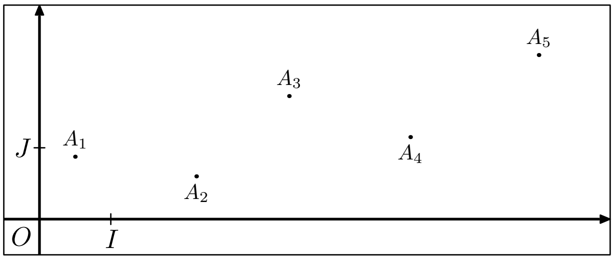
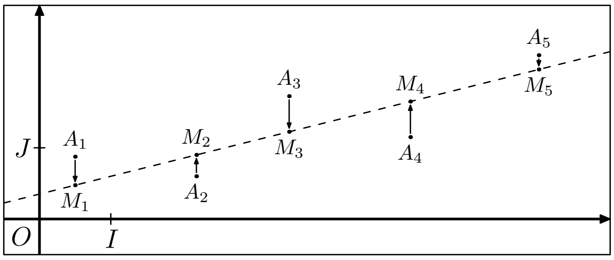
Régression linéaire

A. Introduction :

Soit $(x_i;y_i)_{0\leqslant i\leqslant n}$ une série statistique dont le relevé graphique que les points associés à cette série statistique sont "*presque alignés*" autour d'une droite.



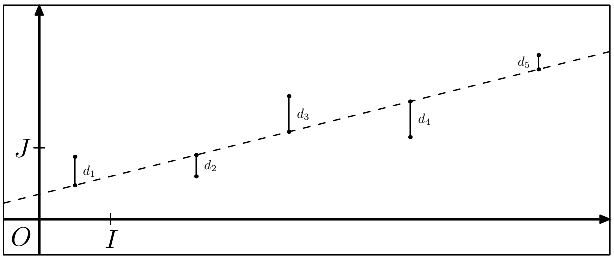
En modélisant ces données par la droite d'équation $y=a{\cdot}x+b$, on obtient les valeurs estimées par notre modèle qui sont les ordonnées des points M_i ci-dessous :



Pour mesurer l'erreur engendrée par cette représentation, on utilise la méthode des moindres carrées :

$$E(a;b)=\sum_{i=1}^{n}\big(y_i-\hat{y}_i\big)^2=\sum_{i=1}^{n}\big(y_i-a{\cdot}x_i-b\big)^2$$

Le choix de cette mesure de l'erreur entre les valeurs de la série statistique $(x_i;y_i)$ et les valeurs prédites $(x_i;\hat{y}_i)$ revient à mesurer la somme des carrés des variations sur l'axe des ordonnées.



On dit que la droite (d) réalise une **régression linéaire au sens des moindres carrés** lorsque les paramètres a et b minimise la valeur prise par la fonction E.

B. Quelques définitions :

Définition :
- Soit $\big(X_i\big)$ une série statistique à une variable.
  - la moyenne, notée $\overline{X}$, de la série $\big(X_i\big)$ a pour valeur :
    $$\overline{X}=\frac{1}{n}{\cdot}\sum_{i=1}^{n}X_i$$
  - la variance, notée $\mathrm{Var}(X)$, de la série $\big(X_i\big)$ a pour valeur : $\mathrm{Var}(X)=\dfrac{1}{n}{\cdot}\displaystyle\sum_{i=1}^{n}\big(X_i-\overline{X}\big)^2$
- Soit $\big(X_i;Y_i\big)$ une série statistique à deux variables. La covariance de la série $\big(X_i;Y_i\big)$, notée $\mathrm{Cov}\big(X;Y\big)$ a pour valeur : $\mathrm{Cov}\big(X;Y\big)=\dfrac{1}{n}{\cdot}\displaystyle\sum_{i=1}^{n}\big(X_i-\overline{X}\big)\big(Y_i-\overline{Y}\big)$

