UE - Sciences Fondamentales - Modélisation Statistique

Projet sur R de Statistique (2A) - 2020

Élèves	ESQUIROL Juliette - PARIZE Antoine
Groupes	4 - 1
Date	25/11/2020

Résolution des 2 derniers exercices du TD2

TABLE DES MATIÈRES

I. Exercice 1 - Ozone

Présentation rapide

Le jeu de données recense 112 observations des 11 variables quantitatives suivantes : la concentration maximale journalière en ozone (**maxO3**, en µg/m3), la concentration en ozone de la veille (**maxO3v**), des variables de nébulosité (**Ne9, Ne12, Ne15**), des variables de température (**T9, T12, T15**), et enfin des variables liées au vent (**Vx9, Vx12, Vx15**).
Ces variables quantitatives sont exprimées en fonction de deux variables qualitatives : la variable **pluie** (i.e, Pluie vs Sec) ainsi que la variable **vent** indiquant la direction du vent (i.e, nord, sud, est, ou ouest).

Le but de cet exercice est de proposer le meilleur modèle de régression multiple capable de prédire la concentration en ozone **maxO3** en fonction des autres variables quantitatives disponibles.

La représentation graphique deux à deux des variables du jeu de données semble indiquer une dépendance linéaire de certaines variables (voir Figure 1).

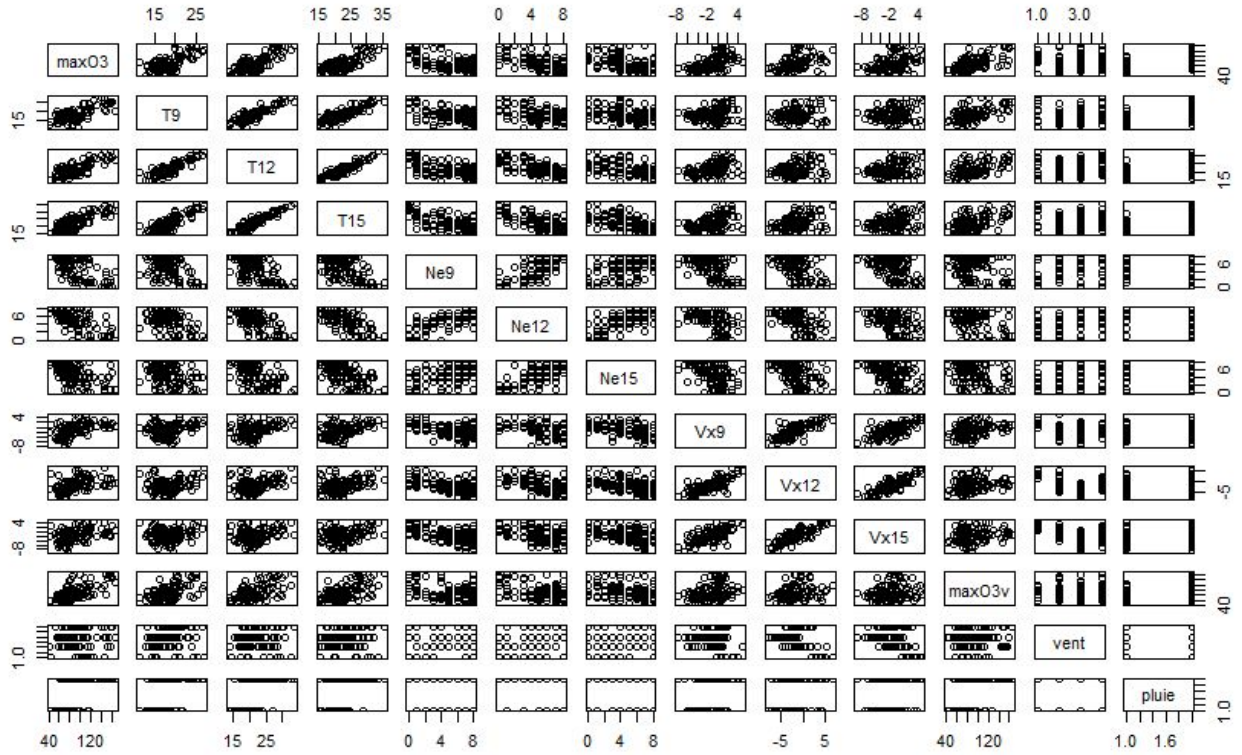


Figure 1 : Représentation graphique 2 à 2 des variables du jeu de données Ozone. Certaines figures semble indiquer une dépendance linéaire entre les variables

ACP sur les données quantitatives

Au vu du grand nombre de variables quantitatives (11) dans ce jeu de données, réaliser une ACP sur ces variables nous semblait être pertinent. En effet, cette ACP pourrait nous renseigner quant à la redondance d'informations entre les différentes variables quantitatives du jeu de données.

L'ACP réalisée sur les variables quantitatives indique que les variables sont globalement bien représentées sur les deux premières composantes principales (squared loadings > 0.5 pour la plupart). Cependant, la variable maxO3v ainsi que les variables de vents ne sont pas très bien représentées sur le plan 1-2 car leur square loadings sont faibles (environ 0.40).

On remarque une redondance d'informations, plusieurs variables semblent fortement corrélées entre elles, formant des groupes de variables similaires (voir Figure 2). On observe aussi que les variables de température et de vent sont positivement corrélées entre elles, mais négativement corrélées aux variables de nébulosité. De plus, la concentration en ozone maxO3 est elle aussi négativement corrélée aux variables de nébulosité. En résumé, plus la nébulosité sera basse, plus la température, le vent et la concentration en ozone seront élevés.

