Projet MAM3 : Régression linéaire

Ben Khalifa Emna, Costantin Perline, Honakoko Giovanni

26/05/2025

Table des matières

1 Théorie

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 : $\forall i \in [\![1, n]\!], y_i \approx f(x_i)$. Pour estimer la fonction f on veut minimiser le risque quadratique :

$$R(g) := \mathbb{E}[(Y - g(X))^2]$$

où Y est la variable réponse et X est la variable explicative. Une estimation de f est :

$$f^* := \underset{g}{\operatorname{argmin}}(g)$$

Cette quantité étant purement théorique on l'a substitue à sa quantité empirique le risque empirique $R_n(g)$ qui s'exprime comme :

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

On supposera que g appartient 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.

1.2 Estimateurs paramétriques $\hat{a}_n$ et $\hat{b}_n$

On pose :

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

Pour cela on cherche les points critiques :

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

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

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

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

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

On a bien :

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

■

Proposition

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