Département de Mathématiques
M1 - MATHÉMATIQUES APPLIQUÉES

MODELES LINEAIRES

Djamila AZZOUZ
Julien PIERROT

Professeur référent
Mme Sylviane WEY
Maître de Conférence à l'Université Le Havre Normandie

Année académique 2021 - 2022

TABLE DES MATIÈRES

TABLE DES FIGURES

ABSTRACT

When studying a phenomenon with one or more explanatory variables $X_1, ...X_p$ and an output variable Y, the purpose of modeling is to express the variable Y as a function of the variables $X_1, ...X_p$ using a mathematical relation. In order to establish this kind of relation, linear regression models are introduced[6] . These models allow to express Y as a function of $X_1, ...X_p$ and are present in various fields such as biological sciences, statistics, environmental sciences and even in the business world [2] . However, it turns out that the explanatory variables can either be quantitative or qualitative, so depending on the type of variables that is wanted for the model, one or another linear regression model will be chosen. Once the model is selected, the next step is to check if it is possible to improve it or if the variables that have been introduced in the model have an influence on the Y variable. Hypothesis testing and the method of variable selection are used in order to answer this question. Finally, to have a better understanding of the models, a few examples of real life applications simulated on the software R [7] will be showcased.

INTRODUTION

Un modèle linéaire est un modèle statistique dont l'objectif est d'exprimer une variable aléatoire Y en fonction de variables explicatives sous forme d'un opérateur linéaire. L'expression de "régression linéaire" a été utilisée pour la première fois dans un article en 1886, par le statisticien britannique Francis Galton, dans lequel il constata un phénomène de "régression vers la moyenne" de la taille des fils en fonction de celle des pères. Les modèles linéaires sont présents dans différents domaines tels que les sciences biologiques, les statistiques, les sciences environnementales et même dans le monde de l'entreprise. Ils ont l'avantage de modéliser beaucoup de phénomènes de manière réaliste afin de faire des prévisions, d'avoir des paramètres faciles à estimer et d'avoir beaucoup d'outils statistiques et informatiques qui leur sont associés.
Dans ce projet de master de mathématiques appliquées, nous verrons, dans un premier temps, toutes les notions de la théorie de probabilités et statistiques que nous avons déjà vu au cours de notre parcours universitaire, et qui nous seront utiles par la suite.
Dans un second temps, nous présenterons toutes les extensions des modèles linéaires à variables quantitatives (linéaire simple, multiple, pôlynomiale) que nous illustrerons avec des exemples d'applications en nous appuyant sur les résultats obtenus avec le logiciel R.
Enfin, nous terminerons avec un modèle linaire qualitatif appelé "Analyse de la variance" (ou ANOVA) que nous illustrerons également avec un exemple d'application avec l'utilisation du logiciel R.

REMERCIEMENTS

La réalisation de ce projet a été possible grâce à la contribution de plusieurs personnes à qui nous voudrions témoigner toute notre gratitude. Nous voudrions, tout d'abord, adresser toute notre reconnaissance à la directrice de ce mémoire, Madame WEY Sylviane, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter nos réflexions et notre curiosité.
On adresse nos sincères remerciements à tous les professeurs, intervenants et toutes les personnes qui par leurs paroles, leurs écrits, leurs conseils et leurs critiques, ont guidé nos réflexions et ont accepté de nous rencontrer et de répondre à nos questions durant nos recherches, plus particulièrement Mr DIARRASSOUBA Ibrahima.
On remercie nos très chers parents, qui ont toujours été là pour chacun de nous. On remercie également nos frères et sœurs, pour leurs encouragements.
Enfin, on remercie nos amis Thomas, Louise, et Ulysse qui ont toujours été là pour nous. Leur soutien inconditionnel et leurs encouragements ont été d'une grande aide.

CHAPITRE

1

RAPPELS ET COMPLÉMENTS

Tout d'abord, nous commençons par un petit rappel sur les lois de probabilités et certaines notions de la statistique inférentielle[1], que nous avons vues au cours de notre formation et que nous utiliserons par la suite pour réaliser notre projet.

I Rappels sur les lois de probabilités

I.1 Variable aléatoire réelle

Définition 1.1. On appelle variable aléatoire toute application mesurable X d'un espace de probabilité (Ω, A, P) dans $\mathbb{R}$ muni de la tribu borélienne. Il s'agit donc tout simplement d'une application mesurable sur un espace de probabilité.

Définition 1.2. Un vecteur aléatoire $X = (X_1, X_2, ...X_n)$ à valeurs dans $\mathbb{R}^n$ est un vecteur gaussien si toute combinaison linéaire de ses composantes suit une loi Normale $\mathcal{N}(0, 1)$.

I.2 Variance et covariance

Définition 1.3. Soit X une variable aléatoire réelle avec $X \in L^2$.
La variance de X est le nombre $V(X) = \|X - E(X)\mathbb{1}\|_{L^2} = E((X - E(X))^2)$

Définition 1.4. Soient X et Y deux variables aléatoires réelles et dans L^2.
La covariance de X et Y est le nombre $Cov(X, Y) =< X - E(X), Y - E(Y) >_{L^2}= E[(X - E(X)(Y - E(Y))]$.

Définition 1.5. Si $X = (X_1, X_2, ..., X_n)$ est un vecteur aléatoire dans $\mathbb{R}^n$ tel que $E(\|X\|^2) = E(X_1^2 + ... + X_n^2) < +\infty$ alors on appelle matrice de covariance de X la matrice carrée symétrique notée Σ définie par :

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & ... & Var(X_1, X_d) \\ Cov(X_2, X_1) & Var(X_2) & ... & Cov(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_d, X_1) & ... & ... & Var(X_n) \end{pmatrix}$$

où $Cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))], \forall (i, j) \in \{1, ..., n\}$

I.3 Moyenne, Espérance

Définition 1.6. Soit $X_1, X_2, ..., X_n$ un échantillon de taille $n \geq 1$ associé à X. On définit la moyenne empirique de cet échantillon, notée $\overline{X_n}$ par :

$$\boxed{\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i}$$

Définition 1.7. Soit X une variable aléatoire réelle. Si X est ou bien ≥ 0 ou bien dans L^1 alors on pose

$$\boxed{E(X) = \int_\Omega X dP}$$

On dit que $E(X)$ est l'espérance de X, ou encore la moyenne de X

Proposition 1.1. *Si $X_1, ..., X_n$ sont des variables aléatoires réelles, alors*

$$E(X_1 + ... + X_n) = E(X_1) + ... + E(X_n).$$

Proposition 1.2. *Si $X_1, ..., X_n$ sont des variables aléatoires indépendantes ou bien ≥ 0, ou bien dans L^1, alors on peut écrire :*

$$\boxed{E(X_1 \times ... \times X_n) = E(X_1) \times ... \times E(X_n)}$$

Définition 1.8. Si $X = (X_1, X_2, ..., X_n)$ est un vecteur aléatoire dans $\mathbb{R}^n$ et si $(X_1, X_2, ..., X_n)$ sont intégrables ou positives alors on appelle espérance de X le vecteur de $\mathbb{R}^n$ égal à

$$E(X) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_d) \end{pmatrix}$$

I.4 Loi normale gaussienne

Définition 1.9. On dit qu'une variable aléatoire X suit une loi Normale ou gaussienne d'espérance m et de variance σ^2, s'il existe $\phi_{(m,\sigma)} : \mathbb{R}^* \longrightarrow \mathbb{R}$ de densité :

$$\boxed{\phi_(m, \sigma)(x) = \frac{1}{\sigma\sqrt{2\pi}} \times \exp(-\frac{(x-m)^2}{2\sigma^2})}$$

Proposition 1.3. *Une variable aléatoire X de loi Normale $\mathcal{N}(m, \sigma^2)$ a pour :*
- *Espérance : $E[X] = m$*
- *Variance : $Var(X) = \sigma^2$*

Proposition 1.4. *X suit une loi Normale $\mathcal{N}(m, \sigma^2) \iff \frac{X-m}{\sigma}$ suit une loi Normale $\mathcal{N}(0, 1)$.*

II Rappels sur la statistique inférentielle

II.1 Lois dérivées de la loi normale

II.1.1 La loi du Chi-deux

Soit $(E, \mathcal{B}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ et soient $X_1, X_2, ..., X_n$ avec $n \geq 1$, des variables aléatoires indépendantes et identiquement distribuées de loi Normale $\mathcal{N}(0, 1)$.

Définition 2.1. La variable aléatoire

$$\boxed{U_n = \sum_{i=1}^{n} X_i^2}$$

suit une loi du $\chi^2(n)$ à n degrés de liberté ayant pour densité :

$$\boxed{f(x) = \frac{x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} \exp\left(-\frac{x}{2}\right)\mathbb{1}_{]0,+\infty[}(x) \text{ où } x \in \mathbb{R}}$$

Remarque 2.1.
- $\Gamma(\alpha) = \int_0^{+\infty} \exp(-x).x^{\alpha-1}dx$ où $\alpha \in \mathbb{R}$ est appelé coefficient de Gamma.
- On a $\Gamma(1/2) = \sqrt{\pi}$ et $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ $\forall\alpha > 0$
- $\chi^2(n) = \gamma(\frac{n}{2}, \frac{1}{2})$ où $\gamma(\alpha, \lambda)$ avec $\alpha > 0$ et $\gamma > 0$ est la loi gamma de densité $f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda.x)\mathbb{1}_{]0,+\infty[}(x)$ où $x \in \mathbb{R}$

Proposition 2.1. *Une variable aléatoire X de loi $\chi^2(n)$ à n degrés de liberté a pour :*
- *Espérance : $E(X) = n$*
- *Variance : $Var(X) = 2n$*

Proposition 2.2. *Si $X_1, X_2, ..., X_n$ sont des variables aléatoires indépendantes et de même loi normale $\mathcal{N}(0, 1)$ alors $X_1^2 + X_2^2 + ... + X_n^2$ suit une loi du $\chi^2(n)$ à n degrés de liberté.*

Nous allons présenter la simulation de la densité de la loi de Chi-deux réalisée avec le langage R.

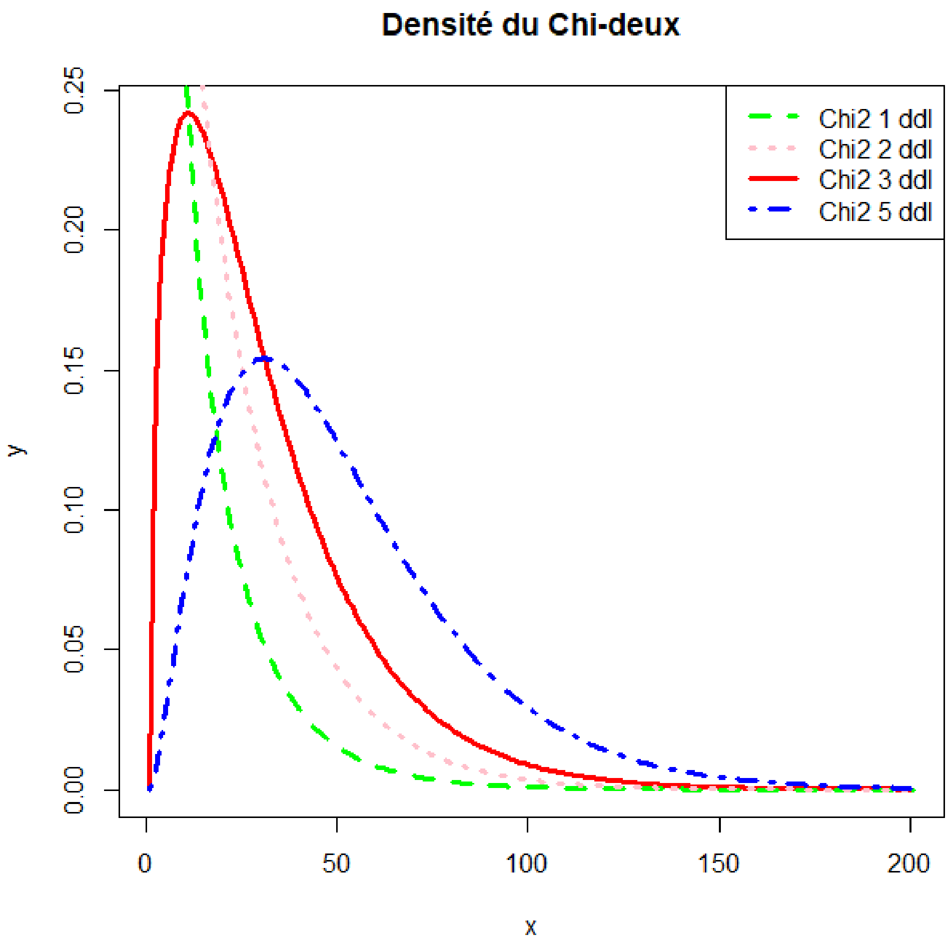

FIGURE 1.1 – Densité de la loi de Chi-deux

II.1.2 Loi de Student

Définition 2.2. La variable aléatoire

$$\boxed{T_n = \frac{X}{\sqrt{\frac{U_n}{n}}}}$$

où X et U_n sont indépendantes, suit une loi de Student à n degrés de liberté ayant pour densité :

$$\boxed{f(x) = \frac{1}{\sqrt{n}B\left(\frac{n}{2},\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad \text{où } x \in \mathbb{R}}$$

Définition 2.3. Une variable aléatoire X de loi de Student à n degrés de liberté a pour :

- Espérance : $E[T_n] = 0$,
- Variance : $Var(T_n) = \frac{n}{n-2}$ où $n > 2$.

Le graphique ci-dessous représente la courbe représentative de la densité de la loi de Student réalisée avec R.

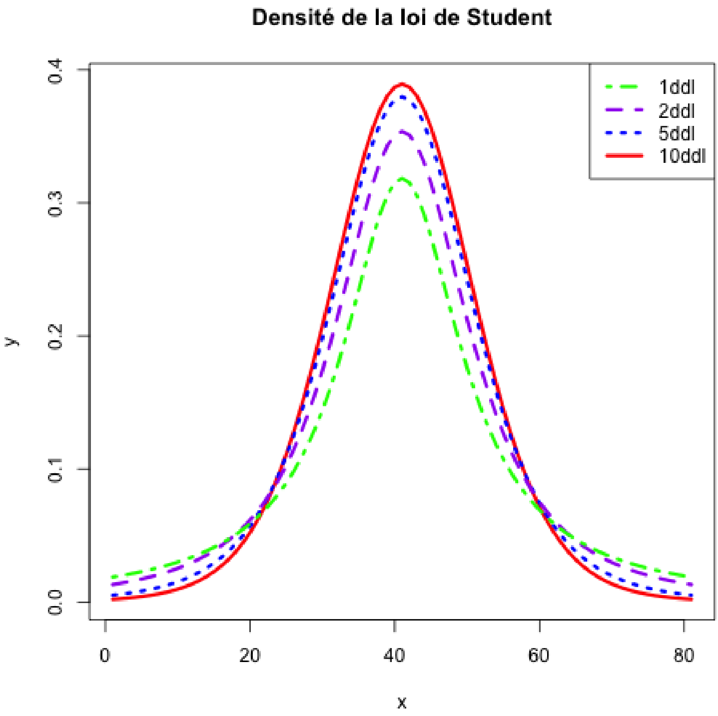

FIGURE 1.2 – Densité de la loi de Student pour différents degrés de liberté

II.1.3 Loi de Fisher-Snédecor

Définition 2.4. Soient $X_m \sim \chi^2(m)$ et $X_n \sim \chi^2(n)$ où X_m et X_n sont deux variables aléatoires indépendantes. Alors la variable aléatoire :

$$\boxed{F_{m,n} = \frac{\frac{X_m}{m}}{\frac{X_n}{n}}}$$

suit une loi de Fisher-Snedecor de degrés de liberté m et n de densité :

$$\boxed{F_{m,n} = \frac{(\frac{m}{n})^{\frac{m}{2}} x^{\frac{m}{2}-1}}{B(\frac{m}{2}, \frac{n}{2})(1 + \frac{m}{n}x)} \mathbb{1}_{]0,+\infty[}(x) \text{ où } x \in \mathbb{R}}$$

Définition 2.5. Une variable aléatoire X de loi de Fisher-Snedecor à m et n degrés de liberté a pour :

— Espérance : $E[F_{m,n}] = \frac{n}{n-2}$ où $n > 2$,

— Variance : $Var(F_{m,n}) = \frac{2n^2(m+n-2)}{m(n-4)(n-2)^2}$ où $n > 4$.

Ci-dessous la courbe de la densité de la loi de Fisher-Snedecor

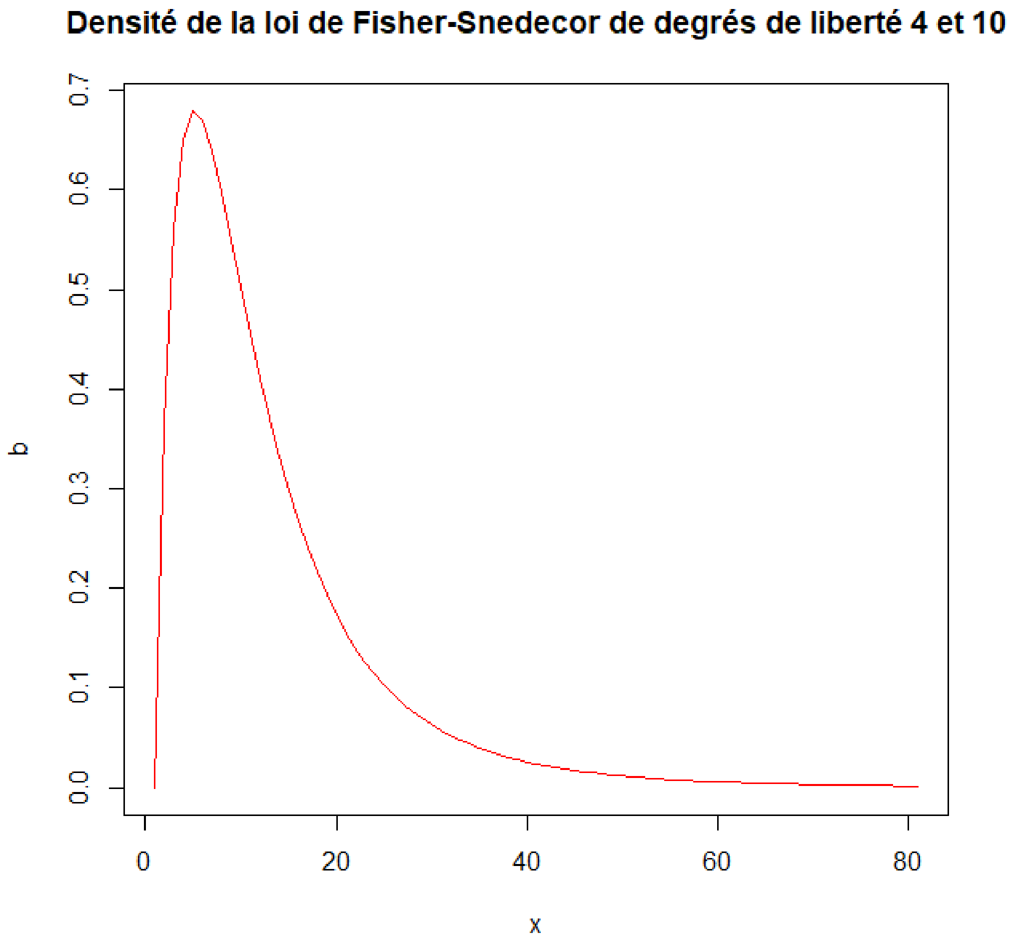

FIGURE 1.3 – Densité de la loi de Fisher-Snedecor de degrés de liberté 4 et 10

II.2 Estimateur

Définition 2.6. Un estimateur du paramètre inconnu θ d'un modèle ou loi de probabilité est une fonction qui fait correspondre à une suite d'observations issues du modèle ou loi de probabilité la valeur $\hat{\theta}$, que l'on nomme estimé ou estimation.

$$\hat{\theta}_n = f(x_1, x_2, ..., x_n)$$

Définition 2.7. Soit $(\Omega, A, \mathcal{P})$ un modèle statistique. Une statistique T est une variable aléatoire sur $(\Omega, \mathcal{A})$ à valeur dans $(E, \mathcal{E})$ dont l'expression ne dépend pas de la famille $\mathcal{P}$.

Définition 2.8. Soient $(\Omega, A, (P_\theta, \theta \in \Theta))$, alors une statistique T est un estimateur sans biais de $g(\theta)$ si :

$$\forall \theta \in \Theta, E_\theta(T) = g(\theta).$$

II.3 La vraisemblance

Soit $(\Omega, A, P)^n = (\Omega^n, A^{\otimes n}, P^{\otimes n})$, le modèle d'échantillon empirique. La vraisemblance est définie par :

$$\boxed{L(\omega_1, ..., \omega_n, \theta) = \prod_{j=1}^{n} f_\theta(\omega_i)}$$

Et sa log-vraisemblance est :

$$\boxed{\log L(\omega_1, ...\omega_n, \theta) = \log(\prod_{j=1}^{n} f_\theta(\omega_i))}$$

II.4 Le maximum de vraisemblance (EMV)

Soit (Ω, A, P) où $P = (P_\theta, \theta \in \Theta)$, un modèle statistique dominé par la mesure μ. On appelle estimateur du maximum de vraisemblance (EMV), pour le paramètre θ et on note $\hat{\theta_n} = \hat{\theta_n}(\omega_1, ..., \omega_n)$, la valeur de θ, si elle existe et si elle est unique, qui rend maximum :

$$\log L(\omega_1, ...\omega_n, \theta) = \log(\prod_{j=1}^{n} f_\theta(\omega_i))$$

où $\hat{\theta_n}$ ne dépend pas de la mesure dominante.
Nous procèdons ainsi pour calculer l'EMV :

- Pour $j = 1....k$, on calcule la quantité : $\dfrac{\partial \log L}{\partial \theta_j}$;
- On résoud le sytème d'inconnues $\theta_1, ..., \theta_k$:
$$\frac{\partial \log L}{\partial \theta_j} = 0 \quad \forall j \in \{1, ..., k\}$$
- On vérifie bien que le θ_j trouvé $\forall j \in \{1, ...k\}$, est bien un maximum. Pour cela il faut calculer la dérivée seconde et vérifier que :
$$\frac{\partial^2 \log L}{\partial^2 \theta_j}|_{\hat{\theta_j}=\theta_j} \leq 0$$

II.5 Variable quantitative, qualitative et exemples

Une variable statistique permet de décrire une caractéristique pour un ensemble d'individus appelé population. Cette caractéristique peut être, par exemple, la couleur des yeux, la taille de chaque individu ou encore le nombre de leurs enfants. En fonction du type de la caractéristique étudiée, il existe deux types de variables statistiques :

- **les variables quantitatives**,
- **les variables qualitatives**.

Comme leurs noms l'indiquent, les variables quantitatives permettent de décrire des quantités comme le nombre d'enfants de chaque individu de la population étudiée ou encore leur âge. En voici un exemple :

Individus	1	2	3	4
Poids	62	80	75	60

Dans ce tableau, on peut voir que la variable quantitative étudiée est le poids sur une population de 4 individus.

Les variables qualitatives, quant à elles, permettent de décrire des qualités comme par exemple, la couleur des yeux des individus ou encore le diplôme possédé. En voici un exemple :

Individus	1	2	3	4	5	6	7
Couleur des yeux	Vert	Marron	Vert	Marron	Marron	Bleue	Bleue

Dans ce tableau, on peut voir que la variable qualitative étudiée est la couleur des yeux sur une population de 7 individus.

CHAPITRE

2

MODÈLE LINÉAIRE QUANTITATIF

Le modèle linéaire de base, qu'on utilise pour analyser une expérience où l'on étudie sur n unités expérimentales les variations d'une variable Y en fonction de facteurs qualitatifs ou quantitatifs (appelé aussi explicatifs), peut s'écrire :

$$Y_i = m_i + \varepsilon_i$$

où :
$i \in [1, n]$: représente le numéro de l'unité expérimentale.
$\forall i \in [1, n] \quad m_i$: est l'espérance de Y_i incluant l'effet de variables explicatives .
$\forall i \in [1, n] \quad \varepsilon_i$: est une variable aléatoire résiduelle, appelée "erreur".

Selon la nature des variables explicatives incluses dans la partie explicative m_i, on distingue deux catégories :

— **cas où les variables explicatives sont quantitatives :** Comme déjà expliqué dans la partie "Rappels" sur la définition d'une variable quantitative, elles nous permettent de définir des modèles appelés "modèle de régression" : simple s'il n'y a qu'une variable explicative, multiple sinon .

— **cas où les variables explicatives sont qualitatives :** appelées facteurs et le modèle ainsi construit est un modèle d'analyse de la variance (**Anova**).

