Codes pour jeu de données Ozone

S. Donnet

26 janvier 2020

1. Les données

(question1 : Exercice du chapitre 1)

```
ozone <- read.table("ozone.txt",header = T,sep = " ")
names(ozone)
```

```
##  [1] "maxO3"  "T9"     "T12"    "T15"    "Ne9"    "Ne12"   "Ne15"
##  [8] "Vx9"    "Vx12"   "Vx15"   "maxO3v" "vent"   "temps"
```

```
nrow(ozone)
```

```
## [1] 112
```

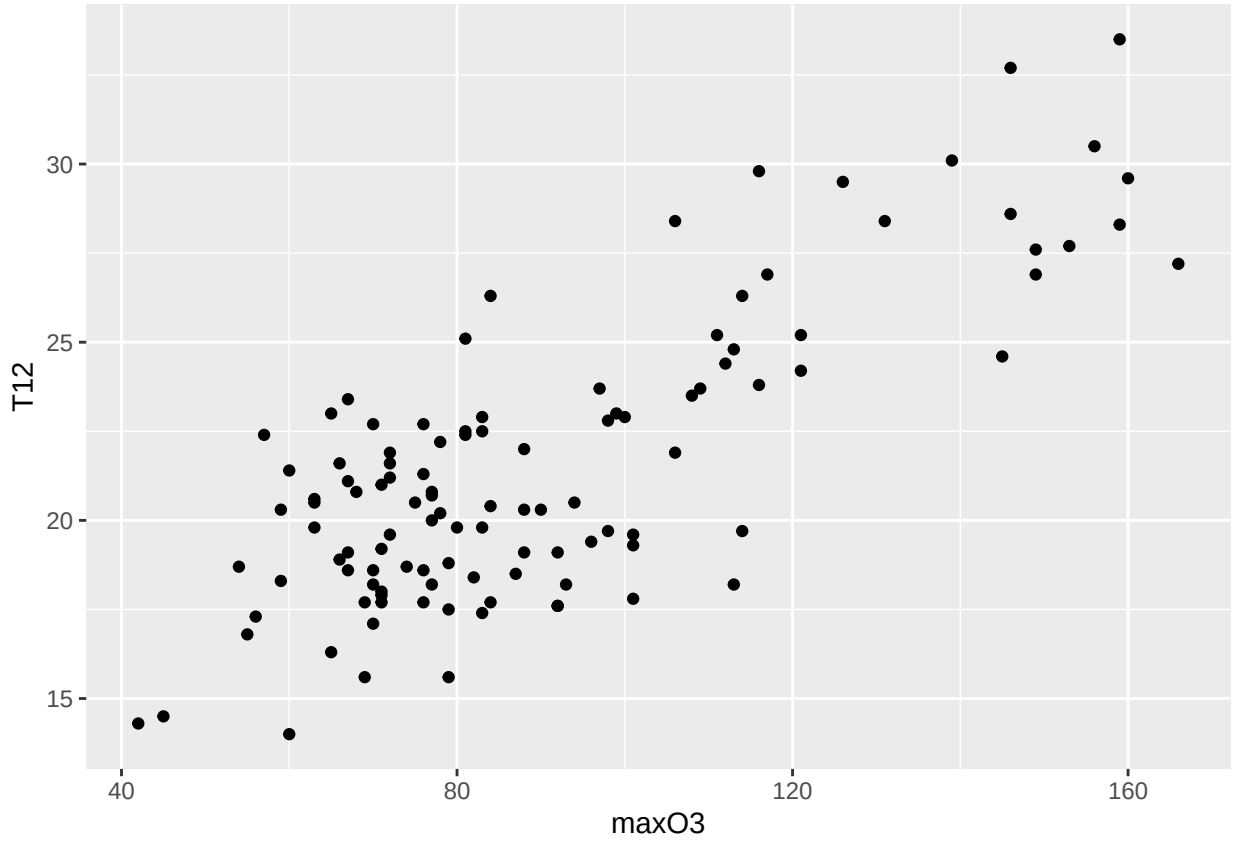
2- Vous semble-t-il opportun d'utiliser le modèle linéaire?

