Analyse de données et apprentissage

Régression linéaire

El-Hadi Djermoune

9 février 2026

Université de Lorraine
Faculté des Sciences et Technologies
M1 Ingénierie de Systèmes Complexes

CRAN UL CNRS – département BioSiS
Bureau: 422, Bât. Henri Poincaré 4e étage
email: el-hadi.djermoune@univ-lorraine.fr

Table des matières

Objectifs

Régression

- Technique utilisée pour la modélisation et l'analyse de données numériques.
- Exploite la relation entre deux ou plusieurs variables de façon à avoir une information sur l'une d'entre elles en connaissant les autres.
- La régression peut être utilisée pour la prédiction, l'estimation, le test d'hypothèses et la modélisation de relations causales.

Objectifs du chapitre

- Comprendre les concepts de base de la régression linéaire du point de vue probabiliste.
- Aborder les principes de l'estimation de paramètres d'un modèle linéaire.

Régression linéaire : pourquoi ?

Vocabulaire

$$\underbrace{Y}_{\text{Variable dépendante}} = \underbrace{X_1 + X_2 + X_3}_{\text{Variables indépendante}}$$

Variable dépendante	Variables indépendante
Variable de bilan	Variables de prédiction
Variable de réponse	Variables explicatives

Pourquoi la régression linéaire ?

- Supposons que l'on souhaite modéliser la variable dépendante Y en termes de trois prédicteurs X_1, X_2 et X_3 :

$$Y = f(X_1, X_2, X_3).$$

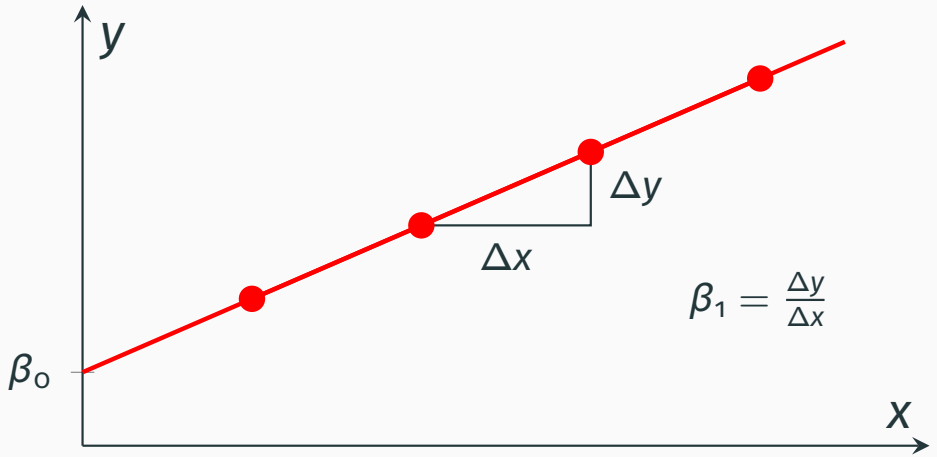
- Généralement, nous n'avons pas suffisamment de données pour estimer directement la fonction f.
- Par conséquent, nous devons supposer que f a une forme simple, comme une relation linéaire :

$$Y = X_1 + X_2 + X_3.$$

Régression linéaire = modèle probabiliste

- Un pan important des mathématiques est dédié à l'étude des variables reliées de façon déterministe :

$$y = \beta_0 + \beta_1 x$$



- Nous sommes intéressés par la compréhension des relations entre variables reliées de façon non déterministes.

Définition d'un modèle linéaire probabiliste

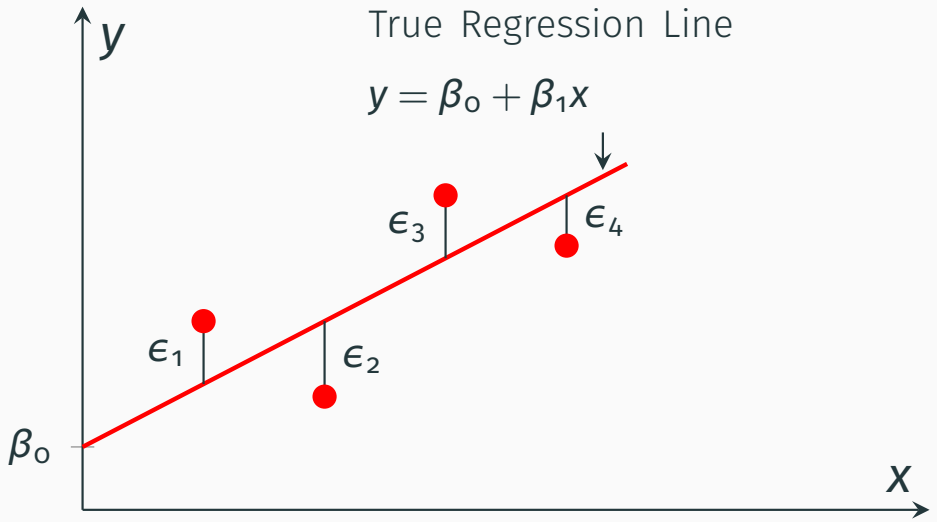
Il existe des paramètres β_0, β_1 et σ^2 tels que, pour toute valeur de x, la variable dépendante y est reliée à x par le modèle :

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

où ϵ est une v.a. supposée suivre une loi normale $\mathcal{N}(0, \sigma^2)$.

