Ateliers L3 CMI

Quentin Hoarau

2026-02-16

Table des matières

Introduction 9
Présentation 9
Notation 9

I Général 10

Projets et présentations 11
Lien d'inscription 11
Résumé des consignes 11
Projet 1 11
Projet 2 11
Projet 3 12

Consignes Projet 1 13
Consignes Générales 13
Partie SIG 13
Partie Calcul Numérique et Econométrie 13
Évaluation 13

Consignes Projet 2 14
Consignes Générales 14
Partie SIG 15
Partie Économétrie 15
Rapport 15
Évaluation 16
Données 16
Liens : 16
Immobiliers 17
Eolien 17
Méthaniseurs 17
Inondations 17
Incendies 17
Installations polluantes 17

Bases de données 18

II Calcul Numérique 19

Introduction 20

TP1 : Commandes de base de R 21
1. Manipulation de vecteurs . . . 21
2. Manipulation de listes . . . 22
3. Manipulation de matrices . . . 22
4. Manipulation de chaînes de caractères . . . 23

TP2 : Tableaux de données 25
1. Les verbes de base de `dplyr` . . . 25
2. Enchaîner des opérations . . . 26
   `group_by` et `summarise` . . . 26
3. Jointures . . . 27
4. Bonus . . . 27

TP3 : Premiers Graphiques 28
Exercice 1 . . . 28
Exercice 2 . . . 29
Exercice 3 . . . 29
Exercice 4 . . . 31
Exercice 5 . . . 31

TP4 : Fonctions et manipulation de données 33
1. Mise en ordre des données . . . 33
   1.1 Représentations multiples d'un même jeu de données . . . 33
   1.2 Colonnes multiples représentant plusieurs variables . . . 34
2. Fonctions en R . . . 34
   2.1 Introduction et exemples . . . 34
   2.2 Arguments et résultat . . . 37
   2.3 Portée des variables . . . 38
   2.4 Les fonctions comme objets . . . 40
3. Fonctions et Dplyr . . . 42
   3.1 Appliquer ses propres fonctions . . . 42
   3.2 `across()` . . . 46
   3.3 Fonctions anonymes et notations abrégées . . . 48
   3.4 `rowwise()` et `c_across()` . . . 52

III SIG 55

Introduction 56

TP1 : Introduction à QGIS 57
1. Prise en main . . . 57
   1.1 Affichage / désaffichage des panneaux . . . 57
   1.2 Utilisation de données vecteur . . . 57
   1.3 Création d'un projet . . . 58
   1.4 Utilisation de l'outil Identifier les entités . . . 58
   1.5 Jointure 1 . . . 58
   1.6 Utilisation d'OpenStreetMap . . . 58
   1.7 Ordre des couches et opacité . . . 58
   1.8 Groupe de couches . . . 59
   1.9 Outil de mesure . . . 59
   1.10 Sélection et export . . . 59
   1.11 Sélection et conditions multiples . . . 59
2. Symbologie 1 : Pays du monde . . . 59
   2.1 Mise en page . . . 60
   2.2 Rendu gradué : Villes du monde . . . 60
   2.2 Symboles proportionnels . . . 60
3. Symbologie 2 : France . . . 60
   3.1 Styles . . . 61
   3.2 Étiquettes . . . 61
   3.3 Mise en page finale . . . 61

TP2 : Traitement sur les données vectorielles 62
1. Données de départ . . . 62
2. Création d'un GeoPackage . . . 62
3. Zone tampon sur l'autoroute A61 . . . 63
   3.1 Vérification du SCR . . . 63
   3.2 Conversion en Lambert 93 . . . 63
   3.3 Sélection et tampon . . . 63
   3.4 Analyse . . . 63
4. Matrice de distance . . . 64
5. Grille hexagonale . . . 64
   5.1 Création . . . 64
   5.2 Nettoyage . . . 64
6. Comptages dans la grille . . . 64
7. Analyse de superposition (Corine Land Cover) . . . 65
   7.1 Sélection des forêts . . . 65
   7.2 Superposition . . . 65
8. Intersection et Group Stats . . . 65
   8.1 Zone tampon des parcs . . . 65
   8.2 Intersection avec CLC . . . 65
   8.3 Tableau croisé dynamique . . . 66

TP3 : SIG avec R 67
1. Premières cartes . . . 67
2. Opérations sur les attributs . . . 68
3. Opération sur les données spatiales . . . 69
3.1 Opérations sur les vecteurs . . . 69
3.2 Opérations sur les rasters . . . 70
4. Opérations sur les géométries . . . 71
4.1 Opérations sur les vecteurs . . . 71
4.2 Opérations sur les rasters . . . 72
5. Application : rapprochement de base par distances . . . 72

IV Econométrie 1 74

Introduction 75

TP1 : Probabilités et Statistiques avec R 76
1. Probabilités avec R . . . 76
1.1 - Échantillonnage . . . 76
1.2 - Fonction de densité de probabilité . . . 76
1.3 - Espérance et Variance . . . 76
1.4 - Distribution Normale Standard . . . 77
1.5 - Distribution du Chi-carré . . . 77
1.6 - Distribution de Student . . . 77
1.7 - Distribution de Fisher . . . 77
2. Statistiques avec R . . . 78
2.1 - Biais . . . 78
2.2 - Efficience d'un estimateur . . . 78
2.3 - Test d'hypothèse . . . 79
2.4 - Test d'hypothèse : valeur-p . . . 79
2.5 - Corrélation . . . 80

TP 2 : Modèle de régression multiple 81

Introduction 82

Rappel sur R 83

Retour sur la régression multiple 84

Exercices 85
bwght . . . 85
hprice1 . . . 86
ceosal2 . . . 86

attend . . . 87
meap93 . . . 87
discrim . . . 88
charity . . . 88
htv . . . 89

TP3 : Analyse des Disparités Scolaires au Collège **91**
0. Installation. . . . 91
1. Données Brevet . . . 91
1.1 Description des données . . . 91
1.2 Evolution temporelle . . . 92
1.3 Variation en coupe . . . 92
2. Données socio-économiques . . . 92
2.1 Description du jeu de données . . . 92
2.2 Transformation du jeu . . . 92
3. Analyse jointe . . . 93
3.1 Jointure . . . 93
3.2 Analyse en coupe . . . 93
3.3 Regressions linéaire . . . 93

TP4 : Etude économétrique de l'Enquête Nationale Transport 2019 **94**
1. Enoncé . . . 94

V Econométrie 2 **95**

Introduction **96**

1 TP 1 : Modèle de régression multiple **97**
1.1 Introduction . . . 97
1.2 Rappel sur R . . . 97
1.3 Retour sur la régression multiple . . . 98
1.4 Exercices . . . 99
1.4.1 bwght . . . 99
1.4.2 hprice1 . . . 100
1.4.3 ceosal2 . . . 100
1.4.4 attend . . . 101
1.4.5 meap93 . . . 101
1.4.6 discrim . . . 102
1.4.7 charity . . . 102
1.4.8 htv . . . 103

2 TP 2 : Régression multiple et variables binaires **105**
2.1 Exercices . . . 105
2.1.1 gpa1 . . . 105
2.1.2 wage2 . . . 105
2.1.3 mlb1 . . . 106
2.1.4 gpa2 . . . 107
2.1.5 charity . . . 107
2.1.6 catholic . . . 108
2.1.7 apple . . . 108

3 Multicolinéarité **110**
3.0.1 Matrice de corrélation . . . 110
3.0.2 VIF . . . 111
3.0.3 Test de Farrar et Glauber . . . 111
3.0.4 Essais sur d'autres jeux de données . . . 111

4 Tests sur la forme fonctionnelle **112**
4.0.1 Test de Ramsey (RESET) . . . 112
4.1 Exercices . . . 113
4.1.1 kielmc . . . 113
4.1.2 wage1 . . . 113
4.1.3 gpa2 . . . 114
4.1.4 hprice1 . . . 114
4.1.5 vote1 . . . 115
4.1.6 hprice1 . . . 115
4.1.7 nbsal . . . 116
4.1.8 bwght2 . . . 116
4.1.9 apple . . . 117

5 Tests sur l'hétéroscédasticité **118**

6 Erreurs standards robustes **119**

7 Moindres carrés quasi-généralisés **120**

8 Exercices **121**
8.1 sleep75 . . . 121
8.2 vote1 . . . 121
8.3 pntsprd . . . 122
8.4 loanapp . . . 122
8.5 gpa1 . . . 123
8.6 meap00 . . . 123
8.7 fertil2 . . . 124
8.8 beauty . . . 124

9 Estimateurs pour les données de panel **125**
9.1 Exercices 125
9.1.1 fertil1 125
9.1.2 cps78_85 126
9.1.3 kielmc 126
9.1.4 injury 127
9.1.5 rental 127
9.1.6 jtrain 127
9.2 murder 128

Introduction

Présentation

Ce livre contient les supports des quatre ateliers spécifiques de la L3 CMI de l'université Paris Nanterre :

- Calcul numérique (S1) – 3 ECTS
- Système d'information géographique (S1) – 3 ECTS
- Econométrie 1 (S1) – 4.5 ECTS
- Econométrie 2 (S2) – 3 ECTS

La quasi-totalité des ateliers se font avec le langage R.

Notation

Toute la notation est en contrôle continu. Les notes viennent de :

- « Participation » qui inclut l'implication en classe, la qualité des rendus des TP et des pénalités pour retards et absences non justifiées
- Présentations en classe
- Projets de fin de semestre :
  - Peuvent être communs à plusieurs matières
  - Sujets donnés ou sujets libres

  - Rapport écrit propre (pas de compilation de code, ou de génération IA)
  - Présentation orale avec slides (sans les lire)

Toutes les projets sont recensés dans ce fichier Excel partagé.

Me contacter à qhoarau@parisnanterre.fr.

Première partie

Général

Projets et présentations

Lien d'inscription

Toutes les projets sont recensés dans ce fichier Excel partagé.

Résumé des consignes

Projet 1

Projet en deux parties :

- Partie 1 (30% du projet) : travail SIG. Réalisation et discussion d'une carte avec QGIS. Export de la carte en pdf/jpeg/… pas de capture d'écran.
- Partie 2 (70% du projet) : travail Calcul Numérique - Econométrie : réalisation d'un travail sur des données trouvées par l'étudiant.

Rendus : slides + script R. A déposer sur CEL.

Présentation : 15 minutes + 5 minutes de question

Date de présentation : mercredi 3 décembre à partir de 14h

Projet 2

Projet en deux parties :

- Partie 1 (40% du projet) : travail SIG
- Partie 2 (60% du projet) : travail Econométrie : réalisation d'un travail sur les données sur des données imposées.

Rendus : slides + rapport + (10 pages de texte à remettre la veille) + script R. Pas de capture d'écran. A deposer sur CEL.

Présentation : 15 minutes + 5 minutes de question

Date de présentation : mardi 16 décembre à partir de 8h30

Projet 3

Ce projet reprend le projet 2 mais avec une perspective d'inférence causale. Il s'agira d'utiliser une méthode de différences en différences échelonnées (*DiD*), voir une *event study*. Le traitement sera défini par le croisement la proximité à une installation (éolienne, usine) ou élément naturel (foret, cours d'eau) après l'apparititon. Par exemple, la proximité à une éolienne après sons ouverture, la proximité à un cours d'eau après une inondation, la proximité à une forêt après un incendie.

Consignes Projet 1

Consignes Générales

Partie SIG

Réalisation d'une carte à partir de données originales trouvées par l'étudiant. La carte doit répondre à une problématique

Exemples : - carte représentant la position de certains restaurants sur une carte chloropèthe sur les revenus

Partie Calcul Numérique et Econométrie

Réalisation d'analyse de données originales trouvées par l'étudiant. Pas nécessairement les mêmes que pour la partie SIG.

Évaluation

— *Rendus :* slides + carte (en pdf/jpg mais pas de capture d'écran)
— *Présentation :* 15 minutes + 5 minutes de question
— **Date de présentation : mercredi 3 décembre à partir de 14h**

Consignes Projet 2

Dans ce projet, vous devez combiner deux jeux de données pour faire une analyse économétrique et SIG. Tous les sujets s'intéressent aux effets de capitalisation dans l'immobilier : comment le marché immobilier valorise-t-il tel ou tel aspect de l'environnement proche du bien ?

Par exemple : - l'installation d'éoliennes, de méthaniseurs, ou de centrale solaire provoquent-elles des nuisances qui entrainent une baisse des prix des maisons ? - Les inondations font-elles perdre de la valeur aux maisons proches des cours d'eau ?

Tout l'enjeu de ces sujets est de quantifier précisément ces effets à l'aide de modèles économétriques du type :

$$log(valeurbien) = 1(distance < Xkm) + controles$$

avec $1(distance < Xkm)$ un dummy valant 1 si le bien est à moins de X km de la nuisance.

Consignes Générales

- Les sujets sont volontairement simples et vagues : il vous appartient de vous les approprier.
- Apporter du contexte avec vos recherches personnelles.
- Décrire les métadonnées : origine, limites, etc.
- Décrire les données : taille, horizon géographique et temporel, grandeurs et catégories contenues. Colonnes utiles vs inutiles. Nombre de données manquantes.
- Vous pouvez rapprocher les données que vous avez.
- Si les données sont trop lourdes, il est accepté de restreindre le jeu de données à une aire temporelle plus réduite.

Partie SIG

- Réaliser des cartes générales (choroplèthes) à l'échelle de la France pour chaque jeu de données.

- Réaliser des cartes combinées lorsqu'un même sujet fait apparaître deux jeux de données.
  - Exemple : une carte « zoom » avec des icônes différentes pour localiser les éoliennes et les maisons vendues à proximité.

- Calculer les distances pour rapprocher les deux bases :
  - Pour chaque bien immobilier, quelle est l'éolienne la plus proche ?

- Calculer les distances en intégrant la dimension temporelle.

- Restreindre le jeu de données aux observations pertinentes (ex. moins de 10 km autour de chaque évènement).
  - Exemple : enlever les maisons vendues à plus de 15 km d'une éolienne qui sera installée en 2024.

Partie Économétrie

- Faire des statistiques descriptives sur chaque jeu de données :
  - Exemple : évolution du nombre d'éoliennes installées/fermées, caractéristiques moyennes des évènements.

- Faire des statistiques descriptives sur la réunion des deux jeux de données :
  - Exemple : évolution du prix en fonction de la distance aux éoliennes.

- Estimer plusieurs régressions sur les prix de l'immobilier **sans** le deuxième jeu de données :
  - Exemple :
    $\log(Prix) = nb\_pieces + surface + departement + anne + ...$
- Estimer plusieurs régressions sur les prix de l'immobilier **avec** le deuxième jeu de données :
  - Exemple :
    $\log(Prix) = (distance < 1km) + nb\_pieces + surface + dep + anne + ...$

Rapport

- Construire une **problématique** :

- Exemple : *Quel est l'impact de l'installation d'éoliennes sur les prix des maisons en zone rurale ?*

- Expliciter le mécanisme à l'œuvre :
  - Exemple : *Une fermeture d'usine entraîne moins de travailleurs (donc baisse de la demande), mais aussi moins de pollution (hausse de la demande).*

- Plan imposé :
  1. Introduction + problématique + contexte + présentation des données
  2. SIG
  3. Économétrie

- Rédaction :
  - Pas moins de **10 pages**.
  - Écrit en **Word ou LaTeX** (pas de *knit* en Markdown).

- Présentation :
  - PowerPoint de **15 minutes** + **5 minutes** de questions.

Évaluation

- *Rendus :* slides et rapport sur CEL
- *Présentation :* 15 minutes + 5 minutes de question
- **Date de rendu du projet : 15 décembre à 16h**
- **Date de présentation : 16 décembre à partir de 8h30**

Données

Liens :

- Immobilier
- Eolien
- Solaire
- Methaniseurs
- Stations de recharge
- Parc véhicules électriques

— Incendies
— Cours d'eau
— Inondations
— Fiscalité locale
— Données Socioéconomiques
— CLC
— BPE
— Installations polluantes

Immobiliers

Lien :

Eolien

Attention aux coordonnées gps des éoliennes (il faut diviser les x /10 ou 100 selon)

Méthaniseurs

Inondations

L'idéal est de rajouter les distances aux cours d'eau pour capturer la vulnérabilité aux inondations.

Incendies

La base de données proposer porte sur les incendies de forets, en particulier dans l'aire Méditerranéenne. L'idéal est de rajouter les distances aux forêts pour capturer la vulnérabilité aux incendies de forêts.

Installations polluantes

Bases de données

Général :

- datagouv : répertoire public français
- datagov US : répertoire public américain
- datagov UK : répertoire public anglais
- kaggle : répertoire public pour data science
- TidyTuesday : projets collaboratifs de la Data Science Learning Community.
- Openintro : données éducatives

Statistiques publiques :

- DBnomics ou voir le package R `dbnomics`
- OCDE ou voir le package R `ocde`
- Eurostat ou voir le package R `eurostat`

Données socioéconomiques :

- Données du livre de Cagé & Piketty : données à la maille commune sur les revenus, education, résultats électoraux…

Données SIG :

- IGN
- Natural Earth

Données Energie/environnement :

- SDES (Ministère Transition Ecologique).
- ODRE (réseaux d'énergie).
- EEA (Agence européenne de l'environnement).
- Données incendies.
- Données prix carburants.

Deuxième partie

Calcul Numérique

Introduction

Le cours en résumé :

— 24 h
— Objectifs : prise en main de R et markdown, statistiques descriptives, graphiques
— Langage : R & Power BI (introduction si a le temps)
— Modalités d'examens :
— Note de participation
— Un projet commun calcul numérique/économétrie avec choix du sujet libre
— Présentation individuelle d'un chapitre d'un livre sur les bonnes pratiques de la dataviz : (Fundamentals of Data Visualization) de Clause Wilke. Inscription (ici).

Le plan :

1. Commandes des bases de R
2. Les tableaux de données en R
3. Les graphiques

En général pour R, regardez ce bon tutoriel sur le `tidyverse`.

TP1 : Commandes de base de R

Pour tout ce TP, essayez de chercher par vous-même les réponses sans utiliser de d'assistant IA. Pour ce faire, passez par google, ou l'outil d'aide de R. En général, regardez ce bon tutoriel sur le `tidyverse`.

1. Manipulation de vecteurs

Soit le vecteur $x = (1, 2, 3, 4, 5)$

1. Créer ce vecteur dans R et le stocker dans un objet que l'on appellera `x` ;
2. Afficher le mode de `x`, puis sa longueur ;
3. Extraire le premier élément, puis le dernier ;
4. Extraire les trois premier éléments et les stocker dans un vecteur que l'on nommera `a`
5. Extraire les éléments en position 1, 3, 5 ; les stocker dans un vecteur que l'on nommera `b`
6. Additionner le nombre 10 au vecteur `x`, puis multiplier le résultat par 2 ;
7. Effectuer l'addition de `a` et `b`, commenter le résultat ;
8. Effectuer l'addition suivante : `x+a`, commenter le résultat, puis regarder le résultat de `a+x` ;
9. Multiplier le vecteur `x` par le scalaire `c` que l'on fixera à 2 ;
10. Effectuer la multiplication de `a` et `b`, commenter le résultat ;
11. Effectuer la multiplication suivante : `x*a`, commenter le résultat ;
12. Récupérer les positions des multiples de 2 du vecteur `x` et les stocker dans un vecteur que l'on nommera `ind`, puis conserver uniquement les multiples de 2 de `x` dans un vecteur que l'on nommera `mult_2` ;
13. Afficher les éléments de `x` qui sont multiples de 3 et multiples de 2 ;
14. Afficher les éléments de `x` qui sont multiples de 3 ou multiples de 2 ;
15. Calculer la somme des éléments de `x` ;
16. Remplacer le premier élément de `x` par un 4 ;
17. Remplacer le premier élément de `x` par la valeur *NA*, puis calculer la somme des éléments de `x` ;
18. Lister les objets en mémoire dans la session R ;
19. Supprimer le vecteur `a` ;
20. Supprimer la totalité des objets de la session.

2. Manipulation de listes

1. Évaluer le code suivant : TRUE+FALSE+TRUE*4 et le commenter ;
2. Évaluer les expressions suivantes : `c(1, 4, TRUE)`, et `c(1, 4, TRUE, "bonjour")`, commenter ;
3. Créer une liste que l'on appellera `l` et qui contient les éléments 1, 4 et TRUE en première, seconde et troisième positions respectivement ;
4. Extraire le premier élément de la liste `l`, et afficher son mode. En faire de même avec le troisième élément, et commenter ;
5. Ajouter un quatrième élément à la liste `l` : "bonjour", puis afficher la structure de `l` ;
6. Retirer le troisième élément de la liste `l` ;
7. Créer une liste de trois éléments : votre nom, votre prénom, et votre année de naissance. Ces trois éléments de la liste devront être nommés respectivement `"nom"`, `"prenom"` et année de naissance. Stocker la liste ainsi créée dans un objet nommé `moi` ;
8. Extraire le prénom de la liste moi de deux manières : en utilisant l'indice, et en utilisant le nommage ;
9. Créer une liste avec la même structure que celle de `moi`, en la remplissant avec les informations d'une autre personne et la nommer `toi` Puis, créer la liste `personnes`, qui contiendra les listes `toi` et `moi` ;
10. Extraire la liste `toi` de personnes (en première position) ;
11. Extraire directement depuis `personnes` le prénom de l'élément en première position.

---

3. Manipulation de matrices

1. Créer la matrice suivante :
$$A = \begin{bmatrix} -3 & 5 & 6 \\ -1 & 2 & 2 \\ 1 & -1 & -1 \end{bmatrix}$$
;
2. Afficher la dimension de A, son nombre de colonnes, son nombre de lignes et sa longueur ;
3. Extraire la seconde colonne de A, puis la première ligne ;
4. Extraire l'élément en troisième position à la première ligne ;
5. Extraire la sous-matrice de dimension 2 x 2 du coin inférieur de A
6. Calculer la somme des colonnes puis des lignes de A
7. Afficher la diagonale de A ;
8. Rajouter le vecteur colonne (1, 2 , 3) à droite de la matrice A et stocker le résultat dans un objet appelé B ;

9. Retirer le quatrième vecteur de B ;
10. Retirer la première et la troisième ligne de B ;
11. Ajouter le scalaire 10 à A ;
12. Ajouter le vecteur colonne (1 2 3) à A ;
13. Ajouter la matrice identité I_3 à A ;
14. Diviser tous les éléments de la matrice A par 2 ;
15. Multiplier la matrice A par le vecteur colonne (1 2 3) ;
16. Afficher la transposée de A ;
17. Effectuer le produit avec transposition $A^t A$.

---

4. Manipulation de chaînes de caractères

Charger le package `tidyverse`, qui contient le package `stringr`.

