TP7B - Statistiques bivariées - Exercices

Dans tout le TP, on importe les modules suivants :

```
import numpy as np
import numpy.random as rd
import numpy.linalg as al
import matplotlib.pyplot as plt
```

1 Méthode des moindres carrés

Exercice 1 ⋆⋆

Télécharger le fichier `TP07b_pib_naissances_2019.csv` sur :
`https://cahier-de-prepa.fr/ecg2-fauriel/download?id=1326`.

1. Utiliser la commande suivante :

```
M = np.genfromtxt("nom_du_fichier.csv", delimiter=",")
```

pour importer les données. Cela donne une matrice M à deux colonnes. La première contient le PIB par habitant (en dollars) et la seconde le nombre de naissances pour 1000 habitants.

2. Tracer le nuage de point. Y a-t-il un lien apparent entre le PIB par habitant et le taux de natalité ?
3. Déterminer la droite des moindres carrés et la tracer. L'approximation linéaire est-elle raisonnable ?
4. Quel est le signe du coefficient directeur de la droite ? Comment l'expliquer ?

2 Covariance, coefficients de corrélations

Exercice 2 - Corrélation $\neq$ causalité ⋆⋆

Voici deux séries de données :

- la dépense des USA sur la science, le spatial et les technologies (en milliards de dollars par an) et le nombre de suicides par pendaison aux USA (par année) :

Année	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Dépense	18.079	18.594	19.753	20.734	20.831	23.029	23.597	23.584	25.525	27.731	29.449
Suicides	5427	5688	6198	6462	6635	7336	7248	7491	8161	8578	9000

- la consommation de fromage par personne aux États-Unis (en livres par personnes et par an) et le nombre de personnes décédées en s'emmêlant dans leurs draps de lits (par année) :

Année	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Consommation	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Morts	327	456	509	497	596	573	661	741	809	717

1. Calculer les coefficients de corrélations linéaires.
2. Que peut-on en déduire ?

3 Pour aller plus loin

Exercice 3 - Existence de la droite des moindres carrés ⋆⋆⋆

On va prouver que la droite des moindres carrés existe et est unique. On reprend les notations :

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \in \mathrm{M}_{n,2}(\mathbb{R}) \quad \text{et} \quad B = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathrm{M}_{n,1}(\mathbb{R}).$$

1. Jusitifier que $\mathrm{rg}(A) \geq 1$. Que signifie le cas $\mathrm{rg}(A) = 1$? On l'exclut pour la suite de l'étude.
   On a donc $\mathrm{rg}(A) \geq 2$.
2. Soit $X = \begin{pmatrix} a \\ b \end{pmatrix}$. Exprimer $\|AX - B\|^2$ en fonction des x_k et y_k.
3. Appliquer le théorème des pseudo-solutions et en déduire l'existence et l'unicité de la droite des moindres carrés.

Exercice 4 - Démonstration de la formule des moindres carrés ⋆⋆⋆⋆⋆

On montre dans cet exercice que l'unique X_0 minimisant $\|AX - B\|$ vérifie ${}^tAAX_0 = {}^tAB$.
On pose $f : \mathrm{M}_{2,1}(\mathbb{R}) \to \mathrm{M}_{n,1}(\mathbb{R})$ définie par $f(X) = AX$. Puisque X_0 existe (voir exercice 3), on a $Y_0 = AX_0$ le projeté orthogonal de B sur $\mathrm{Im}(f)$.

1. Montrer que $Y_0 - B \in \mathrm{Im}(f)^\perp$.
2. En déduire que :
$$\forall X \in \mathrm{M}_{2,1}(\mathbb{R}), {}^tX({}^tAY_0 - {}^tB) = 0.$$
3. Conclure.

4 Travail à préparer pour le prochain TP

Exercice 5 - Sensibilité aux valeurs extrêmes ⋆⋆

On reprend les données de l'exercice 1.

1. Calculer le coefficient de corrélation linéaire.
2. Tracer le nuage de points. Y a-t-il des points qui semblent très éloignés des autres ? Par exemple, y a-t-il des PIB par habitant extrêmes ?
   Y aurait-il une explication pour ces valeurs ? Ces dernières sont-elles pertinentes alors pour notre étude ?
3. Filtrer la matrice M pour que seules les PIB par habitant inférieur à 60000 soient conservés.
4. Calculer le nouveau coefficient de corrélation linéaire.
5. Tracer la droite des moindres carrées calculée à partir du nouveau jeu de données et la superposer aux nuages de points initials. Commenter.