Régression linéaire = modèle probabiliste

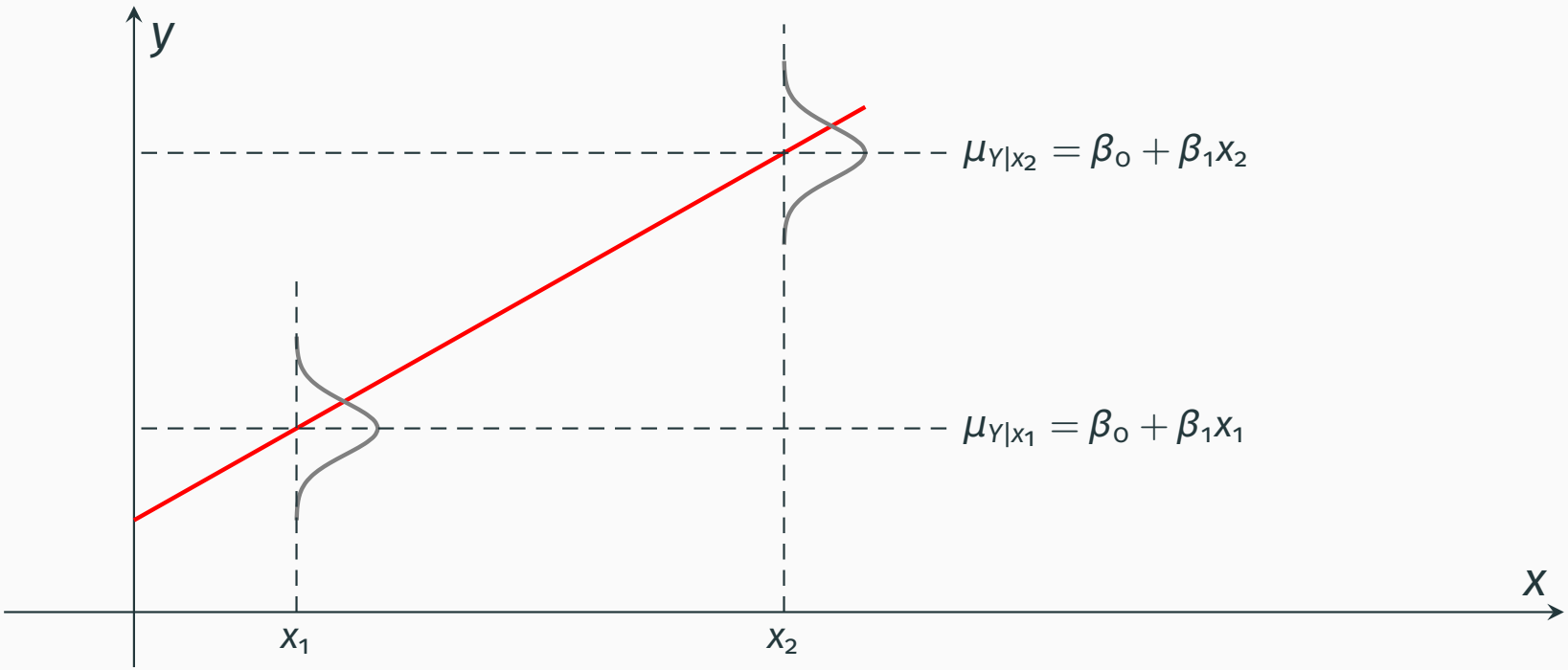
- Modèle : $y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$.



- La valeur moyenne de Y est une fonction linéaire de X, mais pour une valeur donnée x, la variable Y est écartée de sa valeur moyenne d'une quantité aléatoire.
- Formellement, si on note x^* une valeur particulière de la variable indépendante x, notre modèle linéaire dit :

$$\mathbb{E}\{Y|x^*\} = \mu_{Y|x^*} = \text{valeur moyenne de } Y \text{ quand } x = x^*$$
$$\text{var}\{Y|x^*\} = \sigma^2_{Y|x^*} = \text{variance de } Y \text{ quand } x = x^*$$

Interprétation graphique

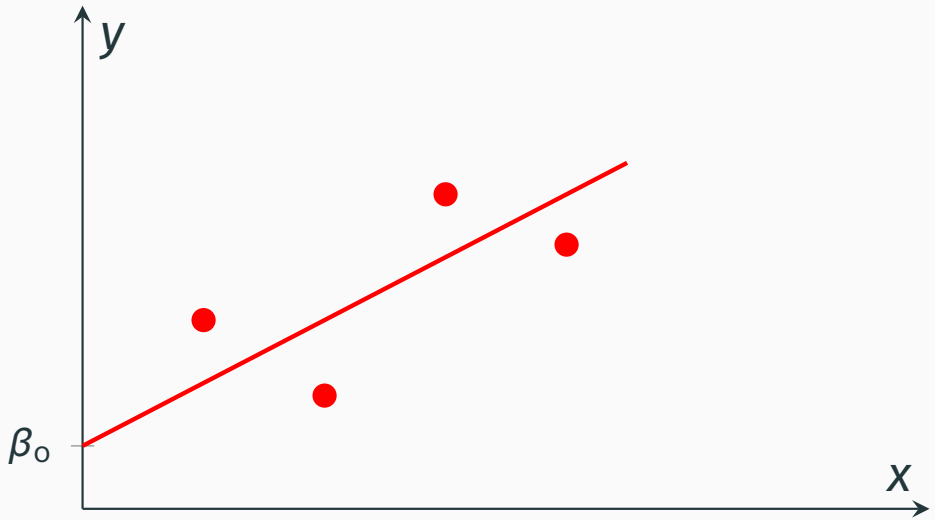


- Par exemple, si x est une taille et y est un poids, $\mu_{Y|x=150}$ est le poids moyen de tous les individus de taille 150 cm dans la population.