1. Créer les objets `a` et `b` afin qu'il contiennent respectivement les chaînes de caractères suivantes : *23 à 0* et *C'est la piquette, Jack !* ;
2. Créer le vecteur `phrases` de longueur 2, dont les deux éléments sont a et b ;
3. À l'aide de la fonction appropriée dans le package stringr, afficher le nombre de caractètres de a, de b, puis appliquer la même fonction à l'objet phrases ;
4. En utilisant la fonction `str_c()`, concaténer a et b dans une seule chaîne de caractères, en choisissant la virgule comme caractère de séparation ;
5. Concaténer les deux éléments du vecteur phrases en une seule chaine de caractères, en les séparant par le caractère de retour à la ligne , puis utiliser la fonction `cat()` pour afficher le résultat ;
6. Appliquer la même fonction que dans la question précédente à l'objet suivant : `c(NA, phrases)` et commenter ;
7. Mettre en majuscules, puis en minuscules les chaînes du vecteur phrases (afficher le résultat, ne pas modifier phrases) ;
8. À l'aide de la fonction word() du package stringr, extraire le mot la, puis Jack de la chaîne `b` ;
9. Même question que la précédente, en utilisant la fonction `str_sub()` ;
10. À l'aide de la fonction `str_detect()`, rechercher si le motif `piqu` puis `mauvais` sont présents dans `b` ;
11. À l'aide de la fonction `str_detect()`, rechercher si le motif `piqu` est présent dans les éléments du vecteur phrases ;
12. À l'aide de la fonction `str_detect()`, rechercher si le motif `piqu` ou le motif `à` sont présents dans les éléments du vecteur phrases ;

13. En utilisant la fonction `str_locate()`, retourner les positions de la première occurence du caractère `a` dans la chaîne `b`, puis essayer avec le caractère `w` pour observer le résultat retourné ;
14. Retourner toutes les positions du motif `a` dans la chaîne `b` ;
15. En utilisant la fonction `str_replace()`, remplacer la première occurence du motif `a`, par le motif `Z` (afficher le résultat, ne pas modifier `phrases`) ;
16. Remplacer toutes les occurences de `a` par `Z` dans la chaîne `b` (afficher le résultat, ne pas modifier `phrases`) ;
17. Utiliser la fonction `str_split()` pour séparer la chaîne `b` en utilisant la virgule comme séparateur de sous-chaînes ;
18. Retirer tous les caractères de ponctuation de la chaîne `b`, puis utiliser la fonction `tr_trim()` sur le résultat pour retirer les caractères blancs du début et de la fin de la chaîne.

TP2 : Tableaux de données

On commence par charger les extensions et les données nécessaires.

```
library(tidyverse)
library(nycflights13)
data(flights)
data(airports)
data(airlines)
```

1. Les verbes de base de dplyr

Exercice 1.1

Sélectionner la dixième ligne du tableau des aéroports (`airports`).

Sélectionner les 5 premières lignes de la table `airlines`.

Sélectionner l'aéroport avec l'altitude la plus basse.

Exercice 1.2

Sélectionnez les vols du mois de juillet (variable `month`).

Sélectionnez les vols avec un retard à l'arrivée (variable `arr_delay`) compris entre 5 et 15 minutes.

Sélectionnez les vols des compagnies Delta, United et American (codes `DL`, `UA` et `AA` de la variable `carrier`).

Exercice 1.3

Triez la table `flights` par retard au départ décroissant.

Exercice 1.4

Sélectionnez les colonnes `name`, `lat` et `lon` de la table `airports`

Sélectionnez toutes les colonnes de la table `airports` sauf les colonnes `tz` et `tzone`

Sélectionnez toutes les colonnes de la table `flights` dont les noms se terminent par "delay".

Dans la table `airports`, renommez la colonne `alt` en `altitude` et la colonne `tzone` en `fuseau_horaire`.

Exercice 1.5

Dans la table `airports`, la colonne `alt` contient l'altitude de l'aéroport en pieds. Créer une nouvelle variable `alt_m` contenant l'altitude en mètres (on convertit des pieds en mètres en les divisant par 3.2808). Sélectionner dans la table obtenue uniquement les deux colonnes `alt` et `alt_m`.

2. Enchaîner des opérations

Exercice 2.1

Réécrire le code de l'exercice précédent en utilisant le *pipe* `%>%`.

Exercice 2.2

En utilisant le *pipe*, sélectionnez les vols à destination de San Francico (code `SFO` de la variable `dest`) et triez-les selon le retard au départ décroissant (variable `dep_delay`).

Exercice 2.3

Sélectionnez les vols des mois de septembre et octobre, conservez les colonnes `dest` et `dep_delay`, créez une nouvelle variable `retard_h` contenant le retard au départ en heures, et conservez uniquement les 5 lignes avec les plus grandes valeurs de `retard_h`.

`group_by` et `summarise`

Exercice 3.1

Affichez le nombre de vols par mois.

Triez la table résultat selon le nombre de vols croissant.

Exercice 3.2

Calculer la distance moyenne des vols selon l'aéroport de départ (variable `origin`).

Exercice 3.3

Calculer le nombre de vols à destination de Los Angeles (code `LAX`) pour chaque mois de l'année.

Exercice 3.4

Calculer le nombre de vols selon le mois et la destination.

Ne conserver, pour chaque mois, que la destination avec le nombre maximal de vols.

Exercice 3.5

Calculer le nombre de vols selon le mois. Ajouter une colonne comportant le pourcentage de vols annuels réalisés par mois.

Exercice 3.6

Calculer, pour chaque aéroport de départ et de destination, la durée moyenne des vols (variable air_time). Pour chaque aéroport de départ, ne conserver que la destination avec la durée moyenne la plus longue.

3. Jointures

Exercice 4.1

Faire la jointure de la table `airlines` sur la table `flights` à l'aide de `left_join`.

Exercice 4.2

À partir de la table résultat de l'exercice précédent, calculer le retard moyen au départ pour chaque compagnie, et trier selon ce retard décroissant.

Exercice 4.3

Faire la jointure de la table `airports` sur la table `flights` en utilisant comme clé le code de l'aéroport de destination.

À partir de cette table, afficher pour chaque mois le nom de l'aéroport de destination ayant eu le plus petit nombre de vol.

Exercice 4.4

Créer une table indiquant, pour chaque vol, uniquement le nom de l'aéroport de départ et celui de l'aéroport d'arrivée.

4. Bonus

Exercice 5.1

Calculer le nombre de vols selon l'aéroport de destination, et fusionnez la table `airports` sur le résultat avec `left_join`. Stocker le résultat final dans un objet nommé `flights_dest`.

TP3 : Premiers Graphiques

Exercice 1

1. Avant toute chose, charger `tidyverse`. Charger aussi le jeu de données `rp2018` dans le package `questionr`. Assigner un dataframe `rp69` comme la restriction de `rp2018` aux départements du Rhône et de la Loire. Faire un nuage de points croisant le pourcentage de sans diplôme (`dipl_aucun`) et le pourcentage d'ouvriers (`ouvr`).

2. Faire un nuage de points croisant le pourcentage de sans diplôme et le pourcentage d'ouvriers, avec les points en rouge et de transparence 0.2.

3. Représenter la répartition du pourcentage de propriétaires (variable `proprio`) selon la taille de la commune en classes (variable `pop_cl`) sous forme de boîtes à moustaches.

4. Représenter la répartition du nombre de communes selon la taille de la commune en classes sous la forme d'un diagramme en bâtons.

5. Faire un nuage de points croisant le pourcentage de sans diplôme et le pourcentage d'ouvriers. Faire varier la couleur selon le département (`departement`).

6. Sur le même graphique, faire varier la taille des points selon la population totale de la commune (`pop_tot`).

7. Enfin, toujours sur le même graphique, rendre les points transparents en plaçant leur opacité à 0.5.

9. Représenter la répartition du pourcentage de propriétaires (variable `proprio`) selon la taille de la commune en classes (variable `pop_cl`) sous forme de boîtes à moustaches. Faire varier la couleur de remplissage (attribut `fill`) selon le département.

10. Représenter la répartition du nombre de communes selon la taille de la commune en classes (variable `pop_cl`) sous forme de diagramme en bâtons empilés, avec une couleur différente selon le département.

11. Faire varier la valeur du paramètre `position` pour afficher les barres les unes à côté des autres.

12. Changer à nouveau la valeur du paramètre `position` pour représenter les proportions de communes de chaque département pour chaque catégorie de taille.

13. Faire un nuage de points représentant en abscisse le pourcentage de cadres (`cadres`) et en ordonnée le pourcentage de diplômés du supérieur (`dipl_sup`). Représenter ce nuage par deux graphiques différents selon le département en utilisant `facet_grid`.

14. Sur le même graphique, faire varier la taille des points selon la population totale de la communes (variable `pop_tot`) et rendre les points transparents.

15. Faire le nuage de points croisant pourcentage de chômeurs (`chom`) et pourcentage de sans diplôme. Y ajouter les noms des communes correspondant (variable `commune`), en rouge et en taille 2.5 :

16. Dans le graphique précédent, n'afficher que le nom des communes ayant plus de 15% de chômage.

Exercice 2

Avant tout, charger le package `tidyverse`.

1. Utiliser la fonction `data()` pour charger en mémoire le jeu de données `economics.` Consulter la page d'aide de ce jeu de données pour prendre connaissance de son contenu ;
2. À l'aide de la fonction `ggplot()`, représenter les dépenses personnelles de consommation (`pce`) en fonction de la date (`date`). Les observations doivent être connectées par une ligne.
3. Modifier le graphique de la question précédente de manière à ce que la couleur de la ligne soit dodger blue et définir la taille de la ligne à 0.5. Stocker le résultat dans un objet que l'on appellera `p_1` ;
4. Ajouter une couche au graphique `p_1` pour modifier les titres des axes (les retirer), et définir le titre suivant : *"Personal Consumption Expenditures (billions dollars)"*.
5. Utiliser la fonction `scale_x_date()` du package `scales` pour modifier l'échelle des abscisses de `p_1`, afin que les étiquettes des marques soient affichées tous les 5 ans ;
6. A l'aide de l'option `date_labels()` de la fonction précédente, modifier le format de ces étiquettes pour que seule l'année des dates s'affiche ;

Exercice 3

1. Utiliser la fonction data() pour charger en mémoire le jeu de données `economics.` Consulter la page d'aide de ce jeu de données pour prendre connaissance de son contenu ;

2. Sélectionner les variables `date`,`psavert` et `uempmed` dans le tableau de données economics et utiliser la fonction `gather()` sur le résultat pour obtenir un tableau dans lequel chaque ligne indiquera la valeur (`value`) pour une variable donnée (`key`) à une date donnée (`date`). Stocker le résultat dans un objet que l'on appellera `df` ;
3. Sur un même graphique, représenter à l'aide de lignes, l'évolution dans le temps du taux d'épargne personnelle (psavert) et de la durée médiane en semaines du chômage (`uempmed`). Stocker le graphique dans un objet que l'on appellera `p_2` ;
4. Modifier le code ayant servi à construire le graphique `p_2` pour que le type de ligne soit différent pour chacune des deux séries représentées. Les deux lignes doivent être tracées en bleu. Stocker le graphique dans un objet que l'on appellera `p_3` ;
5. À présent, modifier le code ayant servi à construire `p_3` pour qu'à la fois la couleur et le type de ligne servent à différencier les deux séries. Stocker le graphique dans un objet que l'on appellera `p_4` ;
6. Modifier le graphique `p_4` en ajoutant une couche d'échelle de couleur pour que le taux d'épargne personnelle (`psavert`) soit représenté en dodger blue, et que la durée de chômage (`uempmed`) soit représentée en rouge. Par ailleurs, retirer le titre de la légende ;
7. Modifier le graphique `p_4` en ajoutant une couche d'échelle de type de ligne pour que le taux d'épargne personnelle (`psavert`) soit représenté par des tirets, et que la durée de chômage (`uempmed`) soit représentée par une ligne pleine. Par ailleurs, retirer le titre de la légende des types de lignes, afin que les légendes de couleur et de type de ligne soient fusionnées ;
8. Créer le tableaux de données df_2, une copie de df, dans lequel la variable key doit être un facteur dont les niveaux sont uempmed et psavert ;
9. Créer le vecteur etiq suivant `etiq <- c("psavert" = "Pers. Saving Rate","uempmed" = "Median Duration of Unemployment (weeks)")` Ce vecteur contient des valeurs d'étiquettes pour la légende du graphique qu'il va falloir créer. Représenter sur un même graphique l'évolution dans le temps du taux d'épargne personnelle et de la durée médiane du chômage en semaines, en s'appuyant sur les données contenues dans le tableau `df_2`. La courbe du taux d'épargne personnelle doit être composée de tirets et être de couleur dodger blue ; la courbe de la durée médiane du taux de chômage doit être une ligne rouge. La légende ne doit pas comporter de titre, et ses étiquettes doivent être modifiées pour que *"Pers. Saving Rate"* s'affiche 'a la place de *"psavert"*, et pour que *"Median Duration of Unemployment (weeks)"* s'affiche à la place de *"uempmed"*. Stocker le graphique dans un objet que l'on appellera `p_5` ; Note : il s'agit de reprendre le code ayant servi à créer le graphique `p_4`, en modifiant légèrement les échelles de couleur et de ligne pour prendre en compte les étiquettes proposées dans le vecteur `etiq`.
10. Modifier `p_5` pour lui ajouter une couche permettant de déplacer la légende en bas du graphique (utiliser la fonction theme()) ;
11. Ajouter une couche au graphique `p_5` qui permet de définir un thème. Utiliser le thème minimal (`theme_minimal()`). Que se passe-t-il pour la légende ? Repositionner la légende en dessous, et retirer les titres des axes ;

12. Sauvegarder le graphique `p_5` au format PDF en précisant une largeur de 12 et une hauteur de 8.

Exercice 4

1. Charger le jeu de données `mpg` contenu dans lepackage ggplot2 en mémoire, puis consulter la page d'aide du jeu de données pour en prendre connaissance ;
2. Représenter à l'aide d'un nuage de points la relation entre la consommation sur autoroute des véhicules de l'échantillon (`hwy`) et la cylindrée de leur moteur (`displ`)
3. Reprendre le code du graphique précédent et modifier la forme des points pour les changer en symbole + ; modifier la couleur des + de manière à la faire dépendre du nombre de cylindres (`cyl`) ;
4. À présent, représenter par des boîtes à moustaches la distribution de la consommation sur autoroute des véhicules (`hwy`) pour chacune des valeurs possibles du nombre de cylindres (`cyl`) ;
5. Charger le jeu de données `economics` contenu dans le package `ggplot2` en mémoire, puis consulter la page d'aide du jeu de données pour en prendre connaissance. Ensuite, ajouter au tableau (les créer) les variables `u_rate` et `e_rate`, respectivement le taux de chômage et le taux d'emploi (on définira le taux de chômage de manière très grossière ici : nombre de personnes non employées sur la population totale) ;
6. Représenter à l'aide de barres l'évolution dans le temps du taux de chômage, et remplir les barres avec la couleur rouge ;
7. Reprendre le code du graphique précédent et ajouter une couche permettant de modifier les limites de l'axe des abscisses pour afficher les valeurs uniquement sur la période "2012-01-01" à "2015-01-01" (utiliser la fonction `coord_cartesian()`). Stocker le graphique dans un objet que l'on appellera p ;
8. Dans le tableau de données economics, sélectionner les variables date, `u_rate` et `e_rate`, puis utiliser la fonction `gather()` pour obtenir un tableau dans lequel chaque ligne correspond à la valeur d'une des variables (taux de chômage ou taux d'emploi) à une date donnée. Stocker le résultat dans un objet que l'on appellera `df_3` ;
9. Utiliser le tableau de données `df_3` pour représenter graphiquement à l'aide de barres les taux de chômage et taux d'emploi par mois sur la période "2012-01-01" à "2015-01-01". Sur le graphique, les barres représentant le taux de chômage et celles représentant le taux d'emploi devront être superposées. Note : il s'agit de modifier légèrement le code ayant permis de réaliser le graphique `p`.

Exercice 5

1. À l'aide de la fonction `WDI` du package `WDI`, récupérer la série fournie par la Banque Mondiale du PIB par tête (*NY.GDP.PCAP.PP.KD*) pour tous les pays disponibles pour

l'année 2010, et stocker ces données dans un tableau que l'on appellera `gdp_capita` ;
2. Dans le tableau `gdp_capita`, modifier la valeur de la variable country pour l'observation de la Slovaquie, pour qu'elle vaille Slovakia au lieu de Slovak Republic ;
3. Filtrer les observations du tableau gdp_capita pour ne conserver que les observations des pays membres de l'Union Européenne dont les noms sont contenus dans le vecteur membres_ue. Stocker le résultat dans un tableau que l'on nommera gdp_capita_eu ;
4. Utiliser le package `rworldmap` pour récupérer les données nécessaires à la réalisation d'une carte du monde ;
5. Afficher une carte du monde à l'aide des fonctions contenues dans le package `ggplot2` ;
6. Modifier les échelles des axes pour faire figurer les méridiens de -60 à 60 par pas de 30 et les parallèles de -180 à 180 par pas de 45. Modifier également la projection cartographique pour choisir la projection orthographique, à l'aide de la fonction `coord_map()` ;
7. Joindre les informations contenues dans le tableau `gdp_capita_eu` au tableau contenant les données permettant la réalisation des cartes ;
8. Réaliser une carte choroplèthe reflétant pour chaque pays membre de l'Union Européenne la valeur du PIB par tête de 2012 ;
9. Modifier les couleurs de l'échelle continue de la carte précédente, pour que les faibles valeurs du PIB par tête soient représentées en jaune, et les valeurs les plus hautes en rouge ;
10. Modifier les ruptures de l'échelle de couleur pour qu'elles aillent de 10000 à 100000 ; modifier également l'étiquette de ces ruptures de sorte que 35000 soit affiché comme 35k, 60000 comme 60k, etc. Enfin, ajouter un titre au graphique et retirer les titres d'axes.

TP4 : Fonctions et manipulation de données

1. Mise en ordre des données

Pour reprendre la définition du manuel de Julien Barnier, le concept de tidy data repose sur trois règles interdépendantes. Des données sont considérées comme *tidy* si :

— chaque ligne correspond à une observation
— chaque colonne correspond à une variable
— chaque valeur est présente dans une unique case de la table ou, de manière équivalente, des unités d'observations différentes sont présentes dans des tables différentes

Dans cette partie, nous allons travailler sur les fonctions du package `tidyr` : `pivot_longer()`, `pivot_wider()`, `separate()` et `unite()`.

Les jeux de données utilisés (`table1`, `table2`, `table3`, `table4a`, `table4b`, `table5`) sont directement disponibles après avoir chargé `tidyverse.`

```
library(tidyverse)

data(table1)
data(table2)
data(table3)
data(table4a)
data(table4b)
data(table5)
```

1.1 Représentations multiples d'un même jeu de données

Avec `table1`, `table2`, et `table3` On vous donne trois représentations du nombre de cas de tuberculose (TB) par pays et par année.

1. Affichez les trois jeux de données et comparez leurs structures.

— Quelles différences remarquez-vous ?
— Quelles sont les variables présentes dans chacun ?

2. Parmi eux, lequel est déjà au format tidy ? Justifiez.
3. Transformez `table2` et `table3` en jeux de données tidy comparables à `table1.` Indice : utilisez `pivot_wider()` et `pivot_longer()`.

1.2 Colonnes multiples représentant plusieurs variables

Les tableaux `table4a` et `table4b` contiennent respectivement le nombre de cas et la population.

2. Fonctions en R

En R, la syntaxe des fonctions est la suivante :

```
ajoute <- function(x,y) {
  res <- x + y
  return(res)
}
```

2.1 Introduction et exemples

Exercice 1.1

Écrire une fonction nommée `perimetre` qui prend en entrée un argument nommé `r` et retourne le périmètre d'un cercle de rayon `r`, c'est-à-dire `2 * pi * r` (`pi` est un objet R qui contient la valeur de π).

Vérifier avec l'appel suivant :

```
perimetre(4)
```

```
perimetre <- function(r) {
  resultat <- 2 * pi * r
  return(resultat)
}
```

Exercice 1.2

Écrire une fonction `etendue` qui prend en entrée un vecteur numérique et retourne la différence entre la valeur maximale et la valeur minimale de ce vecteur.

Vérifier avec l'appel suivant :

```r
etendue(c(18, 35, 21, 40))
```

```r
etendue <- function(v) {
  vmax <- max(v)
  vmin <- min(v)
  return(vmax - vmin)
}
```

Exercice 1.3

Écrire une fonction nommée `alea` qui accepte un argument `n`, génère un vecteur de `n` valeurs aléatoires entre 0 et 1 avec la fonction `runif(n)` et retourne ce vecteur comme résultat.

```r
alea <- function(n) {
  v <- runif(n)
  return(v)
}
```

Modifier la fonction pour qu'elle accepte deux arguments supplémentaires `min` et `max` et qu'elle retourne un vecteur de `n` valeurs aléatoires comprises entre `min` et `max` avec la fonction `runif(n, min, max)`.

```r
alea <- function(n, min, max) {
  v <- runif(n, min, max)
  return(v)
}
```

Modifier à nouveau la fonction pour qu'elle retourne un vecteur de `n` nombres *entiers* aléatoires compris entre `min` et `max` en appliquant la fonction `trunc()` au vecteur généré par `runif()`.

