

Rapport

Contrôle continu sur la régression linéaire



Présenté par :
de Foucaud Tess, Grondin Léa, Lucas Maël, Pelletreau- Duris Tom

Année universitaire 2020 – 2021
Professeur : M. Saracco Jérôme

TABLE DES MATIÈRES

I. Ozone	1
1. ACP des variables quantitatives	1
2. Description de l'impact des valeurs qualitatives sur maxO ₃	2
3. ACP mixte	3
4. Estimation du meilleur modèle	4
5. Conclusion	5
II. Station	6
1. Présentation théorique du jeu de données	6
2. Analyse descriptive du jeu de données	6
3. Estimation du meilleur modèle	7
4. Conclusion explicative	11
Annexes	12
Annexe 1 : Code R de l'exercice Ozone	12
Annexe 2 : Code R de l'exercice Station	12

I. Ozone

Le jeu de données Ozone représente les variables de température à 9h, 12h et 15h, correspondant respectivement à T9, T12 et T15, les variables de nébulosité aux mêmes horaires, correspondant à Ne9, Ne12 et Ne15, les variables de vent, aux mêmes horaires, Vx9, Vx12 et Vx15, ainsi que le maximum journalier de la concentration en ozone maxO3 et celui de la veille maxO3v.

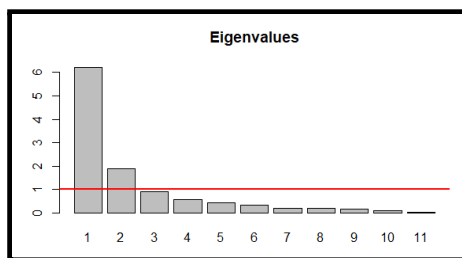
Nous cherchons à savoir s'il existe un lien entre maxO3 et les autres variables.

1. ACP des variables quantitatives

Avant de commencer par faire une ACP des valeurs quantitatives, c'est-à-dire T9, T12, T15, Vx9, Vx12, Vx15, Ne9, Ne12, Ne15, maxO3 et maxO3v, il faut vérifier que les valeurs des différentes variables sont normalement distribuées.

Ainsi, nous avons commencé par faire un test de Shapiro, qui nous a indiqué que toutes les colonnes associées aux variables quantitatives étaient normalement distribuées, sauf pour les variables de vent : Vx9, Vx12 et Vx15. Nous avons donc vérifié ces trois colonnes au moyen du test de Wilcoxon. Grâce à ce dernier, nous avons pu en déduire que l'ensemble des valeurs associées aux variables du jeu de données sont normalement distribuées.

Nous examinons les valeurs propres pour déterminer le nombre de composantes principales à prendre en considération et nous avons décidé de les afficher sous forme d'un graphique, présenté ci-dessous.



```
> barplot(res$eig[,1],main="Eigenvalues",names.arg=1:nrow(res$eig))
> abline(h=1,col=2,lwd=2)
```

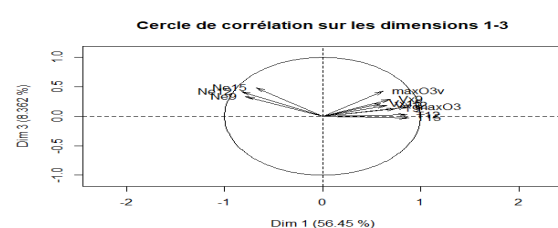
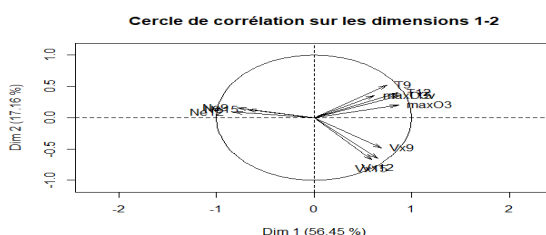
Ainsi, nous pouvons retenir les dimensions dont les valeurs propres (eigenvalues) sont supérieures à 1, c'est-à-dire celles au-dessus de l'axe horizontal rouge. Néanmoins, nous pouvons noter que la valeur propre de la dimension 3 est très proche de 1 et pourrait donc être retenue. Après avoir visualisé les cercles de corrélation et comparé la perte d'informations engendrée par l'une ou l'autre dimension, il semble que garder les dimensions 1 et 2 permette de visualiser un maximum l'information.

En comparant les deux cercles de corrélation, nous pouvons observer que la représentation des variables est plus qualitative pour les dimensions 1 et 2, car les flèches sont plus longues. Nous avons donc choisi ces dimensions pour la suite de l'analyse.

```
> plot(res,axes=c(1,2),choice="cor")#, main="Cercle de corrélation sur les dimensions 1-2")
```

Ainsi, d'après le cercle de corrélation des dimensions 1-2, nous pouvons tout d'abord observer une corrélation positive entre les variables de nébulosité Ne9, Ne12, Ne15 mais aussi une forte corrélation entre les variables de vent Vx9, Vx12 et Vx15 et une autre entre maxO3, maxO3v, et les variables de température T9, T12 et T15.

```
> plot(res,axes=c(1,2),choice="cor")#, main="Cercle de corrélation sur les dimensions 1-2")
> plot(res,axes=c(1,3),choice="cor")#, main="Cercle de corrélation sur les dimensions 1-3")
```



De plus, les variables négativement corrélées sont positionnées sur les côtés opposés de l'origine du graphique, c'est-à-dire dans des quadrants opposés. On peut donc en conclure que les variables de vent et les variables de nébulosité sont fortement corrélées négativement. A l'inverse, les variables maxO3, maxO3v, et celles de température T9, T12 et T15 ne sont pas du tout corrélées avec les variables de vent car les flèches forment sur le cercle un angle proche de 90°.

