

Science des données II : cours 6



Régression linéaire multiple

Philippe Grosjean & Guyliann Engels

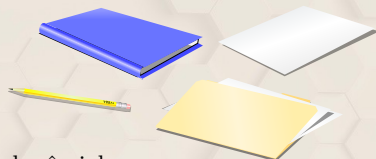
Université de Mons, Belgique
Laboratoire d'Écologie numérique des Milieux aquatiques



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

Objectifs du cours

- Se rappeler la régression linéaire simple
- Découvrir la régression linéaire multiple et polynômiale
- Bien analyser les résidus
- Connaitre et savoir utiliser le critère d'Akaike



Point de départ - Rappel 1

- 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, essayez avec le jeu de données pressure de R).

La régression linéaire simple - Rappel 2

- 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**.

La régression linéaire simple - Rappel 3

■ 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 - Rappel 4

- Notre critère de détermination de la droite : la minimisation de la somme des carrés des résidus => **régression par les moindres carrés**.
- Notez qu'il existe d'autres critères, par exemple, la minimisation de la somme des valeurs absolues des résidus (plus robuste envers les valeurs extrêmes).
- On considère que les résidus ont une **distribution normale** de **moyenne** nulle et d'**écart type** σ constant (homoscédasticité).
- Une fois le modèle **paramétrisé** (parameterized), la droite est définie, et nous pouvons calculer les résidus.

Distribution des résidus - Rappel 5

- 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 - Rappel 6

- 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 α , (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 - Rappel 7

- De même que lors de la décomposition de la variance s^2 dans une ANOVA, on peut **décomposer la variance conditionnelle** $s_{y|x}^2$ 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.

Régression linéaire multiple

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

- L'erreur ϵ suit une **loi Normale** de moyenne nulle et d'écart type constant σ :
 $\epsilon \sim N(0, \sigma)$
- 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

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.

Régression polynomiale

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

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

- 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 polynomiale**.
- 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*.

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**
 - **Graphique quantile-quantile** pour vérifier leur **distribution normale**
 - **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*

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 (nb_{rpar}) 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-vraisemblance*) :

$$AIC = -2 \cdot \log\text{-vraisemblance} + 2 \cdot nb_{rpar}$$