Vérifier le résultat avec :

```r
v <- alea(10000, 1, 6)
table(v)
```

```r
alea <- function(n, min, max) {
  v <- runif(n, min, max + 1)
  v <- trunc(v)
  return(v)
}
```

Exercice 1.4

Écrire une fonction nommée `meteo` qui prend un argument nommé `ville` avec le corps suivant :

```
out <- readLines(paste0("https://v2.wttr.in/", ville, "?A"))
cat(out, sep = "\n")
```

Tester la fonction avec par exemple `meteo("Lyon")` (il est possible que l'affichage dans la console ne soit pas lisible si vous travaillez sous Windows).

```
meteo <- function(ville) {
  out <- readLines(paste0("https://v2.wttr.in/", ville, "?A"))
  cat(out, sep = "\n")
}
```

Exercice 1.5

Soit le code suivant, qui recode une variable du jeu de données `hdv2003` en utilisant `str_to_lower()` puis `fct_recode()` :

```
library(questionr)
library(tidyverse)
data(hdv2003)

hdv2003$hard.rock <- str_to_lower(hdv2003$hard.rock)
hdv2003$hard.rock <- fct_recode(hdv2003$hard.rock, "o" = "oui", "n" = "non")
```

Transformer ce code en une fonction nommée `recode_oui_non`, et appliquer cette fonction à `hard.rock`, `lecture.bd` et `cuisine`.

```
recode_oui_non <- function(var) {
  var_rec <- str_to_lower(var)
  var_rec <- fct_recode(var_rec, "o" = "oui", "n" = "non")
  return(var_rec)
}

hdv2003$hard.rock <- recode_oui_non(hdv2003$hard.rock)
hdv2003$lecture.bd <- recode_oui_non(hdv2003$lecture.bd)
hdv2003$cuisine <- recode_oui_non(hdv2003$cuisine)
```

2.2 Arguments et résultat

Exercice 2.1

Observer le code de la fonction suivante pour comprendre à quoi correspondent chacun de ses trois arguments, puis réordonner et renommer ces arguments de manière plus pertinente :

```r
moyenne_arrondie <- function(d, vecteur_contenant_les_donnees, supprimer_les_na) {
  res <- mean(vecteur_contenant_les_donnees, na.rm = supprimer_les_na)
  res <- round(res, d)
  return(res)
}
```

```r
moyenne_arrondie <- function(v, decimales, na.rm) {
  res <- mean(v, na.rm = na.rm)
  res <- round(res, decimales)
  return(res)
}
```

Donner aux arguments de la fonction une valeur par défaut.

```r
moyenne_arrondie <- function(v, decimales = 2, na.rm = TRUE) {
  res <- mean(v, na.rm = na.rm)
  res <- round(res, decimales)
  return(res)
}
```

Simplifier la fonction en utilisant la syntaxe plus compacte qui ne fait pas appel à `return()`.

```r
moyenne_arrondie <- function(v, decimales = 2, na.rm = TRUE) {
  res <- mean(v, na.rm = na.rm)
  round(res, decimales)
}
```

Exercice 2.2

Simplifier la fonction suivante pour que son corps ne fasse plus qu'une seule ligne :

```r
centrer_reduire <- function(x) {
  res <- x - mean(x)
  res <- res / sd(x)
  return(res)
}
```

```r
centrer_reduire <- function(x) {
  (x - mean(x)) / sd(x)
}
```

Exercice 2.3

Le code suivant permet de déterminer la lettre initiale et la longueur d'un mot.

```r
initiale <- str_sub(mot, 1, 1)
longueur <- nchar(mot)
```

Utiliser ce code pour créer une fonction `caracteristiques_mot()` qui prend un argument `mot` et retourne à la fois son initiale et sa longueur.

```r
caracteristiques_mot("Bidonnage")
```

```r
caracteristiques_mot <- function(mot) {
  initiale <- str_sub(mot, 1, 1)
  longueur <- nchar(mot)
  list(initiale = initiale, longueur = longueur)
}
```

Facultatif : modifier la fonction pour qu'elle retourne un vecteur plutôt qu'une liste, et l'appliquer sur un mot de votre choix. Que constatez-vous ?

Comme les vecteurs atomiques ne peuvent contenir que des données du même type, le nombre correspondant à `longueur` a été converti en chaîne de caractères.

2.3 Portée des variables

Exercice 3.1

En lisant les codes suivants, essayer de prévoir quelle va être la valeur affichée par la dernière ligne. Vérifier en exécutant le code :

```r
f <- function() {
  x <- 3
  x
}

f()
```

```r
f <- function() {
  x
}

x <- 5
f()
```

```r
f <- function(x) {
  x
}

x <- 5
f(30)
```

```r
f <- function(x = 100) {
  x
}

x <- 5
f()
```

```r
f <- function(x = 100) {
  x <- 150
  x
}

x <- 5
f(30)
```

```r
f <- function() {
  x <- 5
}

x <- 1000
f()
x
```

Exercice 3.2

Dans le code suivant, on a essayé de créer une fonction qui modifie un tableau de données passé en argument pour ne conserver que les lignes correspondant aux pommes. Est-ce que ça fonctionne ?

```r
df <- data.frame(
  fruit = c("Pomme", "Pomme", "Citron", "Citron"),
  poids = c(147, 189, 76, 91)
)

filtre_pommes <- function(d) {
  d <- dplyr::filter(d, fruit == "Pomme")
}

filtre_pommes(df)
df
```

Modifier le code pour obtenir le résultat souhaité.

```r
filtre_pommes <- function(d) {
  dplyr::filter(d, fruit == "Pomme")
}

df <- filtre_pommes(df)
df
```

2.4 Les fonctions comme objets

Exercice 4.1

Écrire une fonction nommée `bonjour` qui ne prend aucun argument et affiche juste le texte "Bonjour !" dans la console.

```r
bonjour <- function() {
  cat("Bonjour !")
}
```

Exécuter dans la console les deux commandes suivantes tour à tour :

- `bonjour()`
- `bonjour`

Comprenez-vous la différence entre les deux ?

Copier la fonction dans un nouvel objet nommé `salut`. Exécuter la nouvelle fonction ainsi créée.

```
salut <- bonjour
salut()
```

Exercice 4.2

Construire une fonction `etendue()` qui prend en entrée un vecteur numérique et retourne la différence entre la valeur maximale et la valeur minimale de ce vecteur (vous pouvez récupérer le code de l'exercice 1.2).

```
etendue <- function(v) {
  max(v) - min(v)
}
```

À l'aide de `tapply()`, appliquez la fonction `etendue()` à la variable `age` pour chaque valeur de `qualif` dans le jeu de données `hdv2003`.

```
library(questionr)
data(hdv2003)

tapply(hdv2003$age, hdv2003$qualif, etendue)
```

Réécrire le code précédent en utilisant une fonction anonyme (*ie* en définissant la fonction directement dans le `tapply`).

```
tapply(hdv2003$age, hdv2003$qualif, function(v) {
  max(v) - min(v)
})
```

Exercice 4.3

Exécutez le code suivant. Comprenez-vous les résultats obtenus ?

```
f <- function(y) {
  y * 4
}

body(f)
f(5)

body(f) <- quote(y + 2)
body(f)
f(5)
```

Intuitivement, comprenez-vous à quoi sert la fonction `quote` ?

3. Fonctions et Dplyr

Pour certains des exercices qui suivent on utilisera le jeu de données `starwars`. Le jeu de données contient les caractéristiques de 87 personnages présents dans les films : espèce, âge, planète d'origine, etc.

3.1 Appliquer ses propres fonctions

Exercice 1.1

Créer une fonction `imc` qui prend en argument un vecteur `taille` (en cm) et un vecteur `poids` (en kg) et retourne les valeurs correspondantes de l'indice de masse corporelle, qui se calcule en divisant le poids en kilos par la taille en mètres au carré.

```r
imc <- function(tailles, poids) {
    tailles_m <- tailles / 100
    poids / tailles_m ^ 2
}
```

Utiliser cette fonction pour ajouter une nouvelle variable `imc` au tableau `starwars`.

```r
starwars %>%
    mutate(imc = imc(height, mass))
```

À l'aide de `group_by()` et `summarise()`, utiliser à nouveau cette fonction pour calculer l'IMC moyen selon les valeurs de la variable `species`.

```r
starwars %>%
    group_by(species) %>%
    summarise(
        imc = mean(imc(height, mass), na.rm = TRUE)
    )
```

Exercice 1.2

Toujours dans le jeu de données `starwars`, à l'aide d'un `group_by()` et d'un `summarise()`, calculer pour chaque valeur de la variable `sex` la valeur de l'étendue de la variable `height` du jeu de données `starwars`, c'est-à-dire la différence entre sa valeur maximale et sa valeur minimale.

```
starwars %>%
    group_by(sex) %>%
    summarise(
        etendue_taille = max(height, na.rm = TRUE) - min(height, na.rm = TRUE)
    )
```

En partant du code précédent, créer une fonction `etendue` qui prend en argument un vecteur et retourne la différence entre sa valeur maximale et sa valeur minimale. En utilisant cette fonction, calculer pour chaque valeur de `sex` la valeur de l'étendue des variables `height` et `mass`.

```
etendue <- function(v) {
    max(v, na.rm = TRUE) - min(v, na.rm = TRUE)
}
starwars %>%
    group_by(sex) %>%
    summarise(
        etendue_taille = etendue(height),
        etendue_poids  = etendue(mass)
    )
```

Exercice 1.3

On a vu que la fonction suivante permet de calculer le pourcentage des éléments d'un vecteur de chaînes de caractères se terminant par un suffixe passé en argument.

```
prop_suffixe <- function(v, suffixe) {
    # On ajoute $ à la fin du suffixe pour capturer uniquement en fin de chaîne
    suffixe <- paste0(suffixe, "$")
    # Détection du suffixe
    nb_detect <- sum(str_detect(v, suffixe))
    # On retourne le pourcentage
    nb_detect / length(v) * 100
}
```

Modifier cette fonction en une fonction `prop_prefixe` qui retourne le pourcentage d'éléments commençant par un préfixe passé en argument. *Indication :* pour détecter si une chaîne commence par `"ker"`, on utilise l'expression régulière `"^ker"`.

```
prop_prefixe <- function(v, prefixe) {
    # On ajoute $ à la fin du prefixe pour capturer uniquement en début de chaîne
    prefixe <- paste0("^", prefixe)
    # Détection du motif
    nb_detect <- sum(str_detect(v, prefixe))
    # On retourne le pourcentage
    nb_detect / length(v) * 100
}
```

Utiliser `prop_prefixe` dans un `summarise` appliqué à `rp2018` pour calculer le pourcentage de communes commençant par "Saint" selon le département. Ordonner les résultats par pourcentage décroissant.

```
rp2018 %>%
    group_by(departement) %>%
    summarise(
        prop_saint = prop_prefixe(commune, "Saint")
    ) %>%
    arrange(desc(prop_saint))
```

Créer une fonction `tab_prefixe` qui prend un seul argument `prefixe` et renvoie le tableau obtenu à la question précédente pour le préfixe passé en argument. Tester avec `tab_prefixe("Plou")` et `tab_prefixe("Sch")`

```
tab_prefixe <- function(prefixe) {
    rp2018 %>%
        group_by(departement) %>%
        summarise(
            prop = prop_prefixe(commune, prefixe)
        ) %>%
        arrange(desc(prop))
}
```

Exercice 1.4

Le vecteur suivant donne, pour chacun des neuf principaux films de la saga *Star Wars*, la date à laquelle ils se déroulent dans l'univers de la saga.

```
c(
    "I"    = -32,
    "II"   = -22,
```

```
    "III"  = -19,
    "IV"   =   0,
    "V"    =   3,
    "VI"   =   4,
    "VII"  =  34,
    "VIII" =  34,
    "IX"   =  35
)
```

Dans le jeu de données `starwars`, la variable `birth_year` indique l'année de naissance du personnage en "années avant l'an zéro" (une valeur de 19 signifie donc une année de naissance de -19).

Créer une fonction `age_film` qui prend en entrée un vecteur d'années de naissance au même format que `birth_year` ainsi que l'identifiant d'un film, et calcule les âges à la date du film.

Vérifier avec :

```
age_film(starwars$birth_year, "IV")
```

```
age_film <- function(annees, film) {
    annees_films <- c(
        "I"    = -32,
        "II"   = -22,
        "III"  = -19,
        "IV"   =   0,
        "V"    =   3,
        "VI"   =   4,
        "VII"  =  34,
        "VIII" =  34,
        "IX"   =  35
    )
    annees_naissance <- -annees
    annee_ref <- annees_films[film]
    annee_ref - annees_naissance
}
```

Utiliser la fonction pour ajouter deux nouvelles variables au tableau `starwars` : `age_iv` qui correspond à l'âge (potentiel) au moment du film IV, et `age_ix` qui correspond à l'âge au moment du film IX.

```
starwars %>%
    mutate(
        age_iv = age_film(birth_year, "IV"),
        age_ix = age_film(birth_year, "IX"),
    )
```

3.2 across()

Exercice 2.1

Reprendre la fonction `etendue` de l'exercice 1.2 :

```
etendue <- function(v) {
    max(v, na.rm = TRUE) - min(v, na.rm = TRUE)
}
```

Dans le jeu de données `starwars`, calculer l'étendue des variables `height` et `mass` pour chaque valeur de `sex` à l'aide de `group_by()`, `summarise()` et `across()`.

```
starwars %>%
    group_by(sex) %>%
    summarise(
        across(
            c(height, mass),
            etendue
        )
    )
```

Toujours à l'aide d'`across()`, appliquer `etendue` à toutes les variables numériques, toujours pour chaque valeur de `sex`.

```
starwars %>%
    group_by(sex) %>%
    summarise(
        across(
            where(is.numeric),
            etendue
        )
    )
```

En utilisant `&` et `!`, appliquer `etendue` à toutes les variables numériques sauf à celles qui finissent par “year”.

```
starwars %>%
    group_by(sex) %>%
    summarise(
        across(
            where(is.numeric) & !ends_with("year"),
            etendue
        )
    )
```

Exercice 2.2

Dans le jeu de données `starwars`, appliquer en un seul `summarise` les fonctions `min` et `max` aux variables `height` et `mass`.

```
starwars %>%
    summarise(
        across(
            c(height, mass),
            list(min = min, max = max)
        )
    )
```

Si vous ne l’avez pas déjà fait à la question précédente, modifier le code pour que le calcul des valeurs minimales et maximales ne prennent pas en compte les valeurs manquantes.

```
funs <- list(
    min = function(v) { min(v, na.rm = TRUE) },
    max = function(v) { max(v, na.rm = TRUE) }
)
starwars %>%
    summarise(
        across(
            c(height, mass),
            funs
        )
    )
# Autre possibilité : les arguments supplémentaires passés à across() sont
# transmis aux fonctions appliquées
starwars %>%
```

```
    summarise(
        across(
            c(height, mass),
            list(min = min, max = max),
            na.rm = TRUE
        )
    )
```

Exercice 2.3

Dans le jeu de données `hdv2003`, utiliser `across()` pour transformer les modalités "Oui" et "Non" en `TRUE` et `FALSE` pour toutes les variables de `hard.rock` à `sport`.

```
detecte_oui <- function(v) {
    v == "Oui"
}
hdv2003 %>%
    mutate(
        across(
            hard.rock:sport,
            detecte_oui
        )
    )
```

Ajouter un argument `.names` à `across()` pour que les variables recodées soient stockées dans de nouvelles colonnes nommées avec le suffixe `"_true"`.

```
detecte_oui <- function(v) {
    v == "Oui"
}
hdv2003 %>%
    mutate(
        across(
            hard.rock:sport,
            detecte_oui,
            .names = "{.col}_true"
        )
    )
```

3.3 Fonctions anonymes et notations abrégées

Exercice 3.1

Dans un exercice précédent, on a vu que le code ci-dessous permet de calculer l'étendue des variables `height` et `mass` du jeu de données `starwars`.

```
etendue <- function(v) {
    max(v, na.rm = TRUE) - min(v, na.rm = TRUE)
}

starwars %>%
    group_by(sex) %>%
    summarise(
        across(
            c(height, mass),
            etendue
        )
    )
```

Modifier ce code en supprimant la définition de `etendue` et en utilisant à la place une fonction anonyme directement dans le `across()`.

```
starwars %>%
    group_by(sex) %>%
    summarise(
        across(
            c(height, mass),
            function(v) {
               max(v, na.rm = TRUE) - min(v, na.rm = TRUE)
            }
        )
    )
```

Modifier à nouveau ce code pour utiliser la syntaxe abrégée de type "formule" du *tidyverse*.

```
starwars %>%
    group_by(sex) %>%
    summarise(
        across(
            c(height, mass),
            ~ max(.x, na.rm = TRUE) - min(.x, na.rm = TRUE)
        )
    )
```

Exercice 3.2

Soit le code suivant, qui renomme les colonnes du tableau `starwars` de type liste en leur ajoutant le préfixe "liste_".

```
ajoute_prefixe_liste <- function(nom) {
    paste0("liste_", nom)
}

starwars %>%
    rename_with(ajoute_prefixe_liste, .cols = where(is.list))
```

Réécrire ce code avec une fonction anonyme en utilisant les trois notations :

- classique (avec `function()`)
- formule (du *tidyverse*)
- compacte (à partir de R 4.1)

```
# Classique
starwars %>%
    rename_with(
        function(nom) { paste0("liste_", nom) },
        .cols = where(is.list)
    )
# Formule
starwars %>%
    rename_with(
        ~ paste0("liste_", .x),
        .cols = where(is.list)
    )
# Compacte
starwars %>%
    rename_with(
        \(nom) paste0("liste_", nom),
        .cols = where(is.list)
    )
```

Exercice 3.3

Le code suivant indique, pour chaque région du jeu de données `rp2018`, le nom de la commune ayant la valeur maximale pour les variables `dipl_aucun` et `dipl_sup`.

```
nom_commune_max <- function(valeurs, communes) {
    communes[valeurs == max(valeurs)]
}
```

```r
rp2018 %>%
    group_by(region) %>%
    summarise(
        across(
            c(dipl_aucun, dipl_sup),
            nom_commune_max,
            commune
        )
    )
```

Réécrire ce code en utilisant une fonction anonyme, avec la syntaxe de votre choix (classique, formule ou compacte).

```r
# Classique
rp2018 %>%
    group_by(region) %>%
    summarise(
        across(
            c(dipl_aucun, dipl_sup),
            function(valeurs, communes) { communes[valeurs == max(valeurs)] },
            commune
        )
    )
# Formule
rp2018 %>%
    group_by(region) %>%
    summarise(
        across(
            c(dipl_aucun, dipl_sup),
            ~ .y[.x == max(.x)],
            commune
        )
    )
# Compacte
rp2018 %>%
    group_by(region) %>%
    summarise(
        across(
            c(dipl_aucun, dipl_sup),
            \(valeurs, communes) communes[valeurs == max(valeurs)],
            commune
```

```
        )
    )
```

À l'aide d'une fonction anonyme supplémentaire, modifier le code pour qu'il retourne également, pour les mêmes variables, le nom des communes avec les valeurs minimales.

```
# Formule
rp2018 %>%
    group_by(region) %>%
    summarise(
        across(
            c(dipl_aucun, dipl_sup),
            list(
                max = ~ .y[.x == max(.x)],
                min = ~ .y[.x == min(.x)]
            ),
            commune
        )
    )
# Compacte
rp2018 %>%
    group_by(region) %>%
    summarise(
        across(
            c(dipl_aucun, dipl_sup),
            list(
                max = \(valeurs, communes) communes[valeurs == max(valeurs)],
                min = \(valeurs, communes) communes[valeurs == min(valeurs)]
            ),
            commune
        )
    ) %>% View()
```

3.4 `rowwise()` et `c_across()`

Exercice 4.1

On repart du code final de l'exercice 2.3, qui recodait une série de variables de `hdv2003` en valeurs `TRUE/FALSE` dans de nouvelles variables avec le suffixe `"_true"`.

```r
detecte_oui <- function(v) {
    v == "Oui"
}
hdv2003 <- hdv2003 %>%
    mutate(
        across(
            hard.rock:sport,
            detecte_oui,
            .names = "{.col}_true"
        )
    )
```

Calculer le plus simplement possible une nouvelle variable `total` qui contient, pour chaque ligne, le nombre de valeurs `TRUE` des deux variables `cinema_true` et `sport_true` (si une ligne contient `TRUE` pour ces deux variables, `total` doit valoir 2, etc.)

```r
hdv2003 %>%
    mutate(total = cuisine_true + sport_true)
```

Recalculer la variable `total` pour qu'elle contienne le nombre de `TRUE` par ligne pour les variables `bricol_true`, `cinema_true` et `sport_true`.

```r
hdv2003 %>%
    rowwise() %>%
    mutate(total = sum(cuisine_true, sport_true, bricol_true))
```

Recalculer la variable `total` pour qu'elle contienne le nombre de `TRUE` par ligne pour toutes les variables se terminant par `"_true"`.

```r
hdv2003 %>%
    rowwise() %>%
    mutate(total = sum(c_across(ends_with("_true"))))
```

Reprendre le code précédent pour qu'il puisse s'appliquer directement sur les variables `hard.rock…sport`, sans passer par le recodage en `TRUE`/`FALSE`.

```r
count_oui <- function(v) {
    sum(v == "Oui")
}
```

```
hdv2003 %>%
    rowwise() %>%
    mutate(
        total = count_oui(c_across(hard.rock:sport))
    )
```

Exercice 4.2

Dans le jeu de données `starwars`, la colonne `films` contient la liste des films dans lesquels apparaissent les différents personnages. Cette colonne a une forme un peu particulière puisqu'il s'agit d'une "colonne-liste" : les éléments de cette colonne sont eux-mêmes des listes.

```
head(starwars$films, 3)
```

On essaye de calculer le nombre de films pour chaque personnage avec le code suivant. Est-ce que ça fonctionne ? Pourquoi ?

```
starwars %>%
    mutate(n_films = length(films))
```

Trouver une manière d'obtenir le résultat attendu.

```
starwars %>%
    rowwise() %>%
    mutate(n_films = length(films))
```

Troisième partie

SIG

Introduction

Le cours en résumé :

- 24 h
- objectifs : appréhender les données géographiques et maîtriser les principales opérations sur ce type de données (intersection, distance etc)
- Langage : R & QGIS (introduction)
- modalités d'examens :
  - Note de participation
  - Un projet commun SIG/économétrie avec choix du sujet imposé
  - Présentation en classe d'une carte réalisée avec QGIS : 3mn de présentation

Le plan :

1. Introduction à QGIS
2. Traitement sur les données vectorielles
3. Les SIG avec R

Toutes les données des TP sont disponibles sur ce drive.

TP1 : Introduction à QGIS

1. Prise en main

1.1 Affichage / désaffichage des panneaux

— Fermez les panneaux **Couches** et **Identifier les résultats**.

— Affichez-les de nouveau avec le menu **Vue** > **Panneaux**.

1.2 Utilisation de données vecteur

Toutes les données des TP sont disponibles sur ce drive.

1. Examinez la liste des fichiers du répertoire `Data/ADMIN EXPRESS`.

2. Affichez :
   — `ARRONDISSEMENT.shp`

   — `EPCI.shp`

   — `communes_ara.gpkg`

3. Ouvrez la table attributaire de la couche **ARRONDISSEMENT** :
   — Trier selon la colonne `INSEE_DEP`.

   — Quels sont les arrondissements du département de la Loire (42) ?

4. Supprimez la couche **ARRONDISSEMENT**.

5. Ouvrez les propriétés de la couche **COMMUNE** :
   — Notez le système de coordonnées, la géométrie, la liste des attributs.

6. Ouvrez la table attributaire de la couche **COMMUNE** :
   — Quel est le nombre de communes ?

7. Identifiez la commune de **Saint-Maurice-en-Gourgois** :
   — Dans quel département est-elle ?

   — Chargez la couche `DEPARTEMENT.shp` et identifiez le département 43.

   — Quelle est sa population (champ `POPULATION`) ?

