ECOLE NATIONALE DES SCIENCES APPLIQUEES

OUJDA

PROJET EN MODELISATION STOCHASTIQUE

REALISE PAR:

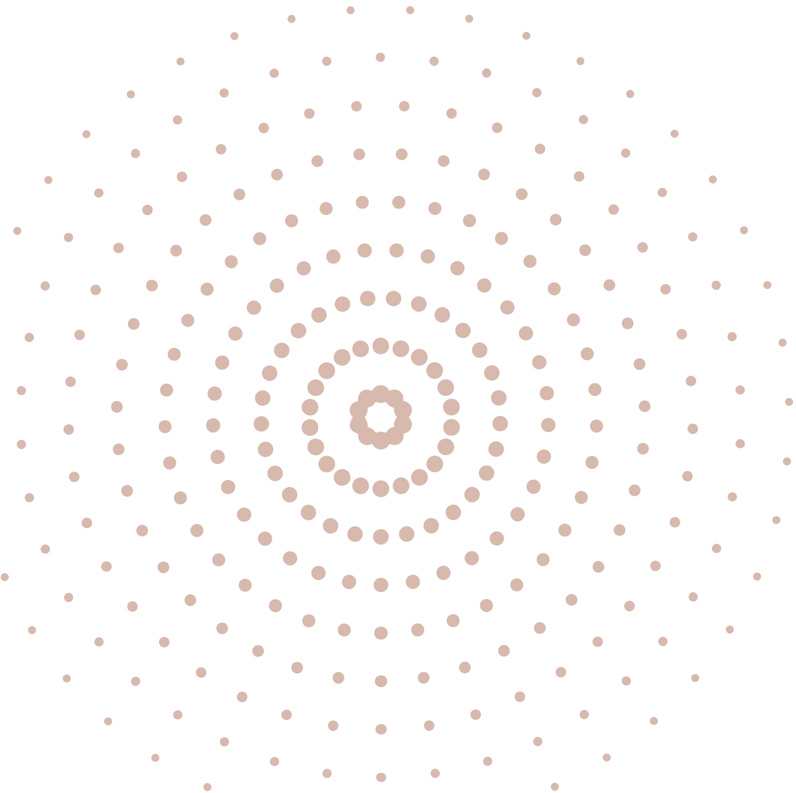
SIDIBE MAMADOU
LAHBOUCHI HANAE
BELAHSEN YMANE

ENCADRE PAR: Mme. RACHIDA ELMEHDI

ANNEE 2023

Abstract:

Cet ensemble de données provient à l'origine de laboratoire de mathématiques appliquées de l'Agrocampus Ouest. L'ozone est un polluant photochimique, dont de nombreux instituts cherchent à prévoir les pics pour prévenir les populations. La pollution automobile, l'absence de vent et la chaleur comptent parmi les facteurs qui accroissent le taux de pollution. Sur la base de certaines mesures incluses dans l'ensemble de données.
Plusieurs contraintes ont été placées sur la sélection de ces instances à partir d'une plus grande base de données.



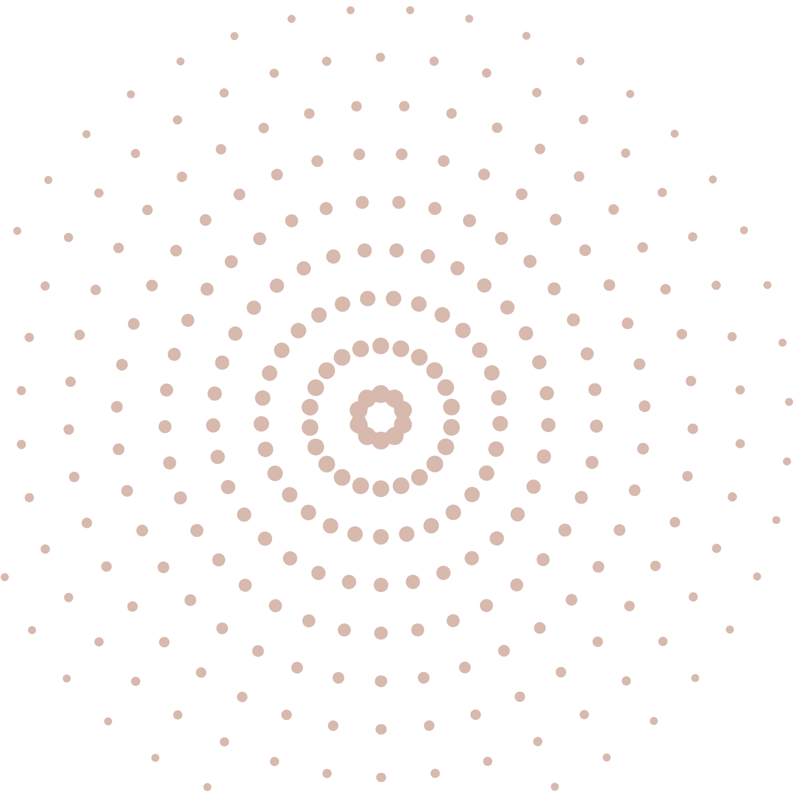
Sommaire:

Introduction:

Ce rapport relate le travail réalisé par le trinôme Mamadou SIDIBE, BELAHSEN Ymane et LAHBOUCHI Hanae dans le cadre de leur projet du module modélisation stocastique, en tant qu'étudiants ingénieurs en Data Science et Cloud Computing, à l'école Nationale des Sciences Appliquées d'Oujda(ENSAO).

Le projet concerne la réalisation d'un modèle prédictif qui cherchent à prévoir les pics pour prévenir les populations ce projet s'orientant via un entraînement préalable d'une régression linéaire multiple.

L'apprentissage supervisé, ou supervised machine learning en anglais, est la technique d'intelligence artificielle retenue pour réaliser cette tâche et rendre ainsi la prédiction simple et rapide.



1 | Premier chapitre

Dans ce chapitre, des informations supplémentaires utiles sont rapportées.

Les chapitres sont généralement subdivisés en trois grandes parties: une partie introductif, une partie théorique et finalement une partie pratique.

1.1 Généralités

L'ozone est un polluant photochimique, dont de nombreux instituts cherchent à prévoir les pics pour prévenir les populations. La pollution automobile, l'absence de vent et la chaleur comptent parmi les facteurs qui accroissent le taux de pollution.

Sur la base de certaines mesures incluses dans l’ensemble de données. Plusieurs contraintes ont été placées sur la sélection de ces instances à partir d’une plus grande base de données.

1.2 Présentation de la base de données

On considère à cet effet un jeu de données issu du Laboratoire de mathématiques appliquées de l'Agrocampus Ouest qui contient 112 données recueillies à Rennes durant l'été 2001. On y trouve les 14 variables suivantes :

- obs : mois-jour ;
- maxO3 : teneur maximale en ozone observée sur la journée (en µ\gr/m3µ\gr/m3) ;
- T9, T12, T15 : température observée à 9 h, 12 h et 15 h ;
- Ne9, 12, Ne15 : nébulosité observée à 9 h, 12 h et 15 h ;
- Vx9, Vx12, Vx15 : composante est-ouest du vent à 9 h, 12 h et 15 h ;
- maxO3v : teneur maximale en ozone observée la veille ;
- vent : orientation du vent à 12 h ;
- pluie : occurrence ou non de précipitations.

On souhaite étudier le lien entre le pic d'ozone journalier et un certain nombre de facteurs potentiellement explicatifs afin de proposer un modèle de régression permettant de prévenir la population.

Chargement du dataset

```
Ozone <- read.table("Ozone.txt", header=TRUE, sep=";", dec=",")
```

Affichage du dataset et vue d'ensemble

```
str(Ozone)
```

```
'data.frame':   112 obs. of  14 variables:
 $ obs   : int  601 602 603 604 605 606 607 610 611 612 ...
 $ maxO3 : int  87 82 92 114 94 80 79 79 101 106 ...
 $ T9    : num  15.6 17 15.3 16.2 17.4 17.7 16.8 14.9 16.1 18.3 ...
 $ T12   : num  18.5 18.4 17.6 19.7 20.5 19.8 15.6 17.5 19.6 21.9 ...
 $ T15   : num  18.4 17.7 19.5 22.5 20.4 18.3 14.9 18.9 21.4 22.9 ...
 $ Ne9   : int  4 5 2 1 8 6 7 5 2 5 ...
 $ Ne12  : int  4 5 5 1 8 6 8 5 4 6 ...
 $ Ne15  : int  8 7 4 0 7 7 8 4 4 8 ...
 $ Vx9   : num  0.695 -4.33 2.954 0.985 -0.5 ...
 $ Vx12  : num  -1.71 -4 1.879 0.347 -2.954 ...
 $ Vx15  : num  -0.695 -3 0.521 -0.174 -4.33 ...
 $ maxO3v: int  84 87 82 92 114 94 80 99 79 101 ...
 $ vent  : chr  "Nord" "Nord" "Est" "Nord" ...
 $ pluie : chr  "Sec" "Sec" "Sec" "Sec" ...
```

1.3 Normalisation base de données

En normalisant les données on aura de meilleur résultats pour la régularisation linéaire et ça nous permettra de réglé le problème de variables qui n'ont pas la même unités

On vas utilisé uniquement les données numerique donc on enlève la variable vent et pluie

```
data = data.frame(Ozone$max03,Ozone$obs, Ozone$max03v, Ozone$Ne12, Ozone$Ne15, Ozone$Ne9, Ozone$T12, Ozone$T15, Ozone$T9,
Ozone$Vx12, Ozone$Vx15, Ozone$Vx9)
```

On procède à la normalisation et on stocke dans un dataframe nomé ozone

```
ozone <- as.data.frame(scale(data))
```

```
str(ozone)
```

```
'data.frame':   112 obs. of  12 variables:
 $ Ozone.max03 : num  -0.1172 -0.2946 0.0602 0.8407 0.1311 ...
 $ Ozone.obs   : num  -1.47 -1.46 -1.45 -1.45 -1.44 ...
 $ Ozone.max03v: num  -0.2324 -0.1263 -0.3031 0.0505 0.8285 ...
 $ Ozone.Ne12  : num  -0.44606 -0.00783 -0.00783 -1.76078 1.30689 ...
 $ Ozone.Ne15  : num  1.359 0.93 -0.356 -2.071 0.93 ...
 $ Ozone.Ne9   : num  -0.3578 0.0275 -1.1286 -1.5139 1.1836 ...
 $ Ozone.T12   : num  -0.749 -0.774 -0.971 -0.452 -0.254 ...
 $ Ozone.T15   : num  -0.9331 -1.0876 -0.6903 -0.0282 -0.4917 ...
 $ Ozone.T9    : num  -0.884 -0.436 -0.98 -0.692 -0.308 ...
 $ Ozone.Vx12  : num  -0.0354 -0.8545 1.2485 0.7005 -0.4805 ...
 $ Ozone.Vx15  : num  0.354 -0.466 0.787 0.54 -0.939 ...
 $ Ozone.Vx9   : num  0.725 -1.183 1.583 0.835 0.271 ...
```

2 | Deuxième chapitre

2.1 Notions Générales:

La modélisation :

La modélisation mathématique est un indispensable outil utilisé dans des domaines divers tels la physique, la biologie, l'ingénierie et même les sciences humaines (psychologie, économie, sciences politiques…). Ainsi, elle est définie comme « l'art (ou la science, selon le point de vue) de représenter ou de transformer une réalité en des modèles abstraits accessibles à l'analyse et au calcul » selon Grégoire Allaire. C'est donc la description d'un système physique en utilisant des concepts, techniques et théories mathématiques afin d'obtenir un modèle adéquat permettant d'effectuer des prédictions ou opérations dans le monde réel. Ces modèles sont hypothétiques, modifiables et adéquats pour certains problèmes dans certaines situations et sont notés par une fonction f inconnue à déterminer par la suite.