I Régression linéaire simple

La réalisation de cette partie est basée sur [3]. Le principe de cette régression en général, est de chercher à expliquer une variable Y, dite variable à expliquer (exogène), en fonction de

la variable explicative X (ou encore endogène).
On dispose donc d'observations de ces variables sur un échantillon de données de n individus, présentés dans le tableau suivant :

	X	Y
I_1	X_1	Y_1
$\vdots$	$\vdots$	$\vdots$
I_n	X_n	Y_n

où X_i, Y_i sont des caractéristiques du i-ème individu $\forall i \in \{1, ..., n\}$ et $I_1 ... I_n$ sont les individus et Y est la variable à expliquer en fonction de X.

— Le but de cette régression consiste donc à ajuster un modèle pour expliquer Y en fonction de X, mais aussi de prédire les valeurs de Y pour de nouvelles valeurs de X. Pour analyser cette relation entre chaque couple (X_i, Y_i), il faut donc chercher une fonction $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ telle que :

$$\boxed{Y_i \simeq f(X_i)} \tag{2.1}$$

— Pour définir $\simeq$, il faut donner un critère évaluant la qualité de l'ajustement de la fonction f aux données et une classe de fonctions C dans laquelle on trouvera f.
Ici on prendra $\quad C = \{f : f(X) = \beta_0 X + \beta_1, (\beta_0, \beta_1) \in \mathbb{R}^2\}$
On peut donc écrire notre problème mathématique sous la forme :

$$\min_{f \in C} \sum_{i=1}^{n} l(Y_i - f(X_i))$$

où n : taille de l'échantillon de données,
$l(\cdot)$: représente une fonction de coût qui peut-être vue comme la distance entre une observation (X_i, Y_i) et son point correspondant dans la droite $(X_i, f(X_i))$. Nous utiliserons le coût quadratique qui s'écrit : $l(f) = f^2$ car il nous permet d'avoir tous les points proches de la droite.
On obtient donc la formule suivante :

$$\boxed{\min_{f \in C} (Y_i - \beta_1 X_i - \beta_2)^2} \tag{2.2}$$

I.1 Modélisation Statistique :

Comme nous l'avons déjà mentionné précédemment, nous allons ajuster les données par une droite. Nous supposons que les données sont sous la forme :

$$\boxed{Y = \beta_0 + \beta_1 X} \tag{2.3}$$

Donc Y_i dépend linéairement de X_i, $\forall i \in \{1, ..., n\}$. Cependant cette liaison sera perturbée par une "erreur" due aux erreurs de mesure. Par conséquent, nous posons le modèle sous la forme :

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i} \tag{2.4}$$

β_j, $j = 0, 1$ sont les paramètres inconnus à estimer,
$\forall i \in [1, n] \quad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$, sont les résidus.
Nous supposons que $E(\varepsilon) = 0$.
Afin d'estimer les paramètres inconnus de ce modèle β_0, β_1, nous allons utiliser la méthode des moindres carrés pour obtenir les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ en minimisant la quantité :

$$\boxed{S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2} \tag{2.5}$$

où $\quad S = \sum_{i=1}^{n} \varepsilon_i^2 \quad$ donc $\quad (\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{min} S(\beta_0, \beta_1)$
L'équation 2.5 est convexe si elle admet un point singulier c'est-à-dire un minimum qui annule ses dérivées partielles.
Ainsi, en minimisant l'erreur de l'équation précédente, nous obtenons donc le système suivant :

$$\begin{cases} \dfrac{\partial S}{\partial \beta_0} = 0 \\ \dfrac{\partial S}{\partial \beta_1} = 0 \end{cases} \tag{2.6}$$

2.6 nous donne

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{2.7}$$

et

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{2.8}$$

2.7 nous donne :

$$\begin{aligned} -\sum_{i=1}^{n} Y_i + n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0 &\Leftrightarrow n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ &\Leftrightarrow n\hat{\beta}_0 = \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} Y_i \\ &\Leftrightarrow \hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{1}{n} \hat{\beta}_1 \sum_{i=1}^{n} X_i \\ &\Leftrightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

Où

$$\begin{cases} \bar{Y} &= \dfrac{1}{n} \sum_{i=1}^{n} Y_i \\ \bar{X} &= \dfrac{1}{n} \sum_{i=1}^{n} X_i \end{cases}$$

sont les moyennes empiriques de X et Y.
L'équation 2.8 nous donne
$-\sum_{i=1}^{n} X_i Y_i + \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0$

En remplaçant $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$ par son expression nous obtenons :

$$(\bar{Y} - \hat{\beta}_1\bar{X})\sum_{i=1}^{n} X_i + \hat{\beta}_1\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_iY_i \Leftrightarrow \bar{Y}\sum_{i=1}^{n} X_i - \hat{\beta}_1\bar{X}\sum_{i=1}^{n} X_i + \hat{\beta}_1\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_iY_i$$
$$\Leftrightarrow \hat{\beta}_1(\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i\bar{Y}$$
$$\Leftrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i\bar{Y}}{\sum_{i=1}^{n} X_i^2 - \bar{X}\sum_{i=1}^{n} X_i}$$

En multipliant par $\sum_{i=1}^{n}(X_i - \bar{X})$, on a :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i(Y_i - \bar{Y})\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n} X_i(X_i - \bar{X})\sum_{i=1}^{n}(X_i - \bar{X})} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Comme nous avons vu à la partie 11 :

$$\begin{cases} Cov(X,Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) \\ Var(X) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{cases}$$

On obtient comme solution de ce système d'équations :

$$\boxed{\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}}$$

Proposition 1.1. *Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais de β_0 resp β_1 qui sont de variance minimale grâce au théorème de Gauss-Markov.*
On a : $E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1$

Enfin nous pouvons donc estimer la droite de régression par la formule suivante : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1\hat{X}$

I.2 Qualité de la régression

Une fois que la droite de régression est estimée, nous allons mesurer la corrélation entre nos variables. Pour cela, nous allons vérifier la qualité de cette régression en utilisant le coefficient de corrélation théorique :

$$R = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Ce coefficient nous permet de comparer la distance de chaque point des données depuis la moyenne de la variable et l'utiliser pour indiquer dans quelle mesure la relation entre les variables suit une ligne imaginaire.

I.2.1 Interprétation du coefficient de corrélation

- Si le coefficient de corrélation R est plus proche de -1 ou 1, alors la corrélation est très bonne. On dit que la relation est linéaire, reportée dans un nuage de points et tous les points des données peuvent être reliés par une ligne droite (ou approchés par cette droite), donc la régression choisie est la meilleure.
- Si les valeurs des deux variables ont tendance à augmenter ou diminuer ensemble, alors le coefficient de corrélation R est positif et dans ce cas, on parle de corrélation positive.
- Si les valeurs de la première augmentent et celles de la deuxième variable diminuent, alors le coefficient de corrélation R est négatif et dans ce cas, on parle de corrélation négative.
- Si $R = 1$ alors la liaison entre nos variables est dite parfaite.De plus comme $R > 0$ la corrélation est positive.
- Si $R = -1$ on a de même une liaison parfaite mais d'une pente négative.

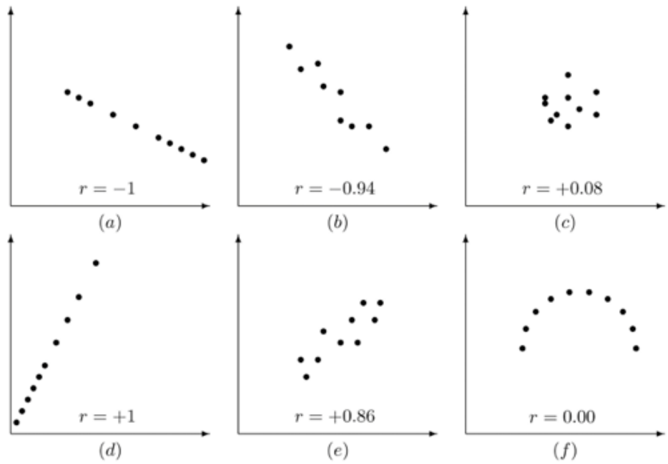

FIGURE 2.1 – Exemples de coefficient de corrélation

I.2.2 Exemple d'applications

Afin de mieux comprendre l'importance de cette régression et comment l'appliquer, nous allons illustrer les résultats obtenus avec l'exemple suivant :
L'association de surveillance de la qualité de l'air **Air Breizh**, mesure la concentration de polluants comme l'ozone ($O3$) ainsi que les conditions météorologiques comme la température, la nébulosité, le vent etc..
Leur objectif est de prévoir la concentration en ozone pour le lendemain afin d'avertir la population en cas de pic de pollution.
Nous souhaitons analyser ici la relation entre le maximum journalier en ozone en (μ/m^3) et les données météorologiques.
Nous disposons de 112 données relevées durant l'été 2001 à Rennes.
On trouve dans le fichier ozone des variables telles que :

- maxO3, qui est la valeur maximale d'ozone observée sur une journée .
- T9, T12 et T15 qui sont les températures prises respectivement à 9 h, 12 h et 15 h ;
- Ne9, Ne12, Ne15 qui sont des nébulosités prises à 9 h, 12 h et 15 h ;
- Vx9, Vx12 et Vx15 qui sont les composantes est-ouest du vent mesurées à 9 h, 12 h et 15 h ;

— maxO3v, qui donne la teneur maximale en ozone observée la veille.
Soient $Y = maxO3$ et $X = T12$. Afin de savoir si la régression linéaire est pertinente, nous devons tracer le nuage de points pour reconnaître le type de la régression.

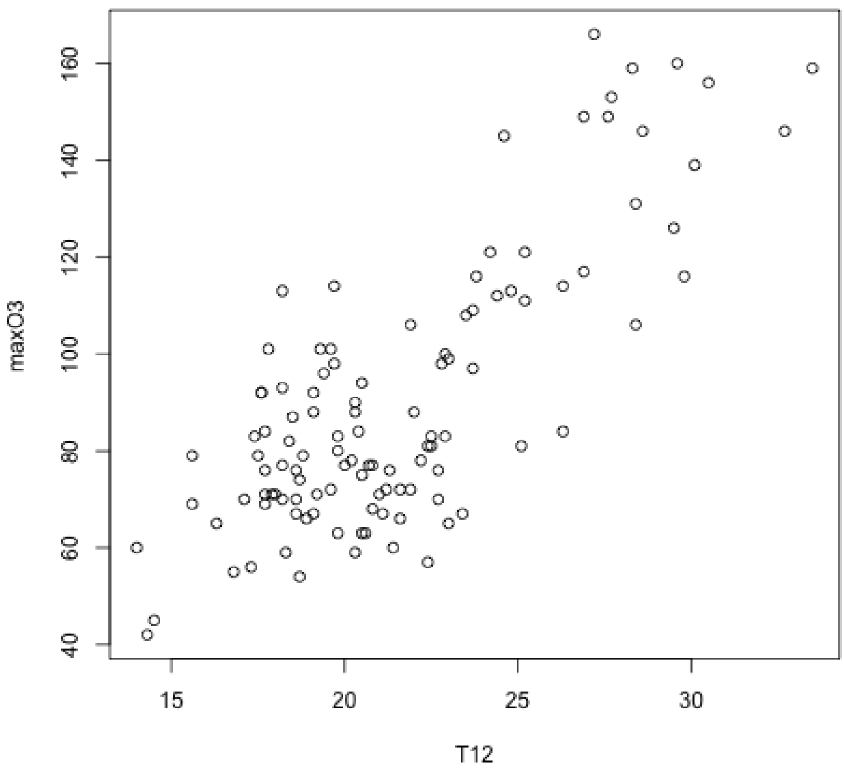

FIGURE 2.2 – Évolution de la concentration en ozone maxO3 en fonction de la température à midi

Comme nous le voyons, les variables X et Y semblent être corrélées. En effet, la tendance ressemble bien à une droite donc une régression linéaire simple paraît pertinente.
Posons

$$Y = \beta_1 X + \beta_0$$

Nous devons estimer les coefficients (β_0, β_1) de la régression à l'aide des formules obtenues précédemment.

Rappelons que : $\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = -27.42 \\ \hat{\beta}_1 = \dfrac{Cov(X,Y)}{Var(X)} = 5.47 \end{cases}$ et $\begin{cases} \bar{Y} = \dfrac{1}{m} \displaystyle\sum_{i=1}^{10} Y_i = 90.30 \\ \bar{X} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{10} X_i = 21.52 \\ Var(X) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{10} X_i^2 - \bar{X}^2 = 16.34 \\ Var(Y) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{10} Y_i^2 - \bar{Y}^2 = 794.51 \end{cases}$

Pour estimer la qualité de cette régression, nous allons calculer le coefficient de corrélation théorique :

$$R = \frac{COV(X,Y)}{\sqrt{VarX \times VarY}} = \mathbf{0.61}$$

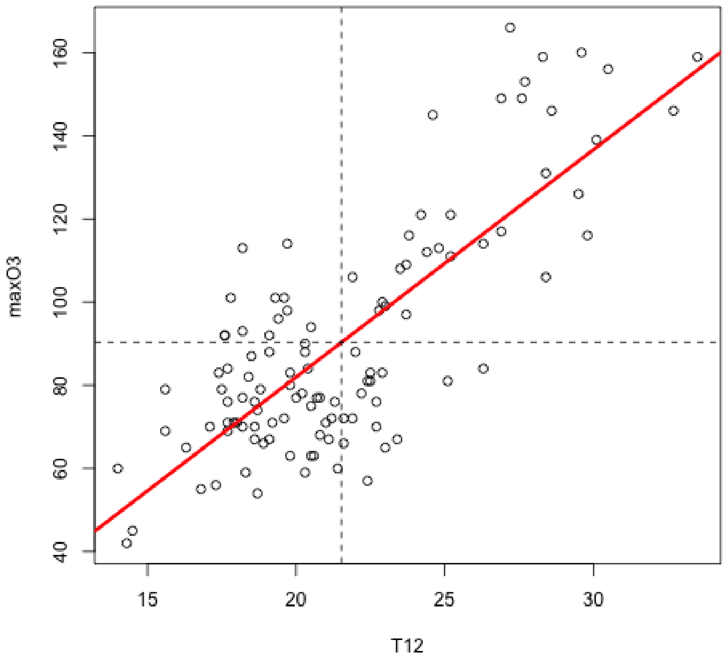

FIGURE 2.3 – Droite de régression linéaire

Nous remarquons que R est positif, donc les variables X et Y semblent être corrélées positivement. De plus, le coefficient de corrélation étant égal à 0.61, ce modèle n'est donc pas le meilleur.
Comme nous l'avons déjà dit, le but de la régression est aussi de prédire les valeurs de Y pour de nouvelles valeurs de X. Nous allons donc chercher à trouver la concentration en ozone maxO3, si la température à midi est de 25 degrés. Donc $X = 25$ et :

$$\hat{Y} = \hat{\beta}_1 \hat{X} + \beta_0 \Leftrightarrow \hat{Y} = \hat{\beta}_1 \times 25 + \beta_0 \Leftrightarrow \hat{Y} = 5.47 \times 25 - 27.42 = 109.33$$

Donc la concentration en ozone sera de $109,33$ pour une température à midi de 25 degrés. On peut vérifier que les résidus suivent une loi normale avec un histogramme classique (cela devrait approximativement dessiner une courbe de Gauss).

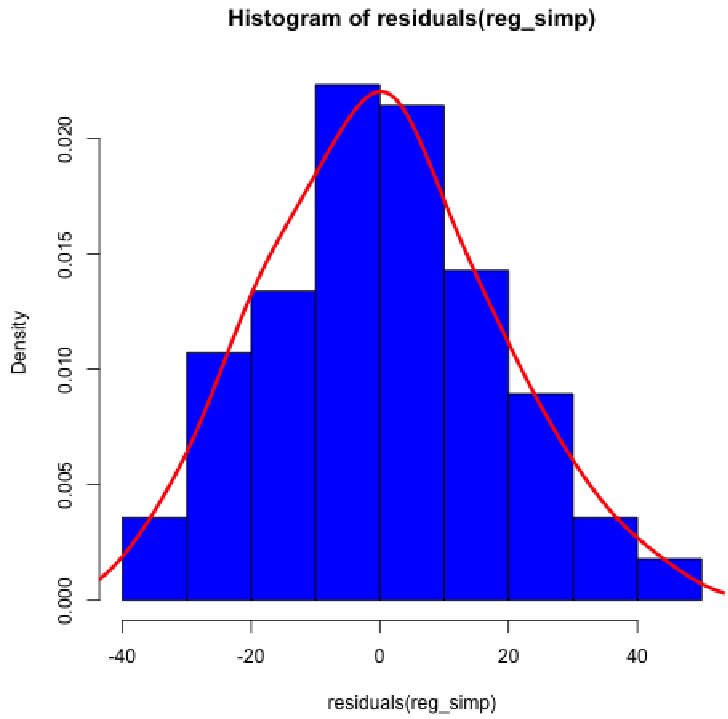

FIGURE 2.4 – Histogramme des résidus

II Régression linéaire multiple

Cette méthode est une généralisation du modèle de régression linéaire simple, où on cherche toujours à exprimer une variable quantitative Y (exogène) en fonction de p variables quantitatives (endogène) $X_1, ..., X_p, p \geq 2$ [4] et [5].
Considérons un échantillon de données de taille n, présenté dans le tableau suivant :

	X	X_1 . . . . X_p
1	Y_1	$X_{1,1} X_{1,2}$ . . . . $X_{1,p}$
2	Y_2	$X_{2,1} X_{2,2}$ . . . . $X_{2,p}$
.	.	. . .
.	.	. . .
.	.	. . .
n	Y_n	$X_{n,1} X_{n,2}$ . . . . $X_{n,p}$

On cherche donc à déterminer une relation linéaire entre Y et les variables X_i, $i = 1, ..., p$ où $Y = (Y_1, ..., Y_n)$.
Nous allons chercher une fonction f telle que : $Y = f(X_i)$ pour $i = 1, ..., p$.
Comme nous l'avons fait dans la régression simple, il nous faut un critère positif évaluant la

qualité de l'ajustement de la fonction f aux données et une classe C où on trouvera f.
On prendra $C = \{f : f(X_1....X_p)\} = \sum_{j=1}^{p} \beta_j X_j + \beta_0,\ \beta_j \in \mathbb{R}\ \forall j \in [1, n]\}$.
On obtient donc notre problème mathématique suivant :

$$\min_{f \in C} \sum_{i=1}^{n} l(Y_i - f(X_{i,1}.......X_{i,p}))$$

n : taille de l'échantillon.
$l(.)$: nous utiliserons toujours le coût quadratique
d'où le problème s'écrit :

$$\min_{f \in C} \sum_{i=1}^{n} (Y_i - (\sum_{j=1}^{p} \beta_j X_i + \beta_0))^2$$

II.1 Modélisation statistique

Comme nous l'avons vu précédemment dans le cas de la régression simple, nous cherchons à déterminer nos coefficients β_i, $i = 0, ..., p$ tels que :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

ou encore

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p} + \varepsilon_i \qquad \forall i \in \{1....n\}$$

où $\forall i \in \{1, ..., n\}$ ε_i : sont des variables aléatoires appelées "résidus" ;
β_0 et β_j ; $j = 1....p$ sont les paramètres à estimer ;
Sous les hypothèses suivantes :

- On suppose que les termes d'erreurs ε_i $\forall i = 1, ..., n$ sont des termes aléatoires indépendants identiquement distribués (iid) tels que : $\begin{cases} E(\varepsilon_i) = 0 & \forall i = 1....n \\ Var(\varepsilon_i) = \sigma^2 & \forall i = 1....n \end{cases}$
  c'est-à-dire : $\varepsilon_i \rightsquigarrow N(0, \sigma^2)$ (ε_i suit une loi normale)
- On suppose aussi que les coefficients $\beta_0, ..., \beta_p$ sont des constantes.

- Sous ces hypothèses du modèle linéaire, on obtient que $(Y_1, ..., Y_n)$ est un échantillon de variables aléatoires indépendantes vérifiant :
- $E[Y_i] = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_p X_{i,p}$.
- $Var[Y_i] = \sigma^2$

On va estimer nos paramètres inconnus $(\beta_0, ..., \beta_p)$ avec la méthode des moindres carrés, en minimisant toujours la quantité :

$$S(\beta_0, ..., \beta_p) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i,1} - \beta_2 X_{i,2} - ...... - \beta_p X_{i,p})^2$$

où $S = \sum_{i=1}^{n} \varepsilon_i^2$
On pose donc :

$$(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p) = \min_{(\beta_0, ..., \beta_p) \in \mathbb{R}^{p+1}} S(\beta_0, ..., \beta_p)$$

Cette fonction est différentiable sur $\mathbb{R}^n$, on peut donc utiliser les méthodes d'optimisation continue pour optimiser le résultat.

Posons : $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$ et $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ et X une matrice de taille $(n \times p + 1)$

$$X = \begin{pmatrix} 1 & X_1^1 & ......X_1^p \\ 1 & X_2^1 & ......X_2^p \\ \vdots & \vdots & \\ 1 & X_n^1 & ......X_n^p \end{pmatrix}$$

On obtient donc $Y = X\beta + \varepsilon$ où $\varepsilon = Y - XB$

Avec ces notions on doit donc déterminer $\hat{\beta}$ estimateur de β tel que :

$$S(\hat{\beta}) = \min_{\beta \in \mathbb{R}^{p+1}} \|\varepsilon\|_2^2 = \min_{\beta \in \mathbb{R}^{p+1}} \|Y - X \times B\|_2^2$$

Nous allons procéder maintenant à la détermination de $\hat{\beta}$.

On a :

$$\begin{aligned} S(\beta) &= \|Y - X\beta\|^2 \\ &= (Y - X\beta)^T \times (Y - X\beta) \\ &= (Y^T - (X\beta)^T)(Y - X\beta) \\ &= (Y^TY) - Y^T(X\beta) - (X\beta)^TY + (X\beta)^T.(X\beta) \end{aligned}$$

On peut remarquer que : $Y^T(X\beta) = ((X\beta)^T.Y)^T = (X\beta)^T.Y$

Donc

$$S(\beta) = Y^T.Y - 2(X\beta)^T.Y + (\beta^TX^T)(X\beta)$$

La condition nécessaire d'optimum est que la dérivée première par rapport à β s'annule, d'où nous obtenons :

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^T.Y + 2X^T(X\beta)$$

Alors il existe un optimum $\hat{\beta}$ qui vérifie :

$$\begin{aligned} -2X^TY + 2X^TX\hat{\beta} = 0 &\Leftrightarrow 2X^TX\hat{\beta} = 2X^TY \\ &\Leftrightarrow \hat{\beta} = (X^TX)^{-1}X^TY \end{aligned}$$

Remarque : X^TX est inversible dès que $n > p + 1$ c'est-à-dire si les colonnes de X sont linéairement indépendantes.

Afin de s'assurer que $\hat{\beta}$ est bien un minimum, il faut que la dérivée seconde soit une matrice définie positive.

Or $\dfrac{\partial^2 S(\beta)}{\partial^2 \beta} = 2X^TX$ et X est de rang n donc X^TX est inversible et n'a pas de valeurs propres nulles.

La matrice X^TX est donc définie.

De plus $\forall Z \in \mathbb{R}^p$,

on a :

$$\begin{aligned} Z^T2X^TXZ = < XZ, 2XZ > &= 2 < XZ, XZ > \\ &= 2\|XZ\|^2 \geq 0 \end{aligned}$$

d'où (X^TX) est bien définie positive, et $\hat{\beta}$ est bien un minimum.
Posons $\hat{Y} = X\hat{\beta}$ où $\hat{Y}$ est le vecteur de valeur produit par le modèle linéaire
On a :

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ &= X(X^TX)^{-1}(X^TY) \\ &= (X(X^TX)^{-1}X^T)Y\end{aligned}$$

Posons $H = X(X^TX)^{-1}X^T$ une matrice de projection orthogonale dans l'espace vectoriel formé par les vecteurs $(X_1, X_2, ..., X_p)$. On peut donc définir $\varepsilon = \hat{Y} - Y$ l'erreur entre $\hat{Y}$ et Y . Une fois que nous trouvons nos estimateurs c'est-à-dire $(\hat{\beta}_0.....\hat{\beta}_p)$, nous devons s'assurer qu'ils admettent de bonnes propriétés au sens statistique.
On peut donc poser deux questions :

- Nos estimateurs sont ils sans biais ?
- Sont-ils de variance minimale dans sa classe d'estimateurs ?

