Scénario: Statistiques élémentaires et concentration d'ozone

Résumé

Initiation à la pratique des techniques élémentaires de statistique par l'étude d'un jeu de données à l'aide du logiciel R. L'objectif est de prévoir la concentration en ozone le lendemain à partir de celle du jour et de mesures météorologiques : description élémentaire, estimation, tests, régression linéaire, analyse de variance (ANOVA), analyse en composantes principales, régression multiple. Si nécessaire, un tutoriel de démarrage avec R est disponible.

1 Introduction

La pollution de l'air constitue actuellement une des préoccupations majeures de santé publique. De nombreuses études épidémiologiques ont permis de mettre en évidence l'influence sur la santé de certains composés chimiques comme le dioxyde souffre (SO2), le dioxyde d'azote (NO2), l'ozone (O3) ou des particules en suspension. Des associations de surveillance de la qualité de l'air (Air Breizh en Bretagne depuis 1994) existent sur tout le territoire métropolitain et mesurent la concentration des polluants. Elles enregistrent également les conditions météorologiques comme la température, la nébulosité, le vent, les chutes de pluie en relation avec les services de Météo France... L'une des missions de ces associations est de construire des modèles de prévision de la concentration en ozone du lendemain à partir des données disponibles du jour : observations et prévisions de Météo France. Plus précisément, il s'agit d'anticiper l'occurrence ou non d'un dépassement légal du pic d'ozone (180 μgr/m3) le lendemain afin d'aider les services préfectoraux à prendre les décisions nécessaires de prévention : confinement des personnes à risque, limitation du trafic routier. Plus modestement, l'objectif de cette étude est de mettre en évidence l'influence de certains paramètres sur la concentration d'ozone (en μgr/m3) et différentes variables observées ou leur prévision. Les 112 données étudiées ont été recueillies à Rennes durant l'été 2001. Elles sont disponibles sur le site du laboratoire de mathématiques appliquées de l'Agrocampus Ouest. Les 13 variables observées sont :

- MaxO3 : Maximum de concentration d'ozone observé sur la journée en μgr/m3
- T9, T12, T15 : Température observée à 9, 12 et 15h
- Ne9, Ne12, Ne15 : Nébulosité observée à 9, 12 et 15h
- Vx9, Vx12, Vx15 : Composante E-O du vent à 9, 12 et 15h
- MaxO3v : Teneur maximum en ozone observée la veille
- vent : orientation du vent à 12h
- pluie : occurrence ou non de précipitations

2 Exploration statistique élémentaire

2.1 Lire les données

Les données sont disponibles dans le répertoire `http://wikistat.fr/data` sous la forme d'un fichier `ozone.csv` construit à partir de Excel en choisissant " ;" comme séparateur et "," comme marque décimale. Télécharger ce fichier dans le répertoire courant de R avant d'exécuter les commandes :

```
# Data frame à partir d'un fichier csv
ozone=read.csv2("ozone.csv")
# vérification
summary(ozone)
# Supprimer la variable inutile "obs"
ozone=ozone[,-1]
```

2.2 Description unidimensionnelle

Décrire chacune des variables en précisant ses caractéristiques :

2.2.1 *Variables quantitatives*

Décrire chaque variable (moyenne, écart-type, quantiles, diagramme boîte, histogramme) afin d'identifier les problèmes potentiels : valeurs atypiques, hétérogénéité des variances, distributions dissymétriques...

```
summary(ozone)
```

```
sapply(ozone, mean)   # moyennes
sapply(ozone, sd)     # écarts-types
boxplot(ozone[,2:4]) # boîtes par groupe
boxplot(ozone[,5:7])
boxplot(ozone[,8:10])
boxplot(ozone[,c(1,11)])
```

Commenter les résultats obtenus.

