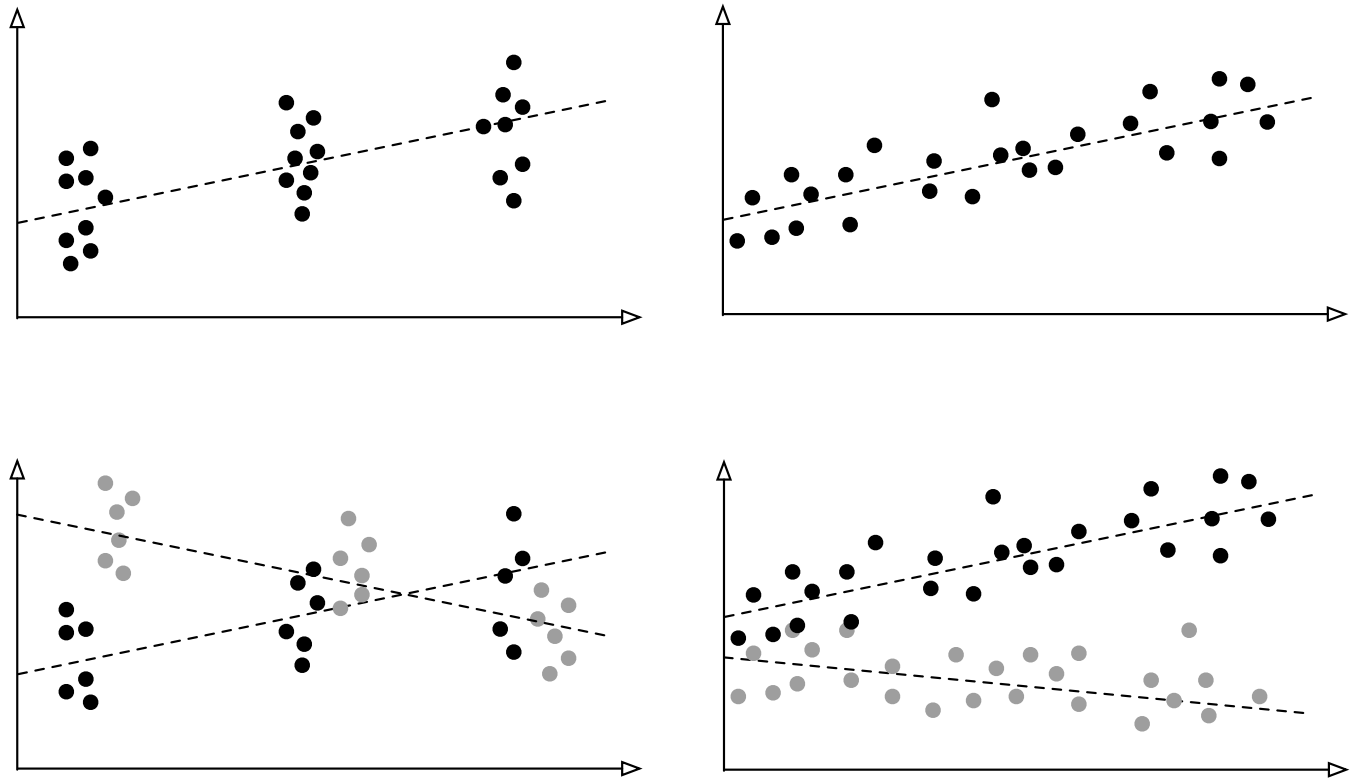
Objectifs

- Dans ce cours, on s'intéressera à l'approche de modélisation par régression linéaire, dans un cadre uni- et multivarié, en montrant le lien avec l'analyse de variance et de covariance.
- Le cas des données corrélées (mesures répétées) sera traité séparément.
- On insistera en particulier sur l'estimation ponctuelle et par intervalles des paramètres des modèles et les procédures de diagnostic des modèles.

Lecture conseillée : Vittinghoff, Glidden, Shiboski, & McCulloch (2005)
(illustrations avec Stata, code R `http://www.aliquote.org/articles/tech/RMB/`)

Un même objectif

Expliquer les variations observées au niveau d'une variable réponse ("dépendante") numérique, y, en fonction de variables prédictrices ("indépendantes"), x_j, pouvant être de nature qualitative ou quantitative.

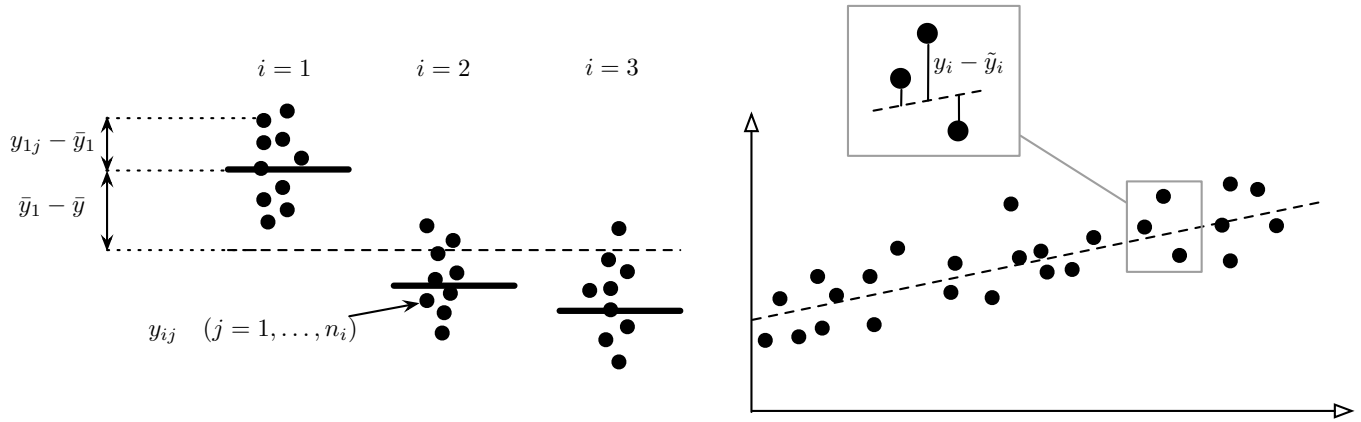


Illustration

L'idée revient toujours à considérer qu'il existe une part systématique et une part aléatoire (résidus) dans ces variations. Le modèle linéaire permet de formaliser la relation entre y et les x_j, en séparant ces deux sources afin d'estimer la contribution relative des x_j dans les fluctuations de y.

$$\text{réponse} = \text{effet prédicteur} + \underbrace{\text{bruit}}_{\text{erreur de mesure, temps d'observation, etc.}}$$

sachant que le modèle théorique relie fonctionnellement la réponse au(x) prédicteur(s) de manière additive : $\mathbb{E}(y \mid x) = f(x; \beta)$.

