Université Abderahmane Mira de Bejaia
Faculté des Sciences Exactes
Département Recherche Opérationnelle

TP Modèles Linéaires

Djamila Ouaret / Faiza Amarouche

Application de régression linéaire simple et multiple sur "ozone" sous R

October 14, 2023

SOMMAIRE

1 Introduction

La régression linéaire est l'une des techniques les plus fondamentales en statistiques et en apprentissage automatique. Elle est utilisée pour modéliser et prédire les relations linéaires entre une variable dépendante continue et une ou plusieurs variables indépendantes.

La régression linéaire simple est le cas où une seule variable indépendante est utilisée pour prédire la variable dépendante. L'objectif est de trouver une relation linéaire qui minimise l'erreur de prédiction entre les valeurs observées et les valeurs prédites par le modèle. La relation linéaire est décrite par une équation de la forme Y = 0 + 1X + , où Y est la variable dépendante, X est la variable indépendante, 0 et 1 sont les coefficients de régression, et est l'erreur résiduelle.

La régression linéaire multiple est une extension de la régression linéaire simple, où plusieurs variables indépendantes sont utilisées pour prédire la variable dépendante. L'équation de régression devient

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

où X_1, X_2, ..., X_n sont les variables indépendantes, β_1, β_2, ..., β_n sont les coefficients de régression correspondants, et ϵ est l'erreur résiduelle.

2 Régression linéaire simple

L'ozone troposphérique est une problématique environnementale de plus en plus répandue à l'échelle de la planète. L'augmentation démesurée de la taille des agglomérations, la croissance démographique et l'augmentation des émissions sont autant de facteurs qui contribuent à cette pollution de nature urbaine.

De ce fait, la pollution de l'air est actuellement une des préoccupations majeures de santé publique. Des associations de surveillance de la qualité de l'air existent sur tout le territoire français et mesurent la concentration des polluants ainsi que les conditions météorologiques, comme la température, la nébulosité, le vent, ...

Dans ce rapport, on dispose d'un jeu de données de concentration d'OZONE dans l'air qui contient 112 données et 13 variables, relevées durant l'été 2001 à Rennes stockées dans le fichier CSV "ozone.csv".

Nous allons commencer par la mise en œuvre de la régression linéaire simple sur le pic d'ozone (maxO3), expliqué par la température à midi (T12).

2.1 Importer les données

On charge nos données sous le nom "ozone" et voyons toutes les variables existantes dans ce jeu de données.

```
ozone=read.csv(file.choose('ozone.csv'), header = TRUE, sep= ";", encoding="UTF-8"
    )
ozone
# Nos variables
names(ozone)
```

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
20010601	87	15.6	18.5	18.4	4	4	8	0.6946	-1.7101	-0.6946	84	Nord	Sec
20010602	82	17	18.4	17.7	5	5	7	-4.3301	-4	-3	87	Nord	Sec
20010603	92	15.3	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82	Est	Sec
20010604	114	16.2	19.7	22.5	1	1	0	0.9848	0.3473	-0.1736	92	Nord	Sec
20010605	94	17.4	20.5	20.4	8	8	7	-0.5	-2.9544	-4.3301	114	Ouest	Sec
20010606	80	17.7	19.8	18.3	6	6	7	-5.6382	-5	-6	94	Ouest	Pluie
20010607	79	16.8	15.6	14.9	7	8	8	-4.3301	-1.8794	-3.7588	80	Ouest	Sec
20010610	79	14.9	17.5	18.9	5	5	4	0	-1.0419	-1.3892	99	Nord	Sec
20010611	101	16.1	19.6	21.4	2	4	4	-0.766	-1.0261	-2.2981	79	Nord	Sec
20010612	106	18.3	21.9	22.9	5	6	8	1.2856	-2.2981	-3.9392	101	Ouest	Sec

Table Chart 1 of 12 Rows per page 10 112 rows

```
'maxO3' · 'T9' · 'T12' · 'T15' · 'Ne9' · 'Ne12' · 'Ne15' · 'Vx9' · 'Vx12' · 'Vx15' · 'maxO3v' · 'vent' · 'pluie'
```

```
# Statistique descriptive des données de l'étude
summary(ozone)
```

```
     maxO3             T9            T12             T15
 Min.   : 42.00   Min.   :11.30   Min.   :14.00   Min.   :14.90
 1st Qu.: 70.75   1st Qu.:16.20   1st Qu.:18.60   1st Qu.:19.27
 Median : 81.50   Median :17.80   Median :20.55   Median :22.05
 Mean   : 90.30   Mean   :18.36   Mean   :21.53   Mean   :22.63
 3rd Qu.:106.00   3rd Qu.:19.93   3rd Qu.:23.55   3rd Qu.:25.40
 Max.   :166.00   Max.   :27.00   Max.   :33.50   Max.   :35.50
      Ne9             Ne12            Ne15            Vx9
 Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :-7.8785
 1st Qu.:3.000   1st Qu.:4.000   1st Qu.:3.00   1st Qu.:-3.2765
 Median :6.000   Median :5.000   Median :5.00   Median :-0.8660
 Mean   :4.929   Mean   :5.018   Mean   :4.83   Mean   :-1.2143
 3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:7.00   3rd Qu.: 0.6946
 Max.   :8.000   Max.   :8.000   Max.   :8.00   Max.   : 5.1962
      Vx12            Vx15            maxO3v          vent
 Min.   :-7.878   Min.   :-9.000   Min.   : 42.00   Length:112
 1st Qu.:-3.565   1st Qu.:-3.939   1st Qu.: 71.00   Class :character
 Median :-1.879   Median :-1.550   Median : 82.50   Mode  :character
 Mean   :-1.611   Mean   :-1.691   Mean   : 90.57
 3rd Qu.: 0.000   3rd Qu.: 0.000   3rd Qu.:106.00
 Max.   : 6.578   Max.   : 5.000   Max.   :166.00
    pluie
 Length:112
 Class :character
 Mode  :character
```

Pour décrire la relation entre la concentration maximale en ozone maxO3 et la température à midi T12, on calcule la corrélation entre les deux variables.

```
# Calculer la corrélation entre 'maxO3' et 'T12'
cor.test(ozone$T12, ozone$maxO3)
```

```
    Pearson's product-moment correlation

data:  ozone$T12 and ozone$maxO3
t = 13.258, df = 110, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7006713 0.8466150
sample estimates:
      cor
0.7842623
```

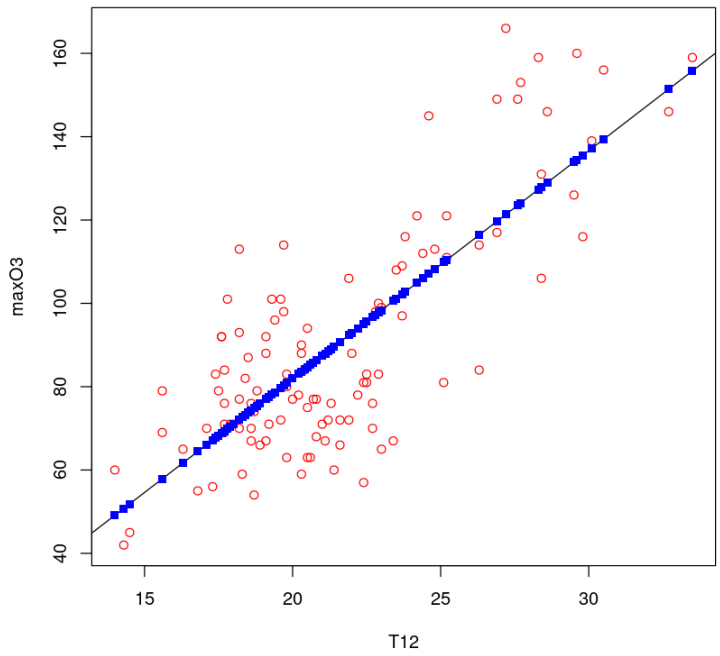
Le test sur le coefficient de corrélation entre maxO3 et T12 donne une p-value $< 2.2 \times 10^{-16}$, inférieure au seuil =0.05. L'hypothèse nulle d'absence de corrélation est donc rejetée au seuil de 5%.