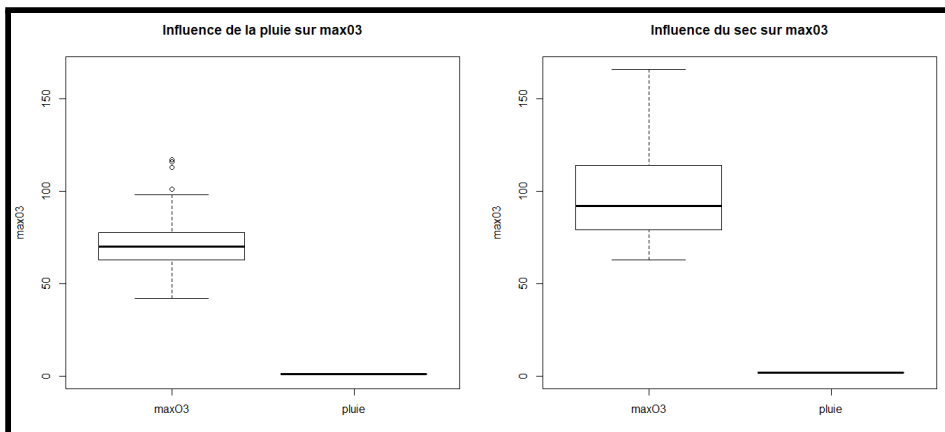
Finalement, d'après la distance des variables à l'origine, nous pouvons en déduire que les variables sont correctement représentées par l'ACP.

2. Description de l'impact des valeurs qualitatives sur maxO3

Dans cette partie, l'objectif est de déterminer les influences séparées des variables "pluie" et "vent" sur maxO3. Pour chaque modalité (c'est-à-dire valeur que peut prendre la variable qualitative), nous avons réalisé un boxplot et comparé ces différents boxplots.

Commençons par la variable "pluie". Celle-ci peut prendre 2 valeurs: Pluie et Sec. Nous avons construit les 2 boxplots suivants:

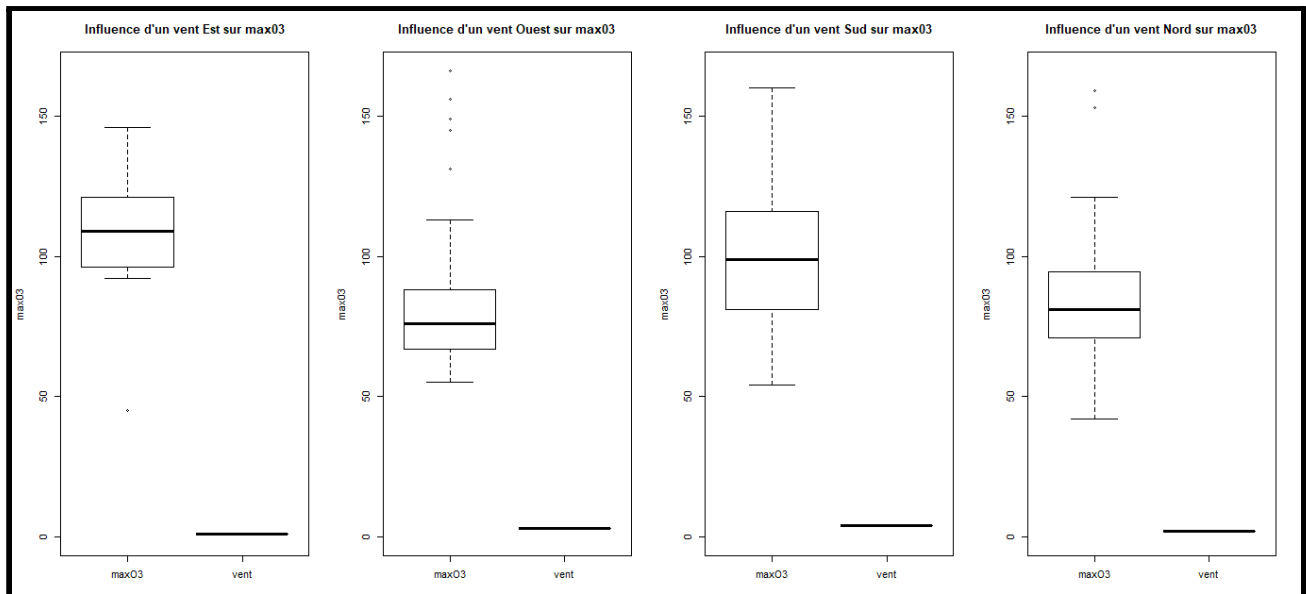
```
> boxplot(dataPluie, main="Influence de la pluie sur maxO3", ylab="maxO3", ylim=c(0,166))
> boxplot(dataSec, main="Influence du sec sur maxO3", ylab="maxO3", ylim=c(0,166))
```



La barre transversale du haut du boxplot de droite est au-dessus de celle de gauche, le maximum que peut prendre maxO3 en condition de sec est donc supérieur à celui qu'il peut prendre en temps de pluie. La moyenne de maxO3 est également légèrement supérieure en condition de sec. On en conclut qu'en moyenne, maxO3 prend de plus grandes valeurs en condition de sec qu'en condition de pluie.