Les deux variables de concentration d'ozone demandent plus d'attention. Il serait intéressant de tester une transformation, souvent le logarithme pour une variable de concentration.

```
hist(ozone$maxO3)
hist(ozone$maxO3v)
hist(log(ozone$maxO3))
hist(log(ozone$maxO3v))
boxplot(log(ozone[,c(1,11)]))
```

Les distributions semblent alors plus symétriques et ne présentent plus de valeurs atypiques.

2.2.2 Variables qualitatives

Fréquences des modalités des variables qualitatives.

```
barplot(table(ozone$pluie))
barplot(table(ozone$vent))
pie(table(ozone$vent))
```

2.3 Description bidimensionnelle

2.3.1 Variables quantitatives

Une matrice de nuages de points donne un aperçu rapide des structures de corrélation :

```
pairs(ozone[,1:11])
plot(maxO3~maxO3v,data=ozone)
plot(log(maxO3)~log(maxO3v),data=ozone)
```

Repérer des groupes de variables très corrélées et la liaison entre les variables principales d'intérêt.

2.3.2 Variables qualitatives

calcul de la table de contingence et graphe des profils colonnes dans un *mosaic plot*.

```
table(ozone$vent,ozone$pluie)
mosaicplot(table(ozone$vent,ozone$pluie))
```

Commenter la liaison entre ces deux variables.

2.3.3 Variables qualitatives et quantitatives

Représenter une possible liaison entre les variables principales et celles qualitatives par des diagrammes boîte.

```
boxplot(maxO3~pluie,data=ozone)
boxplot(maxO3~vent,data=ozone)
```

Commenter.

Bien d'autres options permettent de modifier les apparences des graphiques (titres, légendes...). Consulter l'aide en ligne si nécessaire.

3 Tests de comparaison

Important : Lors de l'exécution de chaque test préciser explicitement :

1. la question posée,
2. l'hypothèse H_0 en relation avec la question et associée au test,
3. la p-valeur calculée et la décision du test,
4. la réponse à la question.

3.1 Cas gaussien

Beaucoup des outils ci-dessous nécessitent de vérifier le caractère gaussien ou non de la distribution. En fait, le nombre important d'observations dans l'échantillon permet de s'affranchir de cette hypothèse mais il est utile de savoir la vérifier et éventuellement de sélectionner la transformation la plus appropriée des données notamment pour les variables de concentration d'ozone.

3.1.1 Normalité d'une distribution : Shapiro-Wilks

La droite de Henri ou graphe quantile-quantile donne déjà un aperçu graphique de la normalité de la distribution avant de calculer le test.

```
# qq-plots
qqnorm(ozone$maxO3)
qqline(ozone$maxO3,col=2)
qqnorm(log(ozone$maxO3))
qqline(log(ozone$maxO3),col=2)
# Test de shapiro-Wilks
shapiro.test(ozone$maxO3)
shapiro.test(log(ozone$maxO3))
```

Le test de Kolmogorov-Smirnov de comparaison à une distribution théorique pourrait également être utilisé (`ks.test`).

Les variables transformées sont ajoutées dans la table.

```
ozone=data.frame(ozone,LmaxO3=log(ozone$maxO3),
   LmaxO3v=log(ozone$maxO3v))
summary(ozone)
```

3.1.2 Intervalle de confiance d'une moyenne : Student

Il est important de savoir estimer l'intervalle de confiance d'une moyenne ; celui-ci permet de tester l'égalité de cette moyenne à une valeur théorique selon l'appartenance ou non de cette valeur à l'intervalle. L'effectif étant suffisamment grand, il n'est pas nécessaire de supposer la normalité des données mais la variable transformée la plus "gaussienne" est choisie. L'intervalle de confiance est calculé par défaut avec un seuil à 95% mais ce paramètre peut être précisé (`conf.level=.95`) de même que la moyenne théorique testée (`mu=0.0`, par défaut à 0).

```
t.test(ozone$LmaxO3, conf.level=.95)
```

3.1.3 Comparaison de deux variances : Fisher

On s'intéresse à l'influence de la présence de pluie sur la concentration en ozone. Tester l'égalité des deux moyennes nécessite de vérifier péralablement plusieurs points :

1. la normalité des distributions dans chaque classe à moins que l'échantillon soit considéré de taille suffisamment grande,
2. le caractère indépendant ou appariés des échantillons,
3. l'égalité ou non des variances à l'intérieure de chaque groupe.