La modélisation statistique :

La modélisation statistique est la représentation de bases de données observées ; qu'on veut étudier ; par des modèles théoriques qui établissent une relation mathématique entre une ou plusieurs variables aléatoires et d'autres variables non aléatoires afin de décrire au maximum ces bases de données. Un modèle statistique est donc un modèle mathématique qui intègre un ensemble d'hypothèses statistiques concernant la génération de données d'échantillonnage (et de données similaires provenant d'une population plus importante). Un modèle statistique représente, souvent sous une forme considérablement idéalisée, le processus générateur de données. De plus, tous les tests d'hypothèses statistiques et tous les estimateurs statistiques sont dérivés de modèles statistiques. Plus généralement, les modèles statistiques font partie des fondements de l'inférence statistique.

La modélisation stochastique VS la modélisation déterministe :

La modélisation déterministe diffère de la modélisation stochastique puisque la première utilise des variables qui ne prennent pas en considération la partie aléatoire dans la détermination du modèle, tandis que la modélisation stochastique présente des données et prédit des résultats qui tiennent compte de certains niveaux d'imprévisibilité ou d'aléatoire et donne une extrême importance à cette erreur statistique qui peut être due à des facteurs inconnus ou impossibles à mesurer.

Les étapes principales de la modélisation :

Pour modéliser, représenter et bien décrire une base de données, il faut suivre ce qui suit:

1. Définir un modèle avec un nombre fini de paramètres (les coefficients)
2. Définir les variables explicatives et celles à expliquer :

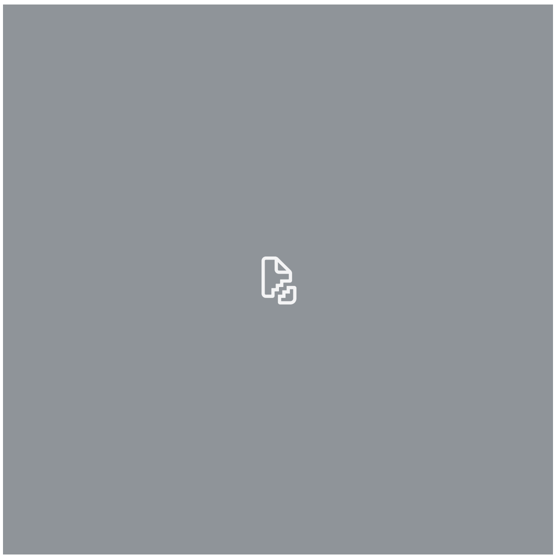
- Variable explicative : c'est une variable non aléatoire et indépendante. Elle est utilisée dans le but d'expliquer, de décrire ou de prédire. En général, les variables explicatives sont indépendantes entre elles.

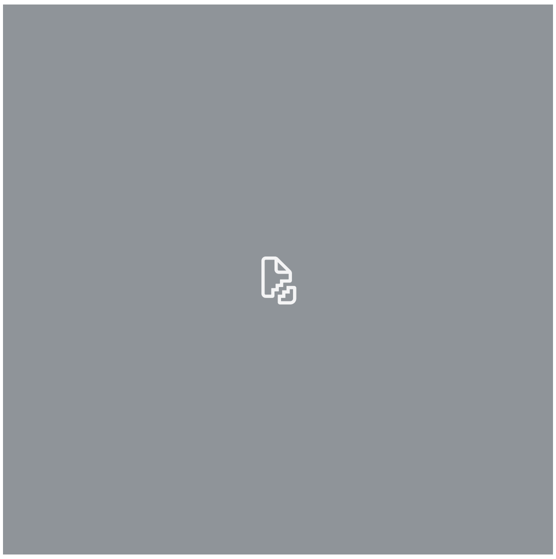
- Variables à expliquer (ou variable de réponse) : c'est une variable qu'on cherche à décrire et à expliquer à partir de variables explicatives. Elle est également une variable dépendante.

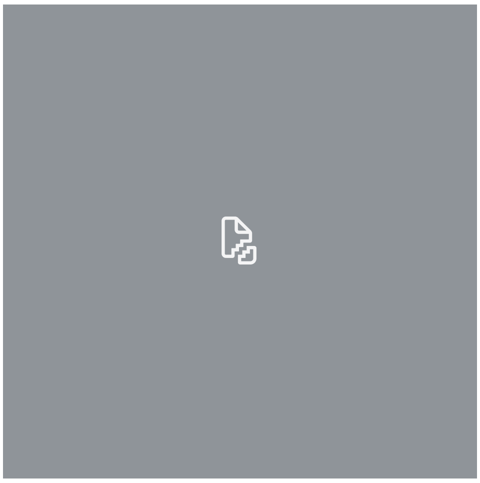
3. Estimer les paramètres du modèle.
4. Vérifier la qualité de l'ajustement du modèle, et comparer les différents modèles, par exemple en découpant les données en un échantillon d'apprentissage et un échantillon de test.
5. Effectuer des prédictions.

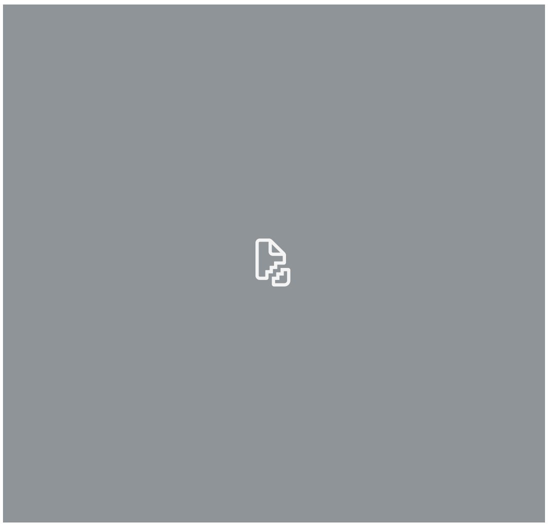
2.2 Description des données

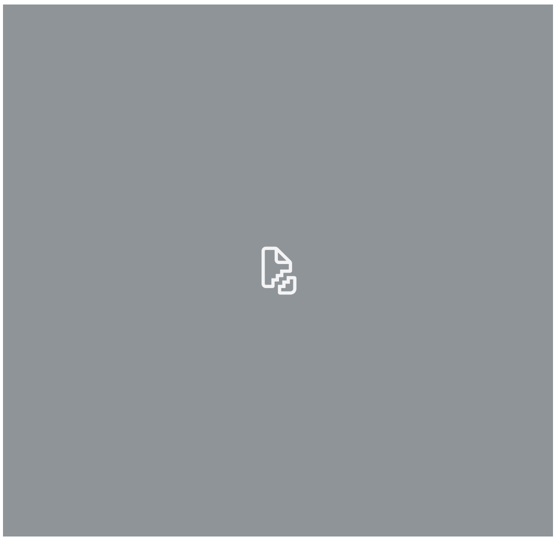
Visualisation de la base de données :Histogramme des variables