Étudions ensuite la variable "vent". Celle-ci peut prendre 4 valeurs: ouest, nord, sud, est. Comme précédemment, nous avons construit 4 boxplots qui permettent de comparer l'influence de ces 4 valeurs sur la variable maxO3 :

```
> boxplot(dataEst, main="Influence d'un vent Est sur maxO3", ylab="maxO3", ylim=c(0,166))
> boxplot(dataOuest, main="Influence d'un vent Ouest sur maxO3", ylab="maxO3", ylim=c(0,166))
> boxplot(dataSud, main="Influence d'un vent Sud sur maxO3", ylab="maxO3", ylim=c(0,166))
> boxplot(dataNord, main="Influence d'un vent Nord sur maxO3", ylab="maxO3", ylim=c(0,166))
```



Sur le graphique ci-dessus, nous voyons qu'un vent Est est celui qui va en moyenne le plus faire augmenter max03: la barre dans le boxplot (donc la moyenne) est légèrement au-dessus de celles des autres boxplots. En moyenne, nous pouvons dire des vents Est et Sud qu'ils sont plus propices à une augmentation de max03 que des vents Nord et Ouest.

3. ACP mixte

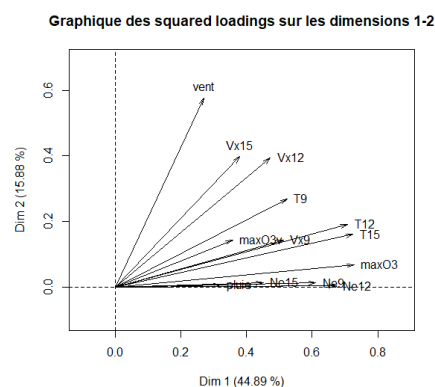
Afin de réaliser l'ACP mixte, nous avons regroupé les données en fonction de leur nature. Nous avons donc fait un tableau de données ne possédant que les valeurs quantitatives, *donneesQuanti* et un comportant les valeurs qualitatives, nommé *donneesQuali*.

Nous avons ensuite stocké l'ACP dans res2.

De la même manière que pour l'ACP classique, nous avons dû déterminer les axes à retenir.

Nous avons décidé de comparer les dimensions 1-2, 1-3 et 1-4 car leur valeur propre respective était de 6.76, 2.38, 1.20 et 1.04. Comme dans la question 2, il s'agit de déterminer quelles dimensions permettent la meilleure représentativité. Nous regardons donc les longueurs des flèches sur les graphiques des squared loadings:

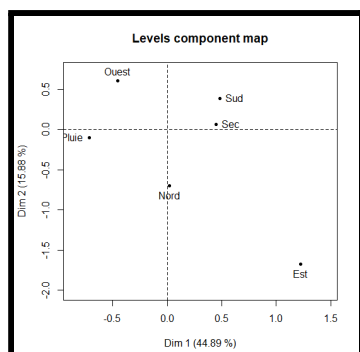
```
> plot(res2, axes=c(1,2), choice="sqload", main="Graphique des squared loadings sur les dimensions 1-2")
```



Nous avons donc décidé de choisir les dimensions 1 et 2, car la qualité de représentation des variables est la plus satisfaisante avec ces dernières. Toutefois il nous est impossible d'émettre des conclusions sur la variable pluie, car celle-ci n'est pas assez représentée.

En plus des conclusions émises à la question 2, il s'agit maintenant de comprendre les corrélations entre le vent et les autres variables. S'il est impossible de mettre toutes ces variables sur un même schéma nous pouvons cependant afficher les variables qualitatives sur un Levels Component Map.

```
> plot(res2, axes=c(1,2), choice="levels")
```



Nous voyons ici, par exemple, que la valeur “Nord” est très proche de l’axe décrivant la dimension 1. Par conséquent, les vents du Nord ont une forte corrélation avec des variables quantitatives bien représentées par l’axe 1: par exemple, maxO3. Toutefois il nous est impossible de dire si cette corrélation est positive ou négative. De la même façon, la valeur “Est” semble très corrélée avec les dimensions 1 et 2. Cependant il reste impossible d’interpréter les valeurs “Pluie” et “Sec” car elles sont mal représentées sur les dimensions 1 et 2 (cf Squared Loadings).