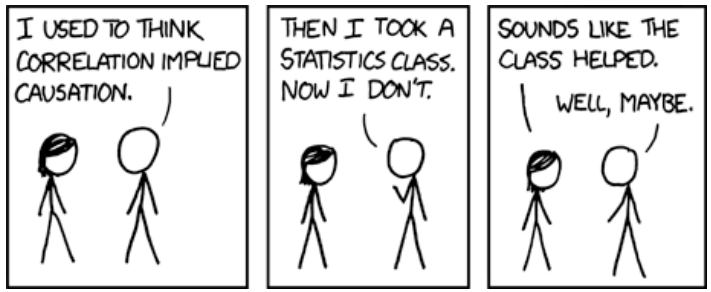


Régression linéaire simple

La régression linéaire permet de modéliser la relation linéaire entre une variable réponse continue et un prédicteur d'intérêt.
Contrairement à l'approche corrélationnelle, dans ce cas les deux variables jouent un rôle asymétrique : la variable explicative ou prédictrice (x) est supposée expliquer une partie des variations observées au niveau de la variable réponse y. On peut vouloir quantifier cette part de variance expliquée, ou encore estimer la contribution de x dans les variations de y.
La linéarité de la relation et l'influence des observations sont deux aspects critiques de la validité des résultats.

Alternatives possibles :

- méthodes résistantes (Huber) ou robustes (LAD, quantile)
- splines ou restricted cubic splines



En détails

Le modèle de régression simple

Soit y_i la réponse observée sur l'individu i, et x_i sa valeur observée pour le prédicteur x. Le modèle de régression linéaire s'écrit

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

où β_0 représente l'ordonnée à l'origine (*intercept*) et β_1 la pente (*slope*) de la droite de régression, et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ est un terme d'erreur (résidus, supposés indépendants).
En minimisant les différences quadratiques entre les valeurs observées et les valeurs prédites (principe des MCO), on peut estimer les coefficients de régression, $\hat{\beta}_0$ et $\hat{\beta}_1$:

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \sum (y_i - \bar{y})(x_i - \bar{x}) / \sum (x_i - \bar{x})^2\end{aligned}$$

Sous H_0, le rapport entre l'estimé de la pente ($\hat{\beta}_1$, de variance $\frac{\text{SSR}/(n-2)}{(n-1)s_x^2}$) et son erreur standard suit une loi de Student à $(n-2)$ degrés de liberté.

Illustration

On simule 10 paires d'observations indépendantes, liées par la relation $y = 5.1 + 1.8 \times x$, à laquelle on rajoute des aléas gaussiens N(0,1).

```
n <- 10
x <- runif(n, 0, 10)
y <- 5.1 + 1.8 * x + rnorm(n)
summary(lm(y ~ x))
```

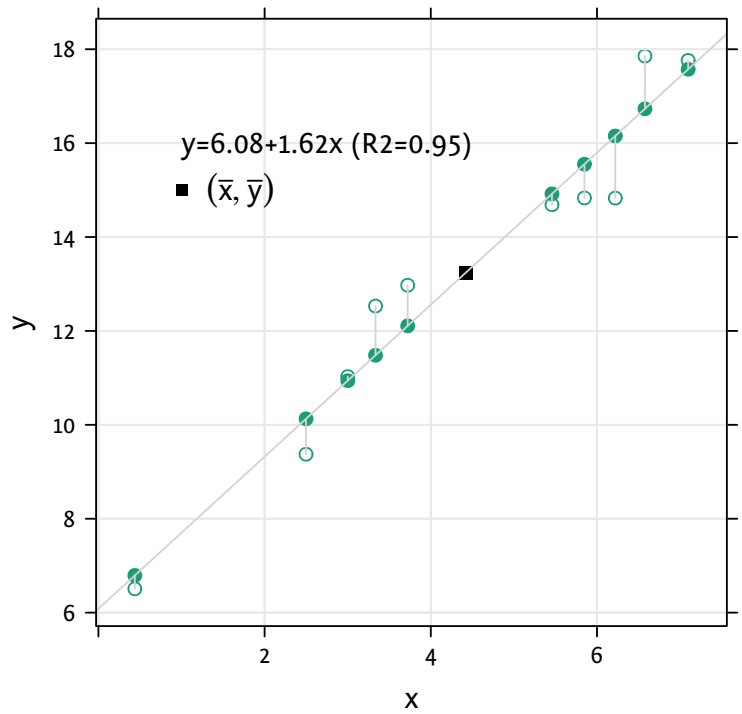
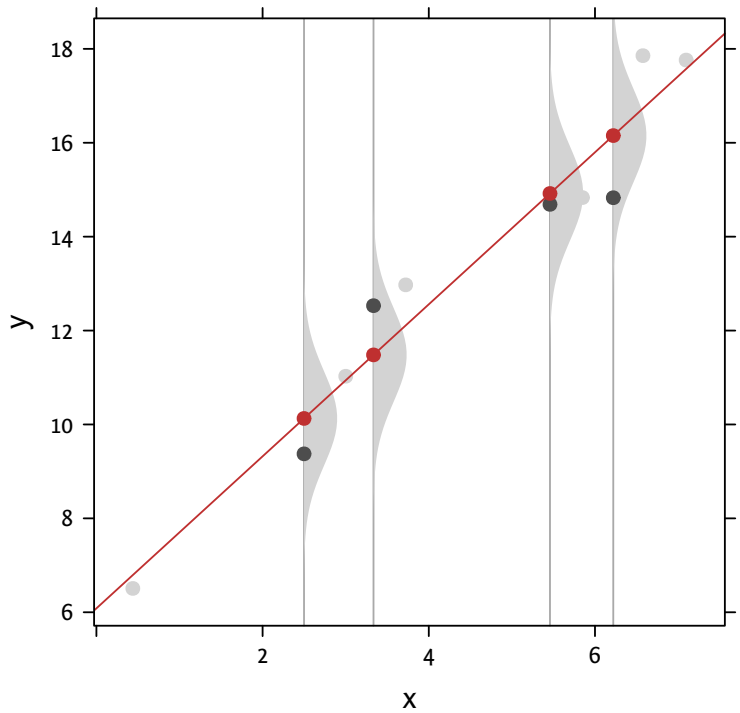


Illustration (2)