OUI

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
gg <- ggplot(ozone, aes(x = maxO3, y = T12)) + geom_point()
gg
```



2. Modèle linéaire avec un unique régresseur

Estimation des paramètres du modèle

(exercice chapitre 1.) 3- Afficher le résumé des informations de la régression. Donner l'EMC $\hat{\beta}$.

```
res_lm <-lm(maxO3 ~T12,data=ozone)
summary(res_lm)
```

```
##
## Call:
## lm(formula = maxO3 ~ T12, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.079 -12.735   0.257  11.003  44.671
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.4196     9.0335  -3.035    0.003 **
## T12           5.4687     0.4125  13.258   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.57 on 110 degrees of freedom
```

```
## Multiple R-squared:  0.6151, Adjusted R-squared:  0.6116
## F-statistic: 175.8 on 1 and 110 DF,  p-value: < 2.2e-16
```

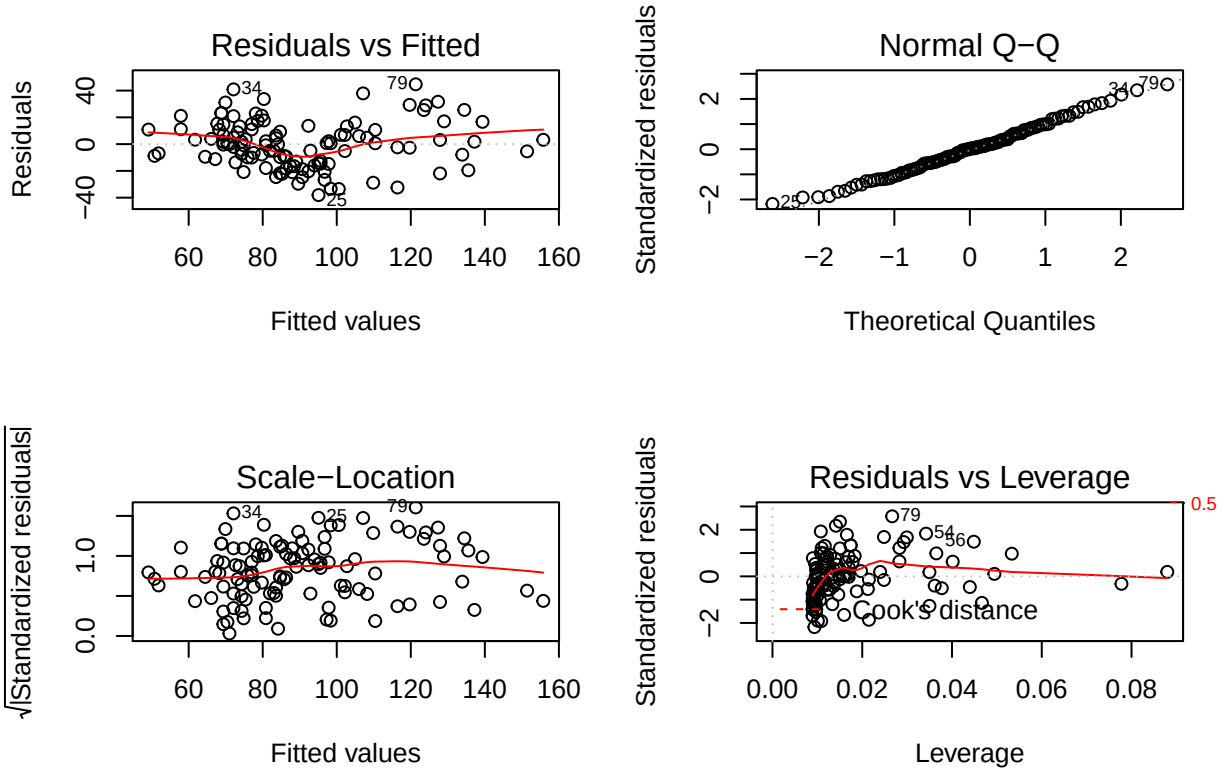
```
coef(res_lm)
```

```
## (Intercept)         T12
##  -27.419636    5.468685
```

Test de la significatité de la variables

(Chap 1, exercice question 4) : Est-ce que la variable `T12` vous semble bien expliquer linéairement la variable `maxO3`?

