

Science des données II : module 3



Modèle linéaire

Philippe Grosjean & Guyliann Engels

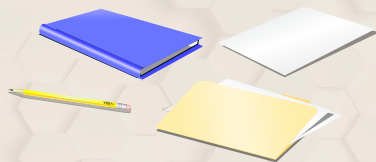
Université de Mons, Belgique
Laboratoire d'Écologie numérique



<http://wp.sciviews.org>
sdd@sciviews.org

Objectifs du cours

- Comprendre le modèle linéaire (ANOVA et régression linéaire tout en un)
- Appréhender la logique des matrices de contraste
- Découvrir l'ANCOVA



ANOVA et régression linéaire

- Nous avons vu que l'ANOVA et la régression linéaire se représentent par des modèles semblables :

$y = \mu + \tau_i + \epsilon$ pour l'ANOVA et

$y = \beta_1 + \beta_2 x + \epsilon$ pour la régression linéaire, avec

$\epsilon \sim \mathcal{N}(0, \sigma)$ dans les deux cas.

- La différence réside dans le type de variable explicative :
 - Variable **qualitative** pour l'ANOVA,
 - Variable **quantitative** pour la régression linéaire.
- Le calcul est, en réalité, identique en interne. Il est donc possible de généraliser ces deux approches en une seule appelée **modèle linéaire**.

Dans R

Une seule fonction fait tout : `lm()`. Les variables qualitatives et quantitatives sont encodées différemment. R est donc capable de déterminer tout seul s'il s'agit d'une ANOVA ou d'une régression.

Modèle linéaire commun

- Comment homogénéiser les deux modèles ANOVA et régression linéaire ?

$y = \mu + \tau_i + \epsilon$ pour l'ANOVA et

$y = \beta_1 + \beta_2 x + \epsilon$ pour la régression linéaire.

- Considérons pour l'ANOVA une variable qualitative à 2 niveaux.
Nous pouvons écrire :

$$y = \mu + \tau_1 I_1 + \tau_2 I_2 + \epsilon$$

avec I_i , une variable **indicatrice** qui prend la valeur 1 lorsque le niveau correspond à i , et 0 dans tous les autres cas.

Modèle linéaire commun (2)

On peut réécrire l'équation comme suit :

$$y = \mu + \tau_1 I_1 + \tau_1 I_2 - \tau_1 I_2 + \tau_2 I_2 + \epsilon$$

- En considérant $\beta_2 = \tau_2 - \tau_1$, cela donne :

$$y = \mu + \tau_1 I_1 + \tau_1 I_2 + \beta_2 I_2 + \epsilon$$

- En considérant $\beta_1 = \mu + \tau_1 = \mu + \tau_1 I_1 + \tau_1 I_2$, on obtient :

$$y = \beta_1 + \beta_2 I_2 + \epsilon$$

... qui est équivalent au modèle de la régression linéaire. Ceci se généralise pour une variable à k niveaux, avec $k - 1$ variables indicatrices au final.

Matrice de contraste

Donc, pour les variables qualitatives, nous considérons un ensemble de variables indicatrices (dans le cas précédent, la moyenne correspondant au premier niveau est considérée comme valeur de référence pour toutes les autres, et les variables indicatrices pour toutes les autres prennent la valeur de 1 séparément à chaque fois que le niveau correspondant est rencontré (k niveaux) :

$$y = \beta_1 + \beta_2 I_2 + \beta_3 I_3 + \dots + \beta_k I_k + \epsilon$$

- Il s'agit de ce qu'on appelle une **matrice de contrastes** de type traitement (voir l'instruction `contr.treatment(4)` pour générer la matrice à 4 niveaux *-en ligne les niveaux, en colonne les valeurs que prennent les I_{k-1} variables indicatrices-*).
- Les contrastes doivent être de préférence **orthogonaux par rapport à l'ordonnée à l'origine**, ce qui signifie que la somme de leurs pondérations doit être nulle pour tous les contrastes définis (*donc, en colonnes*).
- Les contrastes de type traitement ne sont pas orthogonaux!

Autres matrices de contrastes courantes

- Somme à zéro : `contr.sum(4)`
- Matrice de contrastes de Helmert : chaque niveau est comparé à la moyenne des niveaux précédents : `contr.helmert(4)`
- Matrice de contrastes polynomiaux : adapté aux facteurs ordonnés pour lesquels on s'attend à une certaine évolution du modèle du niveau le plus petit au plus grand : `contr.poly(4)`

```
plot(contr.poly(10)[, 1], type = "b")  
plot(contr.poly(10)[, 2], type = "b")  
plot(contr.poly(10)[, 3], type = "b")
```

- Explications supplémentaires sur les variables de type **factor**, ordonnées ou non dans R... Et utilisation dans R Studio
- R utilise par défaut des contrastes de traitement pour les facteurs non ordonnés et des contrastes polynomiaux pour des facteurs ordonnés. voir : `getOption("contrasts")`.

Mélange ANOVA et régression : l'ANCOVA

Variables prédictives : qualitatif + quantitatif => **Analyse de la COVAriance.**

Exemple

Masse de nouveaux nés en fonction du poids de la mère et du fait qu'elle fume ou non.

```
SciViews::R("model", lang = "fr")
babies <- read("babies", package = "UsingR")
# wt = masse du bébé à la naissance en onces et 999 = valeur manquante
# wt1 = masse de la mère à la naissance en livres et 999 = valeur manquante
# smoke = 0 (non), = 1 (oui), = 2 (jusqu'à grossesse),
#       = 3 (plus depuis un certain temps) and = 9 (inconnu)
babies %>.% select(., wt, wt1, smoke) %>.%
  filter(., wt1 < 999, wt < 999, smoke < 9) %>.%
  mutate(., wt = wt * 0.02835) %>.%
  mutate(., wt1 = wt1 * 0.4536) -> Babies
Babies$smoke <- as.factor(Babies$smoke)
# Descriptions graphiques
boxplot(data = Babies, wt ~ smoke)
boxplot(data = Babies, wt1 ~ smoke)
# Modèle linéaire (anciennement ANCOVA)
anova(lm(data = Babies, wt ~ smoke * wt1))
```