Statistiques bivariées

Exercice 1 (Une première régression linéaire)

Le temps de chargement d'une page sur internet dépend de nombreux paramètres, entre autre le nombre d'utilisateurs qui y sont connectés simultanément. Par ailleurs, le temps de chargement influe en retour sur le nombre de visiteurs : plus le temps de chargement est long, plus les utilisateurs sont susceptibles de se diriger vers d'autres ressources. Le responsable d'un site a relevé le nombre d'internautes sur son site en fonction de sa durée de chargement :

Nombre d'internautes connectés (en millier) x_i	0,5	1	2,5	3	4	5	6
Durée de chargement (en secondes) y_i	0,3	0,4	0,6	0,9	1,3	2	2,8

1. Représenter le nuage de points de coordonnées $(x_i; y_i)$ associés à cette série statistique. (on pourra aller sur basthon.fr pour laisser le soin à Python de faire le dessin)
2. Est-il pertinent de réaliser un ajustement linéaire ?
3. Déterminer l'équation $y = ax + b$ de la droite d'ajustement $\mathscr{D}$ obtenue par la méthode des moindres carrés (on pourra de nouveau s'aider de python).
4. Pour la suite, on prendra $y = 0,44x - 0,19$ pour équation de la droite $\mathscr{D}$.
   a. Tracer la droite $\mathscr{D}$.
   b. Avec ce modèle, estimer la durée de chargement pour 8000 personnes connectées. (sur sa feuille)
   c. Une étude indépendante a montré que 60% des internautes cesse de charger une page pour se diriger vers un autre site dès que le temps de chargement dépasse 3,5 secondes.
   Avec le modèle précédent, estimer le nombre de visiteurs sur ce site lorsque la durée de chargement est de 3,5 secondes.
   Combien de visiteurs perdrait-il alors ?

Exercice 2 (Choix du meilleur coefficient de corrélation)

Une entreprise recherche une variable quantitative stratégique S dont elle n'a pas accès en temps réel. Elle souhaite estimer cette variable au moyen d'un "proxy" (*) : une variable quantitative à laquelle elle a accès et dont les variations reflètent le plus fidèlement S. Pour cela, des spécialistes ont considéré trois variables X_1, X_2 et X_3 dont l'historique a montré que les coefficients de corrélation linéaires respectifs avec S est :

$$r_{X_1,S} = 0,78 \qquad r_{X_2,S} = 0,29 \qquad r_{X_3,S} = -0,85$$

Quelle est la variable servant le mieux de proxy ? Justifier.

(*) Un proxy (litt. « mandataire ») est un composant logiciel informatique qui joue le rôle d'intermédiaire en se plaçant entre deux hôtes pour faciliter ou surveiller leurs échanges.

Exercice 3 (Choix du meilleur coefficient de corrélation - autre paradigme)

Un fonds de placement ayant des participations financières dans une entreprise EEE cotée en bourse souhaite '"hedger" ses positions, c'est à dire trouver un placement dont les fluctuations en bourse (positives ou négatives) absorbent le mieux les fluctuations de la valeur de l'action EEE afin de réduire son risque. Elle étudie quatre placements possibles P_1, P_2, P_3 et P_4, dont les valeurs passées ont été comparées à celle de l'action EEE. On donne leurs coefficients de corrélation linéaires respectifs avec la valeur de l'action EEE :

$$r_1 = 0,24 \qquad r_2 = -0,03 \qquad r_3 = -0,18 \qquad r_4 = 0,54$$

Dans quel placement doit-elle investir ? Justifier.

Exercice 4 (Autour de la covariance et la corrélation linéaire)

Soient x et y deux séries statistiques. On pose $x' = 2x$, $y' = 2y$ et $x'' = \frac{1}{2}x$

1. Déterminer la covariance de x' et y' en fonction de celle de x et y.
2. Déterminer la covariance de x'' et y' en fonction de celle de x et y.
3. Quelles sont les séries les plus corrélées entre x et y d'une part, x' et y' d'autre part, et x'' et y' enfin ?