```
par(mfrow=c(2,2))
plot(res_lm)
```



Prédiction

(Chap 1, exercice question 5) 5- Donner la prédiction d'ozone pour une température égale à 27 degrés ainsi que l'intervalle de confiance à 95% correspondant.

```
xnew=c(27)
xnew=data.frame(T12=xnew)
U <- predict(res_lm,new=xnew,se.fit=T)
names(U)
```

```
## [1] "fit"            "se.fit"         "df"             "residual.scale"
```

3. Modèle linéaire à plusieurs variables

Estimation de paramètres

```
attach(ozone)
reg = lm(maxO3~Ne12+Vx12+T12)
summary(reg)
```

```
##
## Call:
## lm(formula = maxO3 ~ Ne12 + Vx12 + T12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.462 -11.448  -0.722   8.908  46.331
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8958    14.8243   0.263   0.7932
## Ne12         -1.6189     1.0181  -1.590   0.1147
## Vx12          1.6290     0.6571   2.479   0.0147 *
## T12           4.5132     0.5203   8.674 4.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.63 on 108 degrees of freedom
## Multiple R-squared:  0.6612, Adjusted R-squared:  0.6518
## F-statistic: 70.25 on 3 and 108 DF,  p-value: < 2.2e-16
```

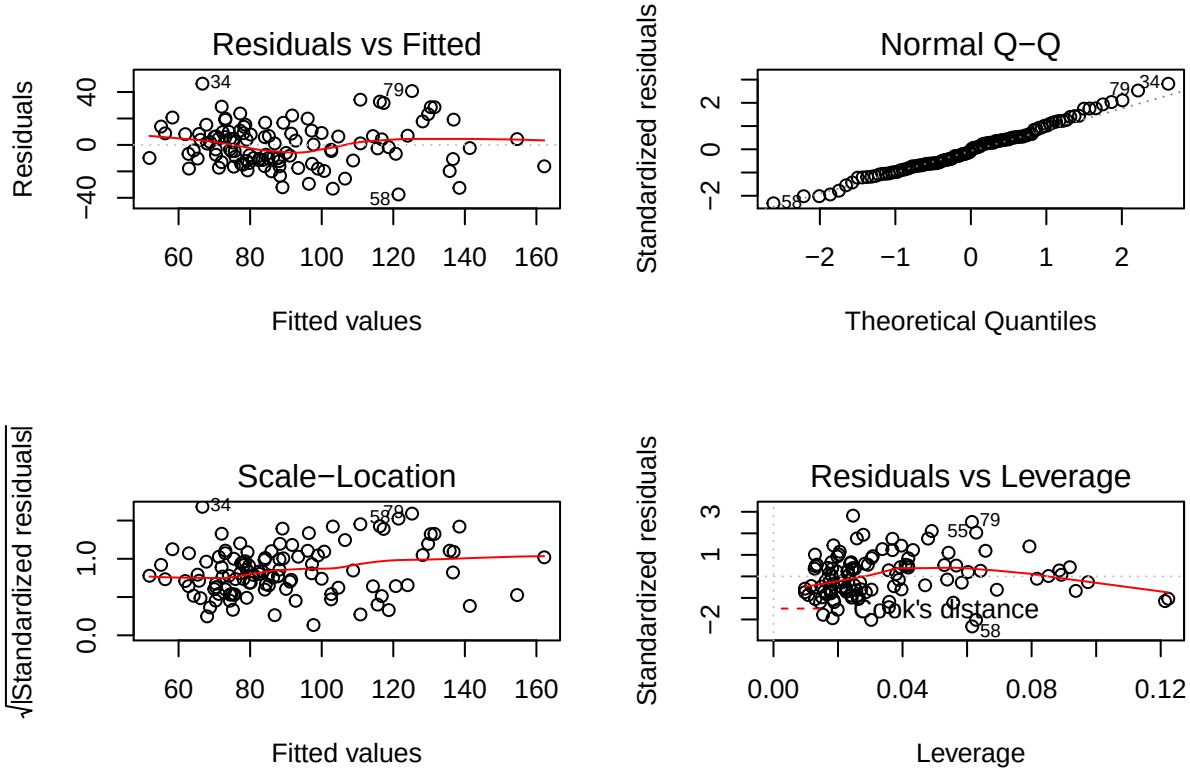
Variance sigma^2

