## TP n° 4 : (TP Noté) Moindre carrés

Pour ce travail vous devez déposer un **unique** fichier sur le moodle du cours HMLA408. Si votre numéro de groupe (`nb_gr`) vaut `nb_gr=E` vous rendrez un fichier d'extension `ext=R`; si votre numéro de groupe vaut `nb_gr=C` ou `nb_gr=D`, vous rendrez un unique fichier d'extension `ext=ipynb`. Pour faciliter la correction, le nom du fichier sera de la forme suivante : `nom_fichier = tp_note_hmla408_gr_nb_gr_prenom_nom.ext`, le tout en minuscule et sans accent ni espace. Vous remplirez votre nom, prénom, le numéro de groupe qui vous concerne (remplacer `nb_gr` par C, D ou E) et l'extension utilisée (remplacer `ext` par `R` ou `ipynb`) de manière adéquate. Un point de malus sera appliqué pour les fichiers dont le nom est erroné.

Vous devez charger votre fichier sur Moodle, avant le vendredi 19/04/2019, 23h55. La note totale est sur **20** points, répartis comme suit :

- qualité des réponses aux questions : **14** pts,
- qualité de rédaction et d'orthographe : **1** pt,
- qualité des graphiques (légendes, couleurs) : **2** pts
- qualité d'écriture du code (noms de variable clairs, commentaires, code synthétique, etc.) : **1** pt
- Rendu reproductible et absence de bug : le code doit s'exécuter sur la machine du correcteur sans manipulation de sa part (par exemple le correcteur n'est pas supposé aller chercher les fichiers sur internet, les enregistrer, etc.). On veillera donc à ce que le chargement des bases de données soit automatisé : **2** pts

Les personnes qui n'auront pas soumis leur devoir sur Moodle avant la limite obtiendront **zéro**.

**Rappel : aucun travail par mail ne sera accepté !**

### Exercice 1. Arbres : taille et volume (5.5pt)

Un étudiant en techniques forestières veut utiliser la régression linéaire pour estimer le volume en bois utilisable d'un arbre debout en fonction de l'aire du tronc mesurée à 25 $cm$ du sol. Il a choisi au hasard 10 arbres et a mesuré, à la base, l'aire correspondante (en $cm^2$). Il a par la suite enregistré, une fois l'arbre coupé, le volume correspondant en $m^3$.

Le fichier `arbres.txt` contient les données. Les variables sont `vol` et `aire` qui représentent respectivement le volume utilisable et l'aire à la base du tronc.

Le but de cet exercice est d'étudier la variable `vol` en fonction de la variable `aire`.

1) (**0.5pt**) Importez automatiquement le jeu de données `arbres.txt` [1] dans une dataframe que vous nommerez `df_arbres`.
2) (**1pt**) Ajustez le modèle linéaire qui explique la variable `vol` (en ordonnée) par la variable `aire` (en abscisse). Donner la valeur de la pente estimée ainsi que celle de l'ordonnée à l'origine. [2]
3) (**1pt**) Représentez le nuage des points du volume (en ordonnée) en fonction de l'aire (en abscisse) ainsi que la droite d'ajustement des moindres carrés. On prendra soin aux légendes, au titre, aux noms des axes, etc.
4) (**0.5pt**) Donnez la proportion de la variance expliquée par le modèle linéaire.
5) (**0.5pt**) Afficher un graphique représentant un estimateur à noyau des résidus studentisés.
6) (**1pt**) Calculez un intervalle de prédiction pour le volume correspondant à une aire de $465cm^2$ pour un niveau de confiance de $1-\alpha$, avec $\alpha$ qui vaut la longueur de la chaîne de caractère `nom_prenom` divisée par 100 (par exemple pour `nom_prenom = Salmon_Joseph`, on trouve $\alpha = 0.13$, soit 13 %).

1. http://josephsalmon.eu/enseignement/datasets/arbres.txt
2. En `Python` on pourra utiliser le package `statsmodel`, *cf.* http://www.statsmodels.org/stable/regression.html, en particulier la formulation de type R permettra de simplifier l'utilisation de cette fonction, voir http://www.statsmodels.org/devel/example_formulas.html. Enfin, en R on pourra utiliser la fonction `lm` *cf.* https://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html

7) (**1pt**) Afficher sur un graphique les données brutes, la droite de régression obtenue par les moindres carrés, la prédiction donnée par le modèle pour la valeur $465cm^2$. Enfin proposera une représentation graphique de l'intervalle de confiance pour le niveau de confiance $\alpha$ calculé à la question précédente.

