Chapitre 2

Méthodes de régression

En sciences, il arrive très souvent d'avoir des données et de vouloir les représenter au mieux par une fonction. On utilise pour faire ça des méthodes de régression. Dans ce chapitre, nous allons expliquer la méthode de régression linéaire pour approcher les données par une droite et, plus généralement, la méthode de régression polynomiale pour approcher les données par un polynôme.

2.1 Idée générale des méthodes de régression

Supposons que nous disposions de données sous la forme de N couples de points

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_N, y_N),$$

issus de mesures expérimentales ou de calculs compliqués. Souvent x est une variable que l'on fixe, et y est une variable que l'on mesure. Pour fixer les idées, prenons l'exemple suivant avec $N = 11$ couples de points :

x	0	1	2	3	4	5	6	7	8	9	10
y	10	8	15	15	30	40	44	60	80	85	110

Ces données sont tracées sur la figure 2.1.

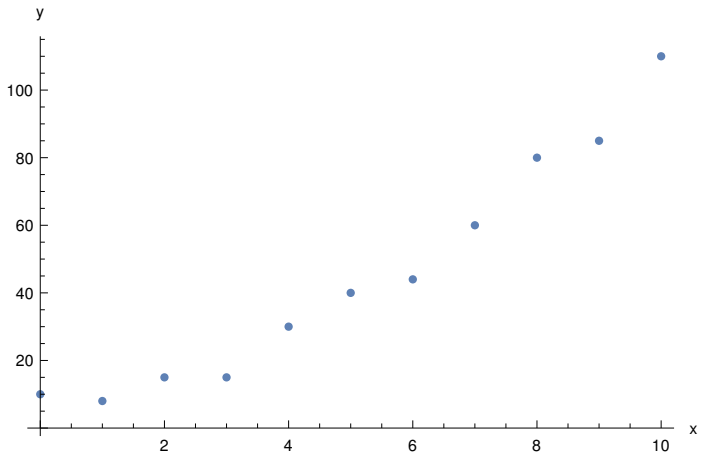

FIGURE 2.1 – Représentation graphique des données.

Dans le problème de la régression, on cherche à trouver une fonction $f : x \mapsto f(x)$, si possible simple, qui permet d'approcher au mieux les données

$$\boxed{y \approx f(x).}$$

On parle aussi d'ajustement de courbe ou, en anglais, de « fitting ».

Approcher les données par une fonction a deux intérêts :
- On détermine ainsi la relation (approximative) existant entre les variables x et y, c'est-à-dire une loi empirique qui permet de rationaliser les données.
- On est alors capable de prédire la valeur de y pour une valeur de x pour laquelle on ne dispose pas de données, ou inversement.

Nous traiterons uniquement les cas simples où l'on choisit pour f :
- une fonction affine (une droite) $f(x) = ax + b$ où a et b sont les paramètres à déterminer ;
- un polynôme de plus haut degré, comme par exemple une parabole $f(x) = ax^2 + bx + c$ où a, b et c sont les paramètres à déterminer.

Les méthodes de régression sont néanmoins beaucoup plus générales que ça. On peut faire des régressions avec des fonctions beaucoup plus compliquées. Par exemple, on peut utiliser une fonction très compliquée de plusieurs variables donnée sous la forme d'un réseau de neurones artificiels (codé dans un ordinateur), et qui contient alors beaucoup de paramètres à déterminer. Il faut déterminer tous ces paramètres avec beaucoup de données. C'est le domaine du « machine learning » ou apprentissage automatique, qui est une branche de l'intelligence artificielle. Ce domaine se développe actuellement rapidement. En chimie par exemple, avec ces techniques et beaucoup de données, on peut essayer d'« apprendre » la fonction reliant la géométrie d'une molécule (natures et positions des atomes) à une propriété physicochimie donnée (par exemple, son enthalpie de formation) afin de pouvoir faire des prédictions rapides sur des nouvelles molécules.