Notons que lorsque toutes les variables sont deux à deux orthogonales, X^TX est une matrice diagonale.
On peut donc calculer l'espérance et la variance de $\hat{\beta}$. On a :

$$\begin{aligned}E(\hat{\beta}) &= E((X^TX)^{-1}X^TY) \\ &= (X^TX)^{-1}X^TE(Y) \qquad or \quad E(Y) = X\beta \\ &= (X^TX)^{-1}X^TX\beta = \beta \qquad car \quad (X^TX)^{-1}X^TX = I_d\end{aligned}$$

$$E(\hat{\beta}) = \beta \text{ donc } \hat{\beta} \text{ est un estimateur sans biais de } \beta$$

Calculons sa variance :

$$\begin{aligned}Var(\hat{\beta}) &= Var((X^TX)^{-1}X^TY) \\ &= ((X^TX)^{-1}X^T)Var(Y)X(X^TX)^{-1} \qquad or \quad Var(Y) = \sigma^2 \\ &= ((X^TX)^{-1}X^T)\sigma^2X(X^TX)^{-1} \\ &= [((X^TX)^{-1}X^T)X(X^TX)^{-1}]\sigma^2 \\ &= (X^TX)^{-1}\sigma^2\end{aligned}$$

D'où

$$\boxed{Var(\hat{\beta}) = (X^TX)^{-1}\sigma^2}$$

Ainsi nous obtenons grâce au théorème de Gauss-Markov que l'estimateur $\hat{\beta}$ est optimal parmi les estimateurs linéaires sans biais de β.

II.2 Qualité de la régression linéaire multiple

On évalue la qualité d'une régression, en mesurant "l'angle" formé par les vecteurs Y et $\hat{Y}$. On introduit les trois valeurs suivantes :

- $SCT = \|Y - \bar{Y}\mathbb{1}\|^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})$ qui n'est autre que la variance totale de Y.
- $SCE = \|Y - \hat{Y}\mathbb{1}\|^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)$ représente la somme des carrés des erreurs.

— $SCR = \|\hat{Y} - \bar{Y}\mathbb{1}\|^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})$ qui est la somme des carrés expliquée par la régression (ou résidu).

Le théorème de Pythagore nous donne :

$$SCT = SCR + SCE$$

On obtient donc la qualité de la régression en calculant :

$$R^2 = \frac{SCE}{SCT}$$

Ou encore on peut l'écrire en fonction des résidus :

$$R^2 = 1 - \frac{SCR}{SCT}$$

Le coefficient $R^2 \in [0, 1]$, appelé coefficient de détermination, représente la proportion de variation totale expliquée par le modèle.

— Plus R^2 est proche de 1, meilleur est l'ajustement ($\hat{Y} \simeq Y$).

Après la construction d'un modèle de régression linéaire multiple, nous nous posons les questions suivantes :

— **Pouvons nous améliorer notre modèle de régression ?**

— **Nos variables explicatives apportent-elles toutes de l'information a notre modèle ?**

Pour cela nous allons introduire les tests d'hypothèses pour vérifier l'importance de nos variables explicatives par rapport à la variable Y.

III Test d'hypothèses sur les coefficients du modèle

Le test d'hypothèses est une démarche qui a pour but de fournir une règle de décision permettant, sur la base des résultats d'échantillon, de faire un choix entre deux hypothèses statistiques. Dans notre cas, il est important d'évaluer la signification de notre modèle.
Pour cela nous allons effectuer deux types de tests : le test de validité global et le test individuel sur les coefficients de notre modèle $(\beta_1, ..., \beta_p)$.

III.1 Cas de la régression multiple

III.1.1 Test de validité global :

L'objectif de ce test est de vérifier si notre modèle est intéressant ou non ?

— Si le modèle n'est pas intéressant, on le traduit par l'hypothèse $H_0 : \beta_j = 0 \;\; \forall j \in \{1, ..., p\}$, c'est-à-dire que le coefficient de détermination vaut zéro ($R^2 = 0$) et cela signifie que notre modèle s'écrit seulement en fonction de β_0 et ε_i ($Y_i = \beta_0 + \varepsilon_i$). Ceci montre que notre modèle n'utilise pas les variables explicatives $(X_1, ..., X_p)$ pour d'écrire Y.

— Si le modèle est intéressant, on le traduit par l'hypothèse $H_1 : \beta_j \neq 0.\ \ \exists j \in \{1, ..., p\}$, c'est-à-dire que $R^2 \neq 0$, et qu'il existe au moins une variable explicative X_j de coefficient β_j qui influe sur Y.

On introduit donc le test suivant :

$$\begin{cases} H_0 : \beta_j = 0 & \forall j \in \{1, ..., p\} \\ H_1 : \beta_j \neq 0 & \exists j \in \{1, ..., p\} \end{cases}$$

Une fois que le test est posé, nous devons chercher une statistique de test.
Or on sait que :
$E\left(\dfrac{SCR}{p}\right) = \sigma^2$ qui est l'espérance du carré moyen du modèle.
Et $E\left(\frac{SCE}{n-p-1}\right) = \sigma^2$ qui est l'espérance du carré moyen de la résiduelle.
Sous H_0, ces deux quantités sont comparables en moyenne. Dans notre test, cela consiste donc à faire une comparaison de variances.
La statistique de test est donc :

$$F_{test} = \frac{SCR/p}{SCE/(n-p-1)} = \frac{\sum\limits_{i=1}^{n}(\hat{Y}_i - \bar{Y})/p}{\sum\limits_{i=1}^{n}(Y_i - \hat{Y}_i)/(n-p-1)}$$

La loi de cette statistique, sous H_0, suit une loi de Fisher à p et $(n-p-1)$ degrés de liberté.

$$F_{test} \rightsquigarrow F(p, n-p-1)$$

Décision :
Si $F_{test} > f_{1-\alpha}(p, n-p-1)$, alors on rejette l'hypothèse H_0 au seuil $\alpha \in [0, 1]$, où $f_{1-\alpha}(p, n-p-1)$ est le quantile d'ordre $1-\alpha$ de la loi $F(p, n-p-1)$.

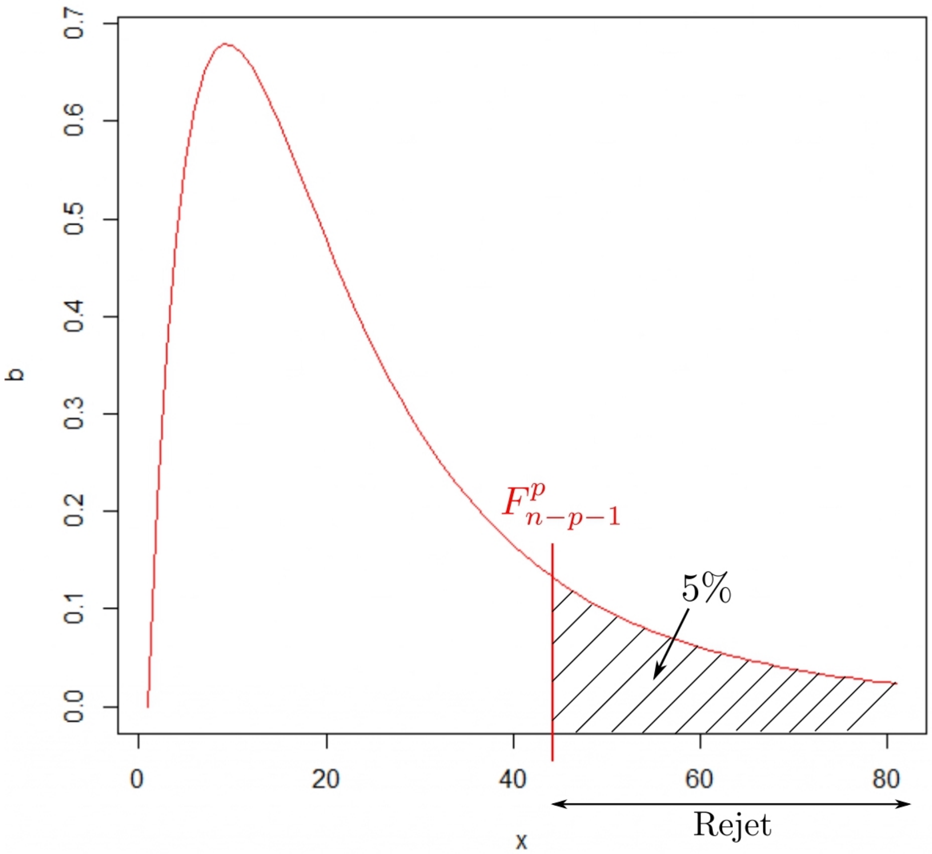

FIGURE 2.5 – Densité de Fisher avec région de rejet.

III.1.2 Test individuel

Il consiste à construire un test, coefficient par coefficient, pour savoir s'il est nul ou pas. On sait que les $\hat{\beta}_j \ \ \forall j \in \{1, ..., p\}$ suivent une loi normale de moyenne β_j et de variance $\sigma^2_{\hat{\beta}_j}$. On centre et on divise par l'écart type.
On obtient :

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \rightsquigarrow N(0,1) \tag{2.9}$$

Remarque 3.1. $\sigma_{\hat{\beta}_j}$ est la vraie valeur de l'écart type de $\hat{\beta}_j$, mais on ne la connaît pas car on a juste les expériences qui nous permettent d'estimer cette valeur. Nous allons remplacer $\sigma_{\hat{\beta}_j}$ par son estimateur $\hat{\sigma}_{\hat{\beta}_j}$.

2.9 Devient :

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \rightsquigarrow T_{(n-p-1)}$$

une loi de student à $(n-p-1)$ degré de liberté.
On pose donc notre test comme suit :

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

— H_0 signifie qu'on va tester si la variable j n'apporte pas d'information supplémentaire importante sachant que les autres variables sont déjà dans le modèle.
— Si $\beta_j = 0$, nous obtenons la statistique de test suivante :

$$T_{test} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

Et $T_{test} \rightsquigarrow T_{n-p-1}$, suit une loi de Student à $(n-p-1)$ degrés de liberté.

Décision

— Si $|T_{test}| > t_{n-p-1}\left(\frac{1-\alpha}{2}\right)$, alors on rejette H_0 au seuil $\alpha \in [0,1]$ où $t_{n-p-1}(\frac{1-\alpha}{2})$ est le quantile d'ordre $\left(\frac{1-\alpha}{2}\right)$.
— Toutefois, il existe d'autres alternatives que les logiciels nous fournissent, pour connaître la crédibilité de notre hypothèse H_0. En effet, ils nous fournissent ce qu'on appelle la $p-value$.
— Si $p-value < \alpha$, où α est le risque fixé à 5% alors on rejette H_0.

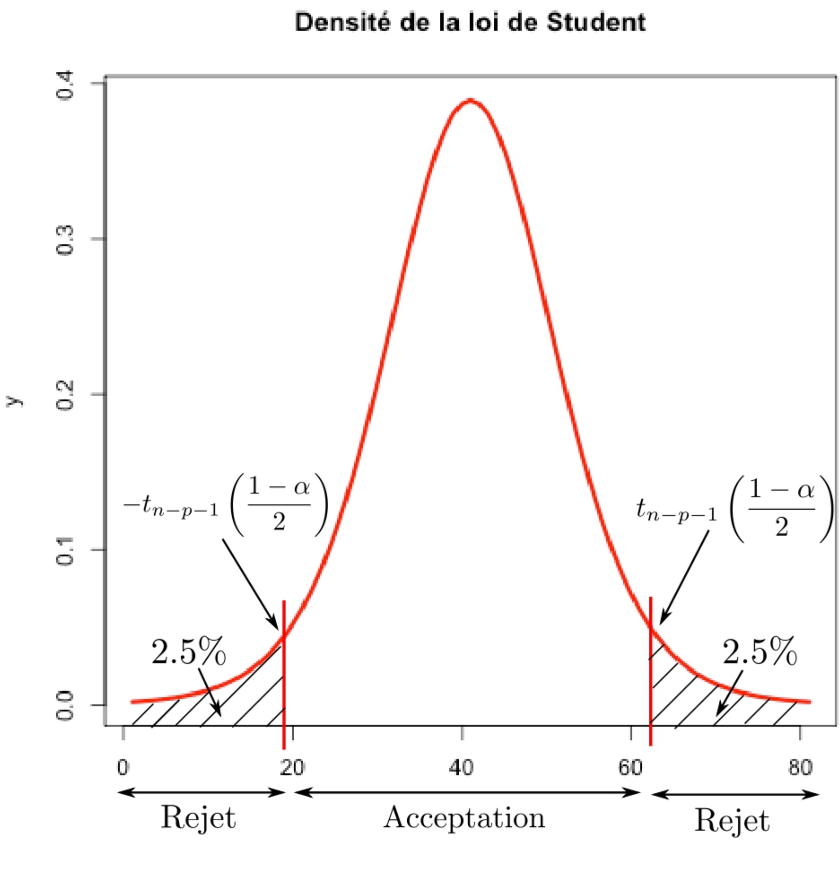

FIGURE 2.6 – Densité de Student avec régions de rejet.

III.2 Exemples d'applications

Reprenons le fichier "ozone" utilisé précédemment, pour illustrer la régression multiple.

Régression linéaire multiple à 2 variables explicatives :

Pour ce premier exemple, nous allons étudier la concentration en oxygène "maxO3" en fonction de la température "T12" et de la vitesse du vent "Vx9". Pour cela, commençons par tracer les nuages de points entre les différentes variables :

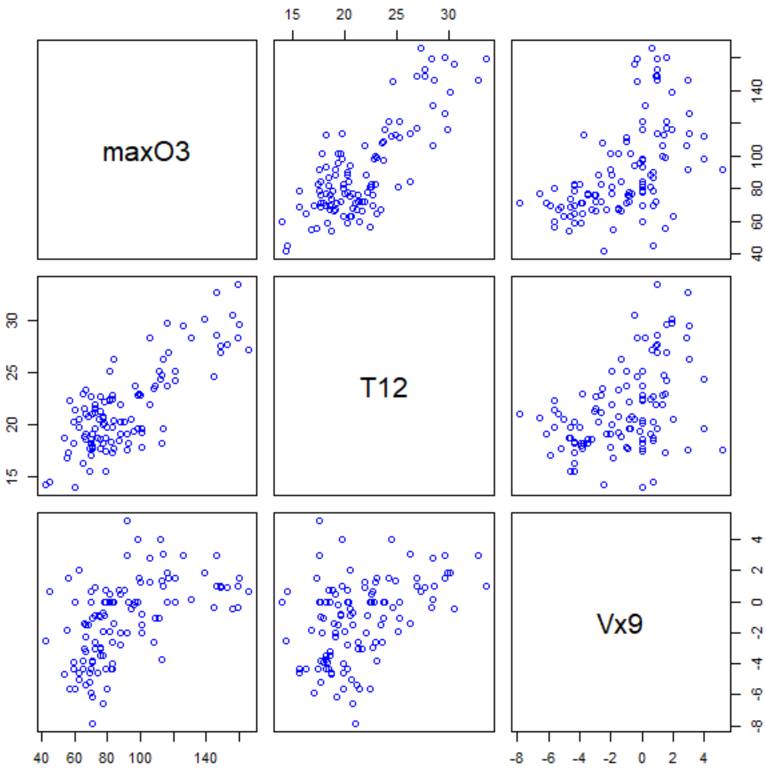

FIGURE 2.7 – Nuage de point des variables "max03", "T12" et "Vx9" prises deux à deux.

Nous remarquons que les variables "maxO3" et "T12" semblent être corrélées car la tendance de ce nuage de points ressemble à une droite. Maintenant, effectuons une régression linéaire multiple pour estimer la variable "maxO3" en fonction des variables explicatives "T12" et "Vx9" et observons ce qu'il se passe :

```
Call:
lm(formula = max03 ~ T12 + Vx9, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-31.441 -11.173   0.357   8.929  44.921

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.3083     9.7931  -0.951 0.343960
T12           4.7684     0.4317  11.046  < 2e-16 ***
Vx9           2.5000     0.6628   3.772 0.000264 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 16.6 on 109 degrees of freedom
Multiple R-squared:  0.6595,    Adjusted R-squared:  0.6533
F-statistic: 105.6 on 2 and 109 DF,  p-value: < 2.2e-16
```

Nous obtenons ainsi le modèle $maxO3 = -9.3083 + 4.7684 * T12 + 2.5 * Vx9$ avec un coefficient de détermination $R^2 = 0.6595$. Etant donné qu'un bon coefficient de détermination est proche de 1, on en conclut que ce modèle est peu satisfaisant et qu'il est possible de faire beaucoup mieux.

Régression linéaire multiple à 3 variables explicatives :

Pour ce second exemple, nous allons étudier la concentration en oxygène "maxO3" en fonction de la variable "T12", de "Vx9" et nous allons ajouter la nébulosité à 15h "Ne15". Commençons par tracer les nuages de points entre les différentes variables :

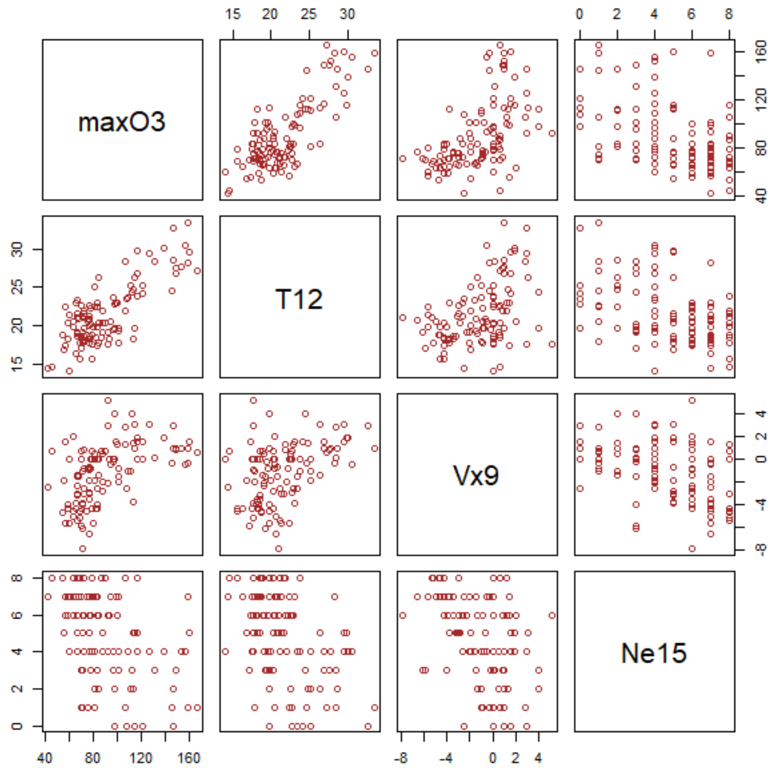

FIGURE 2.8 – Nuage de points des variables "max03", "T12", "Vx9" et "Ne15" prises deux à deux.

Comme précédemment, on remarque que les variables "maxO3" et "T12" semblent être corrélées. Effectuons maintenant une régression linéaire multiple pour estimer la variable "maxO3" en fonction des variables explicatives "T12", "Vx9" et "Ne15".

```
    Call:
lm(formula = max03 ~ T12 + Vx9 + Ne15, data = ozone)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-34.342 -10.947   0.122   8.792  43.675

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0825    11.9716   0.090  0.92812
T12           4.5327     0.4574   9.910  < 2e-16 ***
Vx9           2.2409     0.6816   3.288  0.00136 **
Ne15         -1.1662     0.7815  -1.492  0.13853
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.51 on 108 degrees of freedom
Multiple R-squared:  0.6664,    Adjusted R-squared:  0.6571
F-statistic: 71.91 on 3 and 108 DF,  p-value: < 2.2e-16
```

Ainsi, on obtient le modèle linéaire multiple $maxO3 = 1.0825 + 4.5327 * T12 + 4.5327 * Vx9 - 1.1662 * Ne15$" avec un coefficient de détermination $R^2 = 0.6664$. Ce modèle est donc meilleur que le précédent mais il reste peu satisfaisant.

Régression linéaire multiple à 10 variables explicatives :

Dans ce dernier exemple, nous voulons mesurer la concentration en O3 notée "maxO3" en fonction des 10 variables des données du tableaux "ozone". Commençons d'abord par tracer les nuages de points entre les différentes variables :

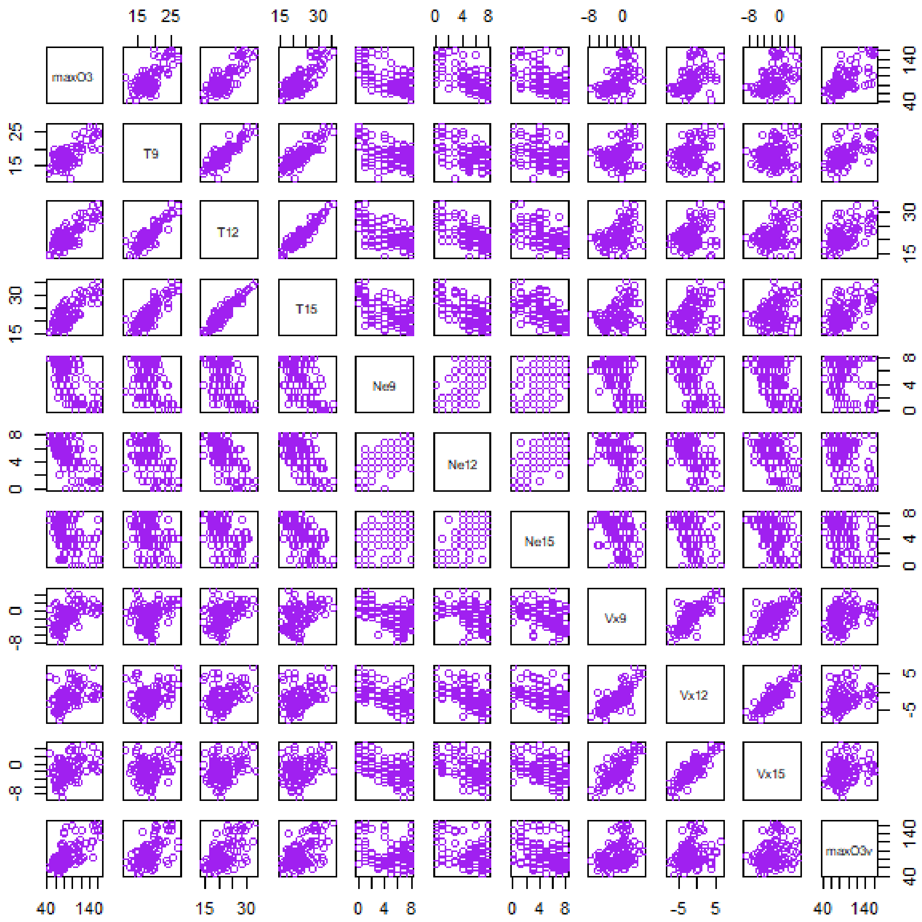

FIGURE 2.9 – Nuage de points des données ozone

Nous remarquons que certaines variables semblent être corrélées entre-elles. En effet, en observant les différents nuages de points, nous pouvons voir, par exemple, que la tendance entre les variables "T12" et "T15" ressemble à une droite. Il en est de même pour les variables "T9" et "T15", "T9" et "T12", "Vx12" et "Vx15", "maxO3" et "T12" et "maxO3" et "T15". Pour confirmer ces corrélations, regardons la matrice de corrélations entre les différentes variables :

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15
maxO3	1.000	0.699	0.784	0.775	-0.622	-0.641	-0.478	0.528	0.431	0.392
T9	0.699	1.000	0.883	0.846	-0.484	-0.472	-0.325	0.251	0.222	0.170
T12	0.784	0.883	1.000	0.946	-0.584	-0.660	-0.458	0.430	0.313	0.271
T15	0.775	0.846	0.946	1.000	-0.586	-0.649	-0.575	0.453	0.344	0.287
Ne9	-0.622	-0.484	-0.584	-0.586	1.000	0.788	0.550	-0.498	-0.529	-0.494
Ne12	-0.641	-0.472	-0.660	-0.649	0.788	1.000	0.710	-0.493	-0.510	-0.432
Ne15	-0.478	-0.325	-0.458	-0.575	0.550	0.710	1.000	-0.401	-0.432	-0.378
Vx9	0.528	0.251	0.430	0.453	-0.498	-0.493	-0.401	1.000	0.750	0.682
Vx12	0.431	0.222	0.313	0.344	-0.529	-0.510	-0.432	0.750	1.000	0.837
Vx15	0.392	0.170	0.271	0.287	-0.494	-0.432	-0.378	0.682	0.837	1.000
maxO3v	0.685	0.582	0.564	0.568	-0.277	-0.362	-0.308	0.340	0.224	0.190

```
       maxO3v
maxO3   0.685
T9      0.582
T12     0.564
T15     0.568
Ne9    -0.277
Ne12   -0.362
Ne15   -0.308
Vx9     0.340
Vx12    0.224
Vx15    0.190
```

Nous avons effectivement un coefficient de corrélation égal à 0.846 entre les variables "T9" et "T12". Nous avons également un coefficient de corrélation égal à 0.837 entre les variables "Vx12" et "Vx15". Comme ces coefficients sont proches de 1 alors la corrélation est confirmée.

```
> ozone=read.table("ozone.txt",header=TRUE)
> reg_mult=lm(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v,data=ozone)
> summary(reg_mult)

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

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

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

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

Ainsi, nous obtenons le modèle linéaire multiple suivant :
$maxO3 = 12.24442 - 0.01901 * T9 + 2.22115 * T12 + 0.55853 * T15 - 2.18909 * Ne9 - 0.42102 * Ne12 + 0.18373 * Ne15 + 0.94791 * Vx9 + 0.03120 * Vx12 + 0.41859 * Vx15 + 0.35198 * maxO3v$.
Nous constatons que le coefficient de détermination R^2 est égal à 0.7638. Rappelons qu'un bon coefficient de détermination doit être proche de 1. Par conséquent, ce modèle décrit la variable quantitative de manière correcte mais il est possible de faire mieux. Par la suite, nous verrons des méthodes de sélections de variables permettant d'améliorer ce modèle afin de mieux décrire la variable quantitative "maxO3" et donc, d'avoir un coefficient de détermination plus proche de 1.

Interprétation des résultats avec les tests :
En utilisant l'exemple présenté dans la régression linéaire multiple, nous allons interpréter les résultats obtenus au sens des tests.
Ci-dessous, les résultats du test de Fisher-Snédécor à 10 variables explicatives et 101 degrés de liberté :

```
> ozone=read.table("ozone.txt",header=TRUE)
> reg_mult=lm(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v,data=ozone)
> summary(reg_mult)

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

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

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

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

En observant la dernière ligne, nous avons :

— la statistique de test ($F_{test} = 32.67$),
— la probabilité critique ($p - value = 2.2e^{-16}$).
On remarque que la probabilité critique est très faible c'est-à-dire que $p-value < 0.05$, alors on rejette sans hésitation l'hypothèse H_0. On peut donc dire qu'il y a un effet d'au moins une variable explicative sur la concentration en ozone($maxO3$) .