C. Droite de régression linéaire :

Théorème : Soit $\big(X_i;Y_i\big)$ une série statistique à deux variables.
Si la variance $\mathrm{Var}(X)$ de la série statistique $\big(X_i\big)$ est non-nulle, la droite de régression linéaire de y en x a pour équation $y=a{\cdot}x+b$ où :

$$a=\frac{\mathrm{Cov}\big(X;Y\big)}{\mathrm{Var}(X)}\quad ;\quad b=\overline{Y}-a{\cdot}\overline{X}$$

D. Démonstration :

Soit $\big(x_i;y_i\big)$ une série à deux variables tels que $\mathrm{Var}(X)\neq 0$.
Considérons la fonction f définie par :

$$f:(a;b)\longmapsto\sum_{i=1}^{n}\big(y_i-a{\cdot}x_i-b\big)^2$$

Proposition : la fonction f est strictement convexe

Preuve : la fonction f est un fonction quadratique *(polynômiale de degré 2)*. On en déduit qu'elle est continue et indéfiniment dérivable. On a les dérivées partielles :

- $\dfrac{\partial f}{\partial b}(a;b)=\displaystyle\sum_{i=1}^{n}-2x_i{\cdot}\big(u_i-a{\cdot}x_i-b\big)$
- $\dfrac{\partial f}{\partial b}(a;b)=\displaystyle\sum_{i=1}^{n}-2{\cdot}\big(y_i-a{\cdot}x_i-b\big)$
- $\dfrac{\partial^2 f}{\partial^2 a}(a;b)=\displaystyle\sum_{i=1}^{n}2{\cdot}{x_i}^2$
- $\dfrac{\partial^2 f}{\partial^2 b}(a;b)=\displaystyle\sum_{i=1}^{n}2=2{\cdot}n$
- $\dfrac{\partial^2 f}{\partial a\,\partial b}(a;b)=\displaystyle\sum_{i=1}^{n}2{\cdot}x_i$

Nous utiliseront la propriété suivante :
f est strictement convexe

$\Longleftrightarrow$ la matrice Hessienne H de f est définie positive

Or, la fonction f admet pour matrice Hessienne :

$$H = \begin{pmatrix} \sum_{i=1}^{n} 2\cdot x_i{}^2 & \sum_{i=1}^{n} 2\cdot x_i \\ \sum_{i=1}^{n} 2\cdot x_i & 2\cdot n \end{pmatrix}$$

On a :

- $\mathrm{Det}(H) = 2n\cdot \sum_{i=1}^{n} 2\cdot x_i{}^2 - \big(\sum_{i=1}^{n} 2\cdot x_i\big)^2$

$$= 4n^2\cdot\Big(\frac{1}{n}\cdot\sum_{i=1}^{n} x_i{}^2 - \Big(\frac{1}{n}\cdot\sum_{i=1}^{n} x_i\Big)^2$$

$$= 4n^2\cdot\Big[\frac{1}{n}\cdot\sum_{i=1}^{n} x_i{}^2 - \Big(\frac{1}{n}\cdot\sum_{i=1}^{n} x_i\Big)^2\Big]$$

$$= 4n^2\cdot \mathrm{Var}(X) > 0$$

- $\mathrm{Tr}(H) = 2n + 2\cdot\sum_{i=1}^{n} x_i{}^2 > 0$

Ces deux résultats montrent que la matrice hessienne de f est définie positive. On en déduit que la fonction f est strictement convexe.

Remarque : Voici une autre démonstration que la matrice H est définie positive.