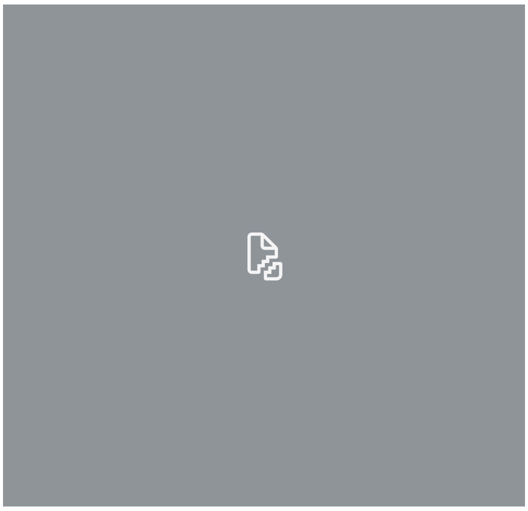


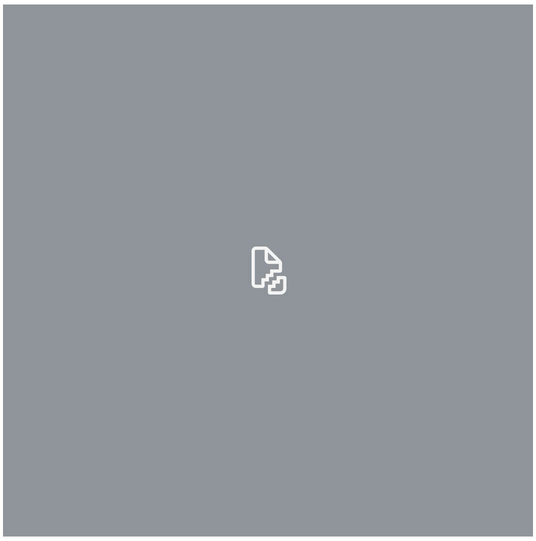


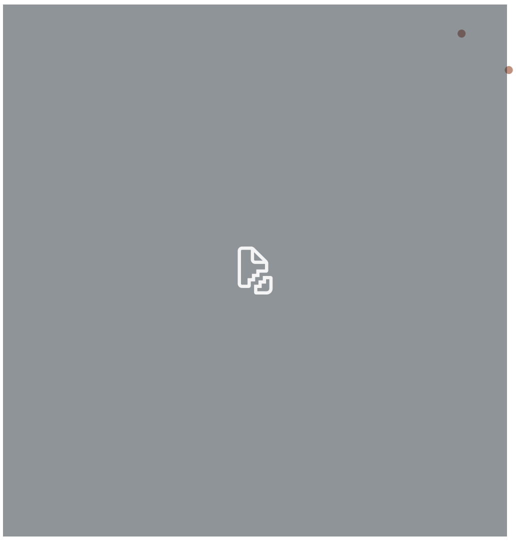


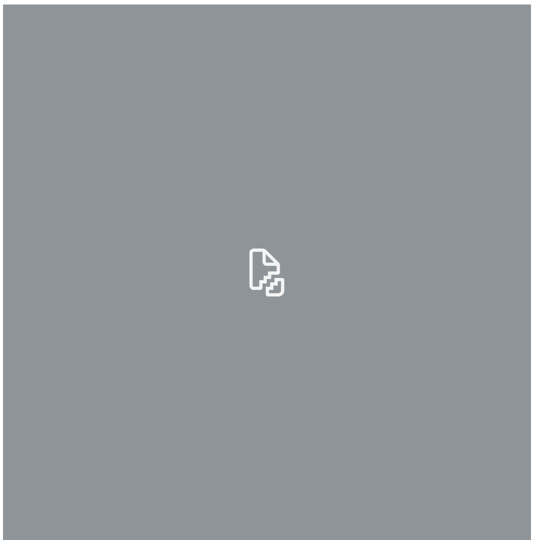


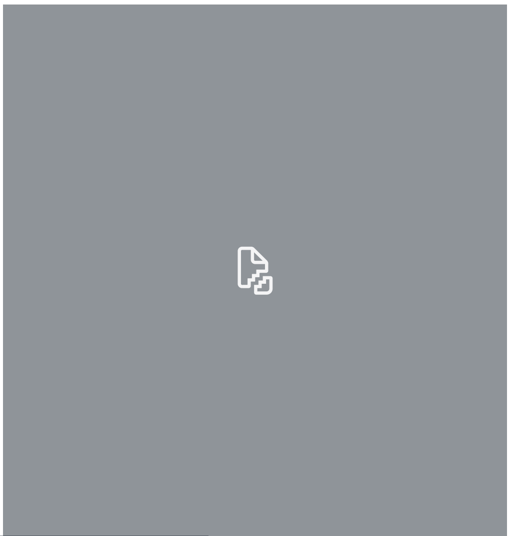


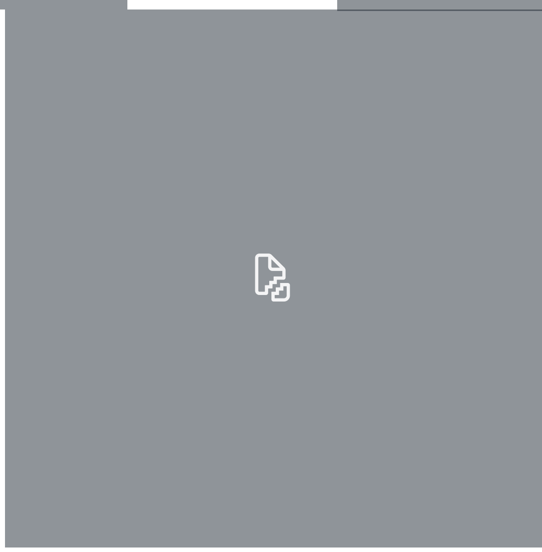






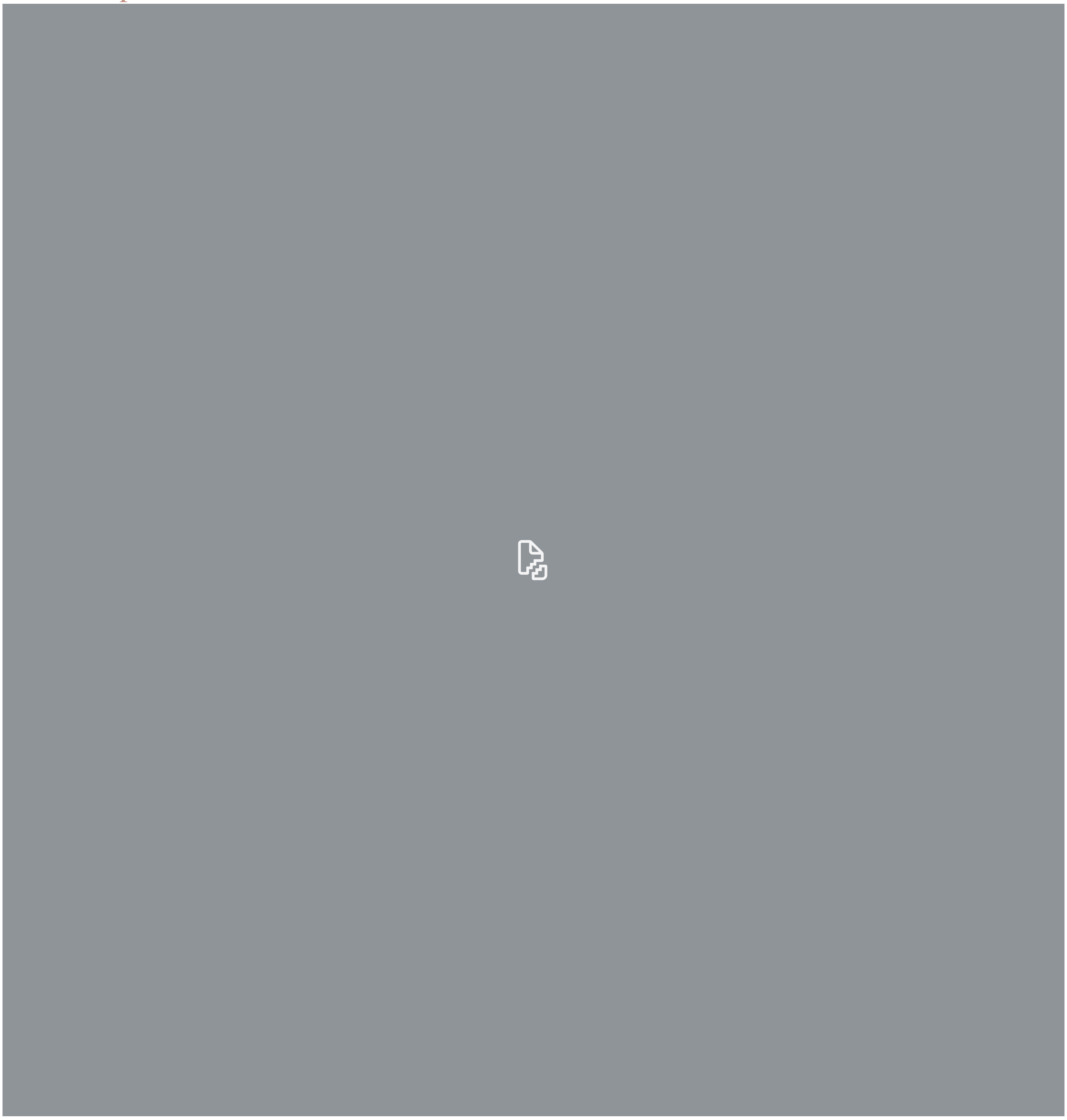






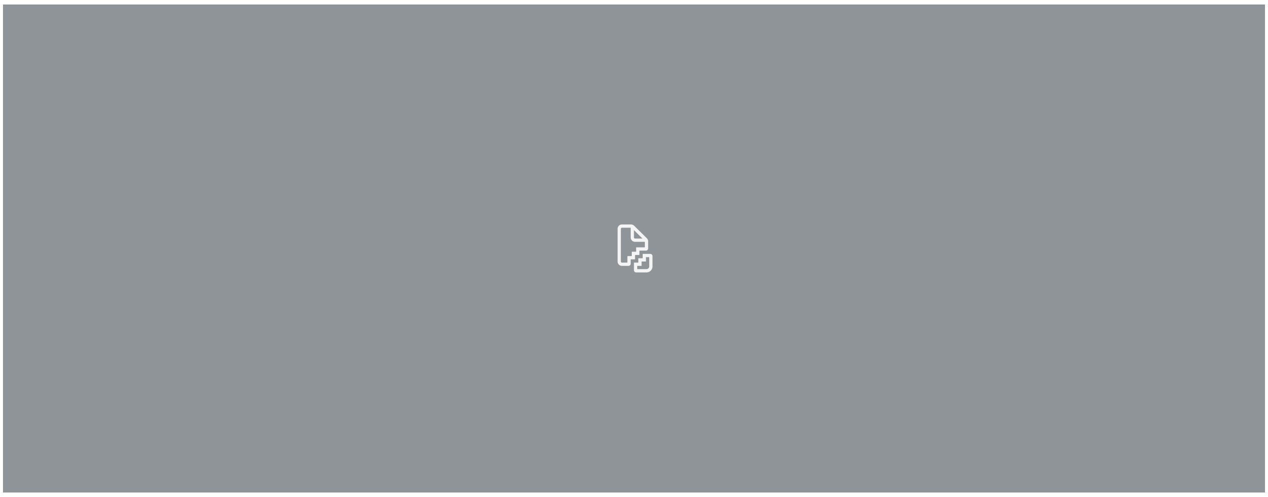
On a exclu les deux variables qualitatives vent et pluie mais on va les intégrer par la suite pour faire l'analyse de la variance covariance ANOVA.

Le multi-plot



Le multi-plot des variables nous indique l’existence d’une tendance entre certaines variables explicatives et principalement entre les trois variables T9, T15 et T12 ansi qu'entre Vx9, Vx12 et Vx15.

Matrice de corrélation des variables



En dressant la matrice de corrélation on constate qu'il y a une forte corrélation entre les variables déja citées mais qui n'est pas égale à 1.

- ### Vérification de la corrélation entre chaque variable

explicative et les charges : Le coefficient de corrélation linéaire R entre les deux variables xj et y :

$$-1 \leqslant R \leqslant 1$$

$|R| > 0.7$: Les variable xj et y sont fortement corrélées.
$|R| > 0.5$: Les variable xj et y sont corrélées.
$|R| < 0.5$: Les variable xj et y sont faiblement corrélées.
$|R| = 0$: Les variable xj et y ne sont pas corrélées.

- ### corrélation avec les observation

```
> cor_obs<-cor(ozone$maxO3,ozone$obs)
> print(cor_obs)
[1] -0.2237124
> #Les variables maxO3 et obs sont faiblement corrélées.
```

- Corrélation avec les température observée à 9 h, 12 h et 15 h

```
> cor_T9<-cor(ozone$maxO3,ozone$T9)
> print(cor_T9)
[1] 0.6993865
> #Les variables maxO3 et T9 sont corrélées.
>
> cor_T12<-cor(ozone$maxO3,ozone$T12)
> print(cor_T12)
[1] 0.7842623
> #Les variables maxO3 et T12 sont fortement corrélées.
>
>
> cor_T15<-cor(ozone$maxO3,ozone$T15)
> print(cor_T15)
[1] 0.77457
> #Les variables maxO3 et T15 sont fortement corrélées.
```

- corrélation avec les nébulosités observées à 9 h, 12 h et 15 h

```
> cor_Ne9<-cor(ozone$maxO3,ozone$Ne9)
> print(cor_Ne9)
[1] -0.6217042
> #Les variables maxO3 et Ne9 sont faiblement corrélées.
> cor_Ne12<-cor(ozone$maxO3,ozone$Ne12)
> print(cor_Ne12)
[1] -0.6407513
> #Les variables maxO3 et Ne12 sont faiblement corrélées.
> cor_Ne15<-cor(ozone$maxO3,ozone$Ne15)
> print(cor_Ne15)
[1] -0.4783021
> #Les variables maxO3 et Ne15 sont faiblement corrélées.
```

- corrélation avec la composante est-ouest du vent à 9 h, 12 h et 15 h

```
> cor_vx9<-cor(ozone$maxO3,ozone$Vx9)
> print(cor_vx9)
[1] 0.5276234
> #Les variables maxO3 et vx9 sont  corrélées.
>
> cor_vx12<-cor(ozone$maxO3,ozone$Vx12)
> print(cor_vx12)
[1] 0.4307959
> #Les variables maxO3 et vx12 sont faiblement corrélées.
>
> cor_vx15<-cor(ozone$maxO3,ozone$Vx15)
> print(cor_vx15)
[1] 0.3918989
> #Les variables maxO3 et vx15 sont faiblement corrélées.
```

- corrélation avec la teneur maximale en ozone observée la veille

```
> cor_maxo3v<-cor(ozone$maxO3,ozone$maxO3v)
> print(cor_maxo3v)
[1] 0.684516
> #Les variables maxO3 et maxo3v sont fortement corrélées.
```

La corrélation entre la cible et une variable descriptive dans un modèle de régression linéaire peut vous indiquer si la variable descriptive est importante pour prédire la cible. Plus la corrélation est forte, plus la variable descriptive est importante pour prédire la cible. Si la corrélation est positive, cela signifie que lorsque la variable descriptive augmente, la cible augmente également. Si la corrélation est négative, cela signifie que lorsque la variable descriptive augmente, la cible diminue.