Estimation des paramètres du modèle

- Les valeurs estimés $\hat{\beta}_0$ et $\hat{\beta}_1$ peuvent être obtenues en utilisant le principe des moindres carrés :

$$J(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2.$$



Le minimum est atteint pour :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

où $\bar{x}$ et $\bar{y}$ sont les valeur moyennes des suites x_n et y_n, et

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

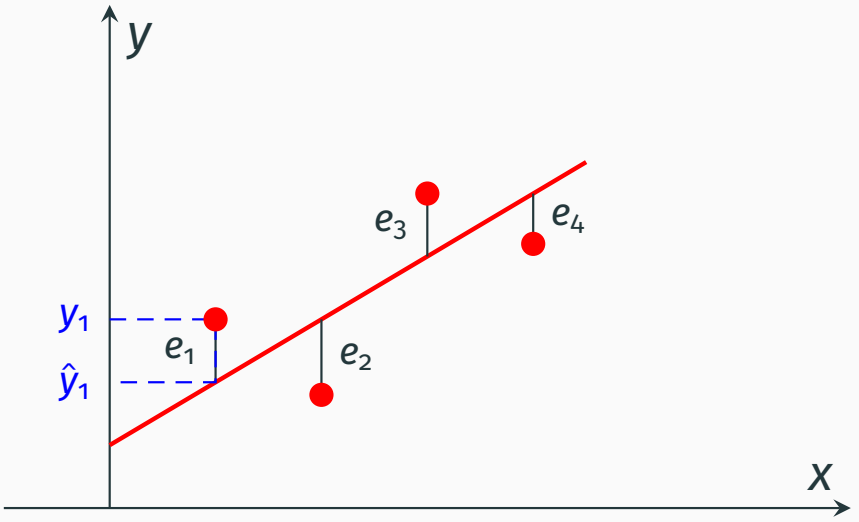
Valeur prédite et valeur résiduelle

- Les valeurs prédites, ou ajustées, sont les valeurs de y calculées par la droite obtenue par les moindres carrés en remplaçant x par $x_1, \ldots, x_n$:

$$\hat{y}_1 = \hat{\beta}_0 - \hat{\beta}_1 x_1$$
$$\hat{y}_2 = \hat{\beta}_0 - \hat{\beta}_1 x_2.$$

- Les résidus sont les écarts entre les valeurs observées et prédites :

$$e_1 = y_1 - \hat{y}_1$$
$$e_2 = y_2 - \hat{y}_2$$



Les résidus sont utiles!

- Ils permettent quantifie la **variance** de l'erreur d'**approximation** de la variable dépendante (SSE = *sum of squares error*) :

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

  c'est l'erreur non expliquée par la régression.
- ... qui, à son tour, permet d'estimer σ^2 (la **variance du bruit** ϵ) :

$$\hat{\sigma}^2 = \frac{SSE}{n-2},$$

- ... et une statistique importante, appelée **coefficient de détermination**, qui mesure la qualité du modèle en prédiction (1 = modèle parfait) :

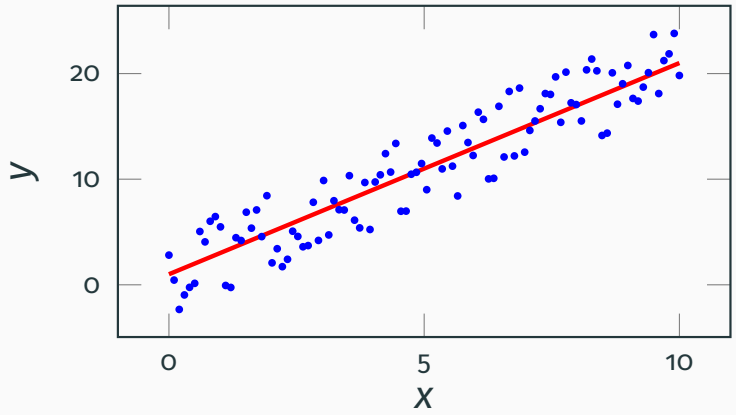
$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x}_i)(y_i - \bar{y}_i)\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

