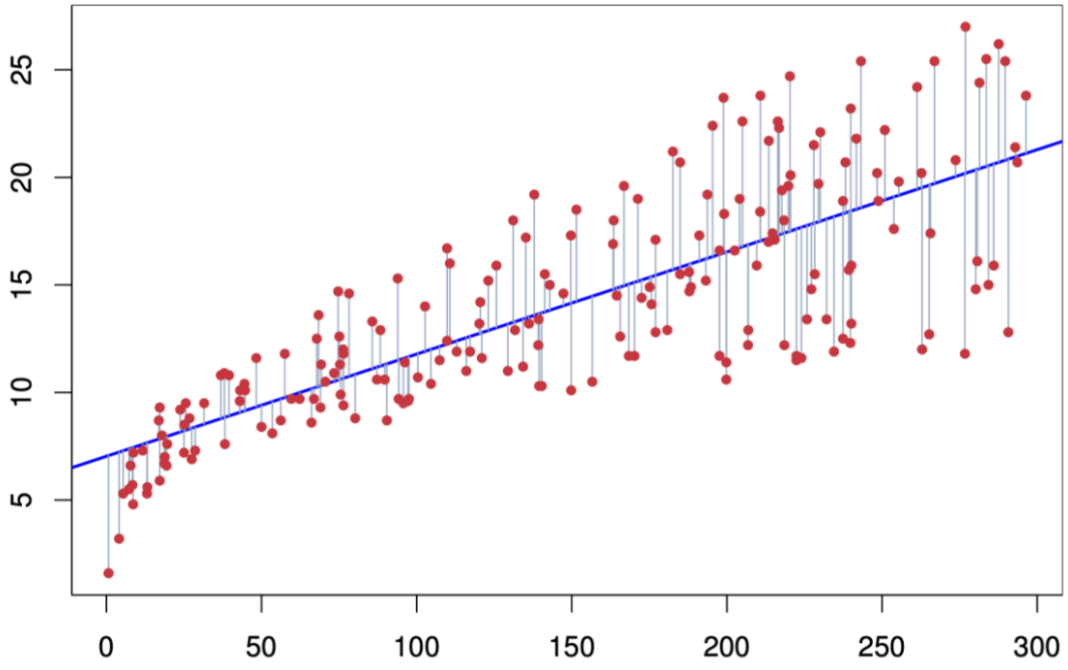
La Régression Linéaire

Travail réalisé par

Adam EL MOUEDDEN

29 octobre 2025



Projet de Mathématiques Appliquées

Analyse des méthodes de régression linéaire et applications pratiques

Table des matières

Introduction

La régression linéaire représente l'une des méthodes statistiques les plus fondamentales et répandues en analyse de données. Des laboratoires de recherche aux salles de marchés financiers, des usines industrielles aux hôpitaux, cet outil universel permet d'établir des relations quantitatives entre variables pour comprendre des phénomènes complexes et prédire des comportements futurs. Son application s'étend de la physique à l'économie, en passant par la médecine et l'ingénierie, faisant d'elle un pilier incontournable de la modélisation prédictive moderne.

L'objectif principal de ce travail est double :

— Comprendre les fondements mathématiques de la régression linéaire

— Citer l'unes de ses applications pratiques dans divers domaines scientifiques

1 Régression Linéaire Simple

1.1 Le but

On dispose d'un ensemble de n observations :

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

où :

— x_i est la variable explicative (ou indépendante),

— y_i est la variable à expliquer (ou dépendante).

On suppose que la relation entre x et y est approximativement linéaire, c'est-à-dire :

$$y_i \approx ax_i + b$$

où :

— a = pente (slope),

— b = ordonnée à l'origine (intercept).

Notre objectif est de trouver les valeurs de a et b qui « collent » le mieux aux données.

1.2 Modèle mathématique

On modélise :

$$y_i = ax_i + b + \varepsilon_i$$

avec :

— ε_i : l'erreur (ou résidu), supposée centrée (moyenne nulle).

$$\text{(forme matricielle : } \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \text{a.} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \text{)}$$

1.3 Méthode des moindres carrés (Least Squares)

On cherche à minimiser la somme des carrés des erreurs :

$$S(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 = \sum_{i=1}^{n} \epsilon_i^2$$

On veut :

$$\min_{a,b} S(a,b) \quad (= \min_{a,b} ||\epsilon||_2^2 \quad avec \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}))$$

1.4 Calcul analytique

On dérive $S(a,b)$ par rapport à a et b et on annule les dérivées :

$$\frac{\partial S}{\partial a} = -2 \sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0$$
$$\frac{\partial S}{\partial b} = -2 \sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

En simplifiant ces deux équations :

$$\sum y_i = a \sum x_i + nb$$
$$\sum x_i y_i = a \sum x_i^2 + b \sum x_i$$

C'est un système linéaire à deux inconnues a, b.