1.3 Création d'un projet

— Enregistrez le projet sous le nom `TP1-1.qgs`.

1.4 Utilisation de l'outil Identifier les entités

— Utilisez l'outil sur Saint-Maurice-en-Gourgois : quelle est la longueur du périmètre ?

— Zoomez sur la dalle.

1.5 Jointure 1

— Joindre la couche **EPCI** à **COMMUNE**.

— Quel est le nom de l'EPCI de Saint-Maurice-en-Gourgois ?

1.6 Utilisation d'OpenStreetMap

1. Créez une connexion XYZ :

— Nom : `OpenStreetMap`
— URL : `https://tile.openstreetmap.org/{z}/{x}/{y}.png`

2. Installez l'extension **QuickMapServices**.

3. Affichez le fond de carte OSM Standard.

1.7 Ordre des couches et opacité

— Classez les couches dans l'ordre : Communes → EPCI → Arrondissement.

1.8 Groupe de couches

— Groupez ARRONDISSEMENT et COMMUNE en **ADMIN**.

1.9 Outil de mesure

— Mesurez la distance maximale de Saint-Maurice-en-Gourgois.

1.10 Sélection et export

— Sélectionnez les communes d'Auvergne-Rhône-Alpes.

— Exportez dans `CommunesEPCI_ARA.gpkg`.

— Créez une couche des seuls EPCI d'Auvergne-Rhône-Alpes.

1.11 Sélection et conditions multiples

— Communes ARA > 1000 habitants → combien ?

— Communes Haute-Loire > 1000 habitants → combien ?

— Exportez les deux sélections.

2. Symbologie 1 : Pays du monde

1. Téléchargez les données physiques et culturelles de **NaturalEarth** à 50m sur ce lien
2. Ouvrez :

— `ne_50m_admin_0_countries`

— `ne_50m_populated_places_simple`

— `ne_50m_geographic_lines`

3. Avec la première couche, créez une symbologie par PIB du pays (`GDP_MD`) en prenant une graduation Jenks.

4. Utilisez palette **Viridis**, projection **World Robinson (EPSG:54030)**.

5. Renommez la couche : *Countries by GDP*.

6. Enregistrez le projet : `TP1_2.qgz`.

2.1 Mise en page

— Créez une mise en page `gpd` en A4 paysage.

— Ajoutez carte, légende, fond gris clair, export en PNG 300 dpi.

2.2 Rendu gradué : Villes du monde

1. Utilisez `pop_max`.

2. Créez 6 classes (Jenks ou seuils manuels).

3. Symbole ponctuel rose, transparence 60 %, taille 0.5–4 mm.

4. Exportez en PNG 300 dpi.

2.2 Symboles proportionnels

— Symbole unique, rose, transparence 60 %.

— Taille proportionnelle `pop_max`.

— Enregistrez le projet.

3. Symbologie 2 : France

Ouvrir :

— `liste_cheflieu.geojson`,
— `occitanie_communes.gpkg`,
— `occitanue_limites.gpkg`,
— `liste-des-gares.geojson`,
— `RéseauFerré.gpkg`

Dans cette parties, il faut faire une carte de l'occitanie en utilisant les 5 jeux de données

3.1 Styles

— Régions en gris clair + bordures blanches.

— Chefs-lieux : symboles catégorisés (`Préfecture région` et `Préfecture`).

— Réseau ferré : catégorisé sur `type_voie`.

— Communes Occitanie : densité de population (Jenks, OrRd, bornes manuelles).

3.2 Étiquettes

— Chefs-lieux avec règles selon statut administratif.

3.3 Mise en page finale

— A4 portrait, titre *Réseau ferré en Occitanie*, carte 1/2 000 000, légende, échelle, nord, sources.

— Ajoutez une carte miniature France + Occitanie.

— Export PNG et PDF.

TP2 : Traitement sur les données vectorielles

1. Données de départ

Dans un nouveau projet, ouvrez les couches :
- `departement_occitanie.gpkg`
- `CLC12_RLRMP_RGF.shp`

Pour visualiser les types d'occupation du sol avec les couleurs standard **Corine Land Cover**, ouvrez les propriétés de la couche et chargez le fichier de style `CLC12.sld` (dans le dossier *FichiersLegende*).

Ajoutez aussi :
- `trace-du-reseau-autoroutier-doccitanie.geojson`
- `dreal-occitanie-mats-eoliens.geojson`

Enregistrez le projet sous le nom `TP2.qgz`.

2. Création d'un GeoPackage

Toutes les couches produites seront enregistrées dans une base unique **GeoPackage**.

1. Créez `TP2_couches.gpkg` (répertoire *data*) en exportant la couche des mâts éoliens.

2. Options :
   — format = *GeoPackage*
   — nom fichier = `TP2_couches.gpkg`
   — nom couche = `Mâts éoliens`
   — SCR = `EPSG:2154`

3. Zone tampon sur l'autoroute A61

3.1 Vérification du SCR

Dans les propriétés de `trace-du-reseau-autoroutier-doccitanie`, identifiez le système de coordonnées.

3.2 Conversion en Lambert 93

Exporte la couche vers `TP2_couches.gpkg` avec :
- nom couche = `Réseau autoroutier`
- SCR = `EPSG:2154`

3.3 Sélection et tampon

1. Ouvrir la table attributaire → champ `num_route`.

2. Sélectionner les tronçons correspondant à **A61**.

3. Créer un tampon de **5000 m** avec options :
   — entités sélectionnées uniquement = Oui

   — Nb segments = 10

   — extrémités = Rond

   — regrouper = Oui

   — sortie = `TP2_couches.gpkg` → `Tampon A61 5000m`

3.4 Analyse

Avec **Compter les points dans les polygones**, combien d'éoliennes se trouvent dans cette zone tampon ?

4. Matrice de distance

Pour chaque mât éolien, calculer la distance avec les **2 plus proches voisins**.
- entrée = `Mâts éoliens`
- identifiant = `id_mat`
- type = *Matrice de distance linéaire (N*k+3)*
- k = 2
- sortie = `TP2_couches.gpkg` → `Calcul Eoliennes 2 voisins`

Inspectez le résultat avec l'outil Identifier : remarquez-vous le type de géométrie ?

5. Grille hexagonale

5.1 Création

1. Dans `departement_occitanie`, sélectionnez l'Aude et zoomez.

2. Avec **Créer une grille** :
   — type = hexagonale

   — étendue = canevas

   — espacement = 5000

   — SCR = EPSG:2154

   — sortie = `TP2_couches.gpkg` → `Grille Aude 5km`

5.2 Nettoyage

Supprimez les hexagones hors de l'Aude :
- sélection par localisation → inverser → supprimer en mode édition.

6. Comptages dans la grille

— **Compter points/polygones** → nb d'éoliennes par maille hexagonale.
  — sortie = `TP2_couches.gpkg` → `Calcul nb éoliennes`

— **Somme longueurs lignes** → total autoroutes par maille.

— sortie = `TP2_couches.gpkg` → `Calcul long autoroutes`

7. Analyse de superposition (Corine Land Cover)

7.1 Sélection des forêts

Sélectionner dans `CLC12_RLRMP_RGF` les polygones dont `CODE_12` commence par “3”. Exporter vers `TP2_couches.gpkg` → `CLC12 Forets Milieux SemiNat`.

7.2 Superposition

Avec **Analyse de superposition** :
- source = `Grille Aude 5km`
- superposition = `CLC12 Forets Milieux SemiNat`
- sortie = `TP2_couches.gpkg` → `Calcul P ForetsSemiNat`

8. Intersection et Group Stats

8.1 Zone tampon des parcs

Créer tampon **2000 m** autour de chaque mât → `Tampon Eol 2000m`. Puis regrouper par `id_parc, n_parc` → `Calcul ParcEol 2000m`.

8.2 Intersection avec CLC

Avec **Intersection** :
- source = `CLC12_RLRMP_RGF`
- superposition = `Tampon Eol 2000m`
- champs conservés : `ID, CODE_12, id_parc, n_parc`
- sortie = `TP2_couches.gpkg` → `Calcul Inter ParcEol CLC12`

8.3 Tableau croisé dynamique

Dans l'extension **Group Stats** :
- Couches = `Calcul Inter ParcEol CLC12`
- Colonnes = `CODE_12`
- Lignes = `id_parc, n_parc`
- Valeurs = `Surface (somme)`

Exporter le tableau et coller dans Excel/Calc.

TP3 : SIG avec R

Chargez les libraries suivantes :

```
#install.packages("spDataLarge", repos = "https://geocompr.r-universe.dev")
#install.packages("remotes")
#remotes::install_github("r-tmap/tmap")

library(tidyverse)
library(sf)
library(stars)
library(terra)
library(spData)
library(spDataLarge)
library(tmap)
library(leaflet)


#remotes::install_github("r-tmap/tmap")
```

1. Premières cartes

1. Décrire l'objet Utilisez `world`. Utiliser `summary()` sur la colonne de géométrie de l'objet `world` inclus dans le package `spData.` Utilisez `ggplot2`. Tracez la carte des continents. Utiliser le `theme_void()`. Tracez le continent asiatique, en filtrant puis appliquant la fonction d'union de formes géométrique `st_union`.

2. Rajouter à la carte des continents des ronds pour chaque pays représentant la racine carré de leur population divisé par 10000. Il faudra pour ça calculer les centroides de chaque pays avec la commande `st_centroid` du package `sf`

3. Tracez la carte de l'Inde dans le continent asiatique. Il faut :

— filtrer et tracer le continent asiatique dans `world` .
— créer un object `india` qui filtre l'Inde dans `world` .
— rajouter la carte le l'Inde en grisant son contour

— créer le centroide de l'inde et rajouter sur la carte une étiquette "Inde" à la coordonnée du centroide du pays

4. Créer un raster de 10x10 pixels avec la commande `rast`, dont les niveaux avec des valeurs aléatoires allant de 0 à 10 (avec la commande `runif`). Tracez ce raster avec `geom_raster`.

5. Chargez le fichier `raster/nlcd.tif` du package `spDataLarge` à l'aide de la commande. Décrire cet objet. Utiliser la fonction `plot`. Enfin, convertir le raster en objet du package `stars` et décrire le résultat.

2. Opérations sur les attributs

Pour ces exercices, nous utiliserons les ensembles de données `us_states` et `us_states_df` du package **spData**.

1. Créez un nouvel objet appelé `us_states_name` qui contient uniquement la colonne `NAME` de l'objet `us_states` en utilisant la syntaxe de base R (`[`) ou tidyverse (`select()`). Quelle est la classe du nouvel objet et qu'est-ce qui le rend géographique ?

2. Sélectionnez les colonnes de l'objet `us_states` contenant les données de population. Obtenez le même résultat en utilisant une autre commande (bonus : essayez de trouver trois façons d'obtenir le même résultat). Indice : essayez d'utiliser des fonctions d'aide, telles que `contains` ou `matches` de **dplyr** (voir `?contains`).

3. Trouvez tous les États ayant les caractéristiques suivantes (bonus : trouvez-les *et* affichez-les) :

— Appartiennent à la région Midwest.
— Appartiennent à la région Ouest, ont une superficie inférieure à 250 000 km^2 *et* en 2015, une population supérieure à 5 000 000 d'habitants (astuce : vous devrez peut-être utiliser la fonction `units::set_units()` ou `as.numeric()`).
— Appartiennent à la région Sud, avaient une superficie supérieure à 150 000 km^2 ou une population totale en 2015 supérieure à 7 000 000 d'habitants.

4. Quelle était la population totale en 2015 dans l'ensemble de données `us_states` ? Quelle était la population minimale et maximale en 2015 ?

5. Combien d'États y a-t-il dans chaque région ?
6. Quelle était la population minimale et maximale en 2015 dans chaque région ? Quelle était la population totale en 2015 dans chaque région ?

7. Ajoutez des variables de `us_states_df` à `us_states` et créez un nouvel objet appelé `us_states_stats`. Quelle fonction avez-vous utilisée et pourquoi ? Quelle variable sert de clé dans les deux ensembles de données ? Quelle est la classe du nouvel objet ?
8. `us_states_df` a deux lignes de plus que `us_states`. Comment pouvez-vous les trouver ? (astuce : essayez d'utiliser la fonction `dplyr::anti_join()`)
9. Quelle était la densité de population en 2015 dans chaque État ? Quelle était la densité de population en 2010 dans chaque État ?
10. Combien la densité de population a-t-elle changé entre 2010 et 2015 dans chaque État ? Calculez le changement en pourcentage et cartographiez-le.
11. Changez les noms des colonnes dans `us_states` en minuscules. (Astuce : les fonctions d'aide - `tolower()` et `colnames()` peuvent aider.)
12. Utilisez `us_states` et `us_states_df` pour créer un nouvel objet appelé `us_states_sel`. Le nouvel objet ne doit contenir que deux variables - `median_income_15` et `geometry`. Changez le nom de la colonne `median_income_15` en `Income`.
13. Calculez le changement du nombre de résidents vivant en dessous du seuil de pauvreté entre 2010 et 2015 pour chaque État. (Astuce : voir `?us_states_df` pour la documentation sur les colonnes du seuil de pauvreté.) Bonus : Calculez le changement en pourcentage des résidents vivant en dessous du seuil de pauvreté dans chaque État.
14. Quelle était la population minimale, moyenne et maximale des personnes vivant en dessous du seuil de pauvreté en 2015 pour chaque région ? Bonus : Quelle est la région où l'augmentation du nombre de personnes vivant en dessous du seuil de pauvreté est la plus importante ?

15. Créez un raster `grain` vide avec neuf lignes et colonnes et une résolution de 0,5 degré décimal (WGS84). Remplissez-le avec des nombres aléatoires. Extraire les valeurs des quatre cellules de coin.
16. Quelle est la classe la plus courante de notre exemple de raster `grain` ?
17. Tracez l'histogramme et la boîte à moustaches du fichier `dem.tif` du package **spDataLarge** (`system.file("raster/dem.tif", package = "spDataLarge")`).

3. Opération sur les données spatiales

3.1 Opérations sur les vecteurs

1. Utiliser les jeux de données `nz` et `nz_height` du package `spData`. Combien de ces points élevés la région de Canterbury contient-elle ?

Bonus : tracez le résultat en utilisant la fonction `plot()` pour montrer toute la Nouvelle-Zélande, la région de `Canterbury` en jaune, les points élevés à Canterbury représentés par des croix rouges (astuce : `pch = 7`) et les points élevés dans d'autres parties de la Nouvelle-Zélande représentés par des cercles bleus. Consultez la page d'aide `?points` pour plus de détails avec une illustration des différentes valeurs `pch`.

2. Quelle région a le deuxième plus grand nombre de points `nz_height`, et combien en a-t-elle ?
3. En généralisant la question à toutes les régions : combien des 16 régions de Nouvelle-Zélande contiennent des points qui appartiennent aux 100 points les plus élevés du pays ? Quelles sont ces régions ?

Bonus : créez un tableau listant ces régions par ordre du nombre de points et leur nom.

4. Le point de départ de cet exercice est de créer un objet représentant l'État du Colorado aux États-Unis. Faites ceci avec la fonction `filter()` (`tidyverse`) et tracez l'objet résultant dans le contexte des États-Unis.

— Créez un nouvel objet représentant tous les États qui se chevauchent géographiquement avec le Colorado et tracez le résultat (astuce : la manière la plus concise de le faire est avec la méthode de sous-ensemble `[`).
— Créez un autre objet représentant tous les objets qui touchent (ont une frontière commune avec) le Colorado et tracez le résultat (astuce : souvenez-vous que vous pouvez utiliser l'argument `op = st_intersects` et d'autres relations spatiales lors des opérations de sous-ensemble spatial en R de base).

Bonus : créez une ligne droite allant du centroïde du district de Columbia près de la côte Est au centroïde de la Californie près de la côte Ouest des États-Unis (astuce : les fonctions `st_centroid()`, `st_union()` et `st_cast()` peuvent aider) et identifiez quels États cette longue ligne est.

3.2 Opérations sur les rasters

5. Utilisez `dem = rast(system.file("raster/dem.tif", package = "spDataLarge"))`, et reclassifiez l'élévation en trois classes : basse (<300), moyenne et haute (>500). Ensuite, lisez le raster NDVI (`ndvi = rast(system.file("raster/ndvi.tif", package = "spDataLarge"))`) et calculez la moyenne du NDVI et de l'élévation pour chaque classe d'altitude.
6. Appliquez un filtre de détection de lignes à `rast(system.file("ex/logo.tif", package = "terra"))`. Tracez le résultat. Astuce : Lisez `?terra::focal()`.

7 Calculez l'indice d'eau normalisé (NDWI ; `(green - nir)/(green + nir)`) d'une image Landsat. Utilisez l'image Landsat fournie par le package **spDataLarge** (`system.file("raster/landsat.tif", package = "spDataLarge")`). Calculez également une corrélation entre le NDVI et le NDWI pour cette région (astuce : vous pouvez utiliser la fonction `layerCor()`).

8. Un message sur StackOverflow montre comment calculer les distances jusqu'à la côte la plus proche en utilisant `raster::distance()`. Essayez de faire quelque chose de similaire mais avec `terra::distance()` : récupérez un modèle numérique d'élévation de l'Espagne et calculez un raster qui représente les distances jusqu'à la côte à travers le pays (astuce : utilisez `geodata::elevation_30s()`). Convertissez les distances résultantes de mètres en kilomètres. Note : il peut être judicieux d'augmenter la taille de cellule du raster d'entrée pour réduire le temps de calcul lors de cette opération (`aggregate()`).
9. Essayez de modifier l'approche utilisée dans l'exercice ci-dessus en pondérant le raster de distance avec le raster d'élévation ; chaque tranche de 100 mètres d'altitude devrait augmenter la distance jusqu'à la côte de 10 km. Ensuite, calculez et visualisez la différence entre le raster créé en utilisant la distance euclidienne (E7) et le raster pondéré par l'élévation.

4. Opérations sur les géométries

4.1 Opérations sur les vecteurs

1. Générez et tracez des versions simplifiées de l'ensemble de données `nz`. Expérimentez avec différentes valeurs de `keep` (allant de 0,5 à 0,00005) pour `ms_simplify()` et `dTolerance` (de 100 à 100 000) pour `st_simplify()`.

— À partir de quelle valeur le résultat commence-t-il à se détériorer pour chaque méthode, rendant la Nouvelle-Zélande méconnaissable ?
— Avancé : Quelle est la différence entre le type de géométrie des résultats de `st_simplify()` par rapport au type de géométrie de `ms_simplify()` ? Quels problèmes cela crée-t-il et comment cela peut-il être résolu ?

2. Dans le premier exercice du chapitre sur les opérations de données spatiales, il a été établi que la région de Canterbury avait 70 des 101 points les plus élevés de Nouvelle-Zélande. En utilisant `st_buffer()`, combien de points dans `nz_height` se trouvent à moins de 100 km de Canterbury ?
3. Trouvez le centroïde géographique de la Nouvelle-Zélande. À quelle distance se trouve-t-il du centroïde géographique de Canterbury ?
4. La plupart des cartes du monde ont une orientation nord en haut. Une carte du monde avec une orientation sud en haut pourrait être créée par une réflexion (l'une des transformations affines non mentionnées dans ce chapitre) de la géométrie de l'objet `world`. Écrivez le code pour le faire. Astuce : vous devez utiliser un vecteur à deux éléments pour cette transformation. Bonus : créez une carte à l'envers de votre pays.
5. Calculez la longueur des lignes de frontières des États-Unis en mètres. Quel État a la frontière la plus longue et lequel a la frontière la plus courte ? Astuce : La fonction `st_length` calcule la longueur d'une géométrie de type `LINESTRING` ou `MULTILINESTRING`. Il

faut aussi transformer la géométrieavec un CRS qui puisse calculer des ditances : ici ESPG=2163.

6. Exécutez le code de la section 5.2.6. En référence aux objets créés dans cette section, sélectionnez le point dans `p` qui est contenu à la fois dans `x` et `y`.

— Utilisez les opérateurs de sous-ensemble de base.
— Utilisez un objet intermédiaire créé avec `st_intersection()`.

6. Exécutez le code suivant :

```
bb = st_bbox(st_union(x, y))
box = st_as_sfc(bb)
set.seed(2024)
p = st_sample(x = box, size = 10)
p_xy1 = p[x_and_y]
plot(box, border = "gray", lty = 2)
plot(x, add = TRUE, border = "gray")
plot(y, add = TRUE, border = "gray")
plot(p, add = TRUE, cex = 3.5)
plot(p_xy1, cex = 5, col = "red", add = TRUE)
text(x = c(-0.5, 1.5), y = 1, labels = c("x", "y"), cex = 3)
```

En référence aux objets créés dans cette section, sélectionnez le point dans `p` qui est contenu à la fois dans `x` et `y`.

— Utilisez les opérateurs de sous-ensemble de base.
— Utilisez un objet intermédiaire créé avec `st_intersection()`.

4.2 Opérations sur les rasters

7. Lisez le fichier srtm.tif dans R (`srtm = rast(system.file("raster/srtm.tif", package = "spDataLarge"))`). Ce raster a une résolution de 0,00083 par 0,00083 degrés. Modifiez sa résolution en 0,01 par 0,01 degrés en utilisant toutes les méthodes disponibles dans le package **terra**. Visualisez les résultats. Pouvez-vous remarquer des différences entre les résultats de ces méthodes de rééchantillonnage ?

5. Application : rapprochement de base par distances

Le but de cet exercice est d'identifier la nature de stations de services. Un jeu de données issu de *https://www.data.gouv.fr/fr/datasets/prix-des-carburants-en-france-flux-instantane-v2-amelioree/* donne les prix des carburants mais on n'a pas d'information sur le type de station

(station d'autoroute, de supermarché…). Le jeu de données magasins d'openstreetmap pourrait permettre d'apporter des informations.

1. Charger les données de `TP3.zip`. Les gpkg s'ouvrent la commande `st_read` du package `sf`.
2. Restreindre la base `magasins` aux types de magasins (`shop`) suivants : "gas","supermarket","convenience", "car_repair","car","mall","convenience;gas"
3. Transformer les deux jeux en sf dataframe en système de coordonnées EPSG 2154. Attention, pour la base `station`, il faut diviser longitude et latitude par 100000.
4. Pour chaque station station, le magasins le plus proche et calculer la distance correspondate.
5. Quelle est la part des magasins à moins de 100 mètres d'une station.
6. Ajouter les attributs `shop` et `operator` pour chaque magasins les plus proche à la base stations.