Application : en utilisant toujours le même exemple, nous allons procéder au test de Student, afin de vérifier le rejet ou l'acceptation de notre hypothèse H_0 qui nous permet de déterminer lesquelles des variables explicatives sont importantes.
Les résultats sont présentés dans l'image ci dessus :

```
> ozone=read.table("ozone.txt",header=TRUE)
> reg_mult=lm(max03~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+max03v,data=ozone)
> summary(reg_mult)

Call:
lm(formula = max03 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 +
    Vx12 + Vx15 + max03v, data = ozone)

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

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.24442   13.47190   0.909   0.3656
T9          -0.01901    1.12515  -0.017   0.9866
T12          2.22115    1.43294   1.550   0.1243
T15          0.55853    1.14464   0.488   0.6266
Ne9         -2.18909    0.93824  -2.333   0.0216 *
Ne12        -0.42102    1.36766  -0.308   0.7588
Ne15         0.18373    1.00279   0.183   0.8550
Vx9          0.94791    0.91228   1.039   0.3013
Vx12         0.03120    1.05523   0.030   0.9765
Vx15         0.41859    0.91568   0.457   0.6486
max03v       0.35198    0.06289   5.597 1.88e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

— Prenons par exemple la température à 9h dont sa probabilité critique vaut 0.98, qui est très élevée et supérieure à 0.05. On en déduit donc qu'elle n'apporte pas d'information complémentaire intéressante, sachant que nous avons déjà dans notre modèle, les autres variables explicatives ($T12, T15, ..., maxOv3$). En effet, si nous construisons notre modèle en fonction de ($T12, T15, ..., maxOv3$), $T9$ ne sera pas importante à rajouter au modèle.

— Nous avons aussi la nébulosité à 9h ($Ne9$) qui apporte une information intéressante même si nous avons les autres variables dans notre modèle car sa $p-value = 0.02 < 0.05$.
— Cependant, la concentration en ozone est une variable qui apporte de l'information très intéressante même si nous avons les autres variables dans notre modèle car sa probabilité critique est très petite ($p - value = 1.88e^{-07} < 0.05$).

Ainsi nous ferons de même pour chaque variable pour voir si on peut la supprimer ou non.

Remarque 3.2. On ne peut pas supprimer toutes les variables simultanément car les tests construits supposent que toutes les autres sont dans le modèle. Par exemple on peut supprimer la variable $T12$ que si $T9$ et $T15$ sont présentes.

III.3 Cas de la régression simple

Rappelons que notre droite estimée est :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Nous allons effectuer un test individuel, qui va nous permettre de savoir s'il existe une relation entre la variable Y et X.
On pose :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Comme nous l'avons vue dans III.1.2 $\hat{\beta}_2 \rightsquigarrow N(\beta_2, \sigma^2_{\hat{\beta}_2})$.
Ainsi

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_j}} \rightsquigarrow T_{(n-2)}$$

Notre statistique de test sera donc :

$$T_{test} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

Décision :

— Si $|T_{test}| > t_{n-2}\left(\frac{1-\alpha}{2}\right)$, alors on rejette H_0 au seuil α.

Observons les résultats obtenus dans la régression simple pour interpréter le test individuel.
Les résultats sont comme suit :

```
> reg_simp=lm(max03~T12,data=ozone)
> summary(reg_simp)

Call:
lm(formula = max03 ~ 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
```

— Comme la probabilité critique de la variable $T12$ est d'une valeur de $2e^{-16}$ qui est inférieure à 0.05, alors on refuse l'hypothèse H_0 c'est à dire que β_2 associé à $T12$ est non nul. On peut dire donc que la température à midi est importante pour notre modèle.

IV Méthodes de sélection des variables

Nous avons vu précédemment que dans une régression multiple, la variable quantitative Y est décrite par p variables explicatives $X_1, ..., X_p$ où $p \geq 2$. Ainsi, nous obtenions le modèle suivant : $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$. Or, rien ne nous assure que toutes les variables explicatives décrivent la variable Y. Le but de cette partie sera donc d'établir les différentes méthodes qui permettent de sélectionner les variables qui décrivent le mieux la variable Y, afin d'obtenir le meilleur modèle à la fois performant (les résidus les plus petits possible) et économique (utiliser le moins de variables explicatives possibles).
Nous allons donc nous intéresser aux méthodes classiques de sélection de modèle.Les principaux critères de sélection sont :

IV.1 La sélection par $\hat{\sigma}^2$

Cette méthode consiste à choisir parmi tous les modèles, le modèle pour lequel $\hat{\sigma}^2(Y)$ est minimum.

IV.2 La sélection par R^2

L'objectif de cette méthode est de comparer le coefficient de corrélation R^2 des différents modèles où R^2 est défini par :

$$R^2(\zeta) = \frac{\|\hat{Y}(|\zeta|) - \bar{Y}\mathbb{1}\|^2}{\|Y - \bar{Y}\mathbb{1}\|^2} \quad = 1 - \frac{SCR(\zeta)}{SCT}$$

On peut remarquer que plus il y a de variables explicatives présentes dans le modèle, plus le coefficient de détermination augmente.

IV.3 La sélection par R^2 ajusté

Cette méthode consiste tout simplement à choisir le modèle dont le coefficient de détermination ajusté noté R_a^2 est maximum ce qui revient à minimiser la SCR divisé par son degré de liberté. Ce dernier diminue à chaque fois qu'on augmente le nombre de variables explicatives. Ce coefficient de détermination ajusté est défini de la manière suivante :

$$\begin{aligned} R_a^2 &= 1 - \frac{n-1}{n-p}(1-R^2) \\ &= 1 - \frac{n-1}{n-p}\frac{SCR}{SCT} \\ &= 1 - \frac{n-1}{SCT}\frac{SCR}{n-p} \end{aligned}$$

où $p \geq 2$ désigne le nombre de variables explicatives dans le modèle.

IV.4 Sélection par PRESS (Prédiction sum of squares)

On choisit le modèle pour lequel PRESS de Allen est minimum :

$$PRESS = \sum_{i=1}^{n} Y_i - \bar{Y}$$

IV.5 Le sélection par C_p de Mallows

Définition 4.1. Le $C_p(\zeta)$ d'un modèle à ζ variables explicatives où $\zeta = \{1, ..., p\}$ est défini par :

$$C_p(\zeta) = \frac{SCR}{\hat{\sigma}^2} - n + 2|\zeta| = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\hat{\sigma}^2} - n + 2p$$

où $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-p-1}$ est un estimateur "naturel" sans biais de σ^2, car on doit diviser par le nombre de données moins le nombre de paramètres à estimer. Ici c'est $n-p-1$.

Dans cette méthode, on sélectionne les modèles qui vérifient la relation suivante :

$$C_p(\zeta)) \leq |\zeta|$$

c'est-à-dire choisir le modèle pour lequel le $C_p(\zeta)$ est minimum.

IV.6 La vraisemblance et pénalisation

Sous l'hypothèse de la normalité des résidus, on peut calculer la log-vraisemblance de l'échantillon de variables aléatoires $(Y_1, ...Y_n)$ indépendantes, qui d'après II.1, suivent une loi normale de moyenne $X\beta$ et de variance σ^2. On a alors d'après II.4 :

La vraisemblance empirique est :

$$L(Y, \beta, \sigma^2) = \prod_{j=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \times \exp\left(\frac{-(Y - X\beta)^2}{2\sigma^2}\right) = \frac{1}{(\sigma\sqrt{2\pi})^n} \times \exp\left(\frac{-(\sum_{i=1}^{n} Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2}{2\sigma^2}\right)$$

Sa log-vraisemblance s'écrit alors sous la forme :

$$\log L(Y, \beta, \sigma^2) = -n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{\|Y - X\beta\|^2}{2\sigma^2} \tag{2.10}$$

Déterminons maintenant le maximum de vraisemblance(EMV)

D'après II.4,nous devons calculer :

$\dfrac{\partial \log L(Y, \beta, \sigma^2)}{\partial \beta}$ et $\dfrac{\partial \log L(Y, \beta, \sigma^2)}{\partial \sigma^2}$

On a alors :

$$\frac{\partial \log L(Y, \beta, \sigma^2)}{\partial \beta} = \frac{1}{2\sigma^2} \times \frac{\partial}{\partial \beta}\left(\|Y - X\beta\|^2\right) \tag{2.11}$$

et

$$\frac{\partial \log L(Y, \beta, \sigma^2)}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\|Y - X\beta\|^2\right) \tag{2.12}$$

Nous devons résoudre donc le système suivant :

$$\frac{\partial \log L(Y, \beta, \sigma^2)}{\partial \beta} = 0$$
$$\frac{\partial \log L(Y, \beta, \sigma^2)}{\partial \sigma^2} = 0$$

A partir de 2.11 nous avons $\hat{\beta}_{mv} = \hat{\beta}$, et à partir de 2.12 nous avons :

$$\frac{-n}{2} = \frac{-1}{2\hat{\sigma}^2_{mv}} \times \|Y - X\beta\|^2$$

D'où :

$$\hat{\sigma}^2_{mv} = \frac{\|Y - X\hat{\beta}_{mv}\|^2}{n} \tag{2.13}$$

Or on sait que $\hat{\sigma}^2 = \dfrac{\|Y - X\hat{\beta}\|^2}{n - p - 1}$, donc $\|Y - X\hat{\beta}_{mv}\| = (n - p - 1)\hat{\sigma}^2$. D'où 2.13 s'écrit :

$$\hat{\sigma}^2_{mv} = \frac{(n-p-1)}{n}\hat{\sigma}^2$$

Afin de vérifier que ces EMV sont maximums, nous devons calculer la dérivée seconde de la log-vraisemblance et s'assurer que :

$$\frac{\partial^2 \log L}{\partial^2 \beta}|_{\hat{\beta}_{mv}=\beta} \leq 0$$

et

$$\frac{\partial^2 \log L}{\partial^2 \sigma^2}|_{\hat{\sigma}^2_{mv}=\sigma^2} \leq 0$$

Le calcul de la log-vraisemblance (évaluée l'EMV) pour le modèle admettant p variables vaut alors :

$$\log L(Y, \hat{\beta}_{mv}, \hat{\sigma}^2_{mv}) = -\frac{n}{2}\log\left(\frac{\|Y - X\hat{\beta}_{mv}\|^2}{n}\right) - \frac{n}{2}\log(2\pi) - \frac{1}{2\hat{\sigma}^2_{mv}}\left(n \times \hat{\sigma}^2_{mv}\right)$$

On obtient donc :

$$\log L(Y, \hat{\beta}_{mv}, \hat{\sigma}^2_{mv}) = -\frac{n}{2}\log\left(\frac{SCR}{n}\right) - \frac{n}{2}\left(1 + \log(2\pi)\right)$$

Choisir un modèle en maximisant la vraisemblance revient á choisir le modèle ayant la plus petite SCR, il faut donc introduire une pénalisation. Afin de minimiser un critère, nous allons travailler avec l'opposée de la log-vraisemblance et les critères s'écrivent :

$$-2\log L(\zeta) + 2|\zeta| f(n)$$

IV.6.1 L'Akaike Information Criterion (AIC)

Cette méthode a été introduite en 1973 par Akaike. On définit l'AIC d'un modèle contenant p variables explicatives où $p \geq 2$ par :

$$AIC(\zeta) = -2\log L(\zeta) + 2|\zeta|$$

Par cette définition, f(n) vaut 1. L'AIC est une pénalisation de la log-vraisemblance, nous obtenons une définition équivalente :

$$AIC(\zeta) = cte + n\log\left(\frac{SCR(\zeta)}{n}\right) + 2|\zeta|$$

Le but de cette méthode est de calculer l'AIC de tous les modèles ayant p variables explicatives et de sélectionner celui qui a le plus faible AIC.

IV.7 Le critère Bayesian Information Criterion (BIC)

Ce critère a été introduit en 1978 par Schwarz. On définit le BIC d'un modèle à p variables explicatives par :

$$BIC(\zeta) = n \log \left(\frac{SCR(\zeta)}{n} \right) + p \log(n) + cste$$

Ce critère étant équivalent à celui de l'AIC, on cherche donc également à minimiser le BIC parmi les modèles candidats à p variables explicatives.

Maintenant que l'on a vu les différentes méthodes qui permettent d'optimiser un modèle linéaire à p variables explicatives, la question que nous nous posons est la suivante : de quelle manière doit-on utiliser ces méthodes ? En effet, lorsque l'on doit chercher le modèle à p variables explicatives qui décrivent le mieux la variable quantitative Y, nous devons utiliser les méthodes d'optimisation sur chaque modèle candidat. Or, le modèle optimisé peut contenir k variables explicatives où $1 \leq k \leq p$ et donc nous nous retrouvons avec $\frac{p!}{k!(p-k)!}$ combinaisons candidates. Si nous devions toutes les tester, on aurait énormément de calculs à faire, ce qui serait très contraignant. Pour palier à ce problème, on utilise des méthodes de recherche pas à pas. Il en existe trois :

— **La méthode ascendante (foward selection) :** cette méthode consiste à partir du modèle le plus simple, et à ajouter au fur et à mesure, une variable explicative au modèle linéaire, afin d'améliorer la condition d'un des critères d'optimalité ci-dessus. L'algorithme s'arrête lorsque toutes les variables explicatives sont intégrées au modèle ou lorsque la dernière variable explicative ajoutée ne permet pas une amélioration de la condition des critères d'optimalités par rapport à la précédente variable ajoutée.

— **La méthode descendante (backward selection) :** a partir du modèle complet, on retire au fur et à mesure la variable explicative la moins informative au modèle linéaire, afin d'améliorer la condition d'un des critères d'optimalité ci-dessus. L'algorithme s'arrête lorsque toutes les variables explicatives sont retirées du modèle ou lorsque la dernière variable explicative retirée ne permet pas une amélioration de la condition des critères d'optimalités par rapport à la précédente variable retirée.

— **La méthode progressive (stepwise selection) :** Cette méthode fonctionne de la même manière que la méthode ascendante sauf qu'en plus, il est possible de supprimer des variables explicatives ajoutées précédemment. En effet, lorsqu'on ajoute plusieurs variables explicatives à un modèle, il arrive parfois que les premières variables explicatives ajoutées ne décrivent plus la variable quantitative. Dans ce cas, afin d'améliorer la condition des critères d'optimalités, il faut donc les supprimer.

Certaines commandes ont étaient extraites du cours [8].

Retour sur l'application : Maintenant que nous avons vu les différentes méthodes de choix de variables, nous allons en appliquer certaines sur l'exemple "ozone" . Commençons d'abord par la méthode du R^2 où on construit tous les sous ensembles possibles et on retient celui pour lequel la probabilité critique du R^2 est la plus petite (on rejette plus fortement l'hypothèse H_0). Nous allons utiliser la fonction "**RegBest**" qui ressort le meilleur modèle à une variable explicative, puis le meilleur à deux variables explicatives jusqu'au meilleur modèle à 10 variables explicatives.

Cette fonction construit le meilleur modèle. Prenons l'exemple d'un modèle à une variable explicative :

— La fonction construit tous les modèles à une variable explicative et elle conserve celui qui a le R^2 le plus grand, et c'est la même chose pour tous les modèles à 2 variables explicatives jusqu'à 10.

— Comme on l'avait mentionné précédemment, le R^2 permet de comparer les modèles avec un même nombre de variables explicatives. Cependant, quand le nombre de variables diffère entre les modèles, nous allons utiliser la $p - value$ sur le test du R^2.

Avec les simulations réalisées avec R (voir annexe 87), les meilleurs modèles à $\{1, ..., 10\}$ variables explicatives sont résumés dans le tableau suivant :

```
$summary
                                R2       Pvalue
Model with 1 variable   0.7796310 6.404709e-38
Model with 2 variables  0.8013090 5.632629e-39
Model with 3 variables  0.8150119 2.013196e-39
Model with 4 variables  0.8252782 1.319017e-39
Model with 5 variables  0.8379452 3.000466e-40
Model with 6 variables  0.8463344 2.075040e-40
Model with 7 variables  0.8478473 1.280292e-39
Model with 8 variables  0.8487821 8.854028e-39
Model with 9 variables  0.8495137 6.107318e-38
Model with 10 variables 0.8495142 5.043277e-37
```

FIGURE 2.10 – Les meilleurs modèles

Comme on le voit sur l'image ci-dessus, plus on augmente le nombre de variables explicatives, plus le R^2 augmente ce qui est logique. Par contre le test de significativité de R^2 diminue entre 1 et 2 variables et aussi entre 2 et 3, 4 et 5 variables puis il augmente à partir de 5 variables explicatives. Le meilleur modèle est alors celui avec 5 variables explicatives qu'on va conserver. Ses caractéristiques sont présentés dans le tableau ci dessus :

```
$best

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.2768 -0.6155  0.0211  0.7290  3.3712 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.304508   1.164615  -0.261 0.794245    
T12          0.739464   0.045871  16.120  < 2e-16 ***
Ne9         -0.197005   0.078526  -2.509 0.013644 *  
Ne12         0.399491   0.094649   4.221 5.19e-05 ***
Vx9         -0.313050   0.072857  -4.297 3.88e-05 ***
Vx12         0.165060   0.068940   2.394 0.018428 *  
maxO3v       0.017655   0.005207   3.391 0.000983 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.259 on 105 degrees of freedom
Multiple R-squared:  0.8463,	Adjusted R-squared:  0.8376 
F-statistic: 96.38 on 6 and 105 DF,  p-value: < 2.2e-16
```

FIGURE 2.11 – Le meilleur modèle à 5 variables explicatives

On voit bien que toutes les $p-value$ des variables explicatives sont inférieures à 0.05, ce qui montre que ces cinq variables décrivent bien la concentration en ozone.
Ainsi, le modèle obtenu est le suivant :

$$maxO3 = -0.30 + 0.73T12 - 0.19Ne9 + 0.39Ne12 - 0.31Vx9 + 0.16Vx12 + 0.01maxO3v$$

R^2 ajusté avec la méthode descendante :

Reprenons le modèle de départ, vu précédemment, qui décrit la variable Y en fonction de toutes les autres :

$$Y = 12.24442 - 0.01901 * T9 + 2.22115 * T12 + 0.55853 * T15 - 2.18909 * Ne9 - 0.42102 * Ne12 + 0.18373 * Ne15 + 0.94791 * Vx9 + 0.03120 * Vx12 + 0.41859 * Vx15 + 0.35198 * maxO3v.$$

D'après les caractéristique de la régression linéaire multiple que nous avions effectuée précédemment avec le logiciel R, ce modèle a un R_a^2 égal à 0.7405. Notre but est d'améliorer ce modèle en supprimant les variables explicatives qui décrivent le moins la variable à estimer Y, afin d'obtenir un modèle linéaire avec le meilleur R_a^2.

— Etape 1 : Regardons tout d'abord les caractéristiques de la régression linéaire multiple du modèle à 10 variables :

```
    Call:
lm(formula = ozone$maxO3 ~ ozone$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
-53.566  -8.727  -0.403   7.599  39.458

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

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

D'après la colonne "Estimate", nous pouvons voir que la variable "T9" est celle qui décrit le moins la variable maxO3. De plus, on remarque également que sa probabilité critique, égale à 0.9866, est la plus élevée parmi les autres variables. Par conséquent, nous allons refaire une régression linéaire multiple, mais cette fois ci, en supprimant la variable "T9" :

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

Residuals:
    Min      1Q  Median      3Q     Max
-53.538  -8.726  -0.398   7.612  39.456

Coefficients:
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  12.25365  13.39469   0.915   0.3624
ozone$T12     2.20940   1.24659   1.772   0.0793 .
ozone$T15     0.55626   1.13114   0.492   0.6239
ozone$Ne9    -2.18538   0.90772  -2.408   0.0179 *
ozone$Ne12   -0.42784   1.30019  -0.329   0.7428
ozone$Ne15    0.18252   0.99531   0.183   0.8549
ozone$Vx9     0.95380   0.83882   1.137   0.2582
ozone$Vx12    0.02726   1.02410   0.027   0.9788
ozone$Vx15    0.41967   0.90897   0.462   0.6453
```

```
ozone$maxO3v  0.35165    0.05958   5.902 4.74e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.29 on 102 degrees of freedom
Multiple R-squared:  0.7638,    Adjusted R-squared:  0.743
F-statistic: 36.66 on 9 and 102 DF,  p-value: < 2.2e-16
```

Ainsi, nous obtenons le modèle :
$maxO3 = 12.25 + 2.20 * T12 + 0.55 * T15 - 2.18 * Ne9 - 0.42 * Ne12 + 0.18 * Ne15 + 0.95 * Vx9 + 0.02 * Vx12 + 0.41 * Vx15 + 0.35 * maxO3v$
avec un R_a^2= 0.743. Ce modèle est donc meilleur que le précédent.

— Etape 2 : En regardant la colonne "Estimate" des caractéristiques de la régression linéaires à 9 variables (voir ci-dessus), nous constatons que la variable "Vx12" est celle qui décrit le moins la variable maxO3. De plus, sa probabilité critique, égale à 0.9788, est la plus élevée. Ainsi, nous allons faire une nouvelle régression linéaire multiple en supprimant la variable "Vx12" :

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

Residuals:
    Min      1Q  Median      3Q     Max
-53.557  -8.738  -0.388   7.588  39.466

Coefficients:
                  Estimate     Std. Error  t value Pr(>|t|)
(Intercept)  12.30906   13.16757   0.935   0.3521
ozone$T12     2.20570    1.23279   1.789   0.0765 .
ozone$T15     0.55833    1.12298   0.497   0.6201
ozone$Ne9    -2.18603    0.90297  -2.421   0.0172 *
ozone$Ne12   -0.43285    1.28026  -0.338   0.7360
ozone$Ne15    0.18270    0.99044   0.184   0.8540
ozone$Vx9     0.96272    0.76521   1.258   0.2112
ozone$Vx15    0.43520    0.69370   0.627   0.5318
ozone$maxO3v  0.35163    0.05929   5.931 4.07e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.22 on 103 degrees of freedom
Multiple R-squared:  0.7638,    Adjusted R-squared:  0.7455
F-statistic: 41.64 on 8 and 103 DF,  p-value: < 2.2e-16
```

Par conséquent, on obtient le modèle :
$maxO3 = 12.3 + 2.2 * T12 + 0.55 * T15 - 2.18 * Ne9 - 0.43 * Ne12 + 0.18 * Ne15 + 0.96 * Vx9 + 0.43 * Vx15 + 0.35 * maxO3v$ avec un R_a^2 égale à 0.7455. Ce modèle est donc meilleur que le précédent.