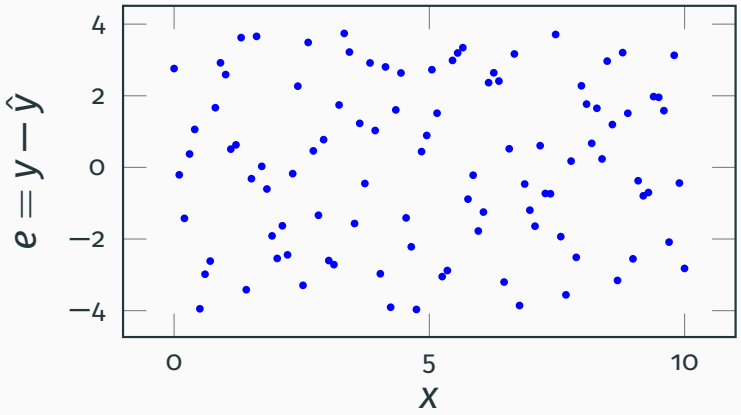
  où SST= *sum of squares total*.
- L'erreur totale (SST) est la somme de SSE et l'erreur de expliquée par la régression (SSR = *sum of squares due to regression*) :

$$SST = SSE + SSR, \qquad SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

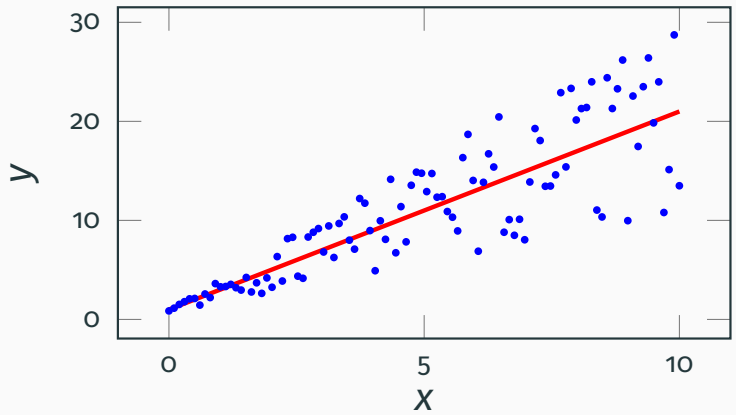
Les résidus sont utiles !

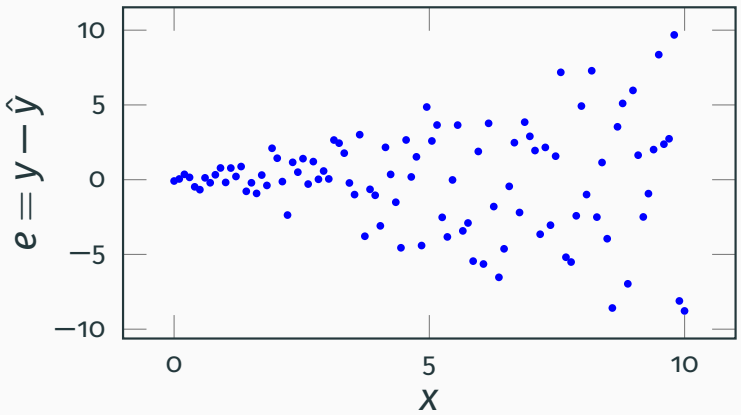
- **Homoscédasticité :** c'est une propriété fondamentale du modèle de la régression linéaire et fait partie de ses hypothèses de base. Elle signifie que la variance des erreurs est la même pour chaque observation ***i***.





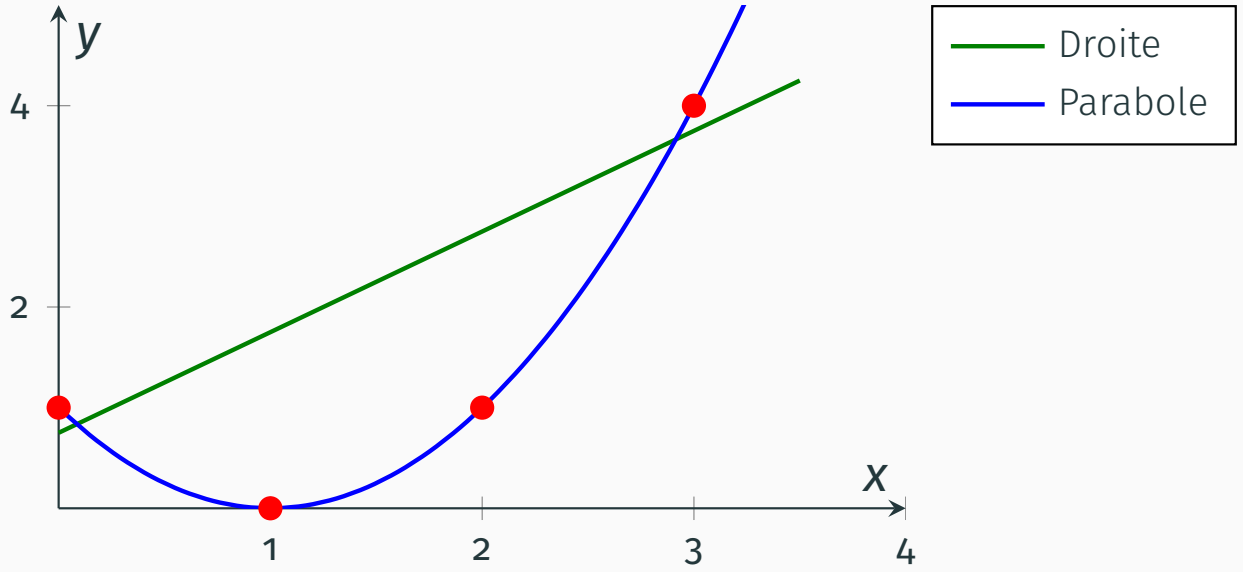
- **Hétéroscédasticité :** la variance (dispersion verticale) change comme le montre la figure ci-dessous. Avant d'utiliser les moindres carrés, les données doivent être transformées pour les rendre homoscédastiques.





