Projet MAM3 : Régression linéaire

Ben Khalifa Emna, Costantin Perline, Honakoko Giovanni

26/05/2025

La régression linéaire est un type de modélisation mathématique permettant de faire des prédictions *(droite rouge)* à partir d'un jeu de données *(nuage de points)* réelles.

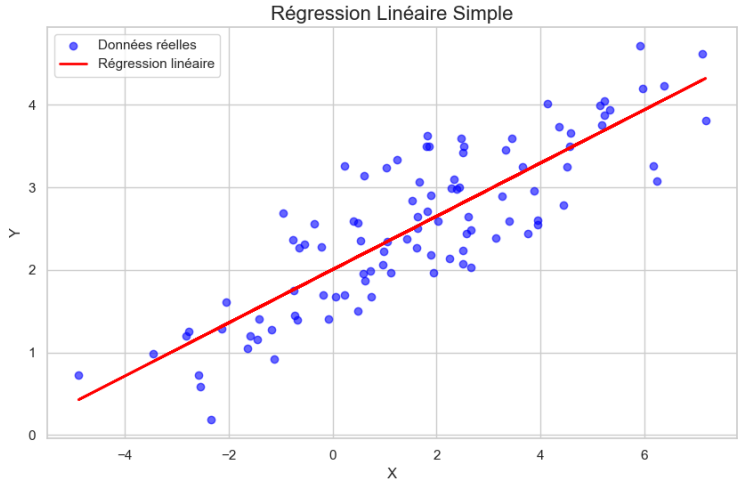

Table des matières

1 Modèle de régression linéaire simple

1.1 Cadre

On se place dans le cadre de la régression linéaire simple où on a une variable réponse et une variable explicative qui sont quantitatives. On dispose de $\mathcal{L} := \left\{(x_i, y_i)_{i\in[\![1,n]\!]}\right\}$ où :

- i représente l'individu considéré
- x_i représente les observations de la variable explicative
- y_i représente les observations de la variable réponse

On cherche f la fonction telle que : $\forall i \in [\![1, n]\!], y_i \approx f(x_i)$. Afin d'estimer f, on veut minimiser une quantité que l'on appelle risque quadratique :

$$R(\tilde{f}) := \mathbb{E}\left[\left(Y - \tilde{f}(X)\right)^2\right]$$

où Y est la variable réponse et X est la variable explicative. La quantité R mesure l'erreur quadratique moyenne entre les valeurs observées et les valeurs prédites.
Cette quantité étant purement théorique, voir quasiment inaccessible en réalité, on l'a substitue à son équivalent empirique :

$$R_n(\tilde{f}) := \frac{1}{n}\sum_{i=0}^{n}\left(Y_i - \tilde{f}(X_i)\right)^2$$

Comme nous l'avons vu dans la vulgarisation en première page, on souhaite que nos prédictions suivent une droite affine. Donc on restreindra $\tilde{f}$ à l'ensemble $\mathcal{F} = \{g : \mathbb{R} \longrightarrow \mathbb{R},\ \ g(x) = ax + b, \forall a, b \in \mathbb{R}\}$.
Dans notre cadre $Y_i = ax_i + b + \varepsilon_i$ où ε_i représente le bruit pour l'individu i.
Les Y_i et les ε_i sont des quantités aléatoires contrairement aux x_i qui sont fixes.

On se ramène à un problème de moindres carrés où nous voulons minimiser la distance qu'il y a entre nos points et l'espace affine engendré par $\tilde{f}$, c'est à dire minimiser $\sum_{i=1}^{n} x_i\left(Y_i - \tilde{f}(x_i)\right)^2 = \sum_{i=1}^{n} x_i(Y_i - ax_i - b)^2$.

1.2 Estimation paramétrique

On pose :

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

On cherche les points critiques :

$$\begin{cases} \dfrac{\partial R_n(g)}{\partial a} = 0 \\ \dfrac{\partial R_n(g)}{\partial b} = 0 \end{cases}$$

$$\begin{cases} -\dfrac{2}{n}\displaystyle\sum_{i=1}^{n} x_i(Y_i - ax_i - b) = 0 \\ -\dfrac{2}{n}\displaystyle\sum_{i=1}^{n} (Y_i - ax_i - b) = 0 \end{cases}$$

$$\begin{cases} \displaystyle\sum_{i=1}^{n} x_i Y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \\ b = \overline{Y}_n - a\overline{b}_n \end{cases}$$

En réinjectant l'expression de b dans la première ligne on obtient :

$$\begin{aligned} \sum x_i Y_i - \overline{Y}_n \sum_{i=1}^{n} x_i - a\overline{x}_n \sum_{i=1}^{n} x_i = 0 & \\ \sum_{i=1}^{n} x_i Y_i - \overline{Y}_n \sum_{i=1}^{n} x_i = a\left[\sum_{i=1}^{n} x_i^2 - \overline{x}_n \sum_{i=1}^{n} x_i\right] & \\ \sum_{i=1}^{n} x_i Y_i - n\overline{y}_n \overline{x}_n = a\left[\sum_{i=1}^{n} x_i^2 - n(\overline{x}_n)^2\right] & \\ a = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{Y}_n \overline{x}_n}{\sum_{i=1}^{n} x_i^2 - n(\overline{x}_n)^2} & \end{aligned}$$

Ainsi on pose :

$$\begin{cases} \hat{a}_n = \dfrac{\sum_{i=1}^{n} x_i Y_i - n\overline{Y}_n \overline{x}_n}{\sum_{i=1}^{n} x_i^2 - n(\overline{x}_n)^2} \\ \hat{b}_n = \overline{Y}_n - a\overline{x}_n \end{cases}$$

■

Proposition

Les estimateurs $\hat{a}_n$ et $\hat{b}_n$ sont sans biais.
Preuve
■

1.3 Propriétés des estimateurs $\hat{a}_n$ et $\hat{b}_n$

Pour nos estimateurs $\hat{a}_n$ et $\hat{b}_n$ on a :

- $\mathbb{V}\left[\hat{b}_n\right] = \dfrac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$ et $\mathbb{V}\left[\hat{a}_n\right] = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$
- $\mathrm{cov}(\hat{a}_n, \hat{b}_n) = \dfrac{-\sigma^2 \overline{x}_n}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$

Preuve

■

Or on ne connaît pas non plus la valeur de σ^2 ce qui nous pousse à utiliser un estimateur d'un tel paramètre. Un estimateur sans biais de cette quantité est $\hat{\sigma}_n^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$, avec $\hat{Y}_i = \hat{a}_n x_i + \hat{b}_n$ la prédiction des Y_i par le modèle de régression linéaire.

1.4 Validité de la régression linéaire

On appelle coefficient de détermination R^2 la quantité définie par :

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}_n\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - Yi\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2} = \left(\mathrm{corr}\left(Y_i, \hat{Y}_i\right)\right)^2$$

Ainsi définit cette quantité mesure la variance des résidus par rapport à la variance notre jeu de donnée et ainsi relate de la dispersion des observations par rapport à la droite de régression.
La validation de notre modèle de régression linéaire est déterminée par les conditions :

- $R^2 \in [0, 1]$
- Test paramétrique sur $\boldsymbol{a}$ concluant sur $\mathcal{H}_1$

Plus R^2 sera proche de 1, plus notre modélisation sera juste. Cependant il est important de pouvoir supposer la normalité du bruit en amont de ces vérifications. De plus cette hypothèse n'est pas vérifiée pour le coefficient de détermination :

2 Modèle de régression linéaire multiple

3 Analyse de la variance