Quatrième partie

Econométrie 1

Introduction

Le cours en résumé :

— 30 h
— objectifs : développer, interpréter et critiquer des modèles économétriques
— modalités d'examens :
— Note de participation
— Un projet commun calcul numérique/économétrie avec choix du sujet libre
— Un projet commun SIG/économétrie avec choix du sujet imposé

Le plan :

1. Statistiques et probabilités en R
2. Régression linéaire multiple
3. Projet 1
4. Projet 2

TP1 : Probabilités et Statistiques avec R

1. Probabilités avec R

1.1 - Échantillonnage

Vous êtes la fée des loteries dans une loterie hebdomadaire, où 6 numéros uniques sur 49 sont tirés.

1. Tirez aléatoirement les numéros gagnants de cette semaine (fixez la graine à 123) en utilisant la fonction `sample`.

1.2 - Fonction de densité de probabilité

Considérez une variable aléatoire X avec une fonction de densité de probabilité (PDF)

$$f_X(x) = \frac{x}{4}e^{-x^2/8}, \quad x \geq 0.$$

1. Définissez la PDF ci-dessus comme une fonction $f()$.
2. Vérifiez si la fonction que vous avez définie est effectivement une PDF (indice : utilisez la fonction `integrate`).

1.3 - Espérance et Variance

Dans cet exercice, vous devez calculer l'espérance et la variance de la variable aléatoire X considérée dans l'exercice précédent.

La PDF $f()$ de l'exercice précédent est disponible dans votre environnement de travail.

1. Définissez une fonction appropriée `ex()` qui s'intègre à l'espérance de X.
2. Calculez l'espérance de X. Stockez le résultat dans `expected_value`.
3. Définissez une fonction appropriée `ex2()` qui s'intègre à l'espérance de X^2.
4. Calculez la variance de X. Stockez le résultat dans `variance`.

1.4 - Distribution Normale Standard

Soit $Z \sim \mathcal{N}(0,1)$.

1. Calculez $\phi(3)$, c'est-à-dire la valeur de la densité de probabilité standard normale en 3.
2. Calculez $ P(|Z| 1.64) $ en utilisant la fonction `pnorm()`.

Indications : en R contient des distributions de probabilités pré-enregistrées (`norm` pour la distribution normale, `chiqsq` pour chi2, `t` pour Student). La syntaxe est la suivante :

— d → densité (ex : `dnorm`)
— p → probabilité (ex : `pnorm`). Pour cette fonction, on utilise souvent l'option `lower.tail=F` pour calculer la probabilité complémentaire.
— q → quantile (ex : `qnorm`)
— r → tirage aléatoire (ex : `rnorm`)

1.5 - Distribution du Chi-carré

1. Soit $W \sim \chi^2_{1,0}$. Tracez la PDF correspondante à l'aide de `curve()`. Spécifiez la plage de valeurs x comme [0,25] via l'argument `xlim.`
2. Soient X_1 et X_2 deux variables aléatoires normalement distribuées indépendantes avec $\mu = 0$ et $\sigma^2 = 15$. Calculez $P(X_1^2 + X_2^2 > 10)$

1.6 - Distribution de Student

1. Soit $X \sim t_{10000}$ et $Z \sim N(0,1)$. Calculez le quantile à 95 % des deux distributions. Que remarquez-vous ?
2. Soit $X \sim t_1$. Générez 1000 nombres aléatoires à partir de cette distribution et attribuez-les à la variable `x`. Calculez la moyenne de l'échantillon de x. Pouvez-vous expliquer le résultat ?

1.7 - Distribution de Fisher

1. Soit $Y \sim F(10,4)$. Tracez la fonction quantile de la distribution donnée à l'aide de la fonction `curve()`.
2. Soit $Y \sim F(4,5)$. Calculez $P(1 < Y < 10)$ en intégrant la PDF avec la fonction `integrate`.

2. Statistiques avec R

2.1 - Biais

On considère l'estimateur alternatif suivant pour μ_Y, la moyenne de Y :

$$\tilde{Y} = \frac{1}{n-1} \sum_{i=1}^{n} Y_i$$

1. Définissez une fonction `Y_tilde()` qui implémente l'estimateur ci-dessus.
2. Tirez aléatoirement 5 observations au hasard à partir de la distribution $N(10, 25)$ et calculez une estimation en utilisant `Y_tilde()`. Répétez cette procédure 10000 fois et stockez les résultats dans `est_biased` en utilisant la fonction `replicate`.
3. Tracez un histogramme de `est_biased`. Ajoutez une ligne verticale rouge à $\mu = 10$ en utilisant la fonction `abline()`.
4. Tirez aléatoirement 1000 observations au hasard à partir de la distribution $N(10, 25)$ et calculez une estimation de la moyenne en utilisant`Y_tilde()`. Répétez cette procédure 10000 fois et stockez les résultats dans `est_consistent.`
5. Tracez un histogramme de `est_consistent`. Ajoutez une ligne verticale rouge à $\mu = 10$ en utilisant la fonction `abline()`.

2.2 - Efficience d'un estimateur

Dans cet exercice, nous souhaitons illustrer le résultat selon lequel la moyenne de l'échantillon :

$$\hat{\mu}_Y = \sum_{i=1}^{n} a_i Y_i$$

avec le schéma de pondération égale $a_i = \frac{1}{n}$ pour $i = 1, ..., n$ est l'estimateur linéaire non biaisé meilleur (BLUE) de μ_Y.

En tant qu'alternative, considérez l'estimateur :

$$\tilde{\mu}_Y = \sum_{i=1}^{n} b_i Y_i$$

où b_i donne aux premières $\frac{n}{2}$ observations un poids plus élevé de 3 que les deuxièmes $\frac{n}{2}$ observations (nous supposons que n est pair pour simplifier).

%Le vecteur de poids w a déjà été défini et est disponible dans votre environnement de travail.

1. Définissez un vecteur de pondération pour une taille d'échantillon `n=100`. Il doit être normalisé.

2. Vérifiez que $\tilde{\mu}_Y$ est un estimateur non biaisé de μ_Y, la moyenne de Y_i.
3. Implémentez l'estimateur alternatif de μ_Y en tant que fonction `mu_tilde()`.
4. Tirez au hasard 100 observations à partir de la distribution $\mathcal{N}(5, 10)$ et calculez les estimations avec les deux estimateurs. Répétez cette procédure 10000 fois et stockez les résultats dans `est_bar` et `est_tilde`. Utilisez la fonction `replicate`.
5. Calculez les variances de l'échantillon de `est_bar` et `est_tilde`. Que pouvez-vous dire sur les deux estimateurs ?

2.3 - Test d'hypothèse

Considérez l'ensemble de données `wage1` du package `wooldridge`. La variable `wage` donne les gains horaires moyens des individus. Nous supposons que les gains horaires moyens `wage` dépassent 10 dollars par heure et souhaitons tester cette hypothèse à un niveau de signification de $\alpha = 0,05$. Veuillez faire ce qui suit :

1. Calculez la statistique de test manuellement et attribuez-la à `tstat`.
2. Utilisez `tstat` pour accepter ou rejeter l'hypothèse nulle.
3. Refaites-le en utilisant l'approximation normale.
4. Calculez la valeur-p manuellement et attribuez-la à `pval` en utilisant l'approximation normale.
5. Utilisez `pval` pour accepter ou rejeter l'hypothèse nulle.
6. Effectuez le test d'hypothèse des questions précédentes en utilisant la fonction `t.test()`.
7. Extrayez la statistique t et la valeur-p de la liste créée par `t.test()`. Attribuez-les aux variables `tstat` et `pvalue`.
8. Vérifiez que l'utilisation de l'approximation normale ici est également valide en calculant la différence entre les deux valeurs-p.

2.4 - Test d'hypothèse : valeur-p

On considère les données CO2 (`data(CO2)`).

1. Tester s'il existe une différence significative dans l'absorption entre les plantes traitées et les plantes non traitées à un niveau de signification de α=0,05.
2. Obtenez l'intervalle de confiance.

2.5 - Corrélation

Charger la librairie `corrgram` et le jeu de données `auto`.

1. Calculez la corrélation simple (linéaire) entre le prix de la voiture (`Price`) et son économie de carburant `MPG` (mesurée en miles par gallon, ou mpg).
2. Utilisez la fonction `cor.test` pour vérifier si le coefficient obtenu est statistiquement significatif au niveau de 5 %.
3. La corrélation simple suppose une relation linéaire entre les variables, mais il peut être utile de relâcher cette hypothèse. Calculez le coefficient de corrélation de Spearman pour les mêmes variables et trouvez sa signification statistique.
4. En R, il est possible de calculer la corrélation pour toutes les paires de variables numériques dans un dataframe en une seule fois. Cependant, cela nécessite d'exclure d'abord les variables non numériques. Créez un nouveau dataframe, `auto_num`, qui ne contient que les colonnes avec des valeurs numériques du dataframe `auto`. Vous pouvez le faire en utilisant la fonction `filter`.
5. Utilisez la fonction `cor` pour créer une matrice de coefficients de corrélation pour les variables du dataframe `auto_num`.
6. La fonction standard `cor.test` ne fonctionne pas avec des dataframes. Cependant, la signification statistique des coefficients de corrélation pour un dataframe peut être vérifiée à l'aide de la fonction `rcorr` du package `Hmisc`. Transformez le dataframe `auto_num` en une matrice (`auto_mat`) et utilisez-le pour vérifier la signification des coefficients de corrélation avec la fonction `rcorr`.
7. Utilisez la fonction `corrgram` du package `corrgram` pour créer un correlogramme par défaut afin de visualiser les corrélations entre les variables du dataframe `auto`.
8. Créez un autre correlogramme qui (1) ne comprend que le panneau inférieur, (2) utilise des diagrammes en camembert pour représenter les coefficients de corrélation et (3) ordonne les variables selon l'ordre par défaut.
9. Créez un nouveau dataframe, `auto_subset`, en sous-échantillonnant le dataframe `auto` pour inclure uniquement les variables `Price`, `MPG`, `Hroom` et `Rseat`. Utilisez le nouveau dataframe pour créer un correlogramme qui (1) affiche les coefficients de corrélation dans le panneau inférieur et (2) montre des diagrammes de dispersion (points) dans le panneau supérieur.
10. Utilisez la fonction `correlations` du package `ggm` pour créer une matrice de corrélation avec à la fois des coefficients de corrélation complets et partiels pour le dataframe `auto_subset`. Trouvez la corrélation partielle entre le prix de la voiture et son économie de carburant.

TP 2 : Modèle de régression multiple

```r
library(tidyverse)
library(wooldridge)
library(AER)
library(stargazer)
library(fixest)
```

Introduction

Lancez **RStudio** et ouvrez un nouveau **R Markdown**. Ce sera votre document de travail durant tout le TP. L'intérêt des Markdown est de pouvoir lancer une multitude de petits scripts successivement. Typiquement, un *chunk* (petit script) par question. Lorsque c'est nécessaire, répondez aux questions entre les chunks.

Créez un fichier partagé "Econométrie-votre nom" sur votre drive fourni de l'université (OneDrive ou Google Drive). Envoyez-moi le lien de ce drive. Ce dossier partagé sera votre dossier de travail pendant toute la durée du cours. Créez un dossier "TP1" et enregistrez-y votre markdown sous le nom "TP1".

Rappel sur R

(i) Chargez le package **AER** (téléchargez-le si besoin). Ouvrez le fichier de données $CASchools$ avec la commande : data(CASchools).

(ii) Définissez les variables $STR = students/teachers$ et $score = (read + math)/2$.

(iii) Affichez les statistiques descriptives de ce jeu de données avec la commande **summary**.

(iv) Estimez le modèle suivant avec la commande **lm** et affichez les résultats de la régression :

$$score = \beta_0 + \beta_1 STR + \beta_2 english + u$$

(v) Reprérentez les résultats précédents à l'aide du **stargazer** et utilisez la commande `stargazer( data , type = 'text')` en remplaçant *data* par les jeux de données précédemment appelés.

(vi) Faites la même chose avec la fonction `feols()` du package `fixest` pour `lm` et `etable()` pour `stargazer`.

Retour sur la régression multiple

La première partie de ce TP vise à reproduire manuellement les fonctions de régression basiques de R. On considère le modèle linéaire suivant :

$$Y = X'\beta + u$$

Le but de cet exercice est de reproduire les résultats de la fonction *lm* de R tels qu'obtenus à la partie précédente.

(i) Rappelez la formule générale de l'estimateur des moindres carrés ordinaires (MCO) $\widehat{\beta}$ sous forme matricielle.

(ii) Proposez une fonction sur R qui prennent en argument un vecteur Y et une matrice X de variables dépendantes et renvoie l'estimateur $\widehat{\beta}$. Appliquez votre fonction de façon à reproduire les résultats de la regression précédentes.

(iii) Rappelez la formule de l'estimateur des moindres carrés ordinaires (MCO) $\widehat{\sigma}^2$ de la variance des résidus et la matrice de variance-covariance associée aux coefficients.

(iv) Modifiez la fonction de la question (ii) pour y intégrer l'estimateur des MCO de la déviation standard des résidus, et l'erreur standard $\widehat{\sigma}_{\beta_i}$ associé à chaque coefficient β_i. Combinez $\widehat{\beta}$ et les $\widehat{\sigma}_{\beta_i}$ dans un dataframe avec comme nom de colonnes "coefficient" et "déviation standard". Appliquez votre fonction de façon à reproduire les résultats de la régression précédentes.

(v) Rappelez la formule du R^2. Modifiez votre fonction pour que le dataframe intègre aussi le R^2. Appliquez votre fonction de façon à reproduire les résultats de la régression précédentes.

(vi) Rappelez les formules des statistiques de Fisher F et de student t_i pour chaque coefficient.

(vii) Modifier la fonction pour y ajouter les tests sur le coefficient. Le résultat de votre fonction reverra un dataframe avec des dans des colonnes séparées :

   — les coefficients de la régression
   — les erreurs standards associés aux coefficients
   — les statistiques des tests des coefficients
   — les valeurs des statistiques de student pour un risque de 1%, 5%, 10%
   — les p-value associées
   — la statistique de Fisher et sa p-value.

(viii) Appliquez votre fonction de façon à reproduire les résultats de la régression précédentes.

Exercices

Ces exercices utilisent les données issus du package `wooldridge.`

bwght

Jeu de données : data(bwght) issu du package wooldridge

Un problème qui intéresse les responsables de santé (et d'autres) est de déterminer les effets du tabagisme pendant la grossesse sur la santé des nourrissons. L'une des mesures de la santé du nourrisson est le poids de naissance ; un poids de naissance trop faible peut exposer le nourrisson au risque de contracter diverses maladies. Étant donné que des facteurs autres que le tabagisme qui influent sur le poids à la naissance sont susceptibles d'être corrélés au tabagisme, nous devons tenir compte de ces facteurs. Par exemple, un revenu plus élevé donne généralement accès à de meilleurs soins prénataux, ainsi qu'à une meilleure nutrition pour la mère. Une équation qui en tient compte est la suivante :

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

Avec $bwght$ le poids du bébé à la naissance, $cigs$ le nombre de cigarette fumées par jour par la mère, et $faminc$ le revenu de la famille.

(i) Quel est le signe le plus probable pour β_2 ?
(ii) Pensez-vous que $cigs$ et $faminc$ sont susceptibles d'être corrélés ? Expliquez pourquoi cette corrélation corrélation pourrait être positive ou négative.
(iii) Maintenant, estimez l'équation avec et sans $faminc$

Présentez les résultats sous forme d'équation, y compris la taille de l'échantillon et le R^2. Discutez de vos résultats, en vous concentrant sur la question de savoir si l'ajout de la $faminc$ modifie sensiblement l'effet estimé de la cigarette sur le poids corporel.

hprice1

Utilisez les données **hprice1** pour estimer le modèle :

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

où *price* est le prix de la maison mesuré en milliers de dollars, $sqrft$ est la surface du logement et *bdrms* le nombre de chambres.

(i) Rédigez les résultats sous forme d'équation.

(ii) Quelle est l'augmentation estimée du prix d'une maison comportant une chambre à coucher de plus, la superficie en pieds carrés étant constante ?

(iii) Quelle est l'augmentation estimée du prix d'une maison avec une chambre supplémentaire de 140 pieds carrés ? Comparez ce résultat à votre réponse dans la partie (ii).

(iv) Quel pourcentage de la variation du prix s'explique par la superficie en pieds carrés et le nombre de chambres à coucher ?

(v) La première maison de l'échantillon a une superficie de \$sqrft = \$2438\$ et un nombre de chambres à coucher $bdrms = 4$\$. Trouvez le prix de vente prédit pour cette maison à partir de la ligne de régression MCO.

(vi) Le prix de vente réel de la première maison de l'échantillon est de 300 000 \$ (donc *price* = 300). Trouvez le résidu pour cette maison. Cela suggère-t-il que l'acheteur a sous-payé ou sur-payé la maison ?

ceosal2

Le fichier **ceosal2** contient des données sur 177 PDG et peut être utilisé pour examiner les effets de la performance de l'entreprise sur le salaire du PDG.

(i) Estimez un modèle reliant le salaire annuel *salary* aux ventes de l'entreprise *sales* et à la valeur marchande *mktval*. Transformer le modèle de façon à obtenir des élasticités constantes pour les deux variables indépendantes. Écrivez les résultats sous forme d'équation.

(ii) Ajoutez $profits$ au modèle de la partie (i). Pourquoi cette variable ne peut-elle pas être incluse sous sous forme logarithmique ? Diriez-vous que ces variables de performance de l'entreprise expliquent la plus grande partie de la variation des salaires des PDG ?

(iii) Ajoutez la variable *ceoten* au modèle de la partie (ii). Quel est le pourcentage de rendement estimé estimé pour une année supplémentaire de mandat du PDG, les autres facteurs étant fixes ?

(iv) Trouvez le coefficient de corrélation de l'échantillon entre les variables log(*mktval*) et $profits$. Ces variables sont-elles fortement corrélées ? Qu'est-ce que cela signifie pour les estimateurs MCO ?

attend

Cet exercice étudie les lien entre entre présence en classe et réussite scolaire. Utilisez les données de **attend** pour cet exercice.

(i) Obtenez les valeurs minimum, maximum et moyenne pour les variables *atndrte* (pourcentage de présence en classe), $priGPA$ (GPA cumulé), et ACT (score ACT).

(ii) Estimez le modèle

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u$$

et écrivez les résultats sous forme d'équation. Interprétez l'ordonnée à l'origine. A-t-il une signification utile ?

(iii) Discutez les coefficients de pente estimés. Y a-t-il des surprises ?

(iv) Quel est *atndrte* prédit si $priGPA = 3.65$ et $ACT = 20$? Que pensez-vous de ce résultat ? Existe-t-il des étudiants dans l'échantillon avec ces valeurs des variables explicatives ?

(v) Si l'étudiant A a une $priGPA$ de 3,1 et un ACT de 21, et que l'étudiant B a une $priGPA$ de 2,1 et un ACT de 26, quelle est la valeur de *atndrte* ? et $ACT = 26$, quelle est la différence prédite dans leurs taux de présence ?

meap93

Utilisez les données de **meap93** pour répondre à cette question.

(i) Estimez le modèle

$$math_{10} = \beta_0 + \beta_1 \log(expend) + \beta_2 lnchprg + u$$

et rapportez les résultats sous la forme habituelle, y compris la taille de l'échantillon et le R-carré. Les signes des coefficients de pente sont-ils ceux que vous attendiez ? Expliquez. (ii) Que faites-vous de l'ordonnée à l'origine que vous avez estimée dans la partie (i) ? En particulier, cela a-t-il un sens de mettre les deux variables explicatives à zéro ? (Indice : rappelez-vous que $\log(1) = 0$). (iii) Exécutez maintenant la régression simple de math10 sur $\log(expend)$, et comparez le coefficient de pente avec l'estimation obtenue dans la partie (i). L'effet de dépense estimé est-il maintenant plus grand ou plus petit que dans la partie (i) ? (iv) Trouvez la corrélation entre $lexpend = \log(expend)$ et *lnchprg*. Son signe vous semble-t-il logique ? (v) Utilisez la partie (iv) pour expliquer vos résultats dans la partie (iii).

discrim

Utilisez les données de **discrim** pour répondre à cette question. Il s'agit de données au niveau du code postal sur les prix de divers articles dans les fast-foods, ainsi que des caractéristiques de la population du code postal, dans le New Jersey et en Pennsylvanie. L'idée est de voir si les restaurants fast-food pratiquent des prix plus élevés dans les zones où la concentration de Noirs est plus importante. (i) Trouvez les valeurs moyennes de *prpblck* et de revenu dans l'échantillon, ainsi que leurs écarts types. Quelles sont les unités de mesure de *prpblck* et du revenu ? (ii) Considérez un modèle pour expliquer le prix du soda, *psoda*, en fonction de la proportion de la population qui est noire et du revenu médian :

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 revenu + u$$

Estimez ce modèle par MCO et rapportez les résultats sous forme d'équation, y compris la taille de l'échantillon et le R-carré. (N'utilisez pas la notation scientifique pour présenter les estimations.) Interprétez le coefficient de *prpblck*. Pensez-vous qu'il soit économiquement important ?

(iii) Comparez l'estimation de la partie (ii) avec l'estimation de régression simple de *psoda* sur *prpblck*. L'effet de discrimination est-il plus important ou plus faible lorsque vous contrôlez le revenu ?

(iv) Un modèle avec une élasticité-prix constante par rapport au revenu pourrait être plus approprié. Présentez les estimations du modèle :

$$log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 log(income) + u$$

Si *prpblck* augmente de 0,20 (20 points de pourcentage), quelle est la variation estimée en pourcentage de *psoda* ? (Indice : la réponse est 2.xx, où vous remplissez le "xx"). (v) Ajoutez maintenant la variable *prppov* à la régression de la partie (iv). Que se passe-t-il ? attendu ?

(vii) Évaluez l'énoncé suivant : "Parce que le $log(income)$ et la variable *prppov* sont si fortement corrélés, ils n'ont rien à faire dans la même régression."

charity

Utilisez les données de **charity** pour répondre aux questions suivantes :

(i) Estimez l'équation :

$$gift = \beta_0 + \beta_1 mailsyear + \beta_2 giftlast + \beta_3 propresp + u$$

par MCO et rapportez les résultats de la manière habituelle, y compris la taille de l'échantillon et le R-carré. Comment le R^2 se compare-t-il à celui de la régression simple qui omet *giftlast* et *propresp* ?

(ii) Interprétez le coefficient de l'année postale. Est-il plus grand ou plus petit que le coefficient de régression simple correspondant ?

(iii) Interprétez le coefficient de *propresp*, en prenant soin de noter les unités de mesure de *propresp*.

(iv) Ajoutez maintenant la variable *avggift* à l'équation. Que devient l'effet estimé de *mailsyear* ?

(v) Dans l'équation de la partie (iv), qu'est-il arrivé au coefficient de *giftlast* ? A votre avis, que se passe-t-il ?

htv

Utilisez les données de **htv** pour répondre à cette question. L'ensemble de données comprend des informations sur les salaires, l'éducation, l'éducation des parents et plusieurs autres variables pour 1 230 hommes actifs en 1991.

