

Option Statistique

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

**Série statistique à une variable :** Un échantillon de taille  $n$ , noté  $\{x_1, \dots, x_n\}$  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 :

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

**Série statistique à deux variables :** Si l'on observe deux quantités simultanément,  $\{x_1, \dots, x_n\}$  et  $\{y_1, \dots, y_n\}$ , 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

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

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

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

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).

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.

**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  $1$  et de médiocre qualité lorsque  $|r_{xy}|$  est proche de  $0$ .

**Valeurs approchées, prévisions :** 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 approchées* de  $y$  les valeurs  $\hat{y}_i := \hat{a}x_i + \hat{b}$ . On utilise aussi les valeurs approchées pour faire des prévisions : si les  $x_i$  sont des dates successives,  $x_1 < \dots < x_n$ , la valeur prédite pour  $y$  à une date future  $x_{n+1}$  est simplement  $\hat{y}_{n+1} = \hat{a}x_{n+1} + \hat{b}$ .