L'estimation ponctuelle égale à 0,7842623 indique **une corrélation relativement forte et positive entre max03 et T12**.

2.2 Tracer le nuage des points et superposer la droite de régression

```r
# Graphique du nuage des points
plot(maxO3~T12, data=ozone, col='red')
# Droite de la regression simple
reg_simple= lm(maxO3~T12, data=ozone)
summary(reg_simple)
# Superposer la droite sur le graphe de nuage des points
abline(reg_simple)
# Afficher les points de données observés pour la variable T12 en fonction des
    valeurs prédites par la régression linéaire
points(ozone$T12,reg_simple$fitted.values,pch=15,col='blue')
```

```
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
```



β_0= -27.420 et β_1= 5.469, cela indique que pour chaque degré supplémentaire de température, on peut s'attendre à ce que la concentration en ozone augmente de 5.469 g/l. Par exemple, avec l'augmentation d'une unité de x (la température à midi) de 15 à 16, on voit bien que l'ozone passe de 54,6109 à 60,0796 degrès soit un 5,4687 de plus, ce qui est bel et bien la valeur de la pente β_1.

Les valeurs de p sont inférieures à 0.05, alors on peut rejeter l'hypothèse nulle, ce qui signifie que tout changement dans la valeur prédictive est **lié** à un changement dans la variable réponse. C'est-à-dire qu'une faible valeur de T12 peut influencer significativement le modèle.

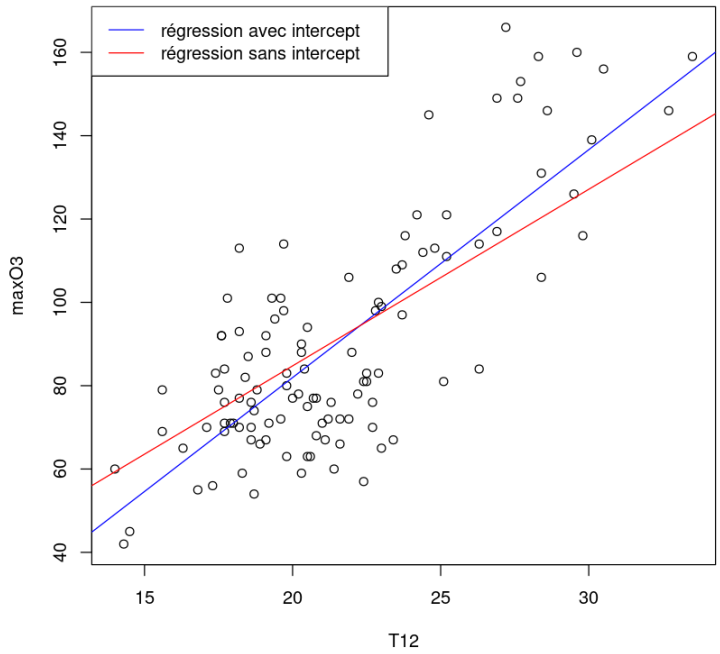
D'où, p = 0.003 pour l'ordonnée à lorigine signifie que la constante (intercept β_0) doit apparaître dans le modèle et $p = 2e - 16$ pour la pente indique **une liaison significative** entre la concentration en ozone (maxO3) et la température à midi (T12).

2.3 Comparer graphiquement la régression avec et sans intercept

```r
# Ajuster le modèle 'sans intercept'
reg_ajusté <- lm(maxO3 ~ T12- 1, data = ozone)
summary(reg_ajusté)
# Tracé du nuage de points entre T12 et maxO3
plot(ozone$T12, ozone$maxO3, xlab = 'T12', ylab = 'maxO3')
# Ajout de la droite de régression linéaire du modèle reg_simple (bleue)
abline(reg_simple, col = 'blue')
# Ajout de la droite de régression linéaire du modèle reg_ajusté (rouge)
abline(reg_ajusté, col = 'red')
```

```r
# Ajout d'une légende en haut à gauche
legend('topleft', c('régression avec intercept', 'régression sans intercept'), lty = 1, col = c('blue', 'red'))
```

```
Call:
lm(formula = maxO3 ~ T12 - 1, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-37.930 -14.290  -2.462   9.605  50.728

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
T12  4.23795    0.07855   53.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.21 on 111 degrees of freedom
Multiple R-squared:  0.9633,    Adjusted R-squared:  0.9629
F-statistic:  2911 on 1 and 111 DF,  p-value: < 2.2e-16
```



- **Comparaison :**

L'analyse met en évidence une relation linéaire positive entre les concentrations maximales de maxO3 et les valeurs de T12. Le nuage de points entre ces deux variables présente une tendance croissante.

L'ajustement d'un modèle de régression linéaire simple sans intercept entre maxO3 (variable dépendante) et T12 (variable explicative) montre une pente positive. Ce résultat traduit le fait qu'une augmentation de T12 est associée à une augmentation de maxO3.

La comparaison graphique avec un modèle de régression linéaire simple avec intercept indique que le modèle sans intercept (forçant une régression passant par 0) s'ajuste bien à la tendance des données.

En conclusion, cette analyse met en évidence et quantifie via un modèle statistique la relation linéaire positive existant entre les concentrations de maxO3 et les températures relevées par la variable T12 dans le jeu de données.

2.4 Les valeurs des estimateurs renvoyées par la fonction lm sont-elles exactes ?

Dans notre jeu de données, nous disposons de 112 observations et nous intéressons au couple $(T12\ ;\ max03) = (xi; yi)$ afin d'estimer les paramètres inconnus β_0 et β_1 de notre modèle.

Ce dernier se résume à notre droite :

$$y_i = \beta_0 + \beta_1 x_i$$

Et, on a:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$$

Calcul manuel:

Nous avons dans la sortie de la commande *summary(ozone)* le résumé statistique de toutes les variables de l'ensemble de données, alors, prenons les moyennes: $\bar{y}$ et $\bar{x}$, telles que: $\bar{y} = 90,30$: La moyenne de la variable maxO3 et $\bar{x} = 21,53$: La moyenne de la variable T12, d'où : $Cov(x,y) = 89,360$ et $Var(x) = 16,34$.

Ce qui nous déduit:

- $\beta_1 = \frac{89,360}{16,34}$ alors, $\beta_1 = 5,469$
- $\beta_0 = 90,30 - 5,469 * 21,53$ donc, $\beta_0 = -27,447$

Sous R:

```
B1=cov(ozone$T12,ozone$max03)/var(ozone$T12)
B0=mean(ozone$max03)-mean(ozone$T12)*(cov(ozone$T12,ozone$max03)/var(ozone$T12))
cat('B0 =',B0)
cat('B1 =',B1)
```

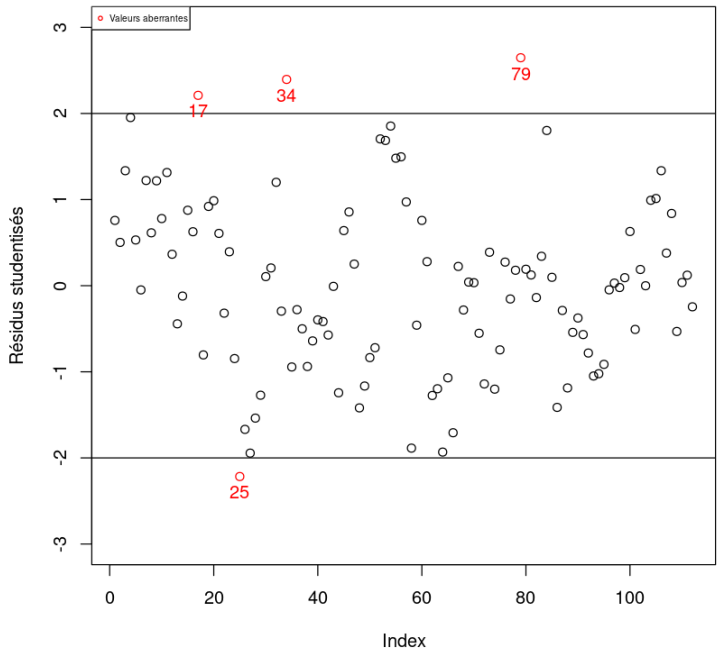
B0 = -27.41964 ; B1 = 5.468685

On conclut explicitement que les valeurs des estimateurs β_0 et β_1 renvoyées par la fonction *lm* sont exactes en les comparant aves les résultats du calcul manuel ou du code R.