(i) Quelle est la fourchette de la variable éducation dans l'échantillon ? Quel pourcentage d'hommes ont terminé leur 12ème année mais pas une année supérieure ? Les hommes ou leurs parents ont-ils, en les hommes ou leurs parents ont-ils, en moyenne, un niveau d'éducation plus élevé ?

(ii) Estimez le modèle de régression

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + u$$

par MCO et présentez les résultats sous la forme habituelle. Quelle est la part de la variation de l'échantillon dans *educ* est expliquée par l'éducation des parents ? Interprétez le coefficient de *motheduc*.

(iii) Ajoutez la variable *abil* (une mesure de l'aptitude cognitive) à la régression de la partie (ii), et rapportez les résultats.

(iv) Ajoutez la variable *abil* (une mesure de l'aptitude cognitive) à la régression de la partie (ii), et présentez les résultats sous forme d'équation. La "capacité" permet-elle d'expliquer expliquer les variations de l'éducation, même après avoir contrôlé l'éducation des parents ? Expliquez.

(v) (Nécessite un calcul) Estimez maintenant une équation où l'aptitude apparaît sous forme quadratique :

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u$$

En utilisant les estimations $\hat{\beta}_3$ et $\hat{b}_4$, utiliser le calcul pour trouver la valeur de *abil*, l'appeler $abil^{\star}$, où *educ* est minimisé. (Les autres coefficients et valeurs des variables d'éducation des parents n'ont pas d'effet ; nous maintenons l'éducation des parents fixe). Remarquez que *abil* est mesuré de manière à ce que des valeurs négatives soient autorisées. Vous pouvez également vérifier que la dérivée seconde est positive et que vous avez bien un minimum.

(v) Argumentez que seule une petite fraction des hommes de l'échantillon a une “capacité” inférieure à la valeur calculée dans la partie (iv). En quoi cela est-il important ?

(vi) Si vous avez accès à un programme statistique qui comprend des capacités graphiques, utilisez les estimations de la partie (iv) pour représenter graphiquement la relation entre l'éducation et l'aptitude prédites. Supposons que la *motheduc* et la $fatheduc$ ont leurs valeurs moyennes dans l'échantillon, 12.18 et 12.45, respectivement.

TP3 : Analyse des Disparités Scolaires au Collège

Les inégalités de performance scolaire sont un sujet récurrent dans les débats sur le système éducatif. Parmi les examens importants en France, le brevet des collèges permet de mesurer les compétences acquises par les élèves à la fin du cycle secondaire. Cependant, les résultats obtenus peuvent varier en fonction de divers facteurs, notamment le contexte socio-économique local.

Ce TP vous propose d'explorer l'influence de facteurs socio-économiques, tels que le revenu médian, le taux de chômage ou encore le niveau d'éducation dans les communes, sur les résultats du brevet des collèges. À travers l'analyse de jeux de données réels, vous serez amenés à identifier des corrélations et à mieux comprendre les déterminants de la performance scolaire.

0. Installation.

Charger les packages `tidyverse`, `stargazer`. ChatGPT ou autre chatbot sont autorisés pour ce TP.

Les données socio-économiques sont bien formatées ici : https://www.unehistoireduconflitpolitique.fr/telecharger.html. Commencer par télécharger les données sur les revenus des communes. On pourra réitérer l'analyse sur les diplômes et les catégories socio-professionnelles.

1. Chercher sur internet et télécharger les données sur les résultats de brevets par établissement.

1. Données Brevet

1.1 Description des données

1. Décrire le jeu de données : colonnes, taille, niveau géographique, horizon temporel…
2. Quelle est la période étudiée ?
3. Combien y a-t-il d'établissements ?

1.2 Evolution temporelle

On veut caractériser les résultas du brevet au niveau national.

1. Créer un fichier de donnée aggrégé par année au niveau national (utiliser `group_by` et `summarize`).
2. Comment semble calculé la colonne `taux_de_reussite`. Tester son intuition une colonne `taux_de_reussite2` et comparer avec `taux_de_reussite`.
3. Faire des graphiques montrant l'évolution des nombres d'inscrits et d'admis.
4. Faire des graphiques montrant le taux annuel d'admis.
5. Faire des graphiques montrant les taux annuels d'admis pour chaque mention.
6. Décrire et interpréter chaque graphiques.

1.3 Variation en coupe

On considère la dernière session reportée par le jeu de donnée.

1. Créer un jeu de donnée filtré sur cette dernière année.
2. Montrer es graphiques en barres pour représenter les différences sur les taux de réussites selon le type d'établissement et le secteur d'enseignement.
3. Faire des classements des dix meilleurs départements selon les différents taux de réussites.

2. Données socio-économiques

2.1 Description du jeu de données

1. Décrire le jeu de données de la même façon que pour le premier jeu. Utiliser les annexes où les données sont décrites.
2. Quelles colonnes (ou ensemble de colonnes) vous semble-t-il pertinent de garder ?

2.2 Transformation du jeu

Transformer ce jeu de données en format "long" avec `pivot_longer`.

3. Analyse jointe .

3.1 Jointure

Pour chaque année et pour chaque établissement, on souhaite avoir les informations socio-économiques de la commune correspondante.

1. Faire la jointure entre les deux jeux de données.
2. Analyser les données manquantes du nouveau jeu de données.

3.2 Analyse en coupe

1. Faire des graphiques par points représentant le revenu moyen de la commune de l'établissement avec ses différents taux de réussites.
2. Faire des graphiques en bar dans lequel par décile de revenu (utiliser la colonne de percentile coté socioeco).

3.3 Regressions linéaire

Pour une année donnée vs toutes taux de reussite en fonction de la taille de la commune, revenus moyen

1. Faites une régression

Notes : - on fait les régressions avec la commande `lm`. - Pour visualiser les régressions, on enregistre les résultats de chaque regressions (eg, `lm1`,`lm2`…) et on visualise avec la commande `stargazer` du package du même nom (eg `stargazer(type="text",lm1,lm2)`).

TP4 : Etude économétrique de l'Enquête Nationale Transport 2019

1. Enoncé

Dans ce TP, nous allons travaillons sur enquête nationale de l'INSEE. Vous aurez la liberté de choisir une question de recherche et de sélectionner les variables qui vous semblent pertinentes (en n'en prenant pas trop tout de même).

Les données sont disponibles ici : sur le site du ministère.

On considère les caractéristiques socio-économiques des ménages suivantes : le revenu, la catégorie socio-professionnelle, le lieux de résidence, la composition du ménage (nombre de personnes, age).

Questions : comment varie les grandeurs suivantes en fonction des grandeurs suivantes :

- les caractéristiques du véhicule : age, motorisation
- distance et nombre de trajets parcourue à velo pour les trajets du quotidient
- distance et nombre de trajets parcourue en voiture en commun pour les trajets du quotidient
- distance et nombre de trajets parcourue en transport en commun pour les trajets du quotidient
- distance et nombre de trajets parcourue en avion pour les voyages

Travail à faire :

1. identifier dans les jeux de données où trouver les informations pertinentes
2. effectuer des statistiques descriptives sur les caractéristiques socio-économiques
3. construire votre jeu de donnée en construisant les grandeurs de transport puis en réalisant en appariemment sur les données socioéconomiques.
4. Effectuer des régressions linéaires en combinant différemment les variables de controlms.

Cinquième partie

Econométrie 2

Introduction

Le cours en résumé :

- 24 h
- objectifs : économétrie avancée : tests statistiques (en parallèle du cours) et introduction à l'inférence causale
- modalités d'examens :
  - Note de participation (10%)
  - Présentations en groupe en classe de chapitre du manuel d'inférence causale de Scott Cunningham (40%)
  - Projet 3 d'économétrie (50%)
  - S'incrire ici pour les projets et présentations : fichier Excel partagé.

Le plan :

1. Régression linéaire multiple et tests
2. Binaires et effets fixes
3. Multicolinéarité
4. Hétéroskédasticité
5. Econométrie de Panel

Présentation :

Dans l'atelier Econométrie 1, nous nous sommes introduit aux concepts élémentaires de l'économétries : régressions linéaires multiples, estimation, significativité. Cet atelier vise à étendre votre connaissance des modèles linéaires sur les cas où les hypothèses de bases ne sont pas respectées (multicolinéarité, hétéroskédasticité etc). En parallèle, il cherche à vous introduire à l'approche de l'inférence causale, qui est le concept empirique dominant en économie et sciences sociales aujourd'hui. Pour ce faire, nous allons faire une lecture collective de l'ouvrage de Scott Cunningham, Causal Inference : the Mixtape. Dans le projet, nous reprendrons le projet 2 et chercherons à appliquer une des méthodes les plus classiques : les différences en différences.

1 TP 1 : Modèle de régression multiple

```
library(tidyverse)
library(wooldridge)
library(AER)
library(stargazer)
library(fixest)

knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, eval=FALSE)
```

1.1 Introduction

Lancez **RStudio** et ouvrez un nouveau **R Markdown**. Ce sera votre document de travail durant tout le TP. L'intérêt des Markdown est de pouvoir lancer une multitude de petits scripts successivement. Typiquement, un *chunk* (petit script) par question. Lorsque c'est nécessaire, répondez aux questions entre les chunks.

Créez un fichier partagé "Econométrie-votre nom" sur votre drive fourni de l'université (OneDrive ou Google Drive). Envoyez-moi le lien de ce drive. Ce dossier partagé sera votre dossier de travail pendant toute la durée du cours. Créez un dossier "TP1" et enregistrez-y votre markdown sous le nom "TP1".

1.2 Rappel sur R

(i) Chargez le package **AER** (téléchargez-le si besoin). Ouvrez le fichier de données `CASchools` avec la commande : `data(CASchools)`.
(ii) Définissez les variables $STR = students/teachers$ et $score = (read + math)/2$.
(iii) Affichez les statistiques descriptives de ce jeu de données avec la commande **summary**.
(iv) Estimez le modèle suivant avec la commande **lm** et affichez les résultats de la régression :

$$score = \beta_0 + \beta_1 STR + \beta_2 english + u$$

(v) Reprérentez els résultats précédents à l'air du **stargazer** et utilisez la commande stargazer(*data* , type = 'text') en remplaçant *data* par les jeux de données précédemment appelés.

(vi) Faites la même chose avec la fonction `feols()` du package `fixest` pour `lm` et `etable()` pour `stargazer`.

1.3 Retour sur la régression multiple

La première partie de ce TP vise à reproduire manuellement les fonctions de régression basiques de R. On considère le modèle linéaire suivant :

$$Y = X'\beta + u$$

Le but de cet exercice est de reproduire les résultats de la fonction *lm* de R tels qu'obtenus à la partie précédente.

(i) Rappelez la formule générale de l'estimateur des moindres carrés ordinaires (MCO) $\widehat{\beta}$ sous forme matricielle.

(ii) Proposez une fonction sur R qui prennent en argument un vecteur Y et une matrice X de variables dépendantes et renvoie l'estimateur $\widehat{\beta}$. Appliquez votre fonction de façon à reproduire les résultats de la regression précédentes.

(iii) Rappelez la formule de l'estimateur des moindres carrés ordinaires (MCO) $\widehat{\sigma}^2$ de la variance des résidus et la matrice de variance-covariance associée aux coefficients.

(iv) Modifiez la fonction de la question (ii) pour y intégrer l'estimateur des MCO de la déviation standard des résidus, et l'erreur standard $\widehat{\sigma}_{\beta_i}$ associé à chaque coefficient β_i. Combinez $\widehat{\beta}$ et les $\widehat{\sigma}_{\beta_i}$ dans un dataframe avec comme nom de colonnes "coefficient" et "déviation standard". Appliquez votre fonction de façon à reproduire les résultats de la régression précédentes.

(v) Rappelez la formule du R^2. Modifiez votre fonction pour que le dataframe intègre aussi le R^2. Appliquez votre fonction de façon à reproduire les résultats de la régression précédentes.

(vi) Rappelez les formules des statistiques de Fisher F et de student t_i pour chaque coefficient.

(vii) Modifier la fonction pour y ajouter les tests sur le coefficient. Le résultat de votre fonction reverra un dataframe avec des dans des colonnes séparées :

— les coefficients de la régression
— les erreurs standards associés aux coefficients
— les statistiques des tests des coefficients
— les valeurs des statistiques de student pour un risque de 1%, 5%, 10%

— les p-value associées
— la statistique de Fisher et sa p-value.

(viii) Appliquez votre fonction de façon à reproduire les résultats de la régression précédentes.

1.4 Exercices

Ces exercices utilisent les données issus du package `wooldridge.`

1.4.1 bwght

Jeu de données : data(bwght) issu du package wooldridge

Un problème qui intéresse les responsables de santé (et d'autres) est de déterminer les effets du tabagisme pendant la grossesse sur la santé des nourrissons. L'une des mesures de la santé du nourrisson est le poids de naissance ; un poids de naissance trop faible peut exposer le nourrisson au risque de contracter diverses maladies. Étant donné que des facteurs autres que le tabagisme qui influent sur le poids à la naissance sont susceptibles d'être corrélés au tabagisme, nous devons tenir compte de ces facteurs. Par exemple, un revenu plus élevé donne généralement accès à de meilleurs soins prénataux, ainsi qu'à une meilleure nutrition pour la mère. Une équation qui en tient compte est la suivante :

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u$$

Avec $bwght$ le poids du bébé à la naissance, $cigs$ le nombre de cigarette fumées par jour par la mère, et $faminc$ le revenu de la famille.

(i) Quel est le signe le plus probable pour β_2 ?
(ii) Pensez-vous que $cigs$ et $faminc$ sont susceptibles d'être corrélés ? Expliquez pourquoi cette corrélation corrélation pourrait être positive ou négative.
(iii) Maintenant, estimez l'équation avec et sans $faminc$

Présentez les résultats sous forme d'équation, y compris la taille de l'échantillon et le R^2. Discutez de vos résultats, en vous concentrant sur la question de savoir si l'ajout de la $faminc$ modifie sensiblement l'effet estimé de la cigarette sur le poids corporel.

1.4.2 hprice1

Utilisez les données **hprice1** pour estimer le modèle :

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u$$

où *price* est le prix de la maison mesuré en milliers de dollars, $sqrft$ est la surface du logement et *bdrms* le nombre de chambres.

(i) Rédigez les résultats sous forme d'équation.

(ii) Quelle est l'augmentation estimée du prix d'une maison comportant une chambre à coucher de plus, la superficie en pieds carrés étant constante ?

(iii) Quelle est l'augmentation estimée du prix d'une maison avec une chambre supplémentaire de 140 pieds carrés ? Comparez ce résultat à votre réponse dans la partie (ii).

(iv) Quel pourcentage de la variation du prix s'explique par la superficie en pieds carrés et le nombre de chambres à coucher ?

(v) La première maison de l'échantillon a une superficie de \$sqrft = \$2438\$ et un nombre de chambres à coucher $bdrms = 4$\$. Trouvez le prix de vente prédit pour cette maison à partir de la ligne de régression MCO.

(vi) Le prix de vente réel de la première maison de l'échantillon est de 300 000 \$ (donc *price* = 300). Trouvez le résidu pour cette maison. Cela suggère-t-il que l'acheteur a sous-payé ou sur-payé la maison ?

1.4.3 ceosal2

Le fichier **ceosal2** contient des données sur 177 PDG et peut être utilisé pour examiner les effets de la performance de l'entreprise sur le salaire du PDG.

(i) Estimez un modèle reliant le salaire annuel *salary* aux ventes de l'entreprise *sales* et à la valeur marchande *mktval*. Transformer le modèle de façon à obtenir des élasticités constantes pour les deux variables indépendantes. Écrivez les résultats sous forme d'équation.

(ii) Ajoutez $profits$ au modèle de la partie (i). Pourquoi cette variable ne peut-elle pas être incluse sous sous forme logarithmique ? Diriez-vous que ces variables de performance de l'entreprise expliquent la plus grande partie de la variation des salaires des PDG ?

(iii) Ajoutez la variable *ceoten* au modèle de la partie (ii). Quel est le pourcentage de rendement estimé estimé pour une année supplémentaire de mandat du PDG, les autres facteurs étant fixes ?

(iv) Trouvez le coefficient de corrélation de l'échantillon entre les variables $\log(mktval)$ et $profits$. Ces variables sont-elles fortement corrélées ? Qu'est-ce que cela signifie pour les estimateurs MCO ?

1.4.4 attend

Cet exercice étudie les lien entre entre présence en classe et réussite scolaire. Utilisez les données de **attend** pour cet exercice.

(i) Obtenez les valeurs minimum, maximum et moyenne pour les variables *atndrte* (pourcentage de présence en classe), *priGPA* (GPA cumulé), et *ACT* (score ACT).

(ii) Estimez le modèle

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u$$

et écrivez les résultats sous forme d'équation. Interprétez l'ordonnée à l'origine. A-t-il une signification utile ?

(iii) Discutez les coefficients de pente estimés. Y a-t-il des surprises ?

(iv) Quel est *atndrte* prédit si $priGPA = 3.65$ et $ACT = 20$? Que pensez-vous de ce résultat ? Existe-t-il des étudiants dans l'échantillon avec ces valeurs des variables explicatives ?

(v) Si l'étudiant A a une *priGPA* de 3,1 et un *ACT* de 21, et que l'étudiant B a une *priGPA* de 2,1 et un *ACT* de 26, quelle est la valeur de *atndrte* ? et $ACT = 26$, quelle est la différence prédite dans leurs taux de présence ?

1.4.5 meap93

Utilisez les données de **meap93** pour répondre à cette question.

(i) Estimez le modèle

$$math_{10} = \beta_0 + \beta_1 \log(expend) + \beta_2 lnchprg + u$$

et rapportez les résultats sous la forme habituelle, y compris la taille de l'échantillon et le R-carré. Les signes des coefficients de pente sont-ils ceux que vous attendiez ? Expliquez. (ii) Que faites-vous de l'ordonnée à l'origine que vous avez estimée dans la partie (i) ? En particulier, cela a-t-il un sens de mettre les deux variables explicatives à zéro ? (Indice : rappelez-vous que $\log(1) = 0$). (iii) Exécutez maintenant la régression simple de math10 sur $\log(expend)$, et comparez le coefficient de pente avec l'estimation obtenue dans la partie (i). L'effet de dépense estimé est-il maintenant plus grand ou plus petit que dans la partie (i) ? (iv) Trouvez la corrélation entre $lexpend = \log(expend)$ et *lnchprg*. Son signe vous semble-t-il logique ? (v) Utilisez la partie (iv) pour expliquer vos résultats dans la partie (iii).

1.4.6 discrim

Utilisez les données de **discrim** pour répondre à cette question. Il s'agit de données au niveau du code postal sur les prix de divers articles dans les fast-foods, ainsi que des caractéristiques de la population du code postal, dans le New Jersey et en Pennsylvanie. L'idée est de voir si les restaurants fast-food pratiquent des prix plus élevés dans les zones où la concentration de Noirs est plus importante. (i) Trouvez les valeurs moyennes de *prpblck* et de revenu dans l'échantillon, ainsi que leurs écarts types. Quelles sont les unités de mesure de *prpblck* et du revenu ? (ii) Considérez un modèle pour expliquer le prix du soda, *psoda*, en fonction de la proportion de la population qui est noire et du revenu médian :

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 revenu + u$$

Estimez ce modèle par MCO et rapportez les résultats sous forme d'équation, y compris la taille de l'échantillon et le R-carré. (N'utilisez pas la notation scientifique pour présenter les estimations.) Interprétez le coefficient de *prpblck*. Pensez-vous qu'il soit économiquement important ?

(iii) Comparez l'estimation de la partie (ii) avec l'estimation de régression simple de *psoda* sur *prpblck*. L'effet de discrimination est-il plus important ou plus faible lorsque vous contrôlez le revenu ?

(iv) Un modèle avec une élasticité-prix constante par rapport au revenu pourrait être plus approprié. Présentez les estimations du modèle :

$$log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 log(income) + u$$

Si *prpblck* augmente de 0,20 (20 points de pourcentage), quelle est la variation estimée en pourcentage de *psoda* ? (Indice : la réponse est 2.xx, où vous remplissez le “xx”). (v) Ajoutez maintenant la variable *prppov* à la régression de la partie (iv). Que se passe-t-il ? attendu ?

(vii) Évaluez l'énoncé suivant : “Parce que le $log(income)$ et la variable *prppov* sont si fortement corrélés, ils n'ont rien à faire dans la même régression.”

1.4.7 charity

Utilisez les données de **charity** pour répondre aux questions suivantes :

(i) Estimez l'équation :

$$gift = \beta_0 + \beta_1 mailsyear + \beta_2 giftlast + \beta_3 propresp + u$$

par MCO et rapportez les résultats de la manière habituelle, y compris la taille de l'échantillon et le R-carré. Comment le R^2 se compare-t-il à celui de la régression simple qui omet *giftlast* et *propresp* ?

(ii) Interprétez le coefficient de l'année postale. Est-il plus grand ou plus petit que le coefficient de régression simple correspondant ?

(iii) Interprétez le coefficient de *propresp*, en prenant soin de noter les unités de mesure de *propresp*.

(iv) Ajoutez maintenant la variable *avggift* à l'équation. Que devient l'effet estimé de *mailsyear* ?

(v) Dans l'équation de la partie (iv), qu'est-il arrivé au coefficient de *giftlast* ? A votre avis, que se passe-t-il ?

1.4.8 htv

Utilisez les données de **htv** pour répondre à cette question. L'ensemble de données comprend des informations sur les salaires, l'éducation, l'éducation des parents et plusieurs autres variables pour 1 230 hommes actifs en 1991.

(i) Quelle est la fourchette de la variable éducation dans l'échantillon ? Quel pourcentage d'hommes ont terminé leur 12ème année mais pas une année supérieure ? Les hommes ou leurs parents ont-ils, en les hommes ou leurs parents ont-ils, en moyenne, un niveau d'éducation plus élevé ?

(ii) Estimez le modèle de régression

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + u$$

par MCO et présentez les résultats sous la forme habituelle. Quelle est la part de la variation de l'échantillon dans *educ* est expliquée par l'éducation des parents ? Interprétez le coefficient de *motheduc*.

(iii) Ajoutez la variable *abil* (une mesure de l'aptitude cognitive) à la régression de la partie (ii), et rapportez les résultats.

(iv) Ajoutez la variable *abil* (une mesure de l'aptitude cognitive) à la régression de la partie (ii), et présentez les résultats sous forme d'équation. La "capacité" permet-elle d'expliquer expliquer les variations de l'éducation, même après avoir contrôlé l'éducation des parents ? Expliquez.