4. Estimation du meilleur modèle

Nous souhaitons savoir quel est le modèle qui explique le plus fidèlement les variations de la concentration maximum journalière en ozone, maxO3, en fonction des autres variables. Nous cherchons donc une relation linéaire entre maxO3 et les autres variables.

```
> res<-lm(maxO3~T9+T12+T15+Vx9+Vx12+Vx15+Ne9+Ne12+Ne15+maxO3v, data=dataRes)
> summary(res)
```

```
Call:
lm(formula = maxO3 ~ T9 + T12 + T15 + Vx9 + Vx12 + Vx15 + Ne9 +
    Ne12 + Ne15 + maxO3v, data = dataACP)

Residuals:
    Min       1Q   Median       3Q      Max
-53.566  -8.727  -0.403   7.599  39.458

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.24442   13.47190   0.909  0.3656
T9          -0.01901    1.12515  -0.017  0.9866
T12          2.22115    1.43294   1.550  0.1243
T15          0.55853    1.14464   0.488  0.6266
Vx9          0.94791    0.91228   1.039  0.3013
Vx12         0.03120    1.05523   0.030  0.9765
Vx15         0.41859    0.91568   0.457  0.6486
Ne9         -2.18909    0.93824  -2.333  0.0216 *
Ne12        -0.42102    1.36766  -0.308  0.7588
Ne15         0.18373    1.00279   0.183  0.8550
maxO3v       0.35198    0.06289   5.597 1.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.36 on 101 degrees of freedom
Multiple R-squared:  0.7638,    Adjusted R-squared:  0.7405
F-statistic: 32.67 on 10 and 101 DF,  p-value: < 2.2e-16
```

Nous avons donc commencé par effectuer un modèle de régression linéaire avec toutes les variables.

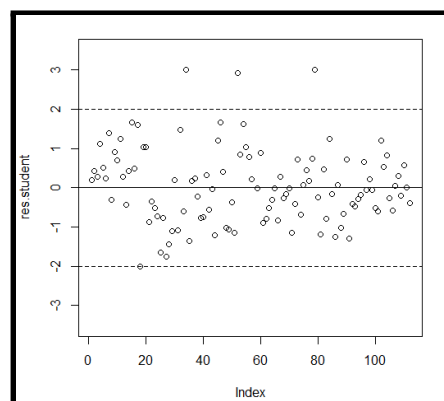
Quand nous avons construit ce modèle, il a été nécessaire de vérifier la normalité des échantillons. En faisant un shapiro.test, nous obtenons une p-value inférieure de 5%: par conséquent, les données suivent une loi normalement distribuée.

Cette normalité est nécessaire pour appliquer l’ensemble des modèles algébriques dans la suite des calculs. Toutefois, nous avons cherché à visualiser ces échantillons normalisés.

Après avoir studentisé l’échantillon, nous obtenons un graphique de la forme suivante:

```
> residus.stud<-rstudent(res)
> plot(residus.stud,ylab="res.student",ylim=c(-3.5,3.5))
> abline(h=c(-2,0,2),lty=c(2,1,2))
```

Or 3 valeurs dépassent largement du cadre [-2,2]. Nous prenons cette tranche car un échantillon dépassant de ce cadre est considéré comme anormal. Or, ces données correspondent au 7 juillet, au 25 juillet et au 25 août: ce sont



des jours de départ en vacances. Ce sont des valeurs qui perturbent le modèle et que l'on peut raisonnablement enlever puisque cette anomalie est expliquée par le fait que la pollution augmente fortement lors des jours de départs en vacances.

Après avoir enlevé ces valeurs, nous créons à nouveau un modèle qui a cette fois une multiple R-squared de 0,7799. Cette dernière valeur représente le pourcentage de variabilité de maxO3 expliqué par le modèle: on veut une valeur qui soit la plus forte possible.

Cependant, il semble que certaines variables présentent une p-value (colonne de droite) largement supérieure à 5%. Par conséquent, nous avons enlevé les variables (une par une) qui n'étaient pas pertinentes pour ce modèle par rapport au critère de la p-value. De cette façon, nous obtenons le modèle suivant:

```
> res<-lm(maxO3~T12+Vx9+Ne9+maxO3v, data=dataRes)
> summary(res)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.47900    10.49963   1.379   0.1709
T12           2.64877     0.45401   5.834 6.17e-08 ***
Vx9           1.28855     0.56424   2.284  0.0244 *
Ne9          -2.68106     0.63153  -4.245 4.76e-05 ***
maxO3v        0.36309     0.05502   6.599 1.78e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.06 on 104 degrees of freedom
Multiple R-squared:  0.7765,    Adjusted R-squared:  0.7679
F-statistic: 90.31 on 4 and 104 DF,  p-value: < 2.2e-16
```

Plusieurs commentaires peuvent être faits. D'abord, la valeur de "Multiple R-Squared" a baissé. Avant d'ôter des variables, cette valeur est de 0.7799. Si cette valeur n'est pas très grande, nous appréhendons ici des phénomènes météorologiques et 77,99% semble être un chiffre satisfaisant. Toutefois sur la deuxième image ce chiffre a baissé jusqu'à 77,74 %, en ayant enlevé quelques variables. Nous avons fait ce choix afin de respecter le principe de parcimonie: il faut un équilibre entre le nombre de variables du modèle et le Multiple R-Squared. Ici, baisser ce dernier de 0.25% pour alléger le modèle de 7 variables nous paraissait être un excellent compromis.

