TP – Régression linéaire multiple

Exercice 1.

On s'intéresse à la base de données `LifeCycleSavings` de `R`.

```
> data(LifeCycleSavings)
```

Cette base contient les taux moyens d'économies (`sr`) de 50 pays entre 1960 et 1970 (source : Belsley, Kuh & Welsch, 1980). Nous souhaitons construire un modèle linéaire explicatif de cette variable en fonction des autres variables à notre disposition.

1. Décrire les différentes variables du paquet. Faire les représentations bivariées des variables.
2. On souhaite construire un modèle de régression du taux d'économie `sr` en fonction des autres variables, en mettant en œuvre une procédure de sélection de variables.
   (a) Appliquer la méthode de sélection de variables descendante à l'aide de la fonction `drop1` puis de la fonction `step`.
   (b) Quel est le modèle final ? On le notera `sr.lm`. Vérifier sa pertinence par rapport au modèle faisant intervenir toutes les variables explicatives.
3. Vérifier les hypothèses du modèle `sr.lm`.
4. Commenter les coefficients obtenus. Les signes semblent-ils cohérents ?

Exercice 2.

On s'intéresse à la base de données `longley` de `R`. Celle-ci contient 7 variables économiques, mesurées annuellement entre 1947 et 1962.

```
> data(longley)
> str(longley)
```

1. Construire un modèle de régression linéaire de la variable `Employment` par rapport aux autres variables. On mettra en œuvre une procédure de sélection de variables ascendante. On pourra pour cela utiliser la fonction `add1`.
2. Mesurer la colinéarité des variables. Vérifier la cohérence du résultat obtenu avec la fonction `cor`.
3. On reprend le modèle complet, sans sélection de variables. Calculer le vecteur de régression $\hat{\beta} = (X^T X)^{-1} X^T Y$ à l'aide de cette formule, en utilisant la fonction `solve`. Comparer avec la valeur donnée par `lm(Employment~.,data=longley)`.

Exercice 3.

On utilise la base `airquality` de R. On souhaite expliquer le taux d'ozone `Ozone` par la vitesse du vent `Wind`, la température `Temp` ou la radiation solaire `Solar.R`.

1. Faire les représentations graphiques bivariées des données. Pourquoi choisit-on de ne pas considérer les variables `month` and `day` comme variables explicatives dans le modèle linéaire ?
2. La commande `table(complete.cases(airquality))` montre la présence de données manquantes. Construire un modèle linéaire gaussien multiple décrivant le taux d'ozone en fonction des 3 autres variables, en retirant les données manquantes.

   ```
   > air.lm <- lm(Ozone ~ Solar.R + Wind + Temp, data=airquality,
   +               na.action=na.exclude)
   ```

   Commenter.

   ```
   > summary(air.lm)
   ```

3. Mettre en place une sélection de variables.
4. Commenter les résultats obtenus.
5. Etudier les résidus du modèle (normalité, homoscédasticité, *etc*).