HAC310X Mathématiques pour la chimie S3

J.L. Ramírez Alfonsín
Institut Montpelliérain Alexander Grothendieck,
Université de Montpellier
Place Eugène Bataillon, 34095, Montpellier

2 septembre 2024

Chapitre 1

Régression linéaire

La régression linéaire est un modèle statistique qui effectue des fonctions prédictives. Pour réaliser des estimations pertinentes, le processus s'appuie sur des valeurs numériques afin de dégager une tendance ou une évolution prévisible. Par le biais d'un dataset, le système permet ainsi de les extrapoler et d'anticiper des valeurs futures.

1.1 Méthode des moindres carrés

Supposons des données expérimentales portées sur un graphique (figure ci-dessous), qui forment un nuage de points. Les points sont numérotés de 1 jusqu'à n. On cherche à déterminer l'équation de la droite qui passe le plus près possible de l'ensemble des points, voir figure 1.1

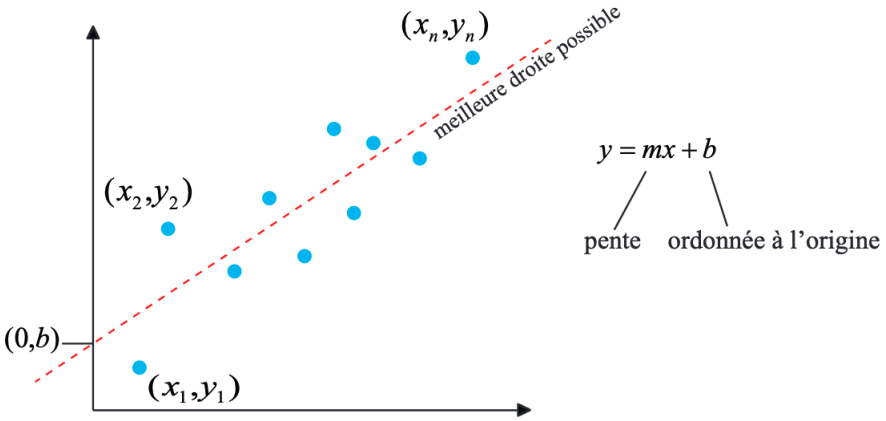

FIGURE 1.1 – Nuage de points avec meilleure droite possible.

On s'intéresse à l'erreur sur l'ordonnée de chaque point par rapport à la meilleure droite

possible. La figure 1.2 montre comment on mesure ces erreurs par rapport à la droite.

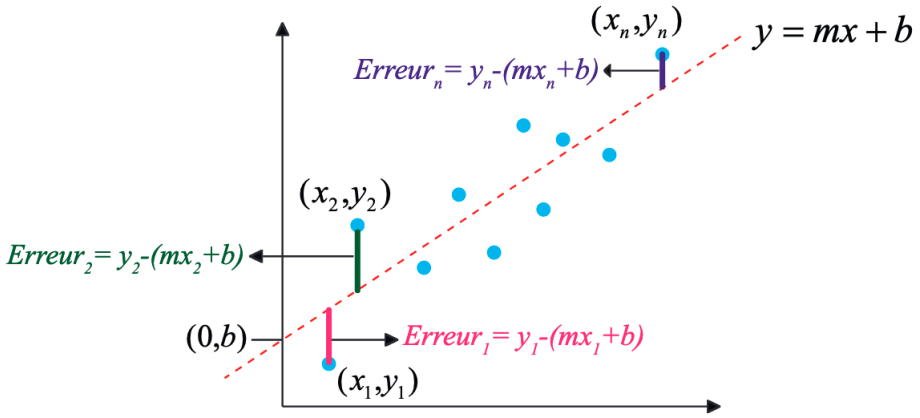

FIGURE 1.2 – Erreurs par rapport à la droite idéale.

Comme certaines des erreurs (ou résidus) sont positives et d'autres négatives, il est préférable d'utiliser les carrés des erreurs pour quantifier l'erreur par rapport à la droite.