```
summary(reg)$sigma^2
```

```
## [1] 276.6734
```

Analyse des résidus

```
par(mfrow=c(2,2))
plot(reg)
```



Prédiction

On a mesuré la valeur des trois variables `T12`, `Ne12` et `Vx12` pour une nouvelle journée : (20,6,-3). Donner le taux d'ozone prévu par le modèle linéaire. (Le code est le même que dans le chapitre 2)

```
xnew = data.frame(t(c(20,6,-3)))
names(xnew) = c("T12","Ne12", "Vx12")
predict(reg,new=xnew)
```

```
##        1
## 79.56015
```

Intervalles de confiance

Sur les paramètres

```
confint(reg,level = 0.95)
```

```
##                   2.5 %    97.5 %
## (Intercept) -25.4886483 33.280203
## Ne12         -3.6368523  0.399082
## Vx12          0.3264694  2.931560
## T12           3.4819098  5.544563
```

Sur les prédictions

```
Nenew=c(2,3)
Vxnew=c(-1,0)
Tnew=c(46,35)
xnew=data.frame(Ne12=Nenew,Vx12=Vxnew,T12=Tnew)
predict(reg,new=xnew,interval="pred",level=0.95)
```

```
##        fit      lwr      upr
## 1 206.6379 166.8820 246.3938
## 2 157.0024 121.8401 192.1647
```

Tests

Test sur chaque paramètres

```
summary(reg)
```

```
##
## Call:
## lm(formula = maxO3 ~ Ne12 + Vx12 + T12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.462 -11.448  -0.722   8.908  46.331
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8958    14.8243   0.263   0.7932
## Ne12         -1.6189     1.0181  -1.590   0.1147
## Vx12          1.6290     0.6571   2.479   0.0147 *
## T12           4.5132     0.5203   8.674 4.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.63 on 108 degrees of freedom
## Multiple R-squared:  0.6612, Adjusted R-squared:  0.6518
## F-statistic: 70.25 on 3 and 108 DF,  p-value: < 2.2e-16
```

Test sur une combinaison de covariables