— Etape 3 : En regardant la colonne "Estimate" des caractéristiques de la régression linéaire à 8 variables, nous constatons que la variable "Ne15" est celle qui décrit le moins la variable maxO3. De plus, sa probabilité critique, égale à 0.8540, est la plus élevée. Ainsi, nous allons faire une nouvelle régression linéaire multiple en supprimant la variable "Ne15"

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

Residuals:
    Min      1Q  Median      3Q     Max
-53.403  -8.637  -0.526   7.569  39.519

Coefficients:
                  Estimate    Std. Error t value  Pr(>|t|)
(Intercept)   12.6524    12.9747   0.975   0.3317
ozone$T12      2.3220     1.0543   2.202   0.0298 *
ozone$T15      0.4458     0.9384   0.475   0.6357
ozone$Ne9     -2.2029     0.8942  -2.464   0.0154 *
ozone$Ne12    -0.2998     1.0527  -0.285   0.7764
ozone$Vx9      0.9693     0.7608   1.274   0.2055
ozone$Vx15     0.4198     0.6855   0.612   0.5416
ozone$maxO3v   0.3514     0.0590   5.956 3.55e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.15 on 104 degrees of freedom
Multiple R-squared:  0.7638,    Adjusted R-squared:  0.7479
F-statistic: 48.03 on 7 and 104 DF,  p-value: < 2.2e-16
```

Par conséquent, on obtient le modèle :
$maxO3 = 12.65 + 2.32 * T12 + 0.44 * T15 - 2.20 * Ne9 - 0.29 * Ne12 + 0.96 * Vx9 + 0.41 * Vx15 + 0.35 * maxO3v$ avec un R_a^2 égal à 0.7479. Ce modèle est donc meilleur que le précédent.

— Etape 4 : En regardant la colonne "Estimate" des caractéristiques de la régression linéaire à 7 variables, nous constatons que la variable "Ne12" est celle qui décrit le moins la variable "maxO3". De plus, sa probabilité critique, égale à 0.7764, est la plus élevée. Nous allons donc faire une nouvelle régression linéaire multiple en supprimant la variable "Ne12".

```
        Call:
lm(formula = maxO3 ~ T12 + T15 + Ne9 + Vx9 + Vx15 + maxO3v, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-52.760  -8.418  -0.919   7.606  39.355

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 10.86699   11.30953   0.961  0.33883
T12          2.36755    1.03753   2.282  0.02451 *
T15          0.44749    0.93426   0.479  0.63295
Ne9         -2.35467    0.71469  -3.295  0.00134 **
Vx9          0.98502    0.75549   1.304  0.19515
Vx15         0.42639    0.68209   0.625  0.53325
maxO3v       0.35185    0.05872   5.992 2.95e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.09 on 105 degrees of freedom
Multiple R-squared:  0.7636,    Adjusted R-squared:  0.7501
F-statistic: 56.52 on 6 and 105 DF,  p-value: < 2.2e-16
```

Ainsi, on obtient le modèle :
$maxO3 = 10.86 + 2.36 * T12 + 0.44 * T15 - 2.35 * Ne9 + 0.98 * Vx9 + 0.42 * Vx15 + 0.35 * maxO3v$
avec un R_a^2 égale à 0.7501. Ce modèle est donc meilleur que le précédent.

— Etape 5 : En regardant la colonne "Estimate" des caractéristiques de la régression à 6 variables, nous constatons que la variable "maxO3v" est celle qui décrirait le moins la variable "maxO3". Or, lorsque l'on regarde les probabilités critiques de chaque variable, nous remarquons que la variable T15 a la plus grande probabilité critique, égale à 0.63295. Nous allons donc faire une nouvelle régression linéaire multiple en supprimant la variable "T15".

```
        Call:
lm(formula = ozone$maxO3 ~ +ozone$T12 + ozone$Ne9 + ozone$Vx9 +
    ozone$Vx15 + ozone$maxO3v)

Residuals:
    Min      1Q  Median      3Q     Max
-52.883  -8.261  -1.156   7.809  40.941

Coefficients:
                  Estimate  Std. Error  t value Pr(>|t|)
(Intercept)  11.40793   11.21202   1.017  0.31125
ozone$T12     2.80740    0.48111   5.835 5.91e-08 ***
ozone$Ne9    -2.38891    0.70852  -3.372  0.00104 **
ozone$Vx9     1.02125    0.74896   1.364  0.17559
ozone$Vx15    0.41655    0.67930   0.613  0.54106
ozone$maxO3v  0.35530    0.05806   6.119 1.61e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.04 on 106 degrees of freedom
Multiple R-squared:  0.7631,    Adjusted R-squared:  0.7519
F-statistic: 68.27 on 5 and 106 DF,  p-value: < 2.2e-16
```

Ainsi, on obtient le modèle :

$maxO3 = 11.40 + 2.82 * T12 - 2.38 * Ne9 + 1.02 * Vx9 + 0.41 * Vx15 + 0.35 * maxO3V$ avec un R_a^2 égale à 0.7519. Ce modèle est donc meilleur que le précédent.

— Etape 6 : En regardant la colonne "Estimate" des caractéristiques de la régression linéaire à 5 variables, nous constatons que la variable "maxO3v" est celle qui décrirait le moins la variable "maxO3". Or, il se trouve que la variable Vx15 a la plus grande probabilité critique, égale à 0.54106. Nous allons donc faire une nouvelle régression linéaire multiple en supprimant la variable "Vx15".

```
        Call:
lm(formula = maxO3 ~ T12 + Ne9 + Vx9 + maxO3v, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-52.396  -8.377  -1.086   7.951  40.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63131   11.00088   1.148 0.253443
T12          2.76409    0.47450   5.825 6.07e-08 ***
Ne9         -2.51540    0.67585  -3.722 0.000317 ***
Vx9          1.29286    0.60218   2.147 0.034055 *
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
```

Ainsi, on obtient le modèle :
$Y = 12.63 + 2.76 * T12 - 2.51 * Ne9 + 1.29 * Vx9 + 0.35 * maxO3v$ avec un R_a^2 égal à 0.7533. Ce modèle est donc meilleur que le précédent.

— Etape 7 : En regardant la colonne "Estimate" des caractéristiques de la régression à 4 variables, nous constatons que la variable "max03v" est celle qui décrirait le moins la variable "maxO3". Or, il se trouve que la variable Vx9 a la plus grande probabilité critique, égale à 0.034. Nous allons donc faire une nouvelle régression linéaire multiple en supprimant la variable "Vx9" :

```
        Call:
lm(formula = maxO3 ~ T12 + Ne9 + maxO3v, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-56.385  -7.872  -1.941   7.899  41.513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.76225   11.10038   0.879    0.381
T12          2.85308    0.48052   5.937 3.57e-08 ***
```

```
Ne9         -3.02423    0.64342  -4.700 7.71e-06 ***
maxO3v       0.37571    0.05801   6.477 2.85e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.23 on 108 degrees of freedom
Multiple R-squared:  0.752,     Adjusted R-squared:  0.7451
F-statistic: 109.1 on 3 and 108 DF,  p-value: < 2.2e-16
```

Ainsi, on obtient le modèle $maxO3 = 9.76 + 2.85 * T12 - 3.02 * Ne9 + 0.37 * maxO3v$ avec un R_a^2 égale à 0.7451. Ce R_a^2 étant inférieur à celui du modèle à 4 variables, on en conclut donc que le meilleur modèle linéaire multiple est :
$$maxO3 = 12.63 + 2.76 * T12 - 2.51 * Ne9 + 1.29 * Vx9 + 0.35 * maxO3v$$

Méthode descendante avec la fonction drop1 :

Cette fonction permet de construire le meilleur modèle en supprimant, à chaque itération, la variable la moins significative, c'est-à-dire, celle ayant le F-value (test de Fisher) le plus petit jusqu'à ce que toutes les probabilités critiques soient strictement inférieures au seuil de 0.05. Les détails de cette méthode sont présentés dans l'annexe (voir p78).
Ainsi, les caractéristiques de notre meilleur modèle sont :

```
    Call:
lm(formula = maxO3 ~ T12 + Vx9 + Ne9 + maxO3v, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-52.396  -8.377  -1.086   7.951  40.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63131   11.00088   1.148 0.253443
T12          2.76409    0.47450   5.825 6.07e-08 ***
Vx9          1.29286    0.60218   2.147 0.034055 *
Ne9         -2.51540    0.67585  -3.722 0.000317 ***
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
```

Par conséquent, le meilleur modèle est :

$$maxO3 = 12.63 + 2.76T12 - 2.51Ne9 + 1.29Vx9 + 0.35maxO3v$$

AIC avec la méthode stepwise : qui est un perfectionnement de la méthode pas a pas ascendante. A chaque étape, une procédure de sélection est effectuée par le critère AIC.
La procédure de sélection est faite comme suit :

- Commencer par le modèle le plus simple $Y = \beta_0 + \varepsilon$.
- Nous allons utiliser la fonction "**step**" qui va nous donner vers la fin des itérations le meilleur modèle.
- A chaque étape, on choisit l'AIC le plus petit (voir annexe p 93), et on rajoute la variable associée à ce dernier, au modèle précédent.

Ainsi le modèle obtenu à la fin de ces itérations est le suivant :

```
Call:
lm(formula = max03 ~ T12 + max03v + Ne9 + Vx9, data = ozone)

Coefficients:
(Intercept)          T12       max03v          Ne9          Vx9
    12.6313       2.7641       0.3548      -2.5154       1.2929
```

FIGURE 2.12 – Meilleur modèle avec la méthode stepwise

Méthode ascendante avec la fonction add1 : Cette fonction nous permet de construire notre meilleur modèle, où à chaque itération, on prend la variable la plus significative (celle ayant une p-value très petite) jusqu'à ce que toutes les p-value des variables qui restent, soient très élevées (dépassent le seuil $\alpha = 0.05$). Les résultats sont présentés dans l'annexe (voir p 95).

Ainsi nous obtenons les caractéristiques de notre meilleur modèle comme suit :

```
> summary(lm5)

Call:
lm(formula = max03 ~ T12 + max03v + Ne9 + Vx9, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-52.396  -8.377  -1.086   7.951  40.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63131   11.00088   1.148 0.253443
T12          2.76409    0.47450   5.825 6.07e-08 ***
max03v       0.35483    0.05789   6.130 1.50e-08 ***
Ne9         -2.51540    0.67585  -3.722 0.000317 ***
Vx9          1.29286    0.60218   2.147 0.034055 *
---
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
```

FIGURE 2.13 – Meilleur modèle avec la fonction add1

On a donc la droite de régression suivante :

$$maxO3 = 12.63 + 2.76T12 - 2.51Ne9 + 1.29Vx9 + 0.35maxO3v$$

On remarque qu'on a obtenu le même modèle que la méthode "**stepwise**", R^2_a et que les fonctions **drop1** et **add1**. Ceci montre la stabilité de notre modèle.

V Régression polynomiale

Cette méthode est un cas particulier de la régression linéaire multiple, où on cherche à exprimer une variable quantitative Y (exogène) en fonction de p puissances d'une seule variable quantitative (endogène) X où $p \geq 1$. De plus, lorsque $p = 1$, on retrouve la régression linéaire simple.

V.1 Modélisation statistique

Comme nous l'avons vu précédemment pour les régressions linéaires simples et multiples, nous cherchons à déterminer les coefficients $\beta_i, i = 0, ..., p$ tels que :

$$P(X) = Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta p X^p = \sum_{i=1}^{p} \beta_i X^i$$

ou encore

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_p X^p + \varepsilon_p \qquad \forall p \in \{1....n\}$$

où ε_i sont des variables aléatoires appelées "résidus".

Pour estimer les paramètres inconnus $(\beta_0, ..., \beta_p)$, nous allons utiliser la méthode des moindres carrés qui consiste à minimiser la quantité :

$$S(\beta_0, ..., \beta_p) = \sum_{i=0}^{N-1} P\,(x_i - y_i)^2 = \sum_{i=0}^{N-1} \left[\left(\sum_{k=0}^{p} \beta_k x_i^k \right) - y_i \right]^2$$

En calculant la dérivée partielle de S par rapport à β_j, on obtient :

$$\frac{\partial S}{\partial \beta_j}(\beta_0 ..., \beta p) = \sum_{i=0}^{N-1} 2x_i^j \left[\left(\sum_{k=0}^{p} \beta_k x_i^k \right) - y_i \right] = 2 \left[\sum_{k=0}^{p} \beta_k \sum_{i=0}^{N-1} x_i^j x_i^k - \sum_{i=0}^{N-1} x_i^j y_i \right]$$

Maintenant, on définit la matrice Q à N lignes et p+1 colonnes par $Q_{ij} = Q^*_{ji} = x_{i-1}^{j-1}$ où i=1,...,N et j=1,...,(p+1) et W, la matrice carrée symétrique de taille p+1 telle que :

$$W_{kk'} = (Q^* Q)_{kk'} = \sum_{i=1}^{N} Q^*_{ki} Q_{ik'} = \sum_{i=0}^{N-1} x_i^{(k-1)+(k'-1)}$$

où k=1,...,(p+1) et k'=1,...,(p+1).

On remarque que :

$$\begin{pmatrix} \frac{\partial S}{\partial \beta_0}(\beta_0 ..., \beta_p) \\ \vdots \\ \frac{\partial S}{\partial \beta_p}(\beta_0 ..., \beta_p) \end{pmatrix} = 2Q^* Q \begin{pmatrix} \beta_o \\ \vdots \\ \beta_p \end{pmatrix} - 2Q^* \begin{pmatrix} y_0 \\ \vdots \\ y_p \end{pmatrix}$$

Posons

$$B = \begin{pmatrix} \beta_o \\ \vdots \\ \beta_p \end{pmatrix}$$

et

$$Y = \begin{pmatrix} y_0 \\ \vdots \\ y_p \end{pmatrix}$$

Alors, pour annuler les dérivées partielles ci dessus, il suffit de résoudre l'équation :

$$Q^*QB = Q^*Y$$

Enfin, posons $A = Q^*Y$, alors on a le système :

$$WB = Y$$

V.1.1 Exemple d'application :

Dans cet exemple, on se propose de mesurer la production (notée PROD=Y) en fonction de la quantité de pesticides (notée Qte=X) utilisée dans un champ. Pour cela, nous disposons d'un ensemble de données que l'on présente ci-dessous :

Qte	0.5	0.6	0.7	0.8	0.9	1.0
Production	4.775	5.070	5.205	5.280	5.345	5.550
Qte	1.1	1.2	1.3	1.4	1.5	1.6
Production	5.595	5.730	5.705	5.720	5.925	5.870
Qte	1.8	1.9	2.0	2.1	2.2	2.3
Production	5.880	6.045	6.100	5.895	6.030	5.855
Qte	2.5	2.6	2.7	2.8	2.9	3.0
Production	5.775	5.870	5.755	5.580	5.695	5.500
Qte	3.2	3.3	3.4	3.5	3.6	3.7
Production	5.180	5.205	5.120	4.975	4.670	4.505
Qte	3.9	4.0				
Production	4.145	3.900				

FIGURE 2.14 – Données de production en fonction de la quantité de pesticides utilisée dans un champ

Commençons tout d'abord par tracer le nuage de points :

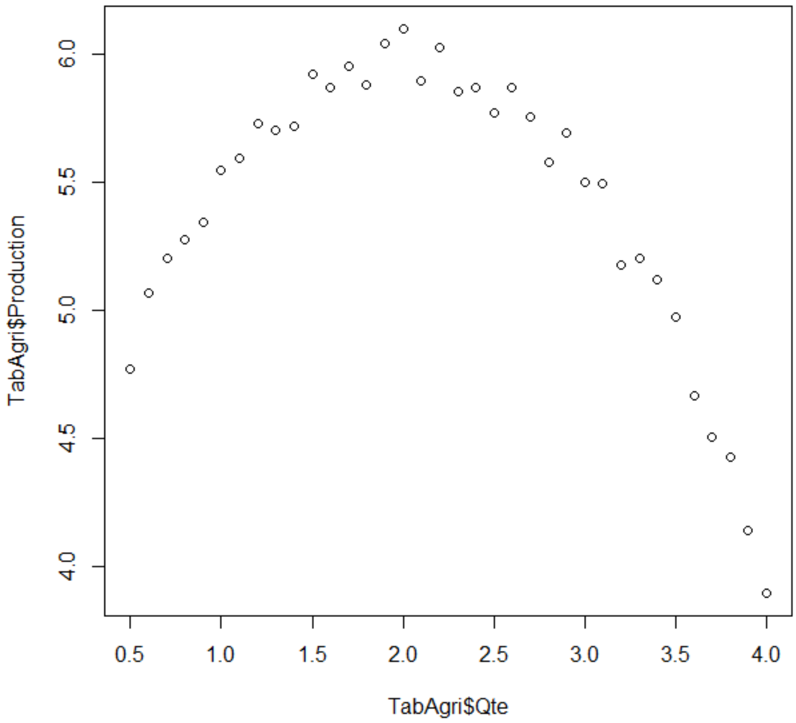

FIGURE 2.15 – Évolution de la production en fonction de la quantité de pesticides dans un champ

Maintenant, effectuons une régression linéaire simple et observons ce qu'il se passe :

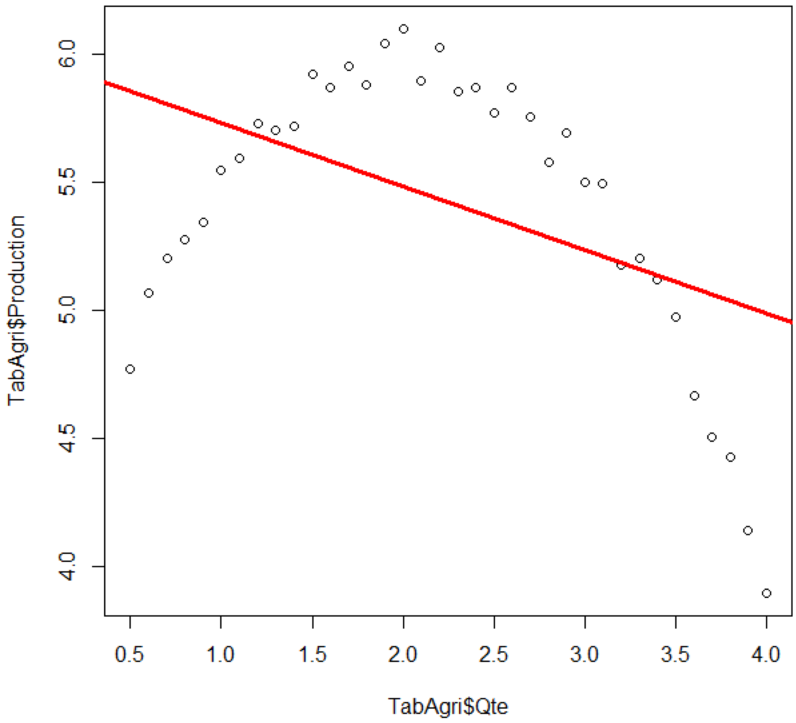

FIGURE 2.16 – Droite de régression linéaire

```
    Call:
lm(formula = Production ~ Qte, data = agri)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08754 -0.41364  0.06681  0.41971  0.61568

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.98110    0.19982  29.933  < 2e-16 ***
Qte         -0.24839    0.08063  -3.081  0.00407 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5026 on 34 degrees of freedom
Multiple R-squared:  0.2182,    Adjusted R-squared:  0.1952
F-statistic: 9.491 on 1 and 34 DF,  p-value: 0.004074
```

Lorsque l'on fait une régression linéaire simple, le coefficient de corrélation R^2 vaut 0.2182 ce qui est très mauvais. En effet, rappelons qu'un bon coefficient de corrélation doit être proche de 1 ou -1. Par conséquent, une régression linéaire simple n'est pas du tout pertinente pour ce

problème, il faut donc chercher un autre type de régression.
En observant la forme du nuage de points tracée ci-dessus, nous remarquons que la tendance ressemble à une cloche, ce qui nous fait penser à la courbe d'un polynôme. Par conséquent, une régression polynomiale semble plus pertinente c'est-à-dire, le modèle sera de la forme $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ...\beta_n X^n$ où β_0, β_1, ..., β_n sont les paramètres que l'on veut estimer où $n \geq 2$. Maintenant, effectuons la régression polynomiale de degré 2.

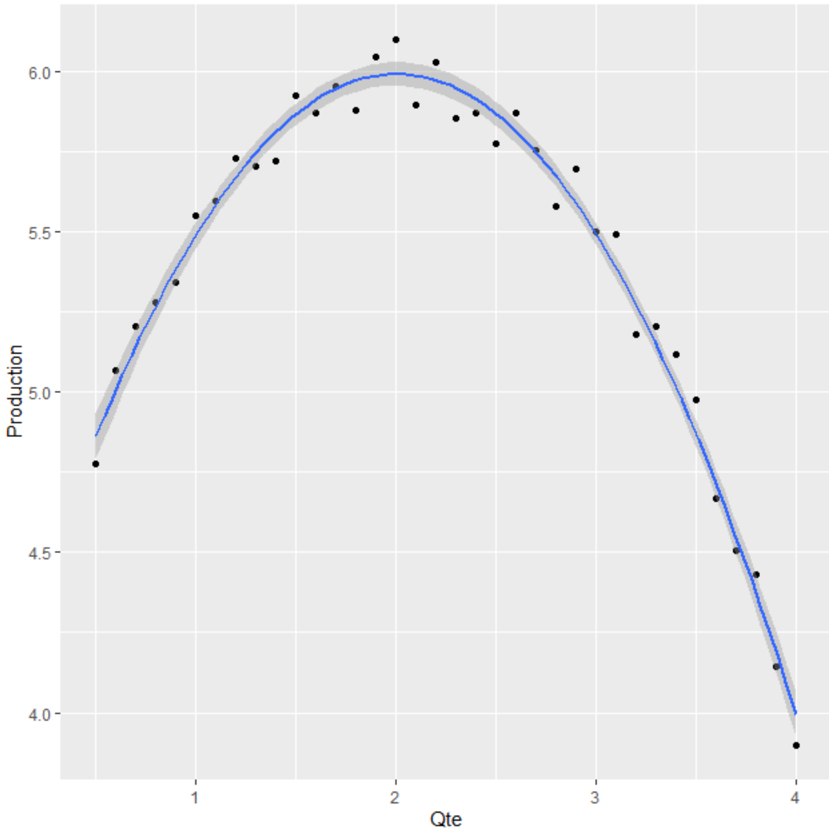

FIGURE 2.17 – Courbe de la régression polynomiale de degré 2

```
    Call:
lm(formula = Production ~ Qte + I(Qte^2), data = TabAgri)

Residuals:
      Min        1Q    Median        3Q       Max
-0.094888 -0.055395  0.006016  0.057806  0.106106

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.98458    0.05883   67.74   <2e-16 ***
Qte          2.00710    0.05876   34.16   <2e-16 ***
I(Qte^2)    -0.50122    0.01279  -39.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07398 on 33 degrees of freedom
```

```
Multiple R-squared:  0.9836,    Adjusted R-squared:  0.9826
F-statistic:   987 on 2 and 33 DF,  p-value: < 2.2e-16
```

Après avoir fait la régression polynomiale de degré 2, on obtient un $R^2 = 0,9836$ ce qui est très satisfaisant et beaucoup mieux que celui de la régression linéaire simple. Ainsi, nous obtenons le modèle suivant :

$$Y = 3.98458 + 2.00710X - 0.50122X^2$$

Maintenant, réalisons une régression polynomiale de degré 3 et observons ce qu'il se passe :

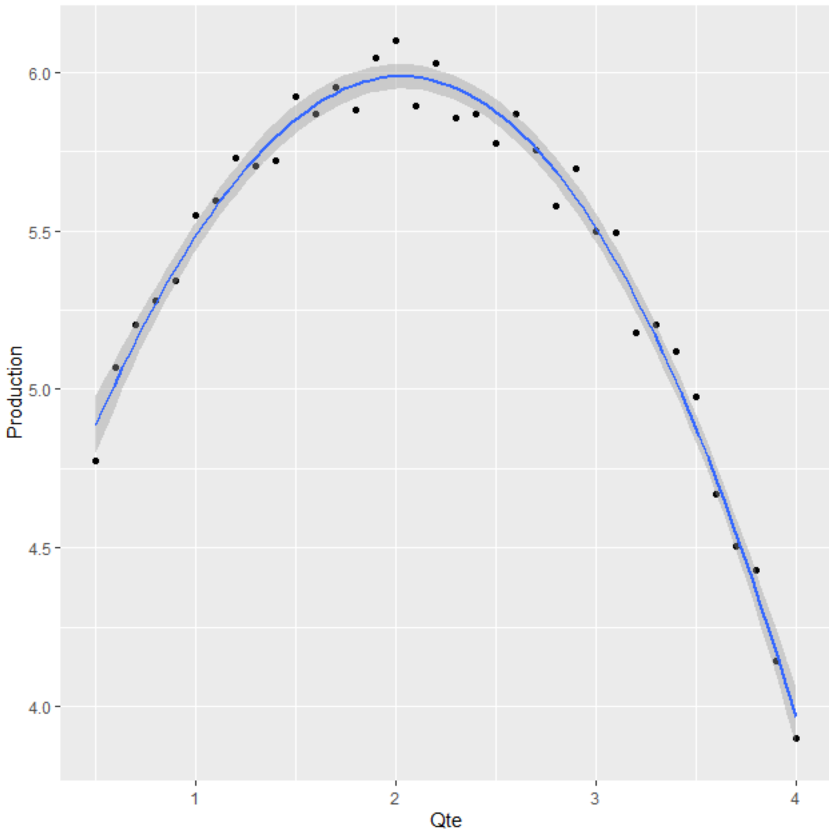

FIGURE 2.18 – Courbe de la régression polynomiale de degré 3

