TD 3
Modèles multiples

Exercice 1 : Colinéarité statistique

On considère un jeu de données $(X_i^{(1)}, X_i^{(2)}, Y_i)$, $i = 1, \ldots, 10$, défini par le modèle suivant,

$$Y_i = X_i^{(1)} + X_i^{(2)} + \varepsilon_i,$$

où $\varepsilon_i \sim \mathcal{N}(0, 1)$, i.i.d. Nous allons montrer les paradoxes possibles de la significativité partielle et le problème de la colinéarité statistique par la simulation suivante.

1. Générer un vecteur `X1` de taille 10 dont les composantes sont aléatoirement distribuées entre 10 et 20 suivant la loi uniforme.

2. Créer un vecteur `X2 = X1`. Modifier légèrement deux des composantes de `X2`, par exemple arrondir `X2[1]` et `X2[2]` en utilisant la fonction `round`.

3. Générer un vecteur gaussien centré et réduit `e` de taille 10.

4. Créer le vecteur `Y=X1+X2+e`.

5. Estimer les coefficients des modèles `Y~X1+X2`, `Y~X1` et `Y~X2` en utilisant la fonction `lm`. Afficher les résultats avec `summary`.

6. Que pensez-vous de ces régressions, les résultats obtenus sont-ils surprenants ?

Exercice 2

Taper les commandes suivantes :

```
i = c(1:100)
X1 = i; X2 = 5*sin(i/2); X3 = log(i); X4 = log(i*i)
Y = rnorm(100, X1-2*X3-5*X4, 4)
reg = lm(Y~X1+X2)
summary(reg)
```

1. Écrire explicitement le modèle simulé en précisant les différentes variables intervenant.

2. Que pensez-vous de cette régression, ces résultats sont-ils surprenants ?

Taper ensuite les commandes :

```
reg2 = lm(Y~X1+X3)
summary(reg2)
```

3. Que pensez-vous de cette nouvelle régression ?

4. Montrer que l'on pouvait s'attendre à la valeur du coefficient de `X3` qui vaut à peu près -12.

5. Que ce serait-il passé si l'on avait ajouté `X4` dans la régression ?