

# Science des données II : module 1



## Régression linéaire

Philippe Grosjean & Guyliann Engels

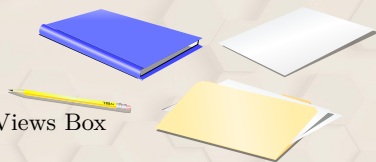
Université de Mons, Belgique  
Laboratoire d'Écologie numérique des Milieux aquatiques



<http://biodatascience-course.sciviews.org>  
[sdd@sciviews.org](mailto:sdd@sciviews.org)

# Objectifs du module

- Retrouver ses marques avec R, RStudio, SciViews Box
- Découvrir la régression linéaire
- Appréhender les différentes formes de régressions linéaires
- Choisir sa régression linéaire de manière judicieuse
- Savoir utiliser les outils de diagnostic de la régression linéaire, en particulier l'analyse des résidus.



## Point de départ - association entre deux variables quantitatives

- Trois niveaux d'association de force croissante: **corrélation, relation et causalité**.
- Deux **coefficients de corrélation différents** : celui de **Pearson** et celui de **Spearman**.
- Nous venons d'aborder la notion de **modèle** dans le cadre de l'ANOVA.
- La **régression linéaire** n'est autre qu'une généralisation du modèle de l'ANOVA qui permet de **représenter, quantifier et analyser une relation linéaire entre deux variables**.

*(si non linéaire, il faut essayer de transformer les données).*

# La régression linéaire simple

- Comme pour la corrélation, la régression nécessite **deux (au moins) variables quantitatives**.
- Contrairement à la corrélation, la régression linéaire simple **ne traite pas les deux variables sur un pied d'égalité** :
  - Une des deux variables est dite **dépendante (dependent)**
  - de l'autre (qui est dite **indépendante**).
- La représentation graphique associée est le **nuage de points** à travers duquel on tracera la droite qui symbolise la régression linéaire considérée.
- Par convention, la **variable indépendante** est toujours présentée en **abscisse** et la **variable dépendante en ordonnée**.

# Cas d'utilisation

## ■ La régression s'applique **dans deux cas** :

- 1] Dans une **expérience**, une variable est fixée par l'expérimentateur et l'autre, appelée **réponse**, est mesurée. Dans ce cas **la variable dépendante est toujours le réponse**.
- 2] Lors d'**observations**, des paires de valeurs sont mesurées pour deux variables, et on suspecte une relation entre elles. Dans ce cas, **il est plus difficile de décider quelle est la variable indépendante et quelle est la variable dépendante** => choix selon le point de vue.

## Exemple

- Circonférence, hauteur et volume de cerisiers : jeu de données **Cerisiers.csv**.
- Étudions ces données (analyse descriptive).
- Question : Y a-t-il des relations linéaires entre les variables mesurées ? Peut-on prédire le volume de bois sur pied à partir du diamètre ou de la hauteur de ces arbres ?

# Test autour de la régression linéaire simple

- Notre critère de détermination de la droite : la minimisation de la somme des carrés des résidus => **régression par les moindres carrés**.
- On considère que les résidus ont une **distribution normale** de **moyenne** nulle et d'**écart type**  $\sigma$  constant (homoscédasticité).
- Une fois le modèle **paramétrisé** (parameterized), la droite est définie, et nous pouvons calculer les **résidus**.

## Distribution des résidus

- Comment calculer l'écart type ?
- On ne le connaît pas... mais on peut l'estimer à partir de l'écart type des résidus  $s_{y|x}$  (analogie avec l'écart type de l'échantillon,  $s_y$ ).

$$s_{y|x} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \quad s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

- Calcul de l'intervalle de confiance sur une valeur prédite par le modèle pour Y :

$$CI_{1-\alpha} = \hat{y}_i \pm t_{\alpha/2}^{n-2} \cdot \frac{s_{y|x}}{\sqrt{n}} \quad CI_{1-\alpha} = \bar{y} \pm t_{\alpha/2}^{n-1} \cdot \frac{s_y}{\sqrt{n}}$$



## Significativité de la pente de la droite

- On considère la droite de régression observée  $Y = aX + b$  comme une estimation de la droite de régression  $Y = \alpha X + \beta$  de la population.
- Le paramètre **a** suit une distribution dans un méta-expérience, et la valeur observée est en fait l'une parmi toutes ses valeurs possibles.
- Si on considère une méta-expérience, on pourra **générer la distribution de a** et donc, inférer à l'aide d'un I.C. sur  $\alpha$ , (pente exacte pour la population).
- Par analogie avec la distribution de la moyenne d'un échantillon, on peut dire que a suit une **distribution de Student** de moyenne a et d'écart type valant :

$$S_a = \frac{s_{y|x}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

D'où on déduit l'IC sur a :

$$CI_{1-\alpha} = a \pm t_{\alpha/2}^{n-2} \cdot S_a$$

*Ce qui nous ramène à un **test de Student classique** pour déterminer la significativité de la pente de la droite (0 compris ou non dans l'IC ?)*

## Coefficient de détermination et ANOVA

- De même que lors de la décomposition de la variance  $s^2$  dans une ANOVA, on peut **décomposer la variance conditionnelle**  $s^2_{y|x}$  liée à la régression linéaire :

$SS(total) = SS(reg) + SS(residus)$  avec :

$$SS(total) = \sum (y_i - \bar{y})^2$$

$$SS(reg) = \sum (\hat{y}_i - \bar{y})^2$$

$$SS(residus) = \sum (y_i - \hat{y}_i)^2$$

- Le **coefficient de détermination**  $R^2 = SS(reg) / SS(total)$ . C'est la **fraction de variance expliquée par le modèle** (valeur comprise entre 0 et 1 ; plus il est élevé, plus la régression explique une part de variance importante).
- Comme pour l'ANOVA, on peut effectuer un **test de la significativité de la régression** car  $MS(reg) / MS(residus)$  suit une distribution  $F$  à respectivement 1 et  $n-2$  ddl.