2.2 Régression linéaire

2.2.1 Description de la méthode de régression linéaire

Dans la régression linéaire, on choisit d'approcher les données par une droite, c'est-à-dire une fonction $f : x \mapsto f(x)$ affine

$$\boxed{f(x) = ax + b,}$$

avec a et b deux paramètres réels à déterminer. La méthode la plus courante pour déterminer ces paramètres est la méthode des moindres carrés.
Pour une valeur x_i dans les données, la valeur de y associée prédite par le modèle est $f(x_i) = ax_i + b$ alors que la valeur réelle dans les données est y_i. Pour ce point, l'erreur commise sur y est donc $(f(x_i) - y_i)$. Dans la méthode des moindres carrés, on choisit les paramètres a et b qui minimisent la somme des carrés de ces erreurs

$$\boxed{F(a,b) = \sum_{i=1}^{N} \left(f(x_i) - y_i\right)^2.}$$

On prend les carrés des erreurs car on n'est pas intéressé par les signes de ces erreurs mais uniquement par leurs valeurs absolues. Il s'agit donc d'un problème d'optimisation (sans contraintes) : celui de chercher le point (a_0, b_0) correspondant au minimum de la fonction

$$\begin{array}{rcl} F: & \mathbb{R}^2 & \to \mathbb{R} \\ & (a,b) \mapsto F(a,b) = \displaystyle\sum_{i=1}^{N} (a\ x_i + b - y_i)^2 . \end{array}$$

2.2.2 Détermination des paramètres a_0 et b_0

On cherche donc (a_0, b_0) comme un point stationnaire de F, c'est-à-dire un point où les dérivées partielles premières de F sont nulles

$$\begin{cases} \dfrac{\partial F}{\partial a}(a_0, b_0) = 0 \\ \dfrac{\partial F}{\partial b}(a_0, b_0) = 0, \end{cases}$$

ce qui donne, après calcul des dérivées partielles,

$$\begin{cases} \displaystyle\sum_{i=1}^{N} 2x_i\,(a_0\ x_i + b_0 - y_i) = 0 \\ \displaystyle\sum_{i=1}^{N} 2\,(a_0\ x_i + b_0 - y_i) = 0. \end{cases}$$

Il s'agit d'un système de deux équations à deux inconnues (a_0 et b_0). Écrivons le plus clairement en factorisant par a_0 et b_0

$$\begin{cases} \left(\displaystyle\sum_{i=1}^{N} x_i^2\right) a_0 + \left(\sum_{i=1}^{N} x_i\right) b_0 = \sum_{i=1}^{N} x_i y_i \\ \left(\displaystyle\sum_{i=1}^{N} x_i\right) a_0 + \left(\sum_{i=1}^{N} 1\right) b_0 = \sum_{i=1}^{N} y_i, \end{cases}$$

En utilisant $\sum_{i=1}^{N} 1 = N$ et en introduisant les quantités suivantes, que l'on peut calculer facilement à partir des données,

$$S_x = \sum_{i=1}^{N} x_i, \quad S_y = \sum_{i=1}^{N} y_i, \quad S_{xx} = \sum_{i=1}^{N} x_i^2, \quad S_{xy} = \sum_{i=1}^{N} x_i y_i,$$

on peut mettre finalement le système sous la forme compacte

$$\begin{cases} S_{xx}\ a_0 + S_x\ b_0 = S_{xy} \\ S_x\ a_0 + N\ b_0 = S_y. \end{cases}$$

Utilisons par exemple la méthode matricielle pour résoudre ce système d'équations. Sous forme matricielle, le système d'équations se réécrit comme

$$\begin{pmatrix} S_{xx} & S_x \\ S_x & N \end{pmatrix} \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} = \begin{pmatrix} S_{xy} \\ S_y \end{pmatrix}.$$

Le déterminant de la matrice 2×2 du système d'équations vaut $S_{xx}N - S_x^2$. Si celui-ci est non nul, on peut introduire l'inverse de la matrice, ce qui permet de résoudre le système d'équations sous la forme

$$\begin{pmatrix} a_0 \\ b_0 \end{pmatrix} = \begin{pmatrix} S_{xx} & S_x \\ S_x & N \end{pmatrix}^{-1} \begin{pmatrix} S_{xy} \\ S_y \end{pmatrix},$$

où l'inverse de la matrice 2×2 est

$$\begin{pmatrix} S_{xx} & S_x \\ S_x & N \end{pmatrix}^{-1} = \frac{1}{S_{xx}N - S_x^2} \begin{pmatrix} N & -S_x \\ -S_x & S_{xx} \end{pmatrix}.$$

