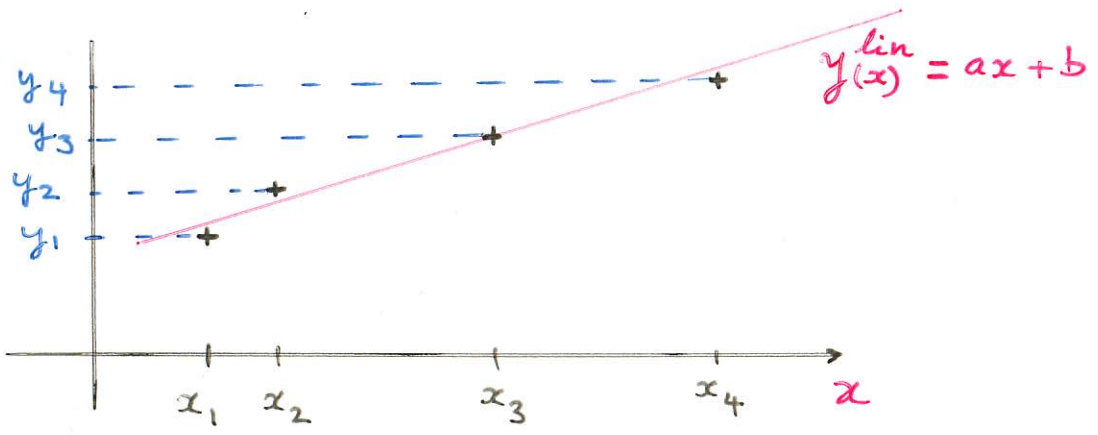
Régression linéaire



- Soient $\{x_i\}_{i=1,N}$ des nombres connus (par exemple les concentrations de solutions étalons)
- Soient $\{y_i\}_{i=1,N}$ les valeurs mesurées pour chaque x_i avec une incertitude δ_i
- Comment trouver la "meilleure" régression linéaire, c'est-à-dire les coefficients a et b tels que la fonction $y^{lin}(x) = ax+b$ est le plus proche possible de y_i lorsque $x = x_i$?
  Autrement dit, comment minimiser $\Delta y_i = y_i - y^{lin}(x_i)$ pour toutes les valeurs de i ?

- On définit la fonction d'erreur comme suit

$$\mathcal{E}(a,b) = \sum_{i=1}^{N} \left(\frac{\Delta y_i}{\delta_i}\right)^2 = \sum_{i=1}^{N} \frac{(y_i - ax_i - b)^2}{\delta_i^2}$$

- Recherche du minimum : pour une fonction $f(x)$ d'une seule variable qui a un minimum en $x = x_0$, le développement limité autour de x_0 s'écrit

$$f(x) = f(x_0) + \left.\frac{df}{dx}\right|_{x=x_0}(x-x_0) + \frac{1}{2}\left.\frac{d^2f}{dx^2}\right|_{x=x_0}(x-x_0)^2 + \cdots$$

si $\left.\frac{df}{dx}\right|_{x=x_0} > 0$, $f(x) < f(x_0)$ pour de très faibles valeurs négatives de $(x-x_0)$

si $\left.\frac{df}{dx}\right|_{x=x_0} < 0$, $f(x) < f(x_0)$ pour de très faibles valeurs positives de $(x-x_0)$

→ absurde !

donc $\boxed{\left.\frac{df}{dx}\right|_{x=x_0} = 0}$

De plus $\left.\frac{d^2f}{dx^2}\right|_{x=x_0} > 0$ (pour avoir un minimum)

- Les valeurs a_0 et b_0 qui minimisent $\mathcal{E}(a,b)$ sont donc telles que

$$\left.\frac{\partial \mathcal{E}(a,b)}{\partial a}\right|_{a_0,b_0} = 0 = \left.\frac{\partial \mathcal{E}(a,b)}{\partial b}\right|_{a_0,b_0}$$

$$\frac{\partial \mathcal{E}(a,b)}{\partial a} = \sum_{i=1}^{N} \frac{1}{\delta_i^2} \times 2 \times (-x_i)(y_i - ax_i - b)$$

$$\frac{\partial \mathcal{E}(a,b)}{\partial b} = \sum_{i=1}^{N} \frac{1}{\delta_i^2} \times 2 \times (-1)(y_i - ax_i - b)$$