**Exercice 2. Prairies et rendement agricole** (**1pt**)

Dans cet exercice, on veut faire une analyse de la variance pour vérifier l'influence du type de sol sur le rendement fourrager. On dispose de 30 observations de parcelles de prairie pour lesquelles on a mesuré la variable `rendement` (en tonnes) et on donne la variable `parcelle` qui indique le type de sol (codé par 1, 2 ou 3).

1) (**0.5pt**) Importez le jeu de données `prairie.txt`[3] dans une table nommée `prairie`
   Expliquez quelle est la différence entre les deux variables `parcelle` et `rendement`.
2) (**0.5pt**) Tracer les diagrammes en violon du rendement des parcelles en fonction du type de sol codé par la variable `parcelle` et commenter.
3) (Question Bonus **1pt**) : proposer une analyse de la variance du modèle et tester l'impact si la variable "`parcelle`" a un impact sur le rendement.

**Exercice 3. Impact d'un traitement sur la croissance des plantes** (**2.5pt**) Une expérience de comparaison de deux traitements de plante repose sur l'étude d'un échantillon de 20 plantes : 10 plantes ont été sélectionnées au hasard pour subir le traitement 1 et les 10 plantes restantes ont subi le traitement 2. On va étudier le poids (en grammes) des plantes après traitement.

1) (**0.5pt**) Créez deux vecteurs `ech1` et `ech2` qui contiennent respectivement les données des poids avec le traitement 1 et le traitement 2 :

   ```
   14.4, 14.7, 13.2, 12.1, 18.7, 15.0, 13.3, 17.8, 16.6, 15.0 (traitement 1)
   25.6, 17.7, 19.0, 26.7, 22.6, 19.1, 22.9, 21.0, 25.7, 23.7 (traitement 2)
   ```

2) (**0.5pt**) On souhaite mettre en place un test de l'égalité des moyennes des deux groupes de plantes qui reçoivent le traitement 1 et 2. Quelles hypothèses devez-vous vérifier au préalable avant de faire ce test ? Les vérifier.
3) (**0.5pt**)Donner un intervalle de confiance bilatéral, au niveau de confiance de $1-\alpha$, (avec $\alpha$ qui vaut la longueur de la chaîne de caractère `nom_prenom` divisé par 100) de la différence des moyennes pour les deux traitements.
4) (**1.pt**) Tester l'égalité des moyennes des deux groupes de plantes qui reçoivent le traitement 1 et 2. Optez pour un test unilatéral qui vous semble le plus pertinent. Donnez $\mathcal{H}_0$, $\mathcal{H}_1$ et la $p$-value de ce test. Conclure.

**Exercice 4. Hospitalisation : Répartition des entrées** (**2.5pt**)

Dans cet exercice on souhaite savoir si les entrées à l'hôpital pour une certaine maladie (la maladie **A**) sont réparties au hasard dans l'année ou bien si certains mois sont plus propices à la maladie **A**. On examine le mois d'entrée d'un échantillon de 120 porteurs de la maladie **A**. Les résultats sont contenus dans le fichier `Hospit.csv`[4].

Écrire un code qui vous permettra de répondre à la question suivante : Peut-on affirmer avec un risque $\alpha$ (qui vaut la longueur de la chaîne de caractère `nom_prenom` divisée par 100) que "les entrées pour la maladie **A** ne se font pas au hasard dans l'année" (donc que "certains mois sont plus ou moins propices à la maladie **A**") ?

Vous inclurez des commentaires dans votre code et vous indiquerez clairement $\mathcal{H}_0$, $\mathcal{H}_1$, le test utilisé et la $p$-value de ce test ainsi que la conclusion que vous pouvez en tirer.

**Exercice 5. Pollution en Occitanie** (**2.5pt**) Reprendre la base de données sur les principaux polluants en Occitanie[5].

3. http://josephsalmon.eu/enseignement/datasets/prairie.txt
4. http://josephsalmon.eu/enseignement/datasets/Hospit.csv
5. http://josephsalmon.eu/enseignement/datasets/Mesure_journaliere_Region_Occitanie_Polluants_Principaux.csv

Proposer une comparaison approfondie du niveau de pollution entre Toulouse et Montpellier sur le niveau de NO2. On pourra s'intéresser au niveau de pollution pour certaines saisons seulement, pour certains jours de la semaine/week-end, pour la journée / pour la nuit, etc. Ces comparaisons seront appuyées d'un argumentaire graphique et statistique (test/intervalle de confiance) de votre choix.