2.5 Tracer le nuage des points des valeurs abérrantes

Vérifions si le jeu de données contient des valeurs abérrantes.

```
# Résidus studentisés
res_stud <- rstudent(reg_simple)
# Identification des valeurs aberrantes
abr <- which(abs(res_stud) > 2)
# Tracé des résidus
plot(res_stud, ylab = "Résidus studentisés",
     ylim = c(-3,3),
     col = ifelse(abs(res_stud) > 2, "red", "black"))
# 2 Lignes horizontales (Seuils)
abline(-2,0)
abline(2,0)
# Légende
legend("topleft", c("Valeurs aberrantes"),
       pch = 1, col = c("red"), cex = 0.5)
# Étiquettage des points aberrants
text(abr, rstudent(reg_simple)[abr], abr, col = "red", pos = 1)
```

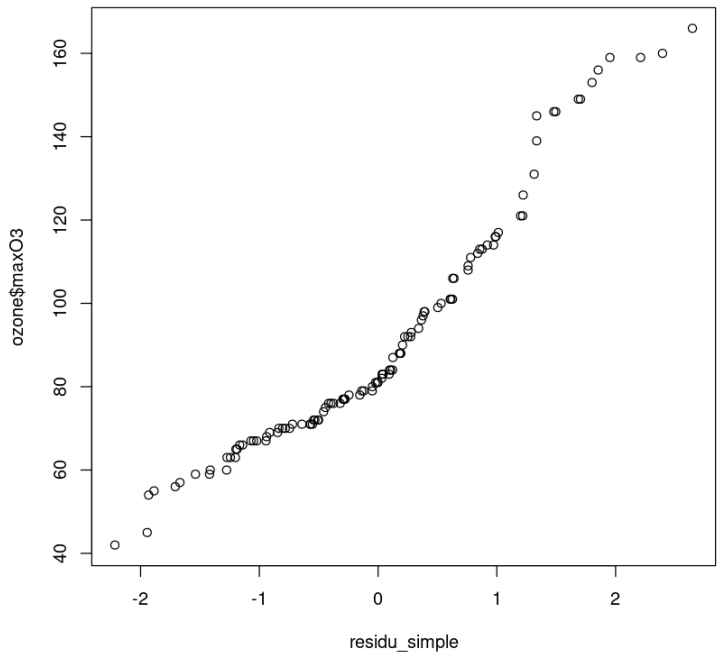


Dans ce cas, on a **quatres** valeurs aberrantes qui se trouvent à l'extérieur de l'intervalle [-2;2] et qui sont: 17, 25, 34 et 79.

2.6 Vérifier avec QQ plot l'hypothèse gaussienne des résidus

On parcourt ce code afin de voir le graphique voulu.

```
# QQ-plot
qqplot(rstudent(reg_simple),ozone$max03)
```



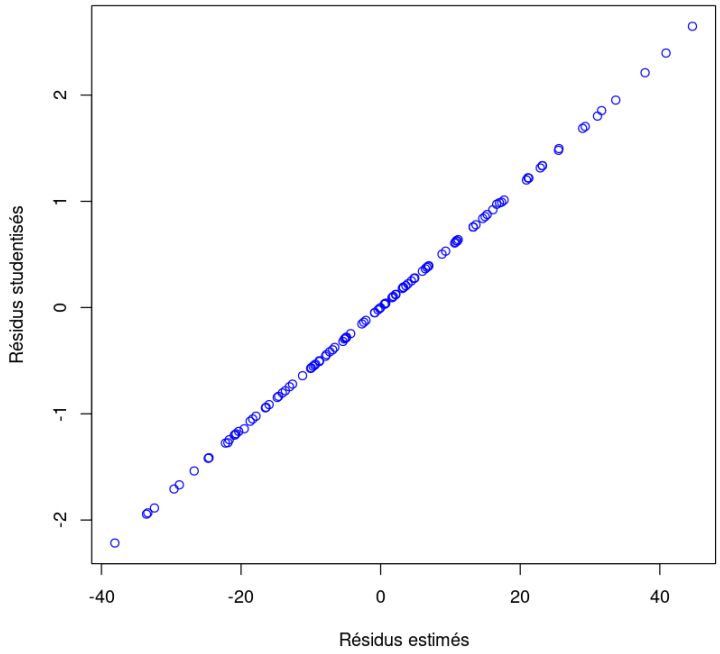
Ici, la fonction *qqplot()* est utilisée pour comparer la distribution des résidus studentisés du modèle de régression (reg-simple) à la distribution des valeurs de la variable maxO3 contenues dans le dataframe ozone.

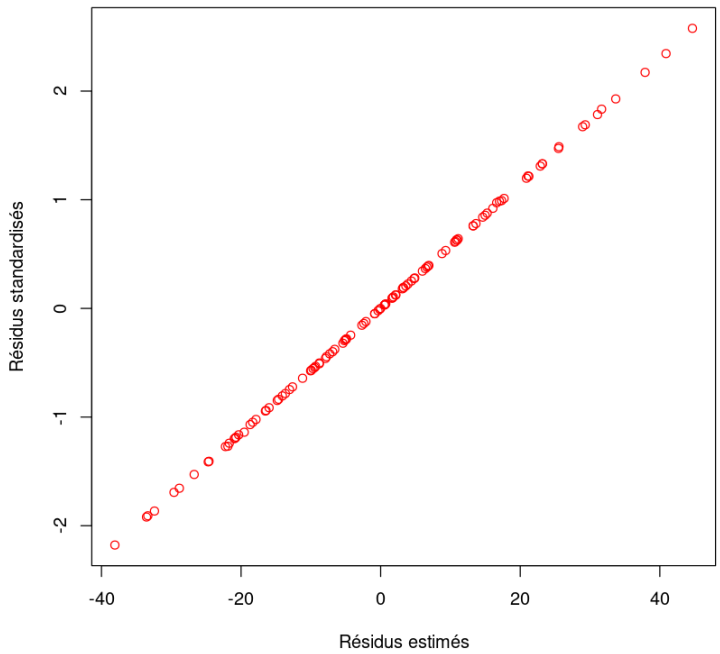
On peut remarquer que les points se distribuent approximativement le long de la diagonale, cela indique que la distribution est gaussienne et les résidus suivent une loi normale. Car, si les résidus s'écartent beaucoup de la diagonale, cela remet en cause l'hypothèse de normalité et la validité du modèle.