En notant $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, on a : $H = 2\cdot X^t\cdot X$

Ainsi, pour tout $\begin{pmatrix} x \\ y \end{pmatrix}$, on a :

$$\begin{pmatrix} x \\ y \end{pmatrix}^t\cdot H\cdot\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}^t\cdot X^t\cdot X\cdot\begin{pmatrix} x \\ y \end{pmatrix}$$

$$= \Big[\begin{pmatrix} x \\ y \end{pmatrix}\cdot X\Big]^t\cdot\Big[X\cdot\begin{pmatrix} x \\ y \end{pmatrix}\Big] = M^t\cdot M \qquad \text{où } M{=}X\cdot u$$

Pour déterminer les valeurs minimisant la fonction f, nous utiliserons le théorème suivant :

Théorème : condition suffisante pour un minimum global
Soit $f:\mathbb{R}^n\longmapsto\mathbb{R}$ une application différentiable et x_0 un point annulant la dérivée de f *(point critique)*.
Si f est strictement convexe alors x_0 est l'unique minimum global de f.

Ainsi, le minimum global $(a\,;b)$ de la fonction f existe si il vérifie : $\dfrac{\partial f}{\partial a}(a,b) = 0$; $\dfrac{\partial f}{\partial b}(a,b) = 0$

- $$\frac{\partial f}{\partial b}(a,b) = 0$$

$$\sum_{i=1}^{n} -2\cdot\big(y_i - a\cdot x_i - b\big) = 0$$

$$-2\cdot\sum_{i=1}^{n}\big(y_i - a\cdot x_i - b\big) = 0$$

$$\sum_{i=1}^{n}\big(y_i - a\cdot x_i - b\big) = 0$$

$$\sum_{i=1}^{n} y_i - a\cdot\sum_{i=1}^{n} x_i - n\cdot b = 0$$

$$n\cdot b = \sum_{i=1}^{n} y_i - a\cdot\sum_{i=1}^{n} x_i$$

$$b = \frac{1}{n}\cdot\sum_{i=1}^{n} y_i - \frac{a}{n}\cdot\sum_{i=1}^{n} x_i$$

$$b = -a\cdot\overline{x} + \overline{y}$$

- La fonction f peut s'exprimer par :

$$f(a,b) = \sum_{i=1}^{n}\big(y_i - a\cdot x_i - b\big)^2$$

$$f(a,b) = \sum_{i=1}^{n}\big[y_i - a\cdot x_i - \big(-a\cdot\overline{x} + \overline{y}\big)\big]^2$$

$$f(a,b) = \sum_{i=1}^{n}\big(y_i - a\cdot x_i + a\cdot\overline{x} - \overline{y}\big)^2$$

$$f(a,b) = \sum_{i=1}^{n}\big[\big(y_i - \overline{y}\big) - a\cdot\big(x_i - \overline{x}\big)\big]^2$$

$$f(a,b) = \sum_{i=1}^{n}\big(y_i - \overline{y}\big)^2 - 2\cdot a\cdot\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big)^2 + a^2\big(x_i - \overline{x}\big)^2$$

$$f(a,b) = \sum_{i=1}^{n}\big(y_i - \overline{y}\big)^2 - 2\cdot a\sum_{i=1}^{n}\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big)^2 + a^2\sum_{i=1}^{n}\big(x_i - \overline{x}\big)^2$$

La dérivée partielle de la f valant 0 lorsque le minimum est réalisé vérifie :

$$\frac{\partial f}{\partial a}(a,b) = 0 - 2\sum_{i=1}^{n}\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big) + 2\cdot a\cdot\sum_{i=1}^{n}\big(x_i - \overline{x}\big)^2$$

$$0 = 2\cdot\Big[a\cdot\sum_{i=1}^{n}\big(x_i - \overline{x}\big)^2 - \sum_{i=1}^{n}\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big)\Big]$$

$$a\cdot\sum_{i=1}^{n}\big(x_i - \overline{x}\big)^2 = \sum_{i=1}^{n}\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big)$$

$$a = \frac{\frac{1}{n}\cdot\sum_{i=1}^{n}\big(y_i - \overline{y}\big)\big(x_i - \overline{x}\big)}{\frac{1}{n}\cdot\sum_{i=1}^{n}\big(x_i - \overline{x}\big)^2}$$

$$a = \frac{\mathrm{Cov}(x_i,y_i)}{\mathrm{Var}(x_i)}$$