1.5 Résolution du système

En notant :

$$\bar{x} = \frac{1}{n} \sum x_i, \quad \bar{y} = \frac{1}{n} \sum y_i$$

On obtient les formules classiques :

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$
$$b = \bar{y} - a\bar{x}$$

1.6 Interprétation géométrique

La droite :

$$y = ax + b$$

est la droite qui minimise la distance verticale entre les points observés et la droite. Graphiquement, c'est la droite « la plus proche » des points selon le critère des moindres carrés.

1.7 Qualité de l'ajustement

On mesure la qualité de la régression par le coefficient de détermination R^2 :

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

où $\hat{y}_i = ax_i + b$.

— Si $R^2 = 1 \rightarrow$ ajustement parfait.

— Si $R^2 = 0 \rightarrow$ la droite ne prédit rien (moyenne constante).

2 Régression linéaire dans le cas général

2.1 Le but

Dans le cadre général de la régression linéaire, on dispose d'un ensemble de n observations

$$(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i), \quad i = 1, 2, \ldots, n,$$

où :

— y_i est la variable dépendante (ou à expliquer),

— x_{ij} désignent les variables explicatives (ou prédicteurs),

— p représente le nombre de variables explicatives.

On suppose qu'il existe une relation linéaire entre y_i et les x_{ij}, modélisée par :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

où :

— $\beta_0, \beta_1, \ldots, \beta_p$ sont les coefficients inconnus du modèle,

— ε_i est un terme d'erreur, supposé centré, c'est-à-dire $\mathbb{E}[\varepsilon_i] = 0$.

Sous forme matricielle, le modèle s'écrit :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

avec :

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

2.2 Méthode de résolution

L'objectif est de déterminer le vecteur $\boldsymbol{\beta}$ qui minimise la somme des carrés des erreurs :

$$S(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

En dérivant cette expression par rapport à $\boldsymbol{\beta}$, on obtient :

$$\nabla_{\boldsymbol{\beta}} S = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

Ce qui conduit aux **équations normales** :

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}.$$

Si la matrice $\mathbf{X}^T\mathbf{X}$ est inversible, la solution du problème des moindres carrés est donnée par :

$$\boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}.$$

2.3 Interprétation géométrique

Sur le plan géométrique, la régression linéaire correspond à la **projection orthogonale** du vecteur $\mathbf{y}$ sur le sous-espace vectoriel engendré par les colonnes de $\mathbf{X}$.

Ainsi :

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

est la projection de $\mathbf{y}$ sur l'espace colonne de $\mathbf{X}$, et le vecteur des résidus

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

est orthogonal à cet espace :

$$\mathbf{X}^T\mathbf{e} = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0.$$

2.4 Qualité de l'ajustement

La qualité de l'ajustement du modèle est évaluée à l'aide du **coefficient de détermination** R^2, défini par :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

où $\bar{y}$ est la moyenne des y_i.

- Si $R^2 = 1$, le modèle explique parfaitement la variabilité des données.
- Si $R^2 = 0$, le modèle n'explique aucune variabilité (les prédictions sont aussi bonnes que la moyenne).

On définit également les décompositions suivantes :

$$\text{SCT} = \sum_i (y_i - \bar{y})^2, \quad \text{SCE} = \sum_i (\hat{y}_i - \bar{y})^2, \quad \text{SCR} = \sum_i (y_i - \hat{y}_i)^2,$$

telles que :

$$\text{SCT} = \text{SCE} + \text{SCR}.$$

2.5 Cas particuliers

- Lorsque $p = 1$, on retrouve le cas de la **régression linéaire simple**, où la relation s'écrit $y = ax + b$.
- Si certaines colonnes de $\mathbf{X}$ sont fortement corrélées, on parle de **multicolinéarité**. Dans ce cas, la matrice $\mathbf{X}^T\mathbf{X}$ devient presque singulière, ce qui rend l'estimation de $\hat{\boldsymbol{\beta}}$ instable.
- Si $n < p$, le système est sous-déterminé : il n'existe pas de solution unique. On doit alors recourir à des méthodes de régularisation (comme la régression ridge ou LASSO).

3 Quelques applications de la régression linéaire :

3.1 En physique-chimie

Outil mathématique fondamental en sciences expérimentales, la régression linéaire permet d'identifier et de quantifier les relations entre grandeurs mesurées. Elle transforme des données expérimentales en lois empiriques, facilitant la détermination de constantes physiques et la validation de modèles théoriques.

3.1.1 Application 1 : Loi de Hooke et détermination de la constante de gravité

À titre d'exemple, considérons une expérience basée sur la **loi de Hooke**, où l'on mesure l'allongement F d'un ressort en fonction de la masse suspendue m.

L'expérience est inspirée de celle décrite sur le site du *Physics Classroom* : `https://www.physicsclassroom.com/class/waves/Lesson-0/Motion-of-a-Mass-on-a-Spring`.

