ECM 3A-DDEFi

Regression linéaire multiple [1]

Objectifs :

- Extraction, lecture et visualisation d'un tableau de données sur R
- Ajustement des données à un modèle de régression linéaire à l'aide de R
- Analyse de la pertinence du modèle choisi

Référence: Régression avec R de Cornillon et Matzner-Lober

1. Récupérer les données dans R en utilisant la commande suivante : `ozone <- read.csv(...)`
   Dans ce fichier, nous disposons de variables climatiques et d'une variable de pollution à l'ozone. la variable **maxO3** représente le maximum journalier d'ozone. Les variables climatiques sont les suivantes: la température, la nébulosité et la projection du le vent sur l'axe sur l'axe Est-Ouest à 9h, 12h et 15h ainsi la teneur en ozone maximale la veille et la pluie.

2. Quelles sont les différentes variables? Quelle est leur nature?
   Indication : Utiliser `names(ozone)` *et* `head(ozone)`.

3. Représenter le maximum d'ozone (**maxO3**) en fonction de la température (**T12**).
   Conseil : Utiliser la fonction `plot`.

4. Même question pour le maximum de l'ozone en fonction du vent et de la pluie. Représenter le vent en fonction de la température.

5. Obtenir les statistiques descriptives du jeu de données `summary(ozone)`

6. Analyser la normalité de la variable **maxO3** à l'aide d'un Q-Q Plot.
   Après une recherche sur le test Shapiro-Wilk (problème de test : hypothèse nulle, alternative; statistique de test, région de rejet), valider le résultat obtenu par le Q-Q Plot à l'aide du test Shapiro-Wilk.

7. Nous allons nous intéresser plus particulièrement à deux variables du jeu de données : la variable **maxO3** et la variable **T12**.

   (a) Nous allons calculer les statistiques élémentaires sur ces deux variables.
   Conseil : Utiliser la commande `summary`

   (b) Représenter le nuage de points $(x_i; y_i)$.
   Conseil : Utiliser la commande `plot`

   (c) Proposer un modèle de régression de **maxO3** par rapport à la variable **T12**.
   Indication : Utiliser la commande `reg1 <- lm(O3~ T12,data=ozone)` *puis la commande* `summary(reg1)`

   (d) Après analyse des résultats, que peut-on déduire?

8. Proposer un modèle de régression de **maxO3** par rapport aux variables **Ne12** et **maxO3v**.
   Indication : Utiliser la commande `reg2 <- lm(O3~Ne12+maxO3v,data=ozone)` *puis la commande* `summary(reg2)`
   Après l'analyse des résultats, que peut-on déduire?

9. Modèle de régression complet.

[1] Version du 11/09/2023

(a) A l'aide de la fonction `lm` estimer les paramètres de la régression de **maxO3** sur les autres variables et afficher les estimateurs.
Indication : `regc<-lm(ozone$maxO3~ . ,data=ozone)` *et* `regc$coefficients`

(b) Extraire les résidus et tracer leur histogramme.
Indication : `regc$residuals`, `hist(regc$residuals)`

(c) A l'aide d'un Q-Q Plot comparer les quantiles des résidus avec les quantiles d'une loi gaussienne.
Indication : `qqnorm(regc$residuals)`, `qqline(res1.regc$residuals)`

(d) Etudier la normalité des résidus à l 'aide d'un test Shapiro-Wilk:

(e) Le test de Kolmogorov-Smirnov de comparaison à une distribution théorique pourrait également être utilisé (`rks.test`). Appliquer le test de Kolmogorov-Smirnov. Conduit-il à la même conclusion que le test de Shapiro-Wilk?

(f) Quel autre test, vu en Approfondissement S7 MIE, pourriez-vous appliquer pour tester la normalité?

(g) Etudier graphiquement l'homoscédasticité des résidus:

```
plot(regc$fitted.values,abs(regc$residuals), col=2)
lines(lowess(regc$fitted.values,abs(regc$residual),f=0.7))
```

(h) Les résidus obtenus ne sont pas de même variance (hétéroscédastiques). Nous allons utiliser alors les résidus studentisés, qui eux sont de même variance.

```
res.simple<-rstudent(regc)
plot(res.simple,pch=15,cex=.5,ylab="Residus",ylim=c(-3,3))
abline(h=c(-2,0,2),lty=c(2,1,2))
```

(i) Appliquer le test Breusch-Pagan (`bptest` sous R), quel est l'objectif de ce test et que peut-on déduire?

(j) Analyser graphiquement la structuration temporelle des résidus?

```
plot(regc$residual,col=2)
ines(lowess(regc$residual,f=0.7),lty=2)
```

(k) Appliquer le test de Durbin-Watson (`dwtest` sous R), quel est l'objectif de ce test et que peut-on déduire?

(l) Appliquer le test de Breusch-Godfrey (`bgtest` sous R), quel est l'objectif de ce test et que peut-on déduire?

(m) Pouvez-vous repérer une structure particulière du nuage ou la présence de "grands" résidus:

```
res.student=rstudent(regc)
ychap=regc$fitted.values,
plot(res.student,ylab="R\' esidus"),
abline(h=c(-2,0,2),lty=c(2,1,2))}
```

(n) Repérer d'éventuels points influents.

```
cook=cooks.distance(regc),
plot(cook~ychap,ylab="Distance de Cook")
abline(h=c(0,1),lty=c(1,2))})
```

(o) Analyser la significativité des variables et proposer les variables pertinentes.
Indication : `summary(regc)`

(p) Nous disposons d'une nouvelle observation de la température **T12** égale à 19 degrés pour le 1er octobre 2001. Prédire le niveau d'ozone pour cette date.

```
  xpredict=19
xpredict<-as.data.frame(xnew)
colnames(xpredict)<-"T12"
predict(regc,xpredict,interval="pred")
```

(q) Représenter sur un même graphique l'intervalle de confiance d'une valeur lissée et l'intervalle de confiance d'une prévision. Pour ce faire il faut calculer ces intervalles pour l'ensemble des points ayant servi à construire la droite de régression. Utiliser la commande suivante:

```
grillex<-seq(min(ozone[,"T12"]),max(ozone[,"T12"]),length=100)
grillex.df<-data.frame(grillex)
dimnames(grillex.df)[[2]]<-"T12"
IC<-predict(regc,new=grillex.df,interval="conf", level=0.95)
ICprev<-predict(regc,new=grillex.df,interval="pred",level=0.95)
plot(maxO3~T12,data=ozone,pch=15,cex=.5)
matlines(grillex,cbind(IC,ICprev[,-1]),lty=c(1,2,2,3,3),col=1)
legend("topleft",lty=2 :3,c("prev","conf"))
```

(r) Analyser la normalité des résidus.

10. Proposer un modèle de régression de **O3** sur les variables **T15**, **Ne12**, **Vx** et **maxO3v**. Remarquer que certaines variables déclarées non-importantes dans le modèle complet sont déclarées importantes dans le modèle intermédiaire.

11. (Comparaison des modèles linéaires). Le modèle linéaire par rapport à quelle variable est-il raisonnable?
Quelle est l'influence de chaque variable?
Indication : Utiliser la fonction `anova(reg1,regc)`
Il existe aussi un package R qui traite du choix des variables: le package **leaps**

```
library("leaps", lib.loc="~/R/lib")
library("lattice")
choix <- regsubsets(maxO3 ~ ., data=ozone, nbest=1,nvmax=11)
plot(choix)
```