```
reg1 <- lm(maxO3 ~ T12)
anova(reg1,reg)
```

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ T12
## Model 2: maxO3 ~ Ne12 + Vx12 + T12
##   Res.Df   RSS Df Sum of Sq    F   Pr(>F)
## 1    110 33948
## 2    108 29881  2    4067.1 7.35 0.001017 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

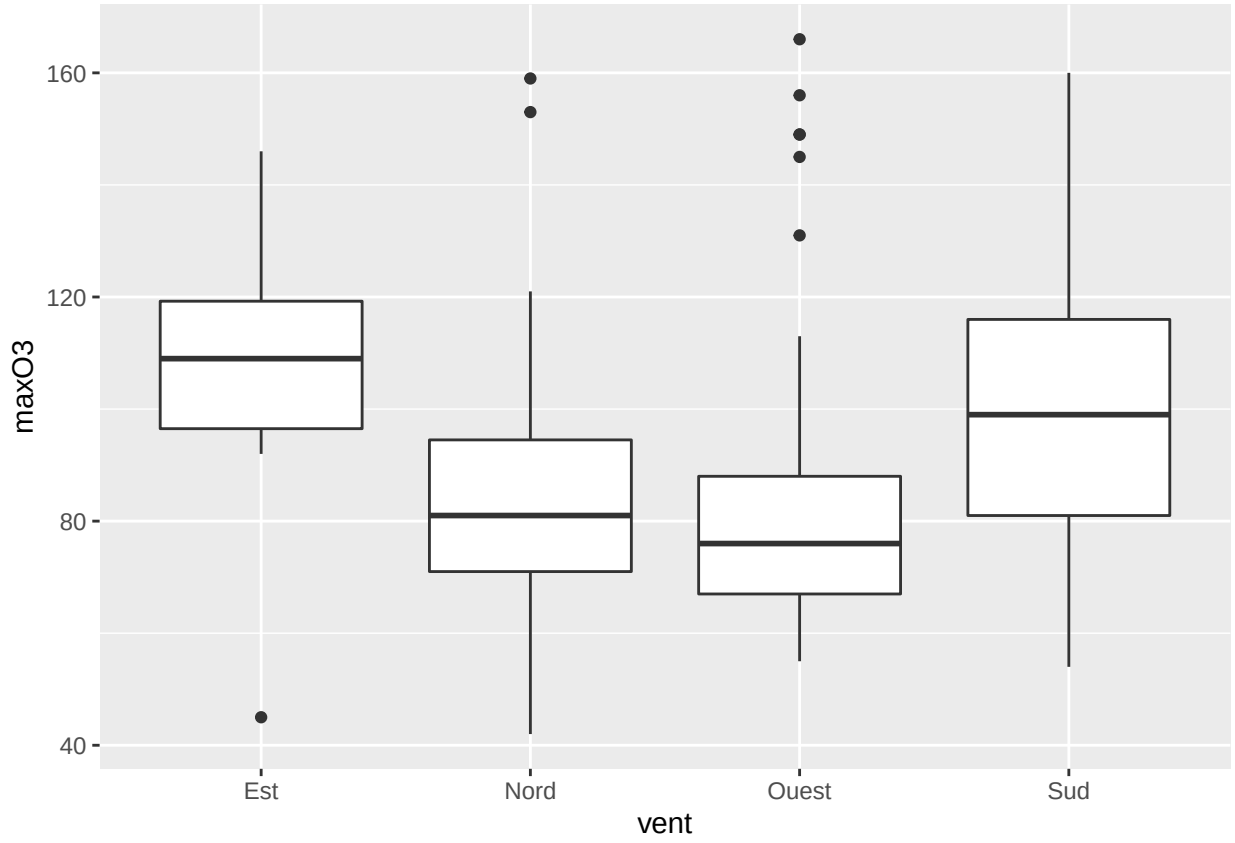
Ajustement