En introduisant les notations

$$S = \sum_{i=1}^{N} \frac{1}{\delta_i^2}$$

$$S_y = \sum_{i=1}^{N} y_i/\delta_i^2$$

$$S_x = \sum_{i=1}^{N} x_i/\delta_i^2$$

$$S_{xx} = \sum_{i=1}^{N} x_i^2/\delta_i^2$$

$$S_{xy} = \sum_{i=1}^{N} x_i y_i/\delta_i^2,$$

on obtient

$$\frac{\partial \mathcal{E}(a,b)}{\partial a} = 2(-S_{xy} + aS_{xx} + bS_x)$$

$$\frac{\partial \mathcal{E}(a,b)}{\partial b} = 2(-S_y + aS_x + bS)$$

d'où

$$\begin{cases} -S_{xy} + a_0 S_{xx} + b_0 S_x = 0 & (1) \\ -S_y + a_0 S_x + b_0 S = 0 & (2) \end{cases}$$

$$\Updownarrow$$

$$\begin{cases} -SS_{xy} + a_0 SS_{xx} + b_0 SS_x = 0 & (1)\times S \rightarrow (3) \\ -S_x S_y + a_0 S_x^2 + b_0 S_x S = 0 & (2)\times S_x \rightarrow (4) \end{cases}$$

$$\Downarrow$$

$$\begin{cases} -SS_{xy} + S_x S_y + a_0 (SS_{xx} - S_x^2) = 0 & (3)-(4) \end{cases}$$

Ainsi

$$\boxed{a_0 = \frac{SS_{xy} - S_x S_y}{SS_{xx} - S_x^2}}$$

De plus $S_x \times (1) - S_{xx} \times (2)$

$$\downarrow$$

$$-S_x S_{xy} + S_x^2 b_0 + S_{xx} S_y - S_{xx} S b_0 = 0$$

d'où

$$\boxed{b_0 = \frac{S_{xx} S_y - S_x S_{xy}}{SS_{xx} - S_x^2}}$$

Incertitudes sur a et b :

soit $\delta a = a - a_0$ et $\delta b = b - b_0$.

$$\chi^2(a,b) = \chi^2(a_0+\delta a, b_0+\delta b) = \sum_{i=1}^{N} \frac{(y_i - a_0 x_i - \delta a x_i - b_0 - \delta b)^2}{\delta_i^2}$$

$$= \sum_{i=1}^{N} \frac{1}{\delta_i^2}\Big[(y_i - a_0 x_i - b_0)^2 + 2(y_i - a_0 x_i - b_0)(-\delta a x_i - \delta b) + (\delta a x_i + \delta b)^2\Big]$$

$$= \chi^2(a_0,b_0) + \delta a \underbrace{\frac{\partial \chi^2(a,b)}{\partial a}\Big|_{a_0,b_0}}_{0} + \delta b \underbrace{\frac{\partial \chi^2(a,b)}{\partial b}\Big|_{a_0,b_0}}_{0}$$

$$+ \sum_{i=1}^{N} \frac{1}{\delta_i^2}(\delta a x_i + \delta b)^2$$

soit $$\chi^2(a,b) = \chi^2(a_0,b_0) + \sum_{i=1}^{N} \frac{\delta a^2 x_i^2}{\delta_i^2} + 2\delta a \delta b \frac{x_i}{\delta_i^2} + \frac{\delta b^2}{\delta_i^2}$$

$$\boxed{\chi^2(a,b) = \chi^2(a_0,b_0) + \delta a^2 S_{xx} + 2\delta a \delta b S_x + \delta b^2 S}$$

L'écart maximal autorisé sur a par rapport à a_0 s'obtient en écrivant $\chi^2(a,b) = \chi^2(a_0,b_0) + S\left[\underbrace{\delta b^2 + 2\frac{S_x}{S}\delta a \delta b}_{(\delta b + \frac{S_x}{S}\delta a)^2 - \frac{S_x^2}{S^2}\delta a^2} + \frac{S_{xx}}{S}\delta a^2\right]$

soit $$\boxed{\chi^2(a,b) = \chi^2(a_0,b_0) + S\left(\delta b + \frac{S_x}{S}\delta a\right)^2 + \left(S_{xx} - \frac{S_x^2}{S}\right)\delta a^2}$$