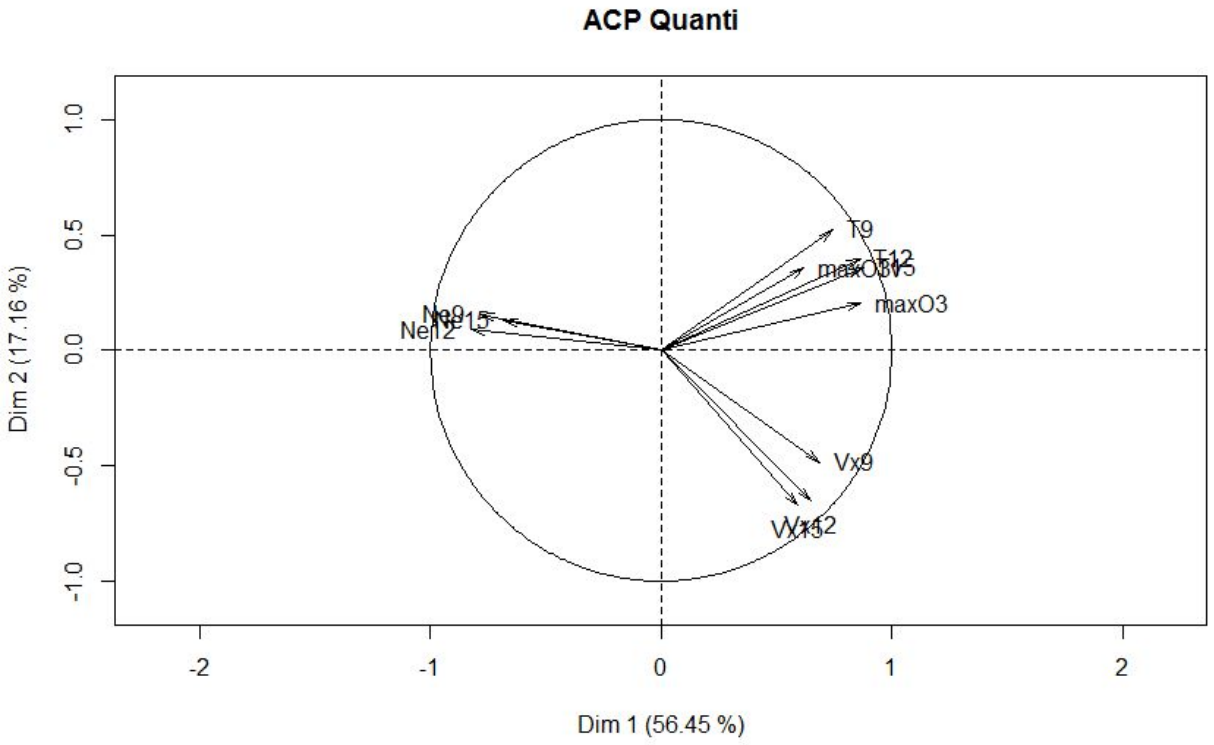


Figure 2 : Cercle des corrélation obtenu à partir de l'ACP sur les variables quantitatives uniquement du jeu de données ozone. On remarque une corrélation entre les variables d'un même type (i.e, Nébulosité, Vent et Température).

Impact des variables vent et pluie sur maxO3

Le jeu de variables ozone possède comme nous l'avons dit plus tôt, deux variables qualitatives : vent à 4 modalités, et pluie à 2 modalités. Nous allons dans ce paragraphe étudier descriptivement l'impact des variables vents et pluie sur la concentration en ozone (Figure 3 et 4).

On remarque qu'il semble y avoir un effet de la direction du vent sur la concentration en ozone. (Figure 3). En effet, les concentrations semblent plus importantes lorsque le vent souffle vers l'est ou le sud comparé au nord et à l' ouest. Il est important de noter la présence de quelques valeurs extrêmes : Pour la modalité "est", une observation est extrêmement faible. Les modalités "nord" et "ouest" présentent toutes les deux quelques observations très hautes en concentration d'ozone.

En ce qui concerne la variable pluie (Figure 4), il semble que la concentration en ozone soit plus importante lorsqu'il fait sec, plutôt que quand il pleut. On notera aussi la présence de quelques observations extrêmes de concentration en ozone pour la modalité "pluie".

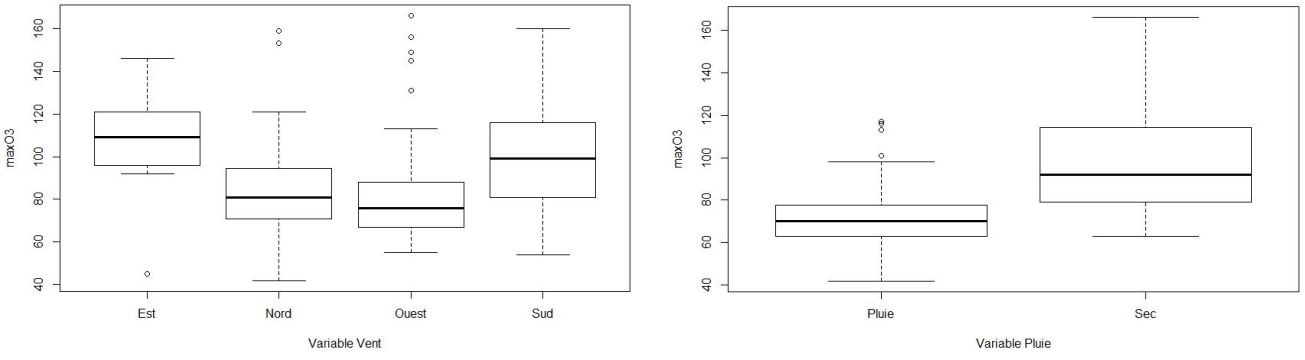


Figure 3 et 4 : Boxplots de la concentration maximale en ozone (maxO3) respectivement en fonction du vent (à gauche) et de la pluie (à droite)

ACP mixte sur le jeu de données entier

Après avoir analysé descriptivement l'effet potentiel des variables Vent et Pluie sur maxO3, nous avons décidé de conduire une ACP mixte sur nos données, l'inclusion de ces variables pouvant potentiellement changer les résultats obtenus. Le cercle des corrélations de cette ACP mixte est visible en Figure 4. Ce dernier est d'ailleurs très similaire à celui obtenu en ACP classique (Figure 2). Les conclusions que nous avons faites dans la partie sur l'ACP classique sont donc applicables à cette ACP mixte.