2.7 Comparer par un graphique les résidus estimés aux résidus standardisés et aux résidus studentisés

```
# Les résidus estimés
residus_estimes <- reg_simple$residuals
# Les résidus standardisés
residus_standardises <- rstandard(reg_simple)
# Les résidus studentisés
residus_studentises <- rstudent(reg_simple)

# Représentation graphique des résidus studentisés en fonction des résidus estimés
    (bleue)
plot(residus_estimes, residus_studentises, col='blue', xlab="Résidus estimés",
    ylab="Résidus studentisés")
# Représentation graphique des résidus standardisés en fonction des résidus
    estimés (rouge)
```

```
11 plot(residus_estimes, residus_standardises, col='red', xlab="Résidus estimés",
       ylab="Résidus standardisés")
```





On voit clairement une similarité des graphiques qui est plutôt un signal positif pour la régression, indiquant que les résidus bruts (estimés) ne présentent a priori pas de problème

majeur de non-normalité.
En d'autres termes, les différentes transformations des résidus (standardisation et studentisation) suivent globalement la même tendance et n'ont dans ce cas que peu d'impact sur leur distribution.

2.8 Marquer les points leviers

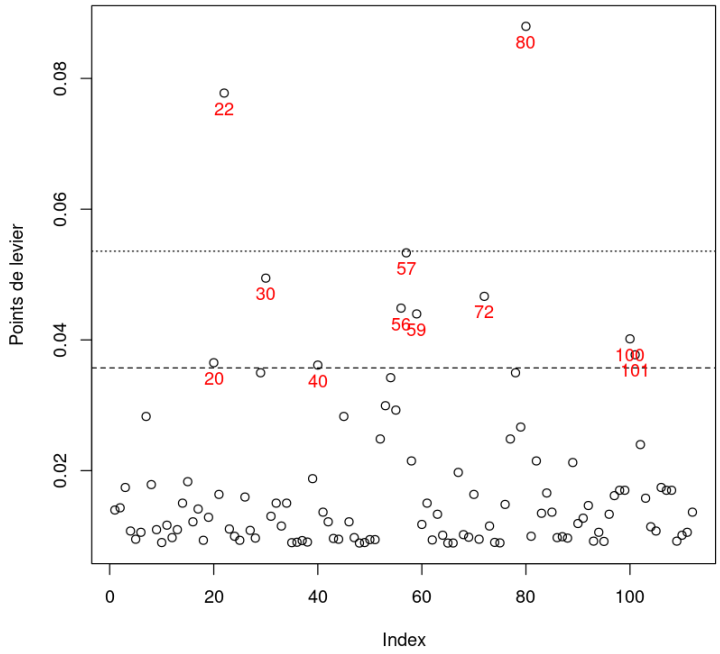
On saisit la ligne de code ci-dessous:

```
# Vérifier l'existance des points leviers
leviers=hatvalues(reg_simple)
leviers
```

```
20010601:  0.0139795884009378 20010602:  0.0143188562636468 20010603:  0.0174299602861664 20010604:  0.0107684564769346
20010605:  0.00950983768503931 20010606:  0.0105725356856 43 20010607:  0.0282951957268396 20010610:  0.0178684684290874
20010611:  0.0109754039660275 20010612:  0.00900536626062821 20010613:  0.0116624066201143 20010614:  0.00975833219620512
20010615:  0.0109754039660275 20010616:  0.0150304720824687 20010617:  0.0183180032698097 20010618:  0.0121755418785122
20010620:  0.0141357343180174 20010621:  0.00934896604033054 20010622:  0.0128684516921354 20010623:  0.0365120546420133
20010624:  0.0163674591908806 20010625:  0.0777576146732268 20010626:  0.0110752349621932 20010627:  0.00996823326506626
20010628:  0.00934896604033054 20010629:  0.0159679383009001 20010630:  0.0108631679348354 20010701:  0.00968744628176796
20010702:  0.0349742241093628 20010703:  0.0494516876065627 20010704:  0.0130279449996 19 20010705:  0.0150304720824687
20010706:  0.0115324491103129 20010707:  0.0150304720824687 20010708:  0.00898744787670104 20010709:  0.009052032820966
20010710:  0.00930544996507884 20010711:  0.00908156861864815 20010712:  0.0187785648083333 20010713:  0.0361511271943375
20010714:  0.0136513472360302 20010715:  0.0121755418785122 20010716:  0.00962857159172155 20010717:  0.00950983768503931
20010718:  0.0282951957268396 20010719:  0.0121755418785122 20010720:  0.00975833219620512 20010721:  0.00893152676642284
20010722:  0.00900536626062821 20010723:  0.00945076608967501 20010724:  0.00945076608967501 20010725:  0.024846399084154
20010726:  0.029939158225828 20010727:  0.0342218488897395 20010728:  0.029263969890814 20010729:  0.0448628091733462
20010730:  0.05332127091356 20010731:  0.0214899510355546 20010801:  0.0439781135801069 20010802:  0.0117775962310747
20010803:  0.0150304720824687 20010804:  0.0094021304761584 20010805:  0.0133341327689239 20010806:  0.0101251668034174
20010807:  0.00893152676642284 20010808:  0.00893743392595927 20010809:  0.0197327679787846 20010810:  0.0102137741964639
20010811:  0.00982232642451644 20010812:  0.0163674591908806 20010813:  0.00950983768503931 20010814:  0.0466652804532288
20010815:  0.0115324491103129 20010819:  0.00902899489877393 20010820:  0.00895692755242948 20010821:  0.0148355358177665
20010822:  0.024846399084154 20010823:  0.0349742241093628 20010824:  0.0266734835287717 20010825:  0.0879667615947204
20010826:  0.00996823326506626 20010827:  0.0214899510355546 20010828:  0.0134800396094737 20010829:  0.0165860240937285
20010830:  0.0136513472360302 20010831:  0.00975833219620512 20010901:  0.00989911949849003 20010902:  0.00968744628176796
20010903:  0.0212467729679716 20010904:  0.0119134609004126 20010905:  0.0127327839281154 20010906:  0.0146691508241571
20010907:  0.00921979615180061 20010908:  0.0105725356856 43 20010912:  0.00917844603504 56 20010913:  0.0133341327689239
20010914:  0.0161805960442115 20010915:  0.0170024788410468 20010916:  0.0170024788410468 20010917:  0.0401630730461627
20010918:  0.0377228254416635 20010919:  0.0239906485726432 20010920:  0.0157861946924959 20010921:  0.0114223790376174
20010922:  0.0107684564769346 20010923:  0.0174299602861664 20010924:  0.0170024788410468 20010925:  0.0170024788410468
20010927:  0.00921979615180061 20010928:  0.0101251668034174 20010929:  0.0105725356856 43 20010930:  0.0136513472360302
```


Et voici le marquage des points leviers dans le nuage des points (T12,maxO3):

```
# Tracé des points de levier en fonction de leur index
plot(1:n, leviers, xlab='Index', ylab='Points de levier')
# Calcul du seuil théorique 1
p = reg_simple$rank  # Nombre de paramètres du modèle
seuil1 = 2*p/n       # Seuil à 2p/n
# Calcul du seuil théorique 2
seuil2 = 3*p/n       # Seuil à 3p/n
# Ajout de lignes horizontales aux seuils
abline(seuil1,0, lty=2)
abline(seuil2,0, lty=3)
# Identification des points au dessus du seuil 1
lev = (1:n)[leviers > seuil1]
# Etiquetage des points de levier au dessus du seuil
text(lev, leviers[lev], lev, col='red', pos=1)
```



Le graphique des points de levier en fonction de leur index montre que 9 points dépassent le seuil théorique de 2p/n, et 2 points dépassent le seuil de 3p/n. Ces points avec un effet de levier élevé sont potentiellement influents sur le modèle.

