TP Statistique en logiciel R Régression linéaire

Nous allons traiter les données journalières de la concentration en ozone en fonction de la température. Les données se trouvent dans le fichier "ozone.txt" que vous pouvez avoir du lien suivant :

<https://r-stat-sc-donnees.github.io/ozone.txt>

La variable à expliquer est la concentration en ozone, notée "maxO3", et la variable explicative est la température à midi, notée "T12".

**Importation de données**

Commencer par extraire les données du fichier "ozone.txt" et les stocker dans une data frame nommée ozone ; Pour cela tester la commande suivante :

**>ozone=read.table("ozone.txt")**

Vous allez rencontrer des problèmes parce que le fichier n’existe pas dans le dossier courant.

Essayer de résoudre ce problème en utilisant les deux fonctions setwd(…)et getwd(). Après avoir manipulé les deux fonctions précédentes, essayer d’intégrer la fonction file.choose().

**> ozone=read.table(file.choose())**

Il faut également gérer certains problèmes lors de l’importation du fichier :

* Notre fichier contient une ligne des entêtes, alors nous devrons ajouter l’option «header=T» parce que par défaut header=F

**> ozone=read.table(file.choose(), header = T)**

* Dans notre fichier, les données sont séparées par des « ; », il faut donc paramétrer le séparateur de données

**> ozone=read.table(file.choose(), header = T , sep = ";")**

* le problème des données manquantes (le mot clé vide)

**> ozone=read.table(file.choose(), header = T , sep = ";", na.strings = "vide")**

* les commentaires (toute chaines de caractères après le caractère "%")

**> ozone=read.table(file.choose(), header = T , sep = ";", na.strings = "vide", comment.char = "%",)**

On calcul ma moyenne de la variable maxO3

**> mean(ozone$maxO3)**

**[1] NA**

La commande **mean(ozone$maxO3)** renvoie **NA** perce que la colonne **maxO3** contientuneligne **NA.** Dans ce cas il faut ajouter l’option na.rm =T

**> mean(ozone$maxO3 , na.rm=T)**

**[1] 90.60185**

Pour calculer la moyenne de T9, vous allez utiliser la fonction mean() :

**> mean(ozone$T9)**

**[1] NA**

**Message d'avis :**

**Dans mean.default(ozone$T9) :**

**l'argument n'est ni numérique, ni logique : renvoi de NA**

Le problème rencontré ici revient au fait que les données de la colonne T9 ne sont pas numériques, car lors de l’importation de données du fichier Ozone.txt nous n’avons pas pris en considération que le séparateur décimal du fichier est "," alors que le séparateur décimal du

logiciel R est ".". Pour convertir le séparateur décimal il fait ajouter l’option dec lors de l’importation

**> ozone=read.table(file.choose(), header = T , sep = ";", na.strings = "vide", comment.char = "%", dec= ",")**

On peut maintenant calculer la moyenne de T9

**> mean(ozone$T9)**

**[1] 18.36071**

**Régression linéaire**

Dans cette partie nous allons importer le fichier ozone original (ozone2.txt)

Représenter graphiquement le nuage de points des données maxO3 en fonction de T9

**>plot(ozone$T12,ozone$maxO3)**

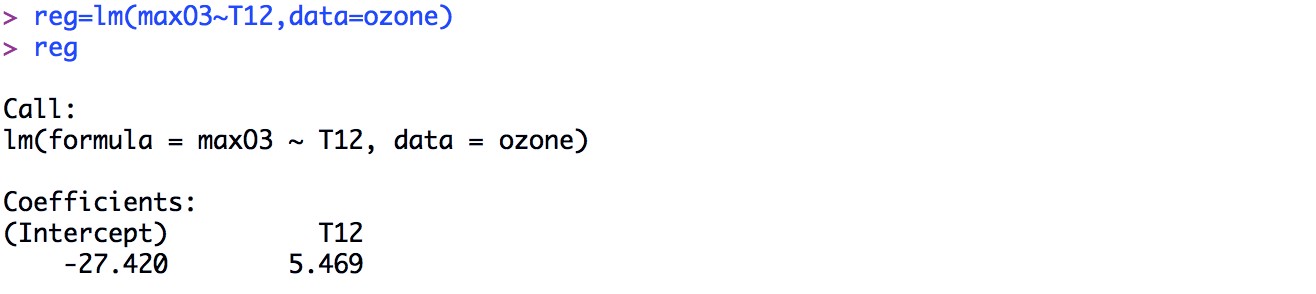
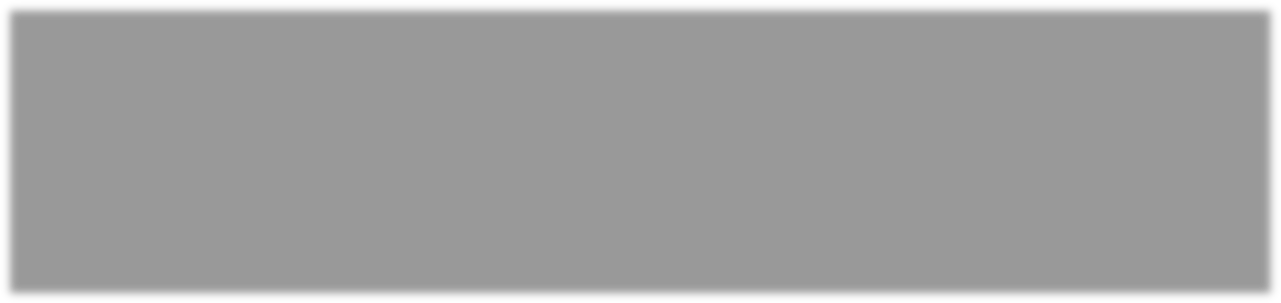
ou bien