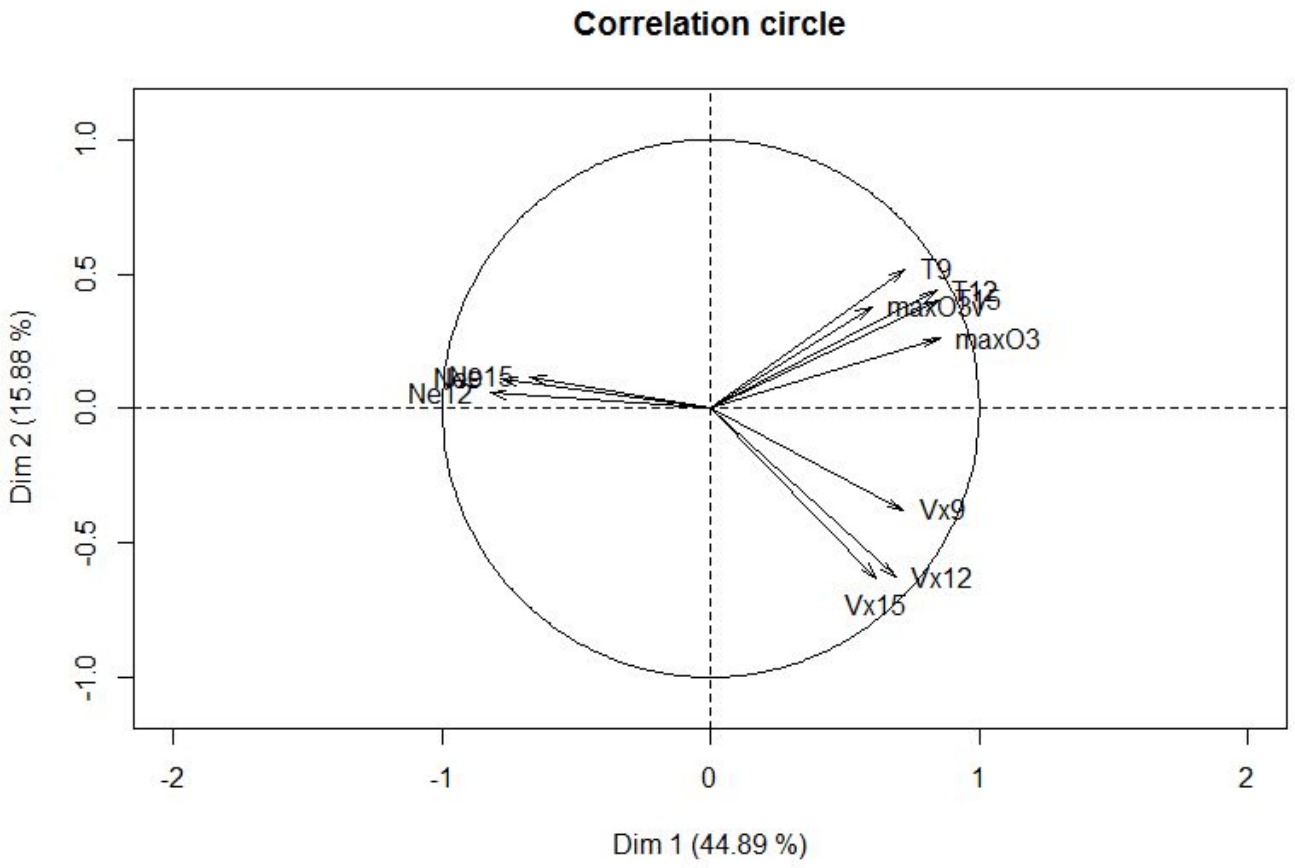


Figure 5 : Cercle des corrélation obtenu après avoir réalisé une ACP mixte sur le jeu de données ozone.

Recherche du meilleur modèle de régression multiple

Compte tenu des ACP réalisées plus tôt et de la redondance d'informations parmi les variables, nous nous attendons à ce que le modèle de régression multiple expliquant le mieux les données comprenne un nombre réduit de variables par rapport à toutes celles disponibles dans le jeu de données initial.

Notre première étape dans la recherche d'un modèle optimal a été la création d'un modèle naïf contenant l'ensemble des 11 variables explicatives. Le modèle obtenu explique environ 74% de la variance des données, mais seulement deux coefficients du modèles sont significativement différents de zéro (Ne9 : $p = 0.02$, maxO3v : $p < 0.001$).

En s'intéressant aux résidus studentisés de notre modèle, on se rend compte que certaines valeurs sont extrêmes par rapport à l'ensemble des données. Cela se traduit par des valeurs bien supérieures à 2 en valeur absolue, mais aussi par une distribution non normale des résidus du modèle (voir Figure 6). Cependant, nous avons choisi de ne pas exclure tout de suite ces valeurs et de voir si un modèle "optimisé" n'aurait pas de tels résidus.