2.9 Analyser la distance de Cook des observations

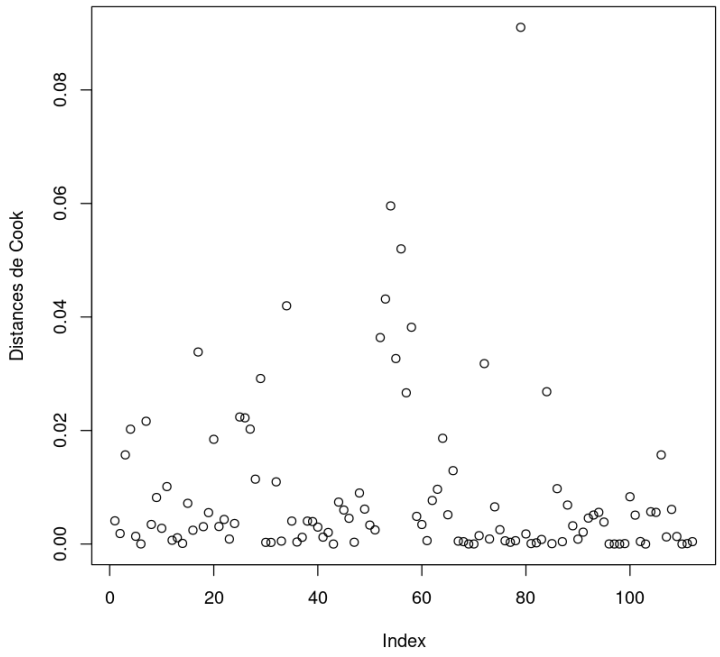
Sous R, on obtient la distance de Cook par la fonction *cooks.distance*.

```
cook=cooks.distance(reg_simple)
cook
```

```
20010601:     0.00408918653573329 20010602:     0.00184731050572267 20010603:     0.0157036693201 65 20010604:     0.0202311629927259
20010605:     0.00136166858501166 20010606:     1.29504909381566e-05 20010607:     0.0216319073789547 20010610:     0.00344744705770462
20010611:     0.00819589358876562 20010612:     0.00277025908126665 20010613:     0.0101207536312791 20010614:     0.000661495700667857
20010615:     0.00109652163254591 20010616:     0.000111794079539862 20010617:     0.00717551581074413 20010618:     0.00243172584515188
20010620:     0.0338285553416731 20010621:     0.00305921743470014 20010622:     0.00553046450257112 20010623:     0.0184489091134327
20010624:     0.00308458354724597 20010625:     0.00432925308687 07 20010626:     0.000874927984835927 20010627:     0.00361552099006556
20010628:     0.0223790416404685 20010629:     0.0222281020252553 20010630:     0.0202449538050 04 20010701:     0.0114254187964979
20010702:     0.0291576436544318 20010703:     0.000291209539752326 20010704:     0.000282121637637 63 20010705:     0.0109530854188215
20010706:     0.000514730341494657 20010707:     0.0419664791729556 20010708:     0.00404448874380 58 20010709:     0.000357327723050524
20010710:     0.00118470172816009 20010711:     0.00404135514383491 20010712:     0.00395577141544445 20010713:     0.00298100916293773
20010714:     0.00121070853852777 20010715:     0.00203458298459526 20010716:     3.18600513714913e-07 20010717:     0.00738717432049722
20010718:     0.00599070609729 19 20010719:     0.00452889720768916 20010720:     0.000312896775636571 20010721:     0.00899087225670057
20010722:     0.00614901971632422 20010723:     0.00333812786539842 20010724:     0.00248760413897949 20010725:     0.0363713210602906
20010726:     0.0431619431029163 20010727:     0.0595671989030089 20010728:     0.0326748993946178 20010729:     0.0519960469071453
20010730:     0.0266413209293377 20010731:     0.0381879982878962 20010801:     0.00487334669652 55 20010802:     0.00343801667417981
20010803:     0.000600093581653987 20010804:     0.00767479516952771 20010805:     0.00964204645899498 20010806:     0.0186314698297323
20010807:     0.00515389748387096 20010808:     0.0129254111414 89 20010809:     0.000507363502932897 20010810:     0.000414549711710327
20010811:     8.7353810346528e-06 20010812:     1.01574016941483e-05 20010813:     0.00147410234059285 20010814:     0.0317849382702789
20010815:     0.000887301983009746 20010819:     0.00655011326543949 20010820:     0.00252135291366663 20010821:     0.000569691816433924
20010822:     0.000305859619106364 20010823:     0.000588109340489449 20010824:     0.0910274316747239 20010825:     0.00177503416572847
20010826:     7.87897170954975e-05 20010827:     0.000210633109069629 20010828:     0.000803476296419307 20010829:     0.0268348621633292
20010830:     6.58618258618 33e-05 20010831:     0.00975271395617414 20010901:     0.000416861780049919 20010902:     0.00687027491849585
20010903:     0.00321183293977904 20010904:     0.000855722869793804 20010905:     0.00209048340037949 20010906:     0.00456584038956178
20010907:     0.00511204955323856 20010908:     0.00558136363696103 20010912:     0.003870464412982 11 20010913:     1.583627823473 53e-05
20010914:     7.61299339395888e-06 20010915:     4.0321141523236e-06 20010916:     7.51767727214779e-05 20010917:     0.00832694905515687
20010918:     0.00509091989743456 20010919:     0.000438986647553955 20010920:     7.35619646981593e-09 20010921:     0.00568512016819358
20010922:     0.00557691326733047 20010923:     0.0157036693201 65 20010924:     0.00125079782801956 20010925:     0.00609650233143995
20010927:     0.00132430533 31595 20010928:     6.854793526662 82e-06 20010929:     8.01045979861869e-05 20010930:     0.000419931203851446
```

Note: Le seuil critique pour la distance de Cook à partir duquel on considère que l'observation est **trop infuente** est le quantile $f_(n, n-p)_{0.5}$. Une distance de Cook en dessous de $f_(n, n-p)_{0.1}$ est considérée comme **souhaitable**. Traçons son graphique:

```r
# Taille des données
n <- length(cook)
# Nombre de paramètres
p <- length(coef(reg_simple))
# Graphique des distances de Cook
plot(1:n, cook, xlab='Index', ylab='Distances de Cook')
# Seuil à 50%
s1 = qf(0.5, p, n-p)
# Seuil à 10%
s2 = qf(0.1, p, n-p)
# Lignes horizontales aux seuils
abline(h = s1, lty=3)
abline(h = s2, lty=2)
```



D'après le graphique ci-haut, les observations de l'analyse de régression n'ont pas une influence excessive sur les estimations des paramètres du modèle, car les distances de Cook sont toutes en dessous du seuil critique $f_(n, n-p)_{0.1}$. Cela n'est qu'une indication que notre modèle de régression est relativement **robuste aux observations atypiques ou influentes**.

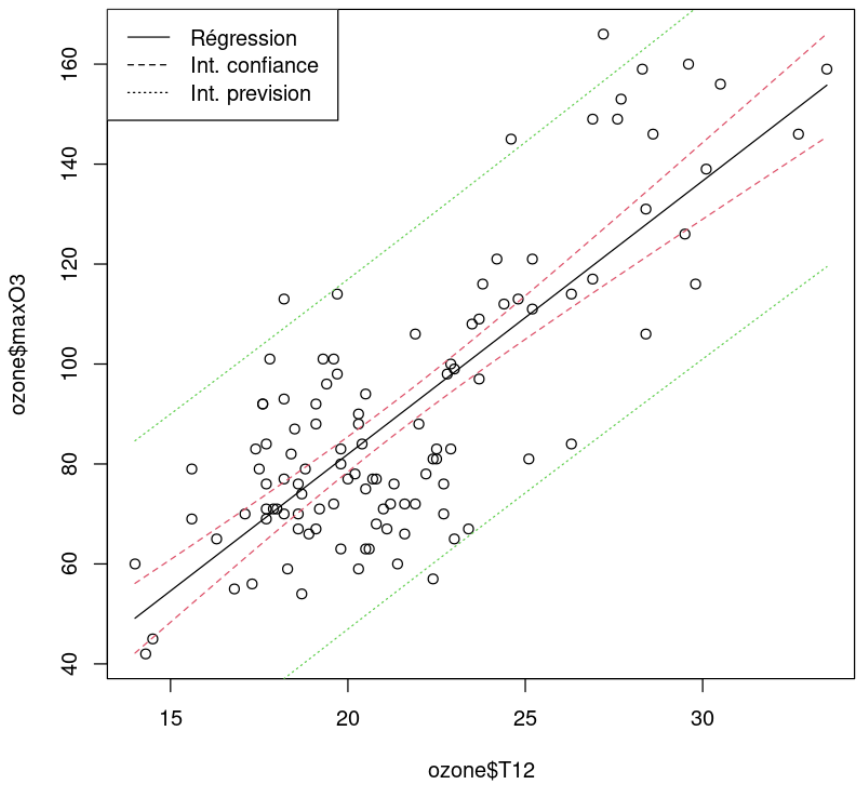
2.10 Intervalle de confiance et prévision

- Ajouter à la figure du nuage de points et de la droite de régression les intervalles de prédiction et de confiance en tout points x_i observé.