Il est important de noter que la corrélation ne signifie pas causalité. Il est possible que d'autres variables soient en jeu et influencent à la fois la variable descriptive et la cible. Il est également important de regarder la force de la corrélation, car une corrélation faible ne sera peut-être pas très utile pour prédire la cible.

2.3 Régression Linéaire Multiple

- Modèle

Le modèle de la régression linéaire multiple s'écrit :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_P x_{ip} + \varepsilon_i \quad , \qquad \begin{array}{l} i = 1,2, \dots, n \\ j = 1,2, \dots, p \end{array}$$

$$\mathrm{y_i} = \beta_0 + \sum_{\mathrm{j=1}}^{\mathrm{P}} \beta_{\mathrm{j}} \mathrm{x_{ij}} + \varepsilon_{\mathrm{i}} \quad , \qquad i = 1,2, \dots, n$$

La forme matricielle du modèle de régression multiple :

$$Y = X\beta + \varepsilon$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1j} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

La taille de la matrice X est : $(n, \ p+1)$

La régression linéaire multiple repose sur les mêmes hypothèses de la régression linéaire simple, plus l'hypothèse d'indépendance entre les variables explicatives qui est exprimée par : $cov(x_j, x_{j'}) = 0 \quad , \quad j = 1,2, \dots, p$ et $j \neq j'$

- Estimation du modèle

Pareil à la régression linéaire simple, l'estimation s'effectue par la méthode des moindres carrées ordinaires qui consiste dans :

$$min\, S\left(\beta_0, \beta_1, \dots, \beta_j, \dots, \beta_P\right) = min \sum_{i=1}^{n} \varepsilon_i^2 = min \ (Y - X\beta)^t \cdot (Y - X\beta)$$

$S(\beta)$ atteint son minimum en $\hat{\beta} = (X^t \cdot X)^{-1} \cdot X^t \cdot Y$

Les valeurs ajustées et les résidus sont trouvés respectivement par :

$$\hat{Y} = X\hat{\beta} = X\,(X^t \cdot X)^{-1} \cdot X^t \cdot Y$$

$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta} = Y - X\,(X^t \cdot X)^{-1} \cdot X^t \cdot Y$$

- Vérification du modèle

- Coefficient de détermination :

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} = 1 - \frac{S_{\hat{\varepsilon}}^2}{S_y^2}$$

avec : $S_{\hat{y}}^2$: La variance de la régression

S_y^2 : La variance totale

- Coefficient de détermination ajusté :

$$R_{adj}^2 = 1 - (1 - R^2)\frac{n-1}{n-P-1}$$

On utilise R^2 ajusté pour éliminer l'effet de surparamétrisation dans la régression multiple. Or R_{adj}^2 tend vers R^2 lorsque n tend vers $+\infty$.

- Tests d'hypothèse sur les paramètres :

1. Test de Student :

$$H_0\colon \beta_j = \beta_{j_0} \quad VS \quad H_1\colon \beta_j \neq \beta_{j_0} \quad , \quad j \in \{0,1,2,\dots,\mathrm{p}\} \quad , \quad \beta_{j_0} = 0$$

La règle de décision de ce test :

Si $|t_{cal}| = \frac{\hat{\beta}_j - \beta_{j_0}}{\hat{s}_{\hat{\beta}_j}} > t_{n-(p+1),1-\frac{\alpha}{2}}$ on rejette H_0 au niveau de risque α.

2. Test Fisher d'hypothèse global :

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta j = \cdots = \beta_P = 0 \quad VS \quad H_1\colon \exists\, j \; tel\; que \;\; \beta_1 \neq 0$$

La règle de décision de ce test :

Si $|F_{cal}| = \left|\frac{\frac{SCReg}{p}}{SCE}\right| > F_{p,\;n-(p+1)}$ on rejette H_0 au niveau de risque α.

- Prédiction

- Prédiction ponctuelle :

Pour une nouvelle valeur x_{i^*} de l'observation i^* :

$$y_{i^*} = \hat{y}(x_{i^*}) = x_{i^*}\hat{\beta}$$

- Prédiction par intervalle :

$$y_{i^*} \in \left[\hat{y}_{i^*} \pm \hat{\sigma}_{\hat{\varepsilon}_{i^*}}\, t_{n-(p+1),1-\frac{\alpha}{2}}\right]$$

$$\Rightarrow \qquad y_{i^*} \in \left[\hat{y}_{i^*} \pm \hat{\sigma}_{\varepsilon}\sqrt{x_{i^*}(X^t \cdot X)^{-1} \cdot x_{i^*}^t + 1}\; t_{n-(p+1),1-\frac{\alpha}{2}}\right]$$

- Point levier

Pour déterminer les points leviers nous rappelons l'expression de la variance des valeurs ajustées :

$$var(\hat{y}) = var\left(X\,(X^t \cdot X)^{-1} \cdot X^t \cdot Y\right)$$

Nous définissons le levier de l'observation i par : $h_i = x_i(X^t \cdot X)^{-1} \cdot x_i^t$

$$var(\hat{y}_i) = h_i\sigma_\varepsilon^2 \quad , \text{ où } \quad 0 \leq h_i \leq 1$$

Donc, plus h_i est élevé, plus le y_i contribue à sa valeur ajustée $\hat{y}_i$. Le levier est important lorsque $h_i > p\,/\,n$ et un point i est considéré généralement un point levier si $h_i > 2p\,/\,n$.

La distance de Cook permet de mesurer l'influence d'un point sur la régression peut être utilisée pour détecter les points aberrants qui peuvent être considérés comme des leviers sur la régression.

- Régression Linéaire Multiple (Application)

Nous avons 12 variables quantitatifs et 2 variables qualitatifs nous allons procéder à une première régression avec les variables qualitatifs (obs,T9,T12,Ne12,Vx12, …) pour construire un modèle de prediction de la variable maxO3 : teneur maximale en ozone observée sur la journée (en µ\gr/m3µ\gr/m3) :

```
str(ozone)
```

```
'data.frame':   112 obs. of  12 variables:
 $ Ozone.maxO3 : num  -0.1172 -0.2946 0.0602 0.8407 0.1311 ...
 $ Ozone.obs   : num  -1.47 -1.46 -1.45 -1.45 -1.44 ...
 $ Ozone.maxO3v: num  -0.2324 -0.1263 -0.3031 0.0505 0.8285 ...
 $ Ozone.Ne12  : num  -0.44606 -0.00783 -0.00783 -1.76078 1.30689 ...
 $ Ozone.Ne15  : num  1.359 0.93 -0.356 -2.071 0.93 ...
 $ Ozone.Ne9   : num  -0.3578 0.0275 -1.1286 -1.5139 1.1836 ...
 $ Ozone.T12   : num  -0.749 -0.774 -0.971 -0.452 -0.254 ...
 $ Ozone.T15   : num  -0.9331 -1.0876 -0.6903 -0.0282 -0.4917 ...
 $ Ozone.T9    : num  -0.884 -0.436 -0.98 -0.692 -0.308 ...
 $ Ozone.Vx12  : num  -0.0354 -0.8545 1.2485 0.7005 -0.4805 ...
 $ Ozone.Vx15  : num  0.354 -0.466 0.787 0.54 -0.939 ...
 $ Ozone.Vx9   : num  0.725 -1.183 1.583 0.835 0.271 ...
```

```
ll1 <- lm(ozone$maxO3~ozone$obs+ozone$T9+ozone$T12+ozone$T15+ozone$Ne9
 +ozone$Ne12+ozone$Ne15+ozone$Vx9+ozone$Vx12+ozone$Vx15+ozone$maxO3v)
summary(ll1)
```

```
Call:
lm(formula = ozone$maxO3 ~ ozone$obs + ozone$T9 + ozone$T12 +
    ozone$T15 + ozone$Ne9 + ozone$Ne12 + ozone$Ne15 + ozone$Vx9 +
    ozone$Vx12 + ozone$Vx15 + ozone$maxO3v)

Residuals:
    Min      1Q  Median      3Q     Max
-53.646  -8.336   0.422   8.167  38.881

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.28274   17.03940   1.308   0.1940
ozone$obs    -0.01380    0.01434  -0.963   0.3380
ozone$T9     -0.42591    1.20229  -0.354   0.7239
ozone$T12     2.41283    1.44722   1.667   0.0986 .
ozone$T15     0.70110    1.15459   0.607   0.5451
ozone$Ne9    -2.21296    0.93890  -2.357   0.0204 *
ozone$Ne12   -0.22476    1.38326  -0.162   0.8713
ozone$Ne15    0.28274    1.00841   0.280   0.7798
ozone$Vx9     0.69004    0.95110   0.726   0.4698
ozone$Vx12    0.17085    1.06553   0.160   0.8729
ozone$Vx15    0.57299    0.92994   0.616   0.5392
ozone$maxO3v  0.34580    0.06324   5.468 3.35e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.37 on 100 degrees of freedom
Multiple R-squared:  0.766,	Adjusted R-squared:  0.7403
F-statistic: 29.76 on 11 and 100 DF,  p-value: < 2.2e-16
```