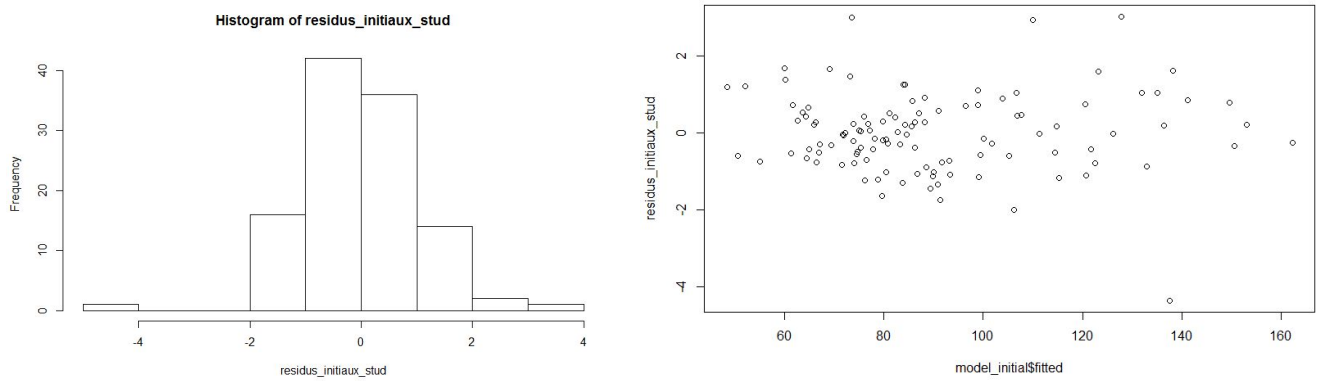


Figure 6 : Histogramme des valeurs des résidus du modèle initial (à gauche), et figure des résidus du modèle initial en fonction des valeurs prédites (à droite). On remarque que la distribution des données n'est pas normale au sens probabiliste du terme.

Nous avons donc choisi de créer un modèle initial prédit par trois variables relatives respectivement au vent (Vx12), à la nébulosité (Ne12) et à la température (T12). Nous avons donc choisi d'utiliser une méthode ascendante pour trouver le meilleur modèle de régression multiple prédisant maxO3. Après utilisation de la fonction *step()*, le modèle optimal pour expliquer les données est le suivant : maxO3 ~ maxO3v + Ne9 + T12 + Vx9. Cependant, ce modèle n'explique qu'à 75% la variance dans les données, un faible gain par rapport au modèle initial (1%). Le critère AIC pour ce modèle est de 696.02.

En nous intéressant à nouveau aux résidus du modèle on s'aperçoit que le problème des résidus persiste. Pour essayer de se débarrasser de ce problème nous avons décidé d'évincer de notre jeu de données les observations à l'origine de ces résidus extrêmes. Ces valeurs sont peut-être tout à fait normales, mais peuvent aussi être dues à des conditions rares de l'environnement lors du recueil des données. C'est pour cette dernière raison que nous avons choisi de les exclure du jeu de données. Au total 4 observations ont été exclues car leurs résidus avoisinaient 3 ou -4, des valeurs trop extrêmes si l'on suppose que les résidus sont distribués normalement. N'exclure que l'observation donc le résidu avoisine 4 n'a pas permis d'obtenir d'aussi bons résultats que lorsque nous avons exclu les 4 observations comme expliqué plus tôt.

Le modèle de régression multiple calculé à partir du nouveau jeu de données obtenu comporte les variables suivantes : maxO3 ~ Vx12 + Ne9 + maxO3v + T12. Ce modèle explique cette fois-ci 82.7% de la variance de la concentration en ozone, un gain de respectivement 6.7% et 7.7% par rapport aux deux précédents modèles. De plus, le critère AIC pour ce modèle est nettement inférieur à ceux des précédents modèles (527.77).

Concernant les résidus de ce modèle, on ne discerne plus de structure anormale, que ça soit dans l'histogramme des résidus ou bien lorsque l'on représente les résidus en fonction des valeurs prédites par le modèle (voir Figure 7). Nous gardons donc ce modèle à 4 variables explicatives comme étant le meilleur modèle pour prédire la concentration maximale journalière en ozone. Nous noterons toutefois qu'un modèle similaire est obtenu autant en utilisant une méthode ascendante qu'une méthode descendante.

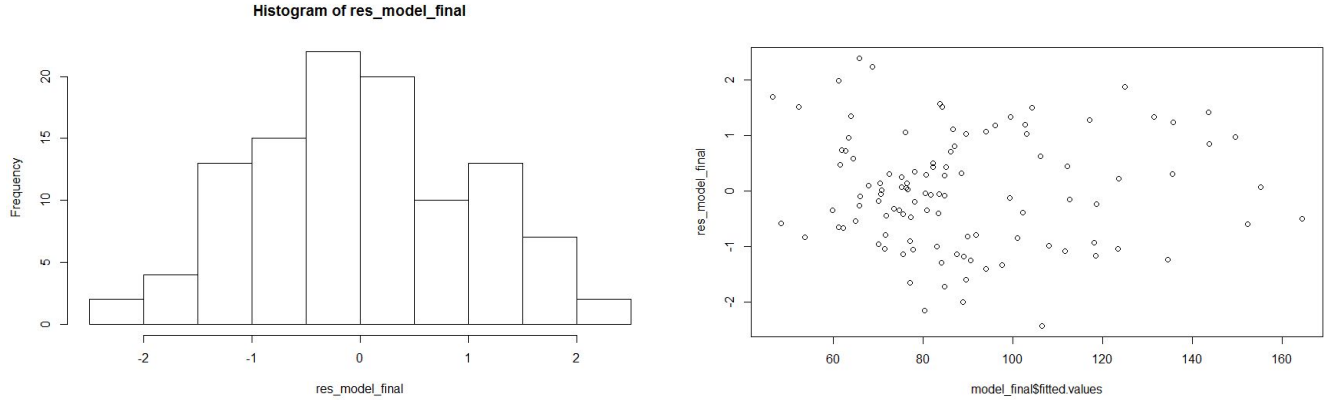


Figure 7 : Histogramme des résidus du modèle final (à gauche), et figure des résidus du modèle final en fonction des valeurs prédites par le modèle (à droite). La structure des résidus semble normale

II. Exercice 2 - Station

II. 1. Analyse du jeu de données

Présentation rapide

Le jeu de données recense les caractéristiques de 45 stations de service situées dans les grands centres urbains.