```
    Call:
lm(formula = Production ~ Qte + I(Qte^2) + I(Qte^3), data = TabAgri)

Residuals:
      Min        1Q    Median        3Q       Max
-0.115027 -0.052990  0.000792  0.065365  0.112612

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.08191    0.11501  35.492  < 2e-16 ***
Qte          1.82354    0.19539   9.333 1.19e-10 ***
I(Qte^2)    -0.40769    0.09581  -4.255  0.00017 ***
```

```
I(Qte^3)    -0.01386    0.01407  -0.985  0.33198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07401 on 32 degrees of freedom
Multiple R-squared:  0.984,     Adjusted R-squared:  0.9825
F-statistic: 657.7 on 3 and 32 DF,  p-value: < 2.2e-16
```

Lorsque l'on fait une régression polynomiale de degré 3, on obtient $R^2 = 0.984$ ce qui très satisfaisant et légèrement mieux qu'une régression polynomiale de degré 2. Ainsi, nous obtenons le modèle suivant :

$$Y = 4,08191 + 1,82354X - 0.40769X^2 - 0.01386X^3$$

Maintenant que l'on a déterminé un modèle linéaire qui décrit efficacement la variable Y, nous allons nous en servir pour prédire les prochaines données. Ainsi, quand $X = 2.5, 3, 3.5, 4$, on obtient les différentes valeurs de Y ci-dessous :

1	2	3	4
5.876204	5.509213	4.876024	3.966244

FIGURE 2.19 – Prédiction sur la quantité de production pour de nouvelles quantités de pesticides utilisées

CHAPITRE

3

MODÈLE LINÉAIRE QUALITATIF

I Analyse de la variance

Nous avons vu précédemment des modèles linéaires qui expriment une variable quantitative en fonction d'une ou plusieurs variables explicatives quantitatives. Cependant, il se trouve que dans certains cas, on a une variable quantitative Y qu'on cherche à exprimer en fonction d'une variable explicative qualitative. C'est pourquoi dans ce chapitre nous allons introduire l'analyse de la variance à un facteur.

I.1 Modélisation

Nous allons modéliser la concentration d'ozone en fonction du vent en provenance de 4 secteurs (EST, OUEST, NORD, SUD), on a donc 4 modalités. Les valeurs des 10 premiers individus sont présentées dans le tableau suivant :

Individus	maxO3	Vent
1	64	E
2	90	N
3	79	E
4	81	N
5	88	O
6	68	S
7	139	E
8	78	N
9	114	S
10	42	O

La représentation graphique des données en utilisant "boxplot sur R" de la variable Y par cellule est la suivante :

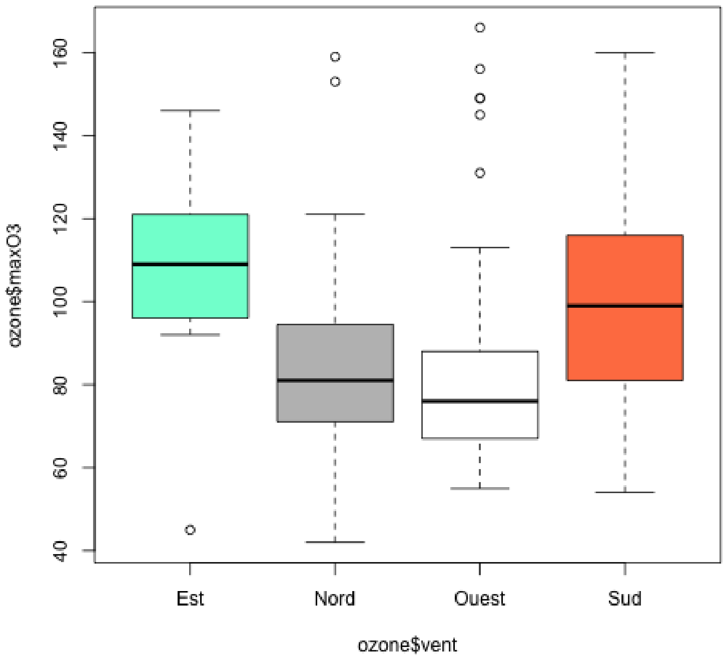

FIGURE 3.1 – Boxplot de la variable maxO3 en fonction du vent (4 modalités)

En observons le graphique, il semblerait que le vent ait une influence sur la concentration en ozone. En effet, lorsque le vent vient de l'EST, la concentration en ozone est très élevée contrairement aux vents venant du NORD ou de l'OUEST. Pour préciser ceci, effectuons l'analyse de la variance a un facteur.
Dans notre cas nous avons :

— $Y = maxO3$ est notre variable à expliquer.
— $A = vent$ est notre variable qualitative.

Comme A est qualitative, nous la remplaçons par $I = 4$ vecteurs : $\mathbb{1}_N$,$\mathbb{1}_S$,$\mathbb{1}_E$,$\mathbb{1}_O$, son codage disjonctif, afin de pouvoir l'intégrer dans un modèle de régression.
Ces 4 vecteurs sont regroupés dans la matrice $A_c = (\mathbb{1}_N, \mathbb{1}_S, \mathbb{1}_E, \mathbb{1}_O)$.
Le modèle de régression s'écrit :

$$Y = \mu\mathbb{1} + A_c\alpha + \varepsilon$$

La variable A engendre une partition des observations en $I = 4$ groupes (appelés aussi cellules). La i-ème cellule est constituée de n_i observations de la variable Y admettant le caractère i de la variable explicative.
Ici $n = 10$ individus où $n = \sum_{i=1}^{I} n_i$. Les données sont regroupées en cellules selon le tableau suivant :

Vent	Nord	SUD	EST	OUEST
maxO3	90 81 78	68 114	64 79 139	88 42

Par convention Y_{ij} correspond au j-ème individu de la cellule i. Les individus ne seront plus numérotés de 1 à n mais suivant le schéma $(1,1),(1,2),.......(I,1),...(I,n_I)$ pour bien insister sur l'appartenance de l'individu à la modalité i qui varie de 1 à n. Avec ces notations, le modèle précédent s'écrit alors sous la forme suivante :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$\forall i \in [1, I]$ et $\forall j \in [1, n_i]$.
Comme nous l'avons fait dans le cas des régressions précédentes, nous devons estimer les coefficients μ et α qui doivent être uniques, il faut donc se donner des contraintes linéaires.
Les plus classiques sont les suivantes :

— choisir $\mu = 0$, cela correspond à supprimer la colonne $\mathbb{1}$ et donc on pose $X = A_c$,
— choisir un des $\alpha_i = 0$,
— choisir $\sum n_i\alpha_i = 0$, la contrainte d'orthogonalité.Lorsque le plan est équilibré (tous les n_i sont égaux), cette contrainte devient $\sum \alpha_i = 0$.

Sous ces contraintes nous avons :

1. Sous $\mu = 0$, qui correspond à $Y_{ij} = \alpha_i + \varepsilon_{ij}$, les estimateurs des paramètres inconuus sont :
$$\hat{\alpha_i} = \bar{Y_i}$$
2. Sous la contrainte $\alpha_1 = 0$, qui correspond à $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, les estimateurs des paramètres inconnus sont :
$$\hat{\mu} = \bar{Y_1}$$
et
$$\hat{\alpha_i} = \bar{Y_i} - \bar{Y_1}$$
La première cellule sert de référence, le coefficient μ est donc égal à la moyenne empirique de la cellule de référence, les $\hat{\alpha_i}$ correspondent à l'effet différentiel entre la moyenne de la cellule i et la moyenne de la cellule de référence.
3. Sous la contrainte $\sum n_i\alpha_i = 0$, les estimateurs des paramètres inconnus sont :
$$\hat{\mu} = \bar{Y}$$
et
$$\hat{\alpha_i} = \bar{Y_i} - \bar{Y}$$

Dans tous les cas, σ^2 est estimé par :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y_i}\right)}{n - I}$$

où $n - I$ représente le nombre de paramètres à estimer.

I.2 Hypothèse gaussienne et test d'influence du facteur

Un des principaux objectifs de l'analyse de la variance est de vérifier si le facteur (dans notre cas c'est "le vent") possède une influence sur la variable expliquer $Y = maxO3$. Pour cela, nous devons aussi introduire l'hypothèse de normalité des résidus ε. Grâce à cette hypothèse

nous pouvons effectuer les tests d'hypothèses énoncés dans la partie III.1.1.
En choisissons le test de validité global, nous avons les hypothèses suivantes :

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = ...\alpha_I \\ H_1 : \alpha_i \neq \alpha_j \quad \exists(i,j) \end{cases}$$

Sous H_0, le modèle s'écrit aussi sous la forme suivante : $Y_{ij} = \mu + \varepsilon_{ij}$. Dans ce cas-là, nous avons notre statistique de test qui suit une loi de Fisher à $I-1$ et $n-I$ degrés de liberté comme dans III.1.1.

$$F_{test} = \frac{SCR/I-1}{SCE/(n-I)} = \frac{\sum_{i=1}^{I} n_i(\bar{Y}_i - \bar{Y})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2/(n-I)}$$

Il faut donc calculer les estimateurs des paramètres inconnus du modèle sous l'hypothèse H_0. Nous avons donc :

$$\hat{\mu} = \bar{Y}$$

et

$$\hat{\sigma_0}^2 = \frac{1}{n-1}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})$$

Maintenant que nous avons trouvé les estimateurs, nous souhaitons savoir l'influence de la variable qualitative sur la variable à expliquer Y. En utilisons la statistique de test énoncé en I.2, on refuse l'hypothèse H_0 ssi :

$$F_{test} > f_{I-1,n-I}(1-\alpha)$$

En refusant l'hypothèse H_0, nous pouvons conclure que la variable qualitative A a une influence sur la variable Y.

I.3 Application

En utilisant l'exemple "ozone", nous allons chercher à exprimer la concentration en ozone en fonction de la variable qualitative "vent".Nous avons donc quatre modalités cf I.1
. Comme on l'a vu précédemment le modèle s'écrit sous la forme :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Estimons ces paramètres en utilisant les contraintes énoncées précédemment :

1. $\alpha_1 = 0$, le logiciel R utilise par défaut $\alpha_1 = 0$ appelée contraste "**treatement**". Cela revient dans notre cas à prendre la cellule "EST" comme cellule de référence.Les résultats de l'analyse sont présentés dans le tableau suivant :

```
> ozone=read.table("ozone.txt",header=TRUE)
> mod1=lm(maxO3~vent,data=ozone)
> summary(mod1)

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
```

Dans l'image ci dessus, on a l'estimateur de μ, noté ici "**Intercept**" qui est la moyenne de la concentration en ozone "maxO3" pour le vent "EST". Les autres valeurs obtenues correspondent aux écarts entre la moyenne de la concentration en ozone de la cellule pour le vent considéré et la moyenne de la concentration de "maxO3" pour le vent d"EST(cellule de référence).
Le modèle obtenu est le suivant :

$$Y = 105.6 - 19.47 * \mathbb{1}_N - 20.900 * \mathbb{1}_O - 3.07 * \mathbb{1}_S$$

Maintenant que nous avons trouvé les estimateurs des paramètres inconnus, nous devons répondre à la question de l'influence du vent. Pour cela nous allons utiliser la fonction "**anova**" sur R . Le résultat est le suivant :

```
> anova(mod1)
Analysis of Variance Table

Response: maxO3
           Df Sum Sq Mean Sq F value  Pr(>F)
vent        3   7586 2528.69  3.3881 0.02074 *
Residuals 108  80606  746.35
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En regardant les résultats obtenus, nous remarquons que la valeur calculée est bien supérieure à la valeur théorique. De plus, la probabilité critique du vent est bien inférieure à 0.05. L'hypothèse H_0 est donc rejetée, ce qui signifie que le vent a bien une influence sur la concentration en ozone.

2. $\sum \alpha_i = 0$, cette contrainte est implémentée sur R :

```
> mod2=AovSum(maxO3~vent,data=ozone)
> mod2
Ftest
             SS  df      MS F value  Pr(>F)
vent       7586   3 2528.69  3.3881 0.02074 *
Residuals 80606 108  746.35
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ttest
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.7382     3.0535 31.0265   <2e-16 ***
vent         10.8618     6.8294  1.5904   0.1147
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nous retrouvons a nouveau le même tableau d'analyse de la variance.
L'effet du vent semble significatif. Nous aurons donc les mêmes estimateurs de nos paramètres inconnus.
Ainsi le modèle obtenu est le suivant :

$$Y = 105.6 - 19.47 * \mathbb{1}_N - 20.900 * \mathbb{1}_O - 3.07 * \mathbb{1}_S$$

Maintenant que l'on sait que le vent a une influence sur la concentration en ozone, essayons de voir si la pluie a également une influence .
Danc ce cas nous avons $I = 2$ (deux modalités : "pluies" et "Sec"). La représentation graphique de la concentration en ozone "maxO3" en fonction de la pluie est la suivante :

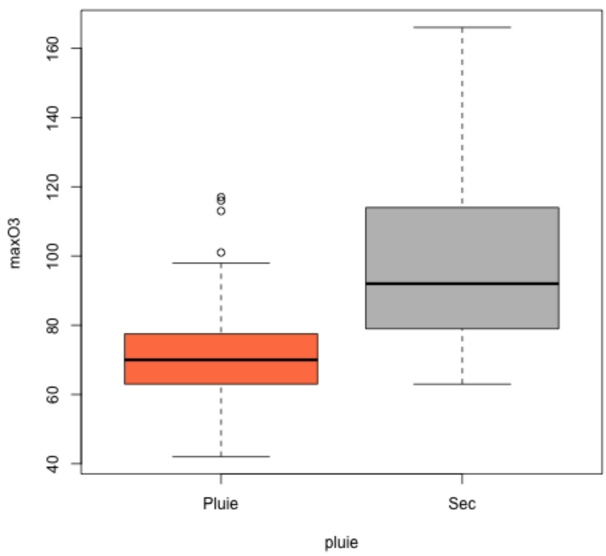

FIGURE 3.2 – Concentration de maxO3 en fonction de la pluie (2 modalités)

D'après l'image ci-dessus, il semblerait que la pluie ait une influence sur la concentration en ozone. En effet, lorsqu'il ne pleut pas, la concentration en ozone est très élevée contrairement

au cas où il pleut .
Nous avons :
$Y = maxO3$ la variable à expliquer.
$A = pluie$ la variable qualitative.
En remplaçant toujours A par son codage disjonctif, nous avons $A_c = (\mathbb{1}_S, \mathbb{1}_P)$.
Le modèle de régression s'écrit sous la forme suivante :

$$Y = \mu\mathbb{1} + A_c\alpha + \varepsilon$$

Nous allons chercher à exprimer la concentration en ozone en fonction de la variable qualitative "pluie", composée de deux modalités.
Estimons ces paramètres en utilisant les contraintes comme nous l'avons fait sur l'influence du vent.

1. Pour $\alpha_1 = 0$ nous avons :

```
> mod=lm(maxO3~pluie,data=ozone)
> summary(mod)

Call:
lm(formula = maxO3 ~ pluie, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-37.841 -17.841  -4.395  11.409  65.159

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   73.395      3.798  19.324  < 2e-16 ***
pluieSec      27.445      4.839   5.672 1.16e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.91 on 110 degrees of freedom
Multiple R-squared:  0.2263,Adjusted R-squared:  0.2192
F-statistic: 32.17 on 1 and 110 DF,  p-value: 1.157e-07
```

   Dans l'image ci-dessus, on a l'estimateur de μ, noté ici "**Intercept**", qui est la moyenne de la concentration en ozone "maxO3" pour la modalité "PLUIE". L'autre valeur obtenue correspond aux écarts entre la moyenne de la concentration en ozone de la cellule pour la modalité considérée et la moyenne de la concentration de "maxO3" pour la modalité "PLUIE"(cellule de référence).
   Le modèle obtenu est le suivant :

$$Y = 73.39 - 27.44 * \mathbb{1}_S$$

   Maintenant que nous avons trouvé les estimateurs des paramètres inconnus, nous devons répondre à la question de l'influence de la pluie. Pour cela nous allons utiliser la fonction "**anova**" sur R comme précédemment. Le résultat est le suivant :

```
> anova(mod)
Analysis of Variance Table
```

```
Response: maxO3
           Df Sum Sq Mean Sq F value    Pr(>F)
pluie       1  19954 19954.2  32.166 1.157e-07 ***
Residuals 110  68238   620.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En regardant les résultats obtenus, nous remarquons que la valeur calculée est bien supérieure a la valeur théorique. De plus, la probabilité critique de la pluie est bien inférieure à 0.05. L'hypothèse H_0 est donc rejetée, ce qui signifie que la pluie a bien une influence sur la concentration en ozone.

2. $\sum \alpha_i = 0$, nous obtenons les résultats suivants :

```
 > mod1=AovSum(maxO3~pluie,data=ozone)
> mod1
Ftest
             SS  df      MS F value    Pr(>F)
pluie     19954   1 19954.2  32.166 1.157e-07 ***
Residuals 68238 110   620.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ttest
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  87.1180     2.4196 36.0058 < 2.2e-16 ***
pluie       -13.7226     2.4196 -5.6715 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Nous retrouvons à nouveau le même tableau d'analyse de la variance.
L'effet de la pluie semble significatif. Nous aurons donc les mêmes estimateurs de nos paramètres inconnus.

CONCLUSION

Les modèles linéaires sont très importants pour l'étude d'un phénomène. En effet, durant la réalisation de ce projet, nous avons pu voir les différents modèles qui nous permettent d'exprimer une variable quantitative Y en fonction d'une ou plusieurs variables explicatives quelles que soient leur type (quantitative ou qualitative). Lorsque les variables explicatives sont quantitatives, nous avons trois modèles de régression linéaire différents : la linéaire simple, multiple et polynômiale. Les variables présentes dans ces modèles de régression décrivent plus ou moins la variable à estimer. C'est à l'aide des tests que l'on vérifie l'influence d'une variable explicative. De plus, ces modèles peuvent être améliorés à l'aide de différentes méthodes de choix de variables telles que AIC, R_a^2, ... Enfin, lorsque les variables explicatives sont qualitatives, on utilise plutôt l'analyse de la variance. Bien entendu, ce rapport ne contient pas tous les modèles linéaires existants. Sachez qu'il existe également un modèle linéaire mélangeant des variables explicatives quantitatives et des qualitatives appelée "Analyse de la covariance" (ANCOVA). Nous espérons que ce projet répondra à vos attentes.

ANNEXE

A

RAPPELS ET COMPLÉMENTS

Commandes R de la densité de la loi du khi-deux

```
{caption = Commande R de la loi chi 2}
>x_dchisq <- seq(0, 20, by = 0.1)
#Courbe de la densite de la loi du Chi-deux a 3 ddl
> b <- dchisq(x_dchisq, df = 3)
> plot(b, type="l", main="Densite du Chi-deux", xlab="x", ylab="y",col="red", lwd=3)
#Courbe de la densite de la loi du Chi-deux a 1 ddl
> c <- dchisq(x_dchisq, df = 1)
> lines(c, type="l", lty=2, col="green", lwd=3)
#Courbe de la densite de la loi du Chi-deux a 2 ddl
> d <- dchisq(x_dchisq, df = 2)
> lines(d, type="l", lty=3, col="pink", lwd=3)
#Courbe de la densite de la loi du Chi-deux a 4 ddl
> e <- dchisq(x_dchisq, df = 5)
> lines(e, type="l", lty=4, col=blue, lwd=3)
> legend(x="topright", legend=c("Chi2 1 ddl","Chi2 2 ddl", "Chi2 3 ddl", "Chi2 5 ddl"),
lty=c(2,3,1,4), col=c("green","pink","red","blue"), lwd=3)
```

Commandes R de la densité de la loi de Student

```
> x=seq(-4,4,0.1)
#Densite de Student a 10 ddl
> density10=dt(x,10)
#Densite de Student a 5 ddl
> density5=dt(x,5)
#Densite de Student a 2 ddl
> density2=dt(x,2)
```

```
#Densite de Student a 1 ddl
> density1=dt(x,1)
> png(file="Densite de Student.png")
> plot(density10,type="l",main="Densite de la loi de Student",xlab="x",ylab="y",
col="red",lwd=3)
> lines(density5,types="l",lty=3,col="blue",lwd=3)
> lines(density2,types="l",lty=2,col="purple",lwd=3)
> lines(density1,types="l",lty=4,col="green",lwd=3)
> legend(x="topright",legend=c("1ddl","2ddl","5ddl","10ddl"),lty=c(4,2,3,1),
col=c("green","purple","blue","red"),lwd=3)
> dev.off()
null device
```

Commandes R de la densité de la loi de Fisher-Snedecor

```
> x_df<-seq(0,8,by=0.1)
> b<-df(x_df,4,10)
> plot(b,type="l",main="Densite de la loi de Fisher-Snedecor de degrés de liberté 4
et 10", xlab="x", ylb="y", col="red", ltw=3)
```

ANNEXE

B

MODÈLE LINÉAIRE QUANTITATIF

Fichier ozone

```
"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"
"20010613" 101 17.3 19.3 20.2 7 7 3 -1.5 -1.5 -0.8682 106 "Nord" "Sec"
"20010614" 90 17.6 20.3 17.4 7 6 8 0.6946 -1.0419 -0.6946 101 "Sud" "Sec"
"20010615" 72 18.3 19.6 19.4 7 5 6 -0.8682 -2.7362 -6.8944 90 "Sud" "Sec"
"20010616" 70 17.1 18.2 18 7 7 7 -4.3301 -7.8785 -5.1962 72 "Ouest" "Pluie"
"20010617" 83 15.4 17.4 16.6 8 7 7 -4.3301 -2.0521 -3 70 "Nord" "Sec"
"20010618" 88 15.9 19.1 21.5 6 5 4 0.5209 -2.9544 -1.0261 83 "Ouest" "Sec"
"20010620" 145 21 24.6 26.9 0 1 1 -0.342 -1.5321 -0.684 121 "Ouest" "Sec"
"20010621" 81 16.2 22.4 23.4 8 3 1 0 0.3473 -2.5712 145 "Nord" "Sec"
"20010622" 121 19.7 24.2 26.9 2 1 0 1.5321 1.7321 2 81 "Est" "Sec"
"20010623" 146 23.6 28.6 28.4 1 1 2 1 -1.9284 -1.2155 121 "Sud" "Sec"
"20010624" 121 20.4 25.2 27.7 1 0 0 0 -0.5209 1.0261 146 "Nord" "Sec"
"20010625" 146 27 32.7 33.7 0 0 0 2.9544 6.5778 4.3301 121 "Est" "Sec"
"20010626" 108 24 23.5 25.1 4 4 0 -2.5712 -3.8567 -4.6985 146 "Sud" "Sec"
"20010627" 83 19.7 22.9 24.8 7 6 6 -2.5981 -3.9392 -4.924 108 "Ouest" "Sec"
"20010628" 57 20.1 22.4 22.8 7 6 7 -5.6382 -3.8302 -4.5963 83 "Ouest" "Pluie"
"20010629" 81 19.6 25.1 27.2 3 4 4 -1.9284 -2.5712 -4.3301 57 "Sud" "Sec"
"20010630" 67 19.5 23.4 23.7 5 5 4 -1.5321 -3.0642 -0.8682 81 "Ouest" "Sec"
"20010701" 70 18.8 22.7 24.9 5 2 1 0.684 0 1.3681 67 "Nord" "Sec"
"20010702" 106 24.1 28.4 30.1 0 0 1 2.8191 3.9392 3.4641 70 "Est" "Sec"
"20010703" 139 26.6 30.1 31.9 0 1 4 1.8794 2 1.3681 106 "Sud" "Sec"
"20010704" 79 19.5 18.8 17.8 8 8 8 0.6946 -0.866 -1.0261 139 "Ouest" "Sec"
"20010705" 93 16.8 18.2 22 8 8 6 0 0 1.2856 79 "Sud" "Pluie"
"20010706" 97 20.8 23.7 25 2 3 4 0 1.7101 -2.7362 93 "Nord" "Sec"
```

