ISEP
IF.1205 – Science et Traitement des données
Mai 2025

PROJET DE SCIENCE DES DONNEES

Analyse en Composantes Principales (ACP) et Régression Linéaire

I. Instructions générales

Le but de ce projet est d'appliquer les méthodes d'analyse en composantes principales et de régression linéaire sur des données réelles. Le projet sera réalisé par équipe de **trois étudiants.** Chaque équipe devra présenter son travail au cours d'une soutenance orale qui aura lieu le **mardi 3 juin prochain**. Vous pouvez travailler en R ou en Python. Les instructions concernant la soutenance orale et l'évaluation sont précisées ci-dessous.

a) Instructions concernant la soutenance orale

La soutenance orale durera environ 15 minutes par groupe, et se décomposera en 10 minutes de présentation et 5 minutes de questions. Les diapositives de la présentation devront contenir les éléments suivants :
- Une page de couverture contenant le prénom, nom de famille, et le numéro d'identification étudiant de tous les membres de l'équipe.
- Un sommaire.
- Une courte introduction.
- Le corps du document (résultats, figures, tables, interprétations, commentaires, ou tout autre élément qui permette de répondre aux questions). Les réponses aux questions posées dans le sujet doivent être clairement indiquées dans cette partie.
- La conclusion.
- Les références.

Votre code R ou Python ne doit pas être inclus dans la présentation. Cependant, vous devez avoir ce code sous la main au moment de la soutenance, pour répondre à toute question éventuelle à ce sujet.
Un seul membre de votre équipe devra déposer votre fichier de présentation au format pdf sur Moodle le **lundi 2 juin midi au plus tard**. Un dépôt Moodle sera créé pour chaque groupe de TD (G7, G8, G9 et G10). Le nom du fichier déposé devra respecter la forme suivante :

NomEtudiant1_NomEtudiant2_NomEtudiant3.pdf

Par ailleurs, chaque équipe devra déposer un fichier csv contenant les réponses quantitatives aux questions posées (indiquées par **[qij]** dans l'énoncé) dans le même dépôt Moodle. Le nom de ce fichier doit être de la forme :

NomEtudiant1_NomEtudiant2_NomEtudiant3.csv

Un modèle pour ce fichier est disponible sur Moodle (**modele.csv**). Un seul membre de votre équipe devra modifier ce fichier csv et le déposer sur Moodle le **lundi 2 juin midi au plus tard**.

b) Instructions concernant l'évaluation

La soutenance orale sera divisée en deux parties : 10 minutes de présentation orale et 5 minutes de questions. L'évaluation sera à la fois collective, notamment pour la qualité et le contenu global de la présentation, mais aussi individuelle.
Chaque étudiant sera donc évalué sur ses interventions au cours des deux parties. Une attention particulière sera portée à la qualité des réponses en termes d'analyse.

II. Analyse de données

a) Le jeu de données

La météorologie est une science qui étudie les phénomènes atmosphériques. L'un de ces enjeux est la prévision localement et à court terme de variables météorologiques quotidiennes - précipitation, ensoleillement, température – mais aussi extrêmes, telles que les inondations, canicules ou cyclones. La fiabilité des prévisions est donc primordiale pour assurer la sécurité des populations, ou dans des secteurs comme le transport, en particulier aérien. Cette fiabilité entraîne aussi des répercussions économiques dans des domaines comme l'agriculture.
Dans ce contexte, le but de ce projet est d'analyser des données météorologiques obtenues en 2024, puis de proposer un modèle afin de prédire les températures mensuelles de 2025.
Les données météorologiques sont disponibles dans les fichiers **data1.csv** et **data2.csv**.

b) Analyse préliminaire : statistiques descriptives

Le fichier **data1.csv** contient des données météorologiques mesurées en France en 2024, et plus particulièrement :

- la température minimale moyenne mesurée en °C,
- la température maximale moyenne mesurée en °C,
- la hauteur de précipitations totale mesurée en mm,
- la durée d'ensoleillement totale mesurée en heures.

Après importation du fichier **data1.csv**, répondez aux questions suivantes :

1. De combien de villes proviennent les données météorologiques [q1a] ? Combien de villes sont concernées par des mesures manquantes [q1b] ? **Si existantes, supprimez ces villes. Elles ne seront pas prises en compte dans la suite de notre analyse.**

2. Quelle est la ville associée à la valeur :
   - minimale de température minimale [q2a],
   - maximale de température minimale [q2b],
   - minimale de température maximale [q2c],
   - maximale de température maximale [q2d],
   - minimale de hauteur de précipitations [q2e],
   - maximale de hauteur de précipitations [q2f],
   - minimale de durée d'ensoleillement [q2g],
   - maximale de durée d'ensoleillement [q2h].

   Indiquez aussi chaque valeur associée [q2a-h]. Commentez.

3. Calculez la variance de :
   - la température minimale [q3a],
   - la température maximale [q3b],
   - la hauteur de précipitations totale [q3c],
   - maximale de durée d'ensoleillement [q3d].
   Commentez.

4. Calculez la moyenne [q4a], la médiane [q4b], et l'écart-type [q4c], de la variable météorologique de variance minimale calculée à la question précédente. Affichez l'histogramme associé. Commentez.

5. Calculez la moyenne [q5a], la médiane [q5b], et l'écart-type [q5c], de la variable météorologique de variance maximale calculée à la question 3. Affichez l'histogramme associé. Commentez.