Pour chaque station il est indiqué :

- les ventes de la station exprimées en milliers de litres (variable ventes comprise entre 203 et 282);
- le nombre de pompes de la station (variable nbpompes comprise entre 3 et 21);
- le nombre de concurrents dans la zone desservie par la station (variable nbconc comprise entre 4 et 23);
- le trafic quotidien exprimé en milliers de voitures (variable trafic comprise entre 8 et 28).

```
mat<-data.frame(station$ventes,station$nbpompes,station$nbconc,station$trafic)
plot(mat)
```

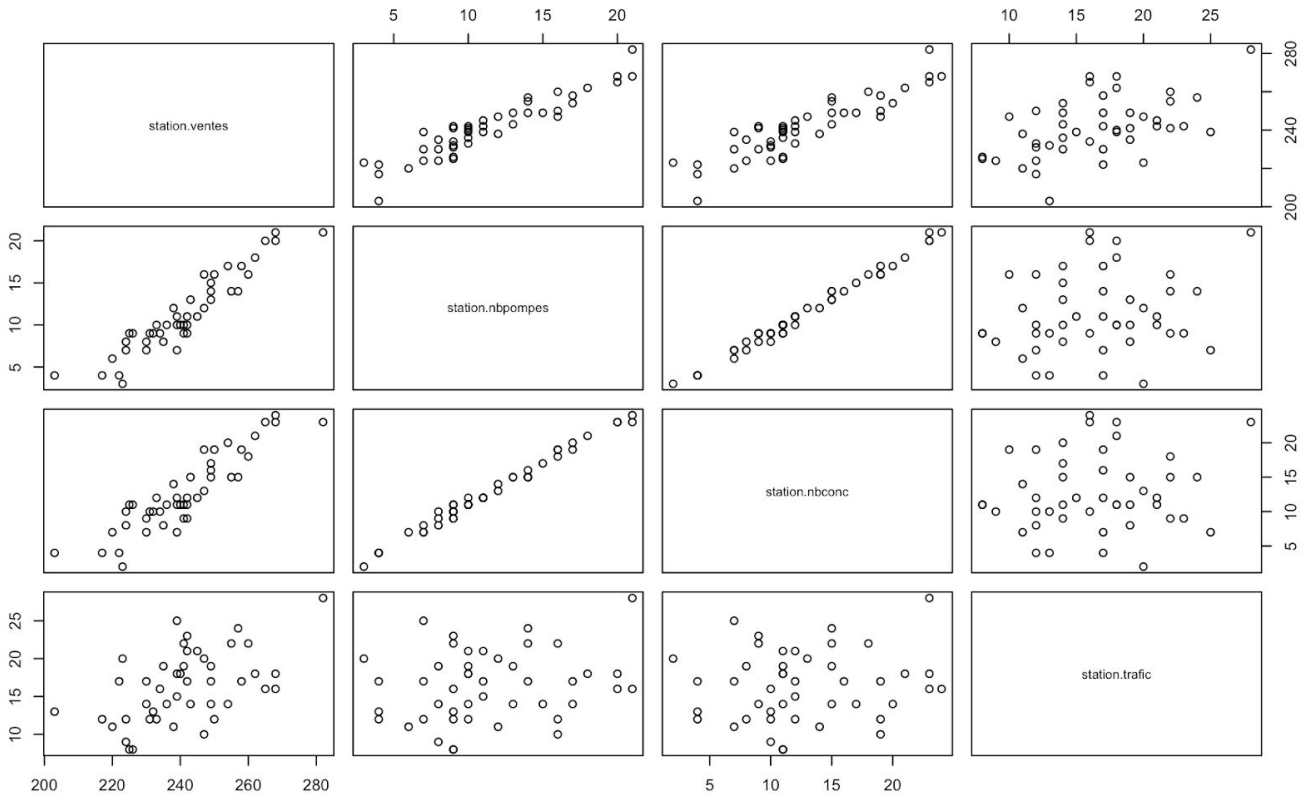


Figure 8 : Représentation graphique 2 à 2 des variables du jeu de données Station. Il semble y avoir une relation linéaire entre : ventes et nbpompes, nbconc et nbpompes, ventes et nbconc.

Analyse en Composantes Principales (ACP)

Nous effectuons une ACP des quatre variables quantitatives du jeu de données. L'espace des variables est donc $\mathbb{R}4$. L'espace des individus est $\mathbb{R}45$ (il y a 45 stations). L'ACP permet d'étudier les éventuelles corrélations entre les variables et d'en tirer des conclusions pour mettre en œuvre un modèle linéaire satisfaisant par la suite.

Après utilisation des commandes res<-PCAmix(station) et res$eig on observe que seule la première valeur propre est strictement supérieure à 1. D'après le critère de Kaiser (valable en ACP centrée réduite), on ne retiendrait donc qu'un seul axe factoriel, qui permet d'expliquer 75,2% de l'inertie. Cependant, la seconde valeur propre est très proche de 1 (elle vaut 0.97) et inclure le second axe permet d'expliquer 99,6% de l'inertie. Observons le cercle des corrélations du plan 1-2 à l'aide de la commande plot(res,choice="cor") (Figure 9). L'ACP nécessite l'installation du package PCAmixdata : if(!require("PCAmixdata")) install.packages("PCAmixdata").

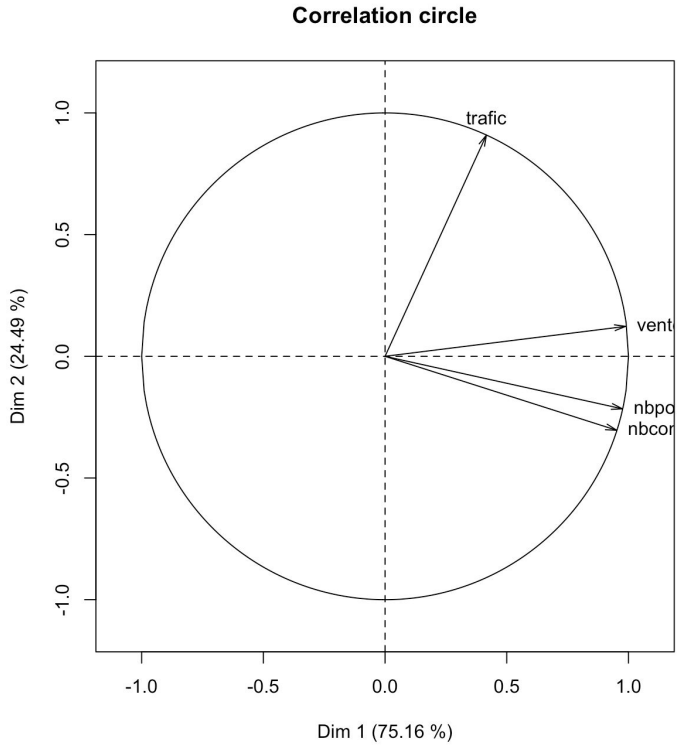


Figure 9 : cercle de corrélation des variables quantitatives du jeu de données station (plan 1-2).