```
"20010707" 113 17.5 18.2 22.7 8 8 5 -3.7588 -3.9392 -4.6985 97 "Ouest" "Pluie"
"20010708" 72 18.1 21.2 23.9 7 6 4 -2.5981 -3.9392 -3.7588 113 "Ouest" "Pluie"
"20010709" 88 19.2 22 25.2 4 7 4 -1.9696 -3.0642 -4 72 "Ouest" "Sec"
"20010710" 77 19.4 20.7 22.5 7 8 7 -6.5778 -5.6382 -9 88 "Ouest" "Sec"
"20010711" 71 19.2 21 22.4 6 4 6 -7.8785 -6.8937 -6.8937 77 "Ouest" "Sec"
"20010712" 56 13.8 17.3 18.5 8 8 6 1.5 -3.8302 -2.0521 71 "Ouest" "Pluie"
"20010713" 45 14.3 14.5 15.2 8 8 8 0.684 4 2.9544 56 "Est" "Pluie"
"20010714" 67 15.6 18.6 20.3 5 7 5 -3.2139 -3.7588 -4 45 "Ouest" "Pluie"
"20010715" 67 16.9 19.1 19.5 5 5 6 -2.2981 -3.7588 0 67 "Ouest" "Pluie"
"20010716" 84 17.4 20.4 21.4 3 4 6 0 0.3473 -2.5981 67 "Sud" "Sec"
"20010717" 63 15.1 20.5 20.6 8 6 6 2 -5.3623 -6.1284 84 "Ouest" "Pluie"
"20010718" 69 15.1 15.6 15.9 8 8 8 -4.5963 -3.8302 -4.3301 63 "Ouest" "Pluie"
"20010719" 92 16.7 19.1 19.3 7 6 4 -2.0521 -4.4995 -2.7362 69 "Nord" "Sec"
"20010720" 88 16.9 20.3 20.7 6 6 5 -2.8191 -3.4641 -3 92 "Ouest" "Pluie"
"20010721" 66 18 21.6 23.3 8 6 5 -3 -3.5 -3.2139 88 "Sud" "Sec"
"20010722" 72 18.6 21.9 23.6 4 7 6 0.866 -1.9696 -1.0261 66 "Ouest" "Sec"
"20010723" 81 18.8 22.5 23.9 6 3 2 0.5209 -1 -2 72 "Nord" "Sec"
"20010724" 83 19 22.5 24.1 2 4 6 0 -1.0261 0.5209 81 "Nord" "Sec"
"20010725" 149 19.9 26.9 29 3 4 3 1 -0.9397 -0.6428 83 "Ouest" "Sec"
"20010726" 153 23.8 27.7 29.4 1 1 4 0.9397 1.5 0 149 "Nord" "Sec"
"20010727" 159 24 28.3 26.5 2 2 7 -0.342 1.2856 -2 153 "Nord" "Sec"
"20010728" 149 23.3 27.6 28.8 4 6 3 0.866 -1.5321 -0.1736 159 "Ouest" "Sec"
"20010729" 160 25 29.6 31.1 0 3 5 1.5321 -0.684 2.8191 149 "Sud" "Sec"
"20010730" 156 24.9 30.5 32.2 0 1 4 -0.5 -1.8794 -1.2856 160 "Ouest" "Sec"
"20010731" 84 20.5 26.3 27.8 1 0 2 -1.3681 -0.6946 0 156 "Nord" "Sec"
"20010801" 126 25.3 29.5 31.2 1 4 4 3 3.7588 5 84 "Est" "Sec"
"20010802" 116 21.3 23.8 22.1 7 7 8 0 -2.3941 -1.3892 126 "Sud" "Pluie"
"20010803" 77 20 18.2 23.6 5 7 6 -3.4641 -2.5981 -3.7588 116 "Ouest" "Pluie"
"20010804" 63 18.7 20.6 20.3 6 7 7 -5 -4.924 -5.6382 77 "Ouest" "Pluie"
"20010805" 54 18.6 18.7 17.8 8 8 8 -4.6985 -2.5 -0.8682 63 "Sud" "Pluie"
"20010806" 65 19.2 23 22.7 8 7 7 -3.8302 -4.924 -5.6382 54 "Ouest" "Sec"
"20010807" 72 19.9 21.6 20.4 7 7 8 -3 -4.5963 -5.1962 65 "Ouest" "Pluie"
"20010808" 60 18.7 21.4 21.7 7 7 7 -5.6382 -6.0622 -6.8937 72 "Ouest" "Pluie"
"20010809" 70 18.4 17.1 20.5 3 6 3 -5.9088 -3.2139 -4.4995 60 "Nord" "Pluie"
"20010810" 77 17.1 20 20.8 4 5 4 -1.9284 -1.0261 0.5209 70 "Nord" "Sec"
"20010811" 98 17.8 22.8 24.3 1 1 0 0 -1.5321 -1 77 "Ouest" "Pluie"
"20010812" 111 20.9 25.2 26.7 1 5 2 -1.0261 -3 -2.2981 98 "Ouest" "Sec"
"20010813" 75 18.8 20.5 26 8 7 1 -0.866 0 0 111 "Nord" "Sec"
"20010814" 116 23.5 29.8 31.7 1 3 5 1.8794 1.3681 0.6946 75 "Sud" "Sec"
"20010815" 109 20.8 23.7 26.6 8 5 4 -1.0261 -1.7101 -3.2139 116 "Sud" "Sec"
"20010819" 67 18.8 21.1 18.9 7 7 8 -5.3623 -5.3623 -2.5 86 "Ouest" "Pluie"
"20010820" 76 17.8 21.3 24 7 5 5 -3.0642 -2.2981 -3.9392 67 "Ouest" "Pluie"
"20010821" 113 20.6 24.8 27 1 1 2 1.3681 0.8682 -2.2981 76 "Sud" "Sec"
"20010822" 117 21.6 26.9 28.6 6 6 4 1.5321 1.9284 1.9284 113 "Sud" "Pluie"
"20010823" 131 22.7 28.4 30.1 5 3 3 0.1736 -1.9696 -1.9284 117 "Ouest" "Sec"
"20010824" 166 19.8 27.2 30.8 4 0 1 0.6428 -0.866 0.684 131 "Ouest" "Sec"
"20010825" 159 25 33.5 35.5 1 1 1 1 0.6946 -1.7101 166 "Sud" "Sec"
"20010826" 100 20.1 22.9 27.6 8 8 6 1.2856 -1.7321 -0.684 159 "Ouest" "Sec"
```

```
"20010827" 114 21 26.3 26.4 7 4 5 3.0642 2.8191 1.3681 100 "Est" "Sec"
"20010828" 112 21 24.4 26.8 1 6 3 4 4 3.7588 114 "Est" "Sec"
"20010829" 101 16.9 17.8 20.6 7 7 7 -2 -0.5209 1.8794 112 "Nord" "Pluie"
"20010830" 76 17.5 18.6 18.7 7 7 7 -3.4641 -4 -1.7321 101 "Ouest" "Sec"
"20010831" 59 16.5 20.3 20.3 5 7 6 -4.3301 -5.3623 -4.5 76 "Ouest" "Pluie"
"20010901" 78 17.7 20.2 21.5 5 5 3 0 0.5209 0 59 "Nord" "Pluie"
"20010902" 76 17.3 22.7 24.6 4 5 6 -2.9544 -2.9544 -2 78 "Ouest" "Pluie"
"20010903" 55 15.3 16.8 19.2 8 7 5 -1.8794 -1.8794 -2.3941 76 "Ouest" "Pluie"
"20010904" 71 15.9 19.2 19.5 7 5 3 -6.1284 0 -1.3892 55 "Nord" "Pluie"
"20010905" 66 16.2 18.9 19.3 2 5 6 -1.3681 -0.8682 1.7101 71 "Nord" "Pluie"
"20010906" 59 18.3 18.3 19 7 7 7 -3.9392 -1.9284 -1.7101 66 "Nord" "Pluie"
"20010907" 68 16.9 20.8 22.5 6 5 7 -1.5 -3.4641 -3.0642 59 "Ouest" "Pluie"
"20010908" 63 17.3 19.8 19.4 7 8 8 -4.5963 -6.0622 -4.3301 68 "Ouest" "Sec"
"20010912" 78 14.2 22.2 22 5 5 6 -0.866 -5 -5 62 "Ouest" "Sec"
"20010913" 74 15.8 18.7 19.1 8 7 7 -4.5963 -6.8937 -7.5175 78 "Ouest" "Pluie"
"20010914" 71 15.2 17.9 18.6 6 5 1 -1.0419 -1.3681 -1.0419 74 "Nord" "Pluie"
"20010915" 69 17.1 17.7 17.5 6 7 8 -5.1962 -2.7362 -1.0419 71 "Nord" "Pluie"
"20010916" 71 15.4 17.7 16.6 4 5 5 -3.8302 0 1.3892 69 "Nord" "Sec"
"20010917" 60 13.7 14 15.8 4 5 4 0 3.2139 0 71 "Nord" "Pluie"
"20010918" 42 12.7 14.3 14.9 8 7 7 -2.5 -3.2139 -2.5 60 "Nord" "Pluie"
"20010919" 65 14.8 16.3 15.9 7 7 7 -4.3301 -6.0622 -5.1962 42 "Ouest" "Pluie"
"20010920" 71 15.5 18 17.4 7 7 6 -3.9392 -3.0642 0 65 "Ouest" "Sec"
"20010921" 96 11.3 19.4 20.2 3 3 3 -0.1736 3.7588 3.8302 71 "Est" "Pluie"
"20010922" 98 15.2 19.7 20.3 2 2 2 4 5 4.3301 96 "Est" "Sec"
"20010923" 92 14.7 17.6 18.2 1 4 6 5.1962 5.1423 3.5 98 "Nord" "Sec"
"20010924" 76 13.3 17.7 17.7 7 7 6 -0.9397 -0.766 -0.5 92 "Ouest" "Pluie"
"20010925" 84 13.3 17.7 17.8 3 5 6 0 -1 -1.2856 76 "Sud" "Sec"
"20010927" 77 16.2 20.8 22.1 6 5 5 -0.6946 -2 -1.3681 71 "Sud" "Pluie"
"20010928" 99 16.9 23 22.6 6 4 7 1.5 0.8682 0.8682 77 "Sud" "Sec"
"20010929" 83 16.9 19.8 22.1 6 5 3 -4 -3.7588 -4 99 "Ouest" "Pluie"
"20010930" 70 15.7 18.6 20.7 7 7 7 0 -1.0419 -4 83 "Sud" "Sec"
```

Régression linéaire simple

```
#Lecture des données contenues dans le Fichier
> ozone=read.table("ozone.txt",header=TRUE)
#Calcul du modele de régression linéaire
> reg_simp=lm(data=ozone,maxO3~T12)
#Moyenne de Y
> mean(ozone$maxO3)
[1] 90.30357
#Moyenne de X
> mean(ozone$T12)
[1] 21.52679
#Variance de Y
> var(ozone$maxO3)
[1] 794.5196
```

```
#Variance de X
> var(ozone$T12)
[1] 16.34036
#Covariance de X et Y
> cov(ozone$maxO3,ozone$T12)
[1] 89.36026
#Affichage des détails de la régression
> summary(reg_simp)

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
#Affichage du  nuage de points et la droite de regression
> png(file="reg_simp.png")
> plot(maxO3~T12,data=ozone)
> abline(reg_simp,col="red",lwd=3)
> dev.off()
quartz
     2
# Affichage de la concentration d’ozone moyenne et la température moyenne a midi
> png(file="reg_simp.png")
> abline(h=mean(ozone$maxO3),lty=2)
> abline(v=mean(ozone$T12),lty=2)
> dev.off()
quartz 1
# Vérifier que les résidus suivent une loi normale
> png(file="residus.png")
> hist(residuals(reg_simp),col="blue",freq=F)
> y=density(residuals(reg_simp))
> m=lines(y,col="red",lwd=3)
> dev.off()
quartz
     2
```

Régression linéaire multiple

```
> ozone=read.table("ozone.txt", header=TRUE)
#On trace les nuages de points
> plot(ozone, col="purple")
#On détermine la matrice de corrélation
> round(cor(ozone),3)
             maxO3     T9    T12    T15    Ne9   Ne12   Ne15    Vx9   Vx12   Vx15
maxO3   1.000  0.699  0.784  0.775 -0.622 -0.641 -0.478  0.528  0.431  0.392
T9      0.699  1.000  0.883  0.846 -0.484 -0.472 -0.325  0.251  0.222  0.170
T12     0.784  0.883  1.000  0.946 -0.584 -0.660 -0.458  0.430  0.313  0.271
T15     0.775  0.846  0.946  1.000 -0.586 -0.649 -0.575  0.453  0.344  0.287
Ne9    -0.622 -0.484 -0.584 -0.586  1.000  0.788  0.550 -0.498 -0.529 -0.494
Ne12   -0.641 -0.472 -0.660 -0.649  0.788  1.000  0.710 -0.493 -0.510 -0.432
Ne15   -0.478 -0.325 -0.458 -0.575  0.550  0.710  1.000 -0.401 -0.432 -0.378
Vx9     0.528  0.251  0.430  0.453 -0.498 -0.493 -0.401  1.000  0.750  0.682
Vx12    0.431  0.222  0.313  0.344 -0.529 -0.510 -0.432  0.750  1.000  0.837
Vx15    0.392  0.170  0.271  0.287 -0.494 -0.432 -0.378  0.682  0.837  1.000
maxO3v  0.685  0.582  0.564  0.568 -0.277 -0.362 -0.308  0.340  0.224  0.190
       maxO3v
maxO3   0.685
T9      0.582
T12     0.564
T15     0.568
Ne9    -0.277
Ne12   -0.362
Ne15   -0.308
Vx9     0.340
Vx12    0.224
Vx15    0.190
maxO3v  1.000
#On effectue la régression linéaire multiple
>reg=lm(formula=ozone$maxO3~ozone$T9+ozone$T12+ozone$T15+ozone$Ne9+ozone$Ne12+
ozone$Ne15+ozone$Vx9+ozone$Vx12+ozone$Vx15+ozone$maxO3v)
#On détermine les caractéristiques de la régression.
> summary(reg)

Call:
lm(formula = ozone$maxO3 ~ ozone$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
-53.566  -8.727  -0.403   7.599  39.458

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.24442   13.47190   0.909   0.3656
ozone$T9     -0.01901    1.12515  -0.017   0.9866
```

```
ozone$T12      2.22115    1.43294   1.550   0.1243
ozone$T15      0.55853    1.14464   0.488   0.6266
ozone$Ne9     -2.18909    0.93824  -2.333   0.0216 *
ozone$Ne12    -0.42102    1.36766  -0.308   0.7588
ozone$Ne15     0.18373    1.00279   0.183   0.8550
ozone$Vx9      0.94791    0.91228   1.039   0.3013
ozone$Vx12     0.03120    1.05523   0.030   0.9765
ozone$Vx15     0.41859    0.91568   0.457   0.6486
ozone$maxO3v  0.35198    0.06289   5.597 1.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

Commandes de la fonction drop1 sur R :

```
> ozone=read.table("ozone.txt", header=TRUE)
> lm1=lm(maxO3~T9+T12+T15+Vx9+Vx12+Vx15+Ne9+Ne12+Ne15+maxO3v,data=ozone)
> drop1(lm1,maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v, test="F")
Single term deletions

Model:
maxO3 ~ T9 + T12 + T15 + Vx9 + Vx12 + Vx15 + Ne9 + Ne12 + Ne15 +
    maxO3v
       Df Sum of Sq   RSS    AIC F value   Pr(>F)
<none>               20827 607.26
T9      1       0.1 20827 605.26  0.0003  0.98655
T12     1     495.5 21323 607.89  2.4027  0.12425
T15     1      49.1 20876 605.52  0.2381  0.62664
Ne9     1    1122.6 21950 611.14  5.4438  0.02162 *
Ne12    1      19.5 20847 605.36  0.0948  0.75884
Ne15    1       6.9 20834 605.30  0.0336  0.85499
Vx9     1     222.6 21050 606.45  1.0796  0.30126
Vx12    1       0.2 20827 605.26  0.0009  0.97647
Vx15    1      43.1 20870 605.49  0.2090  0.64855
maxO3v  1    6459.6 27287 635.51 31.3251 1.88e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> lm2=lm(maxO3~T12+T15+Vx9+Vx12+Vx15+Ne9+Ne12+Ne15+maxO3v,data=ozone)
> drop1(lm2,maxO3~T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v, test="F")
Single term deletions

Model:
maxO3 ~ T12 + T15 + Vx9 + Vx12 + Vx15 + Ne9 + Ne12 + Ne15 + maxO3v
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>               20827 605.26
T12     1     641.4 21469 606.66  3.1412   0.07932 .
```

```
T15     1      49.4 20877 603.52  0.2418   0.62394
Ne9     1    1183.6 22011 609.45  5.7964   0.01786 *
Ne12    1      22.1 20849 603.38  0.1083   0.74278
Ne15    1       6.9 20834 603.30  0.0336   0.85486
Vx9     1     264.0 21091 604.67  1.2930   0.25817
Vx12    1       0.1 20827 603.26  0.0007   0.97882
Vx15    1      43.5 20871 603.49  0.2132   0.64528
maxO3v  1    7112.2 27940 636.16 34.8315 4.735e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> lm3=lm(maxO3~T12+T15+Vx9+Vx15+Ne9+Ne12+Ne15+maxO3v,data=ozone)
> drop1(lm3,maxO3~T12+T15+Ne9+Ne12+Ne15+Vx9+Vx15+maxO3v, test="F")
Single term deletions

Model:
maxO3 ~ T12 + T15 + Vx9 + Vx15 + Ne9 + Ne12 + Ne15 + maxO3v
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>              20827 603.26
T12     1     647.3 21475 604.69  3.2012   0.07652 .
T15     1      50.0 20877 601.53  0.2472   0.62012
Ne9     1    1185.1 22013 607.46  5.8609   0.01723 *
Ne12    1      23.1 20851 601.38  0.1143   0.73598
Ne15    1       6.9 20834 601.30  0.0340   0.85402
Vx9     1     320.1 21148 602.97  1.5829   0.21119
Vx15    1      79.6 20907 601.69  0.3936   0.53182
maxO3v  1    7112.6 27940 634.16 35.1748 4.067e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> lm4=lm(maxO3~T12+T15+Vx9+Vx15+Ne9+Ne12+maxO3v,data=ozone)
> drop1(lm4,maxO3~T12+T15+Ne9+Ne12+Vx9+Vx15+maxO3v, test="F")
Single term deletions

Model:
maxO3 ~ T12 + T15 + Vx9 + Vx15 + Ne9 + Ne12 + maxO3v
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>              20834 601.30
T12     1     971.8 21806 604.40  4.8510   0.02984 *
T15     1      45.2 20880 599.54  0.2257   0.63575
Ne9     1    1215.8 22050 605.65  6.0690   0.01540 *
Ne12    1      16.2 20851 599.38  0.0811   0.77639
Vx9     1     325.2 21160 601.03  1.6232   0.20548
Vx15    1      75.1 20910 599.70  0.3751   0.54157
maxO3v  1    7106.3 27941 632.17 35.4731 3.553e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> lm5=lm(maxO3~T12+T15+Vx9+Vx15+Ne9+maxO3v,data=ozone)
> drop1(lm5,maxO3~T12+T15+Ne9+Vx9+Vx15+maxO3v, test="F")
Single term deletions
```

```
Model:
maxO3 ~ T12 + T15 + Vx9 + Vx15 + Ne9 + maxO3v
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>              20851 599.38
T12     1    1034.0 21885 602.80  5.2072  0.024511 *
T15     1      45.6 20896 597.63  0.2294  0.632953
Ne9     1    2155.5 23006 608.40 10.8549  0.001344 **
Vx9     1     337.6 21188 599.18  1.6999  0.195150
Vx15    1      77.6 20928 597.80  0.3908  0.533245
maxO3v  1    7130.2 27981 630.33 35.9067 2.948e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> lm6=lm(maxO3~T12+Vx9+Vx15+Ne9+maxO3v,data=ozone)
> drop1(lm6,maxO3~T12+Ne9+Vx9+Vx15+maxO3v, test="F")
Single term deletions

Model:
maxO3 ~ T12 + Vx9 + Vx15 + Ne9 + maxO3v
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>              20896 597.63
T12     1    6712.6 27609 626.83 34.0509 5.911e-08 ***
Ne9     1    2241.1 23137 607.04 11.3684  0.001043 **
Vx9     1     366.5 21263 597.58  1.8593  0.175594
Vx15    1      74.1 20970 596.02  0.3760  0.541056
maxO3v  1    7381.5 28278 629.51 37.4444 1.608e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> lm7=lm(maxO3~T12+Vx9+Ne9+maxO3v,data=ozone)
> drop1(lm7,maxO3~T12+Ne9+Vx9+maxO3v, test="F")
Single term deletions

Model:
maxO3 ~ T12 + Vx9 + Ne9 + maxO3v
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>              20970 596.02
T12     1    6650.4 27621 624.88 33.9334 6.073e-08 ***
Ne9     1    2714.8 23685 607.66 13.8522 0.0003172 ***
Vx9     1     903.4 21874 598.75  4.6094 0.0340547 *
maxO3v  1    7363.5 28334 627.73 37.5721 1.499e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> lm7=lm(maxO3~T12+Vx9+Ne9+maxO3v,data=ozone)
> summary(lm7)

Call:
lm(formula = maxO3 ~ T12 + Vx9 + Ne9 + maxO3v, data = ozone)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-52.396  -8.377  -1.086   7.951  40.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63131   11.00088   1.148 0.253443
T12          2.76409    0.47450   5.825 6.07e-08 ***
Vx9          1.29286    0.60218   2.147 0.034055 *
Ne9         -2.51540    0.67585  -3.722 0.000317 ***
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
```

‘Commandes de la méthode R_a^2 à pas descendant

```
    > ozone=read.table("ozone.txt", header=TRUE))
Erreur : ’)’ inattendu(e) dans "ozone=read.table("ozone.txt", header=TRUE))"
> ozone=read.table("ozone.txt", header=TRUE)
> reg1=lm(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v, data=ozone)
> summary(reg1)

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

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

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

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

> reg2=lm(maxO3~T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v, data=ozone)
> summary(reg2)

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

Residuals:
    Min      1Q  Median      3Q     Max
-53.538  -8.726  -0.398   7.612  39.456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.25365   13.39469   0.915   0.3624
T12          2.20940    1.24659   1.772   0.0793 .
T15          0.55626    1.13114   0.492   0.6239
Ne9         -2.18538    0.90772  -2.408   0.0179 *
Ne12        -0.42784    1.30019  -0.329   0.7428
Ne15         0.18252    0.99531   0.183   0.8549
Vx9          0.95380    0.83882   1.137   0.2582
Vx12         0.02726    1.02410   0.027   0.9788
Vx15         0.41967    0.90897   0.462   0.6453
maxO3v       0.35165    0.05958   5.902 4.74e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.29 on 102 degrees of freedom
Multiple R-squared:  0.7638,    Adjusted R-squared:  0.743
F-statistic: 36.66 on 9 and 102 DF,  p-value: < 2.2e-16

> reg3=lm(maxO3~T12+T15+Ne9+Ne12+Ne15+Vx9+Vx15+maxO3v, data=ozone)
> summary(reg3)

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

Residuals:
    Min      1Q  Median      3Q     Max
-53.557  -8.738  -0.388   7.588  39.466

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 12.30906   13.16757   0.935   0.3521
T12          2.20570    1.23279   1.789   0.0765 .
T15          0.55833    1.12298   0.497   0.6201
Ne9         -2.18603    0.90297  -2.421   0.0172 *
Ne12        -0.43285    1.28026  -0.338   0.7360
Ne15         0.18270    0.99044   0.184   0.8540
Vx9          0.96272    0.76521   1.258   0.2112
Vx15         0.43520    0.69370   0.627   0.5318
maxO3v       0.35163    0.05929   5.931 4.07e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.22 on 103 degrees of freedom
Multiple R-squared:  0.7638,    Adjusted R-squared:  0.7455
F-statistic: 41.64 on 8 and 103 DF,  p-value: < 2.2e-16

> reg4=lm(maxO3~T12+T15+Ne9+Ne12+Vx9+Vx15+maxO3v, data=ozone)
> summary(reg4)

Call:
lm(formula = maxO3 ~ T12 + T15 + Ne9 + Ne12 + Vx9 + Vx15 + maxO3v,
    data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-53.403  -8.637  -0.526   7.569  39.519

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.6524    12.9747   0.975   0.3317
T12           2.3220     1.0543   2.202   0.0298 *
T15           0.4458     0.9384   0.475   0.6357
Ne9          -2.2029     0.8942  -2.464   0.0154 *
Ne12         -0.2998     1.0527  -0.285   0.7764
Vx9           0.9693     0.7608   1.274   0.2055
Vx15          0.4198     0.6855   0.612   0.5416
maxO3v        0.3514     0.0590   5.956 3.55e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.15 on 104 degrees of freedom
Multiple R-squared:  0.7638,    Adjusted R-squared:  0.7479
F-statistic: 48.03 on 7 and 104 DF,  p-value: < 2.2e-16

> reg5=lm(maxO3~T12+T15+Ne9+Vx9+Vx15+maxO3v, data=ozone)
> summary(reg5)

Call:
```

```
lm(formula = maxO3 ~ T12 + T15 + Ne9 + Vx9 + Vx15 + maxO3v, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max 
-52.760  -8.418  -0.919   7.606  39.355 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.86699   11.30953   0.961  0.33883    
T12          2.36755    1.03753   2.282  0.02451 *  
T15          0.44749    0.93426   0.479  0.63295    
Ne9         -2.35467    0.71469  -3.295  0.00134 ** 
Vx9          0.98502    0.75549   1.304  0.19515    
Vx15         0.42639    0.68209   0.625  0.53325    
maxO3v       0.35185    0.05872   5.992 2.95e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.09 on 105 degrees of freedom
Multiple R-squared:  0.7636,    Adjusted R-squared:  0.7501 
F-statistic: 56.52 on 6 and 105 DF,  p-value: < 2.2e-16

> reg6=lm(maxO3~T12+Ne9+Vx9+Vx15+maxO3v, data=ozone)
> summary(reg6)

Call:
lm(formula = maxO3 ~ T12 + Ne9 + Vx9 + Vx15 + maxO3v, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max 
-52.883  -8.261  -1.156   7.809  40.941 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.40793   11.21202   1.017  0.31125    
T12          2.80740    0.48111   5.835 5.91e-08 ***
Ne9         -2.38891    0.70852  -3.372  0.00104 ** 
Vx9          1.02125    0.74896   1.364  0.17559    
Vx15         0.41655    0.67930   0.613  0.54106    
maxO3v       0.35530    0.05806   6.119 1.61e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.04 on 106 degrees of freedom
Multiple R-squared:  0.7631,    Adjusted R-squared:  0.7519 
F-statistic: 68.27 on 5 and 106 DF,  p-value: < 2.2e-16

> reg7=lm(maxO3~T12+Ne9+Vx9+maxO3v, data=ozone)
```

```
> summary(reg7)

Call:
lm(formula = maxO3 ~ T12 + Ne9 + Vx9 + maxO3v, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-52.396  -8.377  -1.086   7.951  40.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63131   11.00088   1.148 0.253443
T12          2.76409    0.47450   5.825 6.07e-08 ***
Ne9         -2.51540    0.67585  -3.722 0.000317 ***
Vx9          1.29286    0.60218   2.147 0.034055 *
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

> reg8=lm(maxO3~T12+Ne9+maxO3v, data=ozone)
> summary(reg8)

Call:
lm(formula = maxO3 ~ T12 + Ne9 + maxO3v, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
-56.385  -7.872  -1.941   7.899  41.513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.76225   11.10038   0.879    0.381
T12          2.85308    0.48052   5.937 3.57e-08 ***
Ne9         -3.02423    0.64342  -4.700 7.71e-06 ***
maxO3v       0.37571    0.05801   6.477 2.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.23 on 108 degrees of freedom
Multiple R-squared:  0.752,     Adjusted R-squared:  0.7451
F-statistic: 109.1 on 3 and 108 DF,  p-value: < 2.2e-16
```