On arrive donc aux expressions de a_0 et b_0 :

$$\boxed{\begin{cases} a_0 = \dfrac{NS_{xy} - S_x S_y}{S_{xx}N - S_x^2} \\ b_0 = \dfrac{-S_x S_{xy} + S_{xx} S_y}{S_{xx}N - S_x^2}. \end{cases}}$$

On pourrait vérifier en calculant les dérivées partielles secondes de F (mais on ne le fera pas) que ce point stationnaire (a_0, b_0) correspond bien à un minimum de la fonction F.

Appliquons ces expressions à l'exemple de la partie 2.1. On a $N = 11$ et on calcule $S_x = 55$, $S_y = 497$, $S_{xx} = 385$, $S_{xy} = 3592$. Cela conduit à $a_0 = 1107/110 \approx 10.064$ et $b_0 = -113/22 \approx -5.136$. On a donc trouvé par régression linéaire l'expression de la droite approchant au mieux les données :

$$f(x) = 10.064\, x - 5.136.$$

Cette droite est tracée sur la figure 2.2.

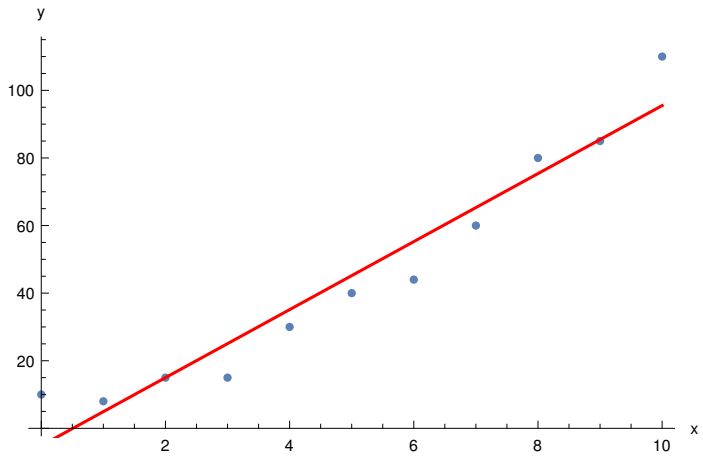

FIGURE 2.2 – Droite déterminée par régression linéaire approchant au mieux les données.

2.2.3 Coefficient de détermination R^2

Pour mesurer la qualité de l'approximation des données par la fonction f de régression, on utilise très souvent le coefficient de détermination, noté R^2, et défini par

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - f(x_i))^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2},$$

où $\bar{y}$ est la moyenne des valeurs y_i définie par

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

Le coefficient de détermination R^2 est toujours compris entre 0 et 1.

Dans la définition de R^2, le terme $\sum_{i=1}^{N} (y_i - f(x_i))^2 / \sum_{i=1}^{N} (y_i - \bar{y})^2$ est le ratio entre la variance des erreurs sur y_i due au modèle de régression et la variance totale des valeurs y_i.

- Si ce terme est petit, cela signifie que le modèle explique bien les variations des valeurs y_i, et alors R^2 est proche de 1. La qualité de la régression est bonne. Dans la limite idéale où les erreurs sur y_i sont toutes nulles (c'est-à-dire que la fonction f passe exactement par tous les points) alors on a la valeur maximale $R^2 = 1$.
- Si ce terme est grand, cela signifie que le modèle n'explique pas bien les variations des valeurs y_i, et alors R^2 est éloigné de 1. La qualité de la régression est mauvaise, et on aurait intérêt à essayer de modéliser les données avec une autre fonction f.

Pour l'exemple de la régression linéaire effectuée dans la partie précédente, on calcule $\bar{y} \approx 45.182$, $\sum_{i=1}^{N} (y_i - f(x_i))^2 \approx 779.191$ et $\sum_{i=1}^{N} (y_i - \bar{y})^2 \approx 11919.6$, ce qui donne un coefficient de détermination de $R^2 \approx 0.935$. Il s'agit d'une régression de qualité acceptable mais pas exceptionnelle.