En appliquant la régression linéaire aux données expérimentales (allongement en fonction de la masse), on obtient le graphe suivant :

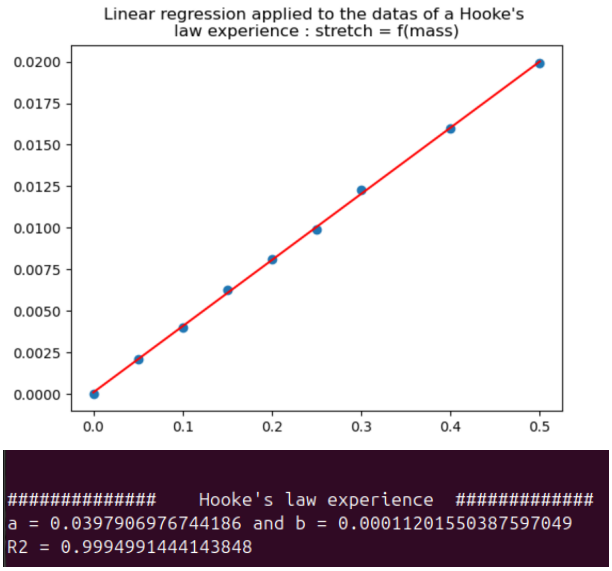


FIGURE 1 – Ajustement linéaire des données expérimentales selon la loi de Hooke

Avec un R2 proche de 1, la linéarité de la relation est clairement établie, validant la Loi de Hooke. Le modèle d'ajustement fournit une pente $a \approx 4 \times 10^{-2}$ et une ordonnée à l'origine b pratiquement nulle, comme attendu par la théorie. Cette pente nous permet d'extraire la constante de raideur du ressort.

3.2 Viscosité d'un liquide ionique : régression linéaire multidimensionnelle

Les données expérimentales ont été obtenues à partir du site `https://www.kaggle.com/datasets/davidescobargarcia/viscosidad-y-densidad-liquido-ionico-bmimpf6/code`. Elles correspondent aux mesures de viscosité (η), de température (T), de pression (P), de volume molaire (V) et de densité (ρ) d'un liquide ionique *[BMIM][PF6]*.

Une régression linéaire multidimensionnelle a été appliquée sur le logarithme de la viscosité en fonction des quatre variables expérimentales. On s'attend à une relation linéaire du type :

$$\ln(\eta) = \beta_0 + \beta_1 T + \beta_2 P + \beta_3 V + \beta_4 \rho$$

```
###############   Viscosity experience   #############

beta_hat = [[-1.49343768e+03]
 [-2.44778537e-02]
 [-1.58401352e-02]
 [ 3.46086222e+00]
 [ 5.70738044e+02]]
R2 = 0.9988391936225872
```

FIGURE 2 – Résultats de la régression linéaire multidimensionnelle appliquée aux données expérimentales.

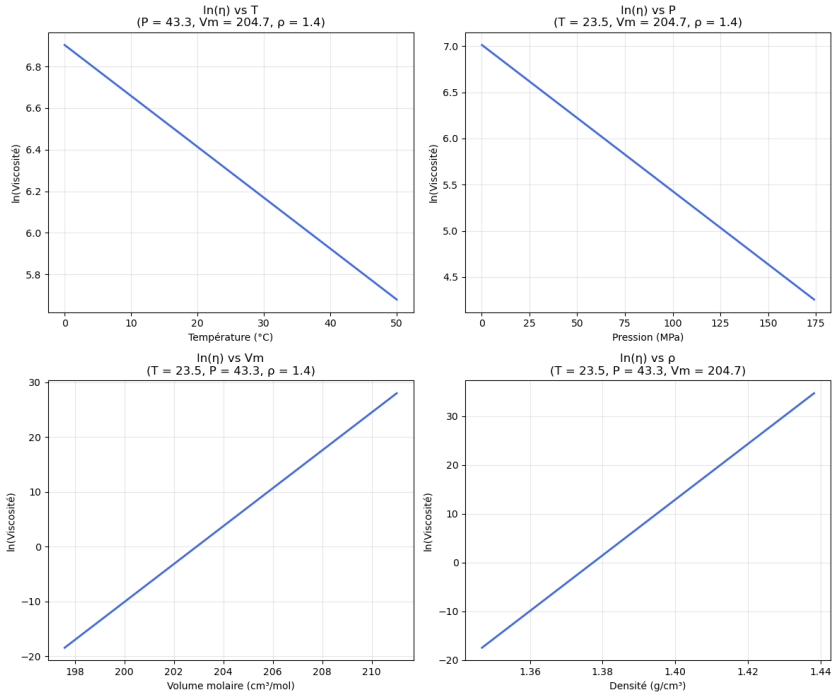


FIGURE 3 – ln(Viscosité) en fonction de chaque paramètre seul, en fixant les autres à leur moyenne.

Les coefficients $\boldsymbol{\beta}$ et le coefficient de détermination R^2 sont présentés sur la figure 3. On observe que $R^2 \approx 0.999$, ce qui indique un ajustement très bon. Les courbes de viscosité prédites en fonction de chaque variable, avec les autres fixées à leur moyenne, confirment la cohérence physique attendue : la viscosité diminue avec la température et la pression et augmente avec la densité et le volume molaire.

Ainsi, le modèle linéaire validé par les données expérimentales confirme la théorie physique de dépendance de la viscosité avec les paramètres thermodynamiques et structuraux du liquide ionique.