On dispose de deux échantillons *indépendants* : les jours de pluie et les jours de temps sec. Testons les autres hypothèses.

```
# Normalité des distributions (facultatif)
shapiro.test(ozone[ozone$pluie=="Pluie","LmaxO3"])
shapiro.test(ozone[ozone$pluie=="Sec","LmaxO3"])
# égalité des variances (test de Fisher)
var.test(LmaxO3~pluie,data=ozone)
```

Commenter les résultats.

3.1.4 Comparaison de deux moyennes

Le test de comparaison des moyennes à utiliser (Student *vs.* Welsh) dépend du résultat précédent concernant l'égalité des variances.

Echantillons indépendants Si les variances sont différentes, il s'agit d'un test de Welch.

```
t.test(LmaxO3~pluie,var.equal=F, data=ozone)
```

Dans le cas où elles sont considérées égales, c'est un test de Student.

```
t.test(LmaxO3~pluie,var.equal=T, data=ozone)
```

Échantillons appariés On se propose d'étudier la persistance moyenne de la concentration en comparant la moyenne du jour avec celle de la veille. La mesure étant observée au même point à deux instants différents, les échantillons sont cette fois appariés.

```
t.test(ozone$maxO3, ozone$maxO3v,paired=TRUE)
```

3.2 Cas non-paramétrique

Si l'hypothèse de normalité des distributions n'est pas vérifiée et si l'échantillon est trop réduit, c'est un test non-paramétrique qu'il faut mettre en œuvre. Les tests non-paramétriques sont basés sur les rangs des observations et donc sur les comparaisons des médianes des échantillons. Une transformation des variables par une fonction monotone (*i.e.* log) qui ne changent pas leur ordonnancement n'a donc pas d'effet sur le calcul d'un test non paramétrique.

3.2.1 Comparaison de deux médianes : Wilcoxon

Echantillons indépendants

```
tapply(ozone$LmaxO3, ozone$pluie, median)
wilcox.test(maxO3 ~ pluie, data=ozone)
```

Échantillons appariés

```
median(ozone$LmaxO3 - ozone$LmaxO3v)
wilcox.test(ozone$LmaxO3, ozone$LmaxO3v,paired=TRUE)
```

Comparer avec les résultats des tests paramétriques.

4 Tests de liaison

4.1 Indépendance de 2 variables qualitatives

Le test du $\chi 2$ est adapté à ce problème.

```
chisq.test(table(ozone$pluie,ozone$vent))
```

Remarque : un avertissement peut signaler que les effectifs théoriques (sous hypothèse d'indépendance) de certaines cellules sont trop faibles pour justifier des propriétés asymtotiques du test du $\chi 2$. Il est dans ce cas nécessaire de regrouper des modalités.

4.2 Une quantitative et une qualitative

L'ANOVA associée à un test de Fisher adapté à cette situation est sans doute le test le plus utilisé ; il revient au test de Student lorsque la variable qualitative n'a que deux modalités. L'ANOVA nécessite de vérifier :

1. le caractère indépendant des échantillons,
2. la normalité des distributions (ou une taille suffisante d'échantillon) dans chaque classe ou plutôt la normalité des résidus au modèle,
3. l'égalité des variances internes à chaque groupe.

Même si la normalité des résidus est vérifiée *a posteriori*, c'est *a priori* qu'il faut prendre en compte ce résultat pour statuer sur la légitimité du test.

Si la normalité n'est pas vérifiée pour un petit échantillon ou si l'agalité des variances n'est pas acceptable, un test non-paramétrique (Kruskal-Wallis) doit être envisagé.