Conclusion

Nos ACP ont permis de représenter correctement l'information, à part en ce qui concerne la variable Pluie. Ainsi nous avons pu montrer des corrélations entre les variables (questions 1 et 3) et maxO3. Pour avoir une meilleure visibilité de l'influence des variables qualitatives sur maxO3, il était nécessaire de les prendre à part en faisant des boxplots (question 2). Enfin, nous avons construit un modèle qui permet d'expliquer à 76% environ la variabilité de maxO3 en fonction de 4 variables (maxO3v, Vx9, Ne9, T12). Ainsi, nous avons une explication satisfaisante de la variabilité de la concentration en ozone, et donc de la pollution, en fonction de quelques paramètres.

II. Station

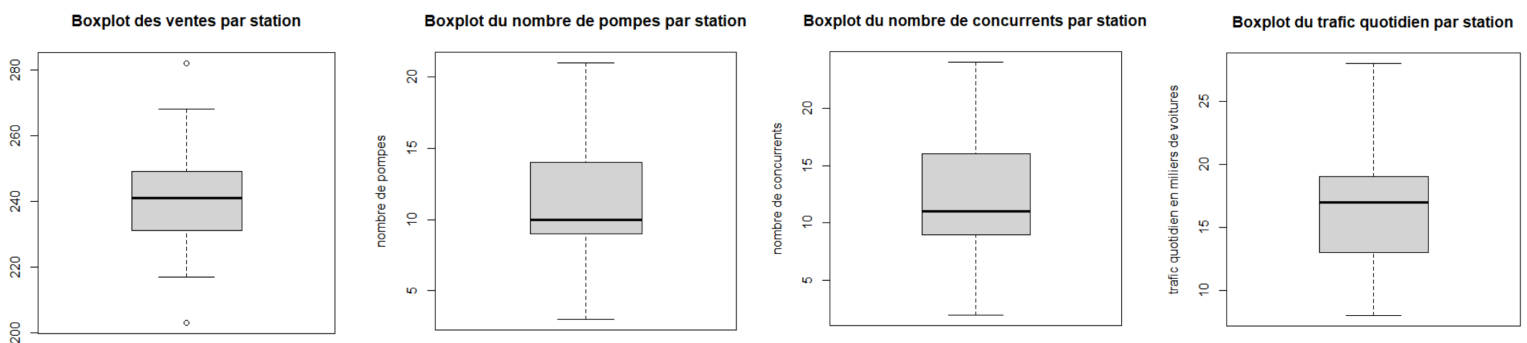
1. Présentation théorique du jeu de données

L'objectif de cet exercice est d'établir un modèle expliquant les ventes des stations services d'un distributeur d'essence dans les grands centres urbains. A partir de ce modèle, nous allons pouvoir expliquer les variations des ventes en fonction d'autres données qui l'influencent.

Le jeu de donnée étudié recense, pour 45 stations, différentes variables :

- ventes : les ventes de la station (en milliers de litres)
- nbpompes : le nombre de pompes de la station
- nbconc : le nombre de concurrents dans la zone desservie par la station
- trafic : le trafic quotidien (en milliers de voitures)

Toutes ces variables sont quantitatives. Voici leur représentation sous forme de boxplots :



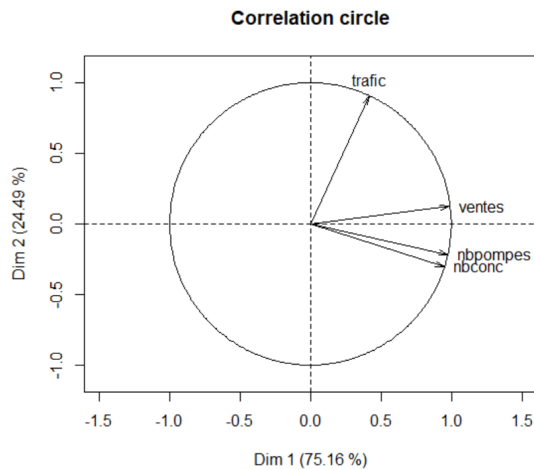
L'objectif de l'étude est donc d'essayer de modéliser les ventes d'une station en fonction des autres variables disponibles, en étudiant toutes les régressions linéaires possibles (simples ou multiples, à une, deux ou trois variables explicatives)

2. Analyse descriptive du jeu de données

Afin d'analyser et de visualiser le jeu de données, nous avons réalisé une ACP sur les quatres variables quantitatives.

D'après le critère de Kaiser, il ne faudrait retenir que les axes associés à des valeurs propre supérieures à 1. Lorsqu'on étudie les eigenvalues, on remarque que seule la dimension 1 possède une valeur propre supérieur à 1. Cependant, la seconde valeur propre étant très proche de 1 (0.98), nous décidons d'inclure également le second axe. Ainsi, le premier axe permet d'expliquer 75,2% de l'inertie, tandis que le 2eme permet d'expliquer 99,6% de l'inertie. Nous étudions donc le cercle des corrélations sur les dimensions 1 et 2.

