Math PT. 2025 - 2026

Mathématiques & Informatique - F. Gaunard
http://frederic.gaunard.com
Lycée Voltaire, Paris 11e.

X
Bonus : Régression linéaire

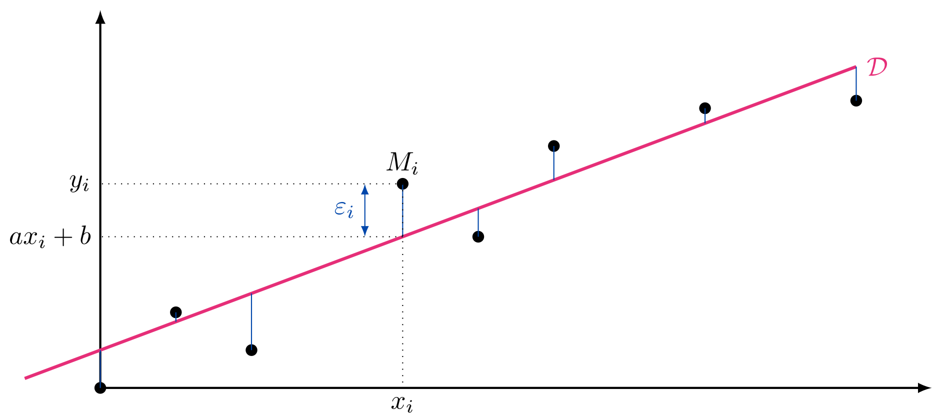
Introduction

Lors d'expériences visant à comparer des données expérimentales à un modèle mathématique, il apparait souvent des erreurs de mesure ; des points $M_i = (x_i, y_i)$ qui devraient être alignés (ou suivre un modèle) ne sont pas alignés, mais sont "presque" sur une même droite $\mathcal{D}$.

On cherche *une* droite $\mathcal{D}$, d'équation $y = ax + b$ qui passe *au plus près* des n points du nuage, c'est à dire, en notant $\varepsilon_i = y_i - (ax_i + b)$ qui *minimise* la quantité

$$\Delta(a, b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2.$$

Effectuer une régression linéaire, ou appliquer la méthode des moindres carrés, revient à déterminer les valeurs de a et de b (et donc l'équation de la droite $\mathcal{D}$) qui minimise Δ.



On propose ci-après de retrouver les formules pour a et pour b avec deux méthodes différentes : une projection orthogonale et une étude de fonction de deux variables.

On renvoie aux **Chapitres 11 & 14** du cours pour les détails théoriques et au **TP n°8** pour la pratique.

Méthode 1 : Projection orthogonale

On équipe $\mathbb{R}^n$ de sa structure euclidienne canonique. Notons $X = (x_1, x_2, ..., x_n)$ le vecteur de $\mathbb{R}^n$ dont les composantes sont les abscisses des points du nuage, $Y = (y_1, ..., y_n)$ celui avec les ordonnées des points du nuage et $\mathbb{1} = (1, 1, ..., 1)$ le vecteur dont toutes les composantes sont égales à 1.
On observe alors que

$$\Delta(a, b) = \|Y - (aX + b\mathbb{1})\|^2.$$

Ainsi, en notant $F = \mathrm{Vect}(X, \mathbb{1})$ le sous-espace de $\mathbb{R}^n$ engendré par X et par $\mathbb{1}$, le cours permet d'affirmer que le minimum cherché est atteint à l'aide du projeté orthogonal $p_F(Y)$ de Y sur F :

$$\min_{a,b\in\mathbb{R}} \Delta(a,b) = \min_{a,b\in\mathbb{R}} \|Y-(aX+b\mathbb{1})\|^2 = \min_{Z\in F} \|Y-Z\|^2 = \|Y-P_F(Y)\|.$$

On sait alors parfaitement obtenir l'expression du projeté orthogonal (sinon, on ira vite relire le chapitre susmentionné). Commençons par obtenir une base orthonormée (b.o.n) de F par le procédé de Gram-Schmidt :

$$\tilde{\mathbb{1}} = \frac{1}{\|\mathbb{1}\|}\mathbb{1} = \frac{1}{\sqrt{n}}\,(1,1,...,1)\,.$$