La p-value du test de Fisher rejette l'hypothèse nulle qui signifie que tous les coefficients de la régression, à l'exception de la constante, sont nuls. De ce fait, les coefficients ne sont pas tous nuls. Les résultats indiquent cependant que les coefficients et des variables **obs, T9, T12, T15, Ne12, Ne15, Vx9, Vx12, Vx15 et l'intercept** ne sont pas significativement différents de 0, nous établissons alors une nouvelle régression sans la variable **Vx12** la moins significative selon l'algorithme backward qui commence par le modèle complet et puis l'élimination des variables les moins significatives.

```
ll2 <- lm(ozone$maxO3~ozone$obs+ozone$T9+ozone$T12+ozone$T15+ozone$Ne9
 +ozone$Ne12+ozone$Ne15+ozone$Vx9+ozone$Vx15+ozone$maxO3v)
summary(ll2)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   22.39856   16.94177   1.322    0.189    
ozone$obs     -0.01349    0.01413  -0.954    0.342    
ozone$T9      -0.37718    1.15763  -0.326    0.745    
ozone$T12      2.36242    1.40583   1.680    0.096 .  
ozone$T15      0.70525    1.14872   0.614    0.541    
ozone$Ne9     -2.20853    0.93396  -2.365    0.020 *  
ozone$Ne12    -0.27269    1.34405  -0.203    0.840    
ozone$Ne15     0.27901    1.00327   0.278    0.782    
ozone$Vx9      0.76033    0.83997   0.905    0.368    
ozone$Vx15     0.66258    0.73978   0.896    0.373    
ozone$maxO3v   0.34514    0.06280   5.496 2.92e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 

Residual standard error: 14.3 on 101 degrees of freedom
Multiple R-squared:  0.7659,	Adjusted R-squared:  0.7428 
F-statistic: 33.05 on 10 and 101 DF,  p-value: < 2.2e-16
```

Nous ôtons encore les variables qui ont un coefficient significativement nul et nous faisons a nouveau le calcul du model jusqu'à obtenir un bon globalement signification et qui satisfait .

```
ll <-lm(ozone$maxO3~+ozone$T12+ozone$Ne9+ozone$Vx9+ozone$maxO3v)
summary(ll)
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    12.63131   11.00088   1.148 0.253443    
ozone$T12       2.76409    0.47450   5.825 6.07e-08 ***
ozone$Ne9      -2.51540    0.67585  -3.722 0.000317 ***
ozone$Vx9       1.29286    0.60218   2.147 0.034055 *  
ozone$maxO3v    0.35483    0.05789   6.130 1.50e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14 on 107 degrees of freedom
Multiple R-squared:  0.7622,	Adjusted R-squared:  0.7533 
F-statistic: 85.75 on 4 and 107 DF,  p-value: < 2.2e-16
```

On continue a supprimer les variables non significatif au fur et à mesure et nous obtenon finalement un résultats statisfaissant avec des varibles signification avec un alpha 5% de risuque au test de Student et de Fisher

- Modèle Obtenu

$$\hat{y}i = \mathbf{12.63131 + 2.76}xi\mathbf{1} + \mathbf{-2.51}xi\mathbf{2} + \mathbf{1.29}xi\mathbf{3} - \mathbf{128.64}xi\mathbf{4} + \mathbf{0.35}xi\mathbf{5}$$

- Les Intervalles de confiance

```
> confint(l1)
                    2.5 %    97.5 %
(Intercept)  -9.17664531 34.439265
ozone$T12     1.82344439  3.704736
ozone$Ne9    -3.85518617 -1.175618
ozone$Vx9     0.09910521  2.486609
ozone$maxO3v  0.24007499  0.469588
```

- Vérification du modèle

R*adj*^{2} = 0.7533 > 0.7 : R*adj*^{2} → 1 : le modèle est bon

3 | Troisième chapitre

3.1 Prédiction

```
## Prédiction
predict(ll, newdata=data.frame(T12=20.5,Ne9=8,Vx9=-0.5,maxO3v=114), se.fit=TRUE,
interval = "prediction", level = 0.95)
```

Ce code R utilise la fonction **predict()** pour prédire une valeur de la variable dépendante (également appelée variable cible ou variable réponse) en utilisant un modèle de régression linéaire appelé "ll". Le modèle "ll" est utilisé pour prédire la valeur de la variable dépendante en fonction de trois variables indépendantes: **T12, Ne9, Vx9, maxO3v**.

La fonction **predict()** prend en entrée un data frame contenant les valeurs des variables indépendantes pour lesquelles vous souhaitez prédire la valeur de la variable dépendante. Dans ce cas, le data frame contient une seule ligne avec les valeurs T12=20.5,Ne9=8,Vx9=-0.5,maxO3v=114 .

La fonction **predict()** calcule également l'intervalle de prédiction, qui est une plage de valeurs dans laquelle la vraie valeur de la variable dépendante se trouve avec une certaine probabilité (définie par le niveau spécifié dans l'argument **level**). Par défaut, le niveau est fixé à 95% (c'est-à-dire que la vraie valeur de la variable dépendante se trouve dans l'intervalle de prédiction avec une probabilité de 95%).

La fonction predict() retourne également la valeur prédite de la variable dépendante et l'erreur standard de la prédiction (stocker dans l'argument **se.fit**).

On obtient le resultat suivant :

$fit

A matrix: 112 × 3 of type dbl

	fit	lwr	upr
1	84.40924	56.24558	112.57289
2	76.18570	47.97915	104.39225
3	89.16429	60.31794	118.01065
4	98.48619	69.94418	127.02820
5	88.97631	60.68142	117.27120
6	78.33266	50.01374	106.65159
7	60.93163	32.67606	89.18719
8	83.55420	55.30044	111.80796
9	88.81804	60.53881	117.09726
10	98.08796	70.05337	126.12254
11	84.04329	55.95999	112.12659
12	87.87053	59.68979	116.05126
13	80.01204	52.00594	108.01815

Ce résultat est le résultat d'un appel à la fonction predict() avec l'option interval="prediction". Il comprend la valeur prédite de la variable dépendante pour chaque ligne du data frame spécifié dans l'argument newdata, ainsi que les limites inférieure et supérieure de l'intervalle de prédiction à un niveau de confiance spécifié (par défaut, 95%).

Voici une description de chaque colonne du résultat:

- fit: valeur prédite de la variable dépendante pour chaque ligne du data frame.
- lwr: limite inférieure de l'intervalle de prédiction pour chaque ligne du data frame.
- upr: limite supérieure de l'intervalle de prédiction pour chaque ligne du data frame.

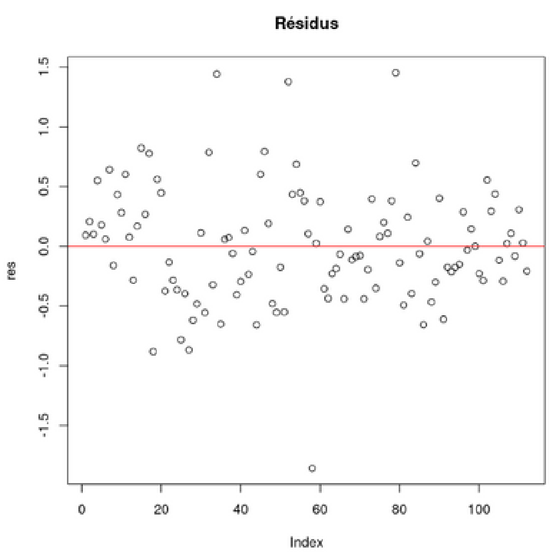
Par exemple, pour la première ligne du résultat, la valeur prédite de la variable dépendante est de 84.40924 et l'intervalle de prédiction est compris entre 56.24558 et 112.57289 (à 95% de confiance). Cela signifie que la vraie valeur de la variable dépendante pour cette ligne se trouve avec une probabilité de 95% entre ces deux valeurs.

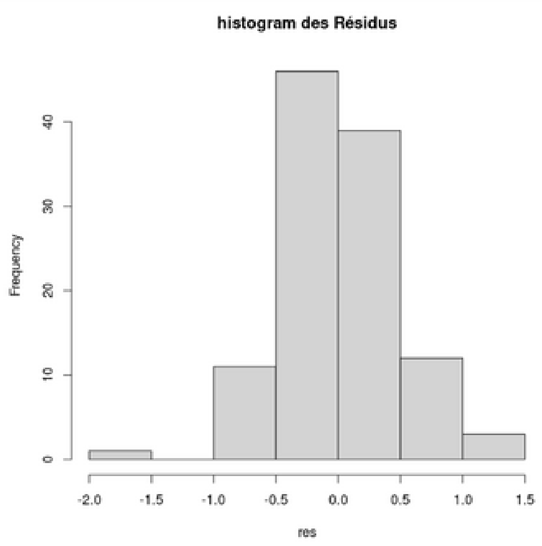
Pour chaque ligne du data frame, la fonction predict() calcule la valeur prédite de la variable dépendante en utilisant le modèle de régression linéaire spécifié et les coefficients de régression estimés pour chaque variable indépendante. Elle calcule également les limites de l'intervalle de prédiction en utilisant l'erreur standard de la prédiction.

3.2 Graphe et analyse

- ### Visualisation des résidus

```
resid(ll)
res<-resid(ll)
plot(res,main="Résidus")
abline(h=0,col="red")
hist(res,main="histogram des Résidus")
```

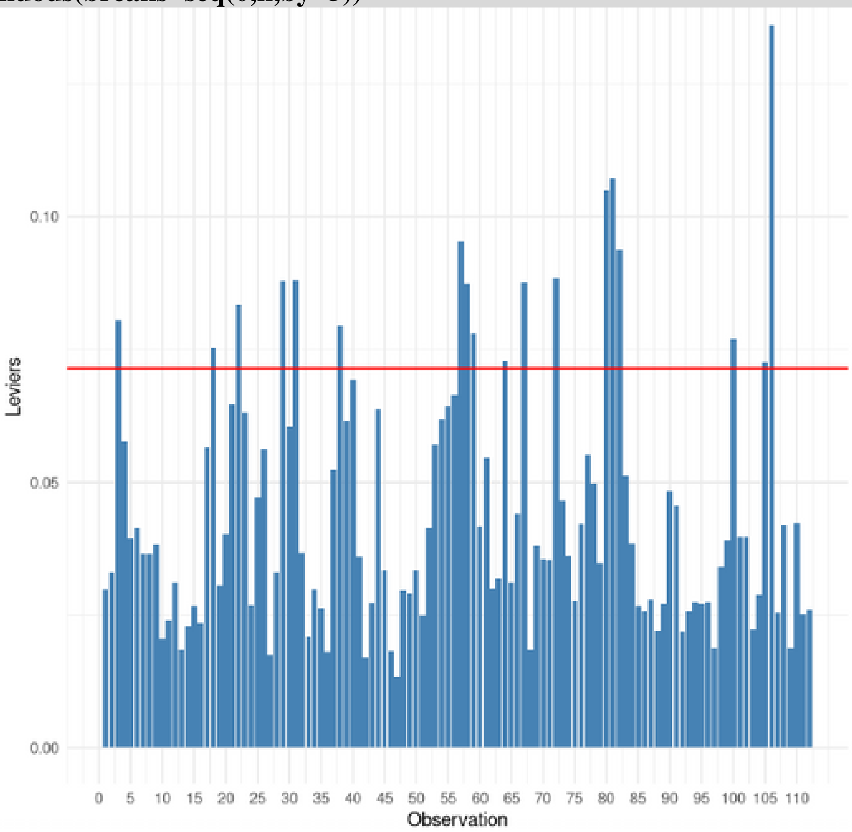




On constate bien que les résidus sont centralisés autour de leur moyenne $E(\varepsilon) = 0$.