6. On s'intéresse aux corrélations linéaires entre les différentes variables météorologiques. Quelles sont les deux variables les plus positivement corrélées [q6a] ? Quelles sont les deux variables les plus négativement corrélées [q6b] ? Quelles sont les deux variables les moins corrélées [q6c] ? Affichez également les valeurs de corrélation associées [q6a-q6c]. Illustrez graphiquement vos résultats, en rajoutant le nom des villes sur chacune des trois figures créées. Commentez.

7. On s'intéresse maintenant aux corrélations linéaires entre les villes. Affichez la matrice de corrélation. Commentez.

c) Analyse en Composante Principales (ACP)

Nous appliquons maintenant une Analyse en Composantes Principales (ACP) à partir des quatre variables météorologiques du jeu de données **data1.csv**. Préalablement, nous centrons et réduisons nos données en utilisant la formule suivante :

$$X_i' = \frac{X_i - \mu}{\sigma}$$

où X_i représente une variable météorologique, μ sa moyenne et σ son écart-type.

8. Appliquez une ACP sur les données météorologiques préalablement centrées et réduites. Affichez les deux premières composantes principales sous forme d'un nuage de points pour visualiser les résultats. Rajoutez le nom des villes sur la figure ainsi créée. Quel est le pourcentage de variance expliquée par la première [q8a] et deuxième composante principale [q8b] ? Affichez aussi ces valeurs sur les axes de votre figure. Commentez.

9. Affichez et commentez le cercle de corrélation.

10. Superposez les résultats de la question 8 et 9 sur une même figure. Pouvez-vous retrouver les résultats de la question 2 à partir de cette figure ?

d) Régression linéaire simple

Nous allons maintenant nous intéresser aux relevés de températures maximales effectués en 2023 et 2024 à Paris, disponibles dans le fichier **data2.csv**. Pour commencer, nous nous focalisons uniquement sur l'année 2024.

11. Affichez l'évolution de la température en 2024 à Paris en fonction du mois. Commentez.

Dans un premier temps, nous cherchons à prédire la température maximale à Paris en janvier 2025 à partir des données obtenues à Paris en 2024. Le modèle de régression linéaire que nous testons s'écrit :

$$\textit{Température maximale} = \beta 0 + \beta 1 * \textit{numéro mois} + \varepsilon \qquad (1)$$

Où *numéro mois* représente la position du mois dans l'année, allant de 0 pour le mois de janvier, à 11 pour le mois de décembre. Nous choisissons d'appliquer la régression linéaire simple en considérant uniquement les n derniers mois, n allant de 1 (modélisation à partir du seul mois de décembre 2024) à 12 (prise en compte des mois de janvier à décembre 2024). Le modèle retenu est celui qui obtient la plus grande valeur de coefficient de détermination R2 ajusté.

12. Appliquez les n régressions linéaires. Quelle est la valeur optimale de n [q12a] ? Quelle est la valeur du coefficient de détermination R2 associé – ajusté [q12b] et non ajusté [q12c] ? Donnez aussi les valeurs de β0 [q12d] et β1 [q12e] prédites par le modèle optimal. Analysez quantitativement et visuellement vos résultats.

13. La température maximale à Paris en janvier 2025 était de 7,5 °C. Quelle est la température prédite pour janvier 2025 par le modèle obtenu à la question précédente [q13a] ? Quel est l'écart entre la température prédite et réelle pour janvier 2025 [q13b] ?

14. Evaluez l'hypothèse de pente nulle pour le coefficient β1 obtenu à la question 12. Quelle est la p-valeur [q14a] obtenue pour le test associé ? Existe-t-il une relation linéaire entre les deux variables en prenant α=5% [q14b] ?

e) Régression linéaire multivariée

Dans un second temps, nous cherchons à prédire la température maximale à Paris en janvier 2025 à partir des données obtenues à Paris en 2024 et en 2023.

15. Superposez l'évolution de température en 2023 et 2024 à Paris en fonction du mois sur la même courbe, en utilisant une couleur différente pour chaque année. Commentez.

Nous faisons l'hypothèse qu'il existe une relation linéaire entre la température maximale mesurée au cours d'un mois donné et celle obtenue aux mois précédents. Nous entraînons notre modèle de régression linéaire multivariée sur les 12 mois de 2024. Pour chaque mois i de 2024, il y a au maximum 12 variables disponibles, correspondant à la température

maximale du mois i - 1 jusqu'à celle du mois i - 12. La variable à prédire correspond à la température maximale du mois i.

16. Combien y a-t-il de combinaisons de variables possibles pour prédire la température d'un mois donné [q16a] ? Calculez celle qui permet de maximiser le coefficient de R2 ajusté. Quel est le nombre de variables ainsi sélectionnées [q16b] ? Affichez le R2 ajusté optimal, les variables sélectionnées et les paramètres associés. Commentez. Existe-t-il une relation linéaire entre les variables sélectionnées en prenant α=5% ?

17. Calculez l'écart entre la température maximale prédite par le modèle de la question précédente pour janvier 2025, et la température maximale mesurée (7,5°C) [q17]. Commentez. Ce modèle peut-il être utilisé pour prédire les températures maximales de février, mars et avril 2025 (les températures mesurées étant respectivement de 8,6°C, 14,6°C et 20 °C) ? Commentez.

III. Références

- https://meteofrance.com/