La régression linéaire pour ajuster des polynômes

- Nous avons vu que la régression linéaire peut être utilisée pour **ajuster une droite** aux données.
- La régression linéaire peut également être utilisée pour **ajuster un polynôme**.



Ajustement par moindres carrés

Que signifie le mot « linéaire »

Le terme « linéaire » dans la régression linéaire signifie que le modèle est linéaire par rapport aux paramètres β_i, i.e. chaque β_i a une puissance 1.

Table des matières

Moindres carrés ordinaires : régression linéaire simple

- On se propose maintenant d'estimer les paramètres du modèle de régression linéaire simple (droite) :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \forall i \in \{1, \dots, n\}.$$

$\beta_0 =$ **ordonnée à l'origine** (*intercept*) ; $\beta_1 =$ **pente** (*slope*).

- Sous une forme vectorielle, nous pouvons écrire :

$$Y = X\beta + \epsilon, \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

- Le but est de trouver le vecteur β telle que le coût quadratique suivant soit minimal :

$$J(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = ||Y - X\beta||_2^2, \qquad ||Y||_2^2 := \sum_{i=1}^{n} y_i^2.$$

Autrement dit, la droite des moindres carrés (*least squares*) minimise la somme des carrés des distances verticales des points (x_i, y_i) du nuage à la droite ajustée $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

Moindres carrés ordinaires : régression linéaire simple

Optimalité

Comme J est une fonction strictement convexe, il existe une seul point (β_0, β_1) qui minimise J. Les valeurs optimales $\hat{\beta}_0$ et $\hat{\beta}_1$ sont celles qui annulent le gradient de J :

$$\frac{\partial J}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{2}$$

$$\frac{\partial J}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \tag{3}$$

- La première équation donne :

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \Longrightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- La deuxième équation donne :

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \quad \Longrightarrow \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Remarques

- La relation $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ montre que la droite des MCO[1] passe par le centre de gravité du nuage $(\bar{x}, \bar{y})$.
- Les expressions obtenues pour $\hat{\beta}_0$ et $\hat{\beta}_1$ montrent que ces deux estimateurs sont linéaires par rapport au vecteur $Y = [y_1, \ldots, y_n]^\top$.
- L'estimateur $\hat{\beta}_1$ peut aussi s'écrire comme suit (exercice !) :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \epsilon_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

1. MCO : moindres carrés ordinaires.

Moindres carrés ordinaires : régression linéaire simple (propriétés)

Sous les hypothèses de centrage ($\mathbb{E}\{\epsilon_i\} = 0$), de décorrélation ($\mathbb{E}\{\epsilon_i \epsilon_j\} = 0, i \neq j$) et d'homoscédasticité ($\text{var}\{\epsilon_i\} = \sigma^2, \forall i$) des erreurs ϵ_i, les propriétés suivantes sont vérifiées.

Propriétés

- $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 et β_1 :

$$\mathbb{E}\{\hat{\beta}_0\} = \beta_0, \quad \mathbb{E}\{\hat{\beta}_1\} = \beta_1.$$

- Les variances de $\hat{\beta}_0$ et $\hat{\beta}_1$ et leur covariance s'écrivent :

$$\text{var}\{\hat{\beta}_0\} = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} = \sigma^2 \left(\frac{1}{n} + \frac{\hat{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right),$$

$$\text{var}\{\hat{\beta}_1\} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \text{cov}\{\hat{\beta}_0, \hat{\beta}_1\} = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- La statistique

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n-2}$$

est un estimateur sans biais de σ^2.

- L'erreur de prévision à un pas $\hat{\epsilon}_{n+1} = (y_{n+1} - \hat{y}_{n+1})$ satisfait les propriétés suivantes :

$$\begin{cases} \mathbb{E}\{\hat{\epsilon}_{n+1}\} = 0, \\ \text{var}\{\hat{\epsilon}_{n+1}\} = \sigma^2 \left(1 + \frac{1}{n} + \frac{x_{n+1} - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \end{cases}$$

Moindres carrés ordinaires : régression linéaire simple (supplément)

Différentiation vectorielle

On peut calculer la différentielle de J par rapport au vecteur β en utilisant les expressions :

$$\frac{\partial A\beta}{\partial \beta} = A^\top \tag{4}$$

$$\frac{\partial \beta^\top A\beta}{\partial \beta} = (A + A^\top)\beta. \tag{5}$$

- Fonction de coût :

$$J(\beta) = ||Y - X\beta||_2^2 = (Y - X\beta)^\top (Y - X\beta) = Y^\top Y - 2Y^\top X\beta + \beta^\top X^\top X\beta.$$

- Différentielle :

$$\frac{\partial J}{\partial \beta} = -2X^\top Y + 2X^\top X\beta = 0.$$

- Vecteur optimal :

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

La matrice $(X^\top X)^{-1} X^\top$ est appelée pseudo-inverse de Moore-Penrose de X.

Régression linéaire simple : exemple

Nombre de divorces en Angleterre et Pays de Galles

Année	1975	1976	1977	1978	1979	1980
Divorces (milliers)	120.5	126.7	129.1	143.7	138.7	148.3

- Modèle (y_i = nombre de divorces, x_i = années) :

$$y_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, 6.$$

- Solution des moindres carrés ordinaires :

$$\beta_0 = -10578, \qquad \beta_1 = 5.42.$$

- Avec ce modèle, on peut prédire le nombre de divorces en 1981 : $\hat{\beta}_0 + \hat{\beta}_1 \cdot 1980 = 153.5$ milliers de divorces.