```
summary(reg)$r.squared
```

```
## [1] 0.6611843
```

4. Modèle d'Anova a un facteur

```
ggplot(ozone,aes(x=vent,y=maxO3))+geom_boxplot()
```




```
reg_reg <- lm(maxO3 ~ - 1 + vent ,data = ozone)
summary(reg_reg)
```

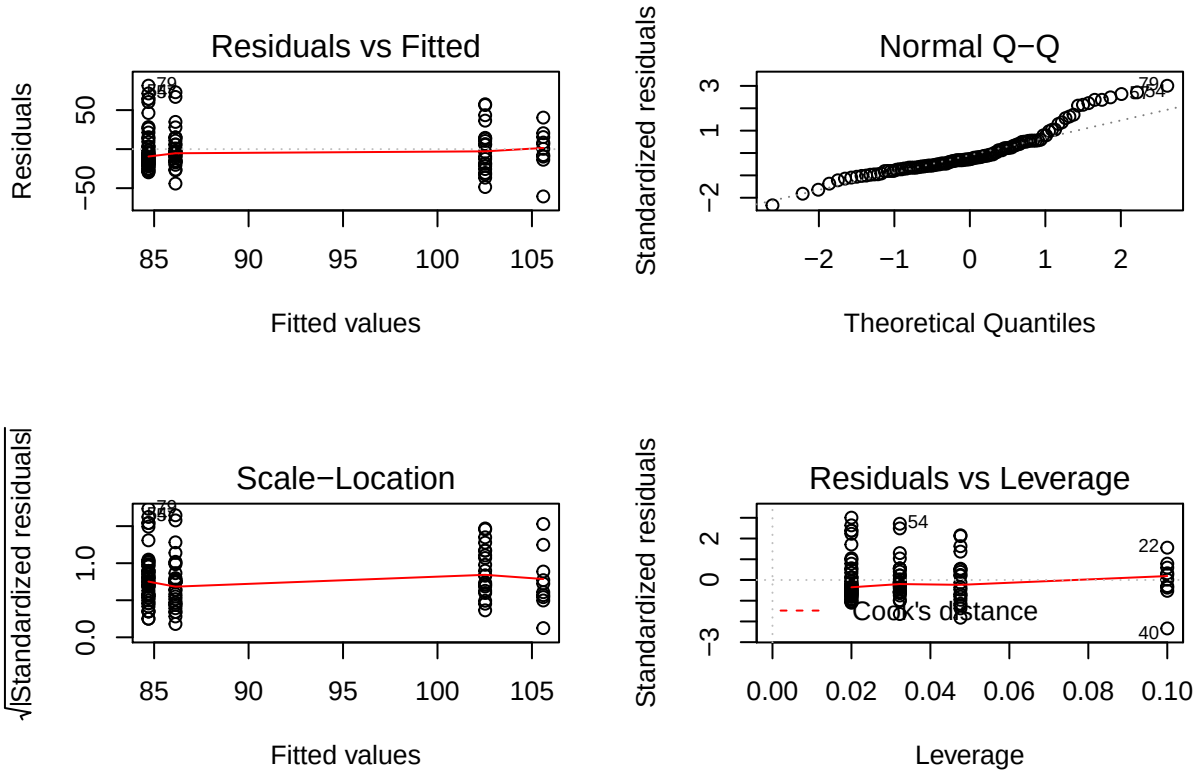
```
##
## Call:
## lm(formula = maxO3 ~ -1 + vent, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.600 -16.807  -7.365  11.478  81.300
##
```

```
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## ventEst    105.600      8.639   12.22   <2e-16 ***
## ventNord    86.129      4.907   17.55   <2e-16 ***
## ventOuest   84.700      3.864   21.92   <2e-16 ***
## ventSud    102.524      5.962   17.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.32 on 108 degrees of freedom
## Multiple R-squared:  0.9195, Adjusted R-squared:  0.9165
## F-statistic: 308.5 on 4 and 108 DF,  p-value: < 2.2e-16
```

```
reg_sing <- lm(maxO3 ~  vent ,data = ozone)
summary(reg_sing)
```

```
##
## Call:
## lm(formula = maxO3 ~ vent, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.600 -16.807  -7.365  11.478  81.300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  105.600      8.639  12.223   <2e-16 ***
## ventNord     -19.471      9.935  -1.960   0.0526 .
## ventOuest    -20.900      9.464  -2.208   0.0293 *
## ventSud       -3.076     10.496  -0.293   0.7700
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.32 on 108 degrees of freedom
## Multiple R-squared:  0.08602,    Adjusted R-squared:  0.06063
## F-statistic: 3.388 on 3 and 108 DF,  p-value: 0.02074
```

```
par(mfrow = c(2,2))
plot(reg_sing)
```



Exercice

1. On considère la régression de `maxO3` sur `Vent` avec la contrainte par défaut (cf sortie précédente). Que peut-on dire à partir des résultats des tests de Student ?

Réponse : La -valeur du test de Student lié au coefficient α_4 est grande : on conserve donc l'hypothèse H_0 correspondante c'est-à-dire qu'on conserve l'hypothèse que $\alpha_4 = 0$ (au niveau usuel 5%). Cela signifie ici que le vent du Sud est considéré comme ayant la même influence sur le taux d'ozone que le vent vent de l'Est. Cela confirme la sensation que l'on a en regardant les boxplots.

Les p-valeurs du vent du Nord et de l'Ouest sont au contraire petites : on considère que le vent du Nord et le vent de L'ouest ont un effet différent du vent d'Est sur le taux d'ozone.

2. En observant les boxplots, on a la sensation que les vents du Nord et de l'Ouest ont le même effet. Faire le test correspondants pour vérifier ces deux hypothèses.

Réponse possible

On fait une régression en prenant le vent du Nord (ou le vent de l'Ouest) comme référence :