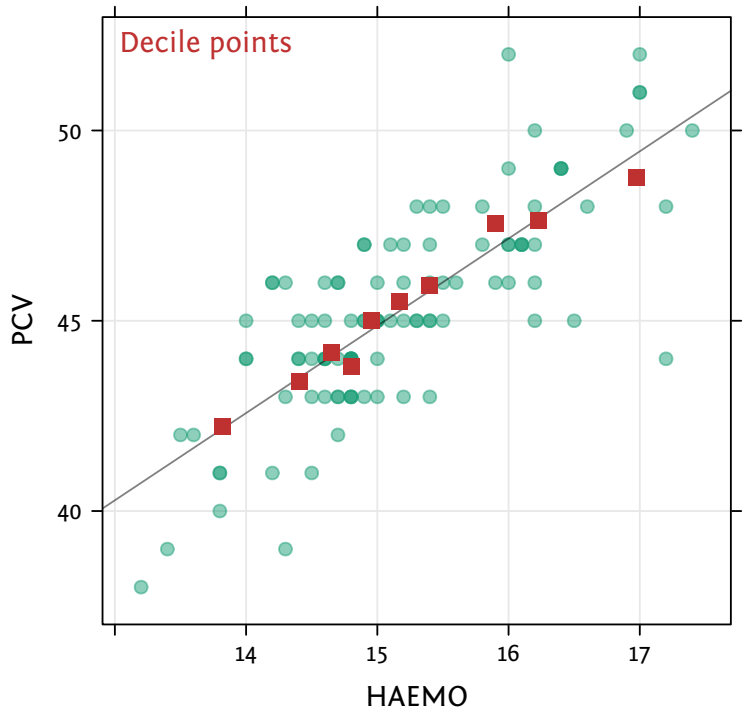
Les valeurs prédites $\tilde{y}$ sont estimées conditionnellement aux valeurs prises par x. L'hypothèse de normalité ($\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$) n'est pas nécessaire pour estimer la pente par MCO, mais seulement pour l'inférence sur les paramètres du modèle.



Application

Health survey of paint sprayers. (Everitt & Rabe-Hesketh, 2001, p. 158)

```
paint <- read.table("PAINT.DAT", header=TRUE)
xyplot(PCV ~ HAEMO, data=paint, type=c("p","r"))
```



Estimation des paramètres du modèle de régression

```
lm.fit <- lm(PCV ~ HAEMO, data=paint)
summary(lm.fit)
confint(lm.fit)
anova(lm.fit)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.4785     2.7982   3.745 0.000301 ***
HAEMO         2.2926     0.1842  12.449  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.688 on 101 degrees of freedom
Multiple R-squared: 0.6054, Adjusted R-squared: 0.6015
F-statistic:   155 on 1 and 101 DF,  p-value: < 2.2e-16
```

Le test pour $H_0 : \beta_1 = 0$ est significatif ($p < 0.001$). Lorsque la concentration en hémoglobine (HAEMO) varie de une unité, le taux hématocrite varie de 2.3 %.

Diagnostic du modèle

L'essentiel de l'activité de diagnostic du modèle (mauvaise spécification et points influents) repose sur l'analyse des résidus, ceux-ci étant accessibles à partir de la commande `resid`.

```
xyplot(resid(lm.fit) ~ HAEMO, data=paint)
xyplot(resid(lm.fit) ~ fitted(lm.fit))
```

Avec la commande `influence`, on dispose également d'un ensemble d'indices statistiques permettant d'évaluer la qualité d'ajustement du modèle ; voir aussi `help(influence.measures)`.

```
influence.measures(lm.fit)
```

```
      dfb.1_  dfb.HAEM     dffit cov.r   cook.d     hat inf
1  -0.272246  0.264677 -0.297215 1.033 4.38e-02 0.04691
2   0.083548 -0.076456  0.143625 1.004 1.03e-02 0.01355
3   0.066030 -0.064081  0.072872 1.063 2.68e-03 0.04282   *
4   0.016322 -0.012385  0.067743 1.021 2.31e-03 0.01004
5   0.002010 -0.001839  0.003455 1.034 6.03e-06 0.01355
6   0.115169 -0.110294  0.139426 1.032 9.75e-03 0.02594
```

Observations extrêmes et influentes

On peut distinguer différents types de points influents :

- "*outlier*" univarié (x ou y)
- point présentant un effet levier ("leverage") : valeur extrêmes sur x
- point ayant un large résidu (en valeur absolue) : valeur extrêmes sur y
- point exerçant un effet sur les coefficients de régression (sensibilité)

La fonction `resid` permet d'obtenir les résidus "simples", $e_i = y_i - \hat{y}_i$, de variance non-constante, mais on montre que celle-ci est fonction de σ_ε^2 :

$$\text{Var}(e_i) = \sigma_\varepsilon^2(1 - h_i),$$

où h_i ("*hat value*") est l'effet levier de la $i$ème observation sur l'ensemble des valeurs ajustées via le modèle de régression.
Les points exerçant un fort effet levier tendent donc à avoir de plus faibles résidus (puisqu'ils "tirent" la droite de régression vers eux).

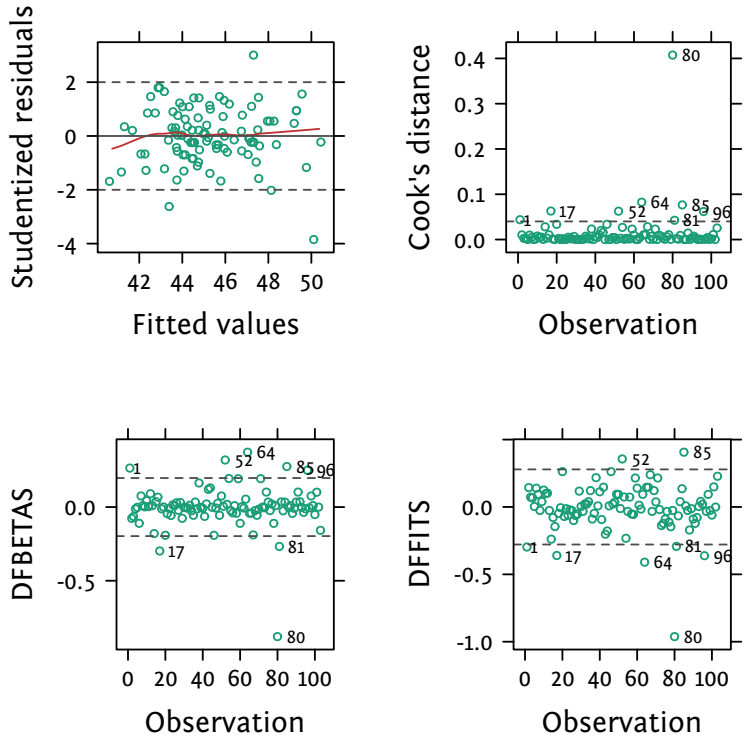
Mesures d'influence et critères d'interprétation

