Science des données II : cours 2

Régression linéaire multiple

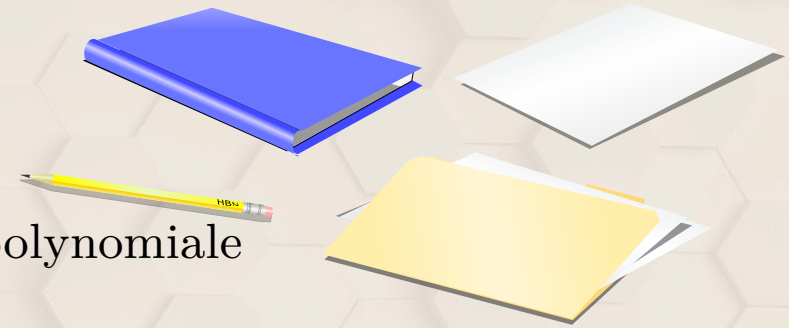
Philippe Grosjean & Guyliann Engels

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

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

UMONS

Objectifs du cours



- Découvrir la régression linéaire multiple et polynomiale
- Bien analyser les résidus
- Connaitre et savoir utiliser le critère d'Akaike

UMONS

Régression linéaire multiple

$$y = \alpha_1.x_1 + \alpha_2.x_2 + ... + \alpha_n.x_n + \beta + \epsilon$$

- L'erreur ϵ suit une loi Normale de moyenne nulle et d'écart type constant σ: $\epsilon \sim \mathrm{N}(0, \sigma)$
- La variance des résidus est constante (homoscedasticité)
- L'erreur est indépendante (problèmes des mesures répliquées dans le temps ou dans l'espace)
- L'**analyse des résidus** permet de vérifier ces différentes conditions, de détecter des valeurs aberrantes, et de mettre en évidence des relations non-linéaires

UMONS

Régression linéaire multiple (2)

- La régression linéaire simple est apparentée à l'ANOVA à 1 facteur (même principe).
- De même, la régression linéaire multiple est apparentée à l'ANOVA à plusieurs facteurs.
- **Une variable réponse** qui dépend de **plusieurs variables indépendantes** simultanément.
- Dans R, la régression multiple est une extension naturelle de la régression linéaire simple. Les mêmes outils sont utilisables. **Les snippets proposent des variantes pour régressions multiples**

Exemple

Le jeu de données *trees* (ou son équivalent *cerisiers*), volume de bois en fonction de la hauteur et du diamètre de l'arbre.

UMONS

Régression polynomiale

- **Rappel:** un polynome est une expression du type (*notez la ressemblance avec l'équation de la régression multiple*):

$$a_0 + a_1.x + a_2.x^2 + ... + a_n.x^n$$

- Un polynome d'**ordre 2** (x élevé jusqu'à la puissance 2) donne une parabole; un polynôme d'**ordre 3** correspond à une courbe en S.
- En considérant les puissances successives de la même variable dans la régression multiple, on obtient une régression polynômiale.
- Ce qui est intéressant : on utilise alors la régression **linéaire** pour ajuster en réalité une **courbe** (parabole, etc.)

Exemple

Utilisons la régression polynomiale sur *cerisiers*.

UMONS

Analyse des résidus

- Utiliser les différentes présentations graphiques pour visualiser graphiquement la distribution des résidus
  - **Résidus en fonction des valeurs prédites**: vue générale et détection de non linéarité et de valeurs extrêmes
  - **Graphqiue quantile-quantile** pour vérifier leur distribution nomale
  - **Racine carré des résidus standardisés en fonction des valeurs prédites** pour vérifier l'homoscédasticité.

Exemple

Illustration de l'utilisation de ces graphiques sur le jeu de données *cerisiers*

UMONS

Critère d'Akaike

- Le R^2 peut servir à quantifier la qualité d'ajustement d'un **modèle linéaire simple**.
- Dans le cas d'un **modèle multiple**, la complexité du modèle est liée au nombre de paramètres à estimer
- Plus un modèle est complexe, plus il est **flexible** et donc, il s'ajuste bien sur les points. Donc, c'est normal que le R^2 **augmente**

=> *mauvais critère pour comparer des modèles de complexité différente*

Le critère d'Akaike

introduit un terme de pénalisation en fonction du nombre de paramètres (*nbrpar*) à prédire qui rétablit l'équilibre (et au lieu d'utiliser le R^2, il utilise une autre descripteur statistique qui quantifie le degré d'ajustement, la *log-vraissemblance*) :

$$\text{AIC} = -2.\text{log-vraisemblance} + 2.\text{nbrpar}$$

UMONS