- ## Calcul du point levier

```
library(ggplot2)
alpha <- 0.05
n <- dim(Ozone)[1]
p <- 4 # Dernier mod?le : ll
analyses <- data.frame(obs=1:n)
analyses$levier <- hat(model.matrix(ll))
seuil_levier <- 2*p/n
ggplot(data=analyses,aes(x=obs,y=levier))+
  geom_bar(stat="identity",fill="steelblue")+
  geom_hline(yintercept=seuil_levier,col="red")+
  theme_minimal()+
  xlab("Observation")+
  ylab("Leviers")+
  scale_x_continuous(breaks=seq(0,n,by=5))
```



Pour sélectionner les points pour lesquels le levier est supérieur au seuil, on exécute ces 2 lignes :

```
idl <- analyses$levier>seuil_levier
idl
analyses$levier[idl]
```

On obtient le resultat suivant :

```
FALSE · FALSE · TRUE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · TRUE
FALSE · FALSE · FALSE · TRUE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · TRUE · FALSE · TRUE · FALSE · FALSE · FALSE · FALSE · FALSE ·
FALSE · TRUE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALS
FALSE · FALSE · TRUE · TRUE · TRUE · FALSE · FALSE · FALSE · FALSE · TRUE · FALSE · FALSE · TRUE · FALSE · FALSE · FALSE · FALSE · TRUE ·
FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · TRUE · TRUE · TRUE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE ·
FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · FALSE · TRUE · FALSE · FALSE · FALSE · FALSE · TRUE · TRUE · FALSE · FALSE ·
FALSE · FALSE · FALSE · FALSE
```

```
0.0804062529974632 · 0.0752043792432835 · 0.0833985566068475 · 0.0878336719009066 · 0.0880626824531707 · 0.0794505937405452 ·
0.095382676203265 · 0.0874358992360933 · 0.0779095203190028 · 0.0728268750586198 · 0.0876475590611728 · 0.0882889428365998 ·
0.104880738440165 · 0.107330565366929 · 0.0938619946863247 · 0.0769694282457611 · 0.0725368199234223 · 0.135904590440385
```

- ## Calcul des résidus studentisés

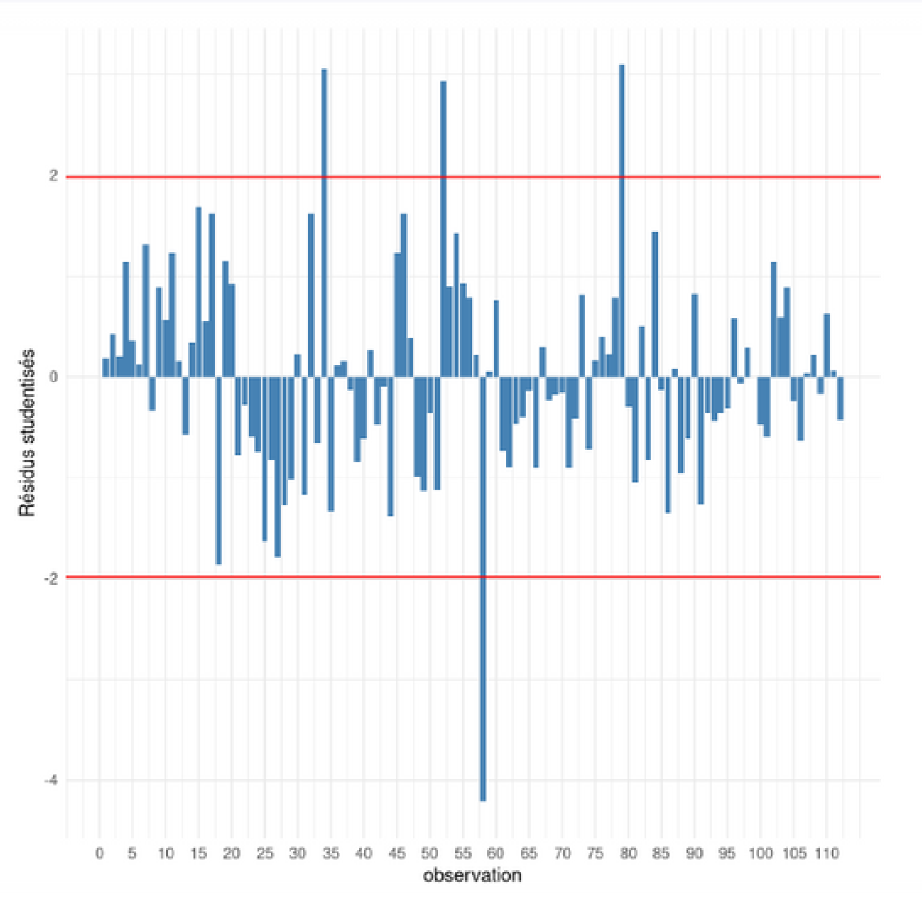
Si l'on souhaite maintenant calculer les résidus studentisés, nous écrivons ceci, sachant que le seuil pour les résidus studentisés est une loi de Student à n-p-1 degrés de liberté :

```
analyses$rstudent <- rstudent(ll)
seuil_rstudent <- qt(1-alpha/2,n-p-1)
```

Visualisons les résidus studentisés :

```
ggplot(data=analyses,aes(x=obs,y=rstudent))+
 geom_bar(stat="identity",fill="steelblue")+
 geom_hline(yintercept=-seuil_rstudent,col="red")+
 geom_hline(yintercept=seuil_rstudent,col="red")+
 theme_minimal()+
 xlab("observation")+
 ylab("Résidus studentisés")+
 scale_x_continuous(breaks=seq(0,n,by=5))
```

Voici le graphes des résidus studentisés :



• Détermination de la distance de cook

Pour trouver la distance de Cook, nous exécutons ceci :

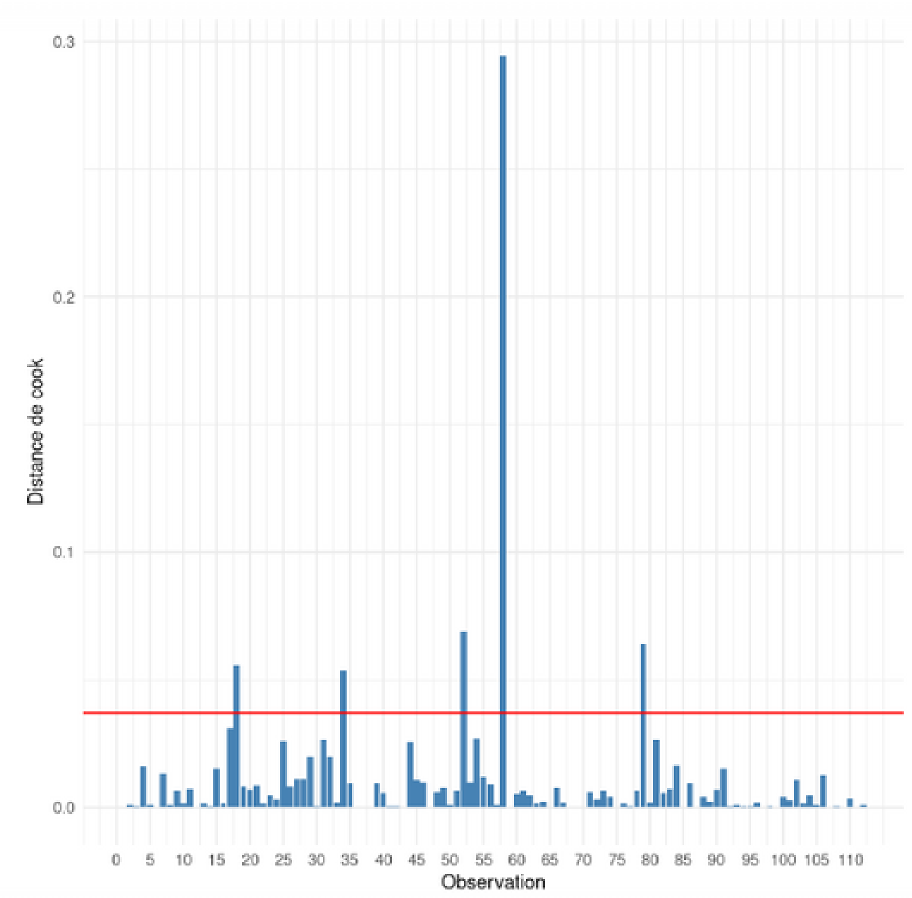
```
influence <- influence.measures(ll)
names(influence)
colnames(influence$infmat)
```

Le seuil de la distance de Cook est de 4/(n-p):

```
analyses$dcook <- influence$infmat[,"cook.d"]
seuil_dcook <- 4/(n-p)
```

On peut détecter les observations influentes comme ceci :

```
ggplot(data=analyses,aes(x=obs,y=dcook))+
 geom_bar(stat="identity",fill="steelblue")+
 geom_hline(yintercept=seuil_dcook,col="red")+
 theme_minimal()+
 xlab("Observation")+
 ylab("Distance de cook")+
 scale_x_continuous(breaks=seq(0,n,by=5))
```



- ## Vérification de la colinéarité des variables

Une autre chose à vérifier est l'éventuelle colinéarité approchée des variables :

```
install.packages("car")
library(car)
vif(ll)
```

Résultat obtenu:

Ici, tous les coefficients sont inférieurs à 10, il n'y a donc pas de problème de colinéarité.

- ## Teste d'homoscédasticité

On peut également tester l'homoscédasticité (c'est-à-dire la constance de la variance) des résidus :

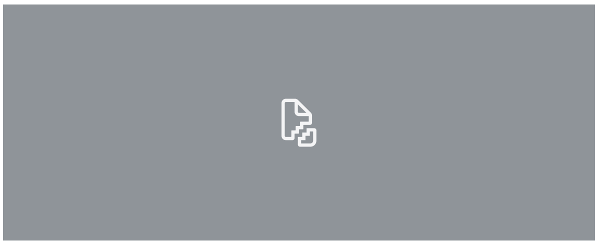
```
install.packages("lmtest")
library(lmtest)
bptest(ll)
```

Résultat obtenu:

La p-valeur ici est supérieure à 5 %, on accepte l'hypothèse H0 selon laquelle les variances sont constantes (l'hypothèse d'homoscédasticité).

- ## Teste de la normalité des résidus

```
shapiro.test(ll$residuals)
```



La p-valeur ici est inférieure à 5 %, on rejette l'hypothèse H0 selon laquelle l'échantillonest issu d'une population normalement distribuée.