On parcourt le code suivant:

```
plot(ozone$maxO3 ~ ozone$T12, ylim = c(min(ozone$maxO3), max(ozone$maxO3)))

seq.x <- seq(min(ozone$T12), max(ozone$T12), length=3*n)
grid.x <- data.frame(seq.x)
dimnames(grid.x)[[2]] <- "T12"

ICconf <- predict(reg_simple, new=grid.x, interval="confidence", level=0.95)
ICprev <- predict(reg_simple, new=grid.x, interval="prediction", level=0.95)

matlines(grid.x, cbind(ICconf, ICprev[,-1]), lty=c(1,2,2,3,3), col= c(1, 2, 2, 3,
    3))

legend("topleft", lty = 1:3, c("Régression","Int. confiance", "Int. prévision") )
```

On obtient ce graphique:



- **Interprétation du graphique**

L'écart des points par rapport à la ligne de régression est en accord avec la discussion précédente sur les résidus et la variance.

Les intervalles des prévision sont toujours plus amples que ceux de confiance.
- Ceci est interprétable comme suit : Les intervalles de confiance concernent l'espérance de

la concentration O3, pour une température T12 donnée, alors que les intervalles de prévision essaient de prendre en compte toute concentration possible dans le seuil fixé.

- **Comparaison entre les deux intervalles**

Remarque: Les observations atypiques (ou valeurs aberrantes) peuvent avoir un impact important sur les intervalles de confiance et de prédiction (seront plus larges).

Dans notre dessin, on observe que l'intervalle de prédiction est plus large que l'intervalle de confiance et cela signifie que l'estimation de la dispersion des valeurs futures est plus large que l'estimation de la dispersion des valeurs moyennes.

De plus, l'intervalle de confiance est dune forme hyperbolique, plus une observation est éloignée du centre de gravité, moins nous avons dinformation sur elle. Lorsque x_i est proche de $\bar{x}$, le terme dominant de la variance est $1/n$, mais dès que x_i séloigne de la moyenne, le terme dominant est le terme au carré.

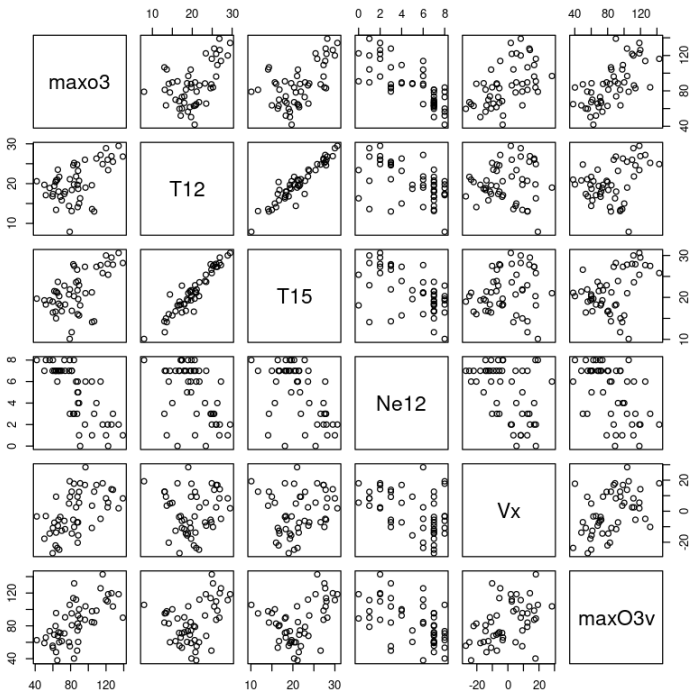
Egalement, l'intervalle de confiance a 95% de chance de contenir la vraie valeur et lintervalle de prédiction a 95% de contenir une observation future.

3 Régression linéaire multiple

Dans cette partie de la régression lineaire multiple, on utilisera des données saisis à la main, composées de 50 observations et 6 variables (MaxO3, maxO3v, Vx, Ne12, T15, T12)

3.1 Saisir les données

```
maxO3<-c(63.6,89.6, 79, 81.2, 88, 68.4, 139, 78.2, 113.8, 41.8, 65, 73, 126.2,
    127.8, 61.6, 63.6, 134.2, 67.2, 87.8, 96.8, 89.6, 66.4, 60, 90.8, 104.2, 70,
    96.2, 65.6, 109.2, 86.2, 87.4, 84, 83, 59.7, 52, 73.8, 120, 122.4, 106.6,
    121.8, 116.2, 81.4, 88.6, 63, 104, 88.4, 83.8, 56.4, 50.4, 79.2)

T12 <- c(13.4, 15, 7.9, 13.1, 14.1, 16.7, 26.8, 14.4, 27.2, 20.6, 21, 17.4, 26.9,
    25.5, 19.4, 20.8, 29.5, 21.7, 19.7, 19, 20.7, 18, 17.4, 16.3, 13.6, 15.8, 26,
    23.5, 26.3, 21.8, 24.8, 25.2, 24.6, 16.8, 17.1, 18, 28.9, 23.4, 13, 26, 24.9,
    18.4, 18.7, 20.4, 19.6, 23.2, 19.8, 18.9, 19.7, 21.1)

T15 <- c(15, 15.7, 10.1, 11.7, 16, 18.1, 28.2, 20.7, 27.7, 19.7, 21.1, 22.8, 29.5,
     27.8, 21.5, 21.4, 30.6, 20.3, 21.7, 21, 22.9, 18.5, 16.4, 18.1, 14.1, 16.7,
    27.3, 23.7, 27.3, 23.6, 26.6, 27.5, 27.9, 19, 18.3, 18.3, 30, 25.4, 14.3, 28,
    25.8, 16.8, 19.6, 16.6, 21.2, 23.9, 20.3, 19.3, 19.3, 21.9)

Ne12 <-  c( 7, 4, 8, 7, 6, 7, 1, 7, 6, 8, 6, 8, 2, 3, 7, 7, 2, 7, 5, 6, 1, 7, 8,
    0, 1, 7, 2, 7, 4, 6, 3, 3, 3, 7, 8, 7, 1, 0, 3, 2, 2, 7, 5, 7, 6, 4, 8, 8, 7,
    3)

Vx <- c(9.35, 5.4, 19.3, 12.6, -20.3, -3.69, 8.27, 4.93, -4.93, -3.38, -23.68,
    -6.24, 14.18, 13.79, -7.39,      -13.79, 1.88, -24.82, 9.35, 28.36, 12.47,
    -5.52, -10.8, 18, 3.55, -12.6, 16.91, -9.35, 16.91, 2.5, -7.09,      -10.15,
    -5.52, -27.06, -3.13, -11.57, 8.27, 5.52, 12.6, 2.5, 18, -14.4, -15.59,
    -22.06, -10.8, -7.2, 17.73,     -14.4, -17.73, 9.26)

maxO3v <- c(95.6, 100.2, 105.6, 95.2, 82.8, 71.4, 90, 60, 125.8, 62.6, 38, 70.8,
    119.8, 103.6, 69.2, 48, 118.6, 60, 74.4, 103.8, 78.8, 72.2, 53.4, 89, 97.8,
    61.4, 87.4, 67.8, 98.6, 112, 49.8, 131.8, 113.8, 55.8, 65.8, 90.4, 111.4,
    118.6, 84, 109.8, 142.8, 80.8, 60.4, 79.8, 84.6, 92.6, 40.2, 73.6, 59, 55.2)

# Créer un dataframe
df_ozone=data.frame(maxO3, T12, T15, Ne12, Vx, maxO3v)
# Visualiser les données
plot(df_ozone)
```



L'objectif de cette appliaction est de modéliser le pic d'ozone journalier en fonction de toutes les autres variables météorologiques; il s'agira là d'une **régression linéaire multiple** avec 5 variables explicatives.