Exercice 5 (Entre corrélation et explication)

Le tableau suivant donne les évolutions, de Mai à Septembre, du nombre de climatiseurs vendus et de noyade par accident dans un secteur littoral.

	Nombre de climatiseurs x_i	Nombre de noyades y_i
Mai	66	1
Juin	88	3
Juillet	90	5
Août	110	8
Septembre	60	0

1. Représenter graphique le nuage de points correspondant.
2. Un ajustement affine semble-t'il pertinent ? Donner l'équation de la droite d'ajustement par moindres carrés.
3. Prévoir le nombre de noyades si en Avril de l'année d'après, 88 climatiseurs sont vendus ?
4. Commenter la relation de causalité entre les variables étudiées. Pouvez-vous proposer une autre explication ? En quoi la régression linéaire n'est-elle pas une explication suffisante ici ?

Exercice 6 (Ajustement linéaire avec Python)

On considère les séries suivantes :

x_i	15	13	5	10	16	7	13	13	16	19
y_i	34	25	12	23	31	17	28	29	28	38

Écrire les lignes de code qui permettent de tracer le nuage de point, le point moyen ainsi que la droite de régression linéaire en confiant toute la partie calculatoire à Python.

Exercice 7 (Compléments en Python - initiation aux bases de données)

Imaginons qu'un fichier nommé appartements.csv(**), représenté ci-contre dans un tableau, permet de collecter les renseignements concernant le prix de vente d'appartements (exprimés en milliers d'euros) et la surface en m^2 à Paris en septembre 2021 :

Surface	Prix
74	796
105	815
17	209
14	180
50	615
73	720
17	275
20	262
38	700
36	320
61	650

Écrire les lignes de code permettant d'importer le fichier csv, d'en extraire les colonnes, de tracer le nuage de points, et le point moyen de cette série double.

Exercice 8 (Ajustement logarithmique)

Le tableau suivant donne l'évolution du chiffre d'affaire (en millions d'euros) d'une entreprise depuis sa création en 2002 :

Année X	2002	2003	2004	2005	2006	2007	2008	2009
Chiffre d'affaire Y	0,7	1,6	2	2,4	2,5	2,8	3	3

On s'intéresse à la série statistique (X, Y).

1. Étudier la pertinence d'un ajustement linéaire.
2. Calculer le coefficient de corrélation de X et exp(Y). Le comparer à celui de X et Y .
3. En déduire un ajustement logarithmique de Y en X.

Exercice 9 (Ajustement carré)

Au cours d'une séance d'essai, un pilote automobile doit, quand il reçoit un signal sonore dans son casque, arrêter le plus rapidement possible son véhicule. Au moment du top sonore, on mesure la vitesse de l'automobile puis la distance nécessaire pour arrêter le véhicule. Pour six expériences, on a obtenu les résultats suivants :

v_i (km/h)	27	43	62	80	98	115
distance y_i d'arrêt (m)	6,8	20,5	35,9	67,8	101,2	135,8

On pose $x_i = v_i^2$ et on considère la série $(x_i; y_i)$.

1. Compléter le tableau

x_i						
y_i	6,8	20,5	35,9	67,8	101,2	135,8

2. Dans un repère orthogonal représenter le nuage de points associé à cette nouvelle série (unités : 1cm pour 1000 en abscisse, et 1 cm pour 10 en ordonnée).
3. **a.** Déterminer, l'équation de la droite de régression de y en x sous la forme $y = mx + p$.
   Tracer cette droite dans le repère précédent.
   b. A l'aide de cette équation, déterminer la valeur estimée de x correspondant à une distance d'arrêt de 180 m, puis la vitesse correspondante du véhicule.
   c. Quelle est la vitesse d'arrêt estimée correspondant à une vitesse de 150 km/h.
   d. Le manuel du code de la route donne, pour calculer la distance d'arrêt, en mètres, la méthode suivante : "Prendre le carré de la vitesse exprimé en dizaines de kilomètres par heure." Comparer le résultat obtenu au c. à celui que l'on obtiendrait par cette méthode.