TP1
Régression linéaire simple (1) : pollution de l'air

Objectifs du TP

— Prise en main d'un jeu de données (analyse descriptive).
— Mise en œuvre de la régression linéaire simple : estimation ponctuelle par moindres carrés et estimation de la variance résiduelle.
— Visualisation du résultat : droite de régression, graphes des résidus.

Pour ceux qui n'ont jamais fait de `R` (et ceux qui ont tout oublié)

— Un bon tutoriel sur R : `https://www.guru99.com/r-tutorial.html` ;
— Demander exercices supplémentaires au prof !

L'aide en ligne est accessible via la commande `help.start()`. On peut également obtenir de l'aide sur une fonction précise avec la commande `help` (par exemple, taper `help(mean)` pour obtenir de l'aide sur la fonction `mean`).

Vous trouverez sur dokeos des aide-mémoire sur les commandes de base.

Quelques liens utiles à propos de R (cliquables sur le pdf, qui est disponible sur dokeos) :
— R pour les débutants, par Emmanuel Paradis.
— Le livre Statistique avec R, de P.A. Cornillon et al, paru aux Presses Universitaires de Rennes, est une excellente référence pour le cours.
— Le site du CRAN : `https://cran.r-project.org/`

Pollution de l'air

La pollution de l'air est un problème de santé publique majeur. De nombreuses études ont démontré l'influence sur la santé de certains composés chimiques, dont l'ozone (O_3), tout particulièrement sur les personnes sensibles (nouveaux-nés, asthmatiques, personnes âgées). Il est donc important de savoir prévoir les pics de concentration de l'ozone.

On sait que la concentration en ozone varie avec la température : plus la température est élevée, plus la concentration en ozone est importante. Cette relation vague doit être précisée en vue de prévoir les pics d'ozone. Dans ce but, l'association Air Breizh (surveillance de la qualité de l'air en Bretagne) mesure depuis 1994 la concentration en O_3 (en μg/L) toutes les dix minutes et obtient donc le maximum journalier de la concentration en O_3, que l'on note désormais `O3`. Air Breizh collecte également des données météorologiques correspondant à ces mesures d'ozone, dont la température à 12h, que l'on note `T12`.

À partir de ces données (rassemblées dans le fichier '`ozone_simple.txt`'), on cherche à expliquer la concentration maximale `O3` à l'aide de la température `T12`.

1 Préliminaire : analyse descriptive

1. Importer les données '`ozone_simple.txt`'.
2. Calculer la moyenne, la médiane, la variance de la concentration maximale `O3` et de la température `T12`.
3. Tracer l'histogramme de la concentration maximale `O3`, puis celui de la température `T12`.
4. Tracer, sous la forme d'un nuage de point, le graphe de la concentration maximale `O3` en fonction de la température `T12`.

2 Calcul des estimateurs des moindres carrés

5. Écrire un modèle de régression linéaire permettant d'expliquer la concentration maximale `O3` en fonction de la température `T12`.
6. Utiliser `R` pour estimer les paramètres de ce modèle :
   (a) d'abord en utilisant les formules du cours,
   (b) ensuite à l'aide de la fonction `lm`.

   Donner les estimations de l'ordonnée à l'origine, la pente et la variance résiduelle.
7. Tracer la droite de régression (superposée avec le nuage de points représentant les observations).
8. Faire un graphe des résidus. Que constate-t-on ?

Questions subsidiaires

9. Chercher une transformation de la variable `T12` qui permette de mieux prévoir `O3` avec un modèle linéaire.
10. Refaire le TP avec la variable explicative `Vx` (composante est-ouest du vent à 12h), à l'aide des données `ozone_simple_Vx.txt`.
11. Refaire le TP avec la variable explicative `Ne12` (nébulosité observée à 12h), à l'aide des données `ozone_simple_Ne12.txt`.

On poursuivra l'analyse de ce premier exemple lors du TP 2.