TP I : Régression linéaire simple

Nous utiliserons le langage `R` pour ce TP. Le corrigé sera à faire sous forme de `.Rmd`.

Exercice 1. Estimateurs des moindres carrés Nous allons traiter les $n = 50$ données journalières de la concentration en ozone en fonction de la température. Les données se trouvent dans le fichier `ozone.txt`. La variable à expliquer est la concentration en ozone, notée `maxO3`, et la variable explicative est la température à midi, notée `T12`.

1. Commencer par représenter les données. On tracera `maxO3` en fonction de `T12`. Une regression linéaire simple semble-t-elle justifiée graphiquement?
2. Effectuer la régression linéaire à l'aide de la commande `lm()` dont on stockera la sortie dans une variable `reg`. Consulter alors les résultats à l'aide de la commande `summary()`. Que représentent les coefficients de la matrice coefficients?
3. Tracer l'estimation de la droite de régression, ainsi qu'un intervalle de confiance à 95% de celle-ci grâce aux commandes suivantes:

```
> plot(maxO3~T12,data=ozone)
> T12 <- seq(min(ozone[,"T12"]), max(ozone[,"T12"]), length=100)
> grille <- data.frame(T12)
> ICdte <- predict(reg, new=grille, interval="confidence", level=0.95)
> matlines(grille$T12, cbind(ICdte), lty=c(1,2,2),col=1)
```

   Décrire précisément ce que font chaque ligne de code. Ce graphique permet de vérifier visuellement l'ajustement des données au modèle de régression proposé. Que remarquez-vous?
4. On s'intéresse à présent à la qualité de prévision du modèle. Pour cela, il faut tracer un intervalle de confiance des prévisions en adaptant le code de la question précédente.
5. On va maintenant calculer les intervalles de confiance des coefficients du modèle de régression. Pour cela, on utilise la fonction `coef()` qui permet d'extraire les estimateurs et leurs écarts types empiriques.

```
> seuil<-qt(0.975, df =reg$df.res)
> beta1min<-coef(resume)[1,1] - seuil * coef(resume)[1,2]
> beta1max<-coef(resume)[1,1] + seuil * coef(resume)[1,2]
> beta2min<-coef(resume)[2,1] - seuil * coef(resume)[2,2]
> beta2max<-coef(resume)[2,1] + seuil * coef(resume)[2,2]
```

   Retrouver ces valeurs à l'aide de la fonction `confint()`. Que remarquez-vous sur l'intervalle de confiance de l'ordonnée à l'origine? Comment l'expliquez-vous?
6. Pour être plus précis et tenir compte de la dépendance entre les deux coefficients, on peut aussi construire une région de confiance. Les commandes suivantes permettent de visualiser la différence entre le rectangle de confiance, simple juxtaposition des deux intervalles de confiance, et la région de confiance. Elles nécessitent l'installation du package `ellipse`.

```
> library(ellipse)
> plot(ellipse(reg, level=0.95), type="l", xlab="beta1", ylab="beta2")
> points(coef(reg)[1], coef(reg)[2], pch=3)
> lines(c(beta1min, beta1min, beta1max, beta1max, beta1min), c(beta2min, beta2max,
↪  beta2max, beta2min, beta2min), lty=2)
```

7. Au vu de la représentation de la concentration d'ozone en fonction de la température, nous souhaitons maintenant modéliser l'ozone par la température au carré. Estimer les paramètres de ce modèle quadratique et le comparer au modèle initial.