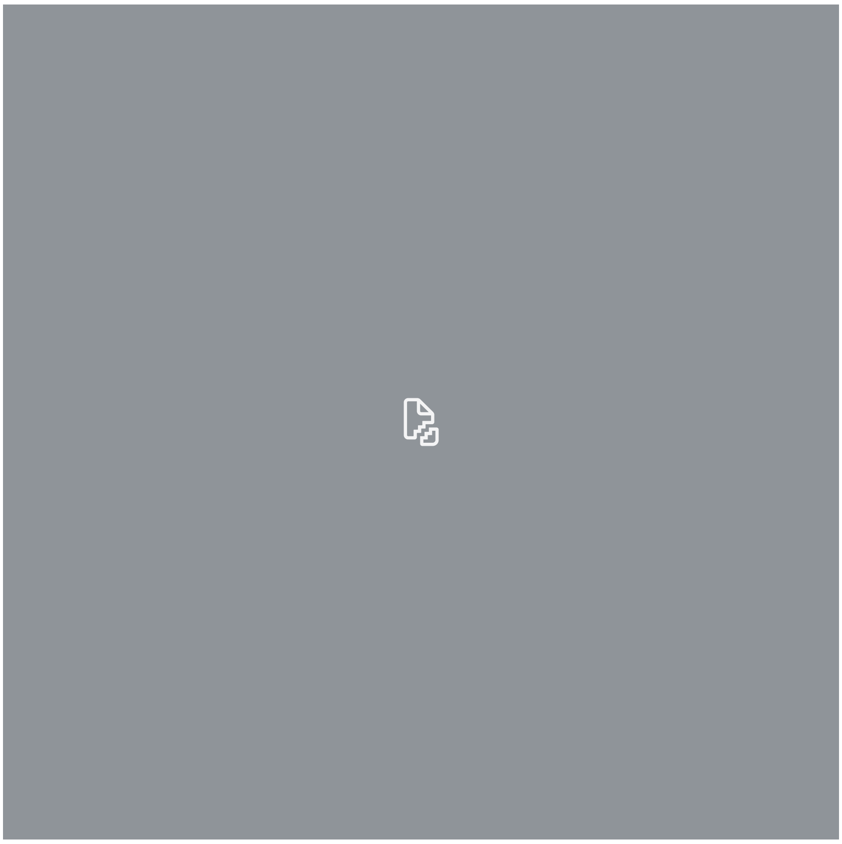
- ## Ellipse de confiance

L'ellipse de confiance est généralement utilisée pour représenter la répartition des données en deux dimensions, c'est-à-dire sur un graphique à deux axes. Elle est construite à partir de la moyenne des données et de leur écart-type. Plus l'écart-type est grand, plus l'ellipse sera étendue et moins le niveau de confiance sera élevé. En revanche, plus l'écart-type est petit, plus l'ellipse sera réduite et plus le niveau de confiance sera élevé.
En résumé, l'ellipse de confiance est un outil utile pour visualiser et quantifier la répartition des données et le niveau de confiance que l'on peut avoir dans les résultats obtenus.

Pour interpréter un graphique d'ellipse de confiance, il est important de se rappeler que l'ellipse représente la répartition des données autour de la moyenne. Plus l'ellipse est petite, plus les données sont concentrées autour de la moyenne et plus le niveau de confiance dans les résultats est élevé. En revanche, plus l'ellipse est grande, plus les données sont dispersées et plus le niveau de confiance est faible.

```
install.packages("ellipse")
library(ellipse)
i=0
j=1
resume <- summary(ll)
plot(ellipse(ll,c(i+1,j+1),level=0.95,type="l",xlab=paste("beta",i,sep=""),
ylab=paste("beta",j,sep=""), ylim=c(9,18), xlim=c(9,18))) #
points(coef(resume)[i+1],coef(resume)[j+1],pch=3)
```

Résultat obtenu:



- la moyenne des données, qui est indiquée par un point au centre de l'ellipse.
- la taille de l'ellipse :elle est grande, cela signifie que les données sont dispersées et que le niveau de confiance est faible.
- la forme de l'ellipse : elle est allongée dans une direction particulière, cela signifie que les données ont une forte tendance à suivre cette direction.
- les limites de l'ellipse : Ils indiquent à quelle distance de la moyenne se trouvent la plupart des données. Par exemple, si l'ellipse est construite avec un niveau de confiance de 95%, alors 95% des données se trouveront à l'intérieur de ses limites.

  - En résumé, pour interpréter un graphique d'ellipse de confiance, il faut tenir compte de la moyenne des données, de la taille de l'ellipse, de sa forme et de ses limites. Ces éléments permettent de comprendre la répartition des données et de quantifier le niveau de confiance dans les résultats.

3.3 Analyse de variance

3.3.1 Analyse de variance à un facteur (ANOVA)

L'analyse de variance (ou ANOVA) à 1 facteur est une méthode statistique permettant de modéliser la relation entre **une variable explicative qualitative** A et **une variable à expliquer quantitative** Y. L'objectif principal étant de comparer les moyennes empiriques de Y pour les modalités de A.

On reprend notre étude, il s'agit d'analyser la relation entre le maximum journalier de la concentration d'ozone et la direction du vent classée en secteurs (Nord, Sud,Est,Ouest). La variable vent du fichier ozone à 4 modalités.

Les diférentes étapes

Importer les données et conserver les données utiles.

```
# Extraction des données utiles
ozone3 <- Ozone[,c('maxO3','vent')]
names(ozone3)
summary(ozone3)
str(ozone3)
```

```
# Extraction des données utiles
ozone3 <- Ozone[,c('max03','vent')]
names(ozone3)
summary(ozone3)
str(ozone3)
```

```
'max03' · 'vent'

     max03             vent
 Min.   : 42.00   Length:112
 1st Qu.: 70.75   Class :character
 Median : 81.50   Mode  :character
 Mean   : 90.30
 3rd Qu.:106.00
 Max.   :166.00

'data.frame':   112 obs. of  2 variables:
 $ max03: int  87 82 92 114 94 80 79 79 101 106 ...
 $ vent : chr  "Nord" "Nord" "Est" "Nord" ...
```

- ### Analyse de la significativité du facteur

```
# Estimation des parametres
reg.aov1 <- lm(maxO3 ~vent,data=ozone3)
# Tableau d'analyse de variance
anova(reg.aov1)
```

L'analyse de variance (ANOVA) est une technique statistique utilisée pour comparer la moyenne de plusieurs groupes. Le tableau de l'ANOVA contient des informations sur la variance expliquée et non expliquée dans les données. Voici comment interpréter les différentes parties d'un tableau d'ANOVA :

- DF (Degrées de liberté) : Il s'agit du nombre de observations indépendantes dans les données.
- Sum Sq (Somme des carrés) : Cette colonne montre la variance expliquée par chaque élément du modèle.
- Mean Sq (Moyenne des carrés) : Cette colonne donne la variance moyenne expliquée par chaque élément du modèle.
- F value (Valeur F) : Cette colonne donne la valeur du test statistique F pour chaque élément du modèle.
- Pr (>F) : Cette colonne donne la p-value associée à la valeur F pour chaque élément du modèle. Si la p-value est inférieure à la seuil de significativité (généralement 0,05), alors l'élément du modèle est considéré comme statistiquement significatif.

Voici un exemple de tableau d'ANOVA :

```
# Tableau d'analyse de variance
anova(reg.aov1)
```

[42]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vent	3	7586.0566	2528.6855	3.3881	0.0207
Residuals	108	80605.622	746.3484	NA	NA

Table Chart 2 rows

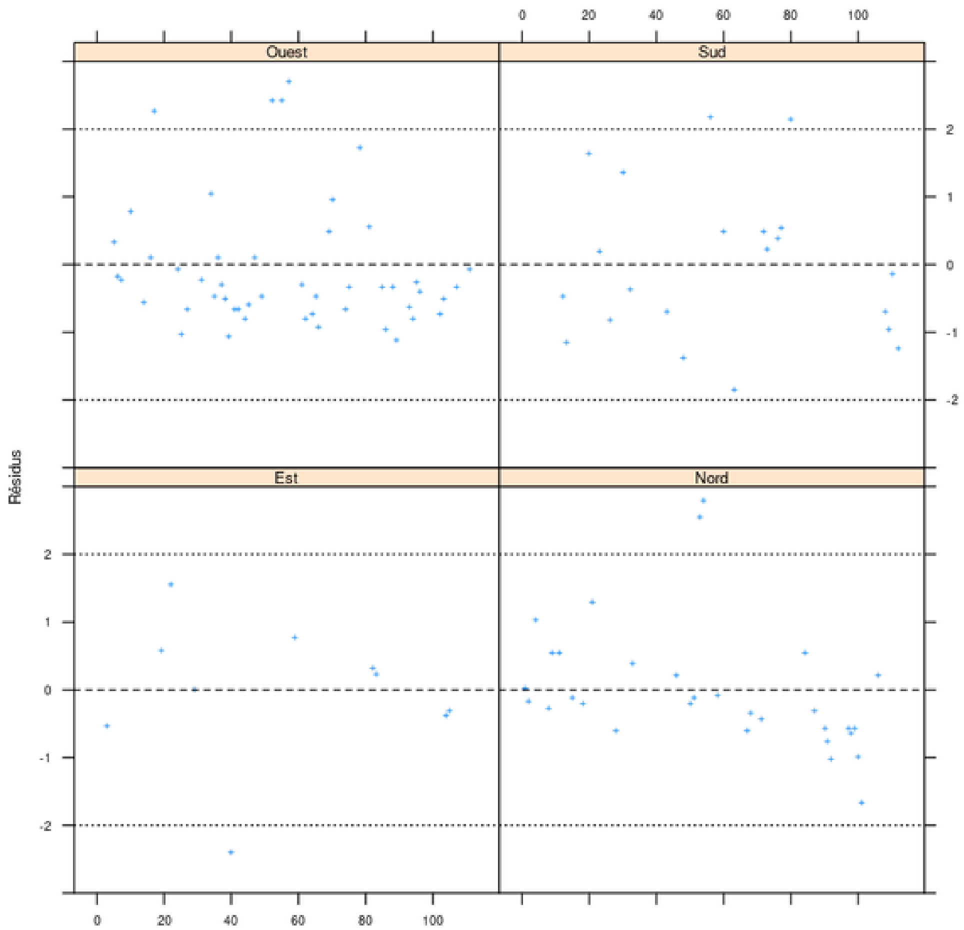
Dans cet exemple, il y a trois éléments dans le modèle (vent) et 108 observations indépendantes (residual). La variance expliquée par le modèle (Sum Sq) est de 7586.0566, ce qui donne une variance moyenne de 2528.6855 (Mean Sq). La valeur F est de 3.3881

et la p-value est de 0.0207, ce qui indique que le modèle est statistiquement significatif.

- ## Analyser les résidus

Même principe que précédemment. Utiliser le package lattice pour représenter les résidus selon les modalités de la variable vent.

```
res.aov1 <- rstudent(reg.aov1) library(lattice) monpanel <- function(...){ panel.xyplot(...)
panel.abline(h=c(-2,0,2),lty=c(3,2,3),...) } trellis.par.set(list(fontsize=list(point=5,text=8)))
xyplot(res.aov1~I(1:112)|vent,data=ozone3,pch="+",ylim=c(-3,3),
panel=monpanel,ylab="Résidus",xlab="")
```



- ## Interprétation des coefficients

Pour préciser comment la direction du vent influe sur le maximum d'ozone, on analyse les coefficients à l'aide du test de student.