En choisissant $\delta b = -\frac{S_x}{S}\delta a$ on réduit l'erreur mais il reste toujours le terme

$$\left(S_{xx} - \frac{S_x^2}{S}\right)\delta a^2$$

Remarque :

$$S_{xx} - \frac{S_x^2}{S} = \frac{S_{xx}S - S_x^2}{S} = -\frac{\Delta}{S}$$

où $$\Delta = S_x^2 - S S_{xx} = \left(\sum_{i=1}^{N} \frac{x_i}{\delta_i^2}\right)\left(\sum_{j=1}^{N} \frac{x_j}{\delta_j^2}\right) - \left(\sum_{i=1}^{N} \frac{1}{\delta_i^2}\right)\left(\sum_{j=1}^{N} \frac{x_j^2}{\delta_j^2}\right)$$

$$= \sum_{i,j=1}^{N} \frac{1}{\delta_i^2 \delta_j^2}\left[x_i x_j - x_j^2\right]$$

$$= \sum_{i,j=1}^{N} \frac{x_i x_j}{\delta_i^2 \delta_j^2} - \frac{1}{2}\sum_{i,j=1}^{N} \frac{x_j^2}{\delta_i^2 \delta_j^2} - \frac{1}{2}\sum_{i,j=1}^{N} \frac{x_i^2}{\delta_i^2 \delta_j^2}$$

Soit $\Delta = \frac{1}{2} \sum_{i,j=1}^{N} \frac{1}{\delta_i^2 \delta_j^2} \left(2x_i x_j - x_j^2 - x_i^2 \right)$

$$\Delta = -\frac{1}{2} \sum_{i,j=1}^{N} \frac{(x_i - x_j)^2}{\delta_i^2 \delta_j^2} < 0$$

- Ainsi l'erreur $\left(S_{xx} - \frac{S_x^2}{S}\right) \delta a^2 = -\frac{\Delta}{S} \delta a^2$ est bien positive et devient significative lorsque $-\frac{\Delta}{S} \delta a^2 \gg 1$

La valeur de δa pour laquelle l'erreur reste acceptable sera donc définie comme $\boxed{\delta a = \sqrt{-\frac{S}{\Delta}}}$ ← interprété comme l'incertitude sur la détermination de a_0.

- Incertitude sur la détermination de b_0 :

$$\mathcal{E}(a,b) = \mathcal{E}(a_0,b_0) + S_{xx} \underbrace{\left[\delta a^2 + 2 \delta a \delta b \frac{S_x}{S_{xx}} \right]} + S \delta b^2$$

$$\left(\delta a + \delta b \frac{S_x}{S_{xx}} \right)^2 - \frac{\delta b^2 S_x^2}{S_{xx}^2}$$

Soit $\mathcal{E}(a,b) = \mathcal{E}(a_0,b_0) + S_{xx} \left(\delta a + \frac{\delta b S_x}{S_{xx}} \right)^2 + \delta b^2 \underbrace{\left(S - \frac{S_x^2}{S_{xx}} \right)}_{\frac{-\Delta}{S_{xx}}}$

En choisissant $\delta a = -\frac{\delta b S_x}{S_{xx}}$ on minimise l'erreur. Pour que cette dernière soit acceptable il faut que $\delta b^2 \times \left(-\frac{\Delta}{S_{xx}}\right) \ll 1$

d'où $\boxed{\delta b = \sqrt{-\frac{S_{xx}}{\Delta}}}$ ← interprété comme l'incertitude sur la détermination de b_0.

- Coefficient de régression linéaire :

Afin de rendre compte de la qualité de la régression linéaire on calcule le coefficient dit de régression linéaire

$$r = \sqrt{1 - \frac{\mathcal{E}(a_0,b_0)}{V_y}}$$

où $V_y = \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{\delta_i^2}$,

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i,$$

$$\mathcal{E}(a_0,b_0) = \sum_{i=1}^{N} \frac{(y_i - a_0 x_i - b_0)^2}{\delta_i^2}.$$

$\bar{y}$ est la valeur moyenne des $\{y_i\}_{i=1,\dots,N}$

V_y indique l'écart des y_i à leur valeur moyenne.

Lorsque la régression est exacte, $\mathcal{E}(a_0,b_0) = 0 \Rightarrow r = 1$