**plot(maxO3~T12,data=ozone)**

Une régression linéaire simple semble-t’elle justifiée graphiquement ? Calculer le coefficient de corrélation linéaire

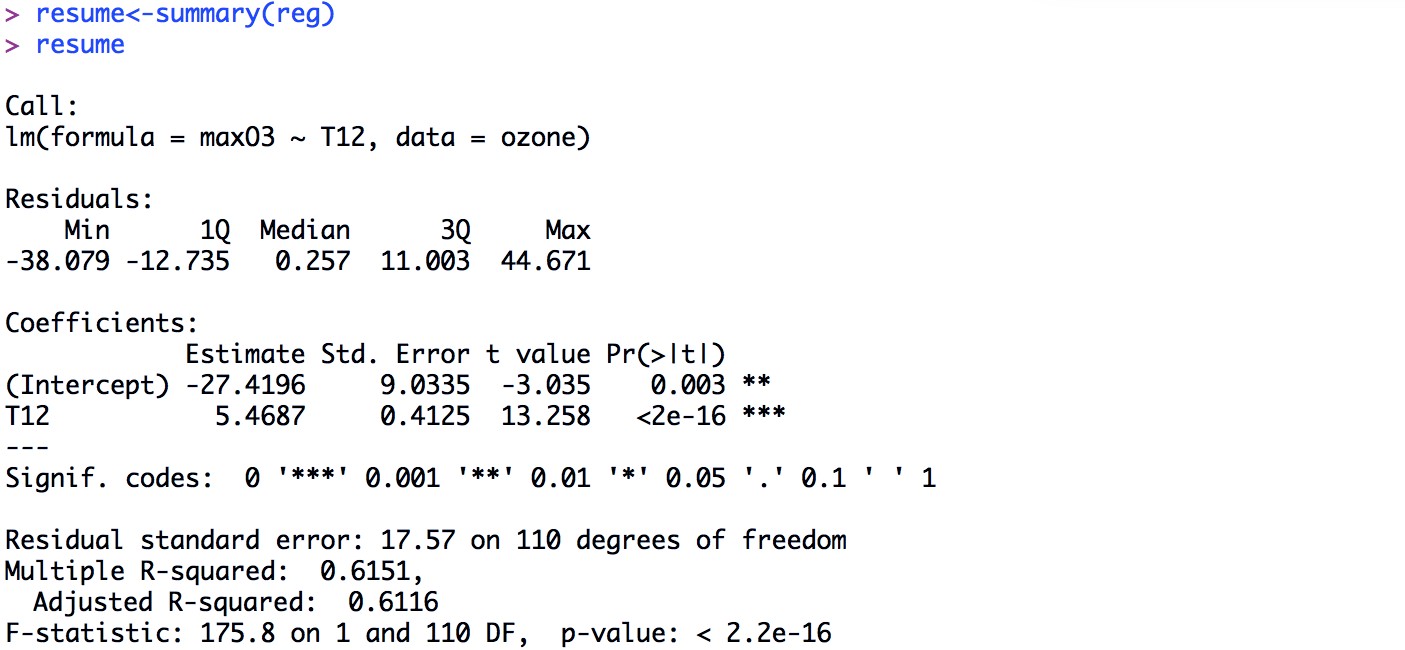
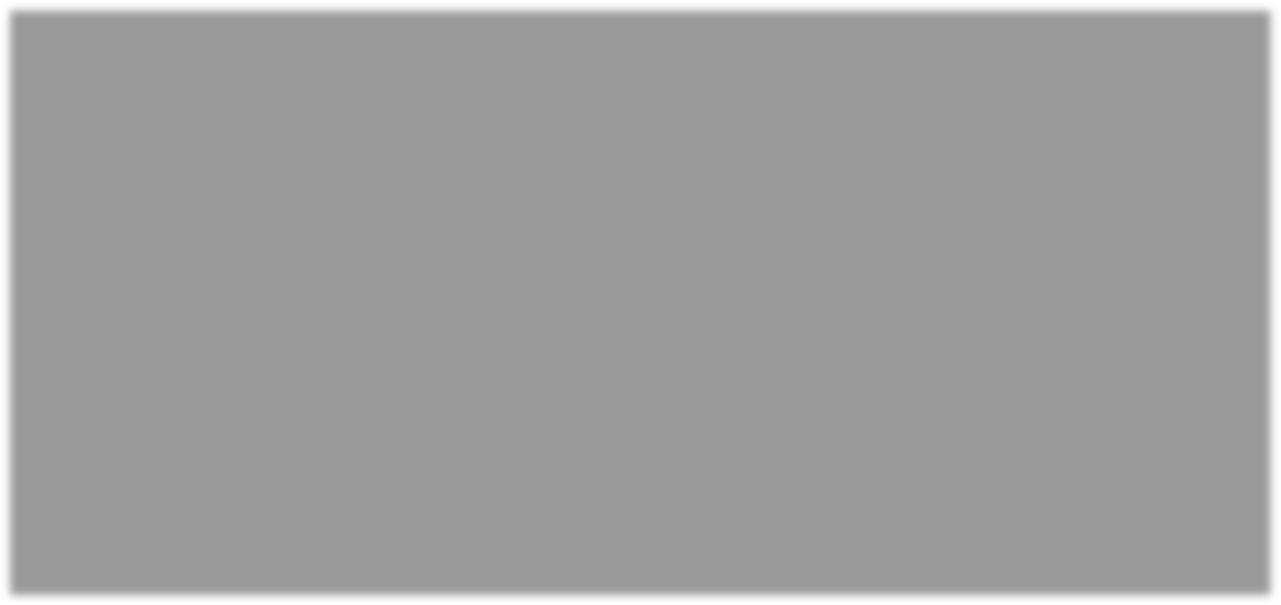
**>cor(ozone$T12,ozone$maxO3)**

