Science des données II : cours 7

Modèle linéaire

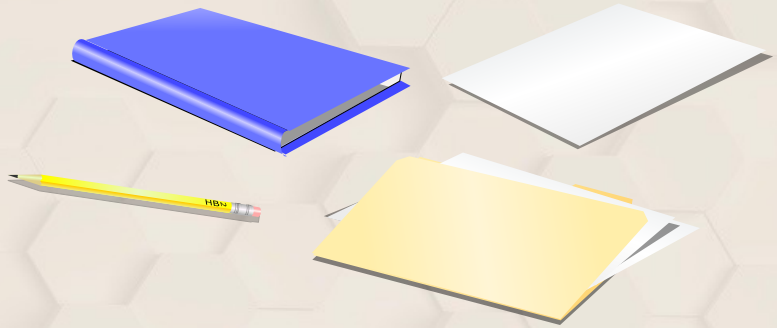
Philippe Grosjean & Guyliann Engels

Université de Mons, Belgique
Laboratoire d'Écologie numérique des Milieux aquatiques

http://biodatascience-course.sciviews.org
sdd@sciviews.org

UMONS

Objectifs du cours



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

UMONS

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.

UMONS

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.

UMONS

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.

UMONS

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) :

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

- 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!**

UMONS

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")`.

UMONS

Mélange ANOVA et régression : l'ANCOVA

Lorsque les variables prédictives sont un mélange de variables qualitatives et quantitatives, on parle d'ANCOVA, ou Analyse de la COVAriance.

Exemple

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

```
data(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)
library("dplyr")
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(wt ~ smoke, Babies)
boxplot(wt1 ~ smoke, Babies)
# ANCOVA
anova(lm(wt ~ smoke * wt1, data = Babies)
```

Modèle linéaire généralisé

- Le modèle linéaire nous a permis de généraliser la régression linéaire multiple (applicable seulement sur des variables quantitatives) à des variables réponses qualitatives grâce aux variables indicatrices Ii.
- Le modèle linéaire généralisé reprend cette idée, mais permet en plus d'avoir d'autres variables réponses que quantitatives (et avec distribution normale des résidus). Dans R, c'est la fonction `glm()`
- Nous rajoutons une fonction de lien $f(y)$ qui transforme la variable initiale en une variable quantitative dont la relation avec les variables explicatives est linéarisée :

$$f(y) = \beta_1 + \beta_2 I_2 + \beta_3 I_3 + ... + \beta_k I_k + \beta_l x_1 + \beta_m x_2 + ... + \epsilon$$

Par exemple

Pour une variable réponse binaire (distribution binomiale), avec une réponse de type logistique

$$y = 1/(1 + e^{-\beta x})$$

la transformation logit est une bonne fonction de lien : $\ln(y/(1-y)) = \beta x$

Modèle linéaire généralisé - application et exemple

Recherche d'effet de variables qualitatives et quantitatives sur une réponse binaire :

```
data(babies, package = "UsingR")
library("dplyr")
# Transformation, et garder aussi la variable 'gestation' en jours
# et 'ht', la taille de la mère en pouces à convertir en m (/ 39.37)
babies %>% select(gestation, smoke, wt1, ht) %>%
  dplyr::filter(gestation < 999, smoke < 9, wt1 < 999, ht < 999) %>%
  # Transformer wt1 en kg et ht en cm
  mutate(wt1 = wt1 * 0.4536) %>% mutate(ht = ht / 39.37) -> Babies_prem
Babies_prem$smoke <- as.factor(Babies_prem$smoke)
# Déterminer quels sont les enfants prématurés (nés avant 37 semaines)
Babies_prem$premat <- as.factor(as.numeric(Babies_prem$gestation < 7*37))
# BMI peut être plus parlant que la masse pour la mère?
Babies_prem %>% mutate(bmi = wt1 / ht^2) -> Babies_prem

# Modèle linéaire généralisé avec fonction de lien de type logit
summary(glm(premat ~ smoke + bmi, family = binomial(link = logit),
  data = Babies_prem))
```