Cours d'analyse de données en géographie
Niveau Master 1 - GEANDO
Séance 7. Régression et corrélation statistique de deux variables

Maxime Forriez[1,a]

[1] Institut de géographie, 191, rue Saint-Jacques, Bureau 105, 75 005 Paris,
[a]maxime.forriez@sorbonne-universite.fr

28 septembre 2025

1 Questions de cours

Les réponses comptent pour 20 % de la note finale du parcours « intermédiaires ».
Les réponses comptent pour 10 % de la note finale du parcours « confirmés ».

1. Quel est l'intérêt de passer des statistiques univariées aux statistiques bivariées ?
2. Quelles différences opérez-vous entre corrélation et correspondances ? Qu'est-ce qu'un rapport de corrélation ?
3. Quelles différences faites-vous entre les valeurs marginales et les valeurs conditionnelles ? Pourquoi distinguer les deux ?
4. Quelles différences faites-vous entre variance et covariance ?
5. Pourquoi mesurer la corrélation ou l'indépendance ?
6. Quel est le principe de la méthode des moindres carrés ? À quoi sert-elle ?
7. Expliquez en un court paragraphe ce qu'est la théorie de la corrélation (simple) ?
8. En quoi consiste le piège de l'autocorrélation ?
9. Expliquez en un court paragraphe ce qu'est une régression linéaire ?
10. Quelle est la différences entre coefficient de corrélation et coefficient de détermination ?
11. Pourquoi faut-il tester les deux droites de régression ?

2 Mise en œuvre avec Python

La sous-partie « Bonus » vous permet d'obtenir des points supplémentaires.

2.1 Objectifs

— Apprendre à gérer les données censurées

— Manipulation les outils de la régression et de la corrélation simple

2.2 Manipulations

Le fichier obtenu compte pour 20 % de la note finale du parcours « intermédiaires ».
Le fichier obtenu compte pour 15 % de la note finale du parcours « confirmés ».
Existe-t-il un lien entre le produit intérieur brut (P.I.B.) et la consommation énergétique ? Le fichier de données que vous allez analyser est issu des données de la Banque mondiale (https://donnees.banquemondiale.org/). Il regroupe deux jeux de données :

— le P.I.B. de chaque territoire (étatique ou non) de 1962 à 2024 en dollars courants (c'est-à-dire sans prendre en compte l'inflation) ;

— la consommation énergétique en kilogrammes équivalent pétrole de 1962 à 2024.

J'ai opéré les principales opérations de nettoyage, et fais en sorte que vous ayez le moins de difficultés possibles à obtenir le résultat recherché.

1. Il existe un décalage entre les données du P.I.B. et de la consommation énergétique. Il faut de fait sélectionner en utilisant `Pandas` les colonnes `PIB_2022` et `Utilisation_d_energie_2022` dans le fichier `pib-vs-energie.csv`.
2. Malheureusement, plusieurs données sont censurées. Il vous faut créer un algorithme qui ne sélectionnera que les couples complets, c'est-à-dire ayant une valeur pour le P.I.B. et la consommation énergétique. Dit autrement, vous devez exclure les données manquantes (aucune valeur pour l'un et l'autre), et les données partielles (une donnée chez l'une, mais pas l'autre).
3. On considère que la variable explicative est la consommation énergétique et la variable à expliquer est le P.I.B. Calculer une régression linéaire simple entre les deux colonnes avec la méthode `scipy.stats.linregress(x, y)` prenant en arguments `x`, la variable à expliquer, et `y`, la variable explicative.
4. Calculer la corrélation simple entre les deux colonnes. Vous pouvez utiliser indifféremment les bibliothèques `Pandas` ou `Scipy`.
5. Faire un graphique de synthèse permettant de visualiser la droite de régression obtenue.
6. Dans votre rapport, vous commenterez votre résultat sous la forme d'un ou deux paragraphes.

2.3 Bonus

Vous avez écrit un algorithme permettant de traiter deux colonnes se rapportant à la même année. Écrivez un algorithme permettant de généraliser votre résultat à toutes les années de 1962 à 2022. N'oubliez pas d'organiser correctement vos fichiers de sortie.