L'angle formé par les variables (nbpomp, nbconc) est faible, ce qui signifie que ces deux variables sont fortement corrélées positivement entre elles (ainsi par exemple une station qui possède un grand nombre de pompes a plutôt tendance à avoir aussi un grand nombre de concurrents). Ces variables sont elles-même bien corrélées à la variable ventes (angle faible). Il semble donc que nbpompes et nbconc apportent la même information. Il sera certainement pertinent d'écarter l'une de ces deux variables du modèle pour éviter une redondance d'information. La variable trafic est très éloignée de nbconc et nbpompes, elle semble donc non corrélée à ces dernières car dépend plutôt de l'axe 2 et non l'axe 1.

Question 1

Expliquer pourquoi le modèle à trois variables ne peut être retenu. Quelle variable faut-il éliminer de ce modèle pour obtenir un modèle à deux variables qui semble mieux convenir ? Justifier votre choix.

```
modele3V<-lm(ventes~nbpompes+nbconc+trafic, data = station)
summary(modele3V)
```

```
Call:
lm(formula = ventes ~ nbpompes + nbconc + trafic, data = station)

Residuals:
     Min       1Q   Median       3Q      Max
-13.1412  -0.2876   0.1360   0.7434   2.0179

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 189.7673     1.6530 114.804  < 2e-16 ***
nbpompes      2.5507     1.3888   1.837   0.0735 .
nbconc        0.2755     1.1504   0.239   0.8119
trafic        1.1592     0.1464   7.920 8.55e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.213 on 41 degrees of freedom
Multiple R-squared:  0.9811,    Adjusted R-squared:  0.9797
F-statistic: 709.1 on 3 and 41 DF,  p-value: < 2.2e-16
```

Figure 10 : Modèle linéaire à 3 variables.

```
par(mfrow=c(1,2))
plot(modele3V$fitted,modele3V$residuals)
abline(h=0,col=2)
```

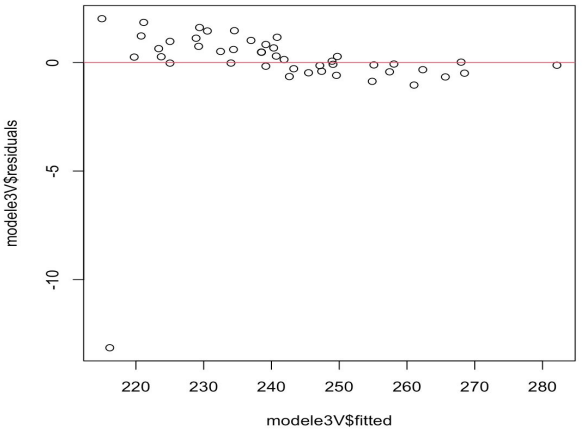


Figure 11 : Représentation des résidus du modèle linéaire à 3 variables en fonction des valeurs prédites.

On obtient une p-value très inférieure à 0.05, ce modèle est donc significativement utile. Le R carré ajusté nous indique que 97,97% de la variance est expliquée. Les coefficients de nbpompes et nbconc ne sont pas significatifs cependant. Ce modèle semble satisfaisant mais cela reste à vérifier avec l'étude des résidus. Grâce à la Figure 11, on observe une structure dans les résidus, le modèle est donc insatisfaisant car l'hypothèse de normalité des résidus n'est pas respectée ici.

Compte tenu du fait que deux des coefficients du modèle ne sont pas significativement différent de zéros, et donc qu'ils ne sont vraisemblablement pas utiles au modèle, nous allons essayer de voir s'il existe un modèle plus satisfaisant que le modèle initial à 3 variables que nous avons construit. En utilisant la fonction drop1(), on se rend compte que l'on pourrait éjecter de notre modèle la variable nbconc car cela nous permettrait d'obtenir un modèle de régression avec un critère AIC plus faible que le modèle actuel (Figure 10). Enlever la variable nbconc de notre modèle paraît logique, car lorsque l'on regarde la représentation graphique de la variable nbpompes en fonction de nbconc (Figure 8), on se rend compte que ces variables sont extrêmement corrélées linéairement entre elles. Elles semblent donc apporter la même information toutes les deux pour le modèle, ce qui n'est pas utile.

```
Single term deletions

Model:
ventes ~ nbpompes + nbconc + trafic
         Df Sum of Sq    RSS     AIC
<none>                  200.74  75.290
nbpompes  1    16.515 217.25  76.848
nbconc    1     0.281 201.02  73.353
trafic    1   307.125 507.86 115.060
```

Figure 12 : Résultat de la commande drop1(modele3V). La variable à retirer est celle dont l'AIC est le moins élevé, il s'agit donc de la variable nbconc.

Question 2.a

Faire une analyse des indices de qualité du modèle à deux variables obtenu à la question précédente et en tirer une conclusion quant à sa validité statistique.

```
modele2V<-lm(ventes~nbpompes+trafic, data = station)
summary(modele2V)
plot(modele2V$fitted,modele2V$residuals)
```

```
par(mfrow=c(1,2))
abline(h=0,col=2)
```

```
Call:
lm(formula = ventes ~ nbpompes + trafic, data = station)

Residuals:
     Min       1Q   Median       3Q
-13.2011  -0.2058   0.0961   0.7826
     Max
  1.9277

Coefficients:
             Estimate Std. Error
(Intercept) 189.99478    1.33748
nbpompes      2.88289    0.07260
trafic        1.12883    0.07258
            t value Pr(>|t|)
(Intercept)  142.06   <2e-16 ***
nbpompes      39.71   <2e-16 ***
trafic        15.55   <2e-16 ***
---
Signif. codes:
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05
  ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.188 on 42 degrees of freedom
Multiple R-squared:  0.9811,    Adjusted R-squared:  0.9802
F-statistic:  1088 on 2 and 42 DF,  p-value: < 2.2e-16
```

Figure 13 : Modèle linéaire à 2 variables.

