Cours d'analyse de données en géographie
Niveau Master 1 - GEANDO
Séance 10. Les méthodes statistiques multivariées explicatives

Maxime Forriez[1,a]

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

30 octobre 2025

Courage, c'est presque fini !

1 Questions de cours

Les réponses comptent pour 25 % de la note finale du parcours « confirmés ».

1. Essayez d'expliquer la généralisation de la corrélation dans un espace à n dimensions ?
2. Comment expliqueriez-vous ce qu'est une régression multiple ?
3. À quoi sert l'analyse de la variance à double entrée ?
4. Pourquoi malgré le fait que l'analyse canonique soit la plus générale, ne sert-elle à rien en analyse de données ?
5. À quoi les méthodes explicatives peuvent-elles être utiles dans des analyses de données géographiques ?

2 Mise en œuvre avec Python

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

2.1 Objectifs

— Apprendre à faire une régression linéaire multiple avec `Pandas`.
— Apprendre à faire une régression linéaire multiple avec `Sklearn`.

2.2 Manipulations

Le fichier obtenu compte pour 25 % de la note finale du parcours « confirmés ».

L'exercice est relativement simple. À la différence des méthodes factorielles, ici, il s'agit juste de déterminer des coefficients. De fait, deux séries de données vous seront proposées.

1. Ouvrir le fichier `temperature.csv`. Le fichier contient 33 villes des États-Unis (champ `Ville`). Les individus sont caractérisés par une variable à expliquer `Température_en_janvier` et trois variables explicatives `Latitude` (en degré décimal), `Longitude` (en degré décimal) et `altitude` (en mètre).
   - Calculer la corrélation entre les quatre variables avec la méthode `corr()` de `Pandas`.
   - Afficher les statistiques descriptives avec `describe()` de `Pandas`.
   - Isoler le champ `Ville` et isoler les données en distinguant `Température_en_janvier`, la variable à expliquer, et `Latitude`, `Longitude` et `altitude`, les variables à expliquer.
   - Avec la bibliothèque `statsmodels.api`, calculer la régression linéaire liant la variable à expliquer avec les variables explicatives.
   - En utilisant les attributs et les méthodes de `statsmodels.api`, calculer les paramètres de la régression linéaires (les coefficients), le coefficient de détermination, le résumé de la régression et les *p-values*.
   - Avec `Scikit-Learn`, isoler les coefficients avec l'attribut `coef_`. Isoler la constante de l'ajustement avec la méthode `intercept_`.
   - Dans votre rapport, commenter et analyser les résultats obtenus.
2. Ouvrir le fichier `geomarketing.csv`. Il s'agit d'un exercice que vous pourriez avoir dans une organisation qui vous emploierez. Il faut expliquer la variable `ca`, le chiffre d'affaires de chaque individu représentant le magasin fictif `Big Bazar`.
   - Une A.C.P. a été, au préalable, réalisée sur l'ensemble des variables. Elle a révélé que les variables significatives étaient :
     - `ca`;
     - `surface_totale`;
     - `potentiel_Z20`;
     - `nb_primaire_Z10`;
     - `nb_primaire_Z20`;
     - `nb_gsa_Z10`;
     - `nb_pharmacie_Z5`;
     - `nb_conc2_Z10`;
     - `nb_conc2_Z20`;
     - `P10_POP_Z15`;

- `P10_MEN_Z10`.

— Isoler `ca` et isoler les neuf variables significatives.

— En utilisant les attributs et les méthodes de `statsmodels.api`, calculer les paramètres de la régression linéaires (les coefficients), le coefficient de détermination, le résumé de la régression et les p-*values*.

— Dans votre rapport, commenter et analyser les résultats obtenus.

2.3 Bonus

À partir du fichier `geomarketing.csv`, opérer la régression avec la totalité des variables, comparer avec les variables sélectionnées par une A.C.P., et, dans votre rapport, commenter et analyser les résultats obtenus.