Commandes de la régression polynômiale sur R

```
> TabAgri = read.table("agri_3.txt",header=TRUE)
```

```
#On trace le nuage de points.
> plot(x=TabAgri$Qte,TabAgri$Production)
#On effectue et trace la régression linéaire simple.
>  regression_lineaire = lm(data = TabAgri, Production~Qte)
> abline(regression_lineaire, col="red", lwd=3)
#On regarde les données de la régression linéaire simple.
> summary(regression_lineaire)
Call:
lm(formula = Production ~ Qte, data = TabAgri)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08754 -0.41364  0.06681  0.41971  0.61568

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.98110    0.19982  29.933  < 2e-16 ***
Qte         -0.24839    0.08063  -3.081  0.00407 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5026 on 34 degrees of freedom
Multiple R-squared:  0.2182,    Adjusted R-squared:  0.1952
F-statistic: 9.491 on 1 and 34 DF,  p-value: 0.004074
#On effectue et trace la régression polynomiale de degré 2.
> regression_poly_degre_2 = lm(data=TabAgri,Production~Qte+I(Qte^2))
> library(tidyverse)
> library(caret)
> ggplot(TabAgri, aes(Qte, Production) ) + geom_point() +
+ stat_smooth(method = lm, formula = y ~ poly(x, 2, raw = TRUE))
> summary(regression_poly_degre_2 )
Call:
lm(formula = Production ~ Qte + I(Qte^2), data = TabAgri)

Residuals:
      Min        1Q    Median        3Q       Max
-0.094888 -0.055395  0.006016  0.057806  0.106106

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.98458    0.05883   67.74   <2e-16 ***
Qte          2.00710    0.05876   34.16   <2e-16 ***
I(Qte^2)    -0.50122    0.01279  -39.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07398 on 33 degrees of freedom
Multiple R-squared:  0.9836,    Adjusted R-squared:  0.9826
```

```
F-statistic:   987 on 2 and 33 DF,  p-value: < 2.2e-16
#On effectue et trace la régression polynomiale de degré 3.
>  regression_poly_degre_3 = lm(data=TabAgri,Production~Qte+I(Qte^2)+I(Qte^3))
> ggplot(TabAgri, aes(Qte, Production) ) + geom_point() +
+ stat_smooth(method = lm, formula = y ~ poly(x, 3, raw = TRUE))
> summary(regression_poly_degre_3 )
Call:
lm(formula = Production ~ Qte + I(Qte^2) + I(Qte^3), data = TabAgri)

Residuals:
      Min        1Q    Median        3Q       Max
-0.115027 -0.052990  0.000792  0.065365  0.112612

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.08191    0.11501  35.492  < 2e-16 ***
Qte          1.82354    0.19539   9.333 1.19e-10 ***
I(Qte^2)    -0.40769    0.09581  -4.255  0.00017 ***
I(Qte^3)    -0.01386    0.01407  -0.985  0.33198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07401 on 32 degrees of freedom
Multiple R-squared:  0.984,     Adjusted R-squared:  0.9825
F-statistic: 657.7 on 3 and 32 DF,  p-value: < 2.2e-16
#Prédictions des prochaines données
>  nouvelles_donnees_Agri = data.frame( Qte = c(2.5,3,3.5,4) )
> prediction_nouvelles_Agri = predict(regression_poly_degre_3,nouvelles_donnees_Agri)
> prediction_nouvelles_Agri

   1              2               3               4
5.876204  5.509213  4.876024  3.966244
```

Méthode de sélection par R^2

```
> RegBest(y=ozone[,2],x=ozone[,-2],nbest=1)
$all
$all[[1]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6100 -0.6892 -0.0199  0.9569  4.2934

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.67731    0.75722   4.856 3.98e-06 ***
T12          0.68210    0.03458  19.727  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.473 on 110 degrees of freedom
Multiple R-squared:  0.7796,Adjusted R-squared:  0.7776
F-statistic: 389.2 on 1 and 110 DF,  p-value: < 2.2e-16


$all[[2]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8562 -0.6348 -0.0155  0.7873  4.3692

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17977    1.24513   0.144 0.885470
T12          0.78205    0.04391  17.811  < 2e-16 ***
Ne12         0.26823    0.07778   3.449 0.000802 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.405 on 109 degrees of freedom
Multiple R-squared:  0.8013,Adjusted R-squared:  0.7977
F-statistic: 219.8 on 2 and 109 DF,  p-value: < 2.2e-16


$all[[3]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0834 -0.5810  0.1249  0.6482  3.1086

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.068439   0.807751   2.561 0.011826 *
T12          0.675692   0.040666  16.616  < 2e-16 ***
Vx9         -0.209117   0.054850  -3.813 0.000229 ***
maxO3v       0.016483   0.005581   2.953 0.003860 **
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.362 on 108 degrees of freedom
Multiple R-squared:  0.815,Adjusted R-squared:  0.8099
F-statistic: 158.6 on 3 and 108 DF,  p-value: < 2.2e-16
$all[[4]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6188 -0.5406  0.0835  0.6459  3.2872

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.140925   1.182544  -0.119  0.90536
T12          0.738978   0.047049  15.707  < 2e-16 ***
Ne12         0.194362   0.077515   2.507  0.01366 *
Vx9         -0.164908   0.056383  -2.925  0.00421 **
maxO3v       0.015659   0.005459   2.868  0.00497 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.329 on 107 degrees of freedom
Multiple R-squared:  0.8253,Adjusted R-squared:  0.8187
F-statistic: 126.4 on 4 and 107 DF,  p-value: < 2.2e-16


$all[[5]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6867 -0.6263 -0.0378  0.6407  3.1428

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.351295   1.156938   0.304 0.761996
T12          0.718907   0.046055  15.610  < 2e-16 ***
Ne9         -0.227886   0.079170  -2.878 0.004835 **
Ne12         0.365064   0.095616   3.818 0.000227 ***
Vx9         -0.197284   0.055703  -3.542 0.000592 ***
maxO3v       0.017504   0.005321   3.289 0.001363 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.286 on 106 degrees of freedom
Multiple R-squared:  0.8379,Adjusted R-squared:  0.8303
F-statistic: 109.6 on 5 and 106 DF,  p-value: < 2.2e-16


$all[[6]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2768 -0.6155  0.0211  0.7290  3.3712

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.304508   1.164615  -0.261 0.794245
T12          0.739464   0.045871  16.120  < 2e-16 ***
Ne9         -0.197005   0.078526  -2.509 0.013644 *
Ne12         0.399491   0.094649   4.221 5.19e-05 ***
Vx9         -0.313050   0.072857  -4.297 3.88e-05 ***
Vx12         0.165060   0.068940   2.394 0.018428 *
maxO3v       0.017655   0.005207   3.391 0.000983 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.259 on 105 degrees of freedom
Multiple R-squared:  0.8463,Adjusted R-squared:  0.8376
F-statistic: 96.38 on 6 and 105 DF,  p-value: < 2.2e-16


$all[[7]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2335 -0.5615  0.0164  0.7648  3.3568

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.393501   1.167710  -0.337   0.7368
T12          0.655579   0.094384   6.946 3.36e-10 ***
T15          0.084870   0.083460   1.017   0.3116
Ne9         -0.191862   0.078676  -2.439   0.0164 *
Ne12         0.399416   0.094634   4.221 5.22e-05 ***
```

```
Vx9          -0.316709   0.072934  -4.342 3.28e-05 ***
Vx12          0.162263   0.068984   2.352   0.0205 *
maxO3v        0.016998   0.005246   3.240   0.0016 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.258 on 104 degrees of freedom
Multiple R-squared:  0.8478,Adjusted R-squared:  0.8376
F-statistic: 82.79 on 7 and 104 DF,  p-value: < 2.2e-16


$all[[8]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2054 -0.5784  0.0481  0.7032  3.6314

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.520908   1.180601  -0.441  0.65998
T12          0.611437   0.109543   5.582 1.94e-07 ***
T15          0.127393   0.099145   1.285  0.20170
Ne9         -0.186352   0.079116  -2.355  0.02039 *
Ne12         0.349141   0.113827   3.067  0.00276 **
Ne15         0.069739   0.087396   0.798  0.42673
Vx9         -0.318197   0.073086  -4.354 3.16e-05 ***
Vx12         0.166258   0.069286   2.400  0.01821 *
maxO3v       0.017090   0.005256   3.251  0.00155 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.261 on 103 degrees of freedom
Multiple R-squared:  0.8488,Adjusted R-squared:  0.837
F-statistic: 72.27 on 8 and 103 DF,  p-value: < 2.2e-16


$all[[9]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1337 -0.5553  0.0687  0.6595  3.5438
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.485494   1.184569  -0.410  0.68278
T12          0.618294   0.110243   5.608 1.76e-07 ***
T15          0.119410   0.100033   1.194  0.23536
Ne9         -0.195090   0.080274  -2.430  0.01683 *
Ne12         0.359122   0.114983   3.123  0.00233 **
Ne15         0.063767   0.088021   0.724  0.47045
Vx9         -0.310013   0.074181  -4.179 6.19e-05 ***
Vx12         0.207188   0.090567   2.288  0.02422 *
Vx15        -0.056608   0.080385  -0.704  0.48291
maxO3v       0.017085   0.005269   3.242  0.00160 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.264 on 102 degrees of freedom
Multiple R-squared:  0.8495,Adjusted R-squared:  0.8362
F-statistic: 63.98 on 9 and 102 DF,  p-value: < 2.2e-16


$all[[10]]

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1322 -0.5580  0.0682  0.6596  3.5428

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4836718  1.1952902  -0.405  0.68659
maxO3       -0.0001487  0.0087997  -0.017  0.98655
T12          0.6186221  0.1124802   5.500 2.88e-07 ***
T15          0.1194931  0.1006460   1.187  0.23791
Ne9         -0.1954145  0.0829312  -2.356  0.02039 *
Ne12         0.3590586  0.1156124   3.106  0.00246 **
Ne15         0.0637939  0.0884697   0.721  0.47253
Vx9         -0.3098711  0.0750186  -4.131 7.47e-05 ***
Vx12         0.2071924  0.0910143   2.276  0.02493 *
Vx15        -0.0565454  0.0808666  -0.699  0.48601
maxO3v       0.0171376  0.0061332   2.794  0.00623 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.27 on 101 degrees of freedom
Multiple R-squared:  0.8495,Adjusted R-squared:  0.8346
F-statistic: 57.02 on 10 and 101 DF,  p-value: < 2.2e-16
$summary
```

```
                               R2       Pvalue
Model with 1 variable   0.7796310 6.404709e-38
Model with 2 variables  0.8013090 5.632629e-39
Model with 3 variables  0.8150119 2.013196e-39
Model with 4 variables  0.8252782 1.319017e-39
Model with 5 variables  0.8379452 3.000466e-40
Model with 6 variables  0.8463344 2.075040e-40
Model with 7 variables  0.8478473 1.280292e-39
Model with 8 variables  0.8487821 8.854028e-39
Model with 9 variables  0.8495137 6.107318e-38
Model with 10 variables 0.8495142 5.043277e-37
$best

Call:
lm(formula = as.formula(as.character(formul)), data = don)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2768 -0.6155  0.0211  0.7290  3.3712

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.304508   1.164615  -0.261 0.794245
T12          0.739464   0.045871  16.120  < 2e-16 ***
Ne9         -0.197005   0.078526  -2.509 0.013644 *
Ne12         0.399491   0.094649   4.221 5.19e-05 ***
Vx9         -0.313050   0.072857  -4.297 3.88e-05 ***
Vx12         0.165060   0.068940   2.394 0.018428 *
maxO3v       0.017655   0.005207   3.391 0.000983 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.259 on 105 degrees of freedom
Multiple R-squared:  0.8463,Adjusted R-squared:  0.8376
F-statistic: 96.38 on 6 and 105 DF,  p-value: < 2.2e-16
```

Méthode stepwise avec le critère AIC

```
>step(lm(maxO3~1,data=ozone),maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v,
direction="both")
Start:  AIC=748.9
maxO3 ~ 1

         Df Sum of Sq   RSS    AIC
+ T12     1     54244 33948 643.98
+ T15     1     52911 35280 648.29
+ T9      1     43138 45053 675.68
+ maxO3v  1     41323 46868 680.10
```

```
+ Ne12    1      36208 51984 691.70
+ Ne9     1      34088 54104 696.18
+ Vx9     1      24551 63640 714.36
+ Ne15    1      20176 68016 721.81
+ Vx12    1      16367 71825 727.91
+ Vx15    1      13545 74647 732.23
<none>                 88192 748.90

Step:  AIC=643.98
maxO3 ~ T12

         Df Sum of Sq   RSS    AIC
+ maxO3v  1       7600 26348 617.59
+ Vx9     1       3919 30029 632.24
+ Ne9     1       3579 30369 633.50
+ Vx12    1       3368 30580 634.28
+ Vx15    1       3070 30878 635.36
+ Ne12    1       2367 31581 637.88
+ Ne15    1       1581 32366 640.63
+ T15     1        890 33058 643.00
<none>                 33948 643.98
+ T9      1         19 33929 645.91
- T12     1      54244 88192 748.90

Step:  AIC=617.59
maxO3 ~ T12 + maxO3v

         Df Sum of Sq   RSS    AIC
+ Ne9     1     4474.5 21874 598.75
+ Vx12    1     2793.4 23555 607.04
+ Vx9     1     2663.0 23685 607.66
+ Vx15    1     2638.9 23709 607.77
+ Ne12    1     2508.0 23840 608.39
+ Ne15    1     1147.7 25200 614.61
<none>                 26348 617.59
+ T15     1      350.0 25998 618.10
+ T9      1      225.2 26123 618.63
- maxO3v  1     7599.8 33948 643.98
- T12     1    20520.3 46868 680.10

Step:  AIC=598.75
maxO3 ~ T12 + maxO3v + Ne9

         Df Sum of Sq   RSS    AIC
+ Vx9     1      903.4 20970 596.02
+ Vx12    1      630.7 21243 597.47
+ Vx15    1      611.0 21263 597.58
<none>                 21874 598.75
```

```
+ T9      1     108.8 21765 600.19
+ T15     1      90.7 21783 600.28
+ Ne15    1      61.1 21812 600.43
+ Ne12    1      60.6 21813 600.44
- Ne9     1    4474.5 26348 617.59
- T12     1    7140.0 29014 628.39
- maxO3v  1    8495.5 30369 633.50

Step:  AIC=596.02
maxO3 ~ T12 + maxO3v + Ne9 + Vx9

         Df Sum of Sq   RSS    AIC
<none>                20970 596.02
+ Vx15    1      74.1 20896 597.63
+ Vx12    1      45.3 20925 597.78
+ T15     1      42.1 20928 597.80
+ Ne12    1      19.0 20951 597.92
+ Ne15    1      16.4 20954 597.94
+ T9      1       0.0 20970 598.02
- Vx9     1     903.4 21874 598.75
- Ne9     1    2714.8 23685 607.66
- T12     1    6650.4 27621 624.88
- maxO3v  1    7363.5 28334 627.73

Call:
lm(formula = maxO3 ~ T12 + maxO3v + Ne9 + Vx9, data = ozone)

Coefficients:
(Intercept)          T12       maxO3v          Ne9          Vx9
    12.6313       2.7641       0.3548      -2.5154       1.2929
```

Méthode ascendante par la fonction add1

```
> lm1=lm(maxO3~1,data=ozone)
> lm1

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

Coefficients:
(Intercept)
       90.3

> add1(lm1,maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v,test="F")
Single term additions

Model:
```

```
maxO3 ~ 1
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>              88192 748.90
T9      1     43138 45053 675.68 105.324 < 2.2e-16 ***
T12     1     54244 33948 643.98 175.764 < 2.2e-16 ***
T15     1     52911 35280 648.29 164.972 < 2.2e-16 ***
Ne9     1     34088 54104 696.18  69.304 2.570e-13 ***
Ne12    1     36208 51984 691.70  76.618 2.769e-14 ***
Ne15    1     20176 68016 721.81  32.630 9.624e-08 ***
Vx9     1     24551 63640 714.36  42.436 2.265e-09 ***
Vx12    1     16367 71825 727.91  25.066 2.123e-06 ***
Vx15    1     13545 74647 732.23  19.960 1.927e-05 ***
maxO3v  1     41323 46868 680.10  96.986 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #La variable la plus significative est T12
> lm2=lm(maxO3~T12,data=ozone)
> add1(lm2,maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v,test="F")
Single term additions

Model:
maxO3 ~ T12
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>              33948 643.98
T9      1      19.1 33929 645.91  0.0614 0.8048168
T15     1     889.9 33058 643.00  2.9342 0.0895638 .
Ne9     1    3578.7 30369 633.50 12.8447 0.0005076 ***
Ne12    1    2366.9 31581 637.88  8.1691 0.0051057 **
Ne15    1    1581.4 32366 640.63  5.3258 0.0229005 *
Vx9     1    3919.1 30029 632.24 14.2256 0.0002639 ***
Vx12    1    3367.5 30580 634.28 12.0031 0.0007604 ***
Vx15    1    3070.1 30878 635.36 10.8377 0.0013410 **
maxO3v  1    7599.8 26348 617.59 31.4397 1.571e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #la variable la plus significative est maxO3v
> lm3=lm(maxO3~T12+maxO3v,data=ozone)
> add1(lm3,maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v,test="F")
Single term additions

Model:
maxO3 ~ T12 + maxO3v
       Df Sum of Sq   RSS    AIC F value    Pr(>F)
<none>              26348 617.59
T9      1     225.2 26123 618.63  0.9309 0.3367776
T15     1     350.0 25998 618.10  1.4540 0.2305228
Ne9     1    4474.5 21874 598.75 22.0925 7.708e-06 ***
Ne12    1    2508.0 23840 608.39 11.3618 0.0010408 **
```

```
Ne15    1    1147.7 25200 614.61  4.9185 0.0286639 *
Vx9     1    2663.0 23685 607.66 12.1430 0.0007131 ***
Vx12    1    2793.4 23555 607.04 12.8082 0.0005183 ***
Vx15    1    2638.9 23709 607.77 12.0205 0.0007564 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #la variable la plus significative est Ne9
> lm4=lm(maxO3~T12+maxO3v+Ne9,data=ozone)
> add1(lm4,maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v,test="F")
Single term additions

Model:
maxO3 ~ T12 + maxO3v + Ne9
       Df Sum of Sq   RSS    AIC F value  Pr(>F)
<none>              21874 598.75
T9      1    108.80 21765 600.19  0.5349 0.46617
T15     1     90.67 21783 600.28  0.4454 0.50597
Ne12    1     60.63 21813 600.44  0.2974 0.58664
Ne15    1     61.12 21812 600.43  0.2998 0.58512
Vx9     1    903.37 20970 596.02  4.6094 0.03405 *
Vx12    1    630.70 21243 597.47  3.1768 0.07753 .
Vx15    1    610.97 21263 597.58  3.0746 0.08239 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #la variable la plus significative est Vx9
> lm5=lm(maxO3~T12+maxO3v+Ne9+Vx9,data=ozone)
> add1(lm5,maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v,test="F")
Single term additions
Model:
maxO3 ~ T12 + maxO3v + Ne9 + Vx9
       Df Sum of Sq   RSS    AIC F value Pr(>F)
<none>              20970 596.02
T9      1     0.018 20970 598.02  0.0001 0.9924
T15     1    42.082 20928 597.80  0.2131 0.6453
Ne12    1    18.999 20951 597.92  0.0961 0.7571
Ne15    1    16.429 20954 597.94  0.0831 0.7737
Vx12    1    45.320 20925 597.78  0.2296 0.6328
Vx15    1    74.125 20896 597.63  0.3760 0.5411
> #toutes les p-value sont très élevés donc on arrête la procédure
```

Qualité de la régression choisie avec add1

```
 > summary(lm5)
Call:
lm(formula = maxO3 ~ T12 + maxO3v + Ne9 + Vx9, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-52.396  -8.377  -1.086   7.951  40.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.63131   11.00088   1.148 0.253443
T12          2.76409    0.47450   5.825 6.07e-08 ***
maxO3v       0.35483    0.05789   6.130 1.50e-08 ***
Ne9         -2.51540    0.67585  -3.722 0.000317 ***
Vx9          1.29286    0.60218   2.147 0.034055 *
---
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
```

ANNEXE

C

ANALYSE DE LA VARIANCE

Boxplot de l'influence du vent sur la concentration en ozone

```
> png(file="boxplot.png")
> boxplot(ozone$maxO3~ozone$vent,col=c("aquamarine","grey","white","coral")
+)
> dev.off()
quartz
     2
```

Boxplot de l'influence de la pluie sur la concentration en ozone

```
> png(file="pluie.png")
> boxplot(maxO3~pluie,data=ozone,col=c("coral","grey"))
> dev.off()
quartz
     2
```

BIBLIOGRAPHIE

[1] A. BERRED. Statistique inférentielle, Cours de statistique inférentielle 2021-2022.

[2] C. Chouquet. Modèles linéaires, Cours de l'univérsité de Paul Sabatier (Toulouse), Année 2009-2010. Consulté le 31.05.2022 sur https ://www.math.univ-toulouse.fr/∼barthe/M1modlin/poly.pdf.

[3] P.A. Cornillon and E. Matzner-Lober. *Régression avec R*. Pratique R. Springer Paris, 2012.

[4] JJ. DAUDIN. Introduction a la régression multiple. Consulté le 25.03.2022 sur https ://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-l-inf-intRegmult.pdf.

[5] Ibrahima. DIARRASSOUBA. Analyse et fouille de données, Cours de M1-Université le Havre Normandie.

[6] M. ETIENNE. Le modèle linéaire et ses extensions, 14 Septembre 2016. Consulté le 25.04.2022 sur http ://moulon.inra.fr/modelstat/supports/ModeleLineaireEt Extensions-compressed.pdf.

[7] Antoine. MASSÉ. Aide à l'utilisation de r, Courbes multiples et régressions. Consulté le 05.05.2022 sur https ://www.overleaf.com/project/621dddcc3d5534ec9777504c.

[8] Sylviane. WEY. Méthodes de selections de variables, Td de licence L3 SV,2020-Université le Havre Normandie.