**ccccccccc :**

**Théorie et applications**

***Travaux pratiques***

L’objectif de ce TP est de construire des modèles sur des données à forte multicolinéarité entre les prédicteurs.

Les données sont issues des résultats des deux tours des élections présidentielles de 2022. L’objectif final est de modéliser les résultats du second tour, de cerner les grandes tendances, de détecter des départements atypiques ou mal reconstitués, etc.

Le jeu de données contient 107 individus qui correspondent aux départements de la métropole, de l’outre-mer et des Français à l’étranger. Les variables correspondent aux nombres d’inscrits, de votants, d’abstentions, de suffrages exprimés, de bulletins blancs ou nuls pour les deux tours, de suffrages remportés par chacun des 12 candidats du premier tour et des deux candidats du second tour. ATTENTION : les analyses sur les quatre questions suivantes seront à effectuer sur l’ensemble des 107 départements et seulement sur la métropole.

***Question 1* :**

Explorer les données avec les statistiques descriptives (univariées et bivariées) et fournir un état des lieux de celles-ci (univariée, scatterplots, cartographie des corrélations), sur les nombres bruts et sur les proportions. Quelles différences de résultats remarquez-vous ?

***Question  2*** :

Sur chacun des deux tours puis sur les deux tours simultanément, quels types d’analyses multidimensionnelles peut-on pratiquer ? Justifier vos choix et appliquer les méthodes retenues.

***Question 3*** :

Commençons par quatre modèles univariés pour estimer les proportions des abstentionnistes, des bulletins blancs ou nuls, des suffrages de chacun des deux candidats retenus, à l’aide des variables du premier tour qui seront également transformées en proportions.

Pour cela, vous appliquerez une régression MCO et vous commenterez les résultats en utilisant les différents outils relatifs à la multicolinéarité vus en cours à l’aide des codes R associés.

Vous appliquerez alors les régressions PCR et PLS à l’aide du package R **plsdepot** et vous commenterez les résultats obtenus, en utilisant un maximum d’aides à l’interprétation pour les deux méthodes afin d’affiner les commentaires dur les résultats.

Vous comparerez les résultats obtenus à l’aide des trois méthodes : MCO, PCR et PLS.

***Question 4*** :

L’objectif est de modéliser l’ensemble des résultats obtenus au second tour à l’aide de ceux du premier tour. La réponse est donc multivariée puisque l’on modélisera les quatre variables ajustées indépendamment dans la question 3. La modélisation sera également à effectuer sur les proportions.

On appliquera la régression PLS. Vous commenterez alors les différents résultats, notamment pour répondre à l’objectif final qui est de cerner les grandes tendances, de détecter des départements atypiques ou mal reconstitués, etc. Et vous comparerez les résultats obtenus à la question 3.