Exercice 5. *(Cinétique chimique).* A 773 K, le cyclopropane se transforme en propène. La mesure de la concentration en cyclopropane c en fonction du temps t donne les résultats suivants :

t (s)	0	300	600	900	1200	1800	2400	3000
c (mmol/L)	1.5	1.24	1.00	0.83	0.68	0.46	0.31	0.21

Si on fait l'hypothèse que cette réaction a une cinétique d'ordre 1 avec une constante de vitesse k, on devrait avoir la loi d'évolution de la concentration suivante :

$$-\frac{dc(t)}{dt} = k\, c(t) \quad \Longrightarrow \ln c(t) = -k\, t + \ln c(0).$$

Pour vérifier cette hypothèse, tracer $\ln c$ en fonction de t et effectuer une régression

linéaire. Déterminer la constante de vitesse k et le coefficient de détermination R^2, et conclure sur la validité de l'hypothèse.

2.3 Régression polynomiale

2.3.1 Description de la méthode de régression polynomiale de degré 2

Dans la régression polynomiale, on choisit d'approcher les données par un polynôme d'un certain degré. Nous traiterons ici le cas d'un polynôme de degré 2, c'est-à-dire une parabole. La fonction $f : x \mapsto f(x)$ est alors

$$\boxed{f(x) = ax^2 + bx + c,}$$

avec a, b et c trois paramètres réels à déterminer.
Nous utilisons ici aussi la méthode des moindres carrés, c'est-à-dire que l'on choisit les paramètres a, b et c qui minimisent la somme des carrés des erreurs entre le modèle et les données. On cherche donc le point (a_0, b_0, c_0) correspondant au minimum de la fonction de trois variables

$$\boxed{\begin{array}{rccl} F: & \mathbb{R}^3 & \to & \mathbb{R} \\ & (a,b,c) & \mapsto & F(a,b,c) = \displaystyle\sum_{i=1}^{N} \left(a\ x_i^2 + b\ x_i + c - y_i\right)^2. \end{array}}$$

2.3.2 Détermination des paramètres a_0, b_0 et c_0

On cherche donc (a_0, b_0, c_0) comme un point stationnaire de F, c'est-à-dire un point où les dérivées partielles premières de F sont nulles

$$\begin{cases} \dfrac{\partial F}{\partial a}(a_0, b_0, c_0) = 0 \\ \dfrac{\partial F}{\partial b}(a_0, b_0, c_0) = 0 \\ \dfrac{\partial F}{\partial c}(a_0, b_0, c_0) = 0, \end{cases}$$

ce qui donne, après calcul des dérivées partielles,

$$\begin{cases} \displaystyle\sum_{i=1}^{N} 2x_i^2 \left(a_0\ x_i^2 + b_0\ x_i + c_0 - y_i\right) = 0 \\ \displaystyle\sum_{i=1}^{N} 2x_i \left(a_0\ x_i^2 + b_0\ x_i + c_0 - y_i\right) = 0 \\ \displaystyle\sum_{i=1}^{N} 2 \left(a_0\ x_i^2 + b_0\ x_i + c_0 - y_i\right) = 0. \end{cases}$$

Il s'agit d'un système de trois équations à trois inconnues (a_0, b_0 et c_0). Écrivons le plus clairement en factorisant par a_0, b_0 et c_0

$$\begin{cases} \left(\sum_{i=1}^{N} x_i^4\right) a_0 + \left(\sum_{i=1}^{N} x_i^3\right) b_0 + \left(\sum_{i=1}^{N} x_i^2\right) c_0 = \sum_{i=1}^{N} x_i^2 y_i \\ \left(\sum_{i=1}^{N} x_i^3\right) a_0 + \left(\sum_{i=1}^{N} x_i^2\right) b_0 + \left(\sum_{i=1}^{N} x_i\right) c_0 = \sum_{i=1}^{N} x_i y_i \\ \left(\sum_{i=1}^{N} x_i^2\right) a_0 + \left(\sum_{i=1}^{N} x_i\right) b_0 + \left(\sum_{i=1}^{N} 1\right) c_0 = \sum_{i=1}^{N} y_i, \end{cases}$$