La régression linéaire assume que la relation entre les variables explicatives et la variable à expliquer (variable numérique continue) va être linéaire, du type : $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...\beta_5 X_5 + \epsilon$

où :

- y: est la variable à expliquer (ici :maxO3).
- X1,...,X5 sont les variables explicatives (ici : maxO3v, Vx, Ne12, T15, T12).
- ϵ: est le terme derreur, assumé être distribué selon une loi Normale $N(0, \sigma^2)$.

On souhaite donc estimer les coefficients du modèle:
$maxO3 = \beta_0 + \beta_1 * T12 + \beta_2 * T15 + \beta_3 * Ne12 + \beta_4 * Vx + \beta_5 * maxO3v + \epsilon$

3.2 Estimer les coefficients de la régression multiple

```
reg_multiple = lm(maxo3 ~ T12 + T15 + Ne12 + Vx + maxO3v)
coef(reg_multiple)
```

```
Call:
lm(formula = maxo3 ~ T12 + T15 + Ne12 + Vx + maxO3v)

Coefficients:
(Intercept)          T12          T15         Ne12           Vx       maxO3v
   62.51384      0.08446      0.93121     -3.94166      0.31653      0.26146
```

L'équation de la droite de régression est donc:
$max03 = 62.51 + 0.08 * T12 + 0.93 * T15 + -3.94 * Ne12 + 0.31 * Vx + 0.26 * maxO3v$

3.3 Intervalle de confiance

```
# Intervalle de confiance (à 95%) des coefficients
confint(reg_multiple)
```

A matrix: 6 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	32.27516866	92.7525159
T12	-2.48979050	2.6587052
T15	-1.55276369	3.4151849
Ne12	-5.98499300	-1.8983294
Vx	-0.02411048	0.6571717
maxO3v	0.07410465	0.4488183

3.4 Test d'hypothèse

```
summary(reg_multiple)
```

```
Call:
lm(formula = maxo3 ~ T12 + T15 + Ne12 + Vx + maxO3v)

Residuals:
    Min      1Q  Median      3Q     Max
-25.673  -8.006   0.501   5.896  25.755

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.51384   15.00405   4.166 0.000142 ***
T12          0.08446    1.27731   0.066 0.947581
T15          0.93121    1.23252   0.756 0.453953
Ne12        -3.94166    1.01388  -3.888 0.000338 ***
Vx           0.31653    0.16902   1.873 0.067761 .
maxO3v       0.26146    0.09296   2.812 0.007318 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13 on 44 degrees of freedom
Multiple R-squared:  0.7275,    Adjusted R-squared:  0.6966
F-statistic:  23.5 on 5 and 44 DF,  p-value: 2.026e-11
```

En utilisant une valeur seuil de 0,05, on rejette l'hypothèse nulle H_0: $B_j = 0$ pour les variables Ne12 et maxO3v, car leurs P-values sont inférieures à 0,05.

Cela signifie que ces variables sont statistiquement significatives à un niveau de confiance de 95%.

Cependant, on ne peut pas rejeter l'hypothèse nulle pour les variables T12 , T15 et Vx , car leurs P-values sont supérieures à 5%.

Cela signifie que ces variables ne sont pas statistiquement significatives à un niveau de confiance de 95 %.

3.5 Qualité du modèle

La qualité du modèle de la régression linéaire multiple est mesurée par le coefficient de détermination (R-Squared ou R^2), qui se définit comme la part de variation dans la variable y qui est expliquée par des variations dans les variables explicatives (souvent exprimé en %).

Formellement: $R^2 = 1 - \frac{SSR}{SST}$

où:

SSR représente la somme des carrés des résidus et **SST** la somme des écarts à la moyenne des valeurs observées.

Plus sa valeur est proche de 1, et plus l'adéquation entre le modèle et les données observées va être forte. Cependant, cette valeur est fortement influencée, entre autres, par le nombre de variables explicatives incluses dans la regression.

Le R^2_{ajust} (Adjusted R-Squared) va alors tenir compte de ce nombre et sera donc plus correct.

Formellement: $R^2_{adjusted} = 1 - \frac{SSR/n-k}{SST/n-1}$.

Alors,

Dans les résultats obtenus avec $summary(reg_multiple)leRcarrest : 0.7275etleRcarrajustestde : 0.6966.Ilestdoncde\mathbf{70\%}, cequiestplusquecorrect.$

Si on compare avec les valeurs trouvées en régression simple qui sont respectivement de 0.6151 et 0.6116, donc inférieur à 70%, on conclut que le modèle de régression multiple est plus ***intéressant****.*

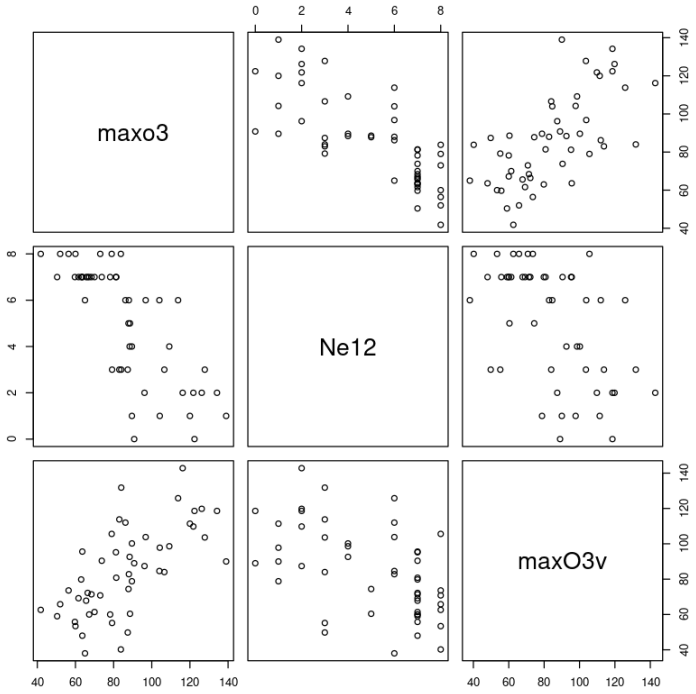
De même, cela indique que la variable T12 seule ne peut pas bien expliquer la variation de maxO3 et que les variables qu'on a rajoutées dans le cas multiple expliquent mieux la variation de maxO3.