```
summary(lm(maxO3~C(vent,base=2),data=ozone))
```

```
##
## Call:
## lm(formula = maxO3 ~ C(vent, base = 2), data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -60.600 -16.807  -7.365  11.478  81.300
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           86.129      4.907  17.553   <2e-16 ***
## C(vent, base = 2)1    19.471      9.935   1.960   0.0526 .
## C(vent, base = 2)3    -1.429      6.245  -0.229   0.8194
## C(vent, base = 2)4    16.395      7.721   2.123   0.0360 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.32 on 108 degrees of freedom
## Multiple R-squared:  0.08602,    Adjusted R-squared:  0.06063
## F-statistic: 3.388 on 3 and 108 DF,  p-value: 0.02074
```

Le vent d'Ouest est la troisième modalité : on a une p-valeur grande (0.8194) donc on conserve H_0 autrement dit le vent d'Ouest n'a pas une influence différente de celle du vent du Nord sur le taux d'ozone.

Pour info, si on utilise le vent d'Ouest comme référence on obtient (sans surprise)

```
summary(lm(maxO3~C(vent,base=3),data=ozone))
```

```
##
## Call:
## lm(formula = maxO3 ~ C(vent, base = 3), data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.600 -16.807  -7.365  11.478  81.300
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           84.700      3.864  21.923   <2e-16 ***
## C(vent, base = 3)1    20.900      9.464   2.208   0.0293 *
## C(vent, base = 3)2     1.429      6.245   0.229   0.8194
## C(vent, base = 3)4    17.824      7.104   2.509   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.32 on 108 degrees of freedom
## Multiple R-squared:  0.08602,    Adjusted R-squared:  0.06063
## F-statistic: 3.388 on 3 and 108 DF,  p-value: 0.02074
```

La même question correspondrait ici à la nullité du coefficient du vent du Nord (numéro 2) : on obtient exactement la même p-valeur (et c'est normal).

Autre solution possible

```
comp.statut = pairwise.t.test(ozone$maxO3,ozone$vent,p.adjust.method = "none")
comp.statut
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ozone$maxO3 and ozone$vent
##
```

```
##       Est   Nord  Ouest
## Nord  0.053 -     -
## Ouest 0.029 0.819 -
## Sud   0.770 0.036 0.014
##
## P value adjustment method: none
```