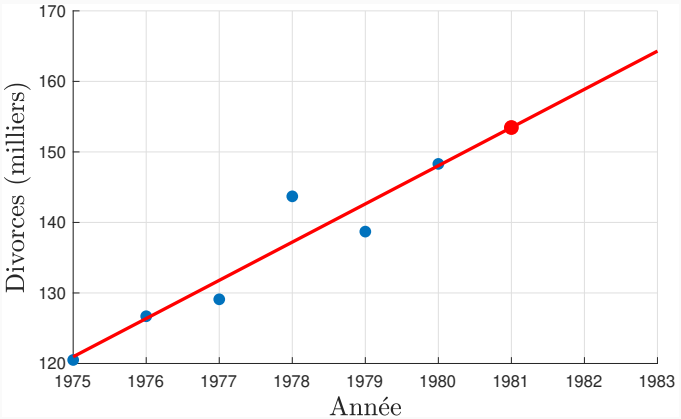


Table des matières

Régression linéaire multiple

- Extension à plus de variables indépendantes :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \ldots, n.$$

Modélisation de la concentration d'ozone dans l'atmosphère (Rennes)

T_{12}	23.8	16.3	27.2	7.1	25.1	27.5	19.4	19.8	32.2	20.7
V	9.25	-6.15	-4.92	11.57	-6.23	2.76	10.15	13.5	21.27	13.79
N_{12}	5	7	6	5	2	7	4	6	1	4
O_3	115.4	76.8	113.8	81.6	115.4	125	83.6	75.2	136.8	102.8

Tab. 10 données journalière de température, de vent, de nébulosité et ozone

T_{12} : température à midi
V : vitesse du vent dans la direction Est-Ouest
N_{12} : nébulosité[2] à midi
O_3 : ozone

$$O_3(i) = \beta_0 + \beta_1 T_{12}(i) + \beta_2 V(i) + \beta_3 N_{12}(i) + \epsilon(i).$$

- Les coefficients partiels de régression β_i quantifient l'effet de l'une des variables indépendantes sur la variable dépendante, en maintenant les autres prédicteurs constants.

2. La nébulosité est la proportion de nuages couvrant le ciel. Elle se mesure en ≪ huitième ≫ du ciel observé (≪ octa ≫), par estimation de la partie couverte.

Régression linéaire multiple

- Représentation vectorielle de la régression :

$$y_i = [1, \quad x_{i1}, \quad \ldots, \quad x_{ip}] \cdot \beta + \epsilon_i.$$

$$Y = X\beta + \epsilon, \qquad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

- Fonction de coût à minimiser (moindres carrés ordinaires) :

$$J(\beta) = ||Y - X\beta||_2^2. \tag{6}$$

- Solution optimale (correspondant à $\frac{\partial J}{\partial \beta} = 0$), en supposant $n \geq p + 1$:

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y. \tag{7}$$

- **Remarques :**
  - les paramètres optimaux sont donnés avec une forme explicite !
  - la solution dans (7) existe si X est de rang plein colonne.
  - si les colonnes de X sont linéairement dépendantes ($\text{rang}(X) < p + 1$), le problème (6) est **mal posé**. Il faut alors recourir à une **régularisation** qui consister à ajouter des contraintes ou des pénalités à la fonction de coût.

Régression linéaire multiple : moindres carrés régularisés

- La régularisation est une technique qui permet de **réduire la variance** du vecteur estimé au prix d'une **augmentation du biais**.
- Il existe plusieurs façon pour régulariser un critère
  - en ajoutant des contraints ($c_1, c_2 \in \mathbb{R}_+$)

$$\text{minimiser } J(\beta) = ||Y - X\beta||_2^2 \quad \text{s.c.} \quad ||\beta||_2^2 < c_2 \qquad \text{(Tikhonov)}$$

$$\text{minimiser } J(\beta) = ||Y - X\beta||_2^2 \quad \text{s.c.} \quad ||\beta||_1 < c_1 \qquad \text{(Lasso)}$$

  - en ajoutant des pénalités ($\lambda_1, \lambda_2 \in \mathbb{R}_+$)

$$\text{minimiser } J(\beta) = ||Y - X\beta||_2^2 + \lambda_2 ||\beta||_2^2 \qquad \text{(Tikhonov)}$$

$$\text{minimiser } J(\beta) = ||Y - X\beta||_2^2 + \lambda_1 ||\beta||_1 \qquad \text{(Lasso)}$$

$$\text{minimiser } J(\beta) = ||Y - X\beta||_2^2 + \lambda_2 ||\beta||_2^2 + \lambda_1 ||\beta||_1 \qquad \text{(Elastic net)}$$

- Les formes contraintes et pénalisées sont équivalentes (les solutions des 2 problèmes d'optimisation sont identiques).
- La solution induite par la régularisation de Tikhonov est appelée *Ridge Regression*.
- Les solutions induites par les normes 2 (Tikhonov) et 1 (Lasso [3]) ont des propriétés différentes : Tikhonov pénalise les grandes valeurs des β_i et Lasso pénalise le nombre de β_i non nuls (parcimonie).

