****

**Examen :** **Modèles de régression linéaire**

**Réalisé par : OUBAHA ANAS**

**Filière : DSE**

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques comme l'ozone (O3). L'objectif de ce travail est de modéliser la valeur des pics d'ozone (O3) en fonction de grandeurs Physiques comme la température et la nébulosité.

On considère à cet effet un jeu de données (voir ci-dessous) contient 102 observations recueillies à Rennes durant l'été 2001 :

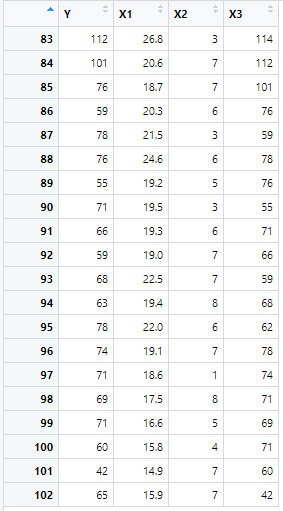
Les 4 variables observées sont :

• Y : Maximum de concentration d'ozone observé sur la journée (Y est la variable à expliquer)

• X1 : Température observée (variable explicative)

• X2 : Nébulosité observée (variable explicative)

• X3 : Teneur maximum en ozone observée la veille (variable explicative)



**Objectif de l’étude :**

Faire toutes les régressions linéaires multiples possibles (avec constante) expliquant la variable Y a l’aide du logiciel R. Puis, Choisir le meilleur modèle parmi ces possibilités en justifiant notre choix.

**Etape 1 : Importer les données**

Pour importer les données enregistre sous format CSV et qu’on va utiliser dans cet étude, on tape le script suivant :

*>> data=read.csv(file.choose(),header = T,sep = ",")*

*>> attach(data)*

*>> view(data)*

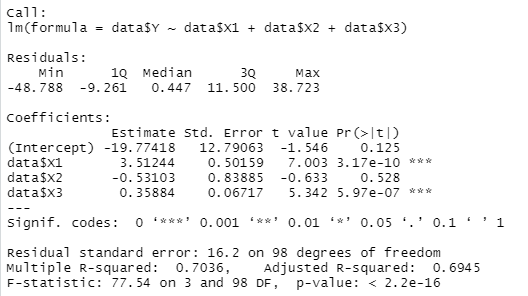
Une aperçue des données a était déjà présentée ci-dessus.

**Etape 2 : Faire une régression avec constante**

Le premier modèle que je vais réaliser nommé « regression1 » est un modèle avec constante. Voici le script correspondant :

*>> regression1<-lm(data$Y∼data$X1+data$X2+data$X3)*

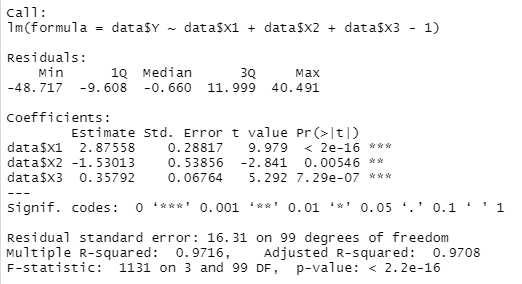
*>> summary(regression1)*

Le résultat de cet régression : **Etape 3 : Faire une régression sans constante**

Le deuxième model « regression2 » concerne un modèle de régression linéaire multiple sans constante.

*>> regression2<-lm(data$Y∼data$X1+data$X2+data$X3-1)*

*>> summary(regression2)*

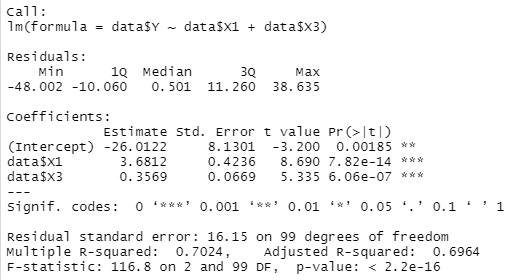
Le résultat de cette régression est le suivant :

**Etape 4 : Choix du modèle le plus pertinent**

Pour choisir le modèle qui représente le plus nos données on va se baser sur plusieurs critères et remarques.

**a- Test de significativité des coefficients :**

Sachant que généralement α utilise dans R est de 5%. Pour le modèle « regression1 » la variable X2 ne vérifie pas la condition P-value ≤ α (car : 0.528 > 0.05) donc on doit accepter l’hypothèse H0. C-à-d, la variable X2 n’est pas significative pour ce modèle.

Si on veut travailler avec le modèle Y = a0 + a1.X1 + a3.X3 et éliminé carrément X2, on aura une significativité globale qui n’est pas meilleur que le modèle « regression1 » même si la significativité individuelle des variables sera bonne comme l’indique la capture suivant :

En revanche, toutes les variables du modèle « regression2 » sont significatives (les P-values ≤ 0.05).

**b- Test du R-carré et du R-carré ajusté :**

-le modèle « regression1 » : R-carré = 0.7036, R-carré ajusté = 0.6945.

-le modèle « regression2 » : R-carré = 0.9716, R-carré ajusté = 0.9708.

On conclut donc que le modèle « regression2 » est plus significatif que le modèle « regression1 », il nous permet d’expliquer plus que 97% de la variance totale des données.

**c- D’autre tests possibles : Le test AIC et BIC**

-le modèle « regression1 » : AIC =863.4802, BIC= 876.6051.

-le modèle « regression2 » : AIC = 861.8965, BIC = 872.3964.

On constat que les indicateurs AIC et BIC pour le modèle « regression2 » sont un peu baisse par rapport a ceux du modèle « regression1 ».

**Conclusion :**

Tous les tests qu’on a faits auparavant citent un modèle qui semble être plus pertinent pour cet étude et c’est le modèle de régression linéaire multiple sans constante. Ce modèle est estimé : Y = 2,8.X1 – 1,5.X2 + 0,3.X3, avec écart-type

σ(a1) = 0.28, σ(a2) = 0.53, σ(a3) = 0.06