On écrira donc que la somme des carrés des erreurs (SCE) vaut

$$SCE = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \cdots + (y_n - (mx_n + b))^2 \qquad (1.1)$$

Pour trouver la meilleure droite possible, il faut déterminer les valeurs de m et de b qui minimisent la SCE.
En développant (1.1) nous obtenons

$$\begin{aligned} SCE \quad &= (y_1^2 + y_2^2 + \cdots + y_n^2) - 2m(y_1x_1 + y_2x_2 + \cdots + y_nx_n) - 2b(y_1 + y_2 + \cdots + y_n) \\ &+m^2(x_1^2 + x_2^2 + \cdots + x_n^2) + 2mb(x_1 + x_2 + \cdots + x_n) + nb^2 \end{aligned}$$

Le premier terme de cette équation peut être réécrit différemment. En effet, la moyenne des carrés des y s'écrit

$$\overline{y^2} = \frac{y_1^2 + y_2^2 + \cdots + y_n^2}{n} \text{ et donc } n\overline{y^2} = y_1^2 + y_2^2 + \cdots + y_n^2$$

Également, le deuxième terme de la SCE, la moyenne des produits des x par y, s'écrit

$$\frac{y_1x_1 + y_2x_2 + \cdots + y_nx_n}{n} = \overline{xy} \text{ et donc } n\overline{xy} = y_1x_1 + y_2x_2 + \cdots + y_nx_n$$

Nous obtenons ainsi que

$$SCE = n\overline{y^2} - 2mn\overline{xy} - 2bn\overline{y} + m^2n\overline{x^2} + 2mbn\overline{x} + nb^2$$

Nous souhaitons donc minimiser la SCE. Il faut donc trouver les valeurs de m et b pour les quelles la SCE présente un minimum (*Attention* : ici les x et les y ne sont pas des variables, leurs valeurs sont connues. Seules m et b sont inconnues.)
Nous avons donc que SCE est une fonction à deux variables. Par exemple, considérons la droite $y = 5x + 3$ et une nuage de points dispersés autour de cette droite, voir figure 1.3.

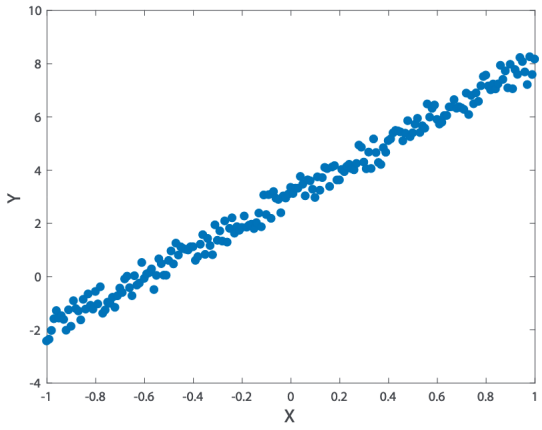

FIGURE 1.3 – Droite et nuage de points.

La figure 1.4 illustre la surface correspondant à SCE qui présente un minimum.

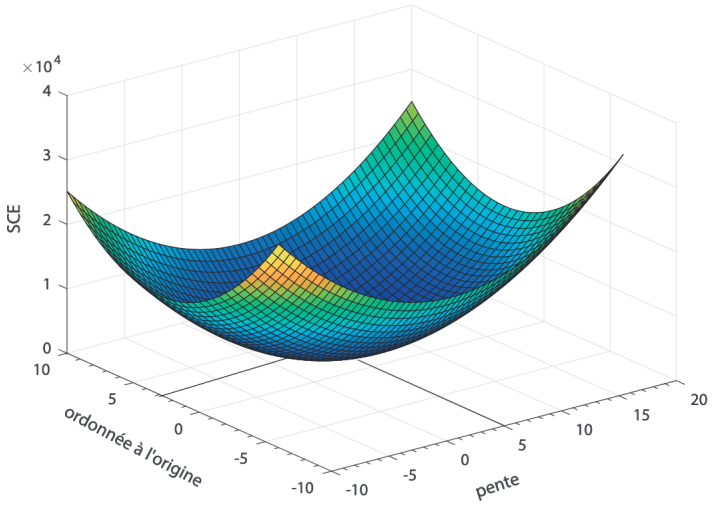

FIGURE 1.4 – Surface en fonction de $m = 5$ et $b = 3$.