Effectuer la régression linéaire à l'aide de la commande



et consulter les résultats à l'aide de la commande

pour tracer la droite de la régression linéaire, nous allons utiliser la fonction abline()



**> abline(lm(maxO3~T12,data=ozone))**

la fonction fitted()permet de donner les valeurs correspondent aux prédictions du modèle

de régression, mais pour les valeurs observées de la variable prédictive.

**> fitted(reg)**

**20010601 20010602 20010603 20010604 20010605 20010606 20010607**

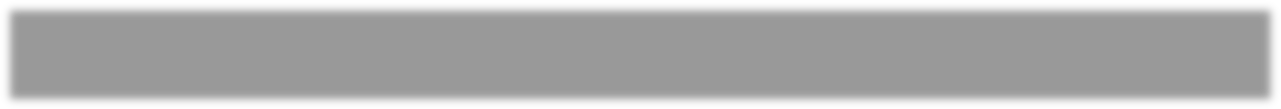
**73.75103 73.20417 68.82922 80.31346 84.68840 80.86032 57.89185**

**…**

la syntaxe ozone$maxO3-fitted(reg) est l’ecart entre la valeur réelle et la valeur

prédite

Comparer le résultat suivant :



et la partie « Residuals » de « resume »

comparer : residuals(reg) et ozone$maxO3-fitted(reg)

et : summary (residuals(reg)) et summary (ozone$maxO3-fitted(reg))

Afin de vérifier visuellement l'ajustement des données au modèle de régression proposé, nous allons tracer l'estimation de la droite de régression, ainsi qu'un intervalle de confiance à 95% de celle-ci grâce aux commandes suivantes

|  |  |  |  |
| --- | --- | --- | --- |
| **fit** | | **lwr** | **upr** |
| **1** | **49.14195 42.16484 56.11906** | | |
| **2** | **50.21912 43.38358 57.05465** | | |
| **3** | **51.29628 44.60144 57.99112** | | |
| **4** | **52.37345 45.81837 58.92853** | | |
| **…** |  | | |

Écrire une fonction en logiciel R qui permet de calculer la valeur prédite de maxO3 selon le model reg à partir de la valeur de T12

**>T12=seq(min(ozone[,"T12"]),max(ozone[,"T12"]),length=100)**

**>grille = data.frame(T12)**

**>ICdte = predict(reg,new=grille,interval="confidence",level=0.95)**

**> ICdte**

**>matlines(grille$T12,cbind(ICdte),lty=c(1,2,2),col=1)**

Contrairement à la fonction residuals(), la fonction rstudent() permet d’obtenir des résidus de même variance. Ce critère est nécessaire pour pouvoir étudier et comparer les résidus.

**> rstudent(reg)**

**20010601 20010602**

**20010603**

**20010604**

**20010605**

**20010606**

**0.7580325419 0.5025954582 1.3353329654 1.9524461622 0.5308438687 -0.0490097183**

**……**