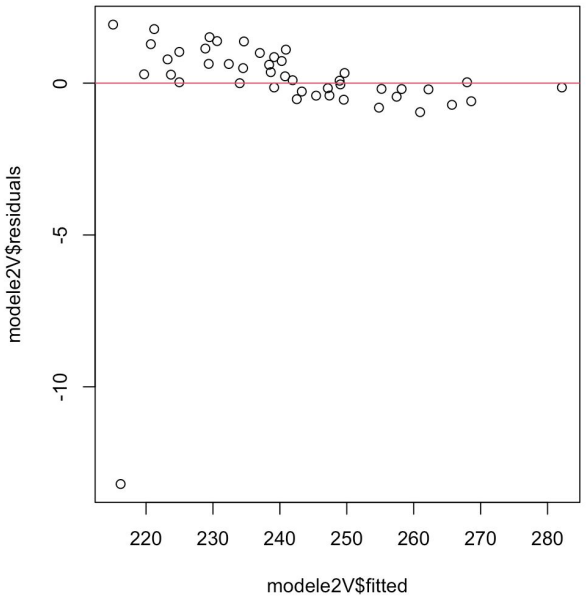


Figure 14 : Représentation des résidus du modèle linéaire à 2 variables en fonction des valeurs prédites.

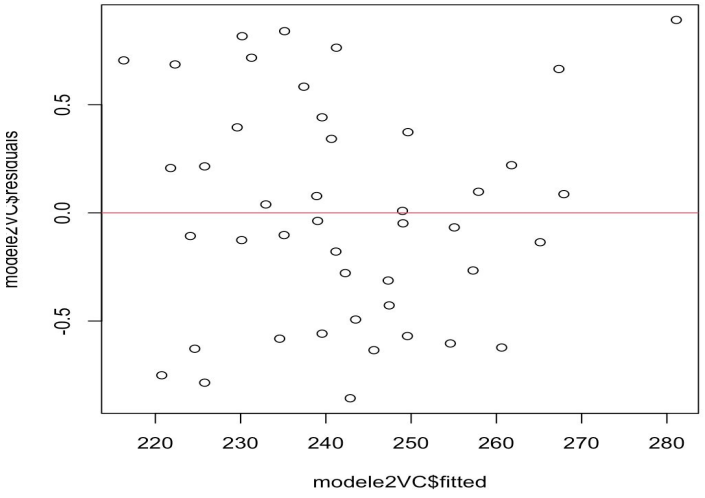
On obtient une p-value très inférieure à 0.05 pour la pertinence du modèle, nous sommes donc en présence d'un modèle utile. Le R carré ajusté nous indique que 98,02% de la variance est expliquée, ce qui veut dire que notre modèle explique très bien les données. Les coefficients de nbpompes et trafic ont une p-value très inférieure à 0.05, ce sont donc 2 variables utiles pour le modèle car leur coefficient est significativement différent de zéro. Ce modèle semble satisfaisant mais cela reste à vérifier avec l'étude des résidus.

Grâce à la Figure 14, on observe une structure dans les résidus. C'est un problème car l'hypothèse de normalité des résidus n'est pas respectée. On remarque cependant qu'une valeur s'écarte énormément des autres, on peut supposer que la retirer du jeu de données règlera le problème. En effet, les modèles de régression multiples sont sensibles aux valeurs extrêmes, par conséquent exclure cette valeur peut être pertinent dans notre cas.

```
station2<-station[-1,]
modele2VC<-lm(ventes~nbpompes+trafic, data = station2)
plot(modele2VC$fitted,modele2VC$residuals)
summary(modele2VC)
```

```
par(mfrow=c(1,2))

abline(h=0,col=2)
```

```
Call:
lm(formula = ventes ~ nbpompes + trafic, data = station2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85703 -0.44413 -0.04286  0.37860  0.89185

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 191.98974    0.32530  590.20   <2e-16 ***
nbpompes      2.77766    0.01763  157.53   <2e-16 ***
trafic        1.09956    0.01721   63.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5178 on 41 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9987
F-statistic: 1.699e+04 on 2 and 41 DF,  p-value: < 2.2e-16
```



<u>Figure 15 :</u> modèle linéaire à 2 variables avec jeu de données modifié.

<u>Figure 16 :</u> Figure des résidus du modèle linéaire à 2 variables (jeu de données corrigé) à 2 variables en fonction des valeurs prédites.

En retirant la première observation du jeu de données, celle qui pose problème, on obtient alors un nouveau jeu de données : station2. En calculant à nouveau notre modèle de régression à deux variables sur ce jeu de données, on obtient une p-value très inférieure à 0.05 pour la pertinence du modèle (Figure 15). Le R carré ajusté nous indique que 99,87% de la variance est expliquée, ce qui est très satisfaisant car ce modèle explique encore mieux la variance des données que ceux précédemment réalisés. Les coefficients de nbpompes et trafic ont une p-value très inférieure à 0.05, ce sont donc 2 variables utiles pour le modèle. On n'observe pas de structure dans les résidus (Figure 16). L'hypothèse de normalité des résidus étant respectée, nous pouvons considérer que ce modèle est satisfaisant.

Question 2.b
Faire une interprétation des coefficients de ce modèle. Que pensez-vous de la validité économique du modèle ? Permet-il de mieux comprendre la réalité ?