Nous cherchons donc à résoudre le système

$$\frac{\partial SCE}{\partial m} = -2n\overline{xy} + 2mn\overline{x^2} + 2bn\overline{x} = 0$$
$$\frac{\partial SCE}{\partial b} = -2n\overline{y} + 2mn\overline{x} + 2bn = 0$$

On peut diviser ces deux équations par $2n$ et obtenir le système

$$\begin{aligned} m\overline{x^2} + b\overline{x} &= \overline{xy} \\ m\overline{x} + b &= \overline{y} \end{aligned} \tag{1.2}$$

Réécrivons la première équation de (1.2) sous la forme

$$m\frac{\overline{x^2}}{\overline{x}} + b = \frac{\overline{xy}}{\overline{x}}$$

et on le soustraite la deuxième équation de (1.2),

$$m\left(\frac{\overline{x^2}}{\overline{x}} - \overline{x}\right) = \frac{\overline{xy}}{\overline{x}} - \overline{y},$$

obtenant la valeur de m

$$m = \frac{\frac{\overline{xy}}{\overline{x}} - \overline{y}}{\frac{\overline{x^2}}{\overline{x}} - \overline{x}}$$

et évidemment, on peut trouver la valeur de b facilement à l'aide de la deuxième équation de (1.2)

$$b = \overline{y} - m\overline{x}.$$

Reprenons maintenant l'expression de m en termes de la variance et la covariance.
On rappelle que la *variance* des x est définie comme :

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\overline{x})^2 = \overline{x^2} - (\overline{x})^2$$

et la *covariance* des x et y est définie par

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

et il est possible de montrer que

$$cov(x,y) = \overline{xy} - \overline{x}\ \overline{y}.$$

Avec ces définitions, on peut donc écrire que

$$m = \frac{cov(x,y)}{\sigma_x^2} \text{ et } b = \overline{y} - m\overline{x}.$$

La droite $y = mx + b$ est appelée *droite des Moindres Carrés* (en abrégé droite des MC).

Remarque 1.1.1 *a) Pour l'expression de m on suppose que le la variance des x est non nul. Or ceci ne peut arriver que si tous les x_i sont égaux, situation sans intérêt pour notre problème et que nous excluons donc a priori dans toute la suite.*

b) La relation $b = \overline{y} - m\overline{x}$ montre que la droite des MC passe par le centre de gravité du nuage $(\overline{x}, \overline{y})$.

1.2 Mesure d'adéquation

Les paramètres étant estimés, l'étape suivante consiste à définir une mesure "raisonnable" de l'adéquation du modèle en fonction des données. Nous allons donc décrire comment est calculé le celèbre coefficient de détermination appelé R^2 qui sert à quantifier la qualité de la régression linéaire. Il faut cependant s'assurer visuellement que le nuage de points suit bien une droite pour apprécier la valeur de ce coefficient.

Soient $e_1, e_2, \ldots, e_n$ les erreurs par rapport à la droite $D : y = mx + b$. Nous avons vu qu'il fallait calculer la somme des carrés des erreurs par rapport à la droite D

$$SCE_D = e_1^2 + e_2^2 + \cdots + e_n^2.$$

Considérons la somme des carrés des erreurs par rapport à la moyenne des y, voir figure 1.5.

$$SCE_{\overline{y}} = (y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \cdots + (y_n - \overline{y})^2.$$

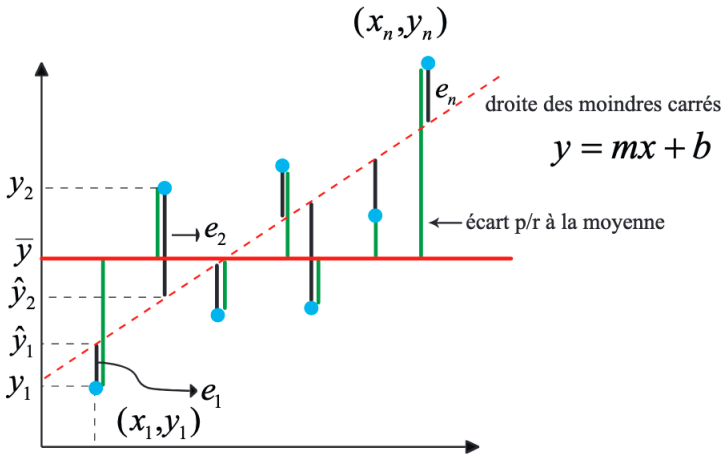

FIGURE 1.5 – Nuage de points avec les écarts par rapport à la moyenne $\overline{y}$ (barres vertes).