- effet levier : large si $h_i > 2(k+1)/n$
- résidus standardisés/studentisés : $e'_i = e_i/s_e\sqrt{1-h_i}$ et $e^*_i = e_i/s_{e(-i)}\sqrt{1-h_i}$, où $s_{e(-i)}$ est une estimation de σ_ε sans l'observation i. Comme $e^*_i \sim t(n-k-2)$, on s'attend à ce que 5 % des valeurs au plus soient situées en dehors de l'intervalle [-2;+2].
- distance de Cook : $D_i = \frac{h_i}{1-h_i} \times \frac{e'^2_i}{k+1}$, considérée large si $D_i > 4/(n-k-1)$.
- DFFITS$= \frac{\hat{y}_i - \hat{y}_i^{(-i)}}{s_{e(-i)}}\sqrt{h_i^{(-i)}}$, considéré large si $> 2\sqrt{p/n}$.
- DFBETAS$= \frac{\hat{\beta}_j - \hat{\beta}_j^{(-i)}}{s_{e(-i)}\sqrt{X'X_{jj}}}$, pour le $j$ème coefficient du modèle (incluant l'intercept), considéré large si $> 2/\sqrt{n}$.

La plupart de ces indices sont estimés en enlevant l'observation en question, et donc permettent d'apprécier son poids dans la qualité de la prédiction ou la stabilité des coefficients de régression. Voir `http://bit.ly/JdaJM3`.

Illustration

Les graphiques suivants visent essentiellement à mettre en évidence de potentielles valeurs extrêmes ou influentes (“*outliers*”). Ils reposent tous sur les données extraites de `influence.measures`. Voir aussi le package car (Fox, 2011) pour d’autres fonctionnalités graphiques.



Prédiction ponctuelle et par intervalle

Les valeurs prédites sont obtenues avec `fitted` ou `predict` (pour de nouvelles observations). Par exemple,

```
fitted(lm.fit)
predict(lm.fit, data.frame(HAEMO=seq(13, 18, by=1)))
```

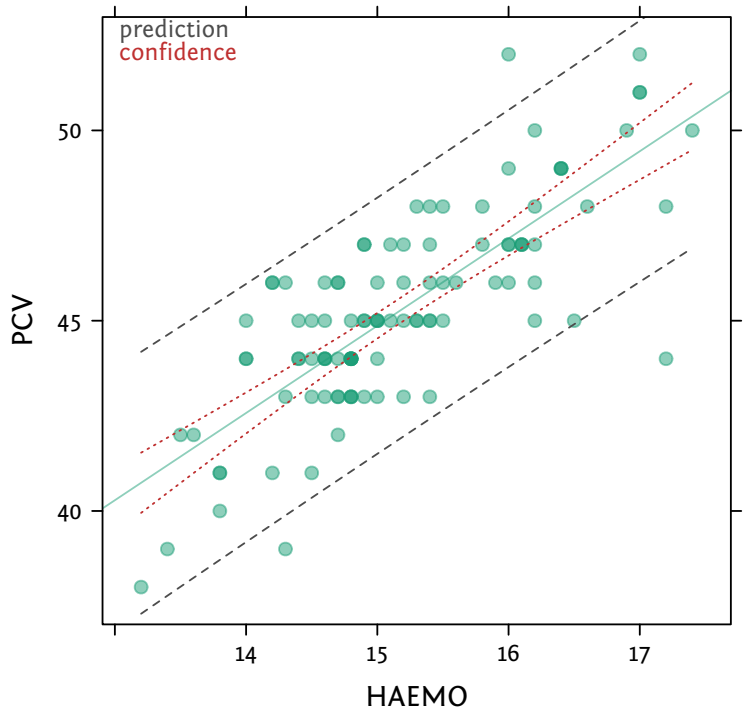
Il existe deux types de prédictions par intervalle (95 %), de la forme

$$\hat{y} \pm t_{n-p-1,1-\alpha/2}\mathsf{SE}.$$

- Prédiction de y : `predict(..., interval="prediction")`
  $\mathsf{SE}(\hat{y}) = s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$
- Prédiction de $\mathbb{E}(y \mid x)$: `predict(..., interval="confidence")`
  $\mathsf{SE}(\mathbb{E}(y \mid x)) = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

Illustration

Les intervalles de confiance de type prédiction sont plus larges que les intervalles associés à la prédiction de valeurs moyennes. Dans les deux cas, leur demi-largeur est toujours plus petite autour du point moyen $(\bar{x}, \bar{y})$.



Voir aussi `xYplot` dans le package `Hmisc`.

Régression vs. ANOVA

La matrice de dessin (*design*) utilisée dans les deux cas est identique :

```
x <- gl(5, 1, 10, labels=letters[1:5])
model.matrix(rnorm(10) ~ x)
```

$\beta_0 + \beta_1 \mathbb{I}(x = b) + \cdots + \beta_4 \mathbb{I}(x = e)$

	Int.	xb	xc	xd	xe
1	1	0	0	0	0
2	1	1	0	0	0
3	1	0	1	0	0
4	1	0	0	1	0
5	1	0	0	0	1
6	1	0	0	0	0
7	1	1	0	0	0
8	1	0	1	0	0
9	1	0	0	1	0
10	1	0	0	0	1

Le modèle s'écrit, pour chaque observation,

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_i$$

Soit les valeurs prédites,

$$\begin{aligned}
\tilde{y}_. &= b_0 + b_1 \times 0 + b_2 \times 0 + b_3 \times 0 + b_4 \times 0 \quad (x = a) \\
\tilde{y}_. &= b_0 + b_1 \times 1 + b_2 \times 0 + b_3 \times 0 + b_4 \times 0 \quad (x = b) \\
&\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
\tilde{y}_. &= b_0 + b_1 \times 0 + b_2 \times 0 + b_3 \times 0 + b_4 \times 1 \quad (x = e)
\end{aligned}$$

c'est-à-dire les moyennes de groupes en ANOVA.

