|  |
| --- |
| Examen blanc  Cours de R & Rstudio (A Maddi)  Durée : 1h  Matériel autorisé : tout sauf ChatGPT |

**Contexte :**

Le baromètre de l'esprit critique collecte des informations sur les attitudes et comportements des individus face à l'information, aux biais cognitifs et à la vérification des faits. Votre mission est d'explorer et d'analyser les données avant un accent particulier sur trois scores :

* score de l’esprit critique (scoreesprit)
* score sur les pratiques d’alimentation (scoreali)
* score autour des questions de confiance en la science (scoreatt)

Le sujet vous est proposé sous forme d’un énoncé complet comportant plusieurs parties allant de la préparation des données jusqu’à l’analyse de régression multivariée.

Vous disposez d’un fichier Excel (baro25\_export.xlsx) contenant les données issues du baromètre de l’esprit critique. Ce jeu de données comporte notamment deux enquêtes (Enquête 1 et Enquête 2), mais vous travaillerez principalement sur l’Enquête 1 (variable baro\_enq1). Les données comportent plusieurs variables socio-démographiques (genre, tranche d’âge, CSP, diplôme, revenu, région, etc.) ainsi que trois scores mesurant différents aspects de l’esprit critique et des pratiques alimentaires/scientifiques.

L’objectif de cet examen est de vous amener à :

1. Préparer et nettoyer les données,
2. Explorer et visualiser les distributions et relations entre variables,
3. Réaliser des analyses de régression pour modéliser et expliquer les scores.

**Partie 1 : Préparation et nettoyage des données**

1. **Chargement des données**
   * Télécharger les données ici : <https://sdrive.cnrs.fr/s/fWfFEoxJjzHeCcy>
   * Charger le fichier Excel (baro25\_export.xlsx) dans R en utilisant le package readxl.
   * Sélectionner les observations correspondant à l’Enquête 1 (filtrage sur la variable echantillon\_main égale à "Oui").
2. **Sélection et renommage**
   * Sélectionnez les variables suivantes :  
     serial\_number, respondent\_id, poids1, scoreatt, scoreali, scoreesprit, regions\_nouvelles\_en\_13, habitez\_vous\_2, genre, tranche\_dage, csp\_2, variable sur les revenus, diplome, autopol.
   * Renommez ces variables afin d’obtenir des noms plus explicites (par exemple : num, id, poids, scoreatt, scoreali, scoreesprit, region, habitat, genre, age, csp, revenu, diplome, autopol).
3. **Recodage et réordonnancement**
   * Recodez les variables qualitatives pour gérer les modalités manquantes (ex. : recoder "NR", "NSP" en valeurs manquantes ou NA pour autopol et diplome).
   * Réordonnez les modalités de variables telles que csp, revenu, diplome, age et region afin d’assurer une interprétation cohérente lors des analyses graphiques et statistiques.
   * Supprimez les observations comportant des valeurs manquantes.

**Partie 2 : Exploration des données**

1. **Statistiques descriptives**
   * Utilisez le package gtsummary pour réaliser un tableau récapitulatif des variables (à l’aide de la fonction tbl\_summary()).
2. **Visualisations**
   * Réalisez des graphiques de densité (par exemple, à l’aide de geom\_density\_ridges\_gradient du package ggridges) pour visualiser la distribution de scoreali selon certaines variables (par exemple, age et autopol).
   * Créez des graphiques exploratoires pour étudier les relations entre les trois scores à l’aide de la fonction ggpairs du package GGally en colorant par la variable genre.

**Partie 3 : Analyse de régression**

Pour chacune des trois variables scores (scoreatt, scoreali et scoreesprit), vous allez :

1. **Modélisation complète**
   * Construisez un modèle de régression linéaire généralisé (glm) en expliquant chaque score en fonction de l’ensemble des variables explicatives suivantes :
     + Les deux scores restants (exemple : pour modéliser scoreatt, utiliser scoreali et scoreesprit comme covariables),
     + genre, age, autopol, revenu, habitat, csp, diplome, region.
   * Utilisez la famille gaussienne et intégrez la variable de poids (poids).
2. **Résumé du modèle**
   * Affichez le résumé du modèle à l’aide de tbl\_regression() et ajoutez la p-value globale avec add\_global\_p().
3. **Visualisation et validation**
   * Visualisez les coefficients du modèle à l’aide de la fonction ggcoef\_model().
   * Effectuez une validation pas à pas (stepwise) pour obtenir un modèle simplifié à l’aide de la fonction step().
   * Comparez les modèles complet et simplifié via la fonction ggcoef\_compare().
   * Vérifiez l’absence de problèmes de multicolinéarité en calculant les Variance Inflation Factors (VIF) avec car::vif() pour chaque modèle.

*Note : Vous répéterez ces étapes pour chacun des trois scores comme variable réponse.*