Nous avons donc que la proportion de la $SCE_{\overline{y}}$ qui n'est pas expliquée par la droite D est

$$\frac{SCE_D}{SCE_{\overline{y}}}.$$

La proportion de la $SCE_{\overline{y}}$ qui est expliquée par D, qui est notée R^2, est donnée par

$$R^2 = 1 - \frac{SCE_D}{SCE_{\overline{y}}} \qquad (\textit{Coefficient de détermination}).$$

Le coefficient de détermination peut être exprimé comme suit

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum(x_i - \bar{x})^2)(\sum(y_i - \bar{y})^2)}}.$$

Exemple 1 *(***Etalonnage pour le dosage du glucose***) Pour pouvoir déterminer la concentration du Glucose à partir de la mesure de l'absorbance, on effectue une calibration préliminaire :* une courbe d'étalonnage du dosage du Glucose.
Une courbe étalon est donc réalisée à partir d'une solution mère que l'on dilue et dont on mesure l'absorbance : on obtient ainsi un nuage de points,

Solutions	*Glucose*(g/L)	*Absorbance*
Blanc	0	0
Dilution 1/4	$0,123$	$0,1$
Dilution 1/2	$0,246$	$0,199$
Dilution 3/4	$0,369$	$0,301$
Non diliue	$0,492$	$0,411$

La figure 1.6 illustre le nuage des points.

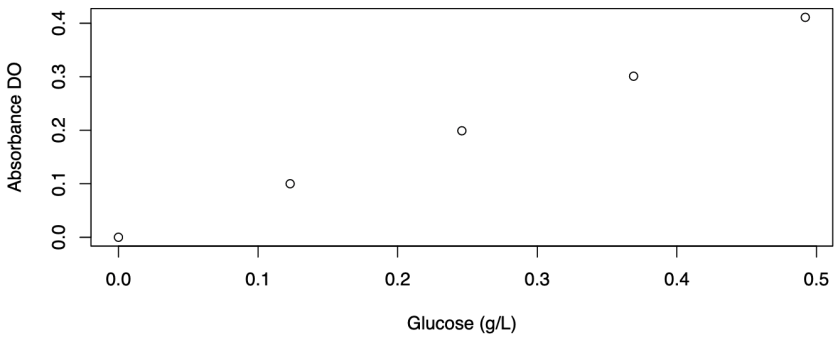

FIGURE 1.6 – Etalonnage de dosage du Glucose.

Nous retrouvons que : $absorbance = 0.832 \times glucose - 0.002$ *et* $R^2 = 0.99956$*, voir figure 1.7.*
Le R^2 *étant très proche de 1, cela signifie que le pouvoir prédictif du modèle est fort.*

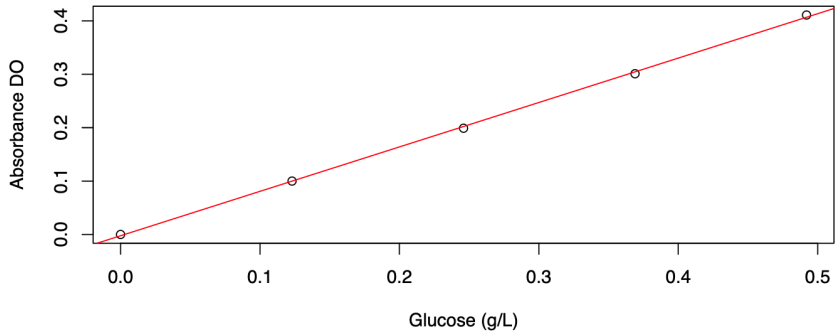

FIGURE 1.7 – Droite de régression entre concentration de Glucose et absorbance.