Consignes

- *Vos calculs et graphiques doivent être réalisés avec R ou Python.*
- *Vous pouvez effectuer ce devoir en groupe.*
- *Vous remettrez un ou des fichiers, selon le logiciel choisi pour traiter les questions.*
  - *Si vous utilisez R ou Python pour effectuer les calculs et un logiciel de composition (Word, Libreoffice, Latex), vous remettrez (1) un unique fichier* `.pdf` *contenant vos réponses et (2) un unique fichier* `.R` *ou* `.py` *contenant le code R ou Python que vous avez utilisé pour effectuer vos calculs.*
  - *Si vous utilisez un environnement de programmation avec sortie enrichie comme RStudio ou JupyterLab, vous remettrez un unique fichier au format R notebook ou IPython notebook ; ce fichier doit contenir vos réponses et le code R ou Python que vous avez utilisé pour effectuer vos calculs.*

  Dans les deux cas, le code doit pouvoir s'exécuter sans erreur sur mon ordinateur (hormis la lecture des données). Par ailleurs, le rendu des équations dans votre document doit être de qualité professionnelle (comme dans l'énoncé de ce devoir, par exemple).
- *Si vous faites le devoir en groupe, le fichier contenant vos réponses doit indiquer le nom de tous les étudiants du groupe. Tous les étudiants d'un même groupe auront la même note.*
- *Le nommage des fichiers remis doit respecter le format suivant :* `VOTRENOM_votreprenom_DM3.ext`, où `ext` $\in$ {`R, py, ipynb, Rmd, pdf`} selon le type du fichier. Si vous faites le devoir en groupe, les nom et prénom utilisés sont ceux d'un des étudiants du groupe.
- Vous déposerez vos fichiers sur la page Moodle du cours (rubrique Exercices et devoirs), avant le 10/11/2025, 23h59.

Questions

Exercice 1. On dispose d'observations expérimentales concernant la consommation de café et d'alcool pour des individus, ainsi que de leur temps de réaction face à un évènement donné. Ces observations sont dans le fichier `temps_reaction.csv` téléchargeable depuis la page Moodle du cours (rubrique "Jeux de données"). La description dans l'ordre des colonnes est la suivante :

- identifiant de l'individu
- 1 si l'individu n'a consommé ni alcool ni café, 0 sinon
- 1 si l'individu a consommé de l'alcool mais pas de café, 0 sinon
- 1 si l'individu a consommé du café mais pas d'alcool, 0 sinon
- 1 si l'individu a consommé de l'alcool et du café, 0 sinon
- temps de réaction

1. Formulez un modèle de régression linéaire destiné à expliquer le temps de réaction par la consommation d'alcool et de café. Proposez une formulation dans laquelle une constante est incluse et une formulation sans constante. Interprétez les coefficients de chaque variable dans chacune des deux formulations.

2. Estimez chacun des deux modèles et commentez vos résultats. Établissez des liens entre les estimations des paramètres dans un modèle et dans l'autre.
3. Pourquoi les coefficients de détermination des deux modèles sont-ils différents ? Dans chacun des deux modèles, expliquez ce qu'il faudrait faire pour pouvoir interpréter ce coefficient correctement.
4. Si on considère le modèle sans constante, exprimez la matrice $\mathbf{X}$ utilisée dans l'expression de l'estimateur MC des coefficients. En effectuant un calcul à la main, démontrez que les estimations des coefficients obtenues dans le modèle sans constante doivent nécessairement coïncider avec des moyennes empiriques. Commentez ce résultat.
5. Dans chacune des deux formulations (avec constante, sans constante), testez l'hypothèse que les consommations de café ou d'alcool n'ont aucun effet sur le temps de réaction.

Exercice 2. Dans le fichier `cheddar.csv` disponible sur Moodle (rubrique "Jeux de données"), on trouve dans les données concernant la qualité gustative de cheddar australien. Les variables sont

— `ID` : identifiant du cheddar
— `TASTE` : note moyenne de dégustation
— `ACETIC` : teneur en acide acétique
— `H2S` : teneur en hydrogène sulfuré
— `LACTIC` : teneur en acide lactique
— `ORG` : 1 si le cheddar est certifié BIO ; 0 sinon

On souhaite étudier le lien entre la qualité gustative d'un cheddar d'une part, et ses propriétés chimiques et son mode de production, d'autre part.

1. Si pour cela on souhaite utiliser un modèle de régression linéaire, écrivez le modèle adéquat.
2. Dans ce modèle, décrivez en détail comment on peut mesurer et décrire l'effet du mode de production (BIO ou pas) sur la variable dépendante.
3. À l'aide de R,
   (a) Estimez les paramètres de ce modèle et interprétez les résultats ;
   (b) quelle est l'estimation de la variance (conditionnelle) de la variable dépendante ?
   (c) que veut dire "F-statistic" ? Comment interpréter sa valeur ?
4. Estimez par moindres carrés un modèle dans lequel on a retiré les variables explicatives dont les coefficients ne sont pas significativement différents de 0. On constate que le coefficient de détermination est plus faible dans l'estimation de ce modèle que dans le modèle initial. Comment pouvez-vous expliquer cela ? Dites notamment si c'est une constatation propre au jeu de données utilisées ou si c'est une propriété qu'on pourrait constater sur n'importe quel jeu de données. Soyez très précis et rigoureux dans vos explications.
5. Estimez par moindres carrés le modèle avec toutes les variables explicatives, mais sans terme constant. Que constatez-vous à propos du coefficient de détermination ? Comment interpréter cela ?
6. Dans le modèle avec constante et avec toutes les variables,
   (a) pour chaque variable explicative, en effectuant un test adéquat au niveau 2%, dites si ses variations provoquent des variations de la variable dépendante ;
   (b) testez au niveau 1% l'hypothèse qu'une augmentation de la teneur en acide acétique a exactement le même effet sur la qualité gustative qu'une baisse identique de la teneur en hydrogène sulfuré.
   (c) Avec le même niveau, testez que le carré de la somme des coefficients de variables chimiques (`ACETIC`, `LACTIC`, `H2S`)) est égale à 10 fois le coefficient de la variable `ORG`.