En utilisant $\sum_{i=1}^{N} 1 = N$ et en introduisant les quantités suivantes, que l'on peut calculer facilement à partir des données,

$$S_x = \sum_{i=1}^{N} x_i, \quad S_y = \sum_{i=1}^{N} y_i, \quad S_{x^2} = \sum_{i=1}^{N} x_i^2, \quad S_{xy} = \sum_{i=1}^{N} x_i y_i,$$
$$S_{x^3} = \sum_{i=1}^{N} x_i^3, \quad S_{x^2y} = \sum_{i=1}^{N} x_i^2 y_i, \quad S_{x^4} = \sum_{i=1}^{N} x_i^4,$$

on peut mettre finalement le système sous la forme compacte

$$\begin{cases} S_{x^4}\ a_0 + S_{x^3}\ b_0 + S_{x^2}\ c_0 = S_{x^2y} \\ S_{x^3}\ a_0 + S_{x^2}\ b_0 + S_x\ c_0 = S_{xy} \\ S_{x^2}\ a_0 + S_x\ b_0 + N\ c_0 = S_y. \end{cases}$$

Utilisons par exemple la méthode matricielle pour résoudre ce système d'équations. Sous forme matricielle, le système d'équations se réécrit comme

$$\begin{pmatrix} S_{x^4} & S_{x^3} & S_{x^2} \\ S_{x^3} & S_{x^2} & S_x \\ S_{x^2} & S_x & N \end{pmatrix} \begin{pmatrix} a_0 \\ b_0 \\ c_0 \end{pmatrix} = \begin{pmatrix} S_{x^2y} \\ S_{xy} \\ S_y \end{pmatrix}.$$

Si le déterminant de la matrice 3×3 est non nul, on peut introduire l'inverse de la matrice, ce qui permet de résoudre le système d'équations sous la forme

$$\begin{pmatrix} a_0 \\ b_0 \\ c_0 \end{pmatrix} = \begin{pmatrix} S_{x^4} & S_{x^3} & S_{x^2} \\ S_{x^3} & S_{x^2} & S_x \\ S_{x^2} & S_x & N \end{pmatrix}^{-1} \begin{pmatrix} S_{x^2y} \\ S_{xy} \\ S_y \end{pmatrix},$$

où l'inverse de la matrice peut être calculé (de préférence avec l'aide d'un ordinateur !) pour chaque application après avoir remplacé les expressions par des valeurs numériques. Encore une fois, on pourrait vérifier en calculant les dérivées partielles secondes de F que ce point stationnaire (a_0, b_0, c_0) correspond bien à un minimum de la fonction F.

Appliquons ces expressions à l'exemple de la partie 2.1. On a $N = 11$ et on calcule $S_x = 55$, $S_y = 497$, $S_{x^2} = 385$, $S_{xy} = 3592$, $S_{x^3} = 3025$, $S_{x^2y} = 29212$, $S_{x^4} = 25333$. Cela conduit à $a_0 = 249/286 \approx 0.871$, $b_0 = 1941/1430 \approx 1.357$ et $c_0 = 103/13 \approx 7.923$. On a donc trouvé par régression polynomiale l'expression de la parabole approchant au mieux les données :

$$f(x) = 0.871\ x^2 + 1.357\ x + 7.923.$$

Cette parabole est tracée sur la figure 2.3.

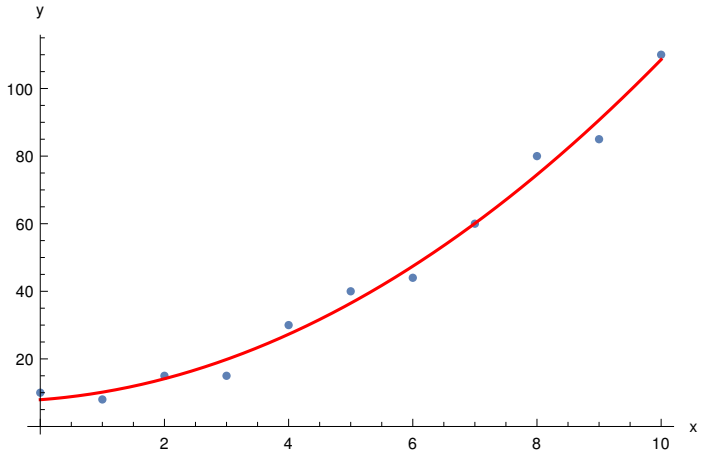

FIGURE 2.3 – Parabole déterminée par régression polynomiale approchant au mieux les données.