	Eigenvalue	Proportion	Cumulative
dim 1	3.006331861	75.15829651	75.15830
dim 2	0.979630649	24.49076623	99.64906
dim 3	0.012707955	0.31769888	99.96676
dim 4	0.001329535	0.03323838	100.00000



Celui-ci nous apporte différentes informations quant à la corrélation des variables. Plus l'angle entre les flèches est faible, plus les variables sont corrélées. On observe que les variables nbpompes et nbconc sont très fortement corrélées entre elles, et sont également nettement corrélées avec "ventes". On peut donc supposer qu'il y a une redondance d'information entre nbpompes et nbconc. L'une de ces deux variables sera probablement à éliminer.

En ce qui concerne "trafic", on remarque qu'il est indépendant de nbpompes et nbconc. Bien que relativement faiblement corrélé à ventes, on peut donc supposer qu'il apportera tout de même des informations supplémentaires intéressantes.

3. Estimation du meilleur modèle

a. Modèle à 3 variables

Le modèle de régression linéaire avec les trois variables nous donne cela :

```
#Modèle de régression linéaire à trois variables
res <- lm(station$ventes~station$trafic+station$nbconc+station$nbpompes, data=matYX)
summary(res)
```

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

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 ***
station$trafic    1.1592     0.1464   7.920 8.55e-10 ***
station$nbconc    0.2755     1.1504    0.239  0.8119
station$nbpompes  2.5507     1.3888    1.837  0.0735 .
---
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
```

On étudie les caractéristiques de ce modèle.

Tout d'abord, les caractéristiques générales montrent une bonne représentativité. On définit H_0 comme l'hypothèse d'absence de significativité globale des variables, c'est à dire qu'au moins une variable n'est pas significativement différente de zéro. On cherche donc à rejeter H_0 . Pour cela il suffit de calculer le F-test, situé sur la dernière ligne de la sortie R. La p-value exprimée sur cette ligne est très faible (inférieure à $2.2 \cdot 10^{-16}$) ce qui se traduit par un fort rejet de H_0 . Le modèle est donc globalement significatif.

De plus, le R^2 ajusté (Adjusted R-squared) calcule la part de variabilité de la variable étudiée qui est expliquée par les variations des variables explicatives. Il est ici de 98% : la variabilité de “ventes” est expliquée à 98% par les variations des autres variables, ce qui est très encourageant.

Dans un deuxième temps, on étudie les caractéristiques des coefficients.

Lorsque l’on cherche à étudier les coefficients, il apparaît qu’ils ne sont pas tous significatifs. Quand on regarde les tests de Student associés aux variables, on cherche à avoir une t-value significativement différente de 0, ou une p-value significativement inférieure à un seuil généralement posé à 5%. Ici, la variable “trafic” est bien significative puisque sa p-value est très faible (de l’ordre de 10^{-10}). Par contre, la variable “nbconc” a une t-value très proche de 0 et une p-value de 81% ce qui est très mauvais. Cela signifie que la variable confirme fortement notre hypothèse H_0 , elle n’est donc pas significative. Dans une moindre mesure, “nbpompes” a une p-value proche de 5% (7%), bien que légèrement supérieure, nous pouvons considérer qu’elle reste significative à 90% grâce au code de significativité.

Dans cette logique, il n’est pas possible de garder notre modèle à trois variables. Si globalement il est très heuristique et remplit nos critères de représentativité, les variables choisies ne sont pas toutes significatives. La variable “nbconc” notamment risque de nuire fortement à la qualité de notre modèle au vu de sa non-représentativité. Il semble donc qu’un modèle à deux variables, contenant “trafic” et “nbpompes” serait plus approprié.

Afin de s’assurer de la validité de notre modèle, il faut également étudier les résidus.

On cherche à s’assurer qu’avec notre modèle,

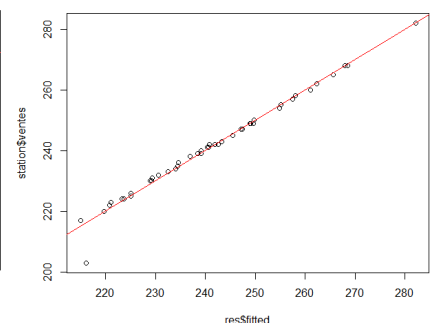
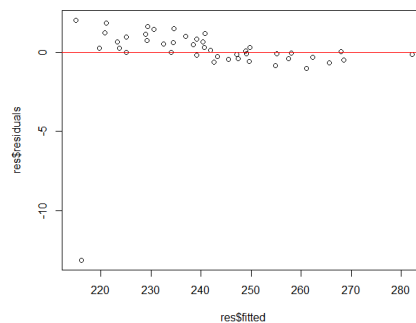
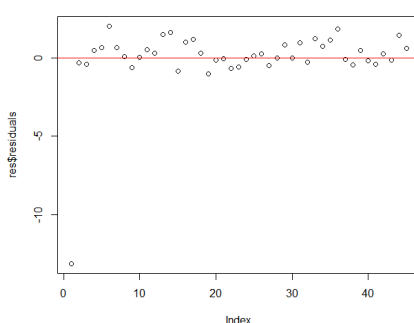
- la moyenne des résidus vaut zéro, c’est à dire qu’en moyenne notre modèle fonctionne bien (normalité des résidus),
- la variance des résidus vaut σ^2 , c’est à dire que les résidus sont bien un bruit et qu’il ne sont pas structurés (homoscédasticité)
- la covariance des résidus entre eux vaut zéro, c’est à dire qu’il n’y a pas de corrélations entre les mesures (indépendance).

On peut représenter graphiquement les résidus, afin d’observer si ces conditions sont (visuellement) respectées. Par exemple, on peut tracer : les résidus en ordonnées (1), résidus en ordonnées, valeurs prédites en abscisse (2), variables expliquées en ordonnées, variables prédites en abscisse (3).