Application

Supposons que seules 10 classes soient disponibles pour la variable HAEMO.

```
haemo.dec <- cut2(paint$HAEMO, g=10)
fm <- PCV ~ haemo.dec
summary(aov.fit <- aov(fm, data=paint))
summary(lm.fit <- lm(fm, data=paint))
.Last.value$sigma^2
grp.means <- tapply(paint$PCV,
                    haemo.dec,
                    mean)
grp.means[2:10] - grp.means[1]
coef(lm.fit)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
haemo.dec     9  416.7   46.30   13.77 8.14e-14 ***
Residuals    93  312.7    3.36
```

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          42.2308     0.5086  83.036  < 2e-16 ***
haemo.dec[14.3,14.6)  1.1692     0.7713   1.516 0.132932
haemo.dec[14.6,14.8)  1.9359     0.7341   2.637 0.009798 **
(...)
haemo.dec[16.5,17.4]  6.5470     0.7952   8.234 1.10e-12 ***

Residual standard error: 1.834 on 93 degrees of freedom
Multiple R-squared: 0.5713, Adjusted R-squared: 0.5298
F-statistic: 13.77 on 9 and 93 DF,  p-value: 8.141e-14
```

La manière dont R code les contrastes (voir `options("contrasts")`) est importante !

```
contr.treatment(10)
contr.sum(10)
contr.helmert(10)
```

Usage des formules sous R

R suit les conventions de notation proposées par Wilkinson & Rogers (1973), discutées dans Chambers & Hastie (1992), pp. 24–30.
Notons `x`, `y`, ... des variables continues, `a`, `b`, ... des variables catégorielles. On utilise "$\sim$" pour dénoter une relation fonctionnelle entre une réponse, y, et

- x, régression linéaire simple
- `x + 0` ou `x -1`, idem avec suppression de l'intercept
- `a + b`, deux effets principaux
- `a * b`, équivalent à `1 + a + b + a:b`, idem avec interaction

Dans le cas des interactions, on a les équivalences suivantes :

- `factor:factor` $\leftrightarrow \gamma_{ij}$ (croisement des niveaux des facteurs)
- `factor:numeric` $\leftrightarrow \beta_j x$ (pente variable selon les niveaux du facteur)
- `numeric:numeric` $\leftrightarrow \beta xz$ (produit élément par élément)

Usage des formules sous R (2)

- `a + b + a:d`, effets principaux pour A et B, et interaction AxD
- `a * b * d - a:b:d`, tous les effets sauf l'interaction AxBxD (inclut les interactions AxB, AxD et BxD)
- `a / b`, équivalent à `1 + a + b %in% a` (relation d'emboitement)
- `a / (b * d)`, tous les termes BxD pour chaque niveau de A
- `.^2`, tous les effets principaux et les interactions de second-ordre

Il est également possible de mettre à jour un modèle existant (en ajoutant ou supprimant des termes) :

```
fm <- y ~ x + a * b
mod1 <- lm(fm, data=dat)
update(mod1, . ~ . - a:b)  # supprime interaction AxB
```

Analyse de variance à deux facteurs

Dans l'analyse de variance à deux facteurs considérés comme des effets fixes, on considère deux variables explicatives catégorielles et une variable réponse continue. À la différence de l'ANOVA à un facteur, l'inclusion d'un second facteur pose la question de la co-relation entre les deux facteurs dans la prédiction de la variable réponse. Cet effet d'interaction peut être ou non l'objet de l'étude, mais le plus souvent il est important de le quantifier et de le tester afin de pouvoir discuter les effets principaux des facteurs d'intérêt.

Traditionnellement, ce type d'analyse se retrouve dans les plans d'expérience (psychologie, agronomie, industrie, etc.) mais reste applicable aux études cliniques. Les conditions d'application de l'ANOVA sont les mêmes que dans le cas à un facteur, en particulier celles concernant les résidus.

La régression logistique ordinale (Armstrong & Sloan, 1989), des techniques de permutation (Anderson & Ter Braak, 2003), ou l'analyse *median polish* (Mosteller & Tukey, 1977) complémentent cette approche paramétrique.

En détails

Le modèle d'ANOVA à deux facteurs

Soit y_{ijk} la $k$ème observation pour le niveau i du facteur A ($i = 1, \ldots, a$) et le niveau j du facteur B ($j = 1, \ldots, b$). Le modèle complet avec interaction s'écrit

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

où μ désigne la moyenne générale, α_i (β_j) l'écart à la moyenne des moyennes de groupe pour le facteur A (B), γ_{ij} les écarts à la moyenne des moyennes pour les traitements $A \times B$, et $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ la résiduelle. Les effets α_i et β_j sont appelés effets principaux, tandis que γ_{ij} est l'effet d'interaction. Les hypothèses nulles associées sont

$$\begin{cases} H_0^A : \alpha_1 = \alpha_2 = \cdots = \alpha_a, & (a-1)\,\text{dl} \\ H_0^B : \beta_1 = \beta_2 = \cdots = \beta_b, & (b-1)\,\text{dl} \\ H_0^{AB} : \gamma_{11} = \gamma_{13} = \cdots = \gamma_{ab}, & (a-1)(b-1)\,\text{dl} \end{cases}$$

Des tests F (CM effets / CM résiduelle) permettent de tester ces hypothèses.

Les SC de type I sont estimées pour A, puis B, et enfin $A \times B$, à la différence des SC de type II/III, mais elles sont égales dans le cas d'un plan complet équilibré.

Application

The effect of Vitamin C on tooth growth in Guinea Pigs. (Bliss, 1952)

Quelques statistiques descriptives (hors `aggregate`) :

```
data(ToothGrowth)
ToothGrowth$dose <- factor(ToothGrowth$dose)
fm <- len ~ supp * dose
replications(fm, data=ToothGrowth)
library(Hmisc)
f <- function(x) apply(x, 2, function(x)
                       c(mean=mean(x), sd=sd(x)))
summary(fm, data=ToothGrowth, fun=f)
```

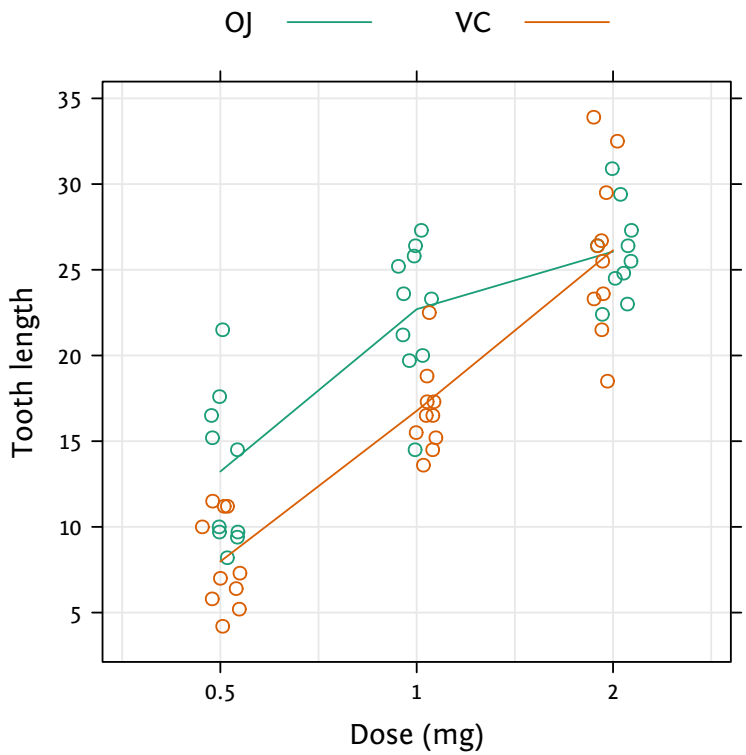
Notons que nous avons délibéremment ignoré le fait que le facteur dose est ordonné. Autre méthode de calcul des moyennes conditionnelles et marginales (ajouter `margins=TRUE`) :