Ensuite on commence par trouver X' orthogonal à $\tilde{\mathbb{1}}$, qu'on normalisera ensuite.

$$X' = X - \langle X,\, \tilde{\mathbb{1}}\rangle\tilde{\mathbb{1}} = (x_1-\overline{x}, ..., x_n-\overline{x})\,,$$

où on a noté $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^n x_i$ la moyenne des x_i. On notera de même $\overline{y}$ la moyenne des y_i. Une fois normalisé, on prend donc

$$\tilde{X} = \frac{1}{\|X'\|}X' = \frac{1}{\sqrt{\displaystyle\sum_{j=1}^n (x_j-\overline{x})^2}}\,(x_1-\overline{x}, ..., x_n-\overline{x})\,.$$

Il suit que :

$$\begin{aligned}
P_F(Y) &= \langle Y|\tilde{\mathbb{1}}\rangle\tilde{\mathbb{1}} + \langle Y|\tilde{X}\rangle\tilde{X}\\
&= (\overline{y}, ..., \overline{y}) + \frac{\displaystyle\sum_{i=1}^n y_i(x_i-\overline{x})}{\displaystyle\sum_{i=1}^n (x_i-\overline{x})^2}\,(x_1-\overline{x}, ..., x_n-\overline{x})\\
&= \frac{\displaystyle\sum_{i=1}^n y_i(x_i-\overline{x})}{\displaystyle\sum_{i=1}^n (x_i-\overline{x})^2} X + \left(\overline{y} - \left(\frac{\displaystyle\sum_{i=1}^n y_i(x_i-\overline{x})}{\displaystyle\sum_{i=1}^n (x_i-\overline{x})^2}\right)\overline{x}\right)\mathbb{1} = aX + b\mathbb{1}.
\end{aligned}$$

En notant $\sigma_x^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^n (x_i-\overline{x})^2 = \dfrac{1}{n}\|X\|^2 - \overline{x}^2$ (on pourra réfléchir au choix de cette notation), on conclut que

$$a = \frac{\displaystyle\sum_{i=1}^n y_i(x_i-\overline{x})}{\displaystyle\sum_{i=1}^n (x_i-\overline{x})^2} = \frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^n y_i x_i - \overline{x}\cdot\overline{y}}{\dfrac{1}{n}\displaystyle\sum_{i=1}^n (x_i-\overline{x})^2} = \frac{\dfrac{1}{n}\langle X|Y\rangle - \overline{x}\cdot\overline{y}}{\sigma_x^2}, \qquad b = a\cdot\overline{y} - \overline{x}.$$

Le point de coordonnées $(\overline{x}, \overline{y})$ s'appelle le **point moyen** du nuage ; on observe que la droite de régression passe par le point moyen.

Méthode 2 : Point critique d'une fonction de deux variables

On montre dans cette section qu'on retrouve les résultats précédents (la valeur de a et b qui minimise Δ) en montrant que Δ présente un minimum (local) en (a,b) qui donc un point critique.
On a:

$$\begin{aligned}
\frac{\partial\Delta}{\partial a}(a,b) &= 2\left(\sum_{i=1}^n x_i^2\right)a + 2\left(\sum_{i=1}^n x_i\right)b - 1\sum_{i=1}^n x_i y_i\\
\frac{\partial\Delta}{\partial b}(a,b) &= 2\left(\sum_{i=1}^n x_i\right)a + 2nb - 2\sum_{i=1}^n y_i
\end{aligned}$$

On trouve le(s) point(s) critique(s) en résolvant un système linéaire, dont on omet les étapes ici

$$(a,b) \text{ point critique de } \Delta \iff \nabla\Delta(a,b) = 0$$
$$\iff \begin{cases} a = \dfrac{\displaystyle\sum_{i=1}^n y_i \sum_{j=1}^n x_j - n\sum_{i=1}^n x_i y_i}{\left(\displaystyle\sum_{i=1}^n x_i\right)^2 - n\displaystyle\sum_{i=1}^n x_i^2} = \dfrac{\frac{1}{n}\langle X|Y\rangle - \overline{x}\cdot\overline{y}}{\sigma_x^2} \\ b = \dfrac{\displaystyle\sum_{i=1}^n y_i - a\sum_{i=1}^n x_i}{n} = \overline{y} - a\cdot\overline{x} \end{cases}$$