Le résultat d'une régression linéaire ANOVA à un facteur (reg.aov1) inclut :

- La p-value de l'ANOVA: cette valeur mesure la probabilité que les différences observées entre les groupes soient dues au hasard. Si la p-value est inférieure à un seuil de signification (généralement 0,05), cela signifie que les différences sont statistiquement significatives et qu'il y a une relation entre la variable indépendante et la variable dépendante.

```
summary(reg.aov1)
```

```
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
```

Ici notre p-value= 0.02074 < 0.05 cela signifie que les différences sont statistiquement significatives et qu'il y a une relation entre la variable indépendante et la variable dépendante.

3.3.2 Analyse de variance avec interaction

C'est une méthode permettant de modéliser la relation entre **une variable quantitative** et **plusieurs variables qualitatives**.
On reprend notre étude, il s'agit d'analyser la relation entre le maximum journalier de la concentration d'ozone et **la direction du vent classée en secteurs** (Nord, Sud, Est, Ouest). et **la précipitation** classée en deux modalités (Sec et Pluie).
Les différentes étapes

- Importer les données et conserver les données utiles.

```
# Extraction des données utiles
ozone4 <- Ozone[,c('maxO3','vent','pluie')]
names(ozone4)
summary(ozone4)
str(ozone4)
```

```
'maxO3' · 'vent' · 'pluie'
```

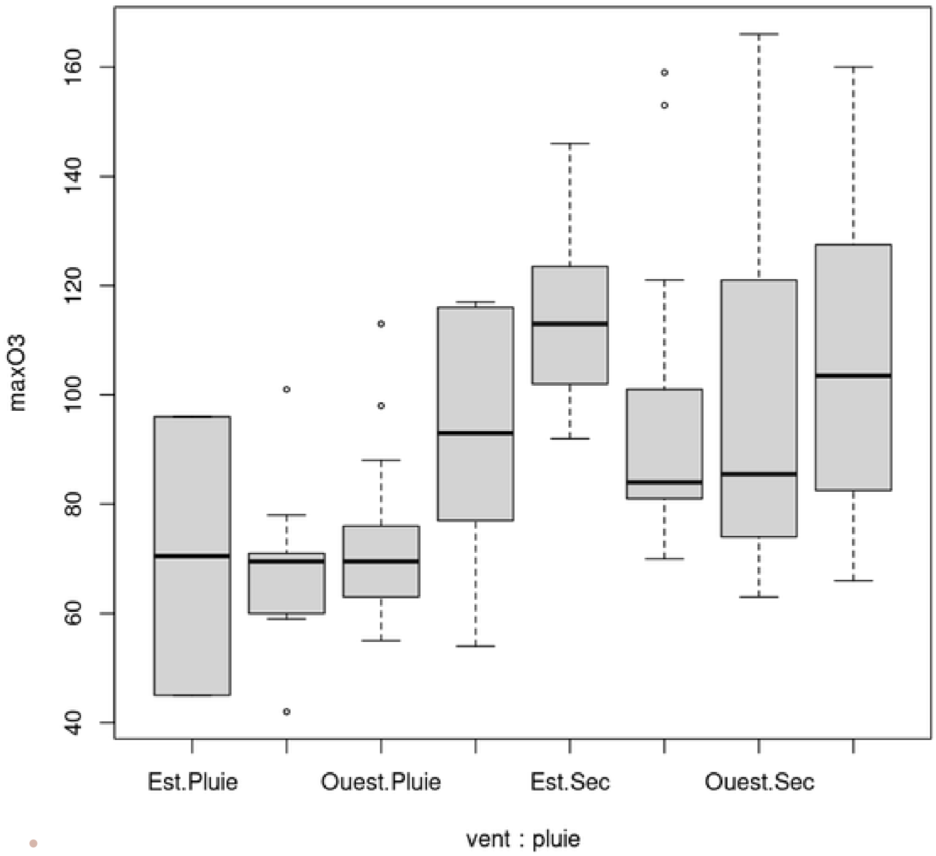
```
     maxO3            vent              pluie          
 Min.   : 42.00   Length:112         Length:112        
 1st Qu.: 70.75   Class :character   Class :character  
 Median : 81.50   Mode  :character   Mode  :character  
 Mean   : 90.30                                        
 3rd Qu.:106.00                                        
 Max.   :166.00                                        
```

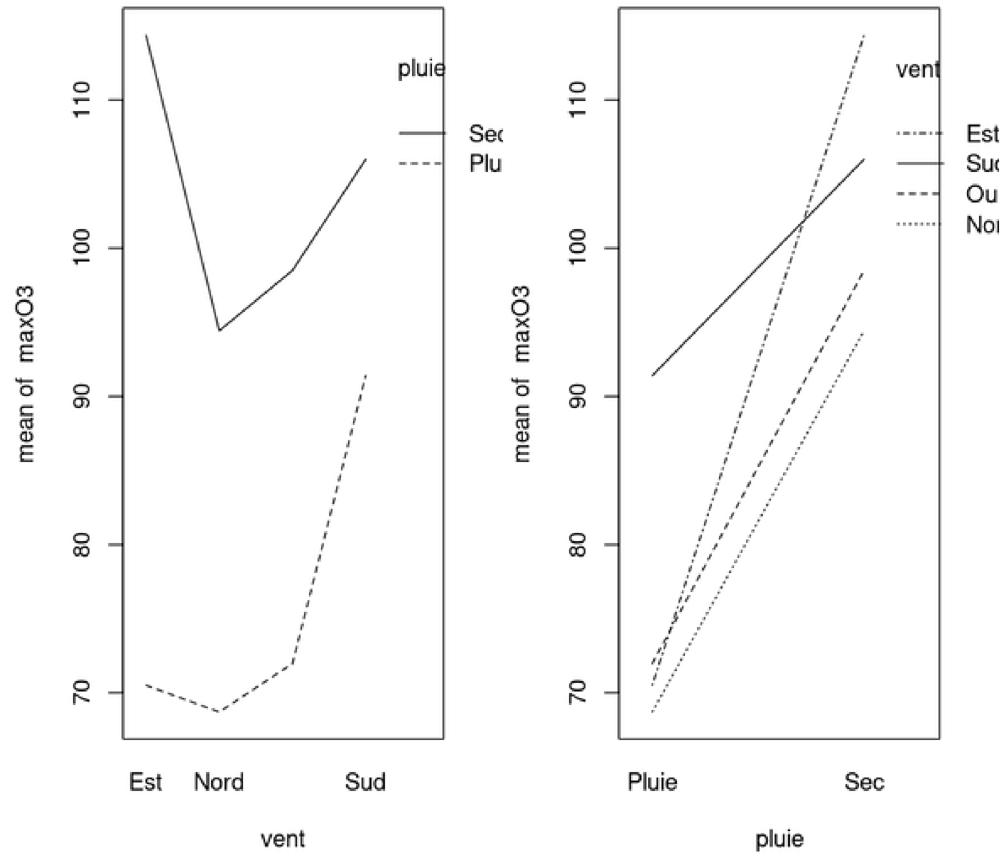
```
'data.frame':	112 obs. of  3 variables:
 $ maxO3: int  87 82 92 114 94 80 79 79 101 106 ...
 $ vent : chr  "Nord" "Nord" "Est" "Nord" ...
 $ pluie: chr  "Sec" "Sec" "Sec" "Sec" ...
```

• Importer les données et conserver les données utiles.

On représente une boîte à moustaches de la variable à expliquer par croisement des modalités des variables explicatives **vent** et **pluie** (4*2).
L'influence conjointe entre les variables **vent** et **pluie** a-t-elle un effet sur la dispersion du maximum de la concentration en ozone.

```
# Representation des donnees
boxplot(maxO3~vent*pluie,data=ozone4,cex=0.5)
# Interaction
par(mfrow=c(1,2))
with(Ozone,interaction.plot(vent,pluie,maxO3))
with(Ozone,interaction.plot(pluie,vent,maxO3))
```





- Choisir le modèle - Estimation des paramètres

```
# Choisir le modèle
mod.int <- lm(maxO3~vent*pluie,data=ozone4)
anova(mod.int)
```

```
# Choisir le modèle
mod.int <- lm(maxO3~vent*pluie,data=ozone4)
anova(mod.int)
```

[51]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vent	3	7586.0566	2528.6855	4.1454	0.0081
pluie	1	16159.4262	16159.4262	26.491	0.000001257
vent:pluie	3	1006.4164	335.4721	0.55	0.6493
Residuals	104	63439.7794	609.9979	NA	NA

Table Chart 4 rows

La p-values des variables vent et pluie sont inférieur à 0.05 ;
Par contre la p-value de l'interception des variables vent et pluie est de 0.6493 **> 0.05**, ce qui indique que le modèle n'est pas statistiquement significatif.

 # Choisir le modèle _
mod.ssint <- lm(maxO3~vent+pluie,data=ozone4) anova(mod.ssint)

```
# Choisir le modèle
mod.ssint <- lm(maxO3~vent+pluie,data=ozone4)
anova(mod.ssint)
```

[52]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
vent	3	7586.0566	2528.6855	4.1984	0.0075
pluie	1	16159.4262	16159.4262	26.8295	0.0000010521
Residuals	107	64446.1957	602.3009	NA	NA

Table Chart 3 rows

les p-value sont inférieur à 0.05 pour toutes les variables du modèle ce qui signifie que ce modèle est statistiquement significatif donc le modèle qu'on choisit.

- ## Interprétation des coefficients

Comme pour les méthodes précédentes, on utilisera la fonction summary pour aider l'interprétation.

summary(mod.ssint)

```
Residuals:
    Min      1Q  Median      3Q     Max 
-42.618 -15.664  -3.712   8.295  67.990 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   85.123      8.710   9.773  < 2e-16 ***
ventNord     -16.333      8.946  -1.826   0.0707 .  
ventOuest    -12.709      8.647  -1.470   0.1446    
ventSud       -2.101      9.431  -0.223   0.8241    
pluieSec      25.597      4.942   5.180 1.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.54 on 107 degrees of freedom
Multiple R-squared:  0.2692,	Adjusted R-squared:  0.2419 
F-statistic: 9.856 on 4 and 107 DF,  p-value: 7.931e-07
```

Ici notre p-value= 7.931e-07 < 0.05 cela signifie que les différences sont statistiquement significatives et qu'il y a une relation entre la variable indépendante et la variable dépendante.

• Bibliographie

- Support de cours

- DATASET:
  jeu de données d'ozone

- https://en.m.wikipedia.org/wiki/Confidence_region

- https://bookdown.org/teodor_tiplica/book_linearrrgression/ValidModel.html

- ANOVA et Analyse
  Support de cours
  https://fr.wikipedia.org/wiki/Analyse_de_la_variance