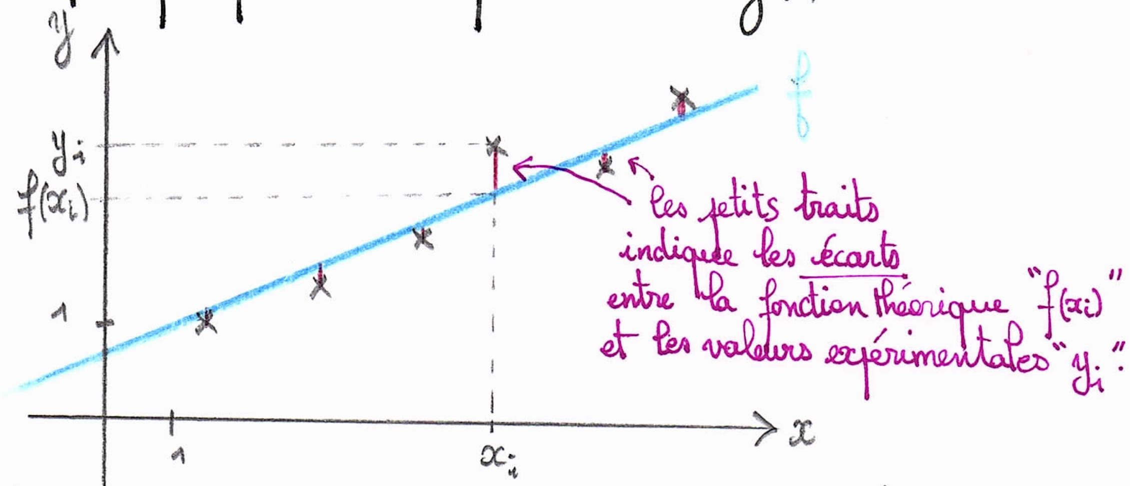
Leçon 40 : Droites de régression

① Présentation du problème

On a un nuage de points $(x_i ; y_i)$ et on cherche une fonction affine ou linéaire $f(x) = mx + p$ dont la courbe représentative passe le plus près possible des points du nuage.



« Le plus près possible » signifie qu'on cherche à rendre minimale la somme des carrés des écarts, c'est-à-dire

$$e = \sum (y_i - f(x_i))^2 = \boxed{\sum (y_i - mx_i - p)^2}$$

C'est la méthode « des moindres carrés ».

← C'est-à-dire des plus petits carrés.

② Cas où l'on cherche f linéaire

C'est-à-dire on a $p = 0$, et la seule inconnue qui reste est m.

Alors

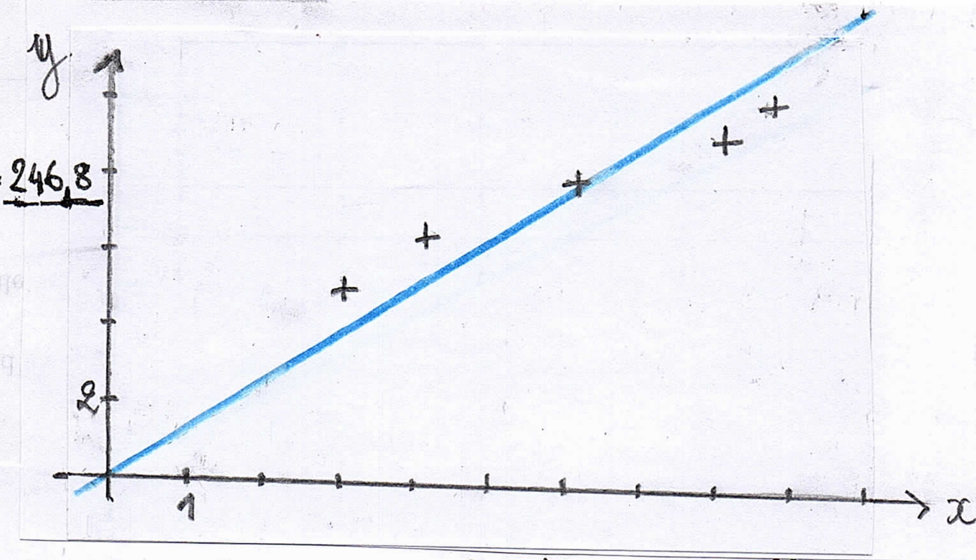
$$e = \sum (y_i - mx_i)^2 = \sum (y_i^2 - 2mx_iy_i + m^2x_i^2)$$

$$= \left(\sum x_i^2\right)m^2 - 2\times\left(\sum x_iy_i\right)m + \left(\sum y_i^2\right).$$

Propriété : e est minimale lorsque $\boxed{m = \dfrac{\sum x_iy_i}{\sum x_i^2}}$.