3. *Least absolute shrinkage and selection operator.*

Régression linéaire multiple : moindres carrés régularisés

Exemple :

- Régularisation de Tikhonov :

$$J(\beta) = ||Y - X\beta||_2^2 + \lambda ||\beta||_2^2$$

- Solution optimale :

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top Y. \tag{8}$$

- La solution optimale possède une forme explicite [4].
- La valeur de λ est choisie par l'utilisateur :
  - une petite valeur engendre un petit bias mais la solution obtenue risque d'être instable ;
  - une grande valeur engendre une plus grande stabilité mais avec un grand biais.

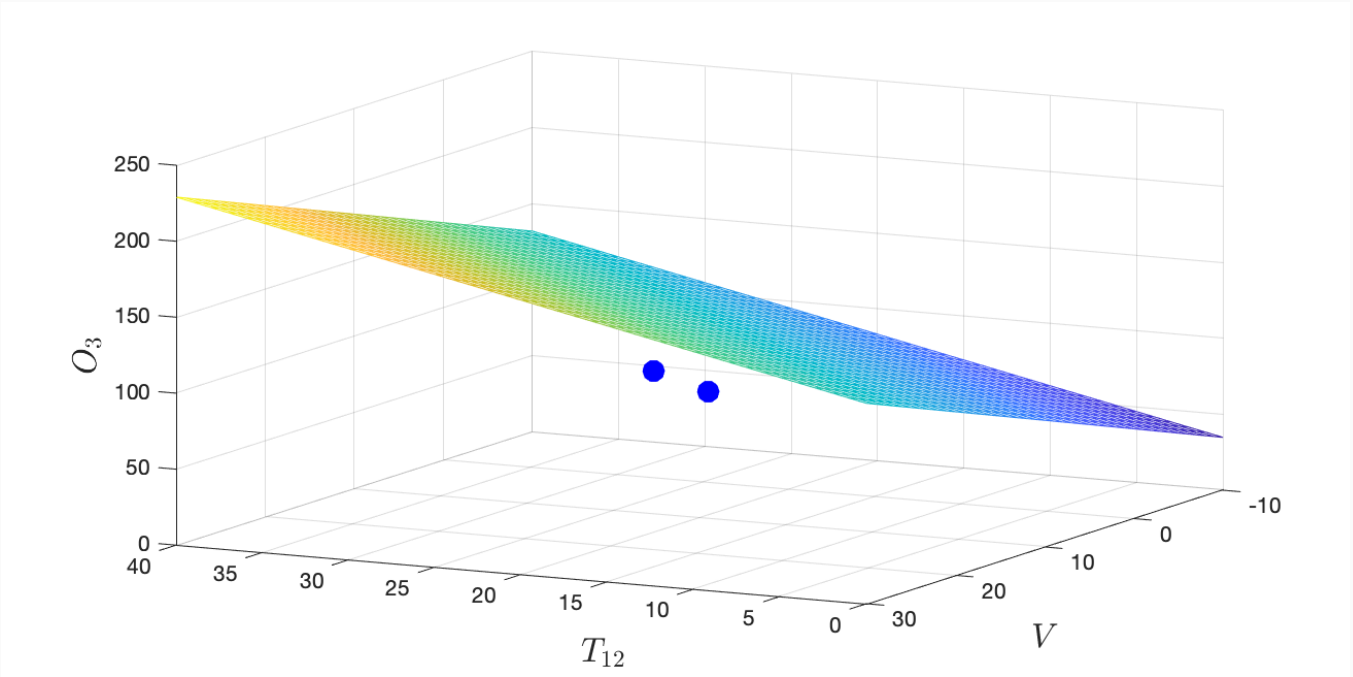
4. Ceci n'est pas le cas si l'on remplace la norme 2 par la norme 1, par exemple.

Régression linéaire multiple : exemple

Modélisation de la concentration d'ozone

Les paramètres du modèle de régression linéaire avec 3 variables explicatives sont :

$$\hat{\beta}_0 = 59.1, \quad \hat{\beta}_1 = 2.43, \quad \hat{\beta}_2 = -0.013, \quad \hat{\beta}_3 = -2.04.$$



Résultat de l'ajustement (O_3) pour $N_{12} = 4$ et différentes valeurs de T_{12} et V. Les 2 points bleus correspondent aux valeurs mesurées ayant contribué à l'ajustement du modèle

Table des matières

Régression linéaire – ajustement de polynômes

- Le modèle polynomial s'écrit :

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \epsilon_i.$$

- Ecriture vectorielle (X est une matrice de Vandermonde) :

$$Y = X\beta + \epsilon, \qquad X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ 1 & x_2 & \cdots & x_2^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^p \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

- Ajustement au sens des moindres carrés :

$$J(\beta) = ||Y - X\beta||_2^2.$$

- Solution optimale (en supposant $n \geq p + 1$) :