(v) (Nécessite un calcul) Estimez maintenant une équation où l'aptitude apparaît sous forme quadratique :

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u$$

En utilisant les estimations $\hat{\beta}_3$ et $\hat{b}_4$, utiliser le calcul pour trouver la valeur de *abil*, l'appeler $abil^\star$, où *educ* est minimisé. (Les autres coefficients et valeurs des variables d'éducation des parents n'ont pas d'effet ; nous maintenons l'éducation des parents fixe). Remarquez que *abil* est mesuré de manière à ce que des valeurs négatives soient autorisées. Vous pouvez également vérifier que la dérivée seconde est positive et que vous avez bien un minimum.

(v) Argumentez que seule une petite fraction des hommes de l'échantillon a une “capacité” inférieure à la valeur calculée dans la partie (iv). En quoi cela est-il important ?

(vi) Si vous avez accès à un programme statistique qui comprend des capacités graphiques, utilisez les estimations de la partie (iv) pour représenter graphiquement la relation entre l'éducation et l'aptitude prédites. Supposons que la *motheduc* et la $fatheduc$ ont leurs valeurs moyennes dans l'échantillon, 12.18 et 12.45, respectivement.

2 TP 2 : Régression multiple et variables binaires

```
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, eval=FALSE)

library(tidyverse)
library(wooldridge)
library(AER)
library(stargazer)
library(fixest)
```

2.1 Exercices

2.1.1 gpa1

i. Estimez le modèle :
$$colGPA = PC + hsGPA + ACT$$
.
ii. Ajoutez les variables `mothcoll` et `fathcoll` et reportez les résultats d'estimation. Qu'observez-vous pour le coefficient de la variable relative à la possession d'un PC ? `PC` apparaît-il toujours statistiquement significatif ?
iii. Testez pour la significativité jointe des coefficients associés aux variables `mothcoll` et `fathcoll` issues de l'équation en (i) en mentionnant les p-valeurs.
iv. Ajoutez `hsGPA^2` au modèle décrit en (i), jugez-vous cette généralisation pertinente ?

2.1.2 wage2

i. Estimez le modèle suivant :

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

et reportez les résultats. Quelle est la différence approximative de salaire mensuel entre les afro-américains et les non afro-américains, toutes choses égales par ailleurs ? Cette différence est-elle statistiquement significative ?

ii. Ajoutez les variables $exper^2$ et $tenure^2$ à l'équation précédente et montrez que les coefficients associés ne sont pas conjointement statistiquement significatifs au seuil de 5 %.
iii. Amendez le modèle original pour permettre aux rendements de l'éducation de dépendre des origines ethniques et testez si effectivement les rendements de l'éducation en dépendent.
iv. À nouveau, partez du modèle de base et amendez le ensuite en autorisant les salaires à différer selon les groupes : mariés et africains américains, mariés et non africains américains, non mariés et africains américains, non mariés et non africains américains. Quelle est la différence de salaire estimée entre les mariés afro-américains et non afro-américains ?

2.1.3 mlb1

Un modèle permettant de caractériser les salaires des joueurs de la ligue majeure de baseball aux États-Unis est donné par :

$$\begin{gathered}
log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr \\
+\beta_6 runsyr + \beta_7 fldperc + \beta_8 allstar + \beta_9 frstbase + \beta_{10} scndbase \\
+beta_{11} thrdbase + \beta_{12} shrtstop + \beta_{13} catcher + u
\end{gathered}$$

avec `outfield` (joueur de champ extérieur) le groupe de référence.

i. Établissez formellement l'hypothèse nulle selon laquelle, toutes choses égales par ailleurs, attrapeurs (`catcher`) et joueurs de champ extérieur gagnent en moyenne le même revenu. Calculez le salaire moyen par groupe.
ii. Testez ensuite cette hypothèse à partir des données contenues dans la base `mlb1` et commentez l'étendue du différentiel de salaire.
iii. Établissez formellement et testez l'hypothèse selon laquelle il n'y a aucune différence de salaire moyen selon les postes occupés, une fois l'ensemble des variables de contrôle prises en compte.
iv. Les résultats tirés des questions (i) et (ii) sont-ils cohérents ? Dans le cas contraire, expliquez le mécanisme sous-jacent.

2.1.4 gpa2

i. On considère le modèle suivant :

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

avec `colgpa` la moyenne des notes obtenues en premier cycle à l'université, `hsize` la taille de la classe au lycée, en centaines, `hsperc` le quantile dans lequel se situe l'étudiant à l'issue de son parcours universitaire, `sat` le score SAT, `female` une variable binaire relative au genre, et `athlete` une variable binaire prenant la valeur un si l'étudiant est un athlète.

Quelles sont vos attentes relativement aux valeurs des différents coefficients du modèle ? Quels sont ceux pour lesquels vous avez des doutes ?

ii. Estimez l'équation mentionnée en question (i) et reportez les résultats de façon standard. Quel est le différentiel de GPA estimé entre les athlètes et les non athlètes ? Cette différence est-elle statistiquement significative ?
iii. Enlevez sat du modèle et ré-estimez l'équation. Quel est maintenant l'effet estimé du statut d'athlète ? Discutez les raisons pour lesquelles ces résultats diffèrent de ceux présentées à la question (ii).
iv. Dans le modèle décrit à la question (i), autorisez l'impact du statut d'athlète à différer selon le genre et testez l'hypothèse nulle d'absence de différence entre les femmes et les hommes athlètes, toutes choses égales par ailleurs.
v. L'effet de sat sur colgpa diffère-t-il selon le genre ? Justifiez votre réponse.

2.1.5 charity

Utilisez les données contenues dans le fichier `charity` pour répondre à cette question. La variable respond est une variable indicatrice égale à un si un individu a répondu par un don à la sollicitation par mail la plus récente d'une association caritative. La variable `resplast` est une variable indicatrice égale à un si un individu a répondu à la sollicitation par mail précédente, `avggift` est la moyenne des dons passés (en florins néerlandais) et `propresp` est le temps relatif passé par la personne à répondre aux sollicitations passées.

i. Estimez un modèle à probabilités linéaires expliquant respond en fonction de `resplast` et `avggift.` Reportez les résultats sous une forme usuelle et interprétez le coefficient de resplast.
ii. La valeur moyenne des dons passés semble-t-elle affecter la probabilité de réponse ?
iii. Ajoutez la variable propresp au modèle et interprétez son coefficient. (Soyez attentif ici : une hausse de une unité de `propresp` est le changement le plus important possible.)
iv. Qu'est-il advenu du coefficient de `resplast` lorsque `propresp` a été ajouté au modèle de régression ? Cela a-t-il du sens ?

v. Ajoutez `mailsyear`, le nombre de mails envoyés par an, au modèle. Evaluez l'effet associé ? Pourquoi n'est-ce sans doute pas une mesure correcte de l'effet causal de l'envoi ciblé de mails (ou « publipostage ») sur la probabilité de réponse ?

2.1.6 catholic

i. Considérez l'échantillon complet et identifiez quel pourcentage d'étudiants sont régulièrement inscrit dans un établissement secondaire d'obédience catholique. Quelle est la moyenne de la variable math12 ?
ii. Réalisez une régression simple de math12 sur cathhs et reportez les résultats selon les normes usuelles. Interprétez vos résultats.
iii. Ajoutez maintenant les variables `lfaminc`, `motheduc`, et `fatheduc` au modèle de régression précédent. De combien d'observations disposez-vous pour cette régression ? Qu'advient-il du coefficient de la variable `cathhs`, apparaît-il significatif ?
iv. Reprenez le cas d'une régression simple de `math12` sur `cathhs`, en restreignant les observations à celles utilisées dans la régression multiple de la question (iii). Certaines de vos conclusions en sont-elles modifiées ?
v. À partir du modèle de régression étudié en question (iii), ajoutez des variables d'interaction entre `cathhs` et chacune des autres variables explicatives du modèle. Ces variables d'interaction sont-elles individuellement et/ou conjointement significatives ?
vi. Qu'advient-il du coefficient de `cathhs` dans la régression de la question (v). Expliquez pourquoi ce coefficient n'est pas très intéressant à étudier.
vii. Calculez l'effet marginal moyen de `cathhs` dans le model estimé en question (v). Comparez les résultats obtenus avec les coefficients de `cathhs` estimés dans la parties (iii) et (v).

2.1.7 apple

i. Définissez une variable binaire $ecobuy = 1$ si $ecolbs > 0$ et $ecobuy = 0$ si $ecolbs = 0$. En d'autres termes, `ecobuy` indique si, à prix donnés, une famille consomme des pommes issues de l'agriculture biologique. Combien de familles affirment acheter des pommes labélisées bio ?
ii. Estimez le modèle à probabilités linéaires suivant :

$$ecobuy = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc + \beta_4 hhsize + \beta_5 educ + \beta_6 age$$

et reportez les résultats sous une forme usuelle. Interprétez avec soin les coefficients associés aux variables de prix.
iii. Y-a-t-il des variables non associées aux prix qui soient conjointement significatives dans le modèle ? (Utilisez les statistiques de Fisher usuelles même si celles-ci ne sont pas valides en présence d'hétéroscédasticité.) Quelle variable explicative autre que les variables de

prix semblent avoir l'effet le plus important sur la décision de consommer des pommes issues de l'agriculture biologique ? Cela vous paraît-il avoir du sens ?

iv. À partir du modèle discuté dans la question (ii), remplacez faminc par `log(faminc)`. Lequel des deux modèles vous semble le mieux expliquer vos données celui introduisant `faminc` ou `log(faminc)` ? Interprétez le coefficient associé à `log(faminc)`.

v. Sur base des estimations réalisées en question (iv), combien des probabilités estimées s'avèrent négatives ? Combien s'avèrent supérieures à l'unité ? En quoi cela devrait-il vous interpeler ?

vi. Revenons à l'estimation de la question (iv), calculez le pourcentage de prédictions correctes pour chacune des valeurs possible de la variable indépendante c'est-à-dire `ecobuy = 0` puis `ecobuy = 1`. Quel résultat s'avère le mieux prédit par le modèle ?

3 Multicolinéarité

#TP3 : Multicolinéarité

```
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE, eval=FALSE)

library(tidyverse)
library(wooldridge)
library(AER)
library(stargazer)
library(fixest)
library(AER)
library(calibrate)
library(PerformanceAnalytics)
library(ISLR)
library(corrplot)
library(mctest)
```

%On considère le modèle général : %

$$Y = X'\beta + u$$

Installer et charger les packages :

- `ISLR` : contient le jeu de données ‘Auto}
- `corrplot` : contient des fonctions de visualisation de matrice de correlation
- `PerformanceAnalytics` : contient des fonctions de visualisation de matrice de corrélation
- `mctest` : contient des fonctions de test de multicolinéarité

3.0.1 Matrice de corrélation

(i) Charger les données `Auto` du package `ISLR`. Estimez le modèle :

$$mpg = \beta_0 + \beta_1 cylinders + \beta_2 horsepower + \beta_3 weight + \beta_4 acceleration + \beta_5 displacement + \beta_6 year + u$$

(ii) Calculer la matrice de corrélation des covariables avec la commande native `cor`.

(iii) Utiliser la commande `corrplot.mixed` issue de `corrplot`. Pour visualiser cette matrice.

(iv) Méthode alternative : utiliser la commande `chart.Correlation(data, histogram=TRUE, pch=19)` issue de `PerformanceAnalytics`.

(v) Que peut-on dire sur le niveau de multicolinéarité ? Commenter.

3.0.2 VIF

(i) Rappelez l'expression du Variance Inflation Factor VIF_i d'une covariable i en fonction de R_i^2, le coefficient de détermination de la régression linéaire $X_i = \gamma X_{-i}$, avec X_{-i} étant tous les covariables autres que X_i.

(ii) Ecrire une fonction `VIF.test(X)` qui associe à une matrice de covariables un vecteur contenant les VIF_i associés à chaque covariable X_i. Indication : il faut faire k régressions et pour chaque régression, récupérer le R^2 à l'aide de la commande `summary(regression)$r.squared`.

(iii) Vérifier votre fonction à l'aide de la fonction `imcdiag` de `mctest`.

3.0.3 Test de Farrar et Glauber

Dans cette section, nous allons recoder le test de Farrar et Glauber. Il s'agit de créer une fonction `fg.test(model)` qui prend un *model* issu d'une régression de *lm* et d'effectuer :

1. Extraire la matrice des covariables en excluant le terme constant
2. Définir la matrice de corrélation de cette matrice et calculer son déterminant D.
3. La deuxième étape de ce test vise à effectuer un test du chi2 sur les hypothèses suivantes : $ `{=tex} \begin{array}{cc} H_0: & D=1 \\ H_1: & D<1 \end{array}` $

Calculer la statistique du Chi-2 associé au test de Farrar et Glauber :

$$\chi^2_{calc} = -log(D)(n - 1 - \frac{1}{6}(2k + 7))$$

avec n est le nombre d'observations et k le nombre de covariables.

4. Calculer la pvalue associée à χ^2_{calc}. Le chi2 du test de Farrar et Glauber suit une loi du chi2 à $\frac{1}{2}k(k + 1)$ degrés de liberté.

3.0.4 Essais sur d'autres jeux de données

Appliquez vos fonctions `VIF.test(X)` et `fg.test(X)` sur les modèles et jeux de données des exercices 1, 2 et 3 du TP1.

4 Tests sur la forme fonctionnelle

#TP 4 : Forme fonctionnelle

On considère le modèle linéaire suivant :

$$Y = X'\beta + u$$

4.0.1 Test de Ramsey (RESET)

Le test de Ramsey cherche à tester si des spécifications non-linéaires n'apportent pas un pouvoir explicatif supérieur au modèle linéaire. Pour ce faire, on test si $(X'\beta)^2$, $(X'\beta)^3$, …, $(X'\beta)^m$ on un effet significatif sur Y.

1. Dans R, partons du jeu de données `hprice1` du package wooldridge, et estimez le modèle :

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms$$

Puis le modèle augmenté :

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + \beta_4 \widehat{price}^2 + \beta_5 \widehat{price}^3$$

avec $\widehat{price}$ la prédiction de $price$ du premier modèle.

2. Effectuez un test de Fisher de restrictions aux coefficients $\beta_4 = \beta_5 = 0$. Construisez ce test manuellement et vérifiez avec les résultats obtenus avec la fonction `linearHypothesis` du package `car`.
3. Généralisez les étapes ci-dessus en construisant une fonction qui prend en argument une régression obtenu avec la fonction `lm` et un entier m qui donne l'exposant maximal. Votre fonction renvoie les résultats du test de Fisher.
4. Vérifiez vos résultats avec la fonction `resettest` du package `lmtest`.

4.1 Exercices

4.1.1 kielmc

Utilisez les données de `kielmc`, uniquement pour l'année 1981, pour répondre aux questions suivantes. Les données concernent les maisons vendues en 1981 à North Andover, Massachusetts ; 1981 est l'année du début de la construction d'un incinérateur d'ordures local.

1. Pour étudier les effets de l'emplacement de l'incinérateur sur le prix des logements, considérez le modèle de régression simple suivant :

$$log(prix) = \beta_0 + \beta_1 log(dist) + u$$

où le prix est le prix du logement en dollars et *dist* est la distance entre la maison et l'incinérateur, mesurée en pieds. En interprétant cette équation de manière causale, quel signe attendez-vous pour β_1 si la présence de l'incinérateur fait baisser les prix des logements ? Estimez cette équation et interprétez les résultats.

2. Au modèle de régression simple de la partie 1. , ajoutez les variables $\log(intst)$, $\log(area)$, $\log(land)$, *rooms*, *baths* et *age*, où *intst* est la distance entre le domicile et l'inter-état, *area* est la superficie de la maison, land est la taille du terrain en pieds carrés, *rooms* est le nombre total de pièces, *baths* est le nombre de salles de bain et *age* est l'âge de la maison en années. Que concluez-vous sur les effets de l'incinérateur ? Expliquez pourquoi 1. et 2. donnent des résultats contradictoires.
3. Ajoutez$log(intst)^2$ au modèle de la partie 2. . Que se passe-t-il maintenant ? Qu'en concluez-vous quant à l'importance de la forme fonctionnelle ?
4. Le carré de $\log(dist)$ est-il significatif lorsque vous l'ajoutez au modèle de la partie 3. ?

4.1.2 wage1

1. Utilisez les MCO pour estimer l'équation suivante

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper2 + u$$

et présentez les résultats en utilisant le format habituel.

2. $exper2$ est-il statistiquement significatif au niveau de 1% ?
3. En utilisant l'approximation

trouvez le rendement approximatif de la cinquième année d'expérience. Quel est le rendement approximatif de la vingtième année d'expérience ? Quel est le rendement approximatif de la vingtième année d'expérience ?

4. A partir de quelle valeur de *exper* l'expérience supplémentaire fait-elle baisser le $\log(wage)$ prédit ? Combien de personnes ont plus d'expérience dans cet échantillon ?

4.1.3 gpa2

1. Estimez le modèle

$$sat = \beta_0 + \beta_1 hsize + b2hsize2 + u$$

où *hsize* est la taille de la classe de diplômés (en centaines), et écrivez les résultats sous la forme habituelle. Le terme quadratique est-il statistiquement significatif ?

2. En utilisant l'équation estimée dans la partie 1. , quelle est la taille "optimale" de l'école secondaire ? Justifiez votre réponse.
3. Cette analyse est-elle représentative des résultats scolaires de tous les élèves de terminale ? lycée ? Expliquez votre réponse.
4. Trouvez la taille optimale estimée de l'école secondaire, en utilisant le $\log(sat)$ comme variable dépendante. variable dépendante. Est-ce très différent de ce que vous avez obtenu dans la partie 2. ?

4.1.4 hprice1

1. Estimez le modèle :

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

%

$$log(price) = \beta_0 + \beta_1 log(lotsize) + b2log(sqrft) + \beta_3 bdrms + u$$

2. Trouvez la valeur prédite de $\log(price)$, lorsque $lotsize = 20{,}000$, $sqrft = 2{,}500$, et $bdrms = 4$. En utilisant les méthodes de la section 6.4, trouvez la valeur prédite du prix pour les mêmes valeurs des variables explicatives. les mêmes valeurs des variables explicatives.
3. Pour expliquer la variation du prix, décidez si vous préférez le modèle de la partie 1. ou le modèle

4.1.5 vote1

1. Considérons un modèle avec une interaction entre les dépenses :

$$voteA = \beta_0 + \beta_1 prtystrA + b2expendA + \beta_3 expendB + \beta_4 expendAexpendB + u$$

Quel est l'effet partiel de $expendB$ sur $voteA$, en maintenant fixes $prtystrA$ et $expendA$? Quel est l'effet partiel de $expendA$ sur $voteA$? Le signe attendu pour β_4 est-il évident ?

2. Estimez l'équation de la partie 1. et présentez les résultats sous la forme habituelle. Le terme d'interaction est-il statistiquement significatif ?
3. Trouver la moyenne de $expendA$ dans l'échantillon. Fixer $expendA$ à 300 (pour 300 000 \$). Quel est l'effet estimé de 100 000 dollars supplémentaires dépensés par le candidat B sur le $voteA$? S'agit-il d'un effet important ?
4. Fixez maintenant $expendB$ à 100. Quel est l'effet estimé de $\Delta expendA = 100$ sur le $voteA$? Cela a-t-il un sens ?
5. Estimez maintenant un modèle qui remplace l'interaction par $shareA$, la part en pourcentage du candidat A dans les dépenses totales de la campagne. Est-il judicieux de maintenir les $expendA$ et les $expendB$ fixes, tout en modifiant la $shareA$?
6. (Nécessite des calculs) Dans le modèle de la partie 5. , trouvez l'effet partiel de $expendB$ sur $voteA$, en maintenant $prtystrA$ et $expendA$ fixes. Evaluez ceci à $expendA = 300$ et $expendB = 0$ et commentez les résultats.

4.1.6 hprice1

1. Estimez le modèle

$$prix = \beta_0 + \beta_1 lotsize + b2sqrft + \beta_3 bdrms + u$$

et présentez les résultats sous la forme habituelle, y compris l'erreur standard de la régression. Obtenez le prix prédit, lorsque vous introduisez $lotsize = 10{,}000$, $sqrft = 2{,}300$, et $bdrms = 4$; arrondissez ce prix au dollar le plus proche.

2. Effectuez une régression qui vous permette d'obtenir un intervalle de confiance de 95 % autour de la valeur prédite dans la partie 1. . Notez que votre prédiction sera légèrement différente en raison des erreurs d'arrondi.
3. Soit $price_0$ le prix de vente futur inconnu de la maison avec les caractéristiques utilisées dans les parties 1. et 2. . Trouvez un intervalle de confiance à 95% pour $price_0$ et commentez la largeur de cet intervalle de confiance.

4.1.7 nbsal

L'ensemble de données `nbsal` contient des informations sur les salaires et des statistiques de carrière pour 269 joueurs de la National Basketball Association (NBA).

1. Estimez un modèle reliant les points par match (*points*) aux années passées dans la ligue (*exper*), l'âge et le nombre d'années passées à l'université (*coll*). Inclure une quadratique dans *exper* ; les autres variables doivent apparaître sous forme de niveau. variables doivent apparaître sous forme de niveau. Rapportez les résultats de la manière habituelle.
2. En gardant les années d'université et l'âge fixes, à partir de quelle valeur d'expérience l'année d'expérience suivante réduit-elle réellement le nombre de points par match ? Cela a-t-il un sens ?
3. Pourquoi pensez-vous que *coll* a un coefficient négatif et statistiquement significatif ? (Indice : les joueurs de la NBA peuvent être recrutés avant la fin de leur carrière universitaire et même directement à la sortie du lycée).
4. Ajoutez une quadratique de l'âge à l'équation. Est-ce nécessaire ? Qu'est-ce que cela semble impliquer sur les effets de l'âge, une fois que l'expérience et l'éducation sont contrôlées ?
5. Régressez maintenant $\log(wage)$ sur *points*, *exper*, *exper* , *age* et *coll*.
6. Testez si l'âge et le col sont conjointement significatifs dans la régression de la partie 5. Qu'est-ce que cela implique quant à la question de savoir si l'âge et l'éducation ont des effets distincts sur le salaire, une fois que la productivité et l'ancienneté sont prises en compte ?

4.1.8 bwght2

1. Estimez le modèle :

$$log(bwght) = \beta_0 + \beta_1 npvis + b2npvis2 + u$$

par la méthode des moindres carrés ordinaires (MCO), et rapportez les résultats de la manière habituelle. Le terme quadratique est-il significatif ?

2. Montrez que, sur la base de l'équation de la partie 1. , le nombre de visites prénatales qui maximise le $\log(bwght)$ est estimé à environ 22. Combien de femmes ont eu au moins 22 visites prénatales dans l'échantillon ?
3. Est-il logique de prédire que le poids de naissance diminue après 22 visites prénatales ? Expliquez pourquoi.