```
library(reshape2)
m <- acast(ToothGrowth, supp ~ dose, mean, value.var="len")
```

Graphique d'interaction

L'inspection visuelle de l'éventuel effet d'interaction peut se faire avec un simple diagramme en lignes. Le non-parallélisme des deux droites VC et OJ suggère que l'effet dose n'est pas le même selon le type de supplément.

```
xyplot(len ~ dose, data=ToothGrowth, groups=supp,
       type=c("p","a"))
```



Analyse du modèle complet

```
aov.fit <- aov(fm, data=ToothGrowth)
summary(aov.fit)
model.tables(aov.fit, type="means", se=TRUE,
             cterms="supp:dose")
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
supp          1  205.4   205.4  15.572 0.000231 ***
dose          2 2426.4  1213.2  92.000  < 2e-16 ***
supp:dose     2  108.3    54.2   4.107 0.021860 *
Residuals    54  712.1    13.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Le test F correspondant à l'interaction supp:dose est significatif à 5 %, suggérant que l'on ne peut se limiter à l'interprétation seule des effets principaux : l'effet dose n'est pas le même selon le type de supplément (VC ou OJ).
Pour quantifier les différences moyennes entre traitements :

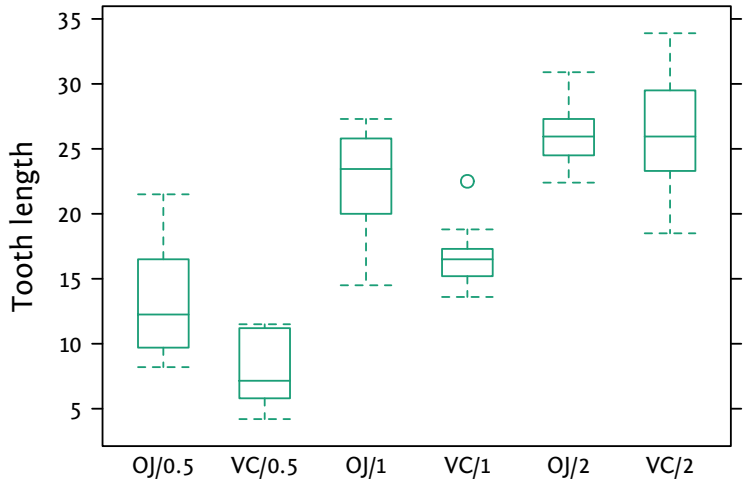
```
apply(m, 2, diff)
```

Voir aussi le package `effects`.

Vérification des conditions d'application

Concernant la normalité, on peut inspecter la distribution des résidus avec un histogramme ou des quantiles. L'homogénéité des variances peut être vérifiée avec des boîtes à moustaches (et un test statistique).

```
qqmath(~ resid(aov.fit))
bwplot(len ~ interaction(supp, dose), data=ToothGrowth)
bartlett.test(len ~ interaction(supp,dose),data=ToothGrowth)
```



Analyse de covariance

L'analyse de covariance consiste à tester différents niveaux d'un facteur en présence d'un ou plusieurs co-facteurs continus. La variable réponse et ces co-facteurs sont supposées reliés, et l'objectif est d'obtenir une estimation des réponses corrigée pour les éventuelles différences entre groupes (au niveau des cofacteurs).

Ce type d'analyse est fréquemment utilisé dans le cas des données pré/post avec des mesures continues et un facteur de groupe, et reste préférable à une simple analyse des scores de différences (Miller & Chapman, 2001; Senn, 2006).

Il existe également des alternatives non-paramétriques (Young & Bowman, 1995), disponibles dans le package `sm` (`sm.ancova`).

En détails

Le modèle d'ANCOVA à un cofacteur

Soit y_{ij} la $j$ème observation dans le groupe i. À l'image du modèle d'ANOVA à un facteur, le modèle d'ANCOVA s'écrit

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij},$$

où β est le coefficient de régression liant la réponse y et le cofacteur x (continu), avec $\bar{x}$ la moyenne générale des x_{ij}, et toujours un terme d'erreur $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. Notons que l'on fait l'hypothèse que β est le même dans chaque groupe. Cette hypothèse de parallélisme peut se vérifier en testant la significativité du terme d'interaction $\alpha\beta$.
La réponse moyenne ajustée pour l'effet du co-facteur numérique s'obtient simplement comme $\bar{\alpha}_i + \hat{\beta}(\bar{x}_i - \bar{x})$, où $\bar{x}_i$ est la moyenne des x dans le $i$ème groupe.

Application

Weight change data for young female anorexia patients. (Hand, Daly, McConway, & Ostrowski, 1993)

```
data(anorexia)
anorexia$Treat <- relevel(anorexia$Treat, ref="Cont")
anorex.aov0 <- aov(Postwt ~ Prewt + Treat, data=anorexia)
anorex.aov1 <- aov(Postwt ~ Prewt * Treat, data=anorexia)
summary(anorex.aov0)
```