Il s'agit des mêmes valeurs que précédemment, on vérifie que c'est un minimum local en explicitant sa matrice hessienne :

$$H_\Delta(a,b) = \begin{pmatrix} 2\displaystyle\sum_{i=1}^n x_i^2 & 2\displaystyle\sum_{i=1}^n x_i \\ 2\displaystyle\sum_{i=1}^n x_i & 2n \end{pmatrix},$$

qu'on remarque indépendante de a et b (bien que ce soit aux valeurs de a et b calculées précédemment qu'elle nous intéresse, et pas ailleurs, car c'est le seul point critique). Par Cauchy-Schwarz,

$$\langle X|\mathbb{1}\rangle^2 = \left(\sum_{i=1}^n x_i\right)^2 < \|X\|^2\|\mathbb{1}\|^2 = n\sum_{i=1}^n x_i^2,$$

et donc

$$\det\left(H_\Delta(a,b)\right) = 4\left(n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2\right) > 0$$

on a bien un extremum et

$$\mathrm{Tr}\left(H_\Delta(a,b)\right) = 2\left(n + \sum_{i=1}^n x_i^2\right) > 0$$

c'est donc bien un minimum (local). Easy.

Coefficient de corrélation linéaire

Définition 1.1. **Coefficient de corrélation linéaire**

On appelle coefficient de corrélation linéaire la quantité $\rho_{X,Y}$ souvent simplement notée ρ définie par

$$\rho = \frac{\frac{1}{n}\langle X|Y\rangle - \overline{x}\cdot\overline{y}}{\|X\|\|Y\|}.$$

Proposition.

Soit ρ le coefficient de corrélation linéaire du couple (X,Y). Alors

i. $\rho \in [-1;1]$;

ii. $\rho = \pm 1$ si et seulement si la régression $Y = aX + b$ est exacte.

☞ Il parait alors assez naturel de penser que si ρ est "assez proche" de 1 (en valeur absolue), l'approximation *affine* pourrait être pertinente.

Si $|\rho|$ est proche de 1 **et qu'on a visualisé une relation linéaire entre les données**, on peut confirmer qu'il y a bien corrélation linéaire entre X et Y.

☞ En sciences humaines, en sciences économiques et en sciences physiques, une valeur de $|\rho|$ de l'ordre de $0,85$ est souvent considérée comme bonne et justifie la pertinence de la régression linéaire.

Régression linéaire avec transformations

Dans certains cas, on peut appliquer le principe de régression linéaire à un couple obtenu par transformées de Y (ou aussi de X) et obtenir une relation de la forme

$$Y \simeq a\varphi(X) + b, \quad \text{ou} \quad \varphi(Y) \simeq a\varphi(X) + b.$$

Considérons un exemple avec des données correspondant à l'évolution du PIB par habitant (en USD) et du pourcentage de la population en zone urbaine de la Norvège, de 1960 à 2020 (source: World Bank Data).

1. Recopier et exécuter les instructions suivantes. Commenter le nuage de points.

```
import pandas as pd
data2=pd.read_csv('http://frederic.gaunard.com/2223/tp2_nor.csv', sep=';')

X=data2['PIB per capita']
Y=data2['Pop urbaine %']

plt.grid()
plt.plot(X,Y, '.') # nuage de points
plt.show()
```

2. Représenter le nuage de points $(\ln(X), Y)$.
3. Calculer le coefficient de corrélation linéaire de Y en $\ln(X)$.
4. Déterminer l'équation de la droite de régression de Y en $\ln(X)$.
5. En déduire qu'on peut supposer que la dépendance entre Y et X est de la forme

$$Y = a\ln(X) + b.$$

6. Représenter le nuage de points précédent sur lequel on fera apparaître la courbe d'équation $y = a\ln(t) + b$.