```
#Etude des residus  
plot(res$residuals)  
abline(h=0,col=2)
```

```
plot(res$fitted,res$residuals)  
abline(h=0,col=2)
```

```
plot(res$fitted,station$ventes)  
abline(0,1,col=2)
```



On observe que les résidus semblent normalement distribués, qu'il n'y a pas de structure particulière et qu'ils sont indépendants entre eux. Ceci à l'exception d'une valeur aberrante. En effet, la première occurrence de nos données montre un écart au modèle de -13,14.

```
> res$residuals
 1      2      3      4
-13.14124168 -0.33098684 -0.40527690 0.48121228
 5      6      7      8
 0.67085485 2.01790987 0.63936636 0.05986721
 9     10     11     12
-0.64744825 0.01651561 0.50941417 0.27642000
13     14     15     16
 1.46661260 1.61116447 -0.86813187 1.01606817
17     18     19     20
 1.16290918 0.29515051 -1.03959397 -0.12574713
21     22     23     24
-0.07008499 -0.66518129 -0.59133867 -0.08339300
25     26     27     28
 0.13599896 0.24906234 -0.49143006 -0.02773086
29     30     31     32
 0.83000640 -0.02544173 0.97226914 -0.28763525
33     34     35     36
 1.22215212 0.74340580 1.11911015 1.84644778
37     38     39     40
-0.11159489 -0.47369702 0.47019340 -0.16999360
41     42     43     44
-0.42989799 0.26366202 -0.14079424 1.45201292
45
 0.59885393
```

On peut penser que cet écart est dû à un phénomène extérieur à nos variables explicatives. Cette influence exogène peut, par exemple, s'expliquer par des travaux dans la zone à proximité de la station n°1, ce qui occulterait sa visibilité et donc ses ventes. Ce pourrait aussi être une station où les clients, toutes choses égales par ailleurs, consomment de plus petites quantités d'essence, par exemple si le prix y est particulièrement cher relativement à d'autres endroits plus accessibles pour ces mêmes clients. C'est souvent le cas pour les stations de centre ville par exemple.

En conséquence, nous avons pris le parti de l'exclure du jeu de données. En temps normal, nous aurions pu en parler avec l'équipe qui a récolté le jeu de données afin d'en savoir un peu plus sur cette station n°1.

On constate qu'avant, le test de Shapiro-Wilk, aussi appelé test de normalité, avait un w de 0.44, ce qui est médiocre. Après avoir enlevé la valeur aberrante, on arrive à une valeur du w à 0.89. Si d'habitude on fixe le niveau de confiance à 95%, au vu du nombre limité de données, cela reste acceptable.

```
> shapiro.test(res$residuals)      > shapiro.test(resModif$residuals)

      shapiro-wilk normality test      shapiro-wilk normality test

data:  res$residuals                data:  resModif$residuals
w = 0.44175, p-value = 7.354e-12    w = 0.88682, p-value = 0.0004388
```

De cette manière, nous pouvons conclure à la bonne normalité des résidus. Cela nous conforte dans l'idée que notre modèle est pertinent.

En conclusion, et avec l'appui des analyses descriptives faites précédemment, nous ne retenons qu'un modèle à deux variables explicatives "nbpompes" et "trafic" et ce, en enlevant la première donnée du jeu de données. D'ailleurs, même avec le retrait de la station n°1, le modèle à trois variables est toujours contestable car la variable "nbconc" est la moins significative des trois.

b. Modèle à 2 variables

Le modèle à deux variables retenu possède de très bonnes caractéristiques :

```
> #Modèle à deux variables explicatives retenues
> res2 <- lm(stationModif$ventes~stationModif$trafic+stationModif$nbpompes, data=matYXModif)
> summary(res2)

call:
lm(formula = stationModif$ventes ~ stationModif$trafic + stationModif$nbpompes,
    data = matYXModif)