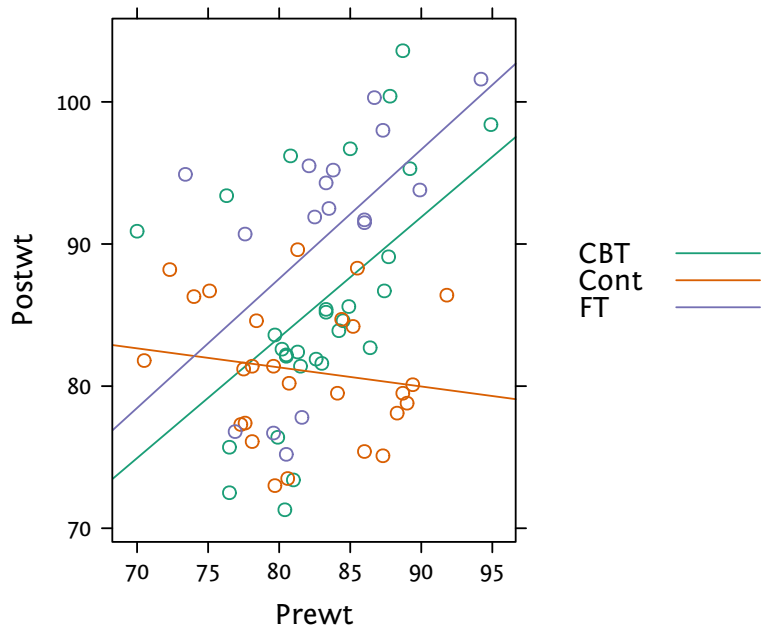
```
            Df Sum Sq Mean Sq F value   Pr(>F)
Prewt        1    507   506.5  10.402 0.001936 **
Treat        2    766   383.1   7.868 0.000844 ***
Residuals   68   3311    48.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En tenant compte du poids initial, le modèle de base (`anorex.aov0`) suggère bien qu'il existe des différences de poids moyens entre les trois groupes. Ce modèle suppose que la relation entre poids avant/après traitement est indépendante du traitement.

Vérification graphique

Un diagramme de dispersion faisant apparaître les différents groupes permet de vérifier visuellement si l'hypothèse de parallélisme est réaliste. Ici, clairement cette hypothèse ne tient pas.

```
xyplot(Postwt ~ Prewt, data=anorexia, groups=Treat,
       aspect="iso", type=c("p","r"))
```



Test de parallélisme

Comparaison avec un modèle incluant l'interaction `Prewt:Treat` :

```
anova(anorex.aov0, anorex.aov1)
```

```
Analysis of Variance Table

Model 1: Postwt ~ Prewt + Treat
Model 2: Postwt ~ Prewt * Treat
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     68 3311.3
2     66 2844.8  2    466.48 5.4112 0.006666 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Le test pour le terme d'interaction indique que celle-ci est significative ; par conséquent, on ne peut pas considérer que la relation entre poids avant et après traitement est la même dans les trois groupes. Il faudra donc décrire les résultats groupe par groupe.

```
summary.lm(anorex.aov1)
```

On peut vouloir plutôt tester $H_0 : \beta = 1$, d'où

```
lm(Postwt ~ Prewt + Treat + offset(Prewt), data=anorexia)
```

Interprétation des coefficients du modèle

Le modèle sans interaction (`coef(anorex.aov0)`) s'écrit

$$\tilde{y}_i = 45.67 + 0.43 \times \text{Prewt}_i + 4.10 \times \mathbb{I}(\text{Treat}_i = \text{CBT}) + 8.66 \times \mathbb{I}(\text{Treat}_i = \text{FT}).$$

Pour les patientes du groupe contrôle, $\tilde{y}_i = 45.67 + 0.43 \times \text{Prewt}_i$, alors que pour celles du groupe FT, $\tilde{y}_i = 45.67 + 0.43 \times \text{Prewt}_i + 8.66$. Ceci correspond bien à l'idée que l'effet de `Prewt` est le même pour toutes les patientes et que le facteur de groupe induit simplement un changement moyen (+4.10 ou +8.66) par rapport au groupe contrôle.

Pour le modèle avec interaction avec `Prewt` centré, on a

$$\begin{aligned} \tilde{y}_i = \quad & 80.99 - 0.13 \times \text{Prewt}_i \\ & + 4.46 \times \mathbb{I}(\text{Treat}_i = \text{CBT}) \\ & + 8.75 \times \mathbb{I}(\text{Treat}_i = \text{FT}) \\ & + 0.98 \times \text{Prewt}_i \times \mathbb{I}(\text{Treat}_i = \text{CBT}) \\ & + 1.04 \times \text{Prewt}_i \times \mathbb{I}(\text{Treat}_i = \text{FT}). \end{aligned}$$

Régression linéaire multiple

L'inclusion de plusieurs prédicteurs conduit à généraliser le modèle de régression simple à la régression multiple et à s'intéresser à l'effet de chaque prédicteur en tenant compte des autres variables dans le modèle : on parlera de l'effet d'un prédicteur à niveau constant des autres prédicteurs. Les coefficients de régression reflète toujours le poids de chaque variable explicative, via le coefficient de corrélation partiel.

La régression multiple pose la question de la sélection des prédicteurs dans le modèle final, et de l'évaluation des effets partiels de chaque prédicteur. L'analyse des résidus du modèle doit tenir compte de l'ensemble des prédicteurs en présence dans le modèle.

Fox (2011) décrit en détails les différentes étapes de construction et de validation d'un modèle de régression multiple. Voir aussi `http://tinyurl.com/carbook`.

En détails

Le modèle de régression multiple

Soit y_i la réponse du $i$ème individu, et $x_{i1}, \ldots, x_{ip}$ ses valeurs observées sur p prédicteurs. Le modèle de régression linéaire (au niveau de ses paramètres) s'écrit

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

où les β_j ($j = 1, \ldots, p$) sont les coefficients de régression qui reflètent le changement observé au niveau de y_i lorsque x_{ij} varie de une unité, les autres prédicteurs étant maintenus constant.
Un tableau d'analyse de variance peut être construit en isolant la variance expliquée par la régression ($\text{CM} = \sum_i (\hat{y}_i - \bar{y})^2/p$) et la résiduelle ($\text{CM} = s^2 = \sum_i (y_i - \hat{y}_i)^2/(n-p-1)$), d'où un test $F(p, n-p-1)$ sur l'égalité de l'ensemble des coefficients de régression (hors intercept),
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p$.
La corrélation R entre les y_i (observés) et les $\hat{y}_i$ (prédits) est appelée coefficient de corrélation multiple, et R^2 reflète la part de variance expliquée par les prédicteurs.
Les coefficients de régression peuvent être testés individuellement en considérant la statistique de test $\hat{\beta}_j / \text{SE}(\hat{\beta}_j) \sim t(n-p)$.

Cas des données corrélées

Tous les modèles discutés précédemment supposaient les observations indépendantes les unes des autres. Dans le cas où la même unité statistique est utilisée pour collecter différentes mesures (mesures répétées), il est important de prendre en compte la corrélation intra-unité induite par ce type de recueil de données. Cela présente l'avantage d'offir un gain de puissance statistique et la possibilité de modéliser la structure de covariance.