Interprétons les coefficients du modele2VC (modèle à 2 variables corrigé, on a supprimé la première observation). On obtient : ventes = 191.98974 + 2.77766*nbpompes + 1.09956*trafic + 0.34

Avec 0.34 comme terme d'erreur.
2.77766 s'interprète comme l'effet marginal d'une pompe supplémentaire sur les ventes. Elle correspond à la variation de 2.77766 unités de ventes de la station induite par la variation d'une unité du nombre de pompes toutes choses égales par ailleurs, c'est-à-dire en prenant en compte le trafic de la station[1].

Le même raisonnement s'applique pour 1.09956. D'après ce modèle, le nombre de ventes d'une station dépend bien plus du nombre de pompes qu'elle possède plutôt que du trafic. En effet, le coefficient de nbpompes et 2,7 fois plus élevé que celui de trafic. La validité économique du modèle semble correcte. Dans la réalité, les deux variables sont liées, une station qui possède un grand nombre de pompes sera plus susceptible d'avoir un

[1] https://www.parisschoolofeconomics.eu/docs/yin-remi/interpretation-des-coefficients.pdf

trafic élevé. En effet, un plus grand nombre de véhicules peuvent faire le plein en même temps donc les bénéfices augmentent. On suppose que cela n'est vrai que jusqu'à un certain nombre de pompes. En effet, il doit y avoir un nombre de pompes à partir duquel le nombre de ventes n'augmente plus dans la réalité. Cependant nous ne pouvons pas vérifier cette hypothèse avec notre modèle et avec ces données.

Question 2.c

Préciser ce que l'on prédirait comme nombre de ventes étant donné un nouveau couple de variables explicatives.

```
nbpompesP<-seq(from=3,to=21,by=1)
traficP<-seq(from=8,to=28,by=1)

Vmoy<-predict(modele2VC,data.frame(nbpompes=nbpompesP, trafic=traficP),interval = "conf", level = 0.95)
round (cbind(nbpompesP,traficP,Vmoy),digit=1)
```

nbpompesP	traficP	fit	lwr	upr
3	8	259.1	258.6	259.6
4	9	260.2	259.7	260.7
5	10	261.3	260.8	261.8
6	11	262.4	262.0	262.9
7	12	263.5	263.1	263.9
8	13	264.6	264.2	265.0
9	14	265.7	265.3	266.1
10	15	266.8	266.4	267.2

Figure 17 : prédiction sur les ventes moyennes avec un intervalle de confiance à 95%.

Après l'affichage du résultat R met en garde quant aux tailles différentes des deux vecteurs nbpompesP et traficP. Nous choisissons ces valeurs puisqu'elles reprennent l'ensemble des valeurs prises par nbPompes et trafic dans le jeu de données. Il est possible de fixer une valeur et d'observer comment se comporter l'autre. Par exemple, remplaçons `traficP<-seq(from=8,to=28,by=1)` par `traficP<-8`. On étudie alors l'évolution de la moyenne des ventes en fonction du nombre de pompes lorsque le trafic est fixé à 8000 voitures par jour. On observe que les moyennes augmentent plus rapidement lorsque le trafic est fixé et que le nombre de pompes augmente.

Question 3

Justifier pourquoi le modèle à deux variables défini à la première question et étudié à la deuxième est-il préférable aux modèles à une seule variable ?

On retient les modèles qui ont le R-squared le plus élevé. Après comparaison, on observe que ce sont les modèles à 2 variables par rapport aux modèles à 1 variable qui possèdent un R^2 le plus élevé.

De plus, nbconc est fortement corrélée positivement avec nbpompes (voir Figures 8 et Figure 9). Ainsi, le nombre de concurrents apporte peu d'informations contrairement à la variable trafic qui apporte vraisemblablement des informations sur la dispersion. La variable trafic va permettre de mieux prédire la vente, puisqu'elle a un impact dessus. Si on la retire du modèle, on perd de l'information pour la prédiction.

Annexes : Code de l’exercice 1

```
library("PCAmixdata")
library("ggplot2")

oz = read.table(file="./data/ozone.txt")
summary(oz)
plot(oz) #plot pour la figure 1, représentation des variables deux à deux

split = splitmix(oz)
oz_quali = split$X.quali
oz_quanti = split$X.quanti

# acp quantitative ozone ----------------------------------------------------

acp_quanti = PCAmix(X.quanti=oz_quanti, ndim=5)
summary(acp_quanti)
plot(acp_quanti, choice="cor", main="ACP Quanti") #cercle des corrélations

# impact vent et pluie sur maxO3 --------------------------------------------------

plot(oz$vent, oz$maxO3, xlab="Variable Vent", ylab="maxO3")
plot(oz$pluie, oz$maxO3, xlab="Variable Pluie", ylab="maxO3")

# acp mix ozone -------------------------------------------------------------

acp_quali = PCAmix(X.quanti = oz_quanti, X.quali = oz_quali, ndim=5)
summary(acp_quali)
plot(acp_quali, choice="cor")
plot(acp_quali, choice="sqload")

# modèle linéaire initial ozone ---------------------------------------------------

model_initial = lm(formula=maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v, data=oz)
summary(model_initial)
residus_initiaux_stud = rstudent(model_initial)
hist(residus_initiaux_stud)
plot(model_initial$fitted, residus_initiaux_stud, ylab="résidus studentisés", xlab="valeurs prédites") #résidus
studentisés

# meilleur modèle sans résidus mauvais ----------------------------------

cond = residus_initiaux_stud > -2 & residus_initiaux_stud < 2
new_oz = oz[cond,] #nouveau df sans les valeurs extrêmes

model_initial = lm(formula=maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v, data=new_oz)
new_model = lm(maxO3~T15+Vx12+Ne12s, data=new_oz)
plot(new_model$fitted, rstudent(new_model)) #plot des résidus studentisés
step_model = step(new_model, ~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v)
model_final = lm(step_model$call, data = new_oz)
summary(model_final)
hist(res_model_final)
plot(model_final$fitted.values, rstudent(model_final))
```