```
pairwise.t.test(ozone$maxO3,ozone$vent,p.adjust.method = "bonferroni")
```

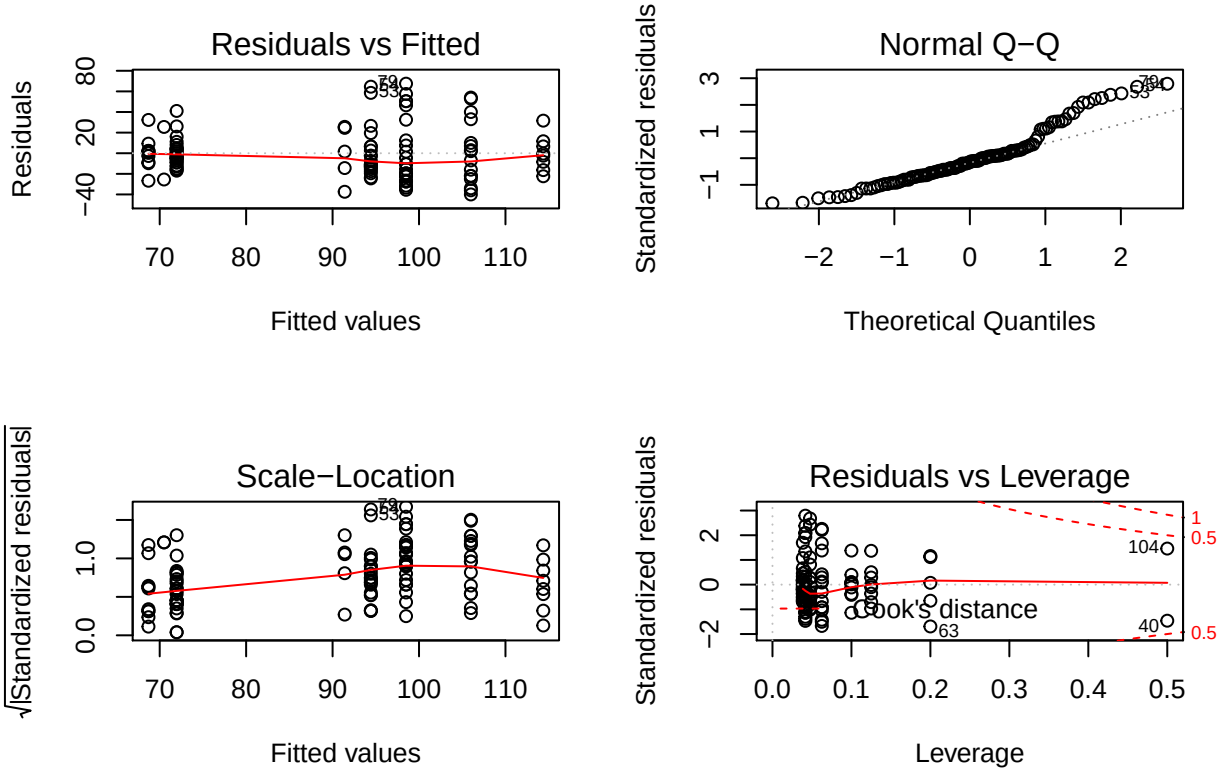
```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ozone$maxO3 and ozone$vent
##
##       Est   Nord  Ouest
## Nord  0.316 -     -
## Ouest 0.176 1.000 -
## Sud   1.000 0.216 0.082
##
## P value adjustment method: bonferroni
```

5. ANOVA a 2 facteurs

```
mod1 <- lm(maxO3~vent*temps,data=ozone)
summary(mod1)
```

```
##
## Call:
## lm(formula = maxO3 ~ vent * temps, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.000 -15.971  -3.462   7.635  67.500
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           70.500     17.464   4.037 0.000104 ***
## ventNord              -1.800     19.131  -0.094 0.925221
## ventOuest              1.462     18.123   0.081 0.935881
## ventSud               20.900     20.664   1.011 0.314161
## tempsSec              43.875     19.526   2.247 0.026749 *
## ventNord:tempsSec    -18.146     21.709  -0.836 0.405138
## ventOuest:tempsSec   -17.337     20.739  -0.836 0.405117
## ventSud:tempsSec     -29.275     23.267  -1.258 0.211138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.7 on 104 degrees of freedom
## Multiple R-squared:  0.2807, Adjusted R-squared:  0.2322
## F-statistic: 5.797 on 7 and 104 DF,  p-value: 1.092e-05
```

```
par(mfrow=c(2,2))
plot(mod1)
```




```
mod2 <- lm(maxO3~vent + temps,data=ozone)
anova(mod1,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ vent * temps
## Model 2: maxO3 ~ vent + temps
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1    104 63440
## 2    107 64446 -3   -1006.4 0.55 0.6493
```

```
mod4 <- lm(maxO3~ temps ,data=ozone)
anova(mod4,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ temps
## Model 2: maxO3 ~ vent + temps
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    110 68238
## 2    107 64446  3    3791.3 2.0982 0.1048
```

```
mod3 <- lm(maxO3~ vent ,data=ozone)
anova(mod3,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ vent
```

```
## Model 2: maxO3 ~ vent + temps
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    108 80606
## 2    107 64446  1     16159 26.829 1.052e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(car)
```

```
## Loading required package: carData
```

```
Anova(mod2)
```

```
## Anova Table (Type II tests)
##
## Response: maxO3
##           Sum Sq  Df F value    Pr(>F)
## vent        3791   3  2.0982    0.1048
## temps      16159   1 26.8295 1.052e-06 ***
## Residuals  64446 107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```