4. Ajoutez l'âge de la mère à l'équation, en utilisant une forme fonctionnelle quadratique. En maintenant *npvis* fixe, à quel âge de la mère le poids de naissance de l'enfant est-il maximisé ? Quelle fraction des femmes de l'échantillon est plus âgée que la moyenne ? Quelle fraction des femmes de l'échantillon est plus âgée que l'âge "optimal" ?
5. Diriez-vous que l'âge de la mère et le nombre de visites prénatales expliquent une grande partie de la variation du log(*bwght*) ?
6. En utilisant des quadratiques pour le *npvis* et l'âge, décidez si l'utilisation du logarithme naturel ou du niveau de *bwght* est meilleure que l'utilisation du logarithme de l'âge.

4.1.9 apple

1. Exécutez la régression *ecolbs* sur *ecoprc* et *regprc* et reportez les résultats. Présentez les résultats sous la forme d'un tableau, y compris le R-carré et le R-carré ajusté. Interprétez les coefficients sur les variables de prix et commentez leurs signes. les variables de prix et commentez leurs signes et leurs amplitudes.
2. Les variables de prix sont-elles statistiquement significatives ? Indiquez les valeurs p pour les tests t individuels. pour les différents tests t.
3. Quel est l'éventail des valeurs ajustées pour *ecolbs* ? Quelle fraction de l'échantillon rapporte $ecolbs = 0$? Commenter.
4. Pensez-vous que l'ensemble des variables de prix explique bien la variation de l'indice *ecolbs* ? Expliquez.
5. Ajoutez les variables $faminc$, *hhsize* (taille du ménage), *educ*, et *age* à la régression de la partie 1. . Trouvez la valeur p pour leur signification conjointe. Qu'en concluez-vous ?
6. Effectuez des régressions simples séparées de *ecolbs* sur *ecoprc*, puis de *ecolbs* sur *regprc*. Comment les coefficients de régression simple se comparent-ils à la régression multiple de la partie 1. ? Trouvez le coefficient de corrélation entre *ecoprc* et *regprc* pour expliquer vos résultats.

5 Tests sur l'hétéroscédasticité

#TP 5 : Heteroscedasticité

%Le but de cette partie est de recoder les fonctions bptest du package `lmtest` et les données de `CASchools`

On considère le modèle linéaire suivant :

$$Y = X'\beta + u$$

Le but de cette section est de recoder une fonction effectuant des tests de Breush-Pagan et de White.

1. La procédure pour ces tests est la suivante :

   1.1. On effectue une régression MCO classique

   1.2 On isole les résidus $\hat{u}$ et applique une régression sur la variance des résidus :

   — Pour le test de Breush-Pagan, on estime le modèle par MCO : $\hat{u}^2 = X'\gamma + \epsilon$
   — Pour le test de White simplifié, on estime le modèle par MCO : $\hat{u}^2 = \gamma_1 \hat{Y} + \gamma_2 \hat{Y}^2 + \epsilon$
   — On effectue un test de Fisher sur la 2e régression. La fonction reportera la p-valeur de ce test.

6 Erreurs standards robustes

1. Rappelez la formule générale des erreurs standars robutes à l'hétéroscédasticité de White.
2. Codez une fonction qui prend en argument un modèle issu de `lm` et renvoie les erreurs standards robustes.
3. Pour calculer les erreurs standards dans R, il faut charger le package `sandwich` et utiliser la fonction `vcovHC`. On peut afficher l'un modèle issu de *lm* dans stargazer avec le code suivant :

7 Moindres carrés quasi-généralisés

On considère le modèle linéaire dans lequel l'hypothèse d'homoscedasticité est levée. On peut utiliser les données `vote1` pour cette partie.

1. Estimer le modèle linéaire $Y = X'\beta + u$ par MCO.
2. Estimer le modèle linéaire $\log(\hat{u}^2) = X'\gamma$
3. Récupérer les valeurs prédites $\hat{g}$ de ce modèle.
4. Estimez l'équation $Y = X'\beta + u$ par moindre carrés pondéré avec les poids $1/exp(\hat{g}/2)$
5. Effectuez une boucle qui répète les étapes 2. à 4. jusqu'à que les écarts entre $1/exp(\hat{g}/2)$ converge vers 10^{-7}.

8 Exercices

Ces exercices utilisent les données issus du package wooldridge. Installez-le puis lancez la commande `library(wooldridge)`.

8.1 sleep75

Considérez le modèle suivant pour expliquer le comportement de sommeil :

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u$$

1. Rédigez un modèle qui permet à la variance de u de différer entre les hommes et les femmes. La variance ne doit pas dépendre d'autres facteurs.
2. Utilisez les données de `sleep75` pour estimer les paramètres du modèle pour l'hétéroscédasticité. (Vous devez d'abord estimer l'équation du sommeil par les MCO pour obtenir les résidus des MCO). La variance estimée de u est-elle plus élevée pour les hommes ou pour les femmes ?
3. La variance de u est-elle statistiquement différente pour les hommes et pour les femmes ?

8.2 vote1

1. Estimez un modèle avec `voteA` comme variable dépendante et $prtystrA$, $democA$, $log(expendA)$, et $log(expendB)$ comme variables indépendantes. Obtenez les résidus des MCO, $\hat{u}_i$, et régressez-les sur toutes les variables indépendantes. Expliquez pourquoi vous obtenez $R^2 = 0$.
2. Calculez maintenant le test de Breusch-Pagan pour l'hétéroscédasticité. Relevez la valeur de la statistique F et indiquez la valeur p.
3. Calculez le cas particulier du test de White pour l'hétéroscédasticité, en relevant à nouveau la valeur de la statistique F. Quelle est la force de la preuve de l'hétéroscédasticité maintenant ?

8.3 pntsprd

1. La variable *sprdcvr* est une variable binaire égale à un si l'écart de points de Las Vegas pour un match de basket universitaire a été couvert. La valeur attendue de *sprdcvr*, soit μ, est la probabilité que l'écart soit couvert lors d'un match choisi au hasard. Testez H0 : $\mu = 0.5$ contre H1 : $\mu \neq .5$ au niveau de signification de 10 % et discutez vos résultats. (Conseil : ceci est facilement réalisable à l'aide d'un test de student en régressant *sprdcvr* sur une ordonnée à l'origine uniquement).
2. Combien de matchs dans l'échantillon de 553 ont été joués sur un terrain neutre ?
3. Estimer le modèle de probabilité linéaire

$$sprdcvr = \beta_0 + \beta_1 favhome + \beta_2 neutral + \beta_3 fav25 + \beta_4 und25 + u$$

et présenter les résultats sous la forme habituelle. (Indiquez les erreurs standard habituelles des MCO et les erreurs standard corrigées de l'hétéroscédasticité). Quelle variable est la plus significative, à la fois d'un point de vue pratique et statistique ?

4. Expliquez pourquoi, sous l'hypothèse nulle H0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, il n'y a pas d'hétéroscédasticité dans le modèle.
5. Utilisez la statistique F habituelle pour tester l'hypothèse de la partie iv). Quelle est votre conclusion ?
6. Compte tenu de l'analyse précédente, diriez-vous qu'il est possible de prédire systématiquement si l'écart de Las Vegas sera couvert en utilisant les informations disponibles avant le match ?

8.4 loanapp

1. Estimez l'équation avec *approve* comme variable d'intérêt et comme variables explicatives : *white* , *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat*1, *mortlat*2, et *vr*.

Calculer les erreurs standard hétéroscédastiques robustes. Comparez l'intervalle de confiance à 95% sur β_{white} avec l'intervalle de confiance non robuste.

2. Obtenez les valeurs ajustées de la régression de la partie 1. Certaines d'entre elles sont-elles inférieures à zéro ? Certaines d'entre elles sont-elles supérieures à un ? Qu'est-ce que cela signifie pour l'application des moindres carrés pondérés ?

8.5 gpa1

1. Utilisez les MCO pour estimer un modèle reliant *colgpa* à *hsGPA*, *ACT*, *skipped* et *PC*. Résidus des MCO.
2. Calculez le cas particulier du test de White pour l'hétéroscédasticité. Dans la régression de $\hat{u}_i^2$ sur $\widehat{colgpa}_i$, $\widehat{colgpa}_i^2$ obtenez les valeurs ajustées $\hat{h}_i$.
3. Vérifiez que les valeurs ajustées de la partie 2. sont toutes strictement positives. Ensuite, obtenez les estimations des moindres carrés pondérés en utilisant les poids $1/\hat{h}_i$. Comparez les estimations des moindres carrés pondérés pour l'effet de l'absence de cours et l'effet de la possession d'un ordinateur avec les estimations des MCO correspondantes. Quelle est leur signification statistique ?
4. Dans l'estimation WLS de la partie 3, obtenez des erreurs standard corrigées de l'hétéroscédasticité. En d'autres termes, tenir compte du fait que la fonction de variance estimée dans la partie 2. puisse être mal spécifiée. Les erreurs standard changent-elles beaucoup par rapport à la partie 3. ?

8.6 meap00

1. Estimez le modèle :

$$math4 = \beta_0 + \beta_1 lunch + \beta_2 \log(enroll) + \beta_3 \log(exppp) + u$$

par MCO et obtenez les erreurs standard habituelles et les erreurs standard entièrement robustes. Comment se comparent-elles en général ?

2. Appliquez le cas particulier du test de White pour l'hétéroscédasticité. Quelle est la valeur du test F ? Qu'en concluez-vous ?
3. Obtenez $\hat{g}_i$ comme valeurs ajustées de la régression $log(\hat{u}_i^2)$ sur $\widehat{math}4_i$, $\widehat{math}4_i^2$ où $\widehat{math}4_i$ sont les valeurs ajustées OLS et $\hat{u}_i$ sont les résidus OLS. Soit $h_i = exp(g_i)$. Utilisez les $\hat{h}_i$ pour obtenir les estimations des MCO. Y a-t-il de grandes différences avec les coefficients des MCO ?
4. Obtenez les erreurs standard pour les MCO qui permettent une mauvaise spécification de la fonction de variance. Sont-elles très différentes des erreurs standard habituelles des MCO ?
5. Pour estimer l'effet des dépenses en $math4$, les MCO ou les MCO semblent-ils plus précis ?

8.7 fertil2

Utilisez les données de `fertil2` pour répondre à cette question.

1. Estimez le modèle

$$enfants = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educ + \beta_4 electric + \beta_5 urban + u$$

et présentez les erreurs standard habituelles et corrigées de l'hétéroscédasticité. Les erreurs standard robustes sont-elles toujours plus grandes que les erreurs standard non robustes ? Les erreurs standard robustes sont-elles toujours plus grandes que les erreurs standard non robustes ?

2. Ajoutez les trois variables muettes religieuses et testez si elles sont conjointement significatives. significatives. Quelles sont les valeurs p pour les tests non robustes et robustes ?
3. A partir de la régression de la partie 2. , obtenez les valeurs ajustées $\hat{y}$ et les résidus, $\hat{u}$. Régressez $\hat{u}^2$ sur $\hat{y}$ et $\hat{y}^2$ et testez la signification conjointe des deux régresseurs. Concluez que l'hétéroscédasticité est présente dans l'équation pour les enfants.
4. Diriez-vous que l'hétéroscédasticité que vous avez trouvée dans la partie 3. est pratiquement importante dans la pratique ?

8.8 beauty

Utilisez les données de `beauty` pour cette question. 1. En utilisant les données regroupées pour les hommes et les femmes, estimez l'équation

$$lwage = \beta_0 + \beta_1 belavg + \beta_2 abvavg + \beta_3 female + \beta_4 educ + \beta_5 exper + \beta_6 exper^2 + u$$

et présentez les résultats en utilisant des erreurs standard corrigées de l'hétéroscédasticité sous les coefficients. Certains des coefficients sont-ils surprenants, que ce soit par leur signe ou par leur ampleur ? Le coefficient de la femme est-il pratiquement grand et statistiquement significatif ?

2. Ajoutez les interactions de la variable "femme" avec toutes les autres variables explicatives dans l'équation de la partie 1. (cinq interactions en tout). Calculer le test F habituel de signification conjointe des cinq interactions et une version corrigée de l'hétéroscédasticité. L'utilisation de la version corrigée de l'hétéroscédasticité modifie-t-elle le résultat de manière importante ?
3. Dans le modèle complet avec interactions, déterminer si celles impliquant les variables de look - femme - *belavg* et femme - *abvavg* - sont conjointement significatives. Leurs coefficients sont-ils pratiquement faibles ?

9 Estimateurs pour les données de panel

#TP 6 : Econométrie de Panel

1. Chargez le package `plm` et chargez le jeu de données `Grunfeld`.
2. Estimez le modèle `inv ~ value + capital` en OLS simple avec la fonction `lm`.
3. Modifier `Grunfeld` et créer les moyennes temporelles pour chaque firme X_{mean} pour chaque variable X, puis créer les variables $X_{fe} = X - X_{mean}$. Estimez le modèle $inv_{mean}\ value_{mean} + capital_{mean}$.
4. Modifier `Grunfeld` et créer les valeurs retardée temporelles pour chaque firme X_{lag} pour chaque variable X, puis créer les variables de différences $X_{fd} = X - X_{lag}$. Estimez le modèle $inv_{fe}\ value_{fe} + capital_{fe}$ avec `lm`. Attention : il faut fixer la constante à 0.
5. Vérifiez les résultats précédents avec la fonction `plm` avec les arguments `model=pooling`, `model=within`, `model=fd`.
6. Même chose avec la fonction `fixest`.

9.1 Exercices

9.1.1 fertil1

i. Estimez le modèle OLS avec `kids` comme variable dépendente : `lm(data=fertil1, kids ~ educ + age + agesq + black + east + northcen + west + farm + othrural + town + y74 + y76 + y78 + y80 + y82 + y84)`
ii. Dans l'équation estimée précédemment, testez si le cadre de vie à l'âge de 16 ans a un effet sur la fertilité (`town`). (Le groupe de référence est la grande ville). Reportez la valeur de la statistique F et sa p-valeur.
iii. Testez si la région (`east`,`northcen`, `west`) du pays dans laquelle on vit à l'âge de 16 ans (le groupe de référence est celui du Sud) a un effet sur la fertilité.
iv. Soit u le terme d'erreur de l'équation de la population. Supposez que la variance de u change au cours du temps (mais pas avec educ, age, etc.). Un modèle tenant compte de cela peut s'écrire ainsi :

$$u^2 = \gamma_0 + \gamma_1 y74 + \gamma_2 y76 + ... + \gamma_6 y84 + v$$

En utilisant ce modèle, testez l'hétéroscédasticité de u. (Indice : votre test F doit avoir 6 000 et 1 122 degrés de liberté).

v. Ajoutez les termes d'interactione entre `educ` et les variables binaires sur les années au modèle estimé dans la question i. Expliquez ce que représentent ces termes. Sont-ils conjointement significatifs ?

9.1.2 cps78_85

i. Estimez le modèle suivant :

$$lwage = y85 + educ + y85 \times educ + exper + exper^2 + union + female + y85 \times female$$

ii. Comment interprétez-vous le coefficient de `y85` dans le modèle ? A-t-il une interprétation intéressante ? (Soyez prudent ici ; vous devez expliquer les termes d'interactions y85 · educ et y85 · female.) iIi. En maintenant fixés les autres facteurs, quel est le pourcentage estimé d'augmentation du salaire nominal d'un homme ayant fait 12 ans d'études ? Proposez une régression permettant d'obtenir un intervalle de confiance pour cette estimation. [Indice : pour obtenir l'intervalle de confiance, remplacez y85 · educ par y85(educ – 12)
iii. Faites une nouvelle estimation du modèle en exprimant tous les salaires en dollars de 1978. En particulier, définissez le salaire réel comme rwage = wage pour 1978 et comme rwage = wage/1,65 pour 1985. Utilisez maintenant `log(rwage)` à la place de `log(wage)`. Quels sont les coefficients qui changent au modèle i. ?

9.1.3 kielmc

Pour cet exercice, utilisez les données du fichier kielmc i. La variable dist est la distance entre chaque maison et le site de l'incinérateur, en pieds. Examinez le modèle suivant :

$$log(price) = \beta_0 + \delta_0 y81 + \beta_1 log(dist) + \delta_1 y81 \cdot log(dist) + u.$$

Si la construction de l'incinérateur réduit la valeur des maisons proches du site, quel est le signe attendu de δ_1 ? Qu'est-ce que cela signifie si $\beta_1 > 0$?

ii. Estimez le modèle de la question (i) et reportez les résultats sous la forme habituelle. Interprétez le coefficient de y81 · log(dist). Que pouvez-vous en conclure ?
iii. Ajoutez `age`, `age2`, `rooms`, `baths`, `log(intst)`, `log(land)`, et`log(area)` à l'équation. Maintenant, que pouvez-vous conclure concernant l'effet de la présence de l'incinérateur sur la valeur des logements ?
iv. Pourquoi le coefficient de log(dist) est-il positif et statistiquement significatif à la question (ii) mais pas à la question (iii) ? Qu'est-ce que cela nous apprend sur les variables de contrôle introduites à la question (iii) ?

9.1.4 injury

i. En utilisant les données du Kentucky, estimez le modèle

$$log(durat) = afchnge + highearn + afchnge * highearn + male + married$$

et y ajouter les variables explicatives jeu entier de variables indicatrices pour les industries et les types d'accidents du travail (les transformer en facteurs au préalables). Comment l'estimation de `afchnge.highearn` varie lorsque ces autres facteurs sont pris en compte ? Est-ce que cette estimation est statistiquement significative ? ii. Comment interprétez-vous les faibles valeurs de R carré de la question (i) ? Cela signifie-t-il que l'équation est inutile ? iii. Estimez le modèle de i. en utilisant les données pour le Michigan. Comparez les estimations sur le terme d'interaction pour le Michigan et le Kentucky. Est-ce que l'estimation pour le Michigan est statistiquement significative ? Justifiez.

9.1.5 rental

Les données pour les années 1980 et 1990 incluent le prix des loyers ainsi que d'autres variables caractérisant les villes dans lesquelles sont établies des universités. L'idée est de voir si la présence d'un nombre élevé d'étudiants fait augmenter les loyers. Le modèle à effets non observés est :

$$log(rentit) = y90t + \log(popit) + log(avgincit) + pctstuit + a_i + u_{it}$$

, où `pop` est la population de la ville, `avginc` est le revenu moyen, et `pctstu` est le nombre d'étudiants exprimé en pourcentage de la population de la ville (au cours de l'année scolaire). i. Estimez l'équation par la méthode des MCO sur données empilées et reportez les résultats. Que pouvez-vous conclure ?

ii. Les écarts-types estimés que vous avez reportés dans la question (i) sont-ils valides ? Justifiez.
iii. Maintenant, exprimez l'équation en différences et estimez-la par les MCO. Comparez votre estimation de β_3 avec celle de la question (ii). Est-ce que la taille relative de la population d'étudiants semble affecter les prix des loyers ?
iv. Obtenez les écarts-types estimés robustes à l'hétéroscédasticité pour l'équation en différences premières de la question (iii). Cela modifie-t-il vos conclusions ?

9.1.6 jtrain

Pour cet exercice, nous utilisons les données contenues dans la base JTRAIN pour étudier l'effet des subventions à la formation professionnelle sur la formation professionnelle par employé. Le modèle de base pour les trois années est donné par :

$$hrsempit = \beta_0 + \delta_1 d88_t + \delta_2 d89_t + \beta_1 grant_{it} + \beta_2 grant_{i,t-1} + \beta_3 log(employ_{it}) + a_i + u_{it}$$

i. Estimez l'équation en utilisant des effets fixes. Combien de firmes sont utilisées dans cette estimation ? Combien d'observations seraient utilisées si chacune des entreprises disposait de données pour l'ensemble des variables explicatives du modèle (et en particulier, `hrsemp`) pour chacune des trois années ?
ii. Interprétez le coefficient relatif à grant et commentez sa significativité.
iii. Cela vous surprend-il que le coefficient associé à grant–1 ne soit pas significatif ? Justifiez.
iv. Les grandes entreprises proposent-elles à leurs employés plus ou moins de temps de formation en moyenne ? De quel ordre sont les différences ? (Par exemple, si une entreprise a 10 % d'employés en plus, quel est le changement relatif dans le nombre d'heures allouées à la formation professionnelle ?)

9.2 murder

Nous utilisons les données étatiques relatives aux taux de criminalité et d'exécution issues du fichier MURDER pour l'exercice qui suit.

i. Considérons le modèle à effets inobservés suivant :

$$mrdrte_{it} = \eta_t + \beta_1 exec_{it} + \beta_2 unem_{it} + a_i + u_{it}$$

avec η_t les différentes variables indicatrices temporelles et ai l'effet inobservé relatif à l'appartenance à un État donné. Si les exécutions passées de personnes jugées coupables de meurtres avaient eu un effet dissuasif, quel aurait dû être le signe de β_1 ? À votre avis, quel devrait être le signe de β_2 ? Justifiez.
ii. En restreignant l'étude aux années 1990 et 1993, estimez l'équation considérée à la question (i) au moyen d'une estimation par les MCO sur les données empilées. Ignorez le problème de corrélation sérielle dans le terme d'erreur composé. Les résultats d'estimation sont-ils en faveur de l'hypothèse d'un effet dissuasif de la peine capitale ?
iii. En vous focalisant toujours sur les années 1990 et 1993, estimez maintenant le modèle au moyen d'effets fixes. Vous pouvez recourir aux différences premières puisque vous n'utilisez que deux années de données. Vos résultats sont-ils maintenant en faveur de l'hypothèse d'un effet dissuasif de la peine capitale ? Ce résultat est-il robuste ?
iv. Calculez les écarts-types estimés robustes à l'hétéroscédasticité des paramètres estimés du modèle décrit en question (ii).
v. Identifiez l'État qui présente le nombre le plus important d'exécutions en 1993. (La variable exec correspond au nombre total d'exécutions réalisées en 1991, 1992, et 1993.) Identifiez le deuxième État en matière d'exécutions en 1993. Combien d'exécutions les séparent ?

vi. Estimez l'équation par la méthode des différences premières, en retirant l'État du Texas de votre ana- lyse. Calculez les écarts-types standard et ceux robustes à l'hétéroscédasticité. Que trouvez-vous ? Que se passe-t-il ?
vii. Utilisez maintenant les trois années de données dont vous disposez et estimez le modèle à effets fixes. Introduisez l'État du Texas dans votre analyse. Discutez la taille et la significativité de l'effet dissuasif de la peine capitale comparativement aux effets mis en exergue dans les questions précédentes lorsque les seules années 1990 et 1993 étaient considérées.