Pour mesurer la qualité de la régression, on peut calculer le coefficient de détermination R^2 (la définition reste la même que celle introduite plus haut). On trouve $R^2 \approx 0.989$. Cette valeur beaucoup plus proche de 1 que celle obtenue avec la régression linéaire confirme qu'une fonction polynomiale de degré 2 est un bien meilleur modèle pour nos données.

Corrigé des exercices

Exercice 5
Tout d'abord, on calcule $\ln c$:

t (s)	0	300	600	900	1200	1800	2400	3000
$\ln c$	0.4055	0.2151	0.0000	-0.1863	-0.3857	-0.7765	-1.1712	-1.5607

Utilisons les mêmes notations utilisées dans le cours : $x = t$ et $y = \ln c$. On veut déterminer par régression linéaire la fonction affine $f(x) = a_0 x + b_0$ approchant au mieux les données : $y \approx f(x)$. Le nombre de points est $N = 8$. On calcule $S_x = 10200$, $S_y \approx -3.4598$, $S_{xx} = 20700000$, $S_{xy} \approx -9456.4924$. Les paramètres a_0 et b_0 recherchés sont déterminés par le système d'équations

$$\begin{cases} 20700000\ a_0 + 10200\ b_0 = -9456.4924 \\ \\ 10200\ a_0 + 8\ b_0 = -3.4598. \end{cases}$$

Cela donne $a_0 = -0.0006557$ et $b_0 = 0.4035$. Revenant maintenant aux notations de l'exercice, la meilleure droite approchant au mieux les données est donc :

$$\ln c = -0.0006557\ t + 0.4035.$$

On a donc une estimation de la constante de vitesse $k \approx 0.0006557\ \text{s}^{-1}$ et du logarithme de la concentration initiale $\ln c(0) \approx 0.4035$, c'est-à-dire $c(0) \approx 1.4970$ mmol/L.
La droite de régression est comparée aux données sur la figure 2.4.

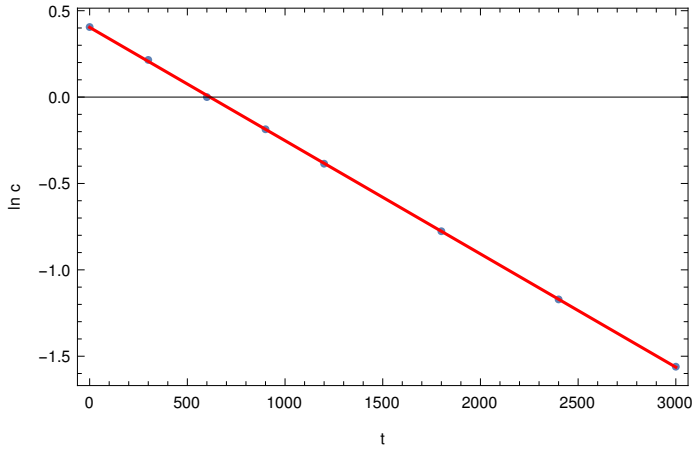

FIGURE 2.4 – Droite de régression $\ln c = -0.0006557\ t + 0.4035$ comparée aux données.

Graphiquement, la régression apparaît comme étant très bonne. Pour confirmer cela de manière plus quantitative, calculons à présent le coefficient de détermination R^2. On calcule $\bar{y} \approx -0.4325$, $\sum_{i=1}^{N} (y_i - f(x_i))^2 \approx 0.0001899$ et $\sum_{i=1}^{N} (y_i - \bar{y})^2 \approx 3.3082$, ce qui donne $R^2 \approx 0.99994$. Il s'agit bien d'une régression d'excellente qualité.
On conclut que les données sont parfaitement en accord avec un modèle cinétique d'ordre 1.