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

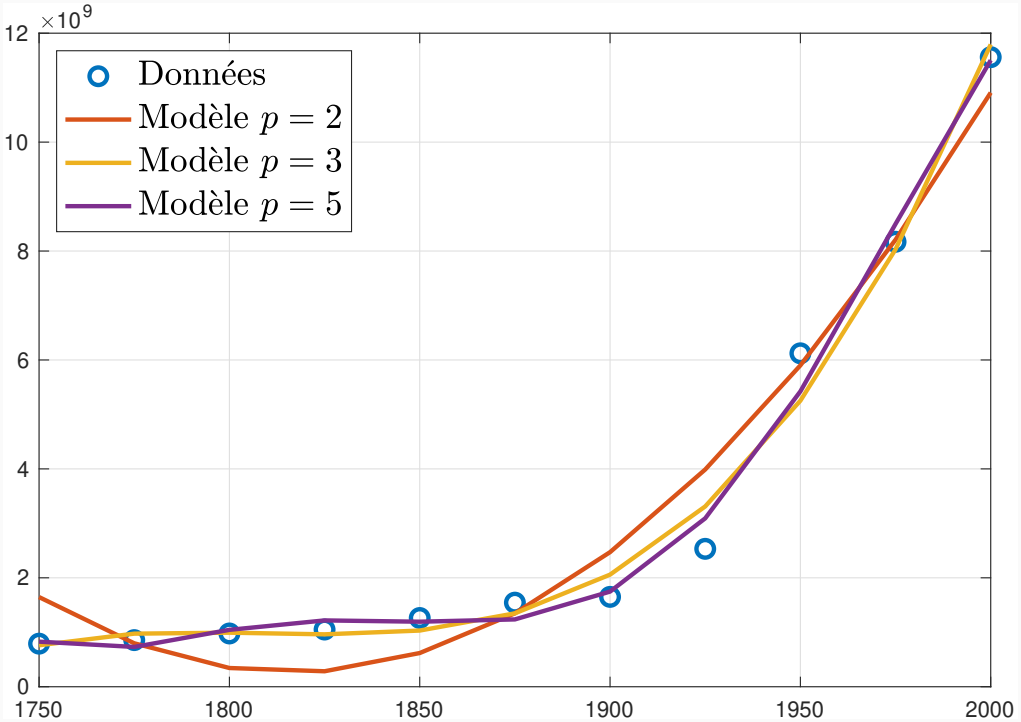
- Si les éléments de X sont grands, il vaut mieux normaliser[5] les données pour éviter de créer une matrice **mal conditionnée**.

5. La normalisation consiste à remplacer les x_i par $(x_i - \bar{x})/\sigma$, où σ est l'écart-type des x_i.

Régression linéaire – ajustement de polynômes : exemple

Evolution de la population mondiale (valeurs un peu exagérées)

Année	1750	1775	1800	1825	1850	1875	1900	1925	1950	1975	2000
Pop. (mill.)	791	856	978	1050	1262	1544	1650	2532	6122	8170	11560

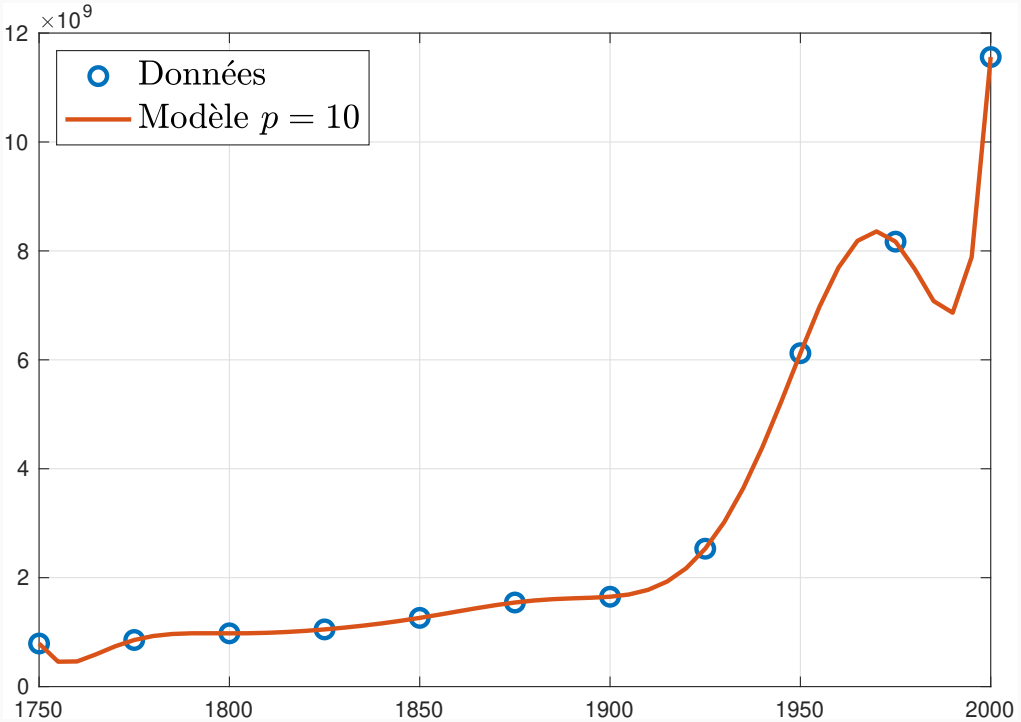


Résultat de l'ajustement pour quelques valeurs de ***p***

Régression linéaire : attention au sur-apprentissage !

Evolution de la population mondiale valeurs un peu exagérées...

Année	1750	1775	1800	1825	1850	1875	1900	1925	1950	1975	2000
Pop. (mill.)	791	856	978	1050	1262	1544	1650	2532	6122	8170	11560



Résultat de l'ajustement pour $p = 10$. **Règle générale** $p \ll n$