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 ***
stationModif$trafic  1.09956    0.01721    63.88  <2e-16 ***
stationModif$nbpompes  2.77766    0.01763   157.53  <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
```

En effet, comme expliqué dans le test à trois variables, la p-value globale du test de Fisher exprimée sur la dernière ligne nous permet d'identifier la significativité du modèle. Cette p-value est très faible (inférieure à 2.2×10^{-16}) ce qui se traduit par un fort rejet de H_0 . Le modèle est donc globalement significatif. De plus, le R^2 ajusté est ici de 99,9% : la variabilité des ventes est donc expliquée à 99,9% par les variations des autres variables. Enfin, chacune des variables explicatives est significative à 99%.

Le modèle que nous avons ici est très fortement heuristique. Il représente très bien la réalité. En d'autres termes, ce modèle a une forte validité économique. Nous pouvons donc faire quelques affirmations : Les ventes d'une station dépendent surtout du nombre de pompes. C'est ce que nous avons déjà vu avec la corrélation. Mais ce que nous pouvons dire c'est que le coefficient B_1 , coefficient de la variable trafic, est d'environ 1. Cela signifie qu'une variation (en millier) du trafic entrainera une variation strictement proportionnelle des ventes. Le coefficient B_2 , coefficient de la variable nbpompes, est d'environ 2,8. Cela signifie qu'une variation unitaire du nombre de pompes entrainera une variation presque trois fois proportionnelle des ventes.

Autrement dit, dans le cas où un propriétaire de station à essence souhaite faire augmenter ses ventes, il peut ouvrir une station dans une zone géographique où le trafic est fort, c'est une bonne assise, mais il obtiendra plus de résultat en augmentant le nombre de pompes dans sa station. Toutes choses égales par ailleurs, ajouter une pompe à sa station aura le même effet sur les ventes qu'une augmentation du trafic quotidien de 2800 voitures. C'est très intéressant ! D'autant plus qu'il est plus difficile pour un propriétaire de station essence d'avoir une influence directe sur le trafic plutôt que sur le nombre de pompes.

Prenons l'exemple où notre propriétaire souhaiterait savoir combien de ventes il peut s'attendre à avoir s'il construit une nouvelle station. Il prévoit un trafic quotidien moyen de 19 000 voitures et une installation avec 15 pompes.

```
$fit
      fit      lwr      upr
1 255.6457 254.2159 257.0756

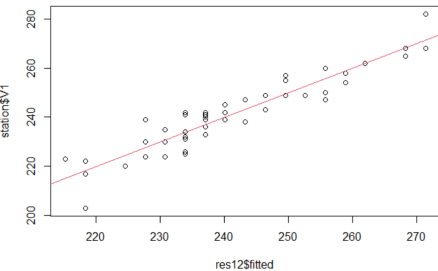
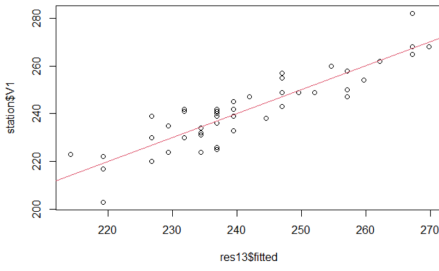
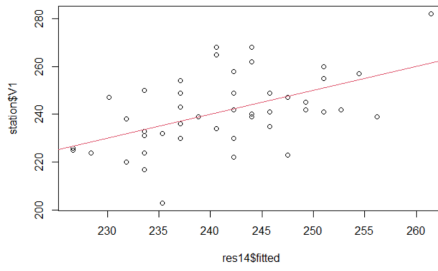
$se.fit
[1] 0.1100446

$df
[1] 41
```

On peut affirmer qu'en moyenne cette station aura des ventes entre 254.1 et 257.2. En effet, la portée de l'incertitude est de plus ou moins 0.110046.

c. Modèle à 1 variable

Nous étudions les régressions linéaires simples entre les ventes et chaque autre variable :

Variable	nbpompes	nbconc	trafic
Régression linéaire			
Adjusted R-squared	0.869	0.8039	0.2533
p-value	< 2.2e-16	< 2.2e-16	0.0002524

On remarque que les modèles de régression linéaire simples obtenus ont des R-squared inférieurs à celui du modèle à 2 variables étudié précédemment. Sachant qu'il est préférable de garder le modèle avec le R-squared le plus élevé, le modèle à 2 variables est le plus pertinent.

4. Conclusion explicative

Pour conclure, nous avons cherché à modéliser les ventes d'une station service en fonction de différentes variables. Pour cela, nous avons étudié et comparé plusieurs modèles de régression linéaire simples ou multiples, à une, deux ou trois variables. Le meilleur modèle retenu expliquant les ventes d'une station service est une régression linéaire multiple avec les variables nbpompes et trafic.

On peut mettre en lumière l'heuristique du modèle. Comme on peut l'observer dans la réalité, un propriétaire de station essence souhaitant construire une nouvelle station essence ou cherchant à augmenter les ventes d'une station existante peut essayer d'influencer trois variables explicatives : le trafic quotidien en millier de voiture, le nombre de pompes et le nombre de concurrents aux alentours. Il peut par exemple choisir la localisation géographique de sa nouvelle station, cela aura pour effet d'influencer le trafic potentiel (si situé sur une route avec beaucoup de passage ou non par exemple) et le nombre de concurrents aux alentours. Mais ces deux variables sont des variables coûteuses car pour les influencer il faut déployer de gros moyens (travaux, construction). Au contraire, l'ajout de pompes à une station est un ajustement moins coûteux. La bonne nouvelle pour le propriétaire est que le nombre de ventes d'une station est plus sensible aux variations unitaire du nombre de pompes qu'aux variations en milliers de voitures du trafic quotidien.

Annexes

Annexe 1 : Code R de l'exercice Ozone

Lien:

<https://docs.google.com/document/d/15AP5hctDLWerSaFJEtLT0InN51LJGhxGstEY1VPgUaM/edit?usp=sharing>

Annexe 2 : Code R de l'exercice Station

Lien:

https://docs.google.com/document/d/1LKkskmXcs4oBLslfandigEWtDwVNpqS_wf0ea4RSR-s/edit?usp=sharing