Ex :

x_i	3	4	6	8	8,5
y_i	5	6,5	8	9	10



$\sum x_i y_i = 3\times5+4\times6{,}5+6\times8+8\times9+8{,}5\times10 = \underline{246{,}8}$

$\sum x_i^2 = 3^2+4^2+6^2+8^2+8{,}5^2 = 197{,}5$

$m = \dfrac{\sum x_i y_i}{\sum x_i^2} \simeq \boxed{1{,}247}$

③ Cas où f est affine

On a alors deux inconnues : m et p. Parmi toutes les fonctions possibles on en cherche une qui passe par le point moyen $G(\bar{x}, \bar{y})$ avec

moyenne des x_i — moyenne des y_i

$$\boxed{\bar{x} = \frac{1}{N} \times \sum x_i} \quad \text{et} \quad \boxed{\bar{y} = \frac{1}{N} \times \sum y_i}.$$

N est le nombre de points

On a alors $f(x) = m\times x + p = m\times x + (\bar{y} - m\times\bar{x}) = \bar{y} + m\times(x - \bar{x})$.

$p = \bar{y} - m\times\bar{x}$ car $f(\bar{x}) = \bar{y}$.

Alors : $e = \sum \left(y_i - (mx_i + p)\right)^2 = \sum \left(y_i - \bar{y} - m\times(x_i - \bar{x})\right)^2$

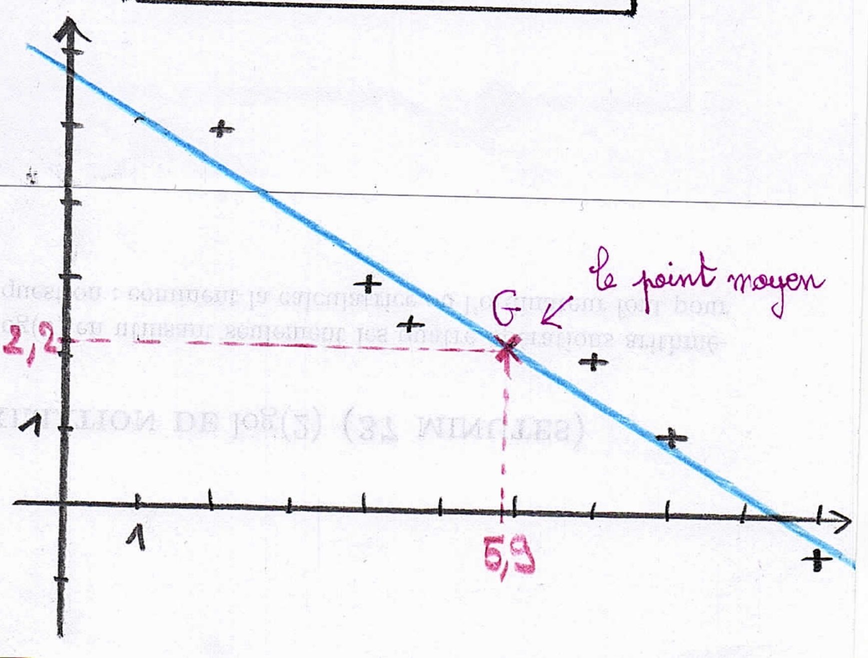
$= \sum \left[(y_i - \bar{y})^2 - 2m(x_i - \bar{x})(y_i - \bar{y}) + (x_i - \bar{x})^2 m^2\right]$

$= \left(\sum (x_i - \bar{x})^2\right) m^2 - 2\times\left(\sum (x_i - \bar{x})(y_i - \bar{y})\right) m + \sum (y_i - \bar{y})^2.$

PROPRIÉTÉ : e est minimale lorsque $\boxed{m = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}}$.

Ex :

x_i	2	4	4,5	7	8	10
y_i	5	3	2,5	2	1	-0,5



$\bar{x} = \frac{1}{6}\times(2+4+4{,}5+7+8+10) \simeq \underline{5{,}9}$

$\bar{y} = \frac{1}{6}\times(5+3+2{,}5+2+1-0{,}5) \simeq \underline{2{,}2}$

$\sum (x_i - \bar{x})(y_i - \bar{y}) = \ldots \simeq -26{,}7\ldots$

$\sum (x_i - \bar{x})^2 = \ldots \simeq 43{,}2\ldots$

$m = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \simeq \underline{-0{,}617\ldots}$