IA School
M2 Data Analyst
Intervenante :
Nouha Othman
2023-2024

Projet de semestre
Analyse de données avec R

1. Informations pratiques :

— À réaliser individuellement
— Le projet doit être implémenté avec R (Le fichier R doit être parmi les livrables.)
— À rendre sur l'espace prévu sur cet effet (classroom) avant le 12 Mai 2024.

2. Objectifs :

Apprendre à utiliser le logiciel R pour analyser des données. Mettre en oeuvre dans R les méthodes de statistique descriptive, décisionnelle, analyse de la variance, régression linéaire, analyse univariée et multivariée de données.

3. Description du projet :

Présentation des données sources et du sujet d'analyse :
Vous devez présenter un sujet d'analyse de votre choix, idéalement (mais pas obligatoirement) , en relation avec votre mission en alternance. Le jeu de données (ou les jeux de données) correspondant à ce sujet doit contenir suffisamment de données pour pouvoir l'analyser avec les différentes approches théoriques vues dans le cours statistiques.
Vous pouvez aussi choisir un dataset disponible dans le RStudio, voir la commande :
> library(help="datasets")

```
AirPassengers            Monthly Airline Passenger Numbers 1949-1960
BJsales                  Sales Data with Leading Indicator
BOD                      Biochemical Oxygen Demand
CO2                      Carbon Dioxide Uptake in Grass Plants
UCBAdmissions            Student Admissions at UC Berkeley
      ....
```

Réalisation du modèle de régression linéaire : L'objectif est de comprendre les relations existantes entre une variable à expliquer et des variables explicatives. Le choix des variables est important.
On propose de faire une régression simple puis, une régression multiple (au moins deux variables explicatives).
Pour chaque modèle, on vous demande de :

1. Décrire les relations statistiques entre les variables explicatives et la variaale dépendante avec les fonctions cor(), cor.test() et le R^2(coefficient de détermination). Analysez la forme de la relation avec le R^2
2. Vérifier les hypothèses de validité du modèle de régression linéaire
3. Valider le modèle de régression
4. Évaluer les points qui ont une grande influence sur la régression afin de les écarter s'il s'agit de points potentiellement aberrants.

Estimation de la pertinence du modèle

— Utilisez un indice qui permet de prouver qu'un modèle de régression est pertinent (RSS, de l'AIC, du BIC, du MSE, ..).

— Validez la pertinence des coefficients par les p-values et R^2 ajusté

Prédictions

Utilisez la commande predict afin de prédire les valeurs possibles à partir d'un modèle de régression.

NB : La commande predict permet de prédire les valeurs possibles à partir d'un modèle de régression.

Cette fonction permet aussi d'anticiper la fiabilité des valeurs.

Pour que predict() fonctionne, les données de prédictions doivent être sous forme de data.frame.

4. Livrables :

Vous devez soumettre une archive .zip ou .rar contenant :

— un script R bien commenté

— Un rapport détaillé au format PDF (de 10 pages maximum) contenant toutes les réponses : l'analyse détaillée avec les commandes saisies, les packages utilisés, des captures des résultas de R et les interprétations des résultats.

5. Évaluation :

— La correction des réponses et le respect des consignes (13 pt).

— La clarté et la pertinence du rapport (7 pt)

— 1 point de bonus pour toute autre amélioration (visualisation des données,...)

Lien utile :

Aide à l'utilisation de R :

blueRégressions linéaires avec R

blueShiny Dashboard