Méthodes mathématiques pour les sciences du vivant

Cours 1 : Droite des moindres carrés (ou droite de régression linéaire)

Série statistique à une variable : Un échantillon de taille n, noté $\{x_1, \ldots, x_n\}$ est une suite de n valeurs prises par une *variable aléatoire* (objet définie en Probabilités modélisant une quantité supposée aléatoire). La *moyenne*, la *variance* et l'*écart type* de l'échantillon sont, par définition :

$$m_x := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad v_x := \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - m_x^2 \qquad s_x := \sqrt{v_x}$$

Série statistique à deux variables : Si l'on observe deux quantités simultanément, $\{x_1, \ldots, x_n\}$ et $\{y_1, \ldots, y_n\}$, on peut représenter les données sous forme d'un *nuage de points* dans le plan $(x_i, y_i)_{i=1..n}$ dont le *centre de gravité* est le point (m_x, m_y) et dont la *covariance* est, par définition :

$$cov_{xy} := \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - m_x m_y.$$

Droite des moindres carrés ou droite de régression linéaire : La droite $y = \hat{a}x + \hat{b}$ obtenue en posant

$$\boxed{\hat{a} := \quad \frac{cov_{xy}}{s_x^2} \quad \text{et} \quad \hat{b} := m_y - \hat{a}m_x}$$

correspond au choix $\hat{a}$ et $\hat{b}$ des nombres a et b qui minimisent la somme des carrés des *résidus* ε_i

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(y_i - (ax_i + b))^2. \tag{1}$$

C'est la "meilleure" approximation linéaire du nuage de points, au sens du critère de minimisation de (1), appelé critère des *Moindres Carrés Ordinaires* (MCO). A noter que cette droite passe par le centre de gravité du nuage de points.

Notons que si, en échangeant les rôles de x et de y, on calcule une approximation linéaire de la forme $x = \hat{a}'y + \hat{b}'$, le critère MCO, $\sum_{i=1}^{n}(x_i - (a'y_i + b'))^2$ dans ce cas, n'est plus le même et la droite obtenue ne coïncide pas, en général, avec la précédente.

Valeurs observées/valeurs prédites par le modèle : Si $y = \hat{a}x + \hat{b}$ est la droite des moindres carrés d'un nuage de points $(x_i, y_i)_{i=1..n}$, on appelle *valeurs prédites* de y *par le modèle* les valeurs $\hat{y}_i := \hat{a}x_i + \hat{b}$. On utilise notamment ces valeurs pour faire des prévisions : si les x_i sont des dates successives, $x_1 < \ldots < x_n$, la valeur prédite pour y à une date future x_{n+1} est simplement $\hat{y}_{n+1} = \hat{a}x_{n+1} + \hat{b}$.

Coefficient de corrélation linéaire : Pour mesurer la qualité de l'approximation d'un nuage $(x_i, y_i)_{i=1..n}$ par sa droite des moindres carrés, on calcule son *coefficient de corrélation linéaire* défini par

$$r_{xy} = \frac{cov_{xy}}{s_x s_y}.$$

C'est un nombre compris entre -1 et $+1$, qui vaut $+1$ (resp. -1) si les points du nuage sont exactement alignés sur une droite de pente a positive (resp. négative). On considère que l'approximation d'un nuage par sa droite des moindres carrés est de bonne qualité lorsque $|r_{xy}|$ est proche de 1 (donc r_{xy} proche de $+1$ ou de -1) et de médiocre qualité lorsque $|r_{xy}|$ est proche de 0.

Attention à ne pas déduire trop hativement une relation de cause à effet d'une corrélation proche de 1 ou bien à écarter l'existence d'une telle relation lorsqu'elle est proche de 0 (présence de variables cachée).

Données éloignées, données influentes : On appelle *donnée éloignée* (*outlier*) un point du nuage situé à l'écart. S'il est éloigné dans la direction de y, il lui correspondra un important résidu. S'il est éloigné dans la direction des x, il peut présenter un très petit résidu et en même temps avoir une grande influence sur les valeurs de $\hat{a}$ et $\hat{b}$ trouvées.

On appelle *donnée influente* un point du nuage dont l'oubli conduirait à une droite des moindre carrée bien différente. C'est souvent le cas des données éloignées dans la direction des x.

Normalité des résidus : On dit que la série statistique à deux variables $\{x_1, \ldots, x_n\}$ et $\{y_1, \ldots, y_n\}$ suit un *modèle linéaire* s'il existe deux nombres $\hat{a}$ et $\hat{b}$ et une suite $\{\varepsilon_1, \ldots, \varepsilon_n\}$ de résidus *indépendants et distribués selon une loi normale centrée* tels que l'on ait pour tout $i = 1, \ldots, n$,

$$y_i = \hat{a}x_i + \hat{b} + \varepsilon_i.$$

Si l'on choisit pour $\hat{a}$ et $\hat{b}$ les coefficients des moindres carrés ordinaires, les résidus sont toujours centrés, par définition. Pour tester leur normalité, on peut tracer une *droite de Henri* (normal quantil plot) de la façon suivante : on commence par *réduire* les données (ici les résidus) c'est-à-dire par les diviser par leur écart type σ_ε. Puis on range les données par ordre croissant, et on note pour chacune d'elle le quantile qu'elle occupe (par exemple, s'il y a 20 données, la plus petite occupe le quantile 5%, la suivante le quantile 10%, ...) et on calcule (à l'aide d'une table ou de la calculette) la valeur correspondant à ce quantile pour la loi normale centrée réduite (par exemple $-1,645$ pour 5% et $-1,282$ pour 10%...). On représente les points ayant ces deux coordonnées et on devrait observer, si les données étaient celles d'un modèle linéaire des points alignés sur une droite $y = x$. En pratique, le modèle sera jugé d'autant plus convenable que les points sont bien alignés.

Calendrier des cours et TD :

13-17/10	Cours 1 + TD 1
20-24/10	Cours 2 + TD 2
27-31/10	TD 3 + TD 4
3-7/11	Cours 3 + TD 5
10-14/11	TD 6 + TD 7
17-21/11	Cours 4 + TD 8 + Controle
24-26/11	TD 9 + TD 10
1-5/12	Cours 5 + TD 11
8-12/12	TD 12 + TD 13
15-19/12	Cours 6 + TD 14 + Controle
5-9/01	TD 15 + TD 16
12-16/01	TD 17 + TD 18