4.2.1 Cas gaussien : ANOVA - Fisher

Le test de Bartlett permet de comparer les variances des groupes dans le cas gaussien ou paramétrique.

```
# test de Bartlett
bartlett.test(LmaxO3 ~ vent, data=ozone)
# ANOVA à un facteur
# estimation des paramètres
res.anova=aov(LmaxO3 ~ vent, data=ozone)
# normalité des résidus au modèle d'ANOVA
qqnorm(res.anova$residuals)
qqline(res.anova$residuals)
shapiro.test(res.anova$residuals)
# Interprétation du test
summary(res.anova)
```

Commenter.

4.2.2 Cas non-paramétrique : Kruskal-Wallis

```
kruskal.test(maxO3 ~ vent, data=ozone)
```

Comparer les résultats.

4.3 Deux variables quantitatives

La régression simple permet de tester l'influence éventuelle d'une variable sur une autre et plus précisément, dans le cas de cet exemple, d'expliquer et même de prévoir la concentration d'ozone en fonction de celle de la veille. La commande `lm` produit un ensemble de résultats sous la forme d'une liste de matrices et vecteurs.

4.3.1 Estimation du modèle

```
# retracer le nuage de point
plot(LmaxO3 ~ LmaxO3v,data=ozone)
# estimation du modèle
res1.reg=lm(LmaxO3 ~ LmaxO3v, data = ozone)
# liste des résultats
names(res1.reg)
```

4.3.2 Diagnostic des résidus

Des graphiques précédents permettent de s'assurer de la validité du modèle ; statuer sur l'homoscédasticité des résidus, leur normalité, la bonne linéarité du modèle.

```
# nuage de point
# normalité des résidus
qqnorm(res1.reg$residuals)
qqline(res1.reg$residuals)
shapiro.test(res1.reg$residuals)
# Repérage d'une structure particulière du nuage
# ou de la présence de "grands" résidus
res.student=rstudent(res1.reg)
ychap=res1.reg$fitted.values
plot(res.student~ychap,ylab="Résidus")
# ajouter des lignes
abline(h=c(-2,0,2),lty=c(2,1,2))
# repérage des points influents
cook=cooks.distance(res1.reg)
plot(cook~ychap,ylab="Distance de Cook")
abline(h=c(0,1),lty=c(1,2))
```

Les résidus sont "grands" si, une fois normalisés ou plutôt "studentisés", ils sont de valeur absolue plus grande que 2. Une observation est influente si elle a un grand résidu est est associée à une grande valeur sur la diagonale de la *hat matrix*. Cela correspond à une valeur élevée (plus grande que 1) de la distance de Cook.

4.3.3 Significativité du modèle

```
summary(res1.reg)
```

Que dire de l'influence de seuil d'ozone de la veille ? Que dire également de la présence d'observations à effet levier potentiel ? Que dire de la qualité d'ajustement de ce modèle et donc de la qualité attendue de la prévision ? Interpréter les tests.

5 ACP et régression multiple

5.1 Analyse en composantes principales

Cette description élémentaire permet de se familiariser avec la structure de corrélation particulière des variables. Il faut sélectionner les seules variables quantitatives et l'ACP est réduite.

```
res.pca=prcomp(ozone[,c(2:10,14,15)],scale=T)
# décroissance des valeurs propres
plot(res.pca)
# parts de variance expliquée
summary(res.pca)
# biplot du premier plan principal
biplot(res.pca)
```

Comment s'interprètent les axes 1 et 2 ?