L'ANOVA dite "à mesures répétées" permet, en incorporant un effet sujet aléatoire, d'incorporer la corrélation intra-unité et de mieux estimer la résiduelle. On fait une hypothèse de symétrie composée selon laquelle la covariance intra-unité est la même pour toutes les unités statistiques. Des résultats équivalents sont obtenus par un modèle linéaire mixte à intercept aléatoire.

En détails

Le modèle mixte à intercept aléatoire

Un modèle linéaire incorporant un terme aléatoire, u, pour les sujets i sur lesquels on collecte j mesures s'écrit

$$y_{ij} = \mu + \alpha_j + u_i + \varepsilon_{ij},$$

où α_j représente l'effet d'un prédicteur, $u_i \sim \mathcal{N}(0, \tau^2)$ et $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ sont indépendants. Au lieu d'estimer les valeurs u_i comme dans un modèle à effets fixes, c'est la variance de la distribution des u_i (τ^2) qui est estimée. Conditionnellement au prédicteur, la variance totale se décompose comme suit :

$$\text{Var}(y_{ij}) = \tau^2 + \sigma^2,$$

et $\tau^2/(\tau^2 + \sigma^2)$, le coefficient de corrélation intraclasse, reflète la corrélation intra-unité, supposée commune à tous les sujets. Cette structure de corrélation est appelée symétrie composée. L'effet fixe du prédicteur peut être testé par comparaison avec un modèle n'incluant pas celui-ci (LRT). Les prédictions sont formées à partir d'une combinaison des effets fixe et aléatoire.

Application

Blood cholesterol concentrations. (Zar, 1998)

```
chs <- read.table("cholesterol.txt", header=TRUE)
chs$Subject <- factor(chs$Subject)  # important
chs <- melt(chs, id.vars="Subject")
aov1 <- aov(value ~ variable + Error(Subject), data=chs)
summary(aov1)
```

```
Error: Subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  6  18731    3122

Error: Within
          Df Sum Sq Mean Sq F value  Pr(>F)
variable   2 1454.0   727.0   12.55 0.00115 **
Residuals 12  695.3    57.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Le terme d'erreur pour tester l'effet intra-sujet est calculé en enlevant la variabilité inter-sujets (SS=18731, soit 96.4 % de la résiduelle d'un modèle à un facteur). Le résultat significatif du test suggère que le niveau de cholestérol varie bien selon le traitement.

Application (2)

Un modèle à effet aléatoire donnera sensiblement le même résultat, tout en permettant d'estimer la corrélation intraclasse :

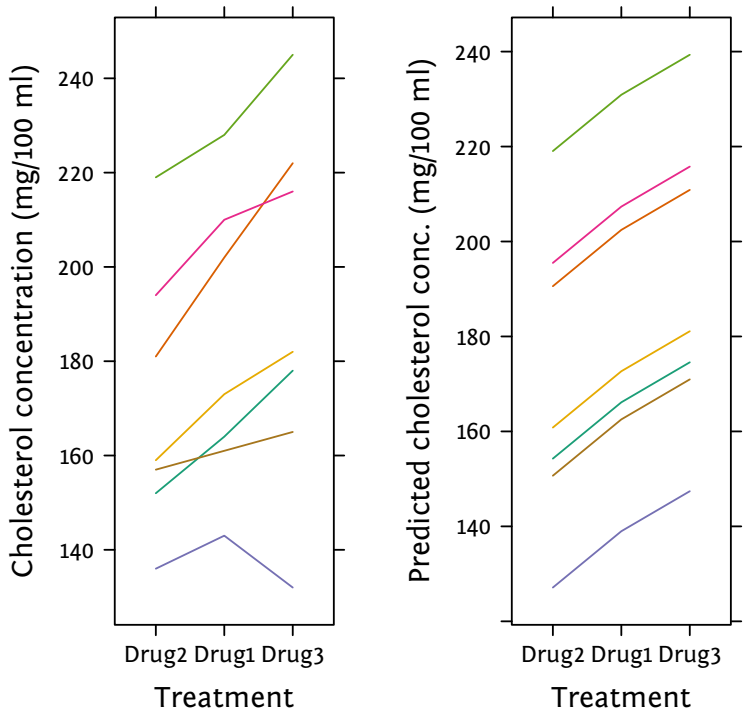
```
library(nlme)
lme1 <- lme(value ~ variable, data=chs,
            random= ~ 1 | Subject)
summary(lme1)
anova(lme1)
31.96^2/(31.96^2+7.61^2)  # ICC
intervals(lme1)
```

```
Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    31.95793 7.612125

Fixed effects: value ~ variable
                   Value Std.Error DF   t-value p-value
(Intercept)   183.00000 12.416889 12 14.737991  0.0000
variableDrug2 -11.85714  4.068852 12 -2.914125  0.0130
variableDrug3   8.42857  4.068852 12  2.071486  0.0605
```

Illustration

On vérifie aisément dans le graphique des valeurs prédites que la corrélation intra-sujet est identique pour tous les sujets et que seuls les niveaux moyens varient entre les sujets.



Index

Bibliographie

Anderson, M. J., & Ter Braak, C. J. F. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation*, *73*, 85–113.

Armstrong, B. G., & Sloan, M. (1989). Ordinal regression models for epidemiological data. *American Journal of Epidemiology*, *129*, 191–204.

Bliss, C. I. (1952). *The Statistics of Bioassay*. Academic Press.

Chambers, J. M., & Hastie, T. J. (Eds.). (1992). *Statistical Models in S*. Wadsworth & Brooks.

Everitt, B., & Rabe-Hesketh, S. (2001). *Analyzing Medical Data Using S-PLUS*. Springer.

Fox, J. (2011). *An R Companion to Applied Regression*. Sage Publications.

Hand, D. J., Daly, F., McConway, K., & Ostrowski, E. (Eds.). (1993). *A Handbook of Small Data Sets*. Chapman & Hall.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding Analysis of Covariance. *Journal of Abnormal Psychology*, *110*, 40–48.

Mosteller, F., & Tukey, J. (1977). *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, *25*, 4334–4344.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch. (2005). *Regression Methods in Biostatistics. Linear, Logistic, Survival, and Repeated Measures Models*. Springer.

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, *22*, 392–399.

Young, S. G., & Bowman, A. W. (1995). Nonparametric analysis of covariance. *Biometrics*, *51*, 920–931.

Zar, J. H. (1998). *Biostatistical Analysis*. Pearson, Prentice Hall.