3.6 Retirer les variables non significatives

D'après les résultats obentenus dans la question 4 par rapport à la P-value de chaque variable, on a constaté que T12, T15 et Vx sont des variables ***non significatives****, on peut donc les* ***retirer****.*

3.6.1 visualisation des variables significatives

```
dff_ozone = data.frame(maxO3, Ne12, maxO3v)
plot(dff_ozone)
```



3.7 Créer le nouveau modèle

On nomme notre nouveau modèle: nv-reg-mult

```
nv_reg_mult = lm(maxo3 ~ Ne12 + maxO3v)
nv_reg_mult
```

```
Call:
lm(formula = maxo3 ~ Ne12 + maxO3v)

Coefficients:
(Intercept)         Ne12       maxO3v
     84.551       -5.378        0.339
```

On redéfinit notre modèle comme suit:

$maxO3 = 84.55 - 5.38 * Ne12 + 0.34 * maxO3v$

3.8 Définir n et k

Dans notre cas:

- *n= 50 observations (n reste comme il est),*
- *k= 2 varibles explicatives; on le trouve avec ces deux lignes de code:*

```
k = nv_reg_mul$rank-1
cat('Nombre de variables explicatives : k=', k)
```

3.9 Calculer les résidus studentisés

Les résidus standardisés sont une mesure importante dans la vérification des hypothèses de base des modèles linéaires. Si les résidus standardisés sont très éloignés de zéro pour certaines observations, cela peut indiquer une violation de ces hypothèses ou la présence de points abérrants dans les données.

La formule pour calculer les résidus studentisés est la suivante :
$\text{résidus}_{s}tudentises = \frac{residu}{(racine(varianceduresidu*(1-influence"pointlevier"))}$

```
# Méthode 1
residuals = residuals(nv_reg_mult)
variance_residuals = var(residuals) * (length(residuals) - 1) / df.residual(nv_reg
    _mult)
residus_studentises1 = residuals / sqrt(variance_residuals*(1 - hatvalues(nv_reg_
    mult)))
residus_studentises1
```

```
1:      -1.18734398365277 2:      -0.552252457210504 3:      0.130757957167447 4:      0.152990520072306 5:
      0.569200152473702 6:      -0.202356983303802 7:      2.24945889476824 8:      0.822729616954312 9:
      1.48086733944581 10:      -1.57981154041567 11:      -0.0126532526377764 12:      0.563300676503301 13:
      0.897637731835963 14:      1.81733802498235 15:      -0.655167003147869 16:      0.0322891866858626 17:
      1.53568664071732 18:      -0.00320424487622826 19:      0.366194500421134 20:      0.704106656017736 21:
      -1.26414502638652 22:      -0.37210044562978 23:      0.0282527396616604 24:      -1.875628505113 25:
      -0.620703562659155 26:      0.17124904650488 27:      -0.546503413634592 28:      -0.320675889088662 29:
      0.949363422701758 30:      -0.308941974131871 31:      0.165231622811397 32:      -2.24874078185593 33:
      -1.80911126085669 34:      -0.460920027690909 35:      -0.891682917795432 36:      -0.281902312838475 37:
      0.233996678922768 38:      -0.182762851978861 39:      0.727837035468222 40:      0.814903575495996 41:
      -0.473651503514237 42:      0.531721209933437 43:      0.787791753831578 44:      -0.819393890979457 45:
      1.71488700900626 46:      -0.448518055180621 47:      2.20210498565613 48:      -0.759604433797985 49:
      -1.24003021987137 50:      -0.615896497591558
```

La commande **rstandard()** en R calcule les résidus standardisés, également appelés résidus studentisés. Ces résidus mesurent la différence entre les valeurs observées et les valeurs prédites par un modèle statistique, divisée par l'estimation de l'écart-type de l'erreur.

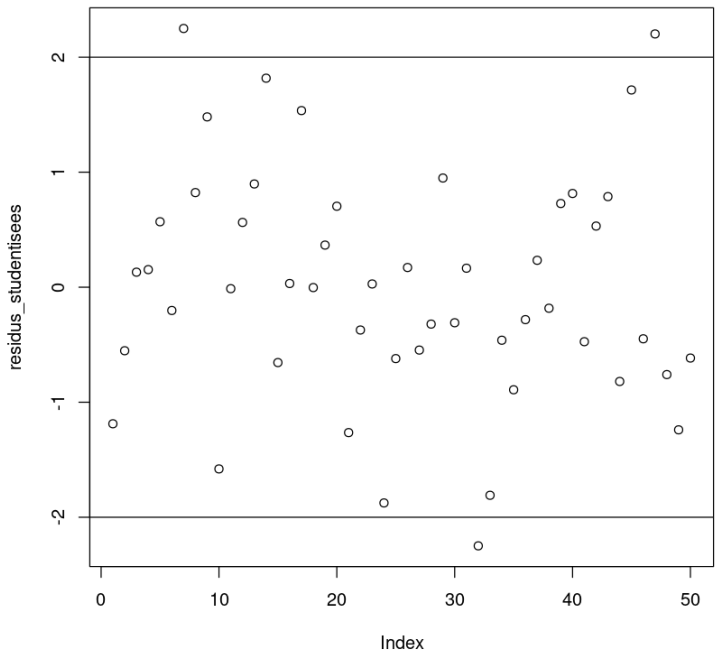
```
# Méthode 2
residus_studentises <- rstandard(nv_reg_mult)
residus_studentises
```

```
1:      -1.18734398365277 2:      -0.552252457210504 3:      0.130757957167447 4:      0.152990520072306 5:
      0.569200152473702 6:      -0.202356983303802 7:      2.24945889476824 8:      0.822729616954312 9:
      1.48086733944581 10:      -1.57981154041567 11:      -0.0126532526377764 12:      0.563300676503301 13:
      0.897637731835963 14:      1.81733802498235 15:      -0.655167003147869 16:      0.0322891866858626 17:
      1.53568664071732 18:      -0.00320424487622826 19:      0.366194500421134 20:      0.704106656017736 21:
      -1.26414502638652 22:      -0.37210044562978 23:      0.0282527396616604 24:      -1.875628505113 25:
      -0.620703562659155 26:      0.17124904650488 27:      -0.546503413634592 28:      -0.320675889088662 29:
      0.949363422701758 30:      -0.308941974131871 31:      0.165231622811397 32:      -2.24874078185593 33:
      -1.80911126085669 34:      -0.460920027690909 35:      -0.891682917795432 36:      -0.281902312838475 37:
      0.233996678922768 38:      -0.182762851978861 39:      0.727837035468222 40:      0.814903575495996 41:
      -0.473651503514237 42:      0.531721209933437 43:      0.787791753831578 44:      -0.819393890979457 45:
      1.71488700900626 46:      -0.448518055180621 47:      2.20210498565613 48:      -0.759604433797985 49:
      -1.24003021987137 50:      -0.615896497591558
```

3.10 Valeurs aberrantes

Afin d'identifier d'éventuelles valeurs abrrantes dans les données, on utilise la représentation graphique suivante:

```r
plot(residus_studentises)
abline(-2,0)
abline(2,0)
```



4 Conclusion

Cette étude avait pour objectif d'étudier les facteurs influençant les concentrations d'ozone troposphérique à partir des données météorologiques et de pollution atmosphérique du jeu ozone.

Différents modèles de régression linéaire simple et multiple ont été ajustés, en prenant la concentration d'ozone maxO3 comme variable dépendante à expliquer par différents prédicteurs environnementaux.

La régression linéaire simple a permis de quantifier les effets d'un facteur clé qui est la température à midi (variable T12), qui ressort comme un déterminant important de niveau d'ozone.

La régression multiple incluant plusieurs prédicteurs météorologiques (vent, humidité, pression, etc) se révèle plus précise, avec un R2 ajusté de 70%. Les validations sur les résidus confirment la qualité de l'ajustement.

Ces résultats illustrent l'intérêt des méthodes de régression linéaire pour modéliser les phénomènes complexes de pollution de l'air à partir de données environnementales. Les modèles développés permettent d'expliquer et de prédire les taux d'ozone de manière fiable dans ce contexte d'étude.

Cette analyse ouvre des perspectives pour affiner ces modèles statistiques et les appliquer à d'autres contextes géographiques et climatiques.