5.2 Régression multiple

5.2.1 Modèle linéaire complet

La régression linéaire simple ne conduit pas à un modèle bien ajusté. Le modèle linéaire multiple ci-dessous, plus complexe, recherche un meilleur ajustement des données.

```
# estimation
res2.reg=lm(LmaxO3 ~ LmaxO3v+T9+T12+T15+Ne9+Ne12+
   Ne15+Vx9+Vx12+Vx15, data = ozone)
# diagnostics
plot(res2.reg)
# résultats
summary(res2.reg)
```

Commenter les résultats obtenus sur la validité du modèle et la qualité de l'ajustement par rapport au modèle précédent. Que dire à propos de la significativité des tests de Student sur la nullité des paramètres ? Que penser alors de la présence de variables présentant de fortes colinéarités ?

5.2.2 Sous-modèle

Une procédure de sélection de modèle non détaillée (*stepwise*) conduit à considérer le modèle ci-dessous :

```
res3.reg=lm(LmaxO3~LmaxO3v+T12+Ne9+Vx9,data=ozone)
# diagnostics
plot(res3.reg)
# résultats
summary(res3.reg)
```

Commenter à nouveau les résultats.

5.2.3 Meilleure prévision

L'objectif est de rechercher le meilleur modèle de prévision de la concentration en ozone. Ceux-ci sont comparés en considérant le PRESS (predicted residual sums of squares) ou *leave one out cross validation*. Une fonction élémentaire est définie pour calculer le PRESS dans le cas élémentaire de la régression linéaire.

```
# définition de la fonction PRESS
press=function(model) {
h=influence(model)$hat
e=influence(model)$wt.res
n=length(e)
sum((e/(1-h))^2)/n
}
# application aux différents modèles
press(res1.reg)
press(res2.reg)
press(res3.reg)
```

Le meilleur modèle de prévision est-il celui qui ajuste le mieux les données ?

Annexe : analyse de covariance et sélection de variables

Ce scénario se limite volontairement aux outils les plus élémentaires. D'autres modèles seraient à tester, notamment une analyse de covariance associant les variables qualitatives au modèle, la présence ou non d'interactions... pour tenter d'améliorer la qualité de prévision. C'est l'objet d'autres scénarios. Pour achever celui-ci dans la logique des outils linéaires mis en œuvre, voici la construction du modèle d'analyse de covariance avec sélection de variables au sens du critère AIC pour optimiser les qualités de prévision. C'est le même type d'algorithme qui a été exécuté pour recherche le sous-modèle "optimal".

L'algorithme de sélection descendante débute par l'estimation du modèle complet. Les interactions ne sont pas prises en compte ; celles avec la pluie ne change rien, elles sont éliminées systématiquement par l'algorithme.

```
res.acova=glm(LmaxO3 ~ T9+12+T15+Ne9+Ne12+
   Ne15+Vx9+Vx12+Vx15+vent+pluie+LmaxO3v,data=ozone)
# Recherche du meilleur modèle au sens
# du critère d'Akaïke par méthode descendante
res.acova.step=step(res.acova,direction="backward")
# paramètres retenus
anova(res.acova.step,test="F")
# Extraction des valeurs ajustées et des résidus
fit.acova=res.acova.step$fitted.values
resid.acova=res.acova.step$residuals
# Graphe des résidus
plot(fit.acova,resid.acova)
```

Ces résultats montrent que l'occurence de pluie est bien à prendre en compte

dans le modèle. Ce modèle est comparé avec le précédent en calculant une estimation de l'erreur de prévision par validation croisée.

```
library(boot) # chargement de la bibliothèque
# validation croisée  10-plis
# meilleur modèle linéaire
res3.reg=glm(LmaxO3~LmaxO3v+T12+Ne9+Vx9,data=ozone)
res4.reg=glm(LmaxO3~LmaxO3v+T12+Ne9+Vx9+pluie,
   data=ozone)
set.seed(111)
cv.glm(ozone, res3.reg, K=10)$delta[1]
set.seed(111)
# modèle d'analyse de covariance
cv.glm(ozone, res.acova.step, K=10)$delta[1]
```

Même si la variable pluie est significative dans le modèle